TW-Net: Transformer Weighted Network for Neonatal Brain MRI Segmentation

Shengjie Zhang, Bohan Ren, Ziqi Yu, Haibo Yang, Xiaoyang Han, Xiang Chen, Yuan Zhou, *Member, IEEE*, Dinggang Shen, *Fellow, IEEE*, and Xiao-Yong Zhang, *Member, IEEE*

***Abstract*—Accurate neonatal brain MRI segmentation is valuable for investigating brain growth patterns and tracking the progression of neurodevelopmental disorders. However, it is a challenging task to use intensity-based methods to segment neonatal brain structures because of small contrast differences between brain regions caused by the inherent myelination process. Although convolutional neural networks offer the potential to segment brain structures in an intensity-independent manner, they suffer from lack of in-plane long-range dependency which is essential for the segmentation. To solve this problem, we propose a novel Transformer-Weighted network (TW-Net) to incorporate in-plane long-range dependency information. TW-Net employs a conventional encoder-decoder architecture with a Transformer module in the middle. The Transformer module uses a rotate-and-flip layer to better calculate the similarity between two patches in a slice to leverage similar patterns of geometrical and texture features within brain structures. In addition, a deep supervision module and squeeze-and-excitation blocks are introduced to incorporate boundary information of brain structures. Compared with state-of-the-art deep learning algorithms, TW-Net outperforms these methods for multiple-label tasks in 2D and 2.5D configurations on two independent public datasets, demonstrating that TW-Net is a promising method for neonatal brain MRI segmentation.**

***Index Terms*—Neonatal brain, segmentation, Transformer, magnetic resonance imaging (MRI)**

I. Introduction

THE newborn period is critical for the development of cognition and motor functions. It has been reported that neurodevelopmental disorders are closely associated with the development of neonatal brain tissues [1]–[5]. Magnetic resonance imaging (MRI) is a noninvasive method that enables the observation of early brain development [6]. The segmentation of the brain MRI plays an important role for quantitative measurement of regional brain structures. Manual segmentation is not only a time-consuming task [7], but also an expertise-demanding task. Therefore, automatic segmentation of neonatal brain MRI is vital for understanding the developmental trajectories of brain maturation.

Many signal intensity-based studies have been conducted to automatically segment three main different tissue classes, including white matter (WM), gray matter (GM), and cerebrospinal fluid (CSF), based on the fact that these tissues show different intensities in MRI images. However, many brain structures are formed during the newborn period, and share a similar intensity distribution in MRI images (Fig. S1), resulting in a challenging task to segment brain structures based on intensity threshold.

Deep learning (DL) methods — in particular convolutional networks such as fully convolutional networks (FCN) [8] and U-Net [9] — have been demonstrated to be promising for segmenting brain structures, overcoming the main drawback of intensity-based methods. Numerous forms of U-Net have been proposed to improve segmentation results [10]–[15]. However, it is still a difficult task for these methods to reliably segment newborn brain MRI due to the following reasons: 1) long-range dependence information is lost; 2) simple feature maps' concatenation cannot leverage enough boundary information from encoder blocks [16].

To address these issues, we present a *Transformer-weighted network* (TW-Net) for the segmentation of multiple neonatal brain structures, including WM, GM, CSF, brainstem, cerebellum, hippocampus, deep gray matter, and ventricle. The proposed TW-Net employs a conventional encoder-decoder architecture with a Transformer module in the middle. The Transformer module divides the feature maps from the encoder into multiple patches to combine the high-level information. Specifically, the architecture has the following features.

