Deep Supervised Attention Networks for Pixel-wise Brain Tumour Segmentation

Wai Po Kevin Teng, Xiongjun Wang

Team project report for
Digital Engineering

Data and Knowledge Engineering Group
Faculty of Computer Science
Otto von Guericke University Magdeburg, Germany
23.01.2021

Deep Supervised Attention Networks for Pixel-wise Brain Tumour Segmentation

Wai Po Kevin Teng, Xiongjun Wang

Data & Knowledge Engineering Group, Faculty of Computer Science
Otto von Guericke University Magdeburg, Germany
{wai.teng, xiongjun.wang}@st.ovgu.de

Abstract. Glioblastoma (GBM) is one of the leading causes of cancer death. The imaging diagnostics are critical for all phases in the treatment of brain tumour. However, manually-checked output by a radiologist has several limitations such as tedious annotation, time consuming and subjective biases, which influence the outcome of a brain tumour affected region. Therefore, the development of an automatic segmentation framework has attracted lots of attention from both clinical and academic researchers. Recently, most state-of-the-art algorithms are derived from deep learning methodologies such as the U-net, attention mechanism and capsule networks. In this paper, we propose multiple model architecture that combines the method of deep supervision and attention mechanism for pixel-wise brain tumour segmentation. We believe that 2D network models are much easier to implement as compared to it's 3D counterpart. Our work aims to exploit the potential of 2D networks as opposed to 3D networks architecture. To validate our work, we set UNet as our baseline model and proposed two novel 2D network architectures as well as one 3D network architecture. Our first proposed model Deep supervised Attention Unet(DAUNet), extends the infamous UNet framework with the addition of attention gates in the skip connection path and deep supervision in the upsampling path. Our second proposed model, multi-scale Self Guided Attention Network(SGANet), attempts to compensate the lack of multi-scale features in the UNet framework by incorporate guided self-attention mechanism and deep supervision for multi-scale features. Our third proposed model, (3D-DAUNet), further the work of our first proposed model(DAUNet), by extending a dimension with 3D convolutional layers. Subsequently, we are able to achieve a low resolution and high resolution feature representations even for small tumour regions for both of our proposed 2D models, namely DAUNet and SGANet. Preliminary results of our proposed models have shown promising results that outperformed our baseline model on mean dice coefficient with training data and validation data, particularly, with our 2D models performing the best. Specifically with DAUNet achieving [0.6736, 0.8608, 0.7042], follow suit, SGANet achieve [0.6568, 0.8717, 0.7045] for Enhancing Tumour (ET), Whole Tumour (WT), and Tumour core (TC) respectively.

Keywords: Brain tumour, Attention network, U-Net, BraTS2020, Multi-scale features, Deep supervision

1 Introduction

Primary malignant brain tumours are among the most dreadful types of cancer, not only because of the dismal prognosis, but also due to the direct consequences on decreased cognitive function and poor quality of life [13]. The most common primary brain tumours in adults are gliomas and meningiomas. For gliomas, the incidence is 6 to 8 in 100,000, with approximately 50% belonging to malignant sub-types. Low grade gliomas (LGG) tend to occur in younger patients, whereas high grade gliomas (HGG) are more frequent in older patients [1]. Grading is performed according to the World Health Organization (WHO) criteria, taking into account the presence of nuclear changes, mitotic activity, endothelial proliferation, and necrosis [35]. In fact, approximately 50% of all gliomas are glioblastoma, which corresponding to WHO grade IV, is the most fatal and most common primary brain neoplasm with an incidence of three to four in 100,000 [18]. The Public Health England estimates the median survival period as 6 months from diagnosis without treatment [8]; however, timely diagnosis and reasonable treatment can effectively help patients increase the probability of survival and prolong the survival time. Imaging is being used to determine the localization, extend, type, and malignancy of the tumour. It provides a lot of important information for diagnosis and planning of treatment. After treatment, imaging is being used to quantify the treatment response and the extent of residual tumour. At follow-up, imaging helps to determine tumour progression and to differentiate recurrent tumour growth from treatment-induced tissue changes [1].Therefore, the imaging diagnostics are critical for all phases in the treatment of a brain tumour. However, manually identifying the brain tumour region by a radiologist has several drawbacks such as laborious annotation, time consuming and subjective uncertainties, so it indicates that brain tumour segmentation are susceptible to human errors. With the advancement of contemporary technologies, the development of an accurate automatic segmentation framework is viewed as the holy grail among domain experts as well as academic researchers. Emerging machine learning methods, especially deep learning, induced more accurate and reliable solutions to increase clinical workflow efficiency, whilst support decision making [43].

Although automatic brain tumour segmentation is attractive, the development of algorithms that are capable of handling segmentation tasks remains challenging, due to the irregular nature of brain tumour. Brain tumour Segmentation (BraTS) Challenge has been released by the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) since 2004, which has always been focusing on the evaluation of state-of-the-art methods for the segmentation of brain tumours in multimodal magnetic resonance imaging (MRI) scans. Convolutional neural network (CNN) was proposed to solve various computer vision tasks, proving its capability and accuracy without compromising performance. CNN has successfully been the de facto state of the art performers in many applications related to images [57]. On the other hand, the availability of a unique dataset of MRI scans of low- and high-grade glioma patients with repetitive manual tumour delineations by several human

experts are scarce. BraTS challenge dataset is based on ample multi-institutional routine clinically-acquired pre-operative multi-modal MRI scans of glioblastoma (GBM/HGG) and lower-grade glioma (LGG) [3] [2] (see Fig.1). As compared to the previous dataset, the BraTS2020 dataset has more routine clinically-acquired 3T multi-modal MRI scans, with accompanying ground truth labels by expert board-certified neurologists [5]. The main task of the BraTS2020 challenge is to develop an automatic method and produce segmentation labels of the different glioma sub-regions with the usage of provided clinically-acquired training data [3] [2].

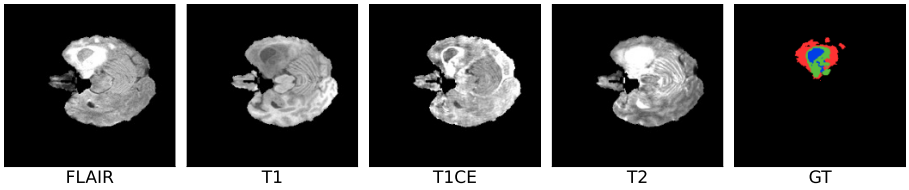


Fig. 1. Multimodal imaging with Flair, t1, t1ce, t2 and annotated ground truth from BraTS2020 training dataset *(from left to right)*. The segmentation are combined to generate the final labels of the tumour sub-regions, each color depicts *Peritumoural Edema(red), Necrotic* and *Non-enhancing tumour Core(blue), Necrotic* and *GD-Enhancing tumour(green)* and*Background(black)*, the main task of this challenge is to develop a method to segment labels of the different glioma sub-regions.

The various methods of automatic image segmentation have been proposed by researchers all over the world [5]. As early solution of image segmentation task, CNN and U-net are the most popular framework structures[48]. For example, Chen et al. [9] proposed an auto-context version of the VoxResNet by combining low-level features, implicit shape information, and high-level context together, which achieved the first place for the 2018 BraTS challenge. On the other hand, Feng et al. [14] developed a 3D U-Net with adaptations in the training and testing strategies, network structures, and model parameters for brain tumour segmentation. Lee et al. [29] proposed a patch-wise U-Net architecture. In this method, the model is used to overcome the drawbacks of conventional U-net with more retention of local information. Subsequently, attention network was another highlighted framework in recent research papers. For example, Oktay et al. [43] proposed a novel attention gate to focus on target structures of varying shapes and sizes. Models trained with attention gate implicitly learn to suppress irrelevant regions in an input image while highlighting salient features. Noori at al. [42] has designed a low-parameter network based on 2D U-Net in which employs an attention mechanism. This technique prevented redundancy for the model by weighting each of the channels adaptively. As opposed to the redundancy from the cascading behavior of U-net for the retrieval of multi-scale features resulting from the input dataset, Sinha and Dolz [49] proposed

multi-scale self guided attention network that provide further insights for the interpretability of the feature maps during the segmentation process.

Our goal were set to evaluate the capability of 2D network structure over 3D network structure without trading off much computational resources for 3D models yet being able to retain substantial accuracy given an image segmentation task. To validate this credibility of a 2D network performance with respect to a 3D model, we extend our best proposed model as shown in Fig. 3 in a 3D form in order to compare the trade off. Therefore the roadmap of our team project is as following, the goal was to enhance the performance of our benchmark, U-net by extending the structure of the model with attention and deep supervised mechanism. To break off the dominance of U-net in an image segmentation task, we incorporate grid-based attention mechanism by [43] in the novel network structure proposed by Sinha and Dolz [49]. Next, we extended our 2D network to 3D network to explain the credibility of the network without compromises.

2 Related Works

2.1 Deep Learning Applications in Medical Image Segmentation

Medical image segmentation has always serve as a crucial task in the realm of healthcare in the wake to pinpoint region of interest of a patient to carry out diagnoses. However, image segmentation is not a trivial task even for domain experts and often prone to human errors, laborious working hours and misinterpretation during the annotation phase of an medical image. Conventional methods for medical image segmentation in computer vision such as watershed transform [47], graph cut [7][6], etc., required feature engineering and prior knowledge of the given data to achieve formidable results. This however does not bridge the gap between domain experts in the realm of healthcare and computational simulation respectively, where the lack of prior knowledge would hinder the advancement of feature extraction.

To tackle this issue, the introduction of end-to-end learning with the implementation of deep neural network, also known as deep learning has, alleviate the need for feature engineering without compromising accuracy of image segmentation. It has been known that artificial neural networks (ANNs) are very flexible, able to model and solve complicated problems. Subsequently, convolutional neural network (CNN) is a particular kind of ANN aimed at preserving spatial relationships in the data, with very few connections between the layers [37]. Convolution layer is the layer to extract features from an input image. Convolution of an image with different filters can perform operations such as edge detection, blur and sharpen by applying filters [45]. A CNN has multiple layers of convolutions and activations, often interspersed with pooling layers, and is trained using backpropagation and gradient descent as for standard artificial neural networks [37]. Simultaneously, with the introduction of convolutional neural network(CNN), CNN has been the fundamental building block for computer vision task. Extending the effectiveness off CNN, most medical image segmentation task has been inspired by fully convolutional neural network(FCN) by Long

et al. [33], where the input data are encoded with multiple CNN layers and aimed to construct the segmentation from the encoded latent space. The improvement of FCN model structure has achieved sufficiently good performance with the introduction of UNet by Ronneberger et al. [48]. Similar to FCN model structure, UNet consisted of an encoding path which contract the input image into latent space and a decoding path which expand the feature maps into the same size as the input image. This way, when "reconstructing" the mask of the image, the network learns to use these features, because the features of the contracting path are concatenated with those of the expanding path[15]. Subsequently, UNet has been the go to benchmark for medical image segmentation due to its computation efficiency and ease of implementation.

To cope up with the nature of medical imaging being volumetric, UNet could be easily extended to accustom volumetric input by replacing the conventional 2D CNN layers into 3D CNN layers. In the work by Miletari et al. [40], the authors proposed VNet for volumetric medical image segmentation. On the other hand, Isensee et al. [24] extended the application of VNet by injecting gradient signals to the layers on the expanding path in the network through deep supervision.

2.2 Attention Mechanism in Medical Image Segmentation

Inspired by human visual biological traits, human beings engage selective focal on salient parts associated with a serial of partial glimpses to visualize better structure of an object [54]. This is because human process data by actively shifting their focus in order to remember specific, related events in the past. Such attention mechanism allows one to follow one thought at a time while suppressing information irrelevant to the task. Therefore, in the context of deep learning, attention mechanism acts as a filter to retain only relevant salient features locally and globally in the network model, which enhance the performance of the network in long-range dependencies. Neural attention models work by receiving input and produces output as usual but it also produces an extra set of outputs used to parameterise an attention model. Referred by Luong et al. [38] in their paper and described by Xu et al. [56]. In their paper, there are enforced by design categorized as hard- and soft-attention: soft attention is when we calculate the context vector as a weighted sum of the encoder hidden states. Hard attention is when, instead of weighted average of all hidden states, we use attention scores to select a single hidden state. Graves et al. [17] stated that, the attention model then operates on some extra data, e.g. image, audio or text, to create a fixed-size glimpse that is passed to the network as an extra input at the next time step. In computer version, attention mechanisms are applied to a variety of problems, including image classifiation [26][53], image segmentation[46], image captioning [36][58] and visual question answering [41][58]. A variant of attention mechanism, self-attention [44][51], enhance computational efficiency of attention model for long-range dependencies via computation of multiple attention heads in parallel. Hu et al.[23] used a channel-wised attention to highlight important feature

dimensions, which was the top-performance in the ILSVRC 2017 image classification challenge. Similarly, non-local self-attention was used by Wang et al.[52] to capture long-range dependencies. Via the implementation of self-attention, the attention heads response to each position by taking the weighted average of a latent space and attend to all positions.

To our knowledge, attention mechanism has been widely applicable in the field of computer vision, but it is less used in the context of medical imaging. Oktay et al. [43] extend the attention mechanism on UNet by incorporating attention gates before concatenating feature maps from the contracting path to the expanding path for pancreas image segmentation. The attention gates serve as a filter to attend towards salient features relevant to the segmentation task. In the work by [59], Zhao et al. proposed spatial-channel attention with Unet for gland segmentation. Spatial attention focus on "where" the informative part of the features is located by extracting spatial features such as contour, edge, etc. Whilst channel attention tend to focus on the "what" aspect of the salient feature by exploiting global information via feature re-calibration. Sinha and Dolz [49] argue that cascading structure, i.e., encoder-decoder architectures (e.g. UNet), would eventually led to the redundant use of information due to the repetitive extraction of similar low-level features at multiple scales. Sinha and Dolz [49] proposed guided self-attention mechanism by adaptively integrate local features with respect to their global dependencies in attempt to provide more semantics.

3 Method

3.1 UNet

In our work, we decide to include conventional image segmentation network architecture, Unet [48] as our baseline model for the comparison with respect to our proposed models. UNet basically is an autoencoder that is used to make sure the latent-representation of the feature map is not completely change from on step to the other. As the latent space is lower-dimensional, only the key information will be extracted. We do not want this information to be changed from one refinement step to the next, we only want small adjustments to be made. There won't be seen in the latent representation. As shown in Fig. 2, the network structure is straight forward with an encoder-decoder architecture, where the input data are downsampled in a cascading fashion, with two 64, 128, 256, 512 and 1024 feature maps on each contracting path respectively. The input images are downscaled by a factor for 2 for each maxpooling layers, where the bottle-neck layer consisted is made up of a dimension of [B,15,15,1024], where B denotes the batch size of an input data. Similarly, in order to perform image segmentation, the latent space is reconstructed to the original image size, i.e. [240,240] for the width and height, via the expanding path. Simultaneously, for each upsampling path, the feature maps are concatenated with the feature maps from the downsampling path on it's symmetrical counterpart via skip connection. This is because there might be information loss from the early stage of the data

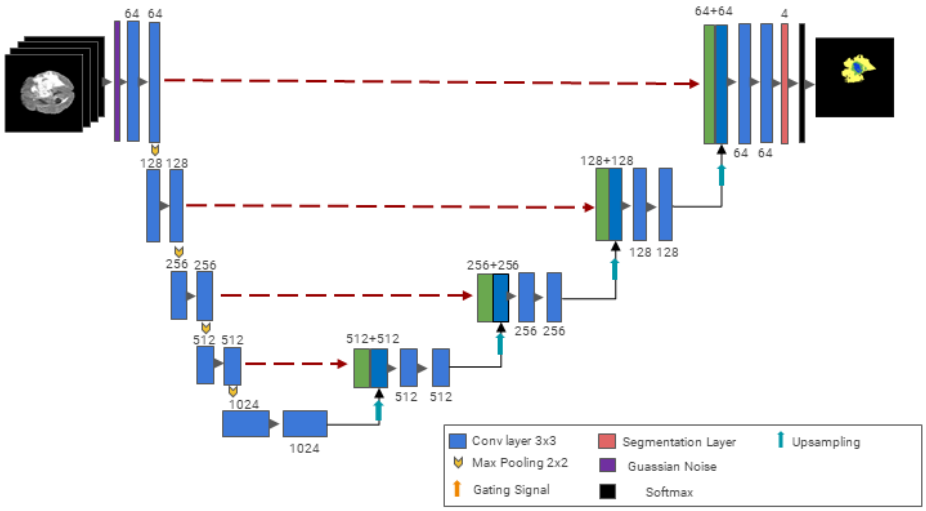


Fig. 2. UNet [48] as benchmark network architecture, with number on the layers indicating filter size.

during the learning process and skip connection would enhance the gradient signal to the feature maps in the earlier stage for better learning.

3.2 Deep Supervised Attention UNet

Overview: In this paper, inspired by [25][28], we proposed a deep supervised U-Attention Net framework for multi-label pixel-wise brain tumour segmentation(Fig. 3). In addition to model structure and parameters modification for U-net, we implemented an attention gate prior to concatenating features from skip connection. Follows, a multistage segmentation layer was added to summarize features element-wise during the upsampling path for network output. Generally, this framework includes three main parts: to speed up computation with equivalent performance of the network in terms of memory, the U-net was chosen as a backbone network structure that learns the high and low-level features, as well as, superposition of multiscale feature maps during upsampling path to enhance gradient signal. However, it has been considered a challenge to reduce false-positive prediction for small objects that show large shape morphology. Conventionally, concatenate operation was implemented with skip-connection directly passing the information from the encoder to the decoder at the same level. Concatenation of these different level feature maps without emphasizing important features brings about redundancies that hinder low-level feature extractions. This may promote errors during the model prediction, leading to wrong segmentation of tumours in the pixel space.

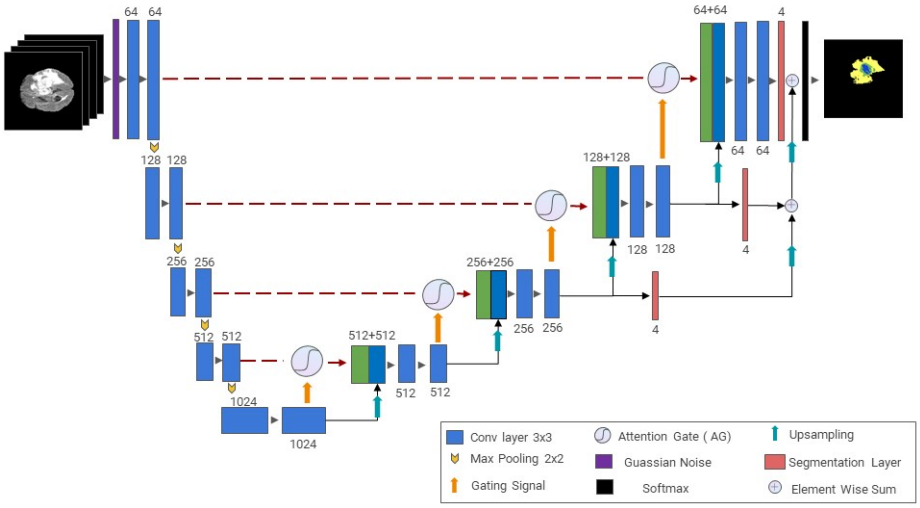


Fig. 3. Proposed Network architecture. Our architecture is inspired by [42] [25][43]. An attention mechanism based 2D U-net integrated with multistage segmentation layer features through deep supervision as proposed network structure. The context pathway aggregates high level information that is subsequently localized precisely in the localization pathway (right). The low level features was filtered by a attention gate before concatenation with the low high features, the segmentation layer combines the multistage feature maps by a element wise summation for final network output as well as propagation.

