AW-Net: A Novel Fully Connected Attention-based Medical Image Segmentation Model

Debojyoti Pal[1], Tanushree Meena[1], Dwarikanath Mahapatra[2], Sudipta Roy[1,*]

[1]Artificial Intelligence & Data Science, Jio Institute, Navi Mumbai-410206, India
[2]Inception Institute of AI (IIAI), UAE

debojyoti.pal@jioinstitute.edu.in; tanushree.meena@jioinstitute.edu.in;
dmahapatra@gmail.com; sudipta1.roy@jioinstitute.edu.in;

Abstract

Multimodal medical imaging poses a unique challenge to the data scientist looking at that data, since it is not only voluminous, but also extremely heterogenous. In this paper, we have proposed a novel fully connected AW-Net which provides a solution to problem of segmenting multimodal 3D/4D medical images by incorporating a novel regularized transient block. The AW-Net uses the concept of stacking of consecutive 2D image slices to extract spatial information for segmentation. Furthermore, dropout layers are incorporated to reduce the computational cost without affecting the accuracy of the output predicted masks. The AW-Net has been tested on benchmark datasets such as BRATS2020 for brain MRI, RSNA2022 cervical spine dataset for spine CT followed by DUKE and QIN dataset for breast MRI and PET respectively. The AW-Net achieves a Dice similarity coefficient (DSC) of 81.3% and 80.5% for breast cancer segmentation from DCE and T1 images, 89.6% as an average of three segmented tumor classes for brain tumor segmentation from BraTS2020 dataset, 93.7% for breast tumor segmentation from breast PET images, and 71.9% for cervical fracture localization on the RSNA 2022 challenge. These evaluation experiments performed on public datasets indicate that the proposed AW-Net is a generalized, reproducible, efficient, and highly accurate model capable of segmenting and localizing anomalies in any multi-modal 3D/4D medical imaging data from small and large data sets. The GitHub link is available at: https://github.com/Dynamo13/AW-Net.

1. Introduction

The last few decades have seen an exponential rise in the usage of multimodal medical imaging. The consequent availability of high-quality data from these imaging modalities has resulted in the rapid development of deep learning (DL) based models[29], [22] in the past two decades. Nevertheless, addressing the reproducibility of deep learning-based models on diverse datasets, especially when achieving highly accurate segmentation masks, is essential prior to their clinical application. The task of detecting and segmenting intricate anomalies such as tumors and target lesions proves to be difficult owing to the diverse nature of these target tissues, coupled with distinct characteristics of multimodal images. As a result, obtaining a substantial number of accurate segmentation masks for 3D datasets is not only difficult, but also time consuming and error prone. In addition, different image modalities undergo different pre-processing techniques and have a wide variation of voxel intensities. Therefore, the development of a single, consistent, and reproducible model for multiorgan segmentation from multimodal imaging is extremely important. The aim of this paper is to develop a generalized segmentation model which provides accurate segmentation masks for multiple regions of interest regardless of the volume of training data. Attention U-Net [23] has shown impressive performance on many benchmark medical datasets but its ability to generalize complex medical tasks such as breast cancer segmentation is still limited. Moreover, the attention mechanism tends to over-fit on complex dataset due to lack of the regularization techniques in the attention mechanism. Therefore a novel, generalized, attention-based segmentation model, namely AW-Net has been proposed to accomplish the task multiorgan segmentation from multimodal medical images. The contribution of the proposed model is summarized as follows:

- A novel regularized transient block (RTB), comprising of regularised convolutional blocks and dropout layers, is introduced between the encoder and decoder path-

ways for better infusion of low-level and high-level feature maps.

- A study of the computational cost of the model has been performed to showcase that the AW-Net is the most cost-effective model among other state-of-the-art models.

- Experiments on multimodal datasets (CT, PET, MRI) have been conducted to validate the effectiveness of the model for segmentation and anomaly localisation of multiple organs.

2. Related Work

2.1. 2D segmentation models

The spatial details contained within the voxels of various imaging modalities significantly impact the quality of the segmented mask. In medical imaging, the accuracy and precision of these segmented masks hold utmost importance, as numerous subsequent decisions rely upon the precise locations provided by these annotations. Fully convolutional models [5], [4] were proposed keeping in mind the challenging nature of this problem. These networks either have stacked a series of convolutional layers together or have used state-of-the art models such as VGG, as a backbone. However, these models use a considerable amount of training data and don't address the problem of segmenting small lesions. To address the problem of misclassification of neighboring pixels, squeeze and excitation blocks [28] were incorporated in different state-of-the-art architectures. However, the squeeze and excitation blocks increased the time-complexity of the model. In addition, the usage of large volume of training data is still an unaddressed problem. Since segmenting anomalies from complex medical images by radiologists are a complicated as well as time-consuming task, it is difficult to obtain a large volume of annotated/segmented masks for training. As a result, these models were not used in real-life scenarios. The U-Net [27] was developed to address the problem of over-fitting and usage of large training data to generate segmentation masks. Over the years, several modifications [33] [37] [16] have been made on the U-Net architecture to achieve precise and accurate results for small lesion segmentation and anomaly localization on multiple medical imaging modalities. One of the most prominent modification is the usage of attention gates [23] [24] and transformers [25] [7] on the baseline U-Net architecture. Attention gates reinforce the model to learn intrinsic anatomical details of the target lesions by representing the extracted features in a higher dimensional space without increasing the model complexity. The incorporation of transformers results in faster convergence of the model by focusing on the regions which are missed by the convolutional neural network layers due to the large kernel

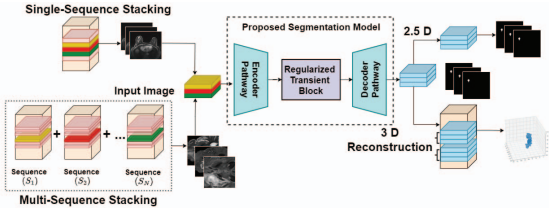

Figure 1: The workflow diagram of the proposed AW-Net.

size of the incorporated layers. Recent developments include the addition of meta-heuristic algorithms such as [26] [35] [10] in DL models to optimize the hyper-parameters involved in the training.

