Multi-Res-Attention UNet : A CNN Model for the Segmentation of Focal Cortical Dysplasia Lesions from Magnetic Resonance Images

Edwin Thomas, S. J Pawan, Shushant Kumar, Anmol Horo, S. Niyas, S. Vinayagamani, Chandrasekharan Kesavadas, and Jeny Rajan

***Abstract*—In this work, we have focused on the segmentation of Focal Cortical Dysplasia (FCD) regions from MRI images. FCD is a congenital malformation of brain development that is considered as the most common causative of intractable epilepsy in adults and children. To our knowledge, the latest work concerning the automatic segmentation of FCD was proposed using a fully convolutional neural network (FCN) model based on UNet. While there is no doubt that the model outperformed conventional image processing techniques by a considerable margin, it suffers from several pitfalls. First, it does not account for the large semantic gap of feature maps passed from the encoder to the decoder layer through the long skip connections. Second, it fails to leverage the salient features that represent complex FCD lesions and suppress most of the irrelevant features in the input sample. We propose Multi-Res-Attention UNet; a novel hybrid skip connection-based FCN architecture that addresses these drawbacks. Moreover, we have trained it from scratch for the detection of FCD from 3 T MRI 3D FLAIR images and conducted 5-fold cross-validation to evaluate the model. FCD detection rate (Recall) of 92% was achieved for patient wise analysis.**

***Index Terms*—Attention gates, focal cortical dysplasia, fully convolutional network, magnetic resonance imaging, respaths, semantic segmentation.**

I. Introduction

FOCAL Cortical Dysplasia (FCD) is a heterogeneous group of disorders due to malformations of cortical development (MCD) and is often associated with hippocampal sclerosis and cortical glioneuronal neoplasms [1]. It is considered to be the most common etiology on the cryptogenic and probable symptomatic epilepsy among adults and children [2].

Several classification systems of FCD have been proposed in the literature [3]–[6]. Palimini's [4] and Barkovich's [5] classification, is based on a two-tier system, while the ILAE classification, adopted the consensus classification system for FCD proposed by Blumcke *et al.* [6], which is composed of a three-tier system and shares many features with [4], [5]. It is the most widely accepted system today and differentiates the types of FCD lesions based on its location, structure, T1/T2/FLAIR signal intensities in Magnetic Resonance Images (MRI), and its adjacent occurrence with other abnormalities.

General features present in MRI images that are used to characterize FCD include cortical thickening, blurring of white-matter (WM), and gray-matter (GM) junction with the abnormal architecture of sub-cortical layer, abnormal sulcal, or gyral pattern and segmental or lobar hypoplasia [7]. FCD can be hard to detect due to its multi-focal occurrence and variable size [8]. Furthermore, the subtle characteristics of these lesions subject them to inter-observer and intra-observer variability. Consequently, automated approaches have gained popularity due to its ability to detect and quantify the extent of FCD lesions consistently. Several approaches for the detection of FCD lesion are proposed in the literature and are broadly classified into symmetry-based, template-based, and feature-based approaches [9]. However, most of these techniques rely on precise hand-crafted feature engineering tasks and are prone to errors. With the emergence of convolutional neural networks (CNNs), classification, segmentation, and related computer vision tasks no longer require domain experts to extract the required features manually. A customized UNet architecture [10] is the current state-of-the-art fully convolutional network (FCN) model proposed for the segmentation of FCD. We have identified several drawbacks in this architecture, and they are elaborated in Section III.

In this paper, we present a new hybrid skip connections based CNN architecture that addresses the drawbacks of [10], which is inspired from [11] and [12] based upon careful considerations about the scope of improvement of these models. We propose the optimal placement of the Attention Gates, the right choice of feature maps for the Gating Signal, and the optimal placement of the ResPaths within the proposed architecture for the segmentation of FCD. The rest of this paper is organized as follows:

Manuscript received April 10, 2020; revised July 16, 2020 and August 19, 2020; accepted September 9, 2020. Date of publication September 15, 2020; date of current version May 11, 2021. This work was supported by the Cognitive Science Research Initiative (Department of Science, and Technology, India) through project funding DST/CSRI/2017/47. *(Corresponding author: Jeny Rajan.)*

Edwin Thomas, S. J Pawan, Shushant Kumar, Anmol Horo, S. Niyas, and Jeny Rajan are with the Department of Computer Science and Engineering, National Institute of Technology Karnataka, Surathkal 575025, India (e-mail: edwinthomas444@gmail.com; pawansnj03@gmail.com; shushantkmr2@gmail.com; horoanmol@gmail.com; niyasknpy@gmail.com; jenyrajan@gmail.com).

S. Vinayagamani and Chandrasekharan Kesavadas are with the Department of Imaging Sciences and Interventional Radiology, Sree Chitra Tirunal Institute for Medical Sciences and Technology, Trivandrum, India (e-mail: dr.vinayagamani@gmail.com; Kesavadaskesav@sctimst.ac.in).

This article has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors.

Digital Object Identifier 10.1109/JBHI.2020.3024188

A literature survey on existing FCD segmentation techniques and Attention-based mechanisms is presented in Section II. Section III describes the motivation behind the proposed architecture and the drawbacks of the baseline model. The proposed network architecture is elucidated in Section IV. Section V presents the experimental setup and results. Finally, Section VI summarizes this paper.

II. Related Work

A. Segmentation of FCD

In this section, we attempt to summarize relevant classical (image processing and machine learning methods) and deep learning-based approaches for the segmentation of FCD. The classical methods were first applied by Boonyapisit *et al.* to correlate the cellular patterns of MCDs, based on the assessment of extra operative electrocorticographic (ECoG) recordings [13]. Two successive deformable models FDM and EDM [14] with level set evolution were proposed to segment FCD by first isolating the lesion based on three prominent discriminative features followed by the examination of the possibility of expansion of the lesion to the cortical boundaries. Following a different approach Colliot *et al.* [15] use a set of computational models based on the blurring of the GM-WM transition region and the dysplastic region's hyperintense signals. They also applied FCD lesion profiling on the images to quantitatively assess the abnormal regions and proved that a large variability exists in degrees of abnormalities. A number of tool-based studies were also conducted [16] that used SPM5 and FSL tools with bias correction of raw FLAIR scans followed by intensity normalization using internal reference regions. However, it looked onto very specific regions and did not use non-sphericity of the lesions during cluster formation. A new method that uses the positive features of both statistical classification and elastic matching methods to enhance the segmentation of the lesions in the WM region was proposed in [17]. In [18], Chin-Ann *et al.* proposed a volume-based discriminative feature analysis technique that identifies features for a group of voxels to capture spatial information. This research work was one of the early methods to have used classical machine learning algorithms to classify FCD.

An automated classifier that depends on surface-based features was introduced by Hong *et al.* [19] to detect FCD type II. On a similar note, another publication by Ahmed *et al.* [20], presented a model that improves the detection of FCD in MRI-negative images by applying a quantitative morphometry approach, which computed five surface-based MRI features such as the cortical thickness, WM and GM contrast, sulcal depth, mean curvature, and Jacobian distortion. Azami *et al.* in [21], designed a machine-learning algorithm based on a one-class SVM classifier for voxel-wise analysis to detect abnormalities in T1 weighted MRI samples of patients. The model uses six textural feature maps, which is then used by the multivariate analysis technique to achieve a good performance in an easier fashion as compared to the Statistical parametric mapping (SPM) analysis technique. In [22], a complex diffusion-based approach was proposed to identify the FCD regions. An artificial neural network-based technique was proposed in [23], which presented a method to detect small FCD lesions on T1 weighted MRI images. It is founded on surface-based features that best describe the morphometric characteristics of small FCD lesions, which were then fed to a simple four-layered artificial neural network. A deep learning-based approach was presented in [10] by Bijay Dev *et al.* This paper used a customized UNet architecture as its base model that was fine-tuned for the FCD dataset. However, the architecture leverages simple skip connections that do not account for the semantic gap between feature maps of corresponding encoder and decoder layers. Experimental results pertaining to the approaches in this literature survey have shown that deep learning approaches are at the forefront of the image segmentation and evaluation tasks. These techniques enable automated optimal feature extraction, which has paved the way for learning more essential features than any manual feature extraction-based methods.