First, we introduce a Transformer module to capture long-range dependence information. In the Transformer module, we employ a *rotate-and-flip* (RF) layer that rotates and flips the patches prior to feeding them to the self-attention mechanism.The capability of the self-attention mechanism relies on the similarity calculation of input patches. The RF layer enhances the similar patterns of geometrical and texture features [17], and hence improves the similarity calculation. Second, we introduce *squeeze-and-excitation* (SE) blocks and a deep supervision module for leveraging boundary information. The SE block is used to aggregate the outputs of decoders with multiple scales. The deep supervision module uses the low-

This work was supported in part by grants from the National Natural Science Foundation of China (81873893,82171903,92043301), Shanghai Municipal Science and Technology Major Project (2018SHZDZX01), and the Office of Global Partnerships (Key Projects Development Fund) at Fudan University. (Corresponding author : Dinggang Shen and Xiao-Yong Zhang)

S. Zhang, Z. Yu, H. Yang, X. Han, X. Chen and X-Y. Zhang are with the Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai, China. MOE Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence, and MOE Frontiers Center for Brain Science, Fudan University, Shanghai, China. (email: xiaoyong_zhang@fudan.edu.cn)

B. Ren is with the Department of School of Cyber Science and Technology, Beihang University, Beijing, China.

Y. Zhou is with the School of Data Science, Fudan University, Shanghai, China. (e-mail:yuanzhou@fudan.edu.cn)

D. Shen is with School of Biomedical Engineering, ShanghaiTech University, Shanghai 201210, China. He is also with Shanghai United Imaging Intelligence Co., Ltd., Shanghai 200230, China, and Shanghai Clinical Research and Trial Center, Shanghai, 201210, China. (e-mail: Dinggang.Shen@gmail.com)

The first two authors contribute equally to this work.

level encoder features with multi-scale decoder features to improve the final segmentation results.

The network architecture is implemented in 2D and 2.5D configurations. We concatenate three slices as input and call it a 2.5D configuration [18], [19]. The proposed method in 2D and 2.5D configurations is evaluated on the multi-structure segmentation task, which aims to segment all the brain regions from input slices simultaneously instead of dividing the segmentation tasks into multiple tasks that each task aims to segment one label.

II. Related Work

There are two kinds of work that are most related to our work: CNN-based models and Transformer-based models.

A. Deep Convolutional Models

Since CNNs are capable of performing pixel-wise prediction by automatically extracting features using trainable filters, numerous works — such as U-Net [9], U-Net++ [10], DenseNet [20] and PSP-Net [21] — employ a multi-path technique based on convolutional operators to extract features at various levels and contextual information. SegNet [22] is similar to U-Net, which leverages pooling indices to conduct feature mapping and achieves promising results. To enhance the performance of convolutional approaches, several algorithms combine convolution with linear layers or a recurrent mechanism to create new architectures. Conv-LSTM [23] optimizes performance by utilizing a long short term memory (LSTM). Deeplabv3+ [24] merges features at several levels using depth-wise convolution and the standard batch normalization (BN) technique. Additionally, (FCN) [8] enables a flexible patch size and sophisticated prediction by utilizing convolutional operators. To address the problem of fuzzy boundary, Zhang *et al.* [25] developed an edge-enhanced network that made use of edge attention to collect edge information from the U-Net encoder. To learn the multi-scale information of encoders, a module for weighted aggressiveness was incorporated to communicate boundary information from the encoder to the decoder.

In summary, the segmentation performance of deep convolutional models suffers from the local receptive field, which hinders the extraction of long-range dependence information from the whole brain structure. To boost the segmentation performance, these models usually rely on increasing the number of convolutional layers. The performance improvement by this operation is limited by the local nature of convolutional layers. Hence, a model with a stronger representation capability is required.