2.2. 2.5D and 3D segmentation models

The 2D models lack the capability to capture the supplementary spatial insights inherent in 3D medical imaging modalities like CT scans, MRI, and PET scans. These imaging techniques consist of series of successive 2D slices stacked together, with each slice containing data about the examined tissue. 3D models [12] [13] facilitate the extraction of the geometry of interest through the process of surface determination. These models process a series of contiguous 3D blocks as input and perform voxel segmentation to provide 3D segmentation masks which permits better representation and quantification of generated segmentation masks for multi-modal images. However, these models take a significant amount of data for training. The computational cost also increases due to the usage of 3D convolution filters. The introduction of attention-based 3D U-Net [8] improved the accuracy of the predicted masks by reducing the number of predicted false positive pixels. The introduction of transformer-based U-Net variants such as 3D Swin U-Net [31] also increased the computational complexity, resulting in the requirement of high-performance computers to train the models. As a result, pseudo 3D models [20] [19] and 2.5D models [36] [11] came into practice. Pseudo 3D models use three individual 2D slices in different orientation planes namely sagittal, axial, and coronal and reconstruct the 3D volumetric image which was then fed into the model. It solved the problem of data availability, but the computational cost was still high. Stacked 2.5D models [17] [32] solved this problem by stacking neighboring 2D slices to obtain spatial information. However, unlike 3D models, only 3 to 5 consecutive images were stacked. These models considered individual 2D slices and performed pixel-level segmentation. These DL based segmentation models have proven to provide better segmentation masks than radiologists in some cases [21].

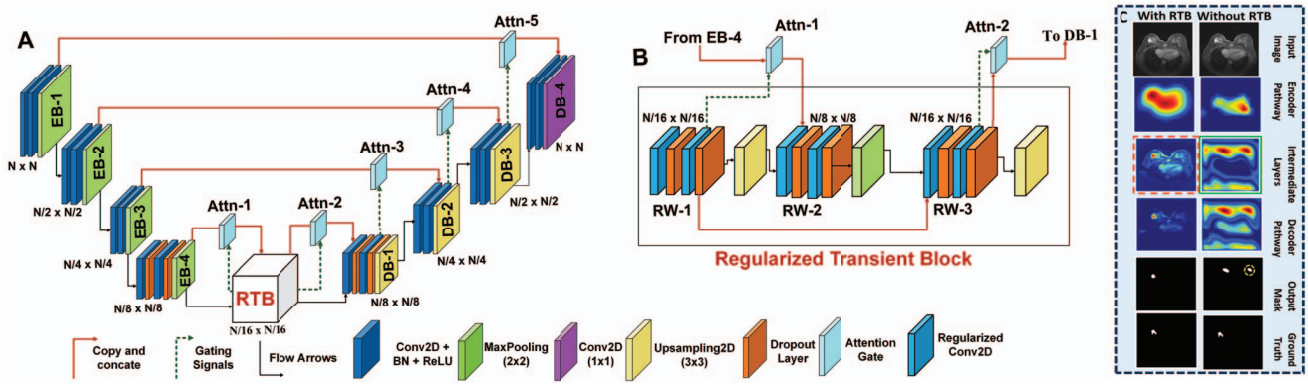

Figure 2: A) The AW-Net architecture for medical image segmentation. The 3D rectangular boxes represent different layers, red solid arrows represent the skip connections and the novel RTB is highlighted in red. B) The inner structure of the RTB block with regularized convolutional block. C) Feature maps showing the effectiveness of RTB (shown as red dotted lines) and the role it plays for the reduction of false positive pixels (shown as yellow dotted circles).

3. Methodology

3.1. Stacked 2.5D slices

The proposed AW-Net has been developed to segment 3D medical images. However, feeding 3D images traditionally results in increased computational cost. A 2.5D stacking algorithm is proposed in this section which takes an entire 3D volume as an input and sequentially processes three slices at a time, with the target slice being the central slice, x_{target} sandwiched between two adjacent ones, x_{left} and x_{right} as illustrated in Figure 1 and formulated in equation 1. Unlike 2.5D stacked models that only consider relevant slices containing the region of interest, this model considers all slices regardless of voxel content. In this paper, we are dealing with multi-modal data i.e., CT, MRI, and PET as well as multi-sequence data i.e., DCE, T1, and FLAIR to name a few. In order to efficiently deal with the variability of the data, the proposed stacking algorithm has two modes: multi-sequence and single-sequence. In a dataset with multiple sequences, like the BraTS2020 for brain MRI, the multi-sequence mode is utilized. This mode processes one slice from each sequence in the dataset and inputs the concatenation of three sequences stacked together for the same position into the model. This process is repeated until the entire volume of data is processed. Conversely, the single-sequence mode would be used for tasks like breast tumor segmentation from the DUKE dataset, where we want to focus on segmenting tumors based on a single sequence. In this scenario, we would concatenate three consecutive slices of information at a time from the entire 3D dataset and feed it to the model sequentially. By leveraging the 3D nature of different data modalities, the stacking approach exploits spatial features from adjacent slides. Although the individual constituent layers are 2D, the stacking method provides efficient training and enables the model to learn the spatial features from adjacent slides.