B. Attention based mechanisms for Image Segmentation

Several attention-based methods have been used across the landscape of semantic segmentation. In this section, we present a brief review of some of the relevant works pertaining to the same. A reverse attention structure [24] was proposed that generated a per-class mask to amplify reverse-class response and learn what is not associated with the region of interest. Alternatively, a feature pyramid network [25] was introduced to generate the attention signal from different pyramid scales, and a global average pooling operation was performed to provide global context as guidance to low-level features as a means to compute category localization details. Dong *et al.* [26] introduced a dual attention model, which is a combination of the position attention module and channel attention module. The output of the two modules was then fused to be used as single attention input. Another automated architecture that leverages deep attention, along with deep supervision, was proposed in [27] for multi-organ segmentation. However, the model added Attention Gates at shallowest depth, which gives attention to naive features. A ranking attention network was proposed in [28], which used an attention module to rank the similarity maps of foreground and background masks for video object detection. Further, these maps and the previous frame's mask were fed into the decoder before the final segmentation. The related approaches using Attention mechanisms were built on different baseline models resonating with the type of concerned datasets. Hence, the dataset played an essential role in guiding the various design decisions adopted in the respective architectures. We also note that the existing work on attention based mechanisms do not specifically address the important design aspect concerning the right alternative of Gating Signals used to guide the activation maps and the placement of Attention Gates in encoder-decoder based semantic segmentation models. In this work, we explore the aforementioned aspects in detail.

III. Motivation

In this section we present few essential design features that we have considered to address the drawbacks of the baseline architecture (Bijay Dev *et al.* [10]). We have drawn inspiration from the existing architectures whose design considerations closely reflects ours.

A. Attention To Salient Features

The Attention Gated networks were introduced to prevent loss of finer features between the intermediate layers and the final layers. Schlemper in [12] proposed attention gated UNet architecture. Similar to the attention layers used in RNNs, image-grid based gating is implemented to allow attention coefficients to keep the focus of the model on specific regions of the image, which is to give more attention to a particular feature like lesion regions. There is a high possibility of a loss of information further up the decoder part in UNet. So, to reduce this, an attention gated signal, which is the context vector represented as a global feature vector, is sent from the coarser layers to the shallow layers along with the longer skip connections which connect encoder and decoder at mirrored depths. The features in the deeper layers are of higher granularity, and hence attention signal is applied to it. So, in this way, the concept of attention helps in directing the model activations to give more importance to salient features and trim the unwanted regions of interest. As the FCD lesions are characterized by their small size, shape variability, and low variance with the surrounding pixels, its boundary delineation a very tedious task. We believe that the concept of attention-based mechanisms shall boost the model's capability to identify the most important features that characterize the lesion, thus improving the quality of segmented FCD lesions.

B. Robustness to Scale and Reduction of Semantic Gap

Medical image segmentation tasks such as those involving segmentation of lesions, organs, tumors, etc., may encounter images that contain varying scales of the areas of interest. Thus, the architectures developed to achieve the above task must be resilient to the examinations of these images to achieve better segmentation. In the literature, several approaches to address this issue are proposed. For instance, the inception block [29], uses kernels of multiple dimensions in parallel to produce feature maps for the same input and concatenates them. Another approach [11], seeks to optimize the inception blocks by linearly concatenating filters and uses short skip connections from intermediate layers within the block before concatenation, hence approximating the effect. On the contrary, the baseline architecture uses a constant kernel dimension within a layer. Therefore, it fails to capture the essence of effective scale-invariant segmentation. We believe that the concept of robustness to scale variability will particularly benefit FCD segmentation considering the variance in the lesion size and shape depending on its stage of development.

Furthermore, the baseline uses simple skip connections to preserve spatial information for the segmentation task. However, [11] and [30], introduced non-linearity in the skip connections to reduce the semantic gap between encoder and decoder features. While the former uses a linearly connected ResPath, the latter uses a nested skip connection and functionally serves the same purpose. It is assumed that the semantic gap increases discrepancy in the learning process and is detrimental to performance. We hypothesize that enabling our model to achieve a lesser semantic gap shall improve its overall performance.

IV. Network Architecture

Having drawn deeper insights into the drawbacks of the baseline architecture, we incorporate three major design features, namely attention to salient features, reduction in the semantic gap, and robustness to scale in the proposed model (Fig. 1) to address the same. The first three sub-sections discuss the various components used in our architecture, followed by an ablation study that elaborates on the optimal placement of these components in the proposed architecture.

A. Multi-Res Block

To ensure scale invariability we have incorporated Multi-Res Blocks from [11]. Consequently, it promises a two-way optimization. First, it reduces the higher memory requirement of 5×5 and 7×7 kernels by using a chain of 3×3 kernels in series. The output from the first, second and third 3×3 filter sets approximate the usage of 3×3, 5×5 and 7×7 kernels respectively. These outputs are concatenated with the feature map obtained after applying a 1×1 convolution to the input. Second, the block increases the number of filters in the successive three layers in steps, to prevent the memory overhead of the filters in previous layers from escalating to the successive layers. It reduces the quadratic effect that is imposed by prior filter dimensions to the current ones.

To substantiate the above point, let us assume without loss of generality a two-layer setup $(L1, L2)$ in which the number of filters in L1 is lesser that or equal to L2. Suppose input feature map to $L1$ is of dimension (a, a, n) and there are l filters of dimension (k, k, n) in this layer, then the output is of dimension $(a + 2p - k + 1, a + 2p - k + 1, l)$ where p is the padding used. The number of filter parameters in this layer is given by Eq. 1. Let $L2$ have $l^{'}$ filters of dimension $(k^{'}, k^{'}, l)$. As the number of filters in $L2$ is at least as many as in $L1$, $l^{'} >= l$. This implies that the number of filter parameters of the $L2$ is greater than or equal to $(k^{'} \times k^{'} \times l \times l)$ as shown in Eq. 2.

$$\text{No. of filter parameters in L1} = k^2 \times n \times l \quad (1)$$

$$\text{No. of filter parameters in L2} \geq (k^{'})^2 \times l^2 \quad (2)$$

Thus the quadratic effect of l, which is the number of filters of the previous layer is propagated on the current layer. This analogy can be extended to any number of layers. Having said that, we believe the Multi-Res block is an apt candidate and we leverage it in the encoder blocks in the proposed architecture as seen in Fig 1.