B. Transformer-based Models

It was recently established that Vision Transformer (ViT) [26] could achieve the same performance as CNN in downstream jobs. ViT applies Transformer directly to full-size images via the global self-attention mechanism. Inspired by Transformer's long-range dependency, some adjustments have been performed to make it suitable for medical image segmentation. TransUNet [27], for example, connects 12 Transformer layers to handle the convolutional encoder's high-level features. TransBTS [28] utilizes three-dimensional convolutional filters to capture local three-dimensional contextual information. Swin-UNet [29] replaces the convolution-based backbones with a pure Transformer operation. ViT improves segmentation performance by computing the similarity between two random patches extracted from MRI slices [30], which has been shown to be very effective for medical image segmentation. Despite the superior performance, transformer-based methods are constrained by the memory available on the GPUs [31], which hinders the application of three-dimensional strategies and in turn results in a loss of information across adjacent MRI slices. Note that both the simple Transformer-Convolution stack and the pure Transformer only models conduct point-to-point matrix calculation [32]. Thus, the post-processing of the patch matrix is of great importance to the segmentation task.

In summary, Transformer-based models only provide the insight that self-attention calculation could increase the segmentation performance in general. However, these models are not designed to handle our task — neonatal brain segmentation — which features low contrast across regions in MRI slices. In conclusion, a task-specific network consisting of convolutional layers and Transformer blocks is expected to solve our problem.

III. Methods

To solve the aforementioned problems, we design the TW-Net with a deep supervision module to extract the boundary information, which helps improve the representation capability of the model. Moreover, we convey the boundary information extracted from the first two layers from the encoder to the decoder, enhancing the feature fusion without leveraging too many convolutional layers. In addition, the Transformer block benefits the capture of long-range dependence information, which is significant for capturing the complete structure of brain regions Many brain regions (*e.g.*, GM) are global structures instead of only occupying a small proportion of the MRI slice. We present TW-Net in details, including the image pre-processing, its network architecture in 2D and 2.5D configurations, and a hybrid loss function.

A. Network Architecture for 2D/2.5D TW-Net

Fig. 1 illustrates the architecture of our proposed TW-Net, which consists of an encoder-decoder convolutional network, with a deep supervision module [33] in the encoder process and SE blocks [34] in the decoder process. A Transformer module is appended at the end of the CNN encoder. The module divides the output of the encoder into several patches and calculates their similarities. We design a RF-layer [35] to rotate and flip the patches for better similarity calculation. Due to the fact that the computational complexity of the Transformer is quadratic with respect to the number of the patches [36], [37], we take a down-sampling strategy to reduce