3.2. Proposed Network

The proposed novel fully connected network, namely AW-Net, is shown in Figure 2. The AW-Net is a modified verision of the attention U-Net architecture which incorporates a novel regularized transient block. The proposed model has A) encoding pathway, B) RTB block C) decoding pathway. Unlike the U-Net architecture, the encoding pathway doesn't connect with the decoding pathway. Instead, the final block of the encoding pathway (EB-4) is connected to the novel RTB. The RTB in conjunction with attention gates is responsible for regularizing and fine tuning the encoded feature maps. This not only prevents overfitting of the model but also forces the model to selectively focus on relevant features. The residual connections in the RTB, combines coarser feature maps with finer feature maps.This allows the network to leverage both low-level and high-level features for segmentation, which can lead to more accurate and robust segmentation results. Attention gates are added to make the model more robust and accurate for small lesion segmentation as shown in Figure 2. The input-output function of the proposed network is represented as:

$$y_{output} = f_{\text{AW-Net}}(x_{left}, x_{target}, x_{right}|\theta) \qquad (1)$$

where y_{output} is the output of the target, $f_{\text{AW-Net}}$ is the proposed model, θ is the model parameter, x_{target} is the central target slice, x_{left} and x_{right} represents the neighboring slices.

3.3. Encoding Pathway

The encoding pathway comprises of four fully connected encoder block connected by the max-pooling layer. The

first three blocks comprises of two convolution layers with a 3×3 filter size stacked together. Each convolution layers are followed by a ReLu activation and a Batch Normalization operation. These convolution layers facilitate the extraction of important features in subsequent blocks. The feature map of the convolution layers increases by a factor of 2 for each subsequent layer from 32 to 256. The resultant neuron n_{conv} after undergoing a dense convolution operation can be represented as:

$$conv(x) = max(0, \sum_{i=1}^{d} w_i x_i + b) \tag{2}$$

where $x \in R^{1\times d}$ represents the individual weights of a d-dimensional input vector, $w \in R^{1\times d}$ represents the weight vector and b represents the bias. An encoder block is connected to the following layer with a max-pooling layer with filter size of 2×2. The final encoder block, however, has a dropout layer as an additional layer to individual convolution layers having a dropout rate of 0.2. It not only prevents the model from over-fitting but also prevents the misclassification of pixels.

3.4. Regularized Transient Block

One of the major problems of attention U-Net is the bottleneck. Bottleneck refers to the narrowest part of the network, where the spatial resolution of the feature maps is reduced. The reduced spatial resolution leads to loss of fine-grained details important for accurate segmentation. The bottleneck of the U-Net architecture is characterized by an increased number of channels, resulting in feature-rich vectors. However, this can lead to two issues: first, an elevated risk of overfitting, and second, an increase in the computational complexity of the model. To solve these problems, we propose a fully connected layer, named regularised transient block (RTB) as a replacement of the bottleneck, shown in Figure 2B. The RTB is comprised of three regularized weight layers, RW-1, RW-2, and RW-3 connected with each other. Individual weight layers comprise of a series of convolutional layer, ReLU activation function followed by a batch normalization operation. Unlike the convolution operations mentioned in Section 3.2, we have used L1 regularizer on the convolutional filter weights shown in equation 3.

$$\theta_{L1} = \gamma \sum_{i=1}^{d} |w_i| \tag{3}$$

$$L_\theta = -\sum_{i=1}^{n} y_i^{GT} log(p_i) + \theta_{L1} \tag{4}$$

where γ is the regularization strength hyperparameter, θ_{L1} is the regularization penalty term, n represents the number of classes, y^{GT} represents the truth label and p_i represents the softmax probability for the i^{th} class. This technique reduces the complexity of the proposed model by adding a penalty term to the loss function that encourages small absolute values for the model weights. The L1 regularization penalty term is added to the cross-entropy loss to form the regularized loss function L_θ, shown in equation 4, which is then minimized during training using an optimization algorithm, Adam. The L1 regularization penalty also optimizes the number of parameters of the model, θ. The regularised convolution operation is followed by a dropout layer having a rate of 0.2. This helps in reducing overfitting and improving the generalization performance of the model.

The encoded feature vector from EB-4 is passed onto RW-1 of the RTB. The regularised feature vector is down-sampled via a max-pooling layer. The resultant vector is simultaneously processed to both RW-2 and attention layer Attn-1. Attn-1 generates an attention co-efficient, α (see equation 6) which can identify salient features and prune irrelevant features which act as noise to the prediction task. This co-efficient is again reintroduced into RW-2. The RTB increases the model depth, thus making it prone to suffer from the vanishing gradient problem. To solve this problem, a skip connection is introduced between RW-1 and RW-3. However, the down sampling operation results in a shape mismatch between the aforementioned regularised weight blocks. As a result, the feature vectors obtained from RW-2 undergoes an up sampling operation. The skip connection allows the model to combine low-level features (RW-1) with high level features (RW-3) making the model more robust. The output of the RTB serves as an input vector for Attn-2. The RTB is finally connected to the decoder pathway via DB-1.

3.5. Decoding Pathway

The decoder pathway consists of four fully connected decoder blocks connected via an up-sampling layer. Each decoder block has an up-sampled vector and an attention vector as an input. The attention vector is basically an output of the attention gate, which is discussed in section 3.6, whereas the up-sampled vector is the output of a transverse convolution layer of filter size 2×2. Both these vectors are concatenated and subsequently serve as an input vector for a series of two convolution layer stacked on top of each other. However, the first decoder block DB-1, comprises of an addition dropout layer after each convolution layer to prevent the model from over-fitting.