B. ResPath

The presence of the semantic gap is one of the significant issues that the baseline architecture failed to address. Among the two notable works to bridge the semantic gap are [11] and [30], we adopt the former as ResPaths have lesser memory

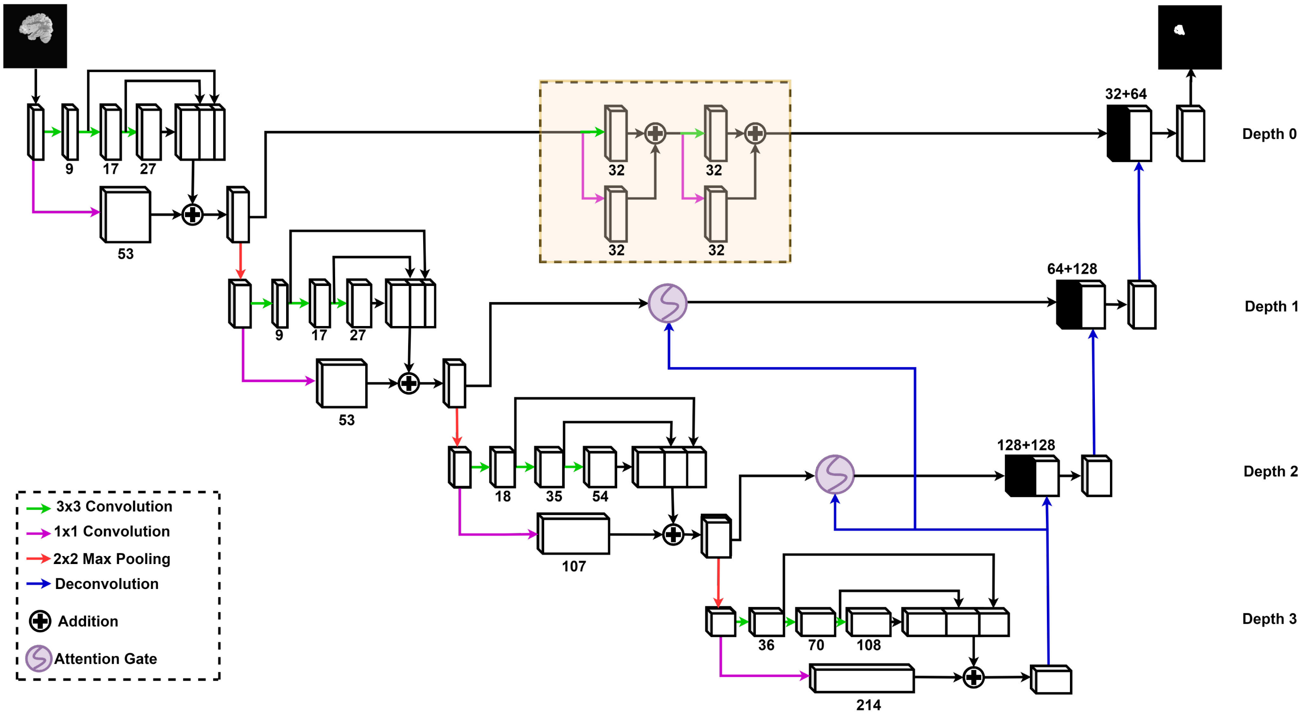

Fig. 1. Block diagram of the proposed Multi-Res-Attention UNet architecture.

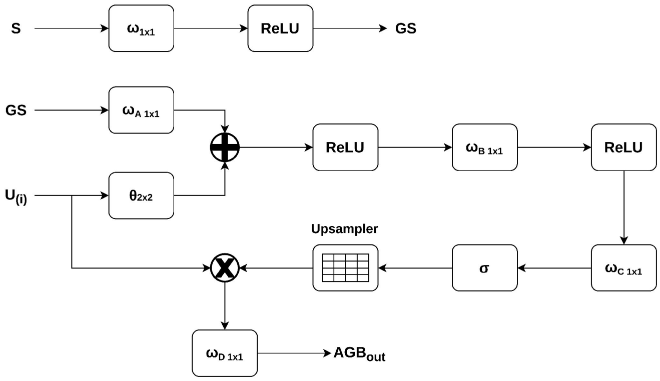

Fig. 2. A schematic diagram showing Attention Gating Signal (GS) and Attention Gating Block structure.

requirements. Taking into account the small size of FCD lesions, we believe that a model of very large depth is undesirable. Using an architecture of lesser depth means that the feature maps that are input to the deepest encoder layer shall contain a sufficiently large patch of the area of interest (due to the lesser number of subsampling operations), adequate for this layer to extract useful information. Consequently, we propose an architecture of depth-3. The features extracted from the depth-0 encoder block (Multi-Res) are low-level features as they are computed in the earlier stages of the network. At the same time, the decoder layer at the mirrored depth-0 receives higher-level feature maps since they go through more processing. This discrepancy between the feature maps is most pronounced across depth-0. Hence, we propose to confine the usage of ResPath in this context to only the shallowest layer of the proposed architecture.

The ResPath across depth-0 (Fig. 1) consists of two layers of 3×3 convolutional filters. The first layer processes the feature map u_i output from the encoder block at depth-0. In parallel to it, another $\omega_{1\times1}$ convolution is performed, and this is added to the above resultant (Res_x, Eq. 3). This is followed by an identical layer, which takes Res_x as input and outputs (Res_y, Eq. 4). $\theta_{X_{3\times3}}$ and $\theta_{Y_{3\times3}}$ represents filters of the first and second layers of the ResPath, respectively. b_x and b_y represents biases corresponding to these layers.

$$Res_x = \theta_{X_{3\times3}}.u_i + \omega_{X_{1\times1}}(u_i) + b_x \tag{3}$$

$$Res_y = \theta_{Y_{3\times3}}.Res_x + \omega_{Y_{1\times1}}(Res_x) + b_y \tag{4}$$

These set of convolutions between the encoder and decoder branch at depth-0, introduces non-linearity in the feature maps and helps in reducing the semantic gap between them.

C. Attention Gating Block

The attention block consists of an Attention Gating Signal (GS) which is taken from the bottom-most Multi-Res block. The feature map s has the coarsest features as compared to the outputs of the other encoder branches at shallower depths and hence serves as the best candidate for this global feature vector. Also, since it learns the finest features, this Attention Gating Signal informs the Attention Gating Block to give a higher focus to relevant features on a global scale. It is passed through a $\omega_{1\times1}$ (ω_{GS}) convolution operation for spatial information extraction. The output is passed through a ReLU activation function. The resulting signal is the Attention Gating Signal. b_{GS} is the bias

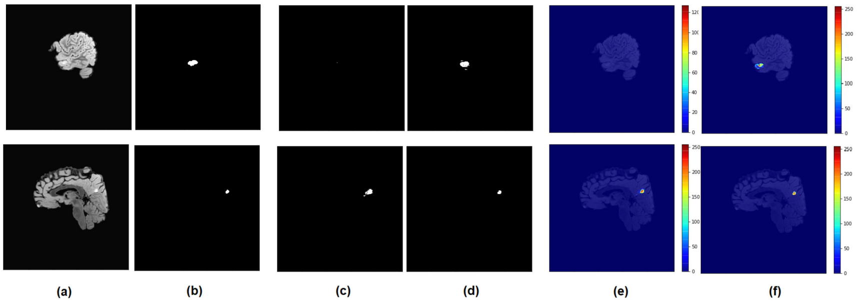

Fig. 3. Ablation Study comparing AB7 and AB5 (a) is the original image. (b) is the ground truth (c,d) is the predicted segmented mask of AB7 and AB5 respectively and (e,f) is the grad cam results of AB7 and AB5 respectively.