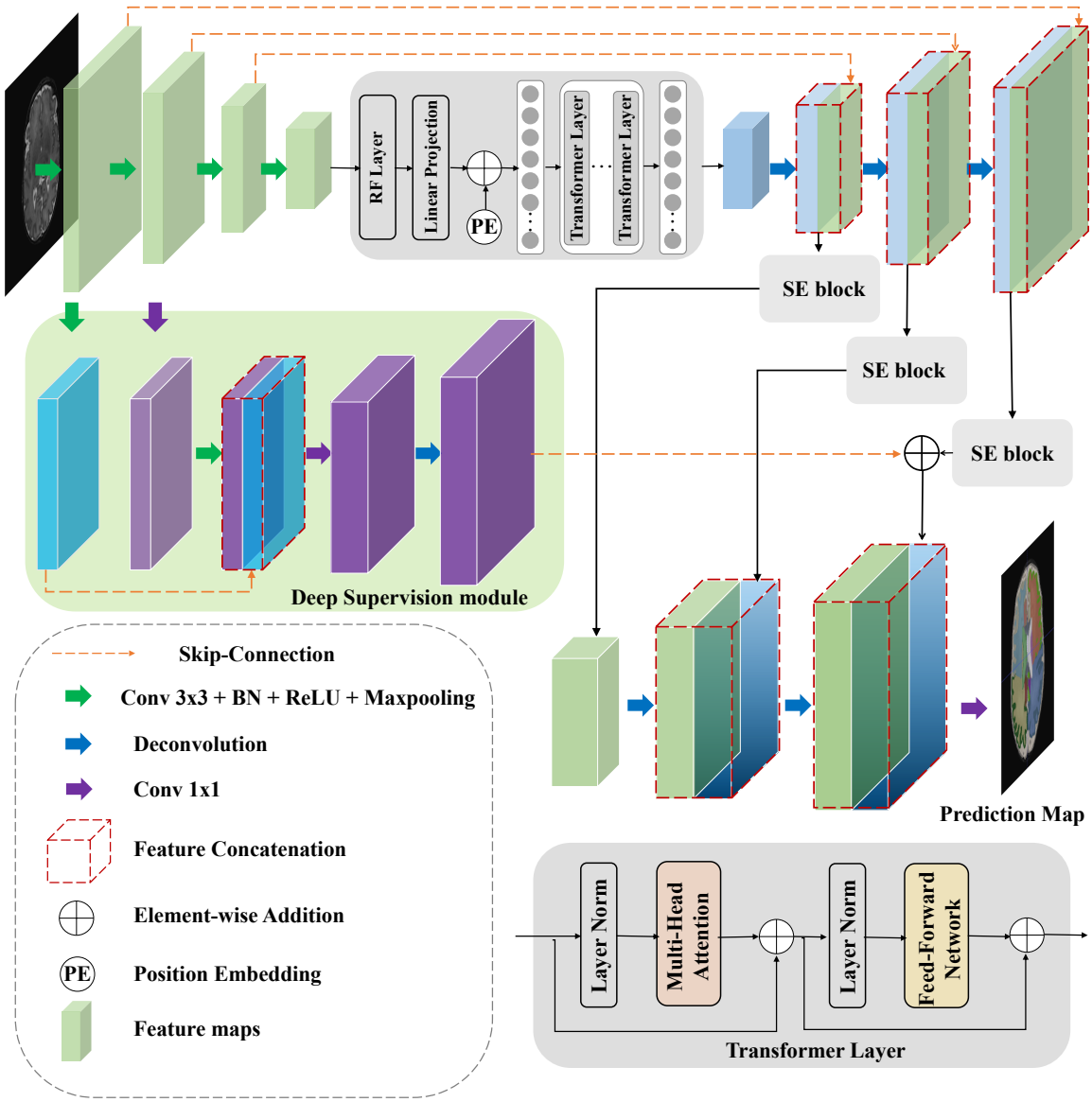

Fig. 1: The overview of the proposed TW-Net. It includes a main network (encoder-decoder with skip connection), an additional network stream (deep supervision module). The bottleneck contains a RF-layer, which rotates or flips the embedded patches before feeding them to the Transformer block. The deep supervision module (in green) is implemented to capture the boundary information and transport it to the decoder blocks

the size of the feature maps from the encoder, hence reduce the length of the patch sequence .

The Transformer module processes the high-level features while keeping the dimension. We upsample the output of the Transformer, hence the feature maps are reshaped to the same size as the second last layer of the encoder. To generate the segmentation results with the same size of the input, three CNN decoders are introduced to conduct feature upsampling operation. Cascaded upsampling is utilized to recover a full resolution segmentation result. In addition, skip-connection is employed to concatenate the features from the encoder and those from the corresponding decoder to create a finer segmentation result.

1) Encoder design: We down-sample the input slice with convolutional blocks. Let $x \in \mathbb{R}^{W\times H\times 3}$ be an input slice with two adjacent slices. The encoder $e : \mathbb{R}^{W\times H\times 3} \rightarrow \mathbb{R}^{W^{(4)}\times H^{(4)}\times C^{(4)}}$ consists of several convolutional blocks,

$$e = e^{(4)} \circ e^{(3)} \circ e^{(2)} \circ e^{(1)}.$$

Each convolutional block $e^{(i)} : \mathbb{R}^{W^{(i-1)}\times H^{(i-1)}\times C^{(i-1)}} \rightarrow \mathbb{R}^{W^{(i)}\times H^{(i)}\times C^{(i)}}$ $(W^{(0)} = W, H^{(0)} = H, C^{(0)} = 3)$ consists of a 3×3 convolutional kernel, batch normalization, ReLU activation, and a max pooling layer with stride being 2. Hence the height $H^{(0)}$ and width $W^{(0)}$ are reduced to $H^{(4)} = H^{(0)}/16$ and $W^{(4)} = W^{(0)}/16$ after 4 convolutional blocks. We denote the output of the encoder by $x^{(4)}$, i.e. $x^{(4)} = e(x)$.