3.6. Attention Gate

Convolution blocks provide semantic information for sufficiently large area of interest. However, when it comes to distinguishing smaller regions of interest with high variability as in case of tumor detection, the spatial information of corresponding voxels plays an important role. The pro-

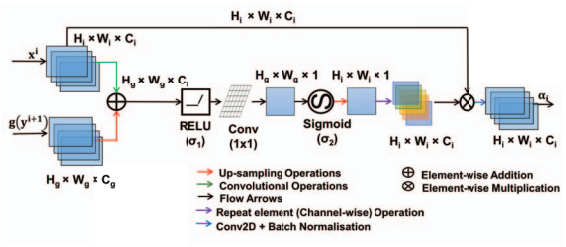

Figure 3: The diagrammatic representation of the attention gate.

posed model, therefore, uses attention gates for better localization of smaller objects. The attention gates make the model learn the important features, voxel-wise, in addition to suppressing irrelevant background pixels. As a result, the model accuracy increases due to the reduction of false positive voxels. The stacking of consecutive image slices helps the attention gates in performing better localization of small target lesions by providing spatial information of the surrounding slices as shown in Figure 3. The attention gates take two vectors as input: a gating-signal vector, x^G and an input vector, x^I. The gating vector x^G contains contextual information which makes the model learn the focus regions on a subset of target tissues. The feature vector x^I and x^G is obtained via skip-connections and gating signals respectively. The gating vector is passed through an upsampling layer whereas the input vector is passed through a convolutional layer of filter size 3×3. The processed vectors are then added and passed onto a ReLU activation function (σ_1). The resultant vector undergoes a linear transformation operation using a convolutional layer of kernel size 1×1 to generate intermediate activation maps q_A. The generated activation maps are then passed through a sigmoid activation function (σ_2). The attention gate is formulated as follows:

$$q_A = \psi(\sigma_1(\sum_{i=1} w_i^I x_i^I + \sum_{j=1}^{n} (x_j^G)^T x_j^G) + b_\sigma) + b_\psi \quad (5)$$

$$\alpha_i = \sigma_2(q_A(x^I, x^G|\theta_A)) \quad (6)$$

where σ_2 corresponds to sigmoid activation function, σ_1 corresponds to the ReLU activation function, θ_A represents the set of parameters for attention gating and ψ represents the linear transformation operation.

3.7. Dataset Description

We evaluated the proposed AW-Net on benchmark datasets such as Duke Breast Cancer [30], QIN (Quantitative Imaging Network) breast data set [18], BraTS2020 dataset [3] and RSNA cervical spine fracture detection challenge [1]. The Duke Breast Cancer consists of DCE (Dynamic Contrast Enhanced) MRI images and T1 MRI images of 529 subjects is selected as breast MRI dataset. The QIN breast data set is used as breast PET dataset. The BraTS2020 dataset containing annotations of enhancing tumor (ET), peritumoral edema (ED) and the necrotic and non-enhancing tumor core (NCR/NET) is considered as brain MRI dataset. Individual multimodal scans are available as NifTI files with a native (T1), post-contrast T1-weighted, T2-weighted and T2 Fluid Attenuated Inversion Recovery (T2-FLAIR) volumes. The RSNA cervical spine fracture detection challenge consists of 1400 CT studies with equal negative studies are considered as a dataset for spine fracture localization. Tumor regions for breast MRI and PET are manually annotated using MicroDICOM viewer and exported as DICOM files. The dataset details are summarized in Table 1.

Table 1: A summarized table of the datasets used.

Dataset	No. of Images	Input Size	Modality
Duke	320	448×448×3	MRI (DCE, T1)
QIN	240	512×512×3	PET
BraTS	600	256×256×3	MRI (T1w, T2, FLAIR)
RSNA	1400	256×256×3	CT

3.8. Implementation Details

The AW-Net is developed in TensorFlow 2.0 and trained using an Intel RTX A4000 chip and an Intel i9 processor. Adam optimizer is used with a learning rate of 0.001 and a decay rate of 0.89. The model is trained for 150 epochs with a batch size of 8. Binary crossentropy is used as a loss function to obtain binary segmentation masks whereas categorical crossentropy is used to obtain multi-class segmentation for the BraTS dataset. The datasets used for the evaluation of the model had different image resolutions and pixel intensities. As a result, individual input slices are normalized. Each dataset has been split into a ratio of 60:20:20 for training, validation, and testing respectively

4. Results And Comparison

4.1. Results

Figure 4 represents the qualitative analysis of the generated predicted mask for different modalities such as MRI (breast and brain), PET (breast), and CT scan (Cervical Spine) from top to bottom.To prove the efficiency and the robustness of the model, Dice Similarity Coefficient (DSC), Average Hausdorff distance (HD), and False Positive Rate (FPR) are used as a performance metrics. The proposed AW-Net achieves an average DSC of more than 80. The maximum (Max), minimum (Min), average (Avg) and stan-

Table 2: Summarized results obtained by the proposed AW-Net.

Body Parts Modality	Sequences	DSC (%)	HD (mm)	FPR (mm)
Breast MRI	DCE	81.3±12.3	10.7±0.7	0.04±0.01
		89.6-61.5	11.3-9.7	0.15-0.01
	T1	80.5±11.5	9.9±0.6	0.04±0.02
		87.5-50.2	11.8-9.1	0.16-0.01
Brain MRI	T1CE+ T2+FLAIR	89.6±8.6	10.5±0.5	0.1±0.08
		96.8-59.5	11.3-8.7	0.18-0.08
Breast PET	-	93.7±11.0	00.5±0.6	0.01±0.05
		96.8-60.3	02.0-0.5	0.14-0.02
Cervical Spine CT	-	71.9±22.1	6.2±1.1	3.5±1.2
		95.0-25.6	9.9-2.2	7.2-2.1

*The shaded row (Max - Min) and next to shaded row (Avg ± SD)