Grid-based attention gate: Inspired by [43], attention gates was introduced before concatenate operation as a filter to suppress low-level feature extractions in irrelevant regions. Attention gated signal was obtained by element wise multiplication between the input signal and grid-based attention coefficient, which learns to focus on wanted region during training phase. The attention network performance could learn to attend specific regions on a pixel level as shown in the architecture in Fig.4.

Furthermore, during the upsampling pathway, the segmentation layers enhanced the output of the previous convolutional layer feature maps created at a different stage. Element-wise summation of feature maps from different levels directly pass local details found in the third lowest resolution segmentation map and the second-lowest resolution segmentation map when upsampling, and summed with the final convolutional layer output. The method was inspired in [34][28], this multistage layer forwards the lowest resolution gradient signals to the network output and back-propagate the signals to the whole upsampling network, which allows the model to localize concurrent brain tumour regions precisely.

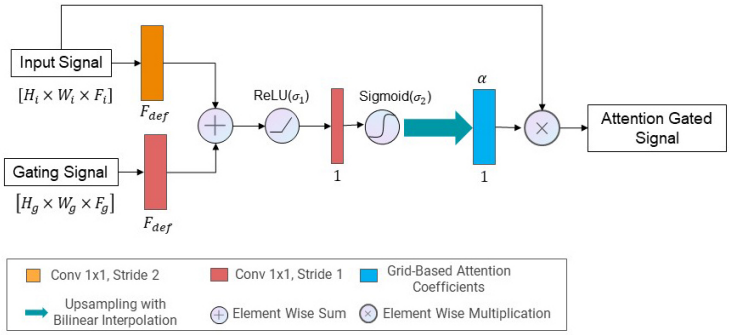


Fig. 4. Attention gate network architecture

Deep Supervision: Through above, the proposed model in this paper implemented a deep supervised U-Attention Net for pixel-wise brain tumour segmentation, which combines the U-net, grid-based attention network and a deep supervised multistage segmentation layer. In this way, we could achieve both the low resolution and high-resolution tumour regions feature representations.

Our proposed backbone network is based on U-Net architecture [48]. The U-Net structure approach allows the network to reintegrate the low-level and high-level features throughout the upsampling and downsampling pathway. Each block in the downsampling pathway has two convolutional layers with a kernel size of 3×3 followed by a max pooling layer with kernel size 2 and stride 2. The activation function is the ReLU(Rectified Linear Unit).

The numbers of filters in five blocks of downsampling pathway are 64, 128, 256, 512 and 1024 respectively, which is consistent with upsampling pathway. These blocks in the downsampling pathway shape will eventually downsample 240×240 input image into a 15×15 size feature maps. Consequently, the same parameters have been applied in the upsampling pathway except an extra convolutional layer, the network output of each deconvolutional layers was concatenated with the output of the attention gate signal from the same level downsampling pathway. In the end, the feature map (15×15) was reconstructed to the original image size(240×240).

During the upsampling process, in order to retain salient features from the lower upsampling layer, a convolutional layer after a deconvolution block was added, which maps the deconvolutional block output$(256 \times 60 \times 60)$ to a$(4 \times 60 \times 60)$ feature map in the third lowest upsampling layer. this feature map was element-wisely summed with the output of the second lowest upsampling convolutional layer$(4 \times 120 \times 120)$. In the last stage, the convolutional layer combines the second lowest upsampling convolutional layer to generate the final output$(4 \times 240 \times 240)$ which is then fed into softmax function for multi-label segmentation.

3.3 Multi-scale Self Guided Attention Network

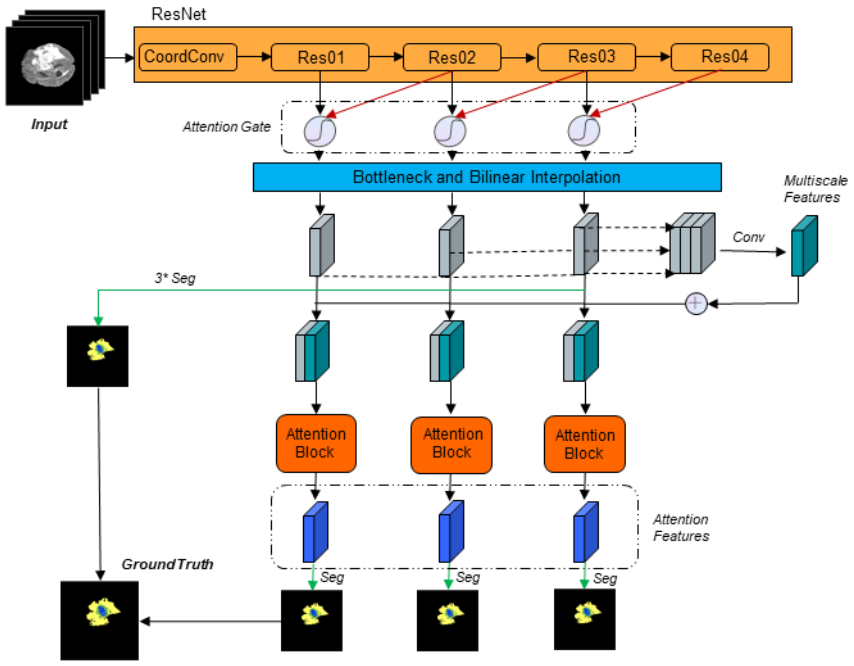


Fig. 5. Proposed Network architecture. Inspired by Sinha and Dolz [49], input images are passed through a backbone network, which in our case ResNet associated with CoordConv layer [31] for better spatial representation. Multi-scale features are then connected to attention gate [43] before inputting into attention block for image segmentation. Deep supervision is implemented to propagate the signal to the shallow layer in the middle part of the network structure.

Overview: In the paper by Sinha and Dolz [49], the authors point out the shortcomings of UNet. [49] mentioned that low-level features are repetitively at multiple scales rendering redundancies while learning features in long-range dependencies. They aimed to overcome this shortcomings in attempt to concatenate multi-scale features into feature maps before passing through guided self-attention mechanism in order to enrich contextual dependencies. Inspired by the attention network used in [16], the building blocks of guided self-attention mechanism consisted of position attention module(PAM) as well as channel attention module(CAM) as shown in Fig. 6. As opposed to the actual network structure in the paper by Sinha and Dolz [49], as shown in Fig. 5, we also incorporate attention gate [43] as explained in section 3.2 for each downscaled feature

maps from the encoding path with the intention to attend salient features from shallow layers. Multi-scale output from the attention gates are interpolated bilinearly to a common size and convolved to obtain a common multi-scale feature map. In this case, the multi-scale feature consisted of low-level detail encoded from early layers and high-level semantics encoded from deeper layers. Subsequently, the multi-scale feature is concatenated with the feature maps at different scale and input into guided attention modules with attention features generated for each scale respectively. A segmentation output is obtained for each attention features respectively for the loss function. In our work, we also incorporate deep supervision by injecting gradient signal from the segmented output derived from the feature maps of multi-scale extraction in the shallow layer as shown in the middle part of Fig. 5.

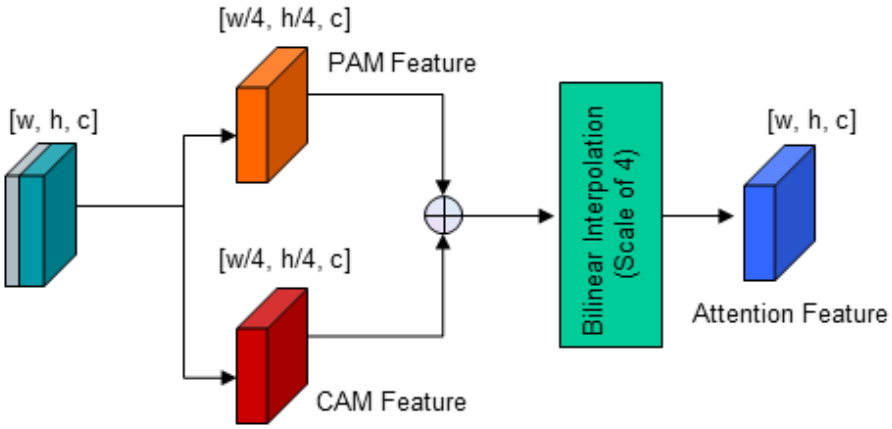


Fig. 6. Guided Self-Attention Mechanism that consisted PAM feature and CAM feature. where W: width, H: height, C: channel. Both PAM and CAM features width and height are downsampled by a scale of 4 due to limited computational resources. Attention feature consisted of concatenated features from PAM and CAM and upsampled to a scale of 4 before output as a segmentation.

Network Backbone: Similar to the implementation in the original paper by Sinha and Dolz [49], the choice for main network backbone in the proposed network architecture in ResNet [19] to retrieve dense local features. In our work, we adopted full pre-activation variant [22] in the order of passing through a normalization layer, next, activation layer and a convolutional layer for better performance. Our choice for normalization layer is group normalization [55] as opposed to batch normalization. As the name indicates, group normalization

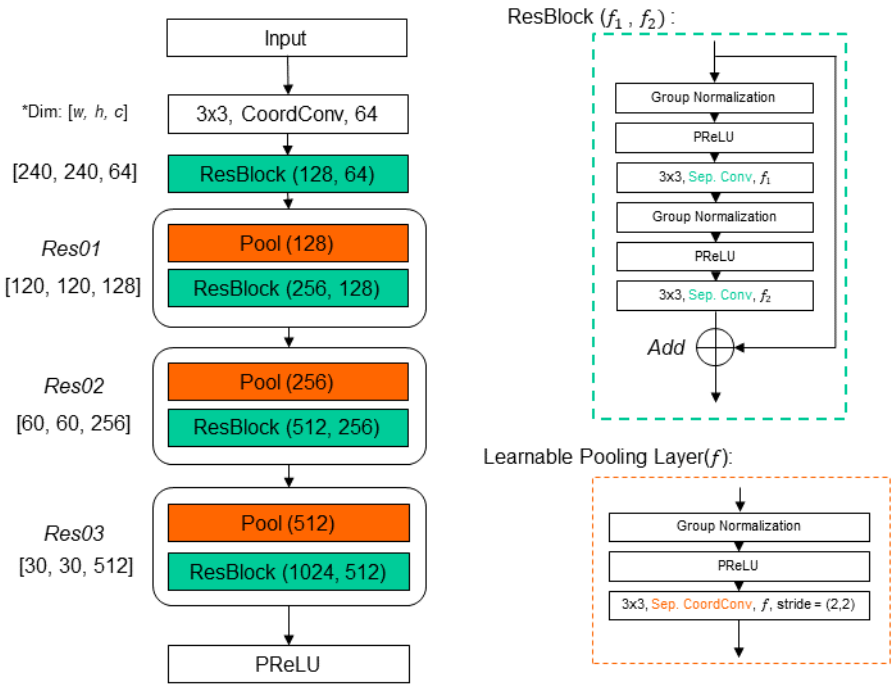


Fig. 7. The backbone of the proposed network architecture comprised of ResNet with 4 Resblock where each Resblock output feature maps of different sizes. To capture the essence of spatial representation in the feature maps, CoordConv [31] is utilized in our work. To speed up computation time, separable convolution [11] was proposed for implementation. Here, the number in the square brackets indicate the dimensionality of the feature maps for each Resblock output. Whereas, the number in the curvy brackets indicate the filter size for ResBlock.

does not relies on the normalization along the batch axis whereas normalize along the channel axis. He et al. [55] suggested that group normalization is much numerical stable for the learning task due to its independence of batch size. Different variant of normalization in the field of deep learning are depicted in Fig. 8.

To enhance the spatial representation of the network, inducing the essence of the pixel-wise segmentation in the pixel space, an additional CoordConv layer, as proposed by [31] solves this problem by supplying the convolution operation with its input coordinates. For this, two extra coordinate channels are concatenated channel-wise to the input data. Coordinate channel i is an $height \times width$ matrix, where elements carry the number of their row. For coordinate channel j, each element carries its column count. A schematic plot of CoordConv layer is shown in 9. For the activation function, we opt for Parametric Rectified Linear Unit(PReLU) [20] as opposed to the popular Rectified Linear Unit(ReLU). The

equation of PReLU are as shown in Eq. (1).

$$f(y_i) = \begin{cases} y_i & \text{if } y_i > 0 \\ a_i y_i & \text{if } y_i \leq 0 \end{cases} \tag{1}$$

Hence, from Eq. (1), we can deduced that PReLU is a kind of LeakyReLU, but a_i is a learnable parameter that does not provide user defined gradient for the function if the function is ≤ 0. The reason for introducing PReLU is to prevent dead function if the input value is ≤ 0. Furthermore, we also incorporate separable convolution in our ResNet as shown in the work by He et al.[11]. Separable convolution would drastically speed up the computational process of the network during the learning task due to the significant reduction of multiplication task. Simultaneously, we adopted learnable pooling layer in our proposed model architecture, rather than the conventional max-pooling layer for downsampling. We hypothesised that maxpooling might left out relevant information during the pooling process. Hence, separable convolution with a kernel size of 3x3 and a stride of 2 is implemented in replace of max-pooling.

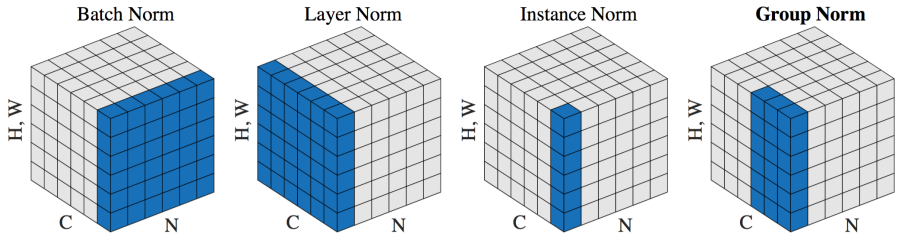


Fig. 8. Variant of normalization layer, with image source derived from the work by Wu and He[55]. Where, C: Channel, N: Number of batch, H: Height, W: Width, respectively.

Position Attention Module(PAM): PAM aims to attend towards spatial representation of features by focusing on 'where' is an informative part located. As depicted in the Fig. 10, the input features are downsampled twice with a learnable pooling layer, i.e. convolution layer with a kernel size of 1x1. This is because our GPU was not able to fit in input features of the original dimension and hence, empirically, downscaling the size of the input image would leverage the memory issue arise from the GPU. Similar to the work by Vaswani et al. [51], the downsampled features are split into multiple stream and most noticeably, the three main attention head, key, query and value head. In the query head, the features are subjected to a convolution layer and reshaped into 2D-array and permutated with the channel axis. Next, the permutated matrix conduct an element-wise multiplication with the convolved matrix from the key head before

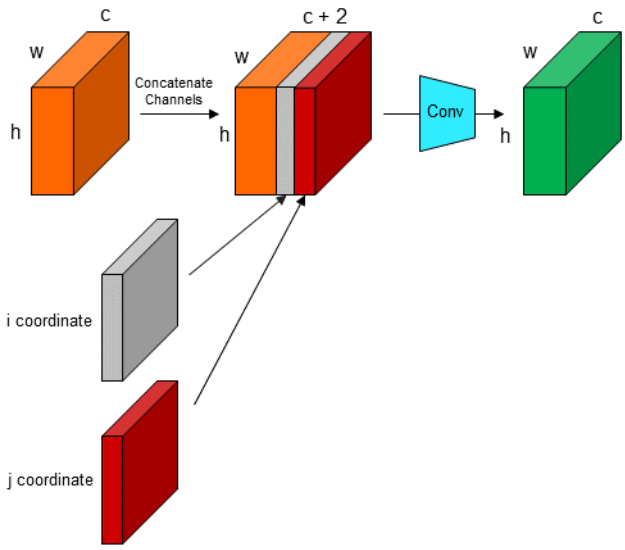


Fig. 9. A variation of convolutional layer, CoordConv [31] by concatenating i coordinates and j coordinates to the feature maps before convolving.

applying a softmax function. These operations would eventually yield a similarity score, denoted as α_p where similar features have higher similarity. The grid-like similarity scores are multiplied with the convolved reshaped feature maps in the value head stream and reshaped to tensor resembling the dimension of the downsampled feature maps [w/4,h/4,c]. After the value head stream, there is a stream where feature maps are pass through a convolution layer of 1x1 with and sigmoid layer which acts as a learnable coefficient, denoted as resembling the dimension of the downsampled feature maps [w/4,h/4,c]. for the PAM. λ_p is multiplied with the feature maps from the Key, Query and Value head and an addition operation is conducted to amplified relevant features.

Channel Attention module(CAM): CAM aims to exploit inter-dependencies betweem features along the channel axes which is size independent and focus on 'what' is an informative part located. Similar to PAM as shown in Fig. 10, CAM also consisted of three main attention heads (Key, Query and Value heads) as depicted in 11. However, unlike PAM, there is no convolution layers in the heads. Instead, the only operations existed in the heads are only reshaping and permutating. The output of key and query head are multiplied element-wise to before undergoing a softmax layer to obtain a similarity score, denoted as α_c.

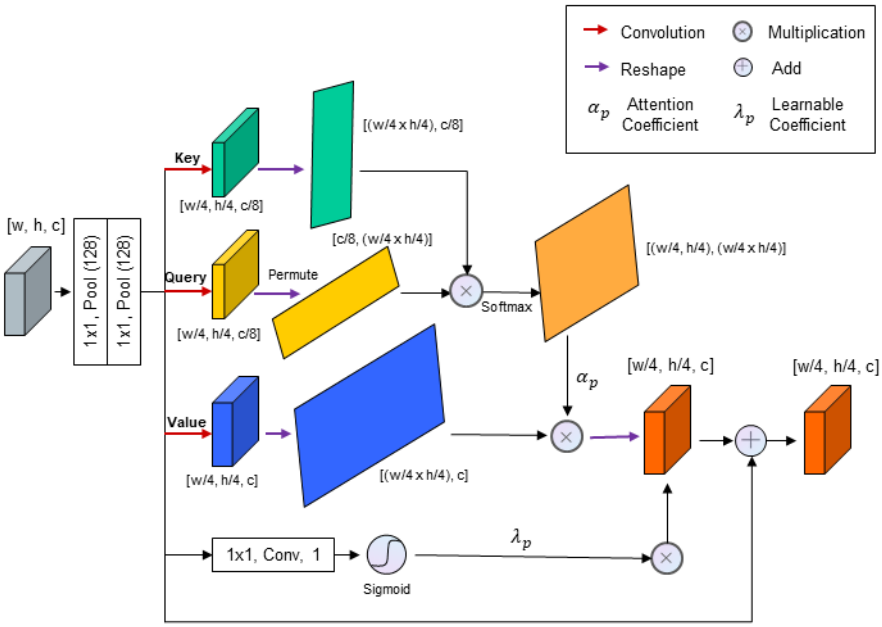


Fig. 10. Schematic plot of Position Attention Module(PAM) for the proposed network structure. Where w: Width, h: Height, c: Channel.