We adopt a 2.5D configuration for the encoder, which uses a 3D kernel at the first block $(e^{(1)})$ followed by several 2D convolutional blocks $(\{e^{(i)} : i = 2, 3, 4\})$ [38]. Since the adjacent image slices in a MRI volume contain strong

correlative spatial features, this strategy helps capture the spatial information without using 3D convolution excessively. After the 3D convolution, we employ a 1×1 convolution to generate feature maps with size of $H^{(1)} \times W^{(1)} \times C^{(1)}$.

2) Transformer with Rotation-Flip and Patch Embedding: Let $x^{(4)} \in \mathbb{R}^{H^{(4)} \times W^{(4)} \times C^{(4)}}$ represent the feature map output from the encoder. The Transformer module first divides $x^{(4)}$ into a sequence of 2D patches $\{\hat{x}_i : i = 1, \ldots, N\}$, where $\hat{x}_i \in \mathbb{R}^{P \times P}$ and $N = \frac{H^{(4)} \times W^{(4)}}{P^2} \times C^{(4)}$ is the number of patches. The patches are input into a RF layer $f_{\text{RF}} : \mathbb{R}^{P \times P} \rightarrow \mathbb{R}^{P \times P}$ to create symmetric features. This layer produces rotated and flipped versions of $\hat{x}_i$ for the Transformer. As shown in Fig. 2, the rotation takes the integral multiple of 90° into consideration. In the implementation, each input patch $\hat{x}_i$ is convolved with 8 filters (4 rotation filters and 4 flipping filters) to produce a 8-channel feature map [39]. Since the Transformer calculates the similarity between elements of a sequence, the RF-layer is used to enhance the capture of long-range dependence information. A max-pooling across 8 channels is used afterwards to produce the output of the RF-layer. We denote the output of this layer as $f_{\text{RF}}(\hat{x}_i)$.

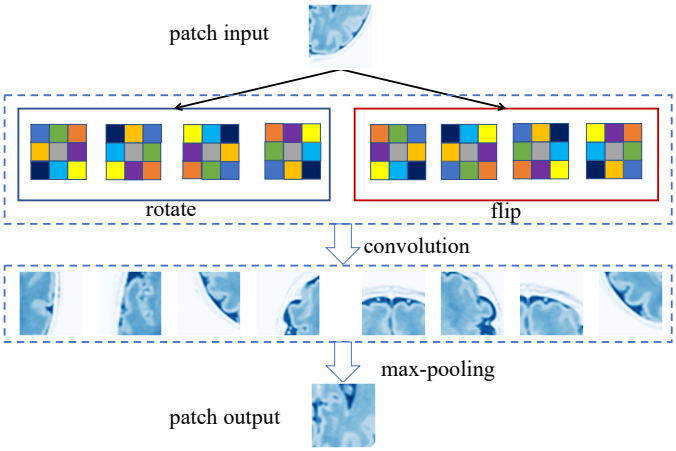

Fig. 2: Illustration of the RF-layer employed in our TW-Net. The first row is the input feature map. The second row shows the 4 rotation and the 4 flipping filters. By convolving the input feature map with the 8 filters, we get 8 different feature maps in the third row. These 8 feature maps are combined to derive the output through a max-pooling procedure.