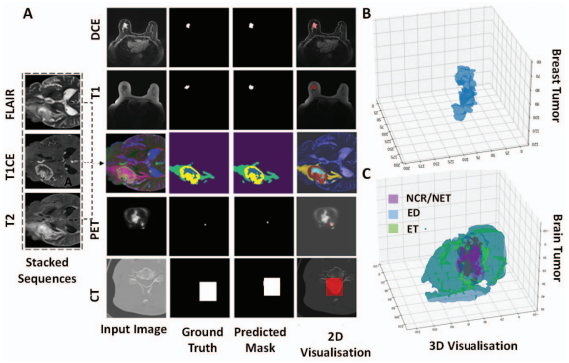

Figure 4: A detailed analysis of the predicted mask generated by the proposed AW-Net with respect to A) 2D visualization of individual segmented masks for a single image slice, B) 3D visualization of predicted breast tumor and C) 3D visualization of brain tumor.

dard deviation (SD) values in terms of DSC, HD and FPR for the different modalities and sequences is shown in Table 2. An average DSC of 81.3, 80.5 and 89.6 is reported for DCE-MRI, T1-MRI, and TICE+T2+FLAIR MRI, respectively. For other modalities such as Breast PET and Cervical Spine CT an average DSC of 93.7 and 71.9 are reported respectively. An average HD of 10.7 and 9.9 is observed for DCE and T1 for breast MRI and 10.5 in brain MRI. Similarly, an average HD of 0.5 and 6.2 is observed for breast PET and cervical spine CT. Table 2 reports an average FPR of 0.04 for breast MRI segmentation, 0.1, 0.01 and 3.5 for brain MRI and breast PET and cervical spine CT, respectively.

4.2. Comparison with other models

A comparative study with other state-of-the-art segmentation models such as U-Net3+ [14], Trans U-Net [25], Swin U-Net [7], Attention UW-Net [24], LinkNet-b7 [2] and FPN [34] and 3D segmentation models such as 3D U-Net [9],3D attention U-Net [15], and 3D LinkNet [6] have been performed with the proposed AW-Net. Table 3 showcases the result of the quantitative analysis done on the aforementioned models with respect to the proposed AW-Net. In addition, the number of parameters for the aforementioned models has also been reported in the last column of Table 3. Figure 5 represents the qualitative analysis of the proposed model with respect to the other compared models in terms of the generated segmented masks.

The AW-Net model proposed in this study shows superior performance compared to other models. It achieves a DSC score that is at least 0.8 higher for NCR/NET, 1.6 higher for ET, and 2.4 higher for ED. It also has the least number of misclassified pixels (FPR) compared to the other models. This is because of the regularised convolutional block and the dropout layers in the RTB. Transformer-based U-Net models perform poorly for breast tumor segmentation, especially for DCE sequence with the trans U-Net and Swin U-Net achieves an average DSC of 76.5 and 70.5, respectively. LinkNet-b7 trails the proposed model by a DSC of 27.9 and 31.0, for small lesion (tumor) segmentation in DCE and T1 sequence respectively. This is due to the bottleneck of the aforementioned model. In comparison to 3D segmentation models like 3D U-Net, 3D attention U-Net, and 3D Link Net, the proposed AW-Net outperforms them by a significant margin of 10.2, 6.4 and 8.5 in terms of DSC. The AW-Net also has the lowest FPR of 0.003 for tumor segmentation from PET imaging. This is because of the additional attention layer and skip connections in the RTB layer. Their additive effect made the model sensitive to adjacent pixels with similar intensity as the tumor cells. The proposed model outperforms the U-Net 3+ for the RSNA cervical spine dataset, with a DSC score above 75. The low average FPR score of 0.6 highlights the effectiveness of the proposed AW-Net in reducing the number of misclassified pixels. The proposed model also has the least number of trainable paramters in comparision to SOTA segmentation models, as is shown in Table 3. AW-Net has 4.68M parameters involved during training due to the regularized convolutional operations in RTB. In summary, the proposed AW-Net is the best-performing model, taking into account all aspects of the comparison.

5. Ablation Study

An ablation study performed on the proposed AW-Net showcases the effectiveness of the RTB with respect to DSC and FPR in Figure 6A and 6B respectively. We vary various

Table 3: Summarized results for segmentation on breast PET, cervical spine CT, breast, and brain MRI in terms of average DSC (in %), FPR (in %) and number of parameters (in M). Yellow shaded cells show the performance of the AW-Net

	PET		CT		MRI				Brain						No. of Parameters (M)
					DCE		T1		NCR/NET		ET		ED		
	DSC	FPR	DSC	FPR	DSC	FPR	DSC	FPR	DSC	FPR	DSC	FPR	DSC	FPR	
Proposed Model	**93.7**	**0.003**	71.9	**3.5**	**81.3**	**0.12**	**80.5**	**0.11**	88.9	**0.17**	**84.4**	**0.19**	**95.6**	**0.13**	**4.68**
Trans U-Net	87.6	0.025	55.4	6.9	76.5	0.25	74.8	0.27	87.7	0.18	81.1	0.22	92.8	0.16	86.70
Swin U-Net	82.8	0.022	45.8	7.2	70.5	0.19	71.2	0.18	88.4	0.19	76.5	0.21	93.2	0.18	9.36
Attention UW-Net	89.3	0.01	68.3	3.8	78.1	0.15	77.6	0.22	82.5	0.14	77.7	0.61	88.5	0.17	16.30
U-Net 3+	79.4	0.07	**72.1**	3.5	73.4	0.17	74.1	0.29	**89.1**	0.18	82.8	0.22	93.2	0.17	7.87
LinkNet-b7	82.1	0.03	41.2	8.3	53.4	0.45	49.5	0.62	88.7	**0.17**	75.7	**0.19**	92.6	**0.13**	72.26
FPN	79.8	0.25	30.5	19.4	71.3	0.14	70.9	0.15	80.1	0.2	54.8	1.3	68.33	2.1	20.30
3D U-Net	83.5	0.11	46.2	5.5	77.5	0.13	76.9	0.08	70.2	0.31	67.5	0.82	87.7	0.45	20.90
3D-Attention U-Net	87.1	0.05	55.1	4.9	78.3	0.7	77.9	0.2	83.8	0.15	72.3	0.26	91.7	0.25	26.73
3D LinkNet	85.2	0.04	44.3	3.7	60.5	0.56	61.3	0.7	75.1	0.22	70.4	0.25	90.1	0.19	20.20