The similarity score is then multiplied with the matrix from the value head and and reshaped resembling the dimension of the downsampled feature maps [w/4,h/4,c]. Up until this stage, there is no learnable parameters in the module of CAM. Inspired by Noori et al. [42], squeeze excitation block was introduced as shown in Fig. 12 to output attention coefficient channel wise, denoted as α_c. In squeeze excitation block, the input features is subjected to Global Average Pooling(GAP) to summarise the features channel-wise and are connected to a fully connected layer with ReLU as activation function and pass through another fully connected layer with sigmoid as activation function. As a result, the sigmoid output weights for each channel of the feature maps and its multiplied with the input features as an output.

3.4 3D Deep Supervised Attention UNet

As the BraST2020 images are all volumetric data, which is abundant in biomedical data analysis. Annotation of such data with segmentation labels causes difficulties, since only 2D slices can be shown on a computer screen. Thus, annotation large volumes in a slice-by-slice manner are tedious. It is inefficient, too, since neighboring slices show almost the same information. Especially for

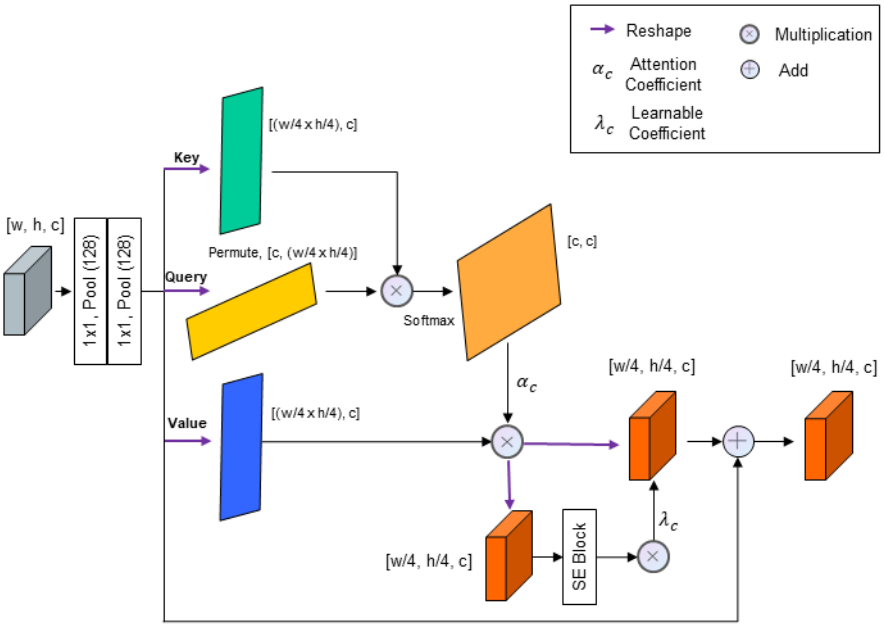


Fig. 11. Schematic plot of Channel Attention Module(CAM) for the proposed network structure. Where w: Width, h: Height, c: Channel. Here SEBlock represents squeeze excitation block as shown in Fig. 12.

learning based approaches that require a significant amount of annotated data, full annotation of 3D volumes is not an effective way to create large and rich training datasets that would generalize well [12]. As elaborated in section **??**, to investigate the capability of our proposed model structure in 3D, we extended our proposed model as shown in Fig. 3 to a 3D version by utilizing 3D convolutional layers and 3D max-pooling layers. In our work the convolutional layers used consisted of kernel size of 3x3x3, whilst the maxpooling layers used have a stride of 2x2x2. Due to limited computation resources, the filter size defined are 8, 16, 32, 64 and 128 respectively for each feature maps in the downsampling path as well as it's symmetrical counter part in the upsampling path.

3.5 Evaluation Metrics

To evaluate the model performance of our proposed method, Dice Score and Hausdorff Distance(HD95) was applied for Enhanced tumour(ET), Whole tumour(WT), and tumour Core(TC) regions. The range of Dice coefficient will be from 0 to 1. For two sets A and B define the Dice coefficient:

$$Dice = \frac{2(|A \cap B|)}{|A| + |B|} \tag{2}$$

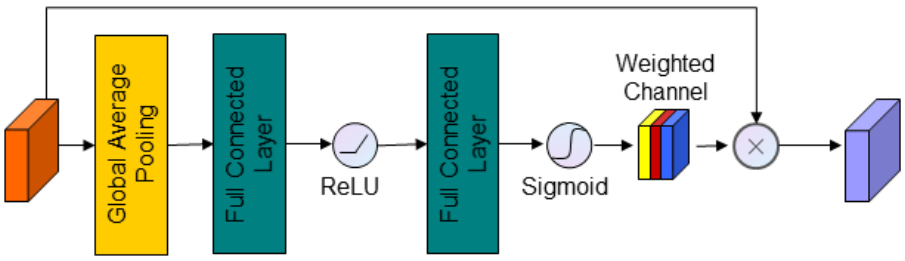


Fig. 12. Squeeze excitation block inspired by [42] as a learnable parameter for CAM.

The Hausdorff Distance (HD) measures how far two subsets of a metric space are from each other and is defined as the longest distance between a point set A and the most adjacent point of set B :

$$HD(A,B) = max\{\sup_{a\in A}\inf_{b\in B} \mathrm{d}(a,b), \sup_{a\in A}\inf_{b\in B} \mathrm{d}(b,a)\} \tag{3}$$

where $d(a,b)$ is the Euclidean Distance between a $\in$ A and b $\in$ B. Taking into inconsistent predictions problems, 95th percentile is used, which in the paper, abbreviated as HD95.

4 Loss Function

4.1 Loss for Unet based network

For the sake of compatibility, network structure that are describe in section 3.1, 3.2, 3.4 utilized the same loss function. The choice of the loss function plays a vital role in determining the proposed model performance as same as the network structure [50]. Conventional loss function for image segmentation task, such as cross entropy, would yield the inclination of the model to learn the label with the highest count and this effect is apparent in the state of heavy class imbalance. To tackle class imbalance problem induced by the nature of the provided label, in this paper, we adapted a loss functions with regularized term implemented by [42]. This loss function combines a Generalized Dice Loss (GDL) [50] together with a Cross-Entropy loss, which adaptively weights the classes to tackle class imbalance whilst accelerate the convergence, respectively.

$$L_{allloss} = L_{GDL}(G,P) + \lambda \times L_{CE}(G,P) \tag{4}$$

Here the λ is empirically set to 1.25. The adaptive weight of the weighted coefficient is derived as the fraction of the total number of labels as the denominator for each class respectively. This would hamper the tendency to learn labels with higher counts while encouraging the learning process of the model on labels with

relatively low counts. Where the weighted coefficient is inversely proportional with the number of label count.

$$L_{ce}(G, P) = wL_{ce}(G, P) \tag{5}$$

$$w = \frac{1}{\sum_{i=0}^{G} n_i} \tag{6}$$

Generalized Dice Loss calculates the intersection of union(IOU) of the segmented output with respect to the given labels for each class respectively.

4.2 Loss for multi-scale self-guided attention network

In our proposed network described in section 3.3, we adopted partial losses calculated in the original implementation by Sinha and Dolz [49]. Similar to the losses explained in section 4.1, we adopted two losses for the proposed network as shown in Fig. 5, i.e. Generalized Dice Loss(GDL) and Cross-Entropy loss. The difference is that because there are multiple outputs from the network, the GDL is computed by first taking the mean of the segmented output, which consisted of 6 images in total, and compare with the IOU of the actual label. On the other hand, cross entropy is computed by applying the mean for the accumulation of the cross entropy loss for each segmented output.

5 Experiments

5.1 Dataset Overview

All BraTS multi-modal scans per subject provide with a T1 weighted, a post-contrast T1-weighted, a T2-weighted and a FLAIR MRI, which was collected with different clinical protocols and various scanners from multiple (n=19) institutions [4]. The imaging from BraTS2020 dataset have been segmented manually by one to four raters, following the same annotation protocol, and their annotations were approved by experienced neuro-radiologists [5][39]. Each tumour was segmented into GD-enhancing tumour (label4), the peritumoural edema (label2), and the necrotic and non-enhancing tumour core (label1) with dimensions of $240 \times 240 \times 155$. The provided data has been preprocessed such as co-registered to the same anatomical template, interpolated to the same resolution ($1mm^3$) and skull-stripped.

5.2 Data Pre-processing

As opposed to conventional RGB images or gray scale images adapt pixel value within the range of $[0, 255]$, MRI intensity values varies throughout different modalities. It is crucial for MRI intensity values to be standardized so that the distribution of the pixel values would be compatible with normal images while feeding it into the model. In our work, we normalize each modality of each patient

respectively by first removing the outliers of the image intensities by purging top and bottom 1% intensities. Follows, the image is subtracted by the mean and dividing it with the standard deviation, which normalized the image to compatible value. Each imaging modalities receive the same normalization treatment. Each imaging modalities for an individual patient are stacked together as the modalities are treated as color channels, where an input image of a patient would output a dimension of [S,W,H,M] where *S: Number of slices, W: Image width, H: Image height, M: Number of modalities*. Subsequently, MRI slices that does not consisted brain region are remove due to redundancy as well as to leverage the computation power of the proposed network.

5.3 Data Augmentation

Data augmentation is an approach to prevent overfitting of the model by acquiring more diversity of data set via geometric, positional, image color and affine transformations. Generally, it is common practice to flip the image horizontally and vertically. In this work, each image has a 50% chance of flipping, vertically(left-right) and horizontally(up-down). Prior to the flipping of images, we also implemented random flip of an angle of $n \times 90°$, where n is defined as a random generated integer. Subsequently, all input images are injected with a Gaussian noise of 0.01 standard deviation inspired in the paper by Noori et al. [42].

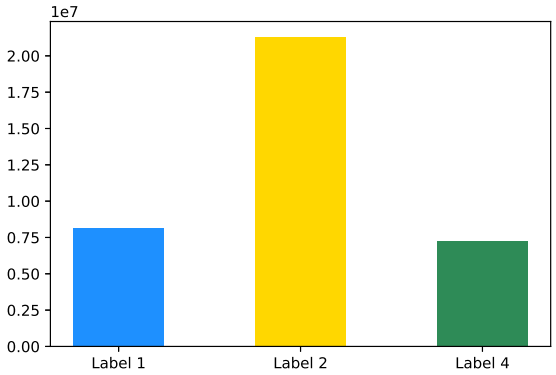


Fig. 13. Distribution of labels for training dataset with *(blue)*: Necrotic and Non-enhancing tumour Core, *(yellow)*: Peritumoural Edema, *(green)*: GD-Enhancing tumour, respectively

5.4 Label Distribution

Consistent with BraTS2020 challenge, the brain tumour segmentation labels are represented discretely with *[0,1,2,4]* corresponding to the *[Background, Non-enhancing tumour Core, Peritumoural Edema, GD-Enhancing tumour]* pixel-wise. Although trivial, the distribution of labels were counted and visualization of the label numbers for each class in the training dataset are plot in Fig.13. This provides a more general view of the label distribution and indication of class imbalance between the labels, specifically label 2 which is the dominant label. It is worth noting that the background labels (label 0) is excluded in Fig. 13 due to scalability.

5.5 Training Procedure

In the case of 2D network models (section 3.1, 3.2, 3.3), during the training process, the input shape of training dataset was set at $240 \times 240 \times 4$ with a batch size of 24 for UNet and DAUNet, as well as batch size of 8 for SGANet. Note that SGANet possess relatively low batch size due to the network capacity being able to fit in the input data. Whereas for 3D network models (section 3.4), the input dataset was set at $155 \times 240 \times 240 \times 4$ and was cropped in the slice axis to 128 due to the nature of downsampling with a scale of 2. The batch size for 3D network models is set at 1 due to limitation of the resources. All batch size are tested extensively for optimization in order to scale down the time cost.

For all network models, Adam optimizer was applied and learning rate was set to 1×10^{-5}. To reduce the over-fitting issues, the dropout value of 0.3 and He-initializer [21] for layer weights are implemented. For stop criteria, the max epoch was empirically set at 30 due to the time and memory limitations. The time-cost for all network models are as shown in Table 5.5. The time-cost for each models are approximately 20, 28, 36 and 8 hours respectively for the training process. There are 369 patients in the training data and 125 patients in the validation dataset from BraTS2020 challenge. Total parameter for each models are (34,514,116), (35,563,044), (2,212,974) and (849,267) respectively for updating when training the model.

Network	Batch Size	Time Cost(Hr.)	Total Parameter
UNet	24	20	34,514,116
DAUNet	24	28	35,563,044
SGANet	8	36	2,212,974
3D-DAUnet	1	8	849,267

Table 1. Network summary from section 3, where the unit for time cost is approximated in [Hour]. Here, DAUnet: Deep Supervised Attention Unet, SGANet: Multi-scale Self Guided Attention Network, 3D DAUnet: 3D Deep Supervised Attention Unet.

6 Results

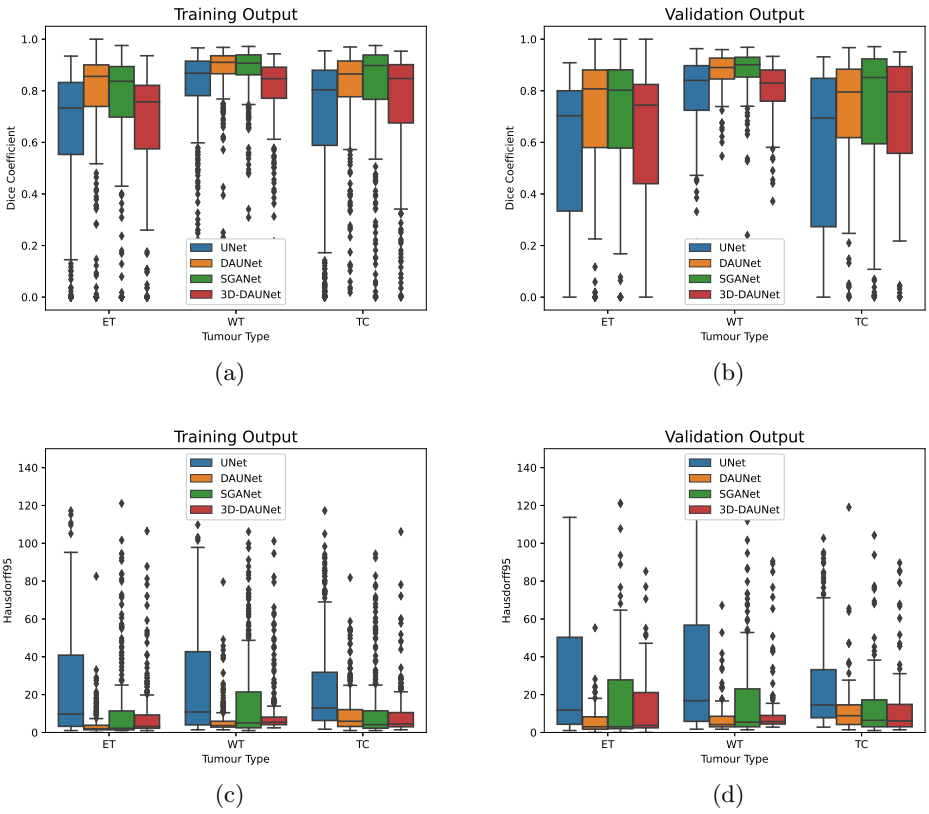


Fig. 14. Boxplot of image segmentation results depicting evaluation metric with respect to tumour type comprising of, (a) dice coefficient for training output, (b) dice coefficient for validation output, (c) Hausdorff95 for training output and (d) Hausdorff95 for validation output respectively. For tumour type, ET: Enhanced Tumour, WT: Whole Tumour, TC: Tumour Core. The abbreviation in the legend represents the model explain in section 3, where, (orange)DAUnet: Deep Supervised Attention UNet, (green)SGANet: Multi-scale Self Guided Attention Network, (red)3D-DAUnet: 3D Deep Supervised Attention Unet.

6.1 Cross Validation and Data Augmentation

In order to exploit the problem of overfitting on our first proposed model, DAUNet as explained in section 3.4. We experimented DAUNet with three vari-

ant of dataset, namely, original implementation(augmentation free), with augmentation and augmentation combined with 8-fold cross validation. The predicted original data and augmented data labels both for training and validation dataset were submitted to the online evaluation system. The results were as shown in Table.2 for training data and Table. 3 for validation data. Consistent with the evaluation metric in section 3.5, Dice coefficient and HD95 were used to evaluate the model performance on three labels: ET(Enhanced Tumour), WT(Whole Tumour), TC(Tumour Core), respectively. To better understand the model prediction performance, the ground truth and prediction label for a patient specimen were visualized in Figure.15 for original training dataset and Figure.16 for original validation data. It is worth noting that for the validation data, only predicted label was visualized since the ground truth is not made available throughout the challenge.

After extensive hyperparameter tuning, DAUNet yield mean dice coefficients of $[0.81, 0.92, 0.86]$ on the original training data and $[0.65, 0.86, 0.66]$ on the validation dataset for enhancing tumour, whole tumour, and tumour core respectively. The results showed a clear over-fitting issue on training dataset predictions from online evaluation system, which eventually, contributed to the low performance of the unseen data(validation data). With the introduction of data augmentation, the augmented training dataset has mean dice coefficients of $[0.75, 0.88, 0.80]$ and $[0.67, 0.86, 0.70]$ on validation dataset for enhancing tumour, whole tumour, and tumour core respectively. To further overcome the overfitting problem occurred in the original implementation, we incorporated data augmentation with 8-fold cross validation. With data augmentation and 8-fold cross validation, our model managed to obtain a mean dice coefficients of $[0.61, 0.81.0.68]$ on training dataset, as well as $[0.58, 0.8, 0.61]$ on validation dataset for enhancing tumour(ET), whole tumour(WT), and tumour core(TC) respectively.

The augmented training data gained a $3 \sim 4\%$ performance boost in mean dice score as compared with the model without augmentation on the validation dataset. Whereas the model performed about 10% better in mean dice score for the data augmented dataset as compared to the data augmented combined with 8-fold cross validation dataset. Clearly, this proves that the data augmentation was an effective method to enhance the model performance in tackling the overfitting issue. As observed from Table 2 and 3, it is observed that enhanced tumour label yield the lowest mean dice score as opposed to whole tumour label and tumour core label. This might be due to the fact that DAUNet does not possessed proficient capability to segment the pixel label from it's neighbour counterparts.

Similar trend could be observed on the measurement of HD95 for all trials. The value of HD95 on augmentation data model$[40.60, 7.94, 15.75]$ are smaller than the original data model$[46.40, 11.01, 16.03]$ as well as the augmented combined with cross-validation data model $[42.76, 19.25, 24.47]$ on validation data for enhancing tumour(ET), whole tumour(WT), and tumour core(TC) respec-

tively. Augmented dataset yield the best results out of the variants in preventing overfitting. Hence, we opted for only data augmentation in our dataset.