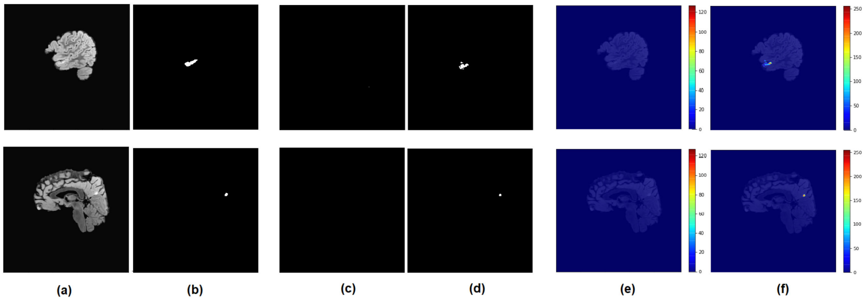

Fig. 4. Ablation Study comparing AB6 and AB4 (a) is the original image. (b) is the ground truth (c,d) is the predicted segmented mask of AB6 and AB4 respectively and (e,f) is the grad cam results of AB6 and AB4 respectively.

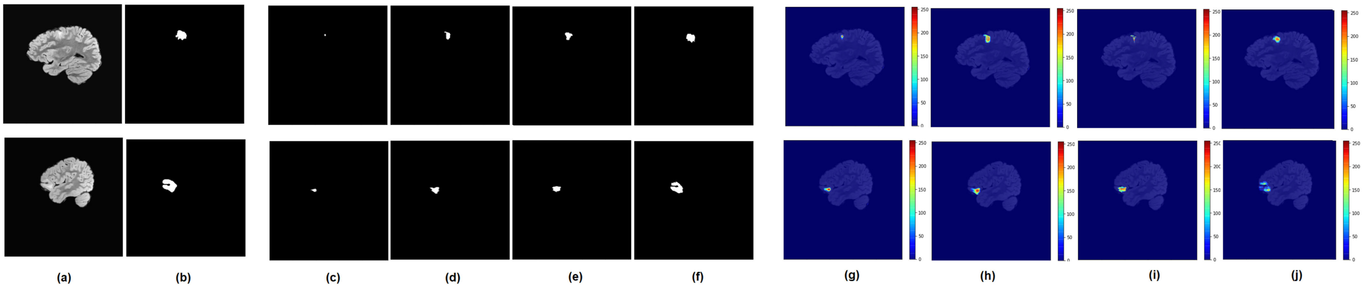

Fig. 5. Results of the semantic segmentation on MRI scans of FCD patients. (a) is the original image, (b) is the ground truth, (c), (d), (e) and (f) is the predicted mask and (g), (h), (i) and (j) is the GradCam activation map of [10], [12], [11] and proposed model respectively.

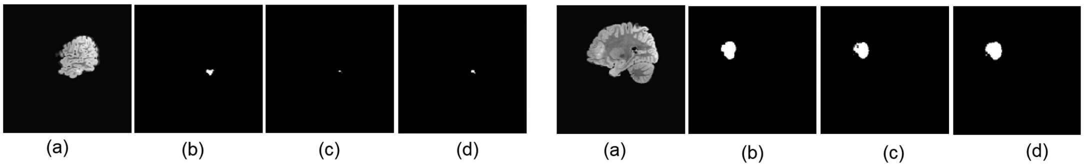

Fig. 6. Results of post-processing on semantic segmentation on MRI scans of FCD patients for the proposed model. (a) Input Image, (b) Ground truth, (c) Prediction mask before post-processing, and (d) Prediction mask after post-processing.

and is represented in Eq. 5.

$$\mathrm{GS} = \mathrm{ReLU}(\omega_{GS}(s) + b_{GS}) \tag{5}$$

This Attention Gating Signal is passed to the Attention Gating Block where the other input u_i is from the Multi-Res block on the encoder branch at depth-1 and depth-2 respectively. Upsampling of GS is done to make the dimension of u_i the same as GS. Inside the Attention Gating Block, GS is passed through a series of convolution operations and activation functions as shown in Fig 2. The output is the pre-attention coefficient, λ_{pre} as given

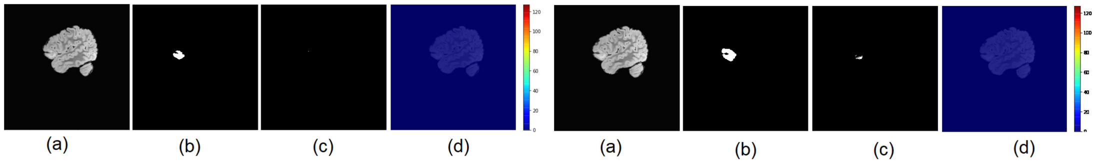

Fig. 7. Results of the semantic segmentation showing the limitations of the proposed model on MRI scans of FCD patients.(a) is the original image, (b) is the ground truth, (c) is the predicted mask and (d) is the GradCam activation map.

TABLE I
ABLATION STUDIES OF PROPOSED ARCHITECTURE FOR SPLIT 1 OF FCD DATA IN PERCENTAGE RATIOS

Study	Description	Dice	Precision	Recall
AB1	$M - RP_2 + AG_2^{[e3]}$	61.41	74.65	52.17
AB2	$M - RP_1 + AG_1^{[e3]}$	61.77	69.43	55.62
AB3	$M - RP_1 - RP_2 + S_1 + AG_2^{[e3]}$	61.14	75.98	51.15
AB4	$M - RP_1 - RP_2 + AG_1^{[e3]} + S_2$	62.49	74.34	53.90
AB5	$M - RP_1 - RP_2 + AG_1^{[e3]} + AG_2^{[e3]}$	63.69	73.43	56.48
AB6	$M - RP_1 + AG_1^{[d2]}$	61.59	73.78	52.85
AB7	$M - RP_1 - RP_2 + AG_1^{[d2]} + AG_2^{[e3]}$	59.11	59.43	58.79
AB8	$M - RP_0 - RP_1 - RP_2 + S_0 + AG_1^{[e3]} + AG_2^{[e3]}$	60.35	73.59	50.69

below:

$$\lambda_{pre} = ReLU(\omega_B(ReLU(\omega_A(GS) + \theta_{2\times 2}.u_i + b_A)) + b_B) \tag{6}$$

λ_{pre} is then convolved again by 1×1 kernel to reduce the dimensionality and comprehend additional spatial information. This is passed through the sigmoid activation function. This makes the attention maps sum to 1 which keeps the output normalized preventing exploding gradient problem. It is then upsampled ($U_{\uparrow}$) to get the attention factor λ (Eq. 7). This λ is then multiplied by the original output feature map from the Multi-Res block to get the final output of Attention Gating Block (AGB_{out}) as described in Eq. 8. In each λ, the extracted complementary information is fused with u_i to determine the output of skip connection.

$$\lambda = U_{\uparrow}(\sigma(\omega_C(\lambda_{pre}) + b_C)) \tag{7}$$

$$AGB_{out} = \omega_D(\lambda \times u_i) + b_D \tag{8}$$

Finally, the resultant feature map gives attention to the salient features of the input image. The output is then concatenated with the feature maps of the previous layer and is upsampled in the decoder blocks. The attention coefficient filters the activations in the backward pass as well as in the forward pass (Eq. 9).

$$\frac{\partial(\lambda \times u_i(\theta))}{\partial\theta} = \lambda.\frac{\partial u_i(\theta)}{\partial\theta} + u_i(\theta).\frac{\partial\lambda}{\partial\theta} \tag{9}$$

This property aids the model in learning the coarsest features specifically and has a significant effect on the weight updation of other kernels. According to Eq. 9, the attention coefficient λ directly affects the derivative in weight updation. The effectiveness of GS from shallower depths reduces as this gradient factor decreases. Attention Gating Block is not used across depth-0 because the upsampling of GS for matching the dimensions of shallower layers that focus on specific pixel information makes it sparser, and hence the information will mislead the Attention Gating Block due to the creation of new artifacts that come from zero-padded strides in this process. As the depth between the Attention Gating Block and GS increases, weight updation gets adversely affected as $\frac{\partial\lambda}{\partial\theta}$ deviates the upsampled GS from the original GS by a huge factor. Due to the above reasons, we propose the use of Attention Gating Block in skip connections of depth-1 and depth-2 that take GS from the deepest layer only. With reference to the Attention mechanism used, it is to be noted that the choice of the feature maps used for the GS and the placement of Attention Gating Blocks are the key differences of the proposed model with respect to [12] and is a crucial design choice which we have investigated based on the Ablation Studies provided in the next section.