Through a linear projection, we could map the patches into a d-dimensional embedding space. Positional embedding from [40] is employed to incorporate the spatial information. The aforementioned process is formulated as follows:

$$z^{(0)} = f_{\text{PE}}(\tilde{x}_1, \ldots, \tilde{x}_N) = [\tilde{x}_1 E; \tilde{x}_2 E; ...; \tilde{x}_N E] + E_{pos}. \quad (1)$$

where $E \in \mathbb{R}^{P^2 \times d}$ is the patch embedding projection ($d = P^2$), and $E_{pos} \in \mathbb{R}^{N \times d}$ denotes the position embedding (d copies of the same position column vector). $\tilde{x}_i \in \mathbb{R}^{1 \times P^2}$ is the flattened and transposed version of $f_{\text{RF}}(\hat{x}_i)$. The output $z^{(0)}$ has a size of $N \times d$.

The Transformer block $f_{\text{T}} : \mathbb{R}^{N \times d} \rightarrow \mathbb{R}^{N \times d}$ consists of L ($L = 6$ by default) layers of multi-head self-attention (MSA) [41] and multi-layer perceptron (MLP) operations. After the stack of $l-1$ Transformer layers, the output of the l-th layer could be computed as follows:

$$z^{(l-\frac{1}{2})} = MSA(LN(z^{(l-1)})) + z^{(l-1)} \quad (2)$$

$$z^{(l)} = MLP(LN(z^{(l-\frac{1}{2})})) + z^{(l-\frac{1}{2})} \quad (3)$$

where $LN(\cdot)$ represents the layer normalization operation and $z^{(l)}$ denotes the output of the l-th Transformer layer. The output of the Transformer block is reshaped to a tensor of size $W^{(4)} \times H^{(4)} \times C^{(4)}$ before feeding it into the decoder.

In summary, the Transformer module is composed by

$$f_{TM} = f_T \circ f_{PE} \circ \boldsymbol{f}_{RF}, \quad (4)$$

where $\boldsymbol{f}_{RF}(x^{(4)}) = (f_{RF}(\hat{x}_1), ..., f_{RF}(\hat{x}_N))$.

3) Decoder design and SE block: The CNN decoder $d : \mathbb{R}^{W^{(4)} \times H^{(4)} \times C^{(4)}} \rightarrow \mathbb{R}^{W^{(1)} \times H^{(1)} \times C^{(1)}}$ consists of 3 deconvolutional blocks,

$$d = d^{(3)} \circ d^{(2)} \circ d^{(1)},$$

each block $d^{(i)} : \mathbb{R}^{W^{(i)} \times H^{(i)} \times C^{(i)}} \rightarrow \mathbb{R}^{W^{(i-1)} \times H^{(i-1)} \times C^{(i-1)}}$ consists of a deconvolutional layer followed by a ReLU activation. The output feature maps of a block are also input into a Squeeze-and-Excitation (SE) block. The SE operation calculates a weight for each channel and multiples the feature map of each channel by this weight to produce weighted feature maps, as shown in Fig. 3. Specifically, global average pooling is first utilized to capture the global context information of the feature maps of the decoder block. Then the feature maps go through two 1×1 convolutional layers with Sigmoid as activation function [42]. These operations are designed to evaluate the channel relevance and generate a weight for each channel. Then the input feature maps are multiplied by the weights to produce an output. In summary, this operation can be seen as a channel-wise attention mechanism. We aggregate the multi-scale information from the decoder blocks by concatenating the (up-sampled) outputs from the SE blocks. The combination of the decoder and the SE blocks outputs a tensor of size $W^{(1)} \times H^{(1)} \times 2C^{(1)}$.

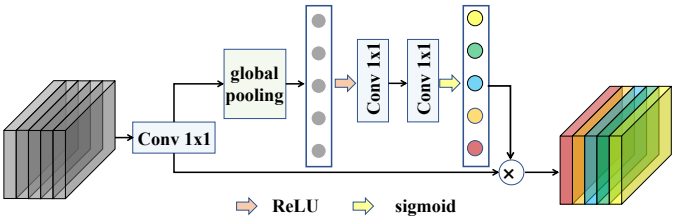

Fig. 3: Illustration of the strategy of a SE block. The global average pooling is first employed to aggregate the global context information, which is followed by two 1×1 convolutional layers with a Sigmoid activation function to create a weight for each channel. After that, these weights are multiplied with the feature maps to obtain more representative features.