	Label	**Dice**			**Hausdorff95**		
		ET	**WT**	**TC**	**ET**	**WT**	**TC**
No Augmentation	Mean	0.81793	0.91904	0.86358	23.49635	5.10187	4.29509
	StdDev	0.23004	0.05521	0.19895	86,11617	11.11984	7.84377
	Median	0.89225	0.93622	0.9372	1.41421	2.23607	1.73205
	25quantile	0.82142	0.90353	0.88291	1.0	1.73205	1.41421
	75quantile	0.93159	0.95362	0.95967	2.0	3.74166	3.0
With Augmentation	Mean	0.75983	0.8854	0.80669	24.72246	6.01101	10.15846
	StdDev	0.25022	0.08867	0.17964	85.92247	7.67428	11.41269
	Median	0.85605	0.91036	0.8652	2.0	3.60555	5.91608
	25quantile	0.73996	0.86622	0.77794	1.41421	2.82843	3.16228
	75quantile	0.90012	0.93592	0.91625	3.74166	5.91608	12.04160
Augmentation + Cross-Validation	Mean	0.61313	0.81364	0.68336	34.88344	17.75722	18.72055
	StdDev	0.2537	0.12823	0.21763	90.8749	19.64714	19.60315
	Median	0.69907	0.85351	0.75798	4.89898	7.54983	11.0
	25quantile	0.52935	0.7734	0.59661	3.16228	4.47214	6.0
	75quantile	0.78024	0.89532	0.84051	13.36599	25.7682	22.60309

Table 2. Dice and HD95 for BraTS 2020 training dataset for DAUNet model, three variant: No Augmentation, With Augmentation and Cross Validation

6.2 Models Output

Based on the evaluation metrics after deploying the models out mentioned in section 3 to the online evaluation platform, we were able summarize the results into boxplot as shown in Fig. 6. Subsequently, the mean, standard deviation, median, 25th quantile and 75th quantile values for each models are shown in Table 4 for training data and Table 5 respectively.

From Table 4, it is apparent that our proposed networks (DAUNet, SGANet and 3D-DAUNet) has better performance as compared to the benchmark model, UNet, with 3D-DAUNet falling a margin behind for the dice score for WT. However, the Hasdorff95 for 3D-DAUNet is relatively low as compared to UNet which indicates that 3D-DAUnet is able to retain the geometric shape of the segmented image. By focusing on the mean value for the evaluation metrics of the models, DAUNet and SGANet have comparative dice scores for the segmented criterion. However, DAUNet has dominant Hausdorff95 advantage as compared to SGANet for the training dataset. The image segmentation output for training dataset with respect to different models are depicted in Fig. 15.

On the other hand, in the case of validation dataset, as observed from Table 5, the domination from DAUNet and SGANet preserve where SAGNet performed slightly better in segmenting WT and TC for dice score. However, DAUNet

	Label	Dice			Hausdorff95		
		ET	**WT**	**TC**	**ET**	**WT**	**TC**
No Augmentation	Mean	0.64974	0.85767	0.66157	46.40101	11.01570	16.03994
	StdDev	0.3058	0.13076	0.31598	113.01136	18.03175	47.18426
	Median	0.79353	0.89633	0.83136	3.16228	4.58258	6.08276
	25quantile	0.5166	0.84108	0.4773	2.0	3.0	3.16228
	75quantile	0.8673	0.92781	0.9053	12.20656	9.43398	13.78405
With Augmentation	Mean	0.67357	0.86084	0.70421	40.60792	7.94162	15.75085
	StdDev	0.30806	0.12601	0.25173	109.01934	10.05485	35.83261
	Median	0.8083	0.8899	0.79739	3.0	4.24264	8.94427
	25quantile	0.58241	0.84655	0.62058	1.73205	3.0	4.24264
	75quantile	0.88178	0.92575	0.88494	8.06226	8.54400	15.0
Augmentation + Cross-Validation	Mean	0.5798	0.80435	0.61458	42.76221	19.25796	24.47344
	StdDev	0.2999	0.14931	0.26964	104.4549	21.29113	38.78857
	Median	0.69605	0.85034	0.71418	5.0	7.81025	14.03567
	25quantile	0.40615	0.78704	0.47814	3.0	4.58258	7.61248
	75quantile	0.79985	0.8914	0.81933	14.42559	28.17978	22.82542

Table 3. Dice and HD95 for BraTS 2020 validation dataset for DAUNet model, three variant: No Augmentation, With Augmentation and Cross Validation

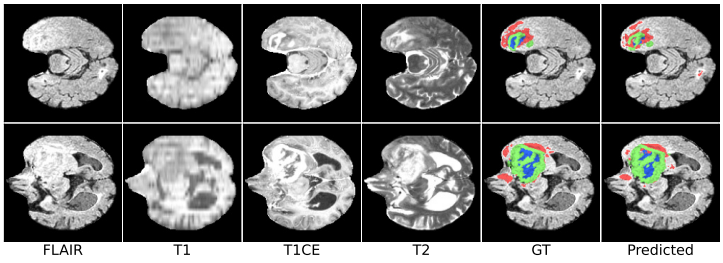
performed better for ET dice score as well as in the Hausdorff95. 3D-DAUNet performed better in all aspect as compared to UNet for the validation dataset. There is still a gap between the dice score as well as Hausdorff95 for all models during the implementation of training dataset and validation dataset, indicating overfitting issue for our model. The image segmentation output for validation dataset with respect to different models are depicted in Fig. 16. Multiple slices of brain tumour segmentation for two patients with respect to various model are shown in Appendix A, Fig. 22 − 29.

	Label	Dice			Hausdorff95		
		ET	**WT**	**TC**	**ET**	**WT**	**TC**
UNet	Mean	0.63354	0.81972	0.67926	46.8488	25.11373	23.53103
	StdDev	0.27738	0.14011	0.28999	95.06432	26.53785	24.11631
	Median	0.73382	0.86818	0.8043	9.43398	10.63015	12.91505
	25quantile	0.5533	0.78122	0.59116	3.25489	4.0	6.32456
	75quantile	0.83195	0.91521	0.87938	40.60788	42.59988	31.78915
DAUNet	Mean	**0.75983**	**0.8854**	0.80669	**24.72246**	**6.01101**	10.15846
	StdDev	0.25022	0.08867	0.17964	85.92247	7.67428	11.41269
	Median	0.85605	0.91036	0.8652	2.0	3.60555	5.91608
	25quantile	0.73996	0.86622	0.77794	1.41421	2.82843	3.16228
	75quantile	0.90012	0.93592	0.91625	3.74166	5.91608	12.04160
SGANet	Mean	0.73344	0.87744	**0.80877**	36.81143	15.60308	11.07561
	StdDev	0.26556	0.09863	0.21387	96.56794	21.16217	16.69531
	Median	0.83676	0.90716	0.89805	2.23607	5.0	4.12311
	25quantile	0.69875	0.86251	0.76845	1.41421	2.33949	2.23607
	75quantile	0.89365	0.9389	0.93869	11.35782	21.40094	11.44552
3D-DAUNet	Mean	0.64768	0.80854	0.75363	34.79334	9.50227	**8.92873**
	StdDev	0.26404	0.12566	0.21943	96.0733	13.29745	11.44526
	Median	0.75849	0.84724	0.84751	3.16228	5.38516	4.47214
	25quantile	0.57546	0.77132	0.67661	2.23607	4.12311	3.0
	75quantile	0.82181	0.89129	0.90125	9.04708	8.12404	10.48809

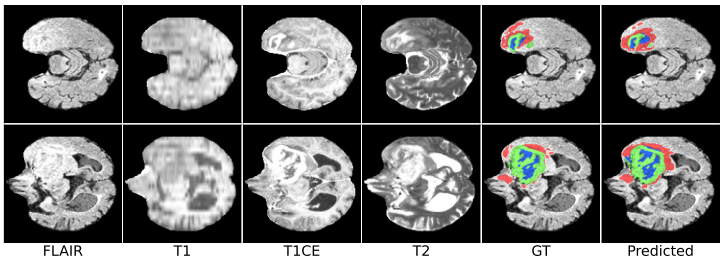
Table 4. Dice and HD95 for BraTS 2020 training dataset for benchmark model and proposed models.

	Label	Dice			Hausdorff95		
		ET	**WT**	**TC**	**ET**	**WT**	**TC**
UNet	Mean	0.55043	0.77359	0.56635	62.03482	31.93506	26.15598
	StdDev	0.31897	0.18658	0.32506	114.2815	29.7414	26.22859
	Median	0.70396	0.84339	0.70239	12.0	16.58312	14.72752
	25quantile	0.33798	0.72932	0.27897	4.3589	5.65685	7.81025
	75quantile	0.80084	0.8972	0.85158	49.74937	56.17829	32.95694
DAUNet	Mean	**0.67357**	0.86084	0.70421	**40.60792**	**7.94162**	**15.75085**
	StdDev	0.30806	0.12601	0.25173	109.01934	10.05485	35.83261
	Median	0.8083	0.8899	0.79739	3.0	4.24264	8.94427
	25quantile	0.58241	0.84655	0.62058	1.73205	3.0	4.24264
	75quantile	0.88178	0.92575	0.88494	8.06226	8.54400	15.0
SGANet	Mean	0.65678	**0.87166**	**0.70454**	51.61144	18.62864	17.3213
	StdDev	0.31808	0.10059	0.30473	112.7819	25.66396	37.62899
	Median	0.8027	0.89954	0.85413	3.0	5.38516	6.40312
	25quantile	0.57842	0.85351	0.60078	2.0	3.0	3.0
	75quantile	0.8822	0.92968	0.92359	24.50508	22.65943	16.58312
3D-DAUNet	Mean	0.60248	0.79398	0.68093	54.07402	12.05111	19.10066
	StdDev	0.30563	0.14588	0.27468	119.2871	17.57802	48.81442
	Median	0.7457	0.82954	0.8002	3.74166	5.74456	6.08276
	25quantile	0.44868	0.76014	0.56009	2.44949	4.47214	3.0
	75quantile	0.83009	0.87991	0.89545	21.09502	9.0	14.59452

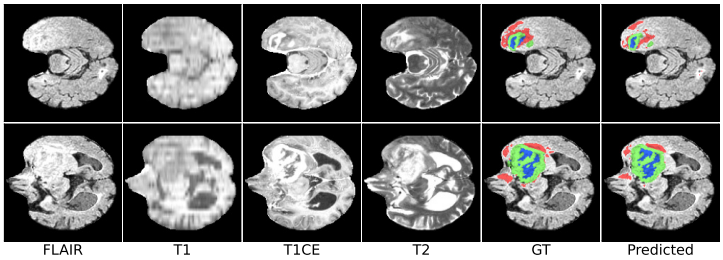
Table 5. Dice and HD95 for BraTS 2020 validation dataset for benchmark model and proposed models.



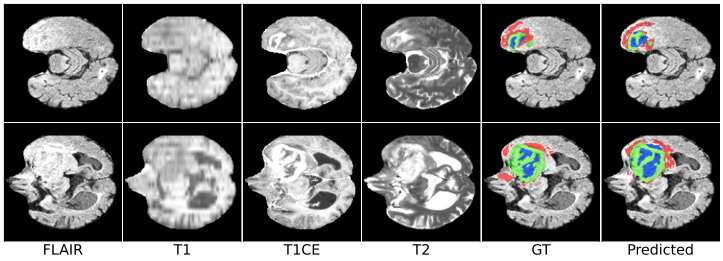
(a) UNet



(b) DAUNet

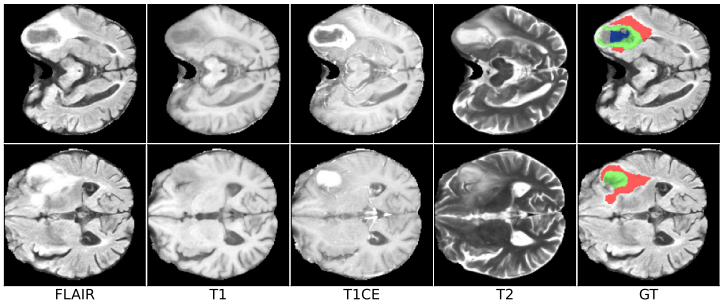


(b) SGANet

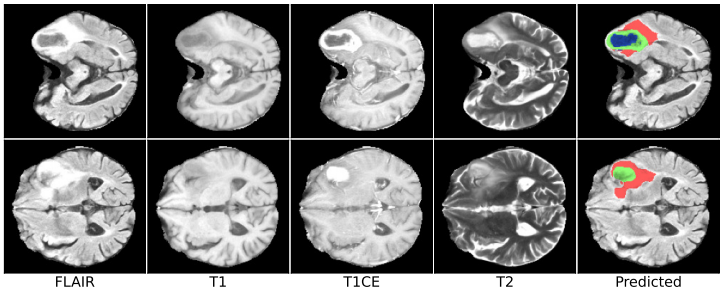


(b) VNet

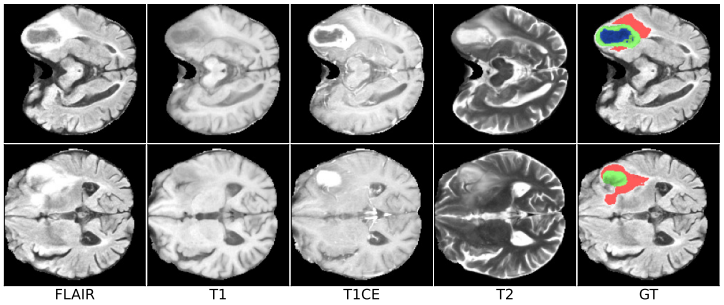
Fig. 15. Segmentation results for BraTS 2020 training dataset, from left to right : FLAIR, T1, T1CE, T2, ground truth(GT),Predicted; Colors: Necrotic and Non-enhancing tumour Core(Blue) Peritumoural Edema(Red) GD-Enhancing tumour(Green). For this instance, the plots are from patient id 338, slice 50 (first row) and 65(second row).



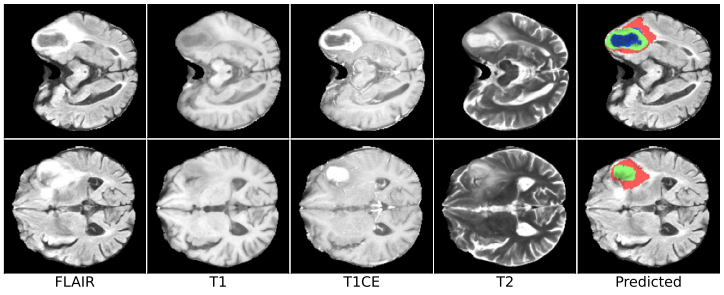
(a) UNet



(b) DAUNet



(b) SGANet



(b) VNet

Fig. 16. Segmentation results for BraTS 2020 validation dataset, from left to right : FLAIR, T1, T1CE, T2, Predicted; Colors: Necrotic and Non-enhancing tumour Core(Blue) Peritumoural Edema(Red) GD-Enhancing tumour(Green). For this instance, the plots are from patient id 24, slice 50 (first row) and 65(second row).

7 Discussion

7.1 Overview

In this section, we try to explain why the model does not perform as expected, the pitfalls and what should have been done to improve the situation. As we all known, Machine Learning Models are like black boxes, which is hard for us to know what happens exactly inside. Especially, deep neural networks are introduced increasingly complex and opaque models with decision boundaries that are extremely hard to understand. Despite many recent developments in explainable AI, there are still enormous challenges for explaining deep neural networks [32]. One of the advantages of attention mechanism is to provide introspection for the model network. In our work, we omitted the introspection of 3D model, 3D-DAUNet due to the nature of 3D being challenging to interpret.

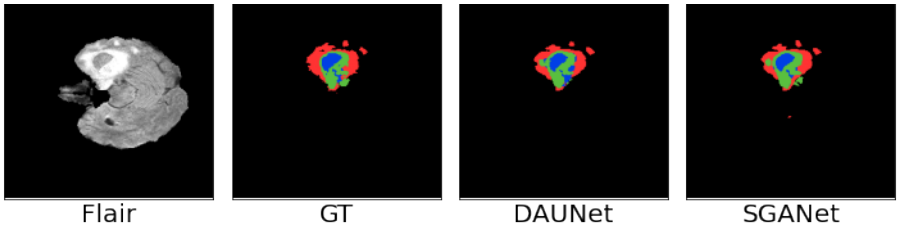


Fig. 17. An exemplary plot of input image (Flair modality) with ground truth and segmented output for DAUNet and SGANet. Colors: Necrotic and Non-enhancing tumour Core(Blue) Peritumoural Edema(Red) GD-Enhancing tumour(Green).

7.2 Introspection

Depicted in Fig. 17, we utilized training dataset as our input data and feed it to our 2D proposed model(DAUNet and SGANet) to retrieved the relevant features for visualization.

DAUNet: In the case of DAUNet, we can visualize the grid-like attention coefficient distribution of the attention gates. Fig. 18 depicted a schematic plots of attention coefficient distribution with respect to the input image shown in Fig. 17. From Fig. 18, particularly for Fig. 18(c) and (d) where the attention gates correspond to the deeper layers in the network in the upsampling path. The attention coefficient with high activations clustered around the area that resembles the ground truth. Whereas, the attention gates in Fig. 18(a) and (b) correspond to the shallow layers in the network in the downsampling path. Here, the attention coefficient has lower activation clustering around the area that resembles the ground truth. From Fig. 18, we can validate that the attention

gate did a relatively decent job in attending towards the features relevant to the ground truth. However, attention gates at the shallow layer tend to pick up more noise for low level features and this might contribute to the lack of performance for the segmentation task during the training process.

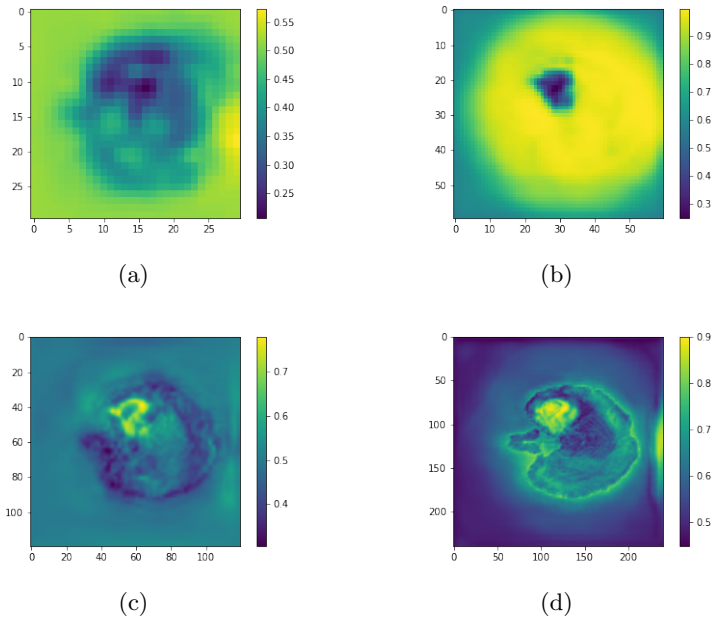


Fig. 18. Visualization of grid-like attention coefficient for attention gates with the attention gate from bottom to top in the network architecture, DAUNet as explain in section 3.2, where, (a) Attention Gate 01 (b) Attention Gate 02 (c) Attention Gate 03 (d) Attention Gate 04, respectively.