D. Ablation Studies

As part of this work, we conducted ablation studies to further confirm the correctness of our hypothesis that guided the design choices that we adopted in the proposed architecture. We prepare eight models starting with the customized depth-3 MultiResUnet model (M) [11] where the components added (+) and deleted (-) to characterize the Ablation study networks are denoted as follows: S_x and RP_x denotes a simple skip connection and Res path resp., used across depth-x and $AG_m^{[ab]}$ represents an Attention Gating Block used as skip connection across depth-m and derives its GS from depth-b encoder block (decoder block, resp.) for a = e (d, resp.).

For demonstration, we have presented the results obtained after training the ablation (AB) models on Split 1 of the FCD dataset (refer Section V for Dataset description and experimental setup). Table I shows the Dice, Recall, and Precision values obtained after testing these models. We observe that AB1 and AB3 produce comparable results, and the same pattern is observed for AB2 and AB4. This confirms our hypothesis that the improvement in performance due to the ResPaths across the deepest layers of the architecture is very less. This calls for a more superior design choice for these two skip connections. Next, we observe an improvement in the overall results in AB4 as compared to AB6. This can be accounted for by the fact that the Gating Signal used in the Attention Gating Block in AB4 uses the coarsest feature maps derived from the deepest encoder layer and is a better substitute than deriving the Gating Signals from the features learned from the shallower depths. The same analogy can be used to explain the trend observed for AB5 and AB7 (resembles the attention mechanism used in [12]). To further emphasize this point, we present the qualitative analysis of few FCD samples using Fig. 3 and Fig. 4. Fig. 3 (row 1(c)) shows a case where AB7 failed to identify the salient features representing the lesion due to the incorrect choice of feature maps provided to the GS and Fig. 3 (row 2,(c)) depicts the case where the non-optimal choice of GS resulted in the model being unable to suppress the irrelevant features that characterize the lesion, consequently over segmenting the lesion area. The same observations can be used to justify the inferior learning capability of AB6 in contrast to AB4, as seen in Fig. 4. Finally, we observe an overall dip in the results of AB8 as compared to AB5. In contrast to the low level feature maps of depth-0 encoder layer, those corresponding to the depth-0 decoder layer are of much higher level and are guided by two Attention Gates between the deepest layers of the architecture. Therefore, the presence of ResPath is most prominent across depth-0 as the corresponding feature maps undergo significantly higher amount of processing in comparison to the other depths of the architecture. We note that the right placement of Attention Gates and ResPaths in AB5 complement each other and attains the best performance.

We observe an incremental trend in the obtained average Dice values by adding the Attention Gates across depth-1 (AB1), depth-1, and depth-2 (AB5) such that the GS is derived from the encoder layer at depth-2 for both the Attention Gates. Consequently, we adopt the design pattern of AB5 in the proposed architecture based on Dice, Precision, and Recall values.

E. Proposed Architecture

The structure of the proposed model is shown in Fig 1. The model is of depth-3 and starts with an input image of dimensions $256 \times 256 \times 1$. The number of feature maps at each depth is 32, 32, 64 and 128 times a factor α with increasing depth respectively. This factor is used to ensure that the trainable parameters for corresponding encoder layers in the proposed model, is comparable to that of the baseline architecture. A value of 1.67 for α results in the proposed model having around 50% of baseline parameters. Multi-Res blocks constitute the encoder side of the network. Each of these blocks until depth-2 is followed by a 2×2 MaxPool layer with stride 2. The skip connection between depth-0 encoder and decoder layers is ResPath.

The output feature map from Multi-Res block at depth-3 is used as the global Gating Signal. The output of the attention blocks are then concatenated to the decoder branch as a skip connection followed by upsampling using transposed convolutions in the decoder side. The upsampling units at depth-1 and depth-2, take input from the below layers and output from the Attention Gating Block at the same level and concatenates them. However, in the layer with depth-0, the output from the previous block and the ResPath is concatenated. This is followed by a 1×1 convolutional kernel with sigmoid activation. All the other layers use the ReLU activation function. The proposed model has 1,036,477 trainable parameters.

The key contributions can be summarized below:

- *The optimal placement of the Respaths:* The amount of semantic gap between corresponding (same depth) encoder and decoder layer feature maps is likely to decrease very slowly as we move towards the succeeding short skip connections. Therefore the Respaths are used only across the first encoder and decoder layers of the architecture due to its prominent effect in the shallower layers.
- *The optimal choice of Feature maps given as input to the Gating Signal:* The fact that the deepest layers contain the coarsest features that characterize the FCD lesions suggests that it could be the most suitable candidate to suppress irrelevant features of the preceding encoder layer while passing the feature maps to the skip connections connected to the decoder layer. Therefore we use the feature maps obtained from the deepest encoder layer as input to all the Gating Signals used within the model.
- *The optimal placement of the Attention Gates:* ResPaths across the deepest layers of the architecture can be replaced by a more superior substitute as the semantic gap is least pronounced at this depth. We propose the use of Attention Gate [12] as a viable replacement of the Res Paths only in the deepest layers of our architecture. In addition to introducing non-linearity to the skip connections, the Attention Gates can also suppress the irrelevant features and pay more attention to the salient features of the concerned data, thus improving the overall performance of the architecture.

V. Experimental Setup and Results

In this section, we describe the FCD dataset, pre-processing and post-processing techniques used for the FCD segmentation task, the optimizers and loss functions and hyper-parameters used for training, the evaluation metric used and the methodology of k-fold cross-validation conducted to evaluate the results. Finally, we present and compare the results obtained after training the baseline model, other standard models, and the proposed architecture on the FCD dataset.

A. Dataset Description

We have conducted all the experiments on the dataset acquired from the Sree Chitra Tirunal Institute for Medical Sciences and Technology (SCTIMST), Trivandrum, India. This study was conducted with the approval of the institutional ethics committee of SCTIMST (No. IEC/1073). The dataset comprises of 3D FLAIR weighted sequence images of 26 patients. These images were acquired on the 3 T scanner (GE Healthcare, UK) and contain 320 sagittal plane slices of thickness 1 mm and pixel spacing of 0.5 mm per patient. The TR/TE/TI/flip angle used was 7200 ms/117.241 ms/1936 ms/90° respectively. We excluded a few initial and final slices of each patient containing empty frames or very few brain pixels (below a threshold) based on the automated thresholding technique.

B. Pre-Processing/ Post-processing

A couple of pre-processing steps were carried out on the raw MRI images before feeding it to the model. The first one is the denoising step, where the MRI images were denoised using the BM3D algorithm [31], [32] for the reduction of noise. This is followed by the skull stripping process in which the skull is stripped from the brain cortex using the FSL-BET toolbox [33]. All the frames were resized to 256×256 and were normalized to have a mean of zero and unit variance.

A thresholding factor of 0.5 was applied to the predicted mask to obtain a binary output. We then pass it through a 3×3 averaging filter and round off the resultant values as part of the post-processing step of the predicted mask. This step aids in reducing the False Negative pixels in the output mask. Moreover, when we compared the post-processed binary output and the labeled ground truth, we found that a 3×3 averaging filter empirically provides better results as compared to 5×5 filters for both the baseline and the proposed architecture. The impact of post-processing on the results is further discussed in sub-section F.