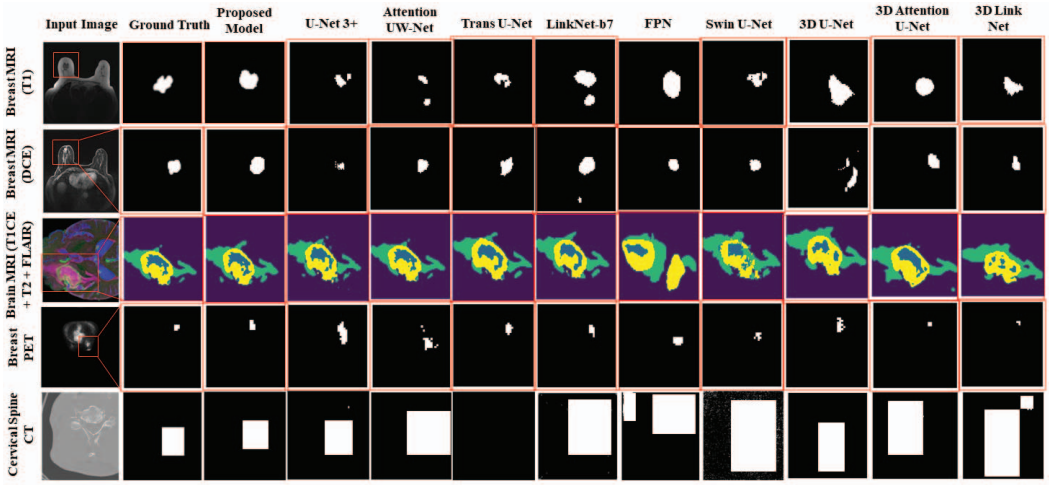

Figure 5: The output masks generated by the proposed model and other compared models: U-Net 3+, Attention UW-Net, Trans U-Net, LinkNet-b7, FPN and Swin U-Net (from left to right) with respect to different modalities: MRI-T1, MRI-DCE, MRI for stacked T1CE, T2 and FLAIR, PET, and CT (from top to bottom).

parameters of the RTB i.e., the regularizer and the dropout rate. We have considered attention U-Net as the baseline model, M. The RTB is represented as a bivariate function RTB(r, L) which takes two parameters, dropout rate r and regularizer function L as input. The proposed model is represented as M + RTB(0.2, L1). In this study we vary the dropout rate from 0.1 to 0.3 and the regularizer between L1 and L2. We have also considered a model variant without the regularizers in the RTB to showcase the importance of regularizers in the transient block. This study has been repeated for different modalities as is showcased in Figure 6A and 6B. The addition of RTB improves the baseline model M, by a DSC of 4.7 for RSNA dataset and an FPR of 1.55 for BraTS dataset. This study establishes the fact that RTB with the combination of L1 regularizer and dropout plays a crucial role in improving model performance and reducing FPR.

6. Data Growth Study

Figure 7 presents a data growth study, where the model's performance is assessed through training on a reduced dataset and evaluating on the remaining data. The primary objective is to analyze how the model performs under these conditions, specifically exploring the impact of data size on its effectiveness. The study aims to uncover valuable insights about the direct relationship between dataset size and model performance.The investigation commences by utilizing 30% of the data for training and reserving 70% for testing. Through a systematic approach, the training data is incrementally increased by 5%, while the testing data is simultaneously reduced by 5%, until reaching a configuration with 80% of the data for training and 20% for testing. This step-by-step analysis is repeated for other segmentation models such as U-Net 3+, Trans U-Net, Attention UW-Net, LinkNet-b7, 3D Attention U-Net, and 3D LinkNet.

This study offers a deeper understanding of the correlation between data size and the overall performance of various 2D and 3D segmentation models. However, it is important to note that configurations with training data less than 30% are not reported due to the limited availability of training data for breast PET and MRI images, which have only 240 and 320 annotated images for the experiment. The proposed model demonstrates remarkable performance in tumor segmentation for breast PET (shown in Fig 7(C)), achieving a DSC of 88% when trained on 30% of the available data. This result is significantly better than other segmentation models like Attention UW-Net, Trans U-Net, and 3D Attention U-Net. For brain tumor detection from the BraTS2020 dataset in Fig 7(B), the proposed model also performs well, obtaining a DSC of 82% when trained on the same 30% of the data. However, in the case of breast MRI tumor segmentation, the AW-Net falls slightly behind the 3D U-Net by 0.9% in DSC when both models are trained on 30% of the data. Nonetheless, AW-Net demonstrates substantial performance improvement after being trained with 60% of the data.On a different note, the RSNA challenge data, used for cervical spine fracture localization, resulted in poorer annotations, leading to a lower DSC of 60% when the model was trained on only 30% of the data. Despite this limitation, the reported DSC values exhibit low variance, even when the model is trained on a small dataset, providing valuable insights into the robustness and effectiveness of the proposed model.

7. Computational Time

To evaluate the efficiency of the model, a study is performed based on the memory involved while training a single epoch (in Mb) with respect to the number of floating-point operations performed per second (FLOPS). Furthermore, test inference time (in seconds) have also been calculated for these segmentation models to induce the implementation of the developed model for real-time applications. The results for the proposed AW-Net with respect to SOTA segmentation models is reported in Table 4. Training AW-Net for a single epoch takes 62.5 Mb and involves 5.7 G of FLOPS which is the least among other SOTA models. The proposed model also has the least test inference time of 0.13 seconds, making it suitable for real-world applications.