SGANet: With similar attention mechanism used in DAUNet, we can visualize the grid-like attention coefficient distribution of the attention gates as shown in Fig.19. Fig.19(a), (b),(c) shows the attention coefficient from the attention gate positioning from left to right in the model structure from Fig. 5, respectively, where the attention gate received feature maps from Res01, Res02 and Res03. As opposed to the attention coefficients distribution explained in the model, DAUNet, SGANet attention gates seem to pick up lower activation signal from the ROI on the first two attention gates, where the attention coefficients seems to activate everywhere in the final attention gate. To further validate the effectiveness of attention mechanism, we exploit the feature maps from the output of PAM and CAM in SGANet as plot in Appendix B and B, Fig. 30, 31, 32, 33, 34, 35. To summarize the output of the feature map, we take the maximum

activation of the attention coefficient with respect to the axis of the filter size (channel axes), such that we obtain a heat map like distribution that explains where the feature maps learnt in the PAM and CAM modules as shown in Fig. 20. In Fig. 20, the first column represents the maximum activation of a feature maps from PAM, whereas the second column represents the maximum activation of a feature maps from CAM. It is observed that PAM provides better activation for the positioning of the region of interest as compared to CAM, but the high activations of the feature maps clustered in other region rather than the ROI. The lack of performance from SGANet might yield from the redundancy of implementation from the complex layer as we can observe from the attention gate in Fig. 19(c) and CAM in Fig. 20(b),(f).

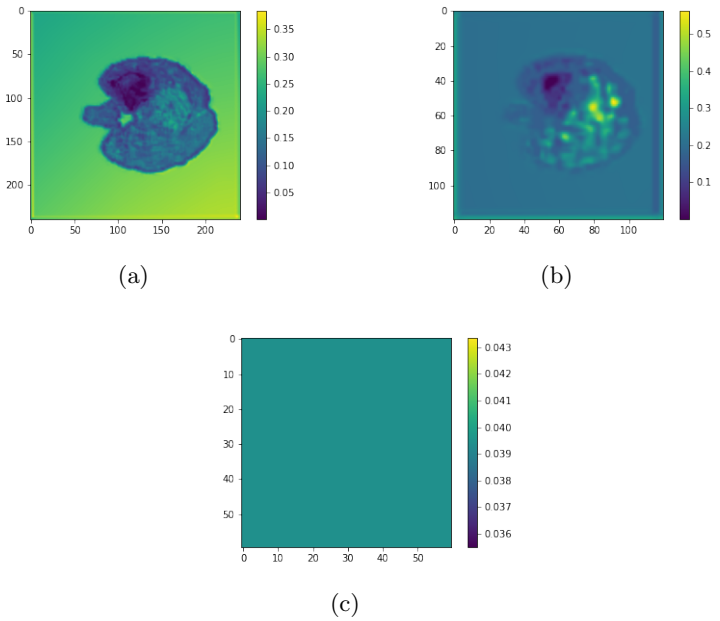


Fig. 19. Visualization of grid-like attention coefficient for attention gates with the attention gate from left to right in the network architecture, SGANet as explain in section 3.3, where, (a) Attention Gate 01 (b) Attention Gate 02 (c) Attention Gate 03, respectively.

7.3 Model Complexity

Model complexity has always been an on going discussion in the field of deep learning. The trade off between model complexity and the accuracy are highly subjective, but model with less complicated model structure and good perfor-

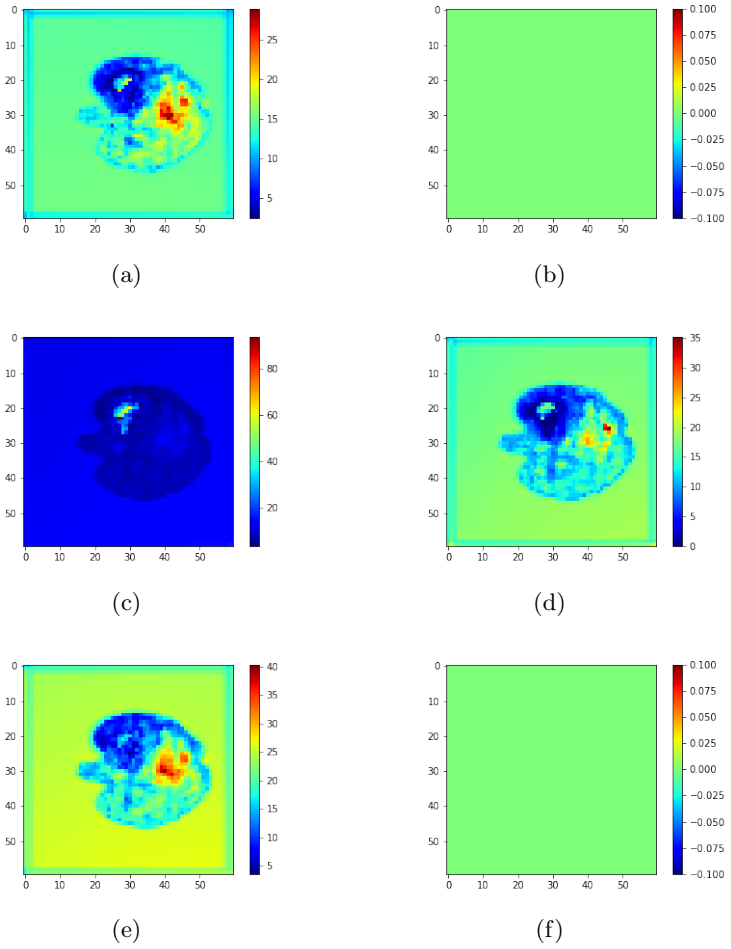


Fig. 20. Visualization of heatmap for the maximum activation of the feature maps along the number of filters for PAM and CAM from SGANet, where, (a) PAM00 (b) CAM00 (c) PAM01 (d) CAM01 (e) PAM02 (f) CAM02, respectively.

mance are desirable. Model complexity also serve as a hindrance for computational resources where GPU VRAM are not able to accustom to the model structure complexity during the training process. In our case, such issue arise during the implementation of SGANet. In the original paper by Sinha and Dolz [49], the feature maps are not downscaled when being fed into PAM and CAM modules. However, in our work, empirically, we acquired to downscale the input features by a scale of 4, such that the GPU is able to proceed without error during the training phase. We argue that the downscaled feature would lead to the information loss during the training phase, hence, resulting to obstructing the potential of the network model as a whole.

The same goes to 3D-DAUNet model structure as explained in section **??**. Generally, 3D structures has always been a hog for computational resources rendering to the low number of filters for each layers in the model structure. Subsequently, we argue that due to the low number of filter size in the layers, the model failed to extract salient features throughout the network and lead to the less credible performance for 3D model. Another factor might be due to the empirical coefficient implemented in the loss function as shown in Eq. 4. The empirical coefficient in the second term of the loss function is designated for 2D network structure, as shown in the paper by Noori et al.[42]. We did not exploit empirically the best empirical coefficient for our 3D model, hence contributing to the not so ideal performance from the 3D model.

7.4 Future Work

Model complexity reduction: as discussed in section 7.2 and 7.3, the redundancy of over-defined filter size in the model structure of SGANet in directly increase the model complexity but does not contribute to the performance of the model structure. This has led to the fatigue of hardware memory such that the feature maps need to be downscaled in the deeper phase of the network to fit in the VRAM capacity of the GPU. Therefore, reducing the model complexity would help to alleviate the need for downscaling in the later phase of the network to prevent information loss for SGAUnet.

Exploiting 3D models: Even though our 3D model does not perform well as compared to our 2D models, DAUNet and SGANet, 3D deep learning network models has always shown promising performance in the realm of medical image segmentation. For BraTS 2019 challenge, Jiang et al.[27] implemented two-stage cascading U-net with 3D convolution that achieved an average dice scores of 0.83267(ET), 0.88796(WT) and 0.83697(TC). In another work by Zhou et al.[60], the authors implemented spatial pyramid pooling with 3D atrous-convolution [10] for brain tumour segmentation.

Choice of loss function: Due to the class imbalance from the class labels, the importance of loss function showed significant importance to penalize dominant class label. Dice loss has proved to be efficient in tackling class imbalance, but

cross entropy has been the Achilles heel of the loss function. In the work by Noori et al.[42], weighted cross entropy has been proposed in the wake to tuple the class imbalance dilemma. The question remains to be empirical towards the best weight for cross entropy. Hence, Lin et al.[30] proposed a variant of cross entropy that aims to help cross entropy to converge faster in the case of class imbalance, known as focal loss. The equation for focal loss is as shown in Eq. 7 and depicted in Fig. 21.

$$f(y_i) = \begin{cases} CE(p_t) = -log(p_t) \\ FL(p_t) = -(1-p_t)^t log(p_t) \end{cases} \tag{7}$$

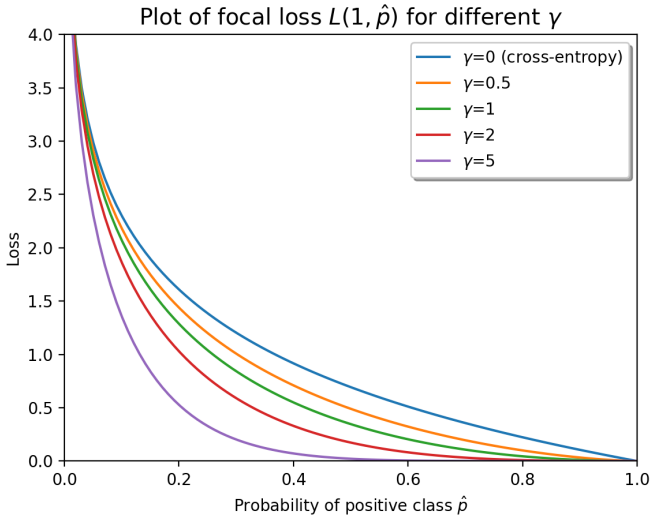


Fig. 21. Schematic plot of focal loss with various γ values, where $FL(p_t) = CE(p_t)$ when $\gamma = 0$. From the dampening of the loss from the increment of γ values, γ reduces the relative loss for well-classified examples, indicating more focus on misclassified examples. (Image source from the work of Lin et al.[30])

8 Conclusion

We have proposed two 2D models and one 3D model, namely, Deep supervised Attention UNet(DAUNet), multi-scaled Self Guided Attention Network(SGANet) and 3D-Deep supervised Attention UNet(3D-DAUNet), framework for pixel-wise brain tumour segmentation, which can focus on brain tumour low resolution and high resolution feature representations with an attention mechanism and a multistage segmentation layer. The result of mean Dice coefficient is $[0.67, 0.86, 0.70]$

for DAUNet, $[0.65, 0.87, 0.70]$ for SGANet and $[0.60, 0.79, 0.68]$ for 3D-DAUNet on validation dataset. For all three models the ET(enhanced tumour) shows a lower segmentation rate compared with other two labels, which this trend could be observed in previous challenges. The most probable reason would be ET was defined as a small region in the label which pose a challenge for the network architecture to pick up relevant signal around the region of interest.

In our proposed networks, 2D model architectures, DAUNet and SGANet outperformed 3D model architectures in all evaluation metrics which correlates with our project goal to validate the effectiveness of deep learning network models in medical image segmentation with 2D models. However, due to the model complexity that contribute to the redundancies of feature maps extraction during the learning phase for SGANet serve as a dilemma for us to push to boundaries for 2D model architecture without shadowing the structure of UNet. Consequently, the extension of UNet as a backbone with the combination of attention mechanism and deep supervision, DAUNet, provides the best result in the BraTS 2020 challenge for brain tumour segmentation. Furthermore, due to time limitation, 3D network architectures for medical image segmentation are yet to be fully exploited. In the future work, we will focus on reducing the model complexity 2D model structures to reduce redundancies, exploit the potential of a 3D model architecture as well as venture with loss function to tackle class imbalance for optimum model performance.

ACKNOWLEDGEMENT

Wai Po Kevin Teng implemented almost all of the code, planned and performed almost all of the experiment for this project. Wai Po Kevin Teng compiled almost all of the project report and generate all of the figures(except those that are sourced from the paper of the cited author for referencing). Xiongjun Wang aided in running limited trials for UNet and DAUNet models. Furthermore, Xiongjun Wang contributed partially to the report, in section 1, section 2, section 3.1, section 3.4 and section 7.1 only. Our team would like to specifically show our gratitude towards our project supervisor M.Sc. Jia Hua Xu for his insightful inputs and selfless guidance.

References

1. Andreas H. Jacobs, L.W.K., Axel Gossmann, Maria A. Rüger, A.V.T.A.T., Herholz, K.: Imaging in neurooncology. NeuroRx **24**(2), 333–347 (2005). https://doi.org/https://doi.org/10.1602/neurorx.2.2.333
2. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., Freymann, J., Farahani, K., Davatzikos, C.: Segmentation labels and radiomic features for the pre-operative scans of the tcga-gbm collection. the cancer imaging archive. Nat Sci Data **4**, 170117 (2017)
3. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., Freymann, J., Farahani, K., Davatzikos, C.: Segmentation labels and radiomic features for

the pre-operative scans of the tcga-lgg collection. The cancer imaging archive **286** (2017)
4. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C.: Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. Scientific data **4**, 170117 (2017)
5. Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R.T., et al., C.B.: Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. CoRR **abs/1811.02629** (2018), http://arxiv.org/abs/1811.02629
6. Boykov, Y., Kolmogorov, V.: An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. IEEE Transactions on Pattern Analysis and Machine Intelligence **26**(9), 1124–1137 (2004). https://doi.org/10.1109/TPAMI.2004.60
7. Boykov, Y.Y., Jolly, M..: Interactive graph cuts for optimal boundary region segmentation of objects in n-d images. In: Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001. vol. 1, pp. 105–112 vol.1 (2001). https://doi.org/10.1109/ICCV.2001.937505
8. Brodbelt, A., Greenberg, D., Winters, T., Williams, M., Vernon, S., Collins, V.P.: Glioblastoma in england: 2007–2011. European Journal of Cancer **51**(4), 533–542 (2015). https://doi.org/10.1016/j.ejca.2014.12.014, https://pubmed.ncbi.nlm.nih.gov/25661102/
9. Chen, H., Dou, Q., Yu, L., Qin, J., Heng, P.A.: Voxresnet: Deep voxelwise residual networks for brain segmentation from 3d mr images. NeuroImage **170**, 446–455 (2018). https://doi.org/10.1016/j.neuroimage.2017.04.041
10. Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. CoRR **abs/1606.00915** (2016), http://arxiv.org/abs/1606.00915
11. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. CoRR **abs/1610.02357** (2016), http://arxiv.org/abs/1610.02357
12. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: Learning dense volumetric segmentation from sparse annotation. CoRR **abs/1606.06650** (2016), http://arxiv.org/abs/1606.06650
13. Dong, H., Yang, G., Liu, F., Mo, Y., Guo, Y.: Automatic brain tumor detection and segmentation using u-net based fully convolutional networks (2017)
14. Feng, X., Tustison, N.J., Patel, S.H., Meyer, C.H.: Brain tumor segmentation using an ensemble of 3d u-nets and overall survival prediction using radiomic features. Frontiers in computational neuroscience **14**, 25 (2020). https://doi.org/10.3389/fncom.2020.00025
15. Fillioux, L.: Using attention for medical image segmentation (2020), https://towardsdatascience.com/using-attention-for-medical-image-segmentation-dd78825eaac6
16. Fu, J., Liu, J., Tian, H., Fang, Z., Lu, H.: Dual attention network for scene segmentation. CoRR **abs/1809.02983** (2018), http://arxiv.org/abs/1809.02983
17. Graves, A., Wayne, G., Danihelka, I.: Neural turing machines. CoRR **abs/1410.5401** (2014), http://arxiv.org/abs/1410.5401
18. Hanif, F., Muzaffar, K., Perveen, K., Malhi, S.M., Simjee, S.U.: Glioblastoma multiforme: A review of its epidemiology and pathogenesis through clinical presentation and treatment. Asian Pacific journal of cancer prevention : APJCP **18**(1), 3–9 (2017). https://doi.org/10.22034/APJCP.2017.18.1.3

19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR **abs/1512.03385** (2015), http://arxiv.org/abs/1512.03385
20. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. CoRR **abs/1502.01852** (2015), http://arxiv.org/abs/1502.01852
21. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. CoRR **abs/1502.01852** (2015), http://arxiv.org/abs/1502.01852
22. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. CoRR **abs/1603.05027** (2016), http://arxiv.org/abs/1603.05027
23. Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E.: Squeeze-and-excitation networks (2019)
24. Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., Maier-Hein, K.H.: Brain tumor segmentation and radiomics survival prediction: Contribution to the brats 2017 challenge, https://arxiv.org/pdf/1802.10508
25. Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., Maier-Hein, K.H.: Brain tumor segmentation and radiomics survival prediction: Contribution to the BRATS 2017 challenge. CoRR **abs/1802.10508** (2018), http://arxiv.org/abs/1802.10508
26. Jetley, S., Lord, N.A., Lee, N., Torr, P.H.S.: Learn to pay attention. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net (2018), https://openreview.net/forum?id=HyzbhfWRW
27. Jiang, Z., Ding, C., Liu, M., Tao, D.: Two-stage cascaded u-net: 1st place solution to brats challenge 2019 segmentation task. In: Crimi, A., Bakas, S. (eds.) Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - 5th International Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Revised Selected Papers, Part I. Lecture Notes in Computer Science, vol. 11992, pp. 231–241. Springer (2019). https://doi.org/10.1007/978-3-030-46640-4_22, https://doi.org/10.1007/978-3-030-46640-4_22
28. Kayalibay, B., Jensen, G., van der Smagt, P.: Cnn-based segmentation of medical imaging data. CoRR **abs/1701.03056** (2017), http://arxiv.org/abs/1701.03056
29. Lee, B., Yamanakkanavar, N., Choi, J.Y.: Automatic segmentation of brain mri using a novel patch-wise u-net deep architecture. PloS one **15**(8), e0236493 (2020). https://doi.org/10.1371/journal.pone.0236493
30. Lin, T., Goyal, P., Girshick, R.B., He, K., Dollár, P.: Focal loss for dense object detection. CoRR **abs/1708.02002** (2017), http://arxiv.org/abs/1708.02002
31. Liu, R., Lehman, J., Molino, P., Such, F.P., Frank, E., Sergeev, A., Yosinski, J.: An intriguing failing of convolutional neural networks and the coordconv solution. CoRR **abs/1807.03247** (2018), http://arxiv.org/abs/1807.03247
32. Liu, S., Kailkhura, B., Loveland, D., Han, Y.: Generative counterfactual introspection for explainable deep learning (2019)
33. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. CoRR **abs/1411.4038** (2014), http://arxiv.org/abs/1411.4038
34. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. CoRR **abs/1411.4038** (2014), http://arxiv.org/abs/1411.4038
35. Louis, D. N., H.E.C., Cairncross, J.G.: Glioma classification: a molecular reappraisal. The American journal of pathology **159**(3), 779–786 (2001). https://doi.org/https://doi.org/10.1016/S0002-9440(10)61750-6
36. Lu, J., Xiong, C., Parikh, D., Socher, R.: Knowing when to look: Adaptive attention via a visual sentinel for image captioning (2017)