C. Training Methodology

The pre-processed dataset is divided based on the 80–20% rule into the train (15% of this data is set as the validation data) and test sets. The training set comprised 4302 frames (18 patients), the validation data had 1075 frames (4 patients), and the testing data consisted of 1194 frames (4 patients). To ensure that the model is not biased towards a particular test set, we also did a 5-fold cross validation. Moreover, we used real-time data augmentation of the training data to reduce the problem of over-fitting. For training all the models we have chosen to represent loss as the summation of Dice Loss (DL)(Eq. 11), [34] and the Binary Cross-Entropy function (BCE)(Eq. 10), [35] similar to [10].

$$\mathrm{BCE} = -\sum_{i=1}^{t}(\alpha_i \log(\beta_\mathrm{i}) + (1-\alpha_\mathrm{i})\log(1-\beta_\mathrm{i})) \tag{10}$$

Here, t represents the total number of pixels in the image. α_i represents the ground-truth value of pixel i and β_i is the predicted value of pixel i of image.

$$\mathrm{DL} = 1 - \frac{\sum_{i=1}^{t}\alpha_i\beta_i + \epsilon}{\sum_{i=1}^{t}\alpha_i + \beta_i + \epsilon} - \frac{\sum_{i=1}^{t}(1-\alpha_i)(1-\beta_i) + \epsilon}{\sum_{i=1}^{t} 2 - \alpha_i - \beta_i + \epsilon} \tag{11}$$

To minimize the BCE loss we have trained the model using the Adam optimizer [36] with default hyper-parameters: $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-8}$ and learning rate $= 1 \times 10^{-3}$.

We used the Xavier uniform initializer [37] to initialize the weights in all the models.

D. Evaluation Metric

We use the Precision, Recall, and Dice coefficient for quantitatively analyzing the performance of the baseline model and the proposed model. Precision is defined as the ratio of True Positives (TP) to the sum of TP and False Positives (FP). Recall is the ratio of TP to the sum of TP and False Negatives (FN). Here, TP is the number of pixels in the output mask that are correctly predicted as lesion pixels, whereas FP is the number of pixels falsely predicted as lesion pixels. FN refers to the number of lesion pixels that are incorrectly predicted by the model as non-lesional pixels. To measure the extent of similarity between the predicted mask and ground truth, we use the Dice coefficient [38] as the evaluation metric, as described in Eq. 12.

$$\text{Dice Coefficient} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{12}$$

TABLE II
PARAMETER DETAILS OF FCN MODELS

Model	Parameters
UNet [10]	1,926,433
MultiResUnet [11]	1,778,658
AttentionUnet [12]	2,364,356
Proposed Model	**1,036,477**

E. k-Fold Cross Validation

We use 5-fold cross-validation to evaluate the performance of the models used within our study. We have divided the dataset into five folds (splits). For each of these folds, we evaluated the models in the following manner. First, we trained all the models from scratch for 250 epochs and recorded the Dice coefficient obtained from the epoch that corresponded to the best validation loss. To ensure the reproducibility of the results, we repeated the above experiment 5 times. Finally, we averaged these 5 Dice coefficient values and recorded it as the result of the fold. Similarly, we conducted experiments for all the five-folds and finally averaged each of the folds' resultant Dice coefficient to present the final result.

F. Results

We analyze the baseline, the proposed model, and the two other closely related standard models [11], [12], both quantitatively and qualitatively. The quantitative evaluation comprises of patient-wise, region-wise, and pixel-wise analysis. We use Precision, Recall, and the Dice coefficient values obtained from the 5-fold cross-validation for this analysis. All the implementations were done using Keras, with Tensorflow as the back-end. Experiments were conducted on a desktop machine with a 64-bit Ubuntu 18.04 OS, Intel Xeon(R) Gold5120 CPU @2.20 GHz 28, 64 GB of RAM and NVIDIA Quadro P5000 GPU with 16 GB dedicated memory.

The proposed model is benchmarked against the baseline model, Attention Unet [12] and MultiResUnet model [11] for the FCD dataset. The standard models were scaled to depth-3 to ensure a fair comparison. Table II presents the total trainable parameters of all the models used in this comparative study. It is noteworthy to mention that the compared models have almost double the number of parameters of that of the proposed model. Table III summarizes the pixel-wise results obtained after performing 5-fold cross-validation for all the models. As

TABLE III
PIXEL-WISE RESULTS OF THE BASELINE, RELATED STANDARD MODELS AND THE PROPOSED MODEL PRIOR TO POST-PROCESSING IN PERCENTAGE RATIOS

Splits	UNet [10]			AttentionUnet [12]			MultiResUnet [11]			Proposed Model		
	Dice	Precision	Recall	Dice	Precision	Recall	Dice	Precision	Recall	Dice	Precision	Recall
Split1	58.89	72.66	49.99	52.57	75.55	41.30	61.07	77.67	51.25	63.69	73.43	56.48
Split2	40.02	80.71	26.83	40.49	84.04	26.83	39.74	79.97	26.49	42.44	84.89	28.31
Split3	66.97	82.25	57.76	75.70	75.49	76.21	70.52	80.91	63.06	72.86	78.81	67.81
Split4	69.72	80.44	62.03	70.64	77.32	65.15	72.79	76.65	69.73	72.07	77.07	68.03
Split5	53.56	83.11	40.30	49.98	75.53	37.79	46.92	83.08	32.84	53.02	79.66	40.59
Average	**57.83**	**79.83**	**47.38**	**57.87**	**77.49**	**49.45**	**58.20**	**79.65**	**48.67**	**60.81**	**78.77**	**52.24**

TABLE IV
PIXEL-WISE RESULTS OF THE BASELINE, RELATED STANDARD MODELS AND THE PROPOSED MODEL AFTER POST-PROCESSING USING 3 × 3 AVERAGE FILTERING IN PERCENTAGE RATIOS

Splits	UNet [10]			AttentionUnet [12]			MultiResUnet [11]			Proposed Model		
	Dice	Precision	Recall	Dice	Precision	Recall	Dice	Precision	Recall	Dice	Precision	Recall
Split1	58.14	59.87	56.96	52.48	76.17	41.05	60.54	64.57	57.48	62.55	59.91	65.63
Split2	42.66	72.06	30.67	40.38	84.30	26.71	41.99	70.97	29.89	45.67	73.67	33.12
Split3	69.34	74.12	66.46	75.74	75.65	76.14	71.73	71.96	72.33	72.54	68.65	77.00
Split4	70.77	71.42	70.67	70.66	77.55	65.01	71.85	67.38	77.45	71.49	66.72	77.33
Split5	57.32	74.85	47.42	50.03	76.31	37.78	52.03	75.75	39.85	56.81	71.57	48.05
Average	**59.64**	**70.46**	**54.44**	**57.86**	**78.00**	**49.34**	**59.63**	**70.13**	**55.40**	**61.81**	**68.10**	**60.23**