8. Conclusion

In this paper, we have proposed a novel AW-Net for the segmentation of multimodal 3D/4D images. The proposed model considers the anatomical features and reduces pixel misclassification by the introduction of RTB. The regularised convolutional layers in the RTB not only reduces the computational complexity but also makes the model robust. Experiments performed on multiple dataset having different modalities and data sequences demonstrate the effectiveness and the generalizability of the model. The model consistently outperforms other benchmark segmentation models in terms of DSC and FPR for public benchmark datasets. Furthermore, the performance analysis with respect to the training data shows that the proposed model works well on smaller datasets. Thus, we conclude that the proposed model is a generalised, cost effective, and robust model.

Table 4: Complexity Analysis of the proposed AW-Net.

Model	FLOPS (G)	Memory Taken (Mb)	Test Inference Time (Sec.)
Proposed Model	**5.7**	**62.5**	**0.13**
Trans U-Net	35.1	240.9	0.35
Swin U-Net	32.8	210.5	0.36
Attention UW-Net	12.5	98.4	0.16
U-Net 3+	132.7	80.3	0.26
LinkNet-b7	29.9	129.4	0.24
FPN	40.4	100.3	0.15
3D U-Net	360.5	319.7	0.52
3D Attention U-Net	250.2	355.4	1.09
3D LinkNet	353.9	325.2	0.52

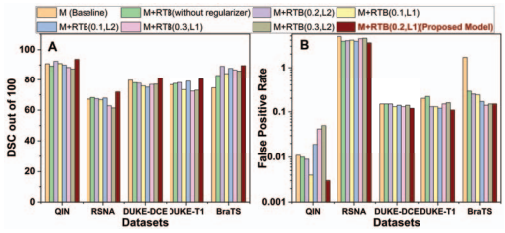

Figure 6: Stacked plot representing A) Average DSC (in %) and B) FPR (in %) of the different models in the ablation study on the proposed AW-Net for different modalities.

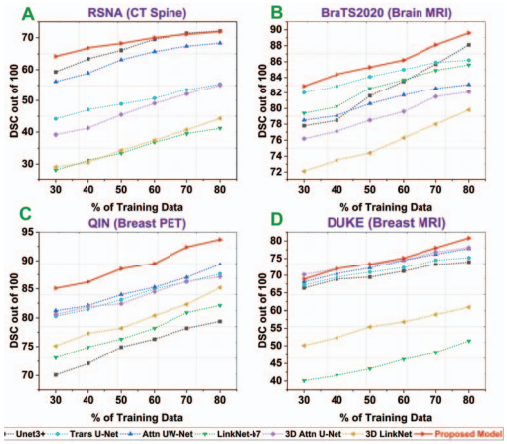

Figure 7: Line plots illustrating the performance of AW-Net for the data growth study with respect to different datasets. A) RSNA. B) BraTS2020. C) QIN. D) DUKE.

References

[1] Errol Colak FelipeKitamura Hui Ming Lin JeffRudie John Mongan Katherine Andriole Luciano Prevedello Michelle Riopel Robyn Ball Sohier Dane Adam Flanders, Chris Carr. Rsna 2022 cervical spine fracture detection, 2022. 5

[2] Cihan Akyel and Nursal Arıcı. Linknet-b7: Noise removal and lesion segmentation in images of skin cancer. *Mathematics*, 10(5):736, 2022. 6

[3] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin Kirby, John Freymann, Keyvan Farahani, and Christos Davatzikos. Segmentation labels and radiomic features for the pre-operative scans of the tcga-lgg collection. *The cancer imaging archive*, 286, 2017. 5

[4] Lei Bi, Dagan Feng, and Jinman Kim. Dual-path adversarial learning for fully convolutional network (fcn)-based medical image segmentation. *The Visual Computer*, 34:1043–1052, 2018. 2

[5] Lei Bi, Jinman Kim, Ashnil Kumar, Michael Fulham, and Dagan Feng. Stacked fully convolutional networks with multi-channel learning: application to medical image segmentation. *The Visual Computer*, 33:1061–1071, 2017. 2

[6] Jun-Xiong Cai, Tai-Jiang Mu, Yu-Kun Lai, and Shi-Min Hu. Linknet: 2d-3d linked multi-modal network for online semantic segmentation of rgb-d videos. *Computers & Graphics*, 98:37–47, 2021. 6

[7] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*, pages 205–218. Springer, 2023. 2, 6

[8] Jianhong Cheng, Jin Liu, Liangliang Liu, Yi Pan, and Jianxin Wang. Multi-level glioma segmentation using 3d u-net combined attention mechanism with atrous convolution. In *2019 ieee international conference on bioinformatics and biomedicine (BIBM)*, pages 1031–1036. IEEE, 2019. 2

[9] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, pages 424–432. Springer, 2016. 6

[10] Karuppaiah Geetha, Veerasamy Anitha, Mohamed Elhoseny, Shankar Kathiresan, Pourya Shamsolmoali, and Mahmoud M Selim. An evolutionary lion optimization algorithm-based image compression technique for biomedical applications. *Expert Systems*, 38(1):e12508, 2021. 2

[11] Lin Han, Yuanhao Chen, Jiaming Li, Bowei Zhong, Yuzhu Lei, and Minghui Sun. Liver segmentation with 2.5 d perpendicular unets. *Computers & Electrical Engineering*, 91:107118, 2021. 2

[12] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584, 2022. 2

[13] Chao Huang, Hu Han, Qingsong Yao, Shankuan Zhu, and S Kevin Zhou. 3d u 2-net: A 3d universal u-net for multi-domain medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II*, pages 291–299. Springer, 2019. 2