37. Lundervold, A.S., Lundervold, A.: An overview of deep learning in medical imaging focusing on MRI. Zeitschrift für Medizinische Physik **29**(2), 102–127 (May 2019). https://doi.org/10.1016/j.zemedi.2018.11.002, https://doi.org/10.1016/j.zemedi.2018.11.002
38. Luong, M.T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation (2015)
39. Menze, B.H., Jakab, A., Bauer, e.a.: The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging **34**(10), 1993–2024 (2015). https://doi.org/10.1109/TMI.2014.2377694
40. Milletari, F., Navab, N., Ahmadi, S.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. CoRR **abs/1606.04797** (2016), http://arxiv.org/abs/1606.04797
41. Nam, H., Ha, J.W., Kim, J.: Dual attention networks for multimodal reasoning and matching (2017)
42. Noori, M., Bahri, A., Mohammadi, K.: Attention-guided version of 2d unet for automatic brain tumor segmentation. In: 2019 9th International Conference on Computer and Knowledge Engineering (ICCKE). pp. 269–275. IEEE (10/24/2019 - 10/25/2019). https://doi.org/10.1109/ICCKE48569.2019.8964956
43. Oktay, O., Schlemper, J., Le Folgoc, L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., Glocker, B., Rueckert, D.: Attention u-net: Learning where to look for the pancreas, https://arxiv.org/pdf/1804.03999
44. Parikh, A.P., Täckström, O., Das, D., Uszkoreit, J.: A decomposable attention model for natural language inference. CoRR **abs/1606.01933** (2016), http://arxiv.org/abs/1606.01933
45. Prabhu, R.: Understanding of convolutional neural network (cnn) — deep learning (2018), https://medium.com/@RaghavPrabhu/understanding-of-convolutional-neural-network-cnn-deep-learning-99760835f148
46. Ren, M., Zemel, R.S.: End-to-end instance segmentation with recurrent attention (2017)
47. Roerdink, J.B., Meijster, A.: The watershed transform: Definition, algorithms and parallelization strategies. Fundamenta Informaticae 41 pp. 187–228 (2001)
48. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation (2015)
49. Sinha, A., Dolz, J.: Multi-scale guided attention for medical image segmentation. CoRR **abs/1906.02849** (2019), http://arxiv.org/abs/1906.02849
50. Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Cardoso, M.J.: Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. CoRR **abs/1707.03237** (2017), http://arxiv.org/abs/1707.03237
51. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. CoRR **abs/1706.03762** (2017), http://arxiv.org/abs/1706.03762
52. Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X.: Residual attention network for image classification (2017)
53. Wang, X., Peng, Y., Lu, L., Lu, Z., Summers, R.M.: Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays (2018)
54. Woo, S., Park, J., Lee, J., Kweon, I.S.: CBAM: convolutional block attention module. CoRR **abs/1807.06521** (2018), http://arxiv.org/abs/1807.06521
55. Wu, Y., He, K.: Group normalization. CoRR **abs/1803.08494** (2018), http://arxiv.org/abs/1803.08494

56. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention (2016)
57. Yamashita, R., Nishio, M., Do, R.K.G., Togashi, K.: Convolutional neural networks: an overview and application in radiology. Insights into imaging **9**(4), 611–629 (2018)
58. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering (2016)
59. Zhao, P., Zhang, J., Fang, W., Deng, S.: SCAU-Net: Spatial-Channel Attention U-Net for Gland Segmentation. Frontiers in Bioengineering and Biotechnology **8**, 670 (2020). https://doi.org/10.3389/fbioe.2020.00670, https://www.frontiersin.org/article/10.3389/fbioe.2020.00670
60. Zhou, Z., He, Z., Jia, Y.: Afpnet: A 3d fully convolutional neural network with atrous-convolution feature pyramid for brain tumor segmentation via MRI images. Neurocomputing **402**, 235–244 (2020). https://doi.org/10.1016/j.neucom.2020.03.097, https://doi.org/10.1016/j.neucom.2020.03.097

Appendix A

Image Segmentation Output

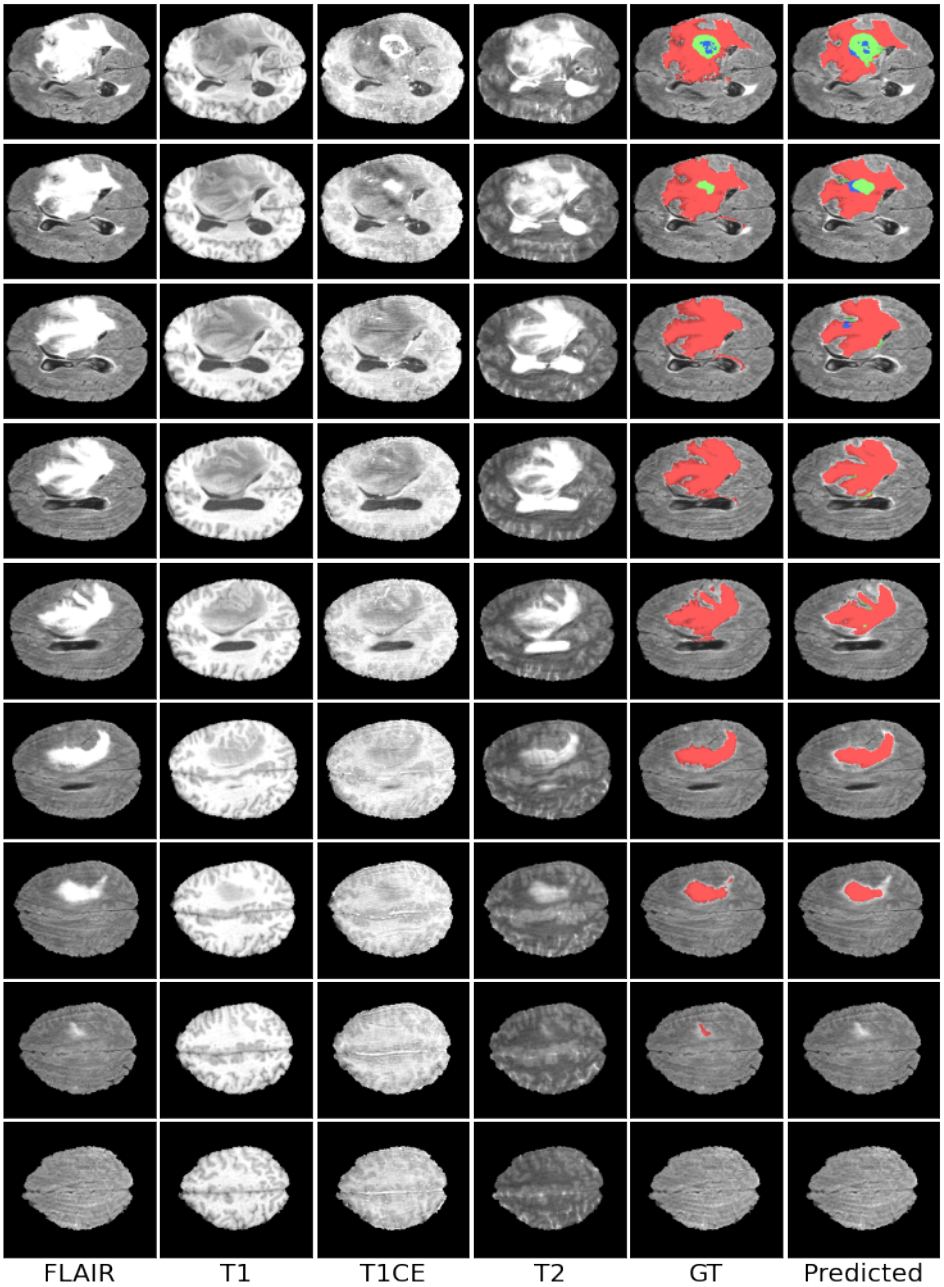


Fig. 22. UNet: Segmentation results for BraTS 2020 training dataset, from left to right : FLAIR, T1, T1CE, T2, Predicted; Colors: Necrotic and Non-enhancing tumour Core(Blue), Peritumoural Edema(Red), GD-Enhancing tumour(Green). For this instance, the plots are from patient id 001.

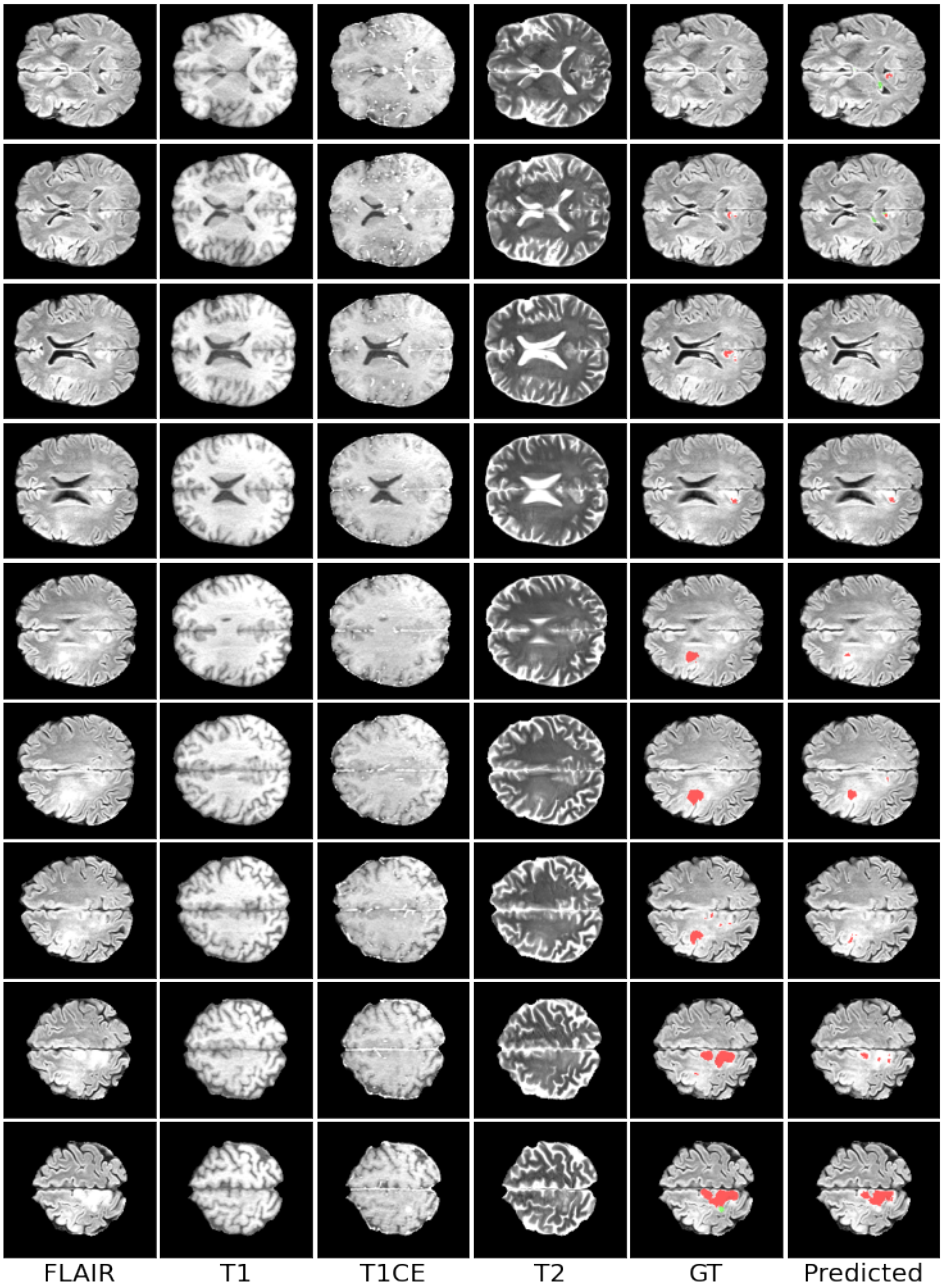


Fig. 23. UNet: Segmentation results for BraTS 2020 training dataset, from left to right : FLAIR, T1, T1CE, T2, Predicted; Colors: Necrotic and Non-enhancing tumour Core(Blue), Peritumoural Edema(Red), GD-Enhancing tumour(Green). For this instance, the plots are from patient id 110.

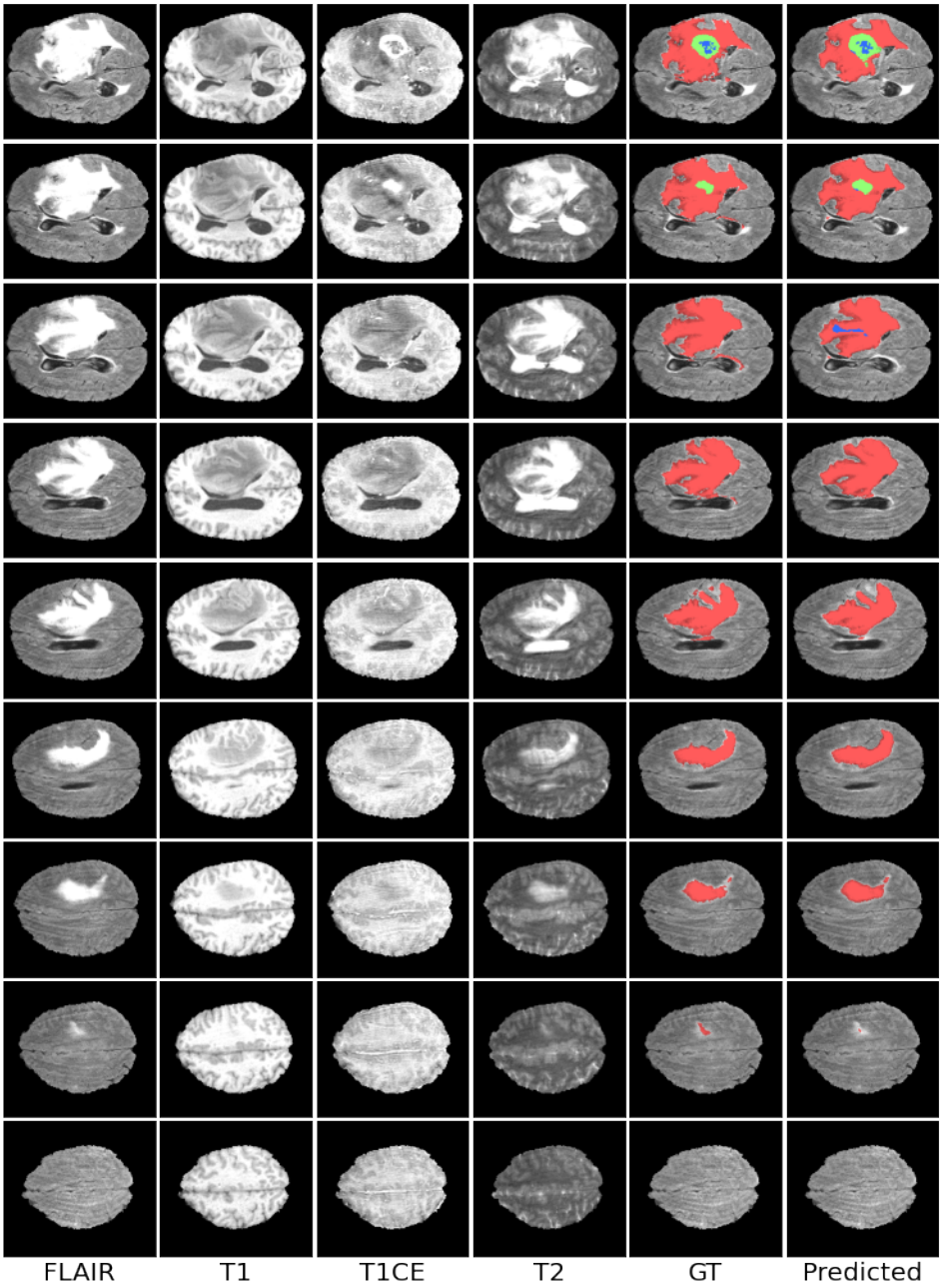


Fig. 24. DAUNet: Segmentation results for BraTS 2020 training dataset, from left to right : FLAIR, T1, T1CE, T2, Predicted; Colors: Necrotic and Non-enhancing tumour Core(Blue), Peritumoural Edema(Red), GD-Enhancing tumour(Green). For this instance, the plots are from patient id 001.

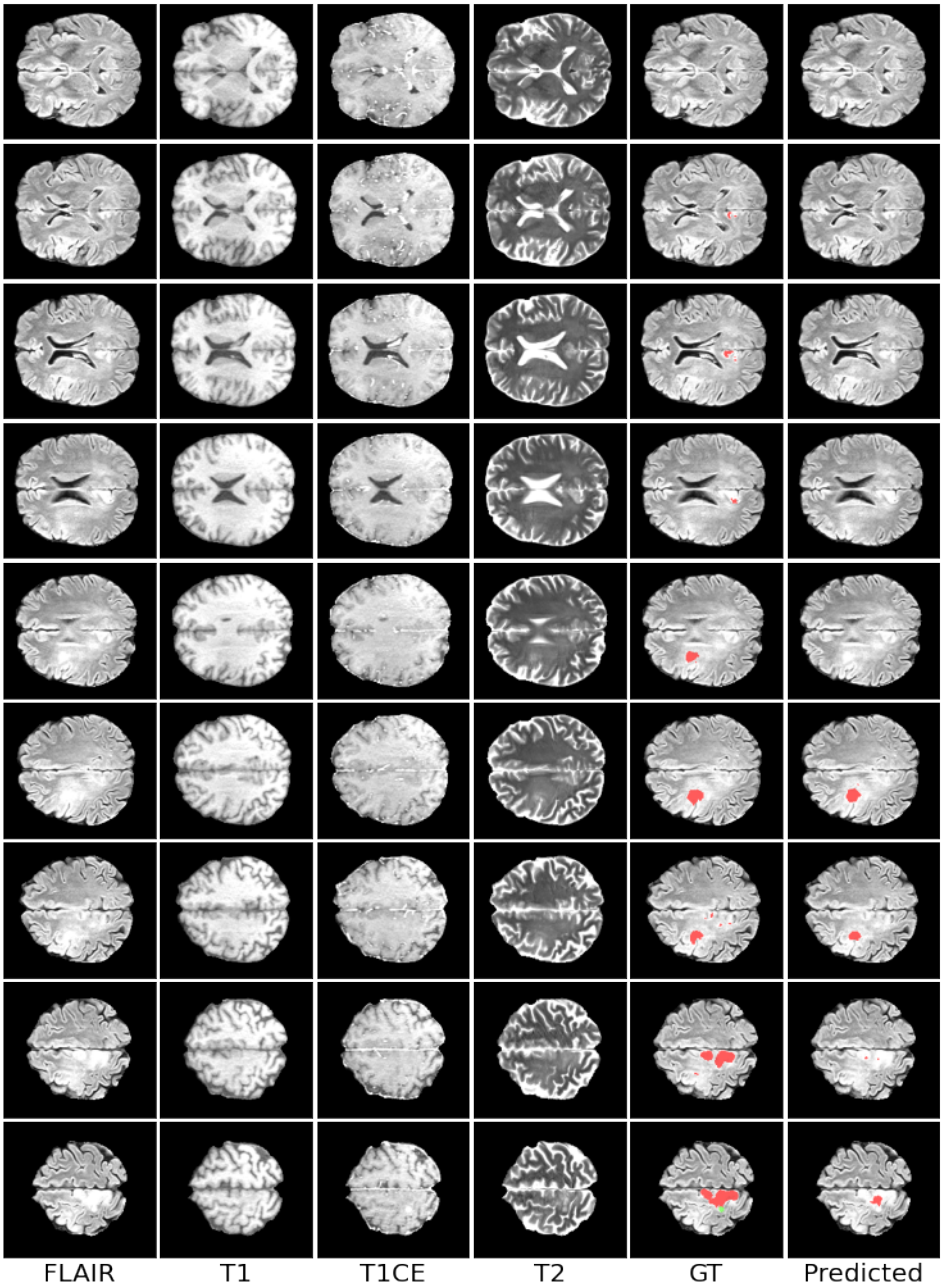


Fig. 25. DAUNet: Segmentation results for BraTS 2020 training dataset, from left to right : FLAIR, T1, T1CE, T2, Predicted; Colors: Necrotic and Non-enhancing tumour Core(Blue), Peritumoural Edema(Red), GD-Enhancing tumour(Green). For this instance, the plots are from patient id 110.