TABLE V
REGION-WISE RESULTS OF THE BASELINE, RELATED STANDARD MODELS AND THE PROPOSED MODEL IN PERCENTAGE RATIOS

Splits	UNet [10]			AttentionUnet [12]			MultiResUnet [11]			Proposed Model		
	Dice	Precision	Recall	Dice	Precision	Recall	Dice	Precision	Recall	Dice	Precision	Recall
Split1	63.57	91.44	48.86	61.67	94.17	46.08	63.86	95.36	48.35	77.91	94.00	66.58
Split2	44.25	78.50	31.61	45.46	77.64	32.32	43.29	76.45	30.54	57.04	74.21	46.79
Split3	85.54	95.54	77.50	90.01	95.51	85.16	83.93	96.31	74.53	84.81	90.47	79.84
Split4	88.68	96.86	82.10	87.27	96.68	79.72	88.08	94.85	82.38	87.27	93.25	82.10
Split5	75.60	97.26	62.35	66.71	94.31	52.10	70.41	96.82	55.68	71.05	87.92	60.12
Average	**71.53**	**91.92**	**60.48**	**70.22**	**91.66**	**59.08**	**69.91**	**91.96**	**58.30**	**76.62**	**87.97**	**67.09**

observed, the proposed model outperforms the baseline model in most of the splits and achieves an overall absolute improvement of 2.98% in Dice value. It attains an absolute increase of 4.86% in Recall rate and a comparable Precision value to the baseline architecture. Moreover, it can be clearly observed that the proposed model outperforms the other two standard models by achieving an absolute improvement of 2.61% from [11] and 2.94% from [12]. We note that the performance of all models is considerably low in Split2 and Split5 as compared to the others. This can be attributed to the fact that these splits contain more negative frames with a very less cross-sectional area of the brain. There exists an imbalance in the proportion of positive to negative samples in these splits, which impacts the training process adversely. This, in turn, reduces the Recall rate and consequently the Dice value decreases. From Table III, it can be noted that the variance between the Recall and Precision values are very high due to the under segmentation over the lesion boundaries. The FCD features get diluted over these boundaries as a result of which the models often fail to identify such pixels as positive. This could account for the comparatively less Recall obtained.

In addition, an upswing in the overall results is detected after post-processing, as seen in Table IV. The post-processing is mainly targeted to improve the Recall performance without creating many false positives. It uses a 3 × 3 average filtering over the prediction mask from the segmentation model. This helps to reduce small holes in the predicted mask, and slightly dilate the predicted mask boundaries. However, Precision values slightly reduce after post-processing due to the increase in the number of false positives. While analyzing Table III and IV, it can be concluded that in most of the cases, the post-processing helps to improve the overall segmentation performance in terms of Recall and Dice.

We have presented the region-wise and patient-wise analysis in Table V and Table VI, respectively. The region-wise analysis generalizes the frame-wise parameter evaluation and is a good measure to validate how well the proposed model makes frame-wise predictions. Evidently, the proposed model is

TABLE VI
PATIENT-WISE RESULTS OF THE BASELINE, RELATED STANDARD MODELS AND THE PROPOSED MODEL IN PERCENTAGE RATIOS

Splits	**Recall**			
	UNet [10]	**AttentionUnet [12]**	**MultiResUnet [11]**	**Proposed**
Split1	60.00	55.00	50.00	90.00
Split2	40.00	55.00	40.00	75.00
Split3	100.00	100.00	100.00	100.00
Split4	100.00	100.00	100.00	100.00
Split5	95.00	100.00	80.00	95.00
Average	**79.00**	**82.00**	**74.00**	**92.00**

more robust as it achieves a region-wise absolute improvement of 5.09%, 6.71% and 6.40% in Dice value in comparison to the baseline, [11] and [12] models respectively. The proposed architecture also attains the highest overall Recall rate of 92% for patient-wise analysis. We conclude from the above results that it is able to capture more positive FCD frames as compared to the other models in our study. However, it slightly falls short of the other models in terms of identifying the true extent of FCD lesions within frames. The above two observations can be explained by taking into account the design considerations of the proposed model. The presence of Attention Gates with optimally derived GS allowed our architecture to learn the salient features, which in turn enabled it to recognize more positive FCD frames. As the trainable parameters of the baseline and the other standard models are almost double the number of the proposed model (Table II), they demonstrate a slightly better ability to learn the extent of the FCD lesion given a positively identified FCD frame. However, the presence of the Multi-Res blocks and the careful placement of the ResPath and Attention Gating Blocks enables the proposed model to effectively capture the features that identify the diverse FCD regions using a lesser number of parameters as compared to the standard models and achieve a comparable Precision value. Hence, we observe a clear trade-off between the obtained Precision values and the model complexity. Although a better trend in Recall rates is observed for attention-based models in our comparative study, the proposed model achieves the highest Recall rate, which can be explained again due to the right design choices for optimally choosing the Gating Signal and the placement of Attention Gates (Section IV - D).

Fig 5. shows the qualitative analysis of all the models on selected FCD frames before post-processing. In Fig. 5 (row 1), we observe that [12] fails to capture the characteristics of the FCD lesion accurately and under segments the area of interest. While [11] discovers a greater area of the affected region, it fails to delineate the true extent of the lesion boundary accurately. Although [10], [12] and [11] correctly detects the lesion in (Fig. 5 (row 2)), they segment only a small portion of the FCD lesion. The proposed model combines the best of both models by optimizing [12] (optimal choice of GS), and [11] (Multi-Res blocks and optimal placement of Respaths) architectures and obtains a comparatively superior segmentation result (Fig. 5(f)).

Fig. 6 presents the post-processed frames corresponding to two segmented masks of the proposed model. The first case depicts the scenario in which post-processing slightly dilates the predicted lesion by increasing the boundary of the lesion centered around the area of interest. Consequently, a slight increase in the True Positives is observed. In the second case, it fills the holes in the predicted mask and reduces the False Negatives. The limiting results of the proposed model are shown in Fig 7. In the first scenario, it fails to detect the presence of the FCD lesion, and in the second one, it is unsuccessful in segmenting out the true extent of the lesion. This behavior can be ascribed to the high similarity between the lesion pixels and its surrounding regions.

Grad-CAM [39], which is a Gradient-based Localization technique, is used to produce a heatmap image that indicates the regions of the input image whose change would most contribute towards maximizing the output of the layer being analyzed. Fig. 5(j) depicts the heatmap corresponding to the final layer for selected slices of the dataset as input for the proposed model. Higher intensity of the heatmap is concentrated on the FCD region, and the remaining regions are blurred out subsequently. This behavior is justified due to the optimal placement of attention blocks in the proposed architecture.

To further examine the robustness of the proposed model in comparison to the other standard models used in this study, we train all the architectures from scratch and test it on the publicly available Low-Grade Glioma Segmentation Dataset and the results are depicted in Section I of the Supplementary Material.

VI. CONCLUSION

This work presents a hybrid skip connection based architecture for the semantic segmentation of the FCD dataset. We started by contemplating over the drawbacks of the baseline architecture and came up with three potential flaws, namely, its inability to capture scale variability, the semantic gap between feature maps of shallowest and deepest layers, and lack of attention to salient features. We address the aforementioned issues in the proposed model by the optimal placement of Respaths, Attention Gates, and the usage of Multi-Res blocks. We conducted both quantitative and qualitative analysis to assess the performance of our model. Experimental evaluations show that the proposed model outperforms the baseline model and other standard models used in our study, with only 50% of the other models' total parameters. Experimental results underpin the superiority of the proposed architecture over the baseline model and pave the way for future research on the applications of hybrid skip connections. We believe the proposed model can also be considered a strong candidate for the segmentation of other types of lesions in the brain and also serve as a useful tool that can assist neuro-radiologists in screening patients with intractable epilepsy.

REFERENCES

[1] S. H. Kim and J. Choi, "Pathological classification of focal cortical dysplasia (FCD): Personal comments for well understanding fcd classification," *J. Korean Neurosurg. Soc.*, vol. 62, no. 3, pp. 288–295, 2019.