[14] Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Jian Wu. Unet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1055–1059. IEEE, 2020. 6

[15] Mobarakol Islam, VS Vibashan, V Jeya Maria Jose, Navodini Wijethilake, Uppal Utkarsh, and Hongliang Ren. Brain tumor segmentation and survival prediction using 3d attention unet. In *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 5*, pages 262–272. Springer, 2020. 6

[16] Yang Lei, Sibo Tian, Xiuxiu He, Tonghe Wang, Bo Wang, Pretesh Patel, Ashesh B Jani, Hui Mao, Walter J Curran, Tian Liu, et al. Ultrasound prostate segmentation based on multidirectional deeply supervised v-net. *Medical physics*, 46(7):3194–3206, 2019. 2

[17] Jingyuan Li, Guanqun Liao, Wenfang Sun, Ji Sun, Tai Sheng, Kaibin Zhu, Karen M von Deneen, and Yi Zhang. A 2.5 d semantic segmentation of the pancreas using attention guided dual context embedded u-net. *Neurocomputing*, 480:14–26, 2022. 2

[18] Xia Li, Richard G Abramson, Lori R Arlinghaus, Anuradha Bapsi Chakravarthy, Vandana G Abramson, Melinda Sanders, and Thomas E Yankeelov. Data from qin-breast. *The Cancer Imaging Archive*, 2016. 5

[19] Sun'ao Liu, Hai Xu, Yizhi Liu, and Hongtao Xie. Improving brain tumor segmentation with dilated pseudo-3d convolution and multi-direction fusion. In *MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part I 26*, pages 727–738. Springer, 2020. 2

[20] Tao Liu, Yun Tian, Shifeng Zhao, XiaoYing Huang, Yang Xu, Gaoyuan Jiang, and Qingjun Wang. Pseudo-3d network for multi-sequence cardiac mr segmentation. In *Statistical Atlases and Computational Models of the Heart. Multi-Sequence CMR Segmentation, CRT-EPiggy and LV Full Quantification Challenges: 10th International Workshop, STACOM 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Revised Selected Papers 10*, pages 237–245. Springer, 2020. 2

[21] Avetisian Manvel, Kokh Vladimir, Tuzhilin Alexander, and Umerenkov Dmitry. Radiologist-level stroke classification on non-contrast ct scans with deep u-net. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, Oc-*

tober 13–17, 2019, Proceedings, Part III 22, pages 820–828. Springer, 2019. 2

[22] Tanushree Meena and Sudipta Roy. Bone fracture detection using deep supervised learning from radiological images: A paradigm shift. *Diagnostics*, 12(10):2420, 2022. 1

[23] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018. 1, 2

[24] Debojyoti Pal, Pailla Balakrishna Reddy, and Sudipta Roy. Attention uw-net: A fully connected model for automatic segmentation and annotation of chest x-ray. *Computers in Biology and Medicine*, 150:106083, 2022. 2, 6

[25] Olivier Petit, Nicolas Thome, Clement Rambour, Loic Themyr, Toby Collins, and Luc Soler. U-net transformer: Self and cross attention for medical image segmentation. In *Machine Learning in Medical Imaging: 12th International Workshop, MLMI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 12*, pages 267–276. Springer, 2021. 2, 6

[26] Mamoona Riaz, Maryam Bashir, and Irfan Younas. Meta-heuristics based covid-19 detection using medical images: A review. *Computers in Biology and Medicine*, page 105344, 2022. 2

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 2

[28] Abhijit Guha Roy, Nassir Navab, and Christian Wachinger. Concurrent spatial and channel 'squeeze & excitation'in fully convolutional networks. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*, pages 421–429. Springer, 2018. 2

[29] Sudipta Roy, Tanushree Meena, and Se-Jung Lim. Demystifying supervised learning in healthcare 4.0: A new reality of transforming diagnostic medicine. *Diagnostics*, 12(10):2549, 2022. 1

[30] Ashirbani Saha, Michael R Harowicz, Lars J Grimm, Connie E Kim, Sujata V Ghate, Ruth Walsh, and Maciej A Mazurowski. A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 dce-mri features. *British journal of cancer*, 119(4):508–516, 2018. 5

[31] Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20730–20740, 2022. 2

[32] Girindra Wardhana, Hamid Naghibi, Beril Sirmacek, and Momen Abayazid. Toward reliable automatic liver and tumor segmentation using convolutional neural network based on 2.5 d models. *International journal of computer assisted radiology and surgery*, 16:41–51, 2021. 2

[33] Xiao Xiao, Shen Lian, Zhiming Luo, and Shaozi Li. Weighted res-unet for high-quality retina vessel segmentation. In *2018 9th international conference on information technology in medicine and education (ITME)*, pages 327–331. IEEE, 2018. 2

[34] Hang Xu, Lewei Yao, Wei Zhang, Xiaodan Liang, and Zhenguo Li. Auto-fpn: Automatic network architecture adaptation for object detection beyond classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6649–6658, 2019. 6

[35] Li Xu, Yuqing Hou, Fengjun Zhao, and Jinniu Bai. Medical ct image enhancement system based on swarm intelligence optimization algorithm. In *Cyber Security Intelligence and Analytics: The 4th International Conference on Cyber Security Intelligence and Analytics (CSIA 2022), Volume 1*, pages 1035–1042. Springer, 2022. 2

[36] Huahong Zhang, Alessandra M Valcarcel, Rohit Bakshi, Renxin Chu, Francesca Bagnato, Russell T Shinohara, Kilian Hett, and Ipek Oguz. Multiple sclerosis lesion segmentation with tiramisu and 2.5 d stacked slices. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22*, pages 338–346. Springer, 2019. 2

[37] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pages 3–11. Springer, 2018. 2