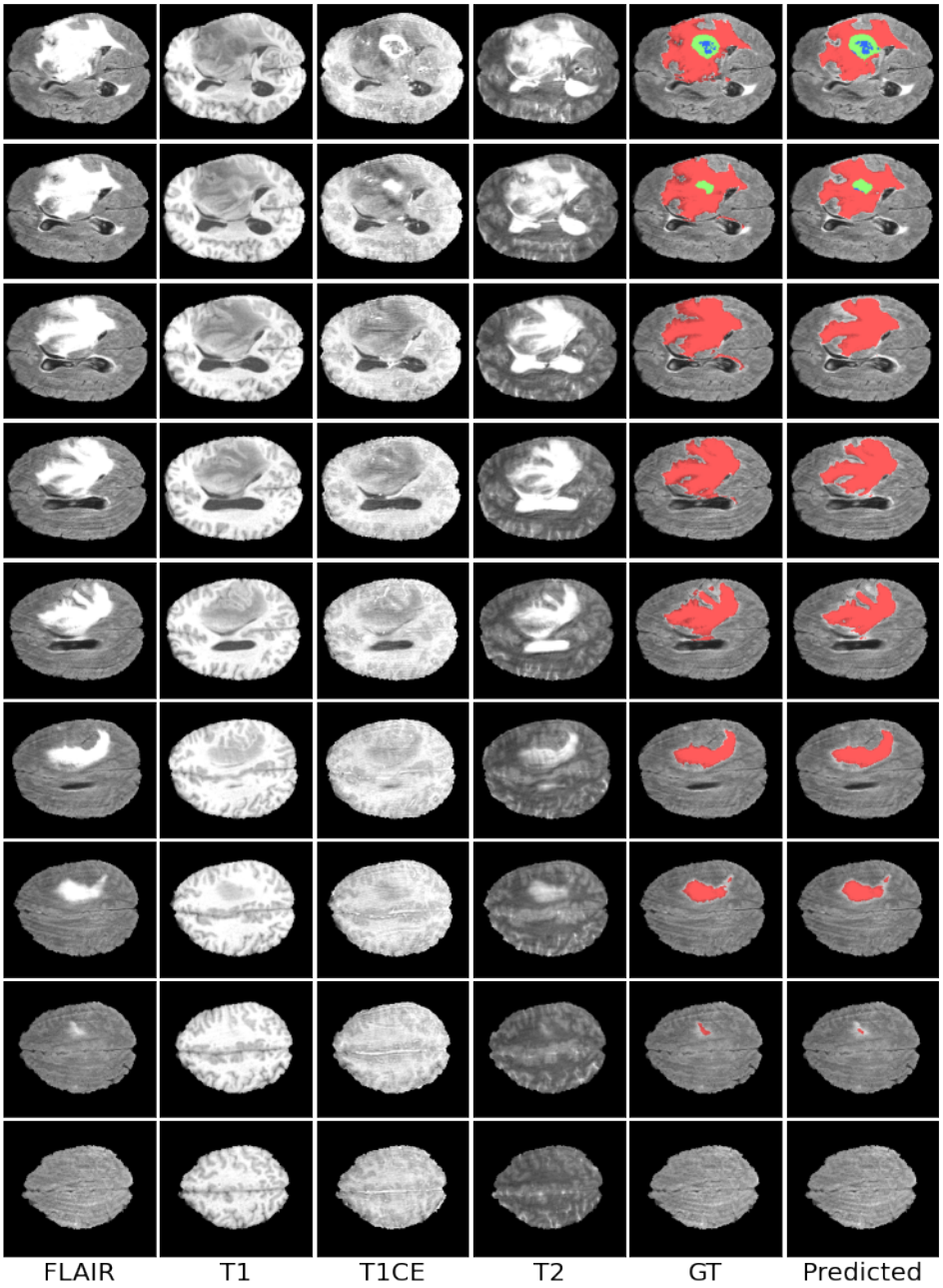


Fig. 26. SGANet: Segmentation results for BraTS 2020 training dataset, from left to right : FLAIR, T1, T1CE, T2, Predicted; Colors: Necrotic and Non-enhancing tumour Core(Blue), Peritumoural Edema(Red), GD-Enhancing tumour(Green). For this instance, the plots are from patient id 001.

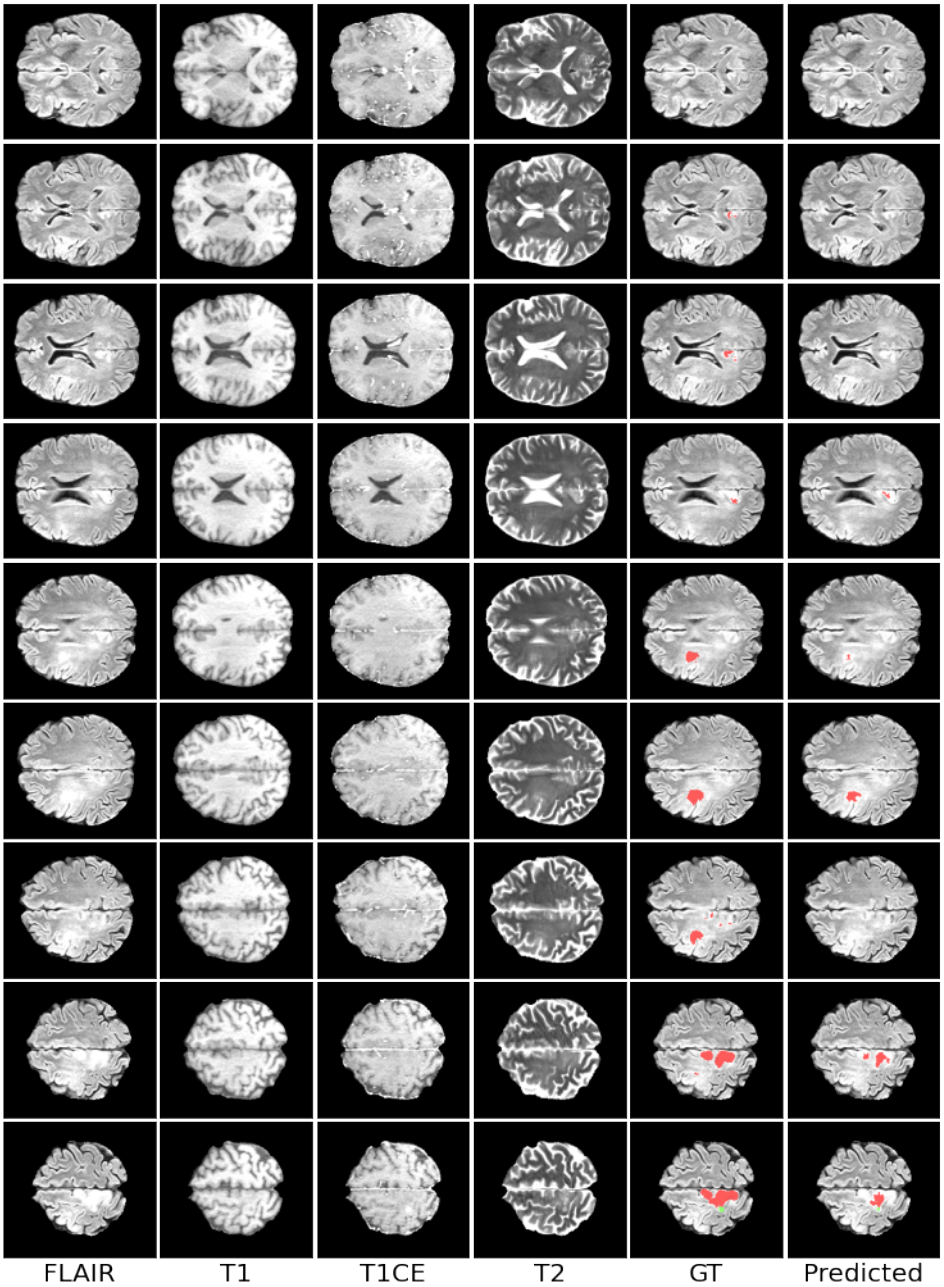


Fig. 27. SGANet: Segmentation results for BraTS 2020 training dataset, from left to right : FLAIR, T1, T1CE, T2, Predicted; Colors: Necrotic and Non-enhancing tumour Core(Blue), Peritumoural Edema(Red), GD-Enhancing tumour(Green). For this instance, the plots are from patient id 110.

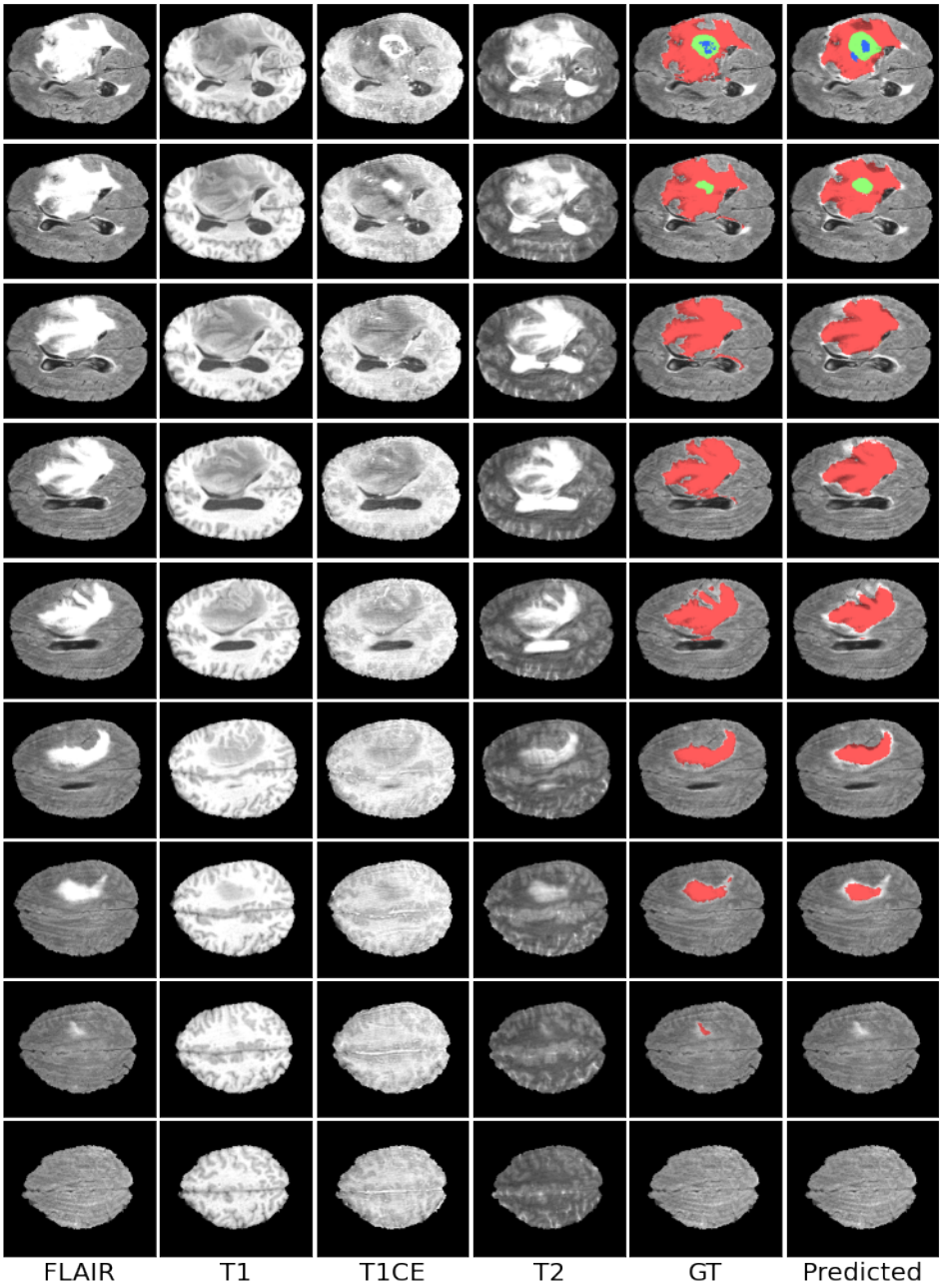


Fig. 28. 3D-DAUNet: Segmentation results for BraTS 2020 training dataset, from left to right : FLAIR, T1, T1CE, T2, Predicted; Colors: Necrotic and Non-enhancing tumour Core(Blue), Peritumoural Edema(Red), GD-Enhancing tumour(Green). For this instance, the plots are from patient id 001.

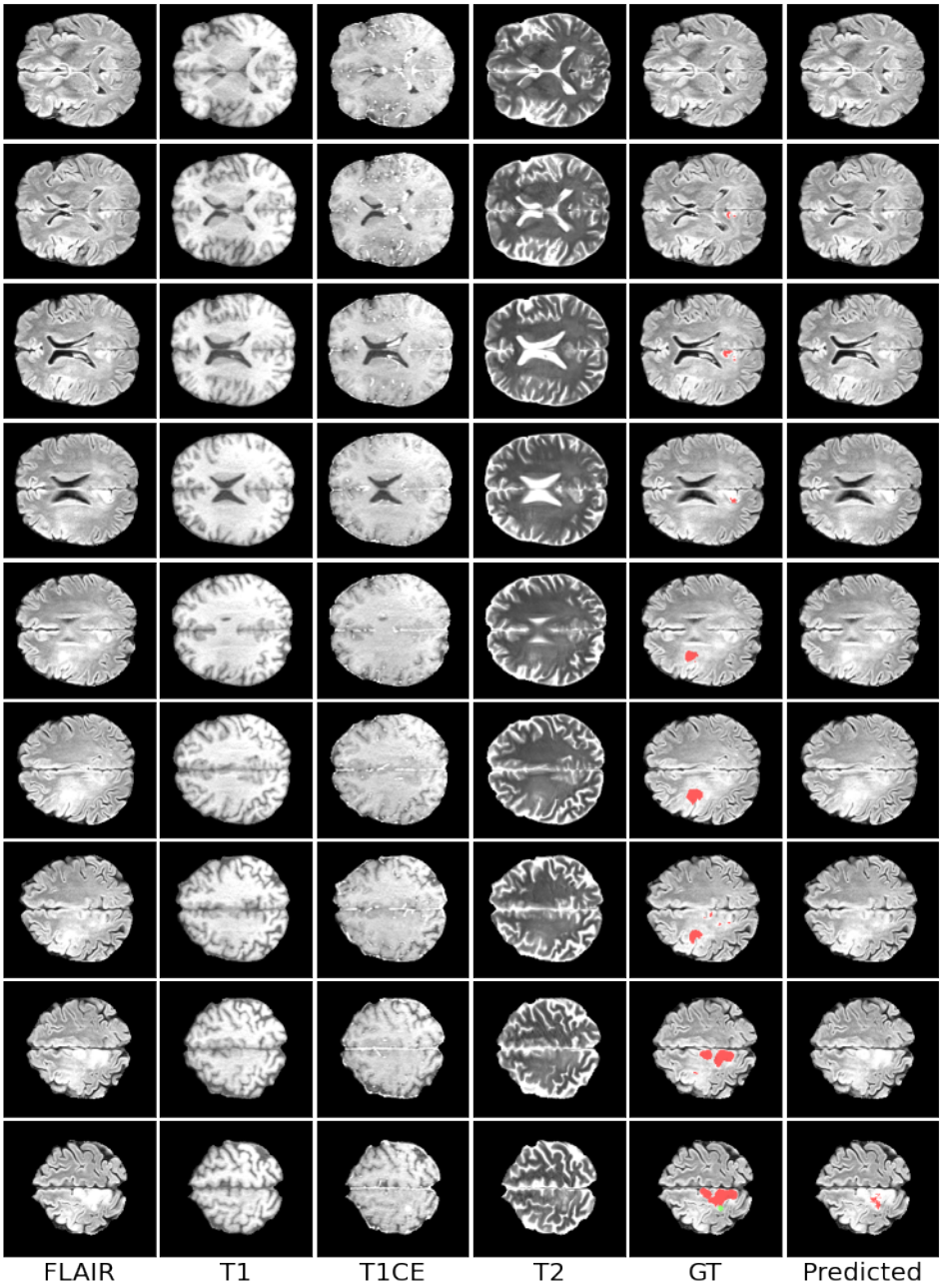


Fig. 29. 3D-DAUNet: Segmentation results for BraTS 2020 training dataset, from left to right : FLAIR, T1, T1CE, T2, Predicted; Colors: Necrotic and Non-enhancing tumour Core(Blue), Peritumoural Edema(Red), GD-Enhancing tumour(Green). For this instance, the plots are from patient id 110.