[2] S. K. Lee and D.-W. Kim, "Focal cortical dysplasia and epilepsy surgery," *J. Epilepsy Res.*, vol. 3, no. 2, pp. 43–47, 2013.

[3] D. Taylor, M. Falconer, C. Bruton, and J. Corsellis, "Focal dysplasia of the cerebral cortex in epilepsy," *J. Neurol. Neurosurg. Psychiatry*, vol. 34, no. 4, pp. 369–387, 1971.

[4] A. Palmini *et al.*, "Terminology and classification of the cortical dysplasias," *Neurology*, vol. 62, no. 6 suppl 3, pp. S2–S8, 2004.

[5] A. Barkovich, R. Kuzniecky, G. Jackson, R. Guerrini, and W. Dobyns, "A developmental and genetic classification for malformations of cortical development," *Neurology*, vol. 65, no. 12, pp. 1873–1887, 2005.

[6] I. Blümcke *et al.*, "The clinicopathologic spectrum of focal cortical dysplasias: A consensus classification proposed by an ad hoc task force of the ilae diagnostic methods commission 1," *Epilepsia*, vol. 52, no. 1, pp. 158–174, 2011.

[7] P. Widdess-Walsh, B. Diehl, and I. Najm, "Neuroimaging of focal cortical dysplasia," *J. Neuroimag.*, vol. 16, no. 3, pp. 185–196, 2006.

[8] S. Fauser *et al.*, "Multi-focal occurrence of cortical dysplasia in epilepsy patients," *Brain*, vol. 132, no. 8, pp. 2079–2090, 2009.

[9] C.-A. Yang, M. Kaveh, and B. J. Erickson, "Automated detection of focal cortical dysplasia lesions on t1-weighted MRI using volume-based distributional features," in *Proc. IEEE Int. Symp. Biomed. Imag.: From Nano to Macro*, 2011, pp. 865–870.

[10] K. B. Dev, P. S. Jogi, S. Niyas, S. Vinayagamani, C. Kesavadas, and J. Rajan, "Automatic detection and localization of focal cortical dysplasia lesions in MRI using fully convolutional neural network," *Biomed. Signal Process. Control*, vol. 52, pp. 218–225, 2019.

[11] N. Ibtehaz and M. S. Rahman, "MultiresuNet: Rethinking the u-net architecture for multimodal biomedical image segmentation," *Neural Netw.*, vol. 121, pp. 74–87, 2020.

[12] J. Schlemper *et al.*, "Attention gated networks: Learning to leverage salient regions in medical images," *Med. Image Anal.*, vol. 53, pp. 197–207, 2019.

[13] K. Boonyapisit *et al.*, "Epileptogenicity of focal malformations due to abnormal cortical development: Direct electrocorticographic–histopathologic correlations," *Epilepsia*, vol. 44, no. 1, pp. 69–79. 2003. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1528-1157.2003.08102.x

[14] O. Colliot, T. Mansi, N. Bernasconi, V. Naessens, D. Klironomos, and A. Bernasconi, "Segmentation of focal cortical dysplasia lesions on MRI using level set evolution," *NeuroImage*, vol. 32, no. 4, pp. 1621–1630, 2006.

[15] O. Colliot, S. B. Antel, V. B. Naessens, N. Bernasconi, and A. Bernasconi, "In vivo profiling of focal cortical dysplasia on high-resolution MRI with computational models," *Epilepsia*, vol. 47, no. 1, pp. 134–142.

[16] N. K. Focke, S. B. Bonelli, M. Yogarajah, C. Scott, M. R. Symms, and J. S. Duncan, "Automated normalized flair imaging in MRI-negative patients with refractory focal epilepsy," *Epilepsia*, vol. 50, no. 6, pp. 1484–1490. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1528-1167.2009.02022.x

[17] P. Subashini and M. S. Jansi, "A study on detection of focal cortical dysplasia using MRI brain images," *J. Comput. Appl.*, vol. 4, no. 1, pp. 23–27, 2011.

[18] C.-A. Yang, M. Kaveh, and B. J. Erickson, "Automated detection of focal cortical dysplasia lesions on t1-weighted MRI using volume-based distributional features," in *Proc. IEEE Int. Symp. Biomed. Imag.: From Nano to Macro*, 2011, pp. 865–870.

[19] S.-J. Hong, H. Kim, D. Schrader, N. Bernasconi, B. C. Bernhardt, and A. Bernasconi, "Automated detection of cortical dysplasia type ii in MRI-negative epilepsy," *Neurology*, vol. 83, no. 1, pp. 48–55, 2014.

[20] B. Ahmed *et al.*, "Cortical feature analysis and machine learning improves detection of MRI-negative focal cortical dysplasia," *Epilepsy Behav.*, vol. 48, pp. 21–28, 2015.

[21] M. El Azami, A. Hammers, N. Costes, and C. Lartizien, "Computer aided diagnosis of intractable epilepsy with MRI imaging based on textural information," in *Proc. Int. Workshop Pattern Recognit. Neuroimag.*, 2013, pp. 90–93.

[22] J. Rajan, K. Kannan, C. Kesavadas, and B. Thomas, "Focal cortical dysplasia (fcd) lesion analysis with complex diffusion approach," *Computerized Med. Imag. Graph.*, vol. 33, no. 7, pp. 553–558, 2009.

[23] P. Besson, O. Colliot, A. Evans, and A. Bernasconi, "Automatic detection of subtle focal cortical dysplasia using surface-based features on MRI," in *Proc. 5th IEEE Int. Symp. Biomed. Imag.: From Nano to Macro*, 2008, pp. 1633–1636.

[24] Q. Huang *et al.*, "Semantic segmentation with reverse attention," In *Proceedings of the British Machine Vision Conference (BMVC)*, T. K. Kim, S. Zafeiriou, G. Brostow and K. Mikolajczyk, Eds., BMVA Press, Sep. 2017, pp. 18.1–18.13.

[25] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid attention network for semantic segmentation," in *BMVC*, p. 285, 2018.

[26] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3146–3154.

[27] X. Dong *et al.*, "Synthetic MRI-aided multi-organ segmentation on male pelvic CT using cycle consistent deep attention network," *Radiotherapy Oncol.*, vol. 141, pp. 192–199, 2019.

[28] Z. Wang, J. Xu, L. Liu, F. Zhu, and L. Shao, "Ranet: Ranking attention network for fast video object segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 3978–3987.

[29] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. 30st AAAI Conf. Artif. Intell.*, 2017, pp. 4278–4284.

[30] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Proc. Deep Learn. Med. Image Anal. Multimodal Learn. Clin. Decis. Support*. Springer, 2018, pp. 3–11.

[31] M. Maggioni and A. Foi, "Nonlocal transform-domain filter for volumetric data denoising and reconstruction," *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 119–133, 2013.

[32] M. Maggioni and A. Foi, "Nonlocal transform-domain denoising of volumetric data with groupwise adaptive variance estimation," in *Proc. SPIE*, vol. 8296, p. 82960O, Feb. 2012.

[33] S. M. Smith, "Fast robust automated brain extraction," *Hum. Brain Mapp.*, vol. 17, no. 3, pp. 143–155, 2002.

[34] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Proc. Deep Learn. Med. Image Anal. Multimodal Learn. Clin. Decis. Support*. Springer, 2017, pp. 240–248.

[35] J. Shore and R. Johnson, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy," *IEEE Trans. Inf. Theory*, vol. 26, no. 1, pp. 26–37, 1980.

[36] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations (ICLR)*, Y. Bengio and Y. LeCun, eds., San Diego, CA, USA, May 7–9, 2015.

[37] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.

[38] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[39] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.