Appendix B

PAM Feature Maps

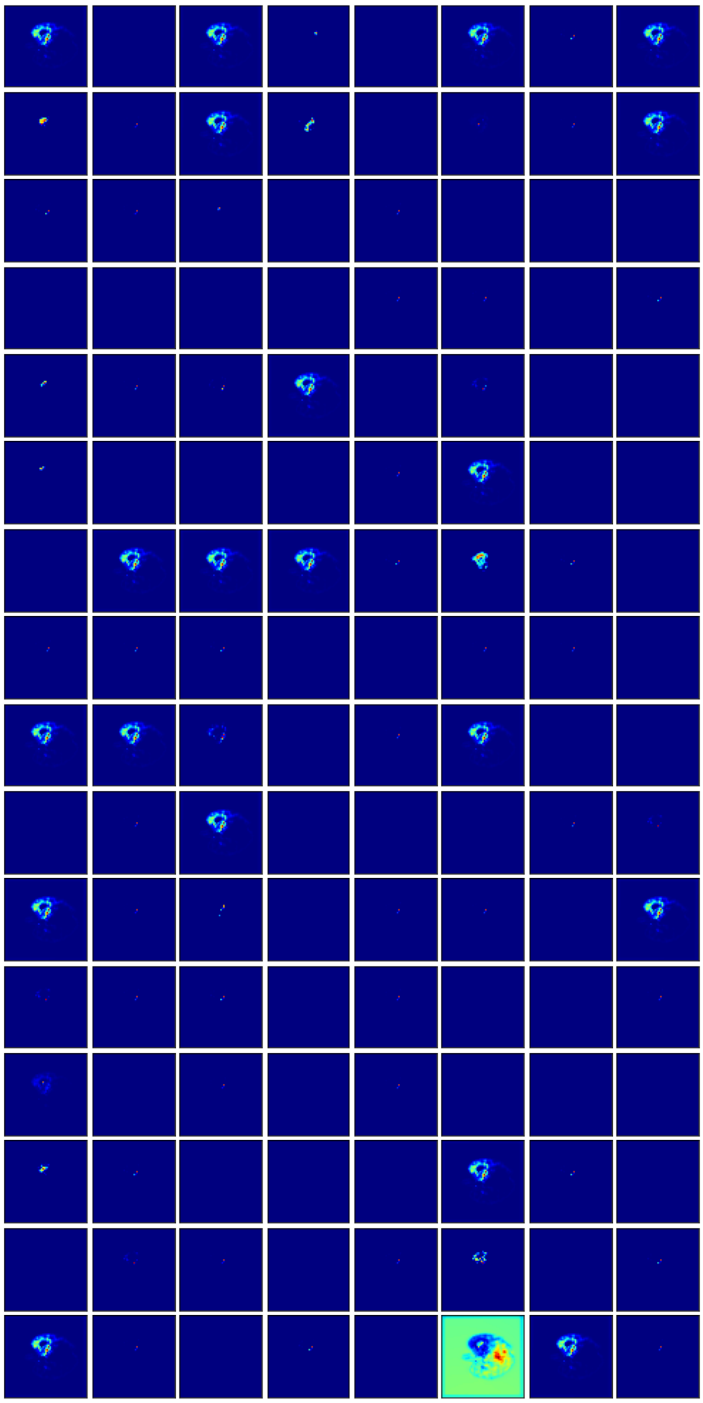


Fig. 30. PAM00 feature maps (leftmost).

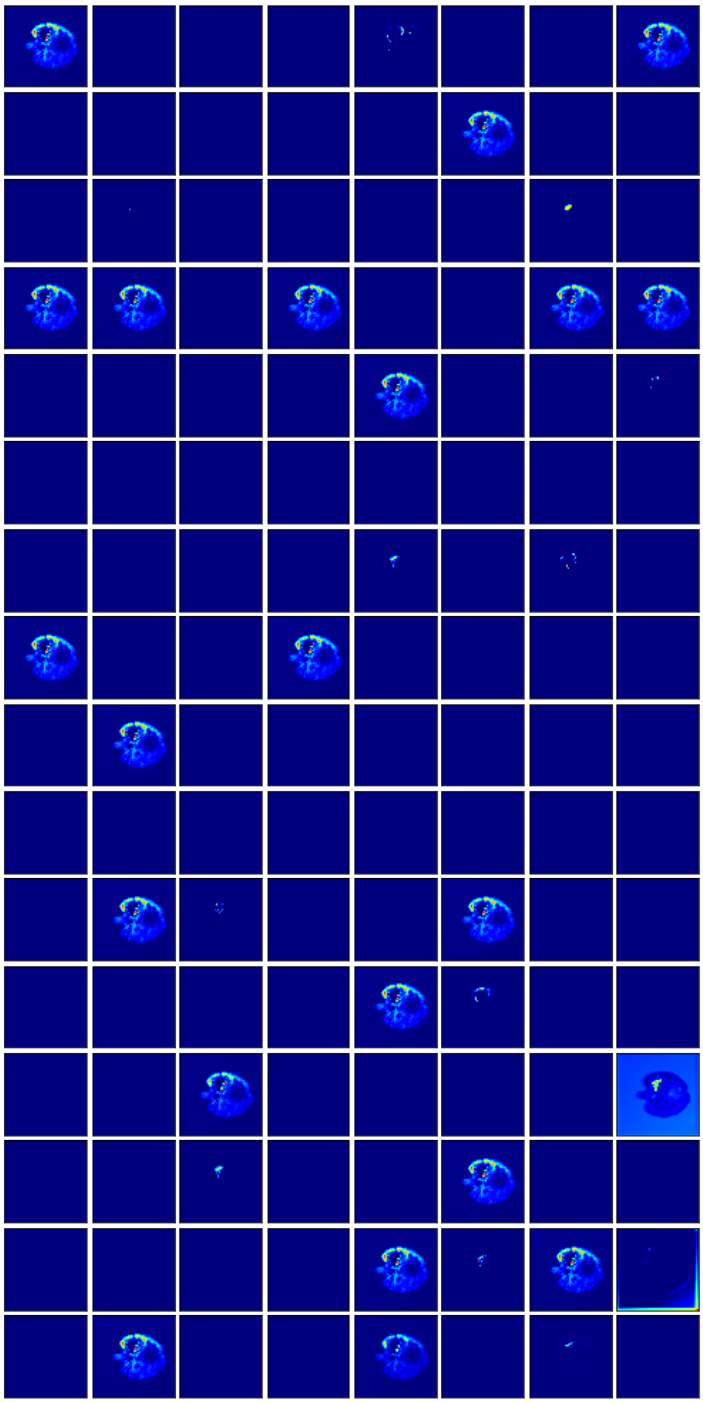


Fig. 31. PAM01 feature maps (mid).

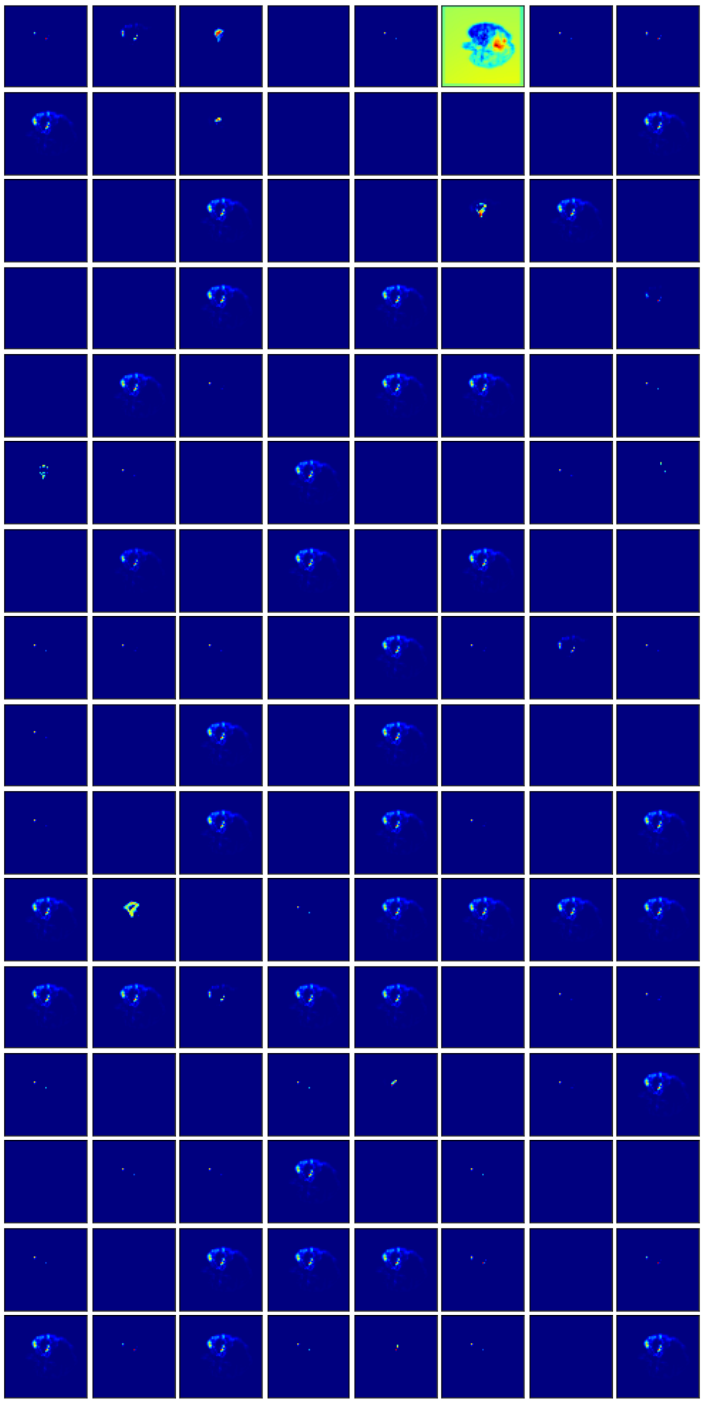


Fig. 32. PAM02 feature maps (rightmost).

Appendix C

CAM Feature Maps

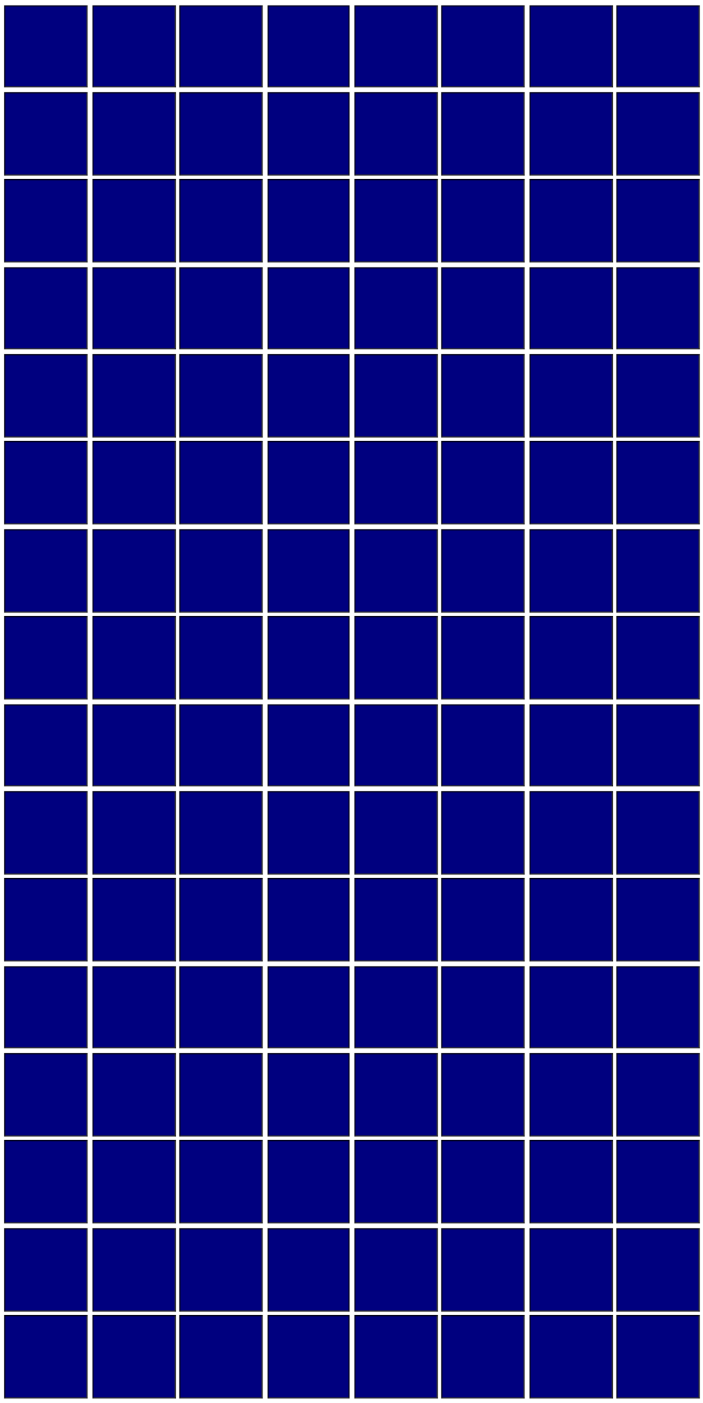


Fig. 33. CAM00 feature maps (leftmost).

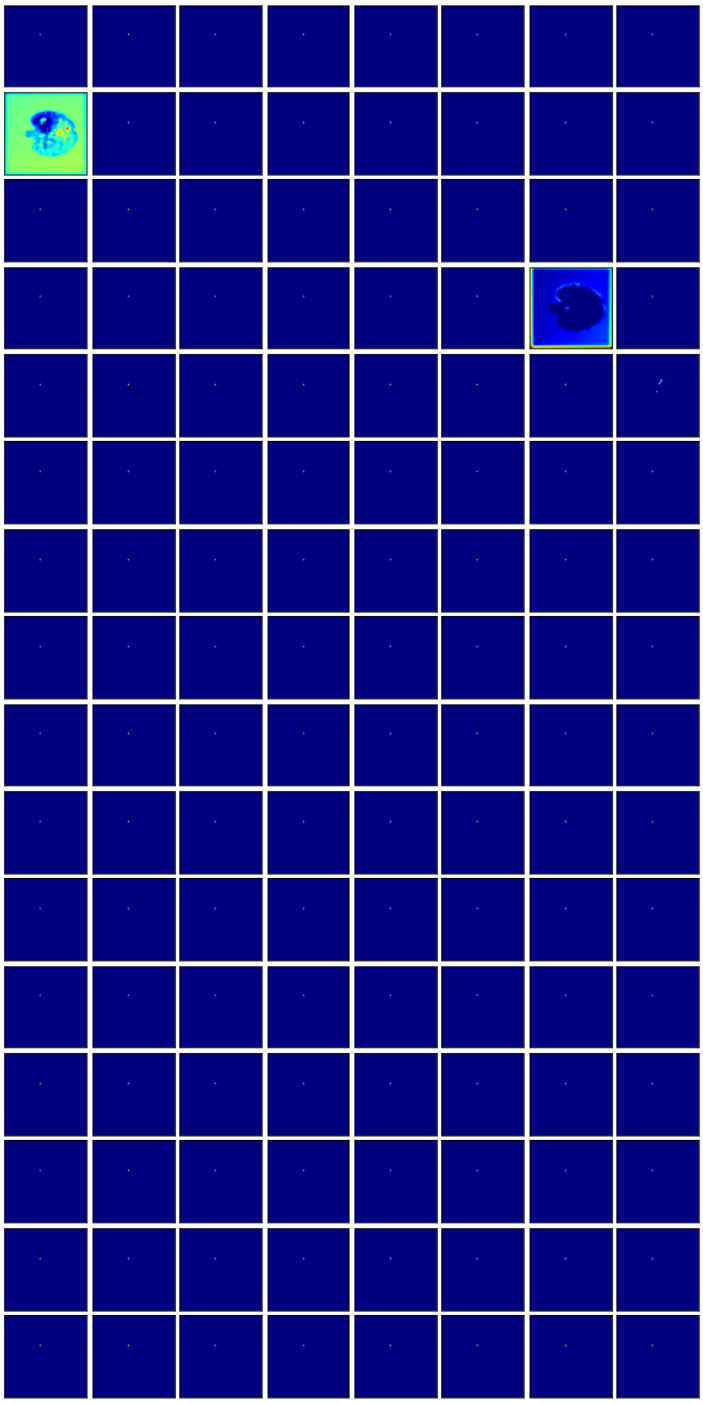


Fig. 34. CAM01 feature maps (mid).

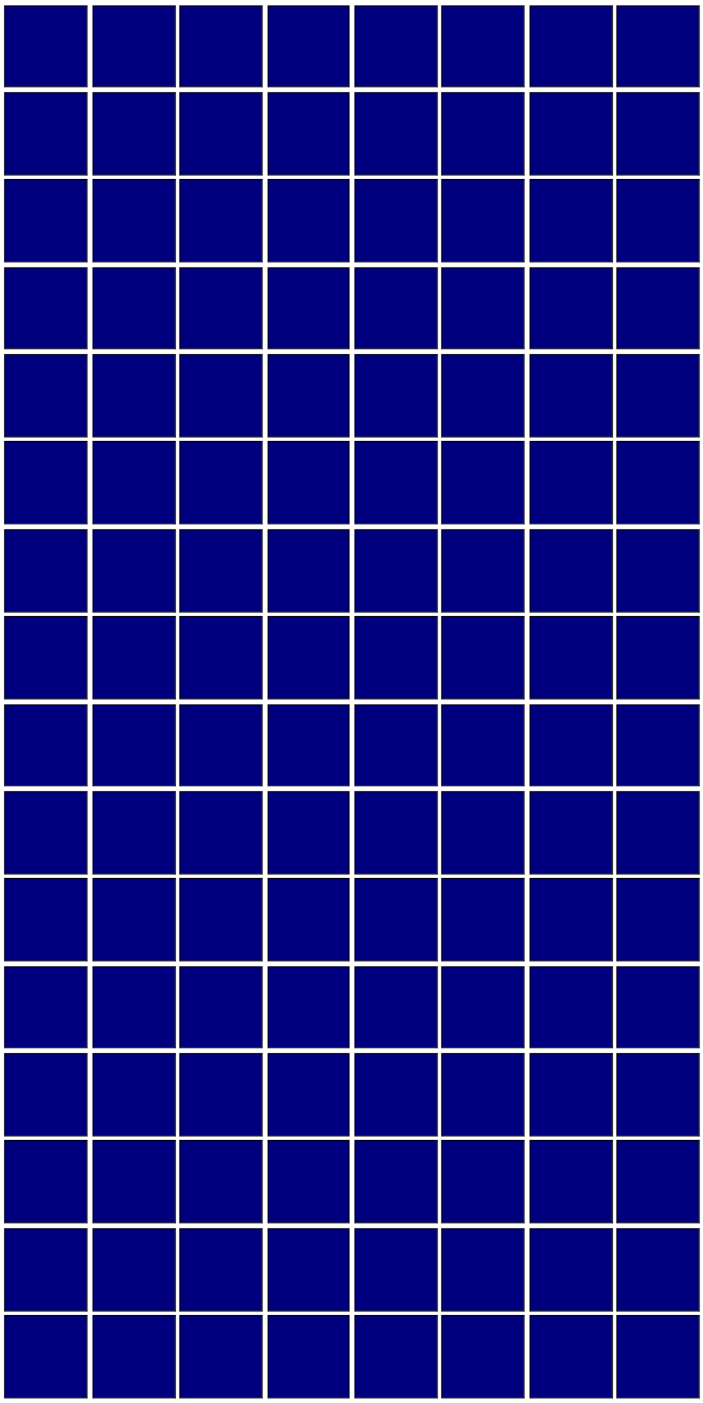


Fig. 35. CAM02 feature maps (rightmost).