4) Deep supervision module: As shown in Fig. 1, the green region denotes the deep supervision module $f_{DS}: \mathbb{R}^{W^{(1)}\times H^{(1)}\times C^{(1)}} \times \mathbb{R}^{W^{(2)}\times H^{(2)}\times C^{(2)}} \rightarrow \mathbb{R}^{W^{(1)}\times H^{(1)}\times 2C^{(1)}}$, which is leveraged to extract the boundary information during segmentation. According to the previous research [25], [43], low-level features contain sufficient boundary information. Thus we select feature maps from the first two encoder blocks to provide the fused feature map with fine-grained constraints. In the deep supervision block, the feature maps from the second encoder block are upsampled to the same size as the first encoder block. And they are both fed to the 1×1 and 3×3 convolution layers. After convolution, the concatenated feature maps are applied to guide the boundary segmentation in the decoding path.

The output of the deep supervision block is added to the output of the SE block of the decoder. Then 3×3 deconvolution and 1×1 convolution are employed to generate the prediction map.

B. Hybrid Loss Function

We design a hybrid loss to effectively train our proposed TW-Net, which consists of a First-Order Gradient (FOG) loss [44], a topological loss [45], and a segmentation loss to constrain the geometry and feature balance.

1) First Order Gradient (FOG) loss: Let $s(x) \in \mathbb{R}^{W\times H\times C}$ denote the output probability map given input x and $g \in \mathbb{R}^{W\times H\times C}$ denote ground truth, and C is the number of classes, $v = (v_x, v_y, v_z) \in \Omega$ represents the spatial position vector and $\Omega \subset \mathbb{R}^3$ is the image domain. Then we could apply a 3D geometric loss that constrains the gradients between the predicted map and ground truth to be similar, where the gradient is denoted by $\nabla s(x)_v = [\left(\frac{\partial}{\partial v_x}, \frac{\partial}{\partial v_y}, \frac{\partial}{\partial v_z}\right) s(x)]_v$. Since $s(x)$ corresponds to a single slice, the derivative w.r.t. the z-axis is calculated by find its adjacent slices. The FOG loss is defined as follows:

$$L_{FOG}(x) = \frac{1}{|\Omega|} \sum_{v\in\Omega} \|\nabla s(x)_v - \nabla g_v\|^2 \tag{5}$$

2) Topological Loss: To balance the feature map between encoders and decoders, we employ a topological loss [45] to bring it closer to that desired topology. The loss is defined as follows:

$$\begin{aligned} L_{\text{Topo}}(x) = &\sum_{l=1}^{L_e} (1 - \|f_p(x^{(l)}) - x^{(L_e)}\|^2) + \\ &\sum_{l=L_e+1}^{L_e+L_d} \|f_{up}(x^{(l)}) - x^{(L_e+L_d)}\|^2 \end{aligned} \tag{6}$$

where f_p and f_{up} stand for maxpooling and uppooling respectively (such that the sizes are compatible), L_e / L_d denotes the number of encoder / decoder blocks, and $x^{(l)}$ / $x^{(L_e+l)}$ represents the output of the lth encoder / decoder block, i.e.

$$x^{(l)} = (e^{(l)} \circ \cdots \circ e^{(1)})(x) \tag{7}$$

$$x^{(L_e+l)} = (d^{(l)} \circ \cdots \circ d^{(1)} \circ f_{TM} \circ e)(x) \tag{8}$$

As one feature map will go through 4 encoder blocks and 3 decoder blocks, the weights among all levels should be balanced for more representative features.

3) Total Loss: The main loss term targets at minimizing the distance between the prediction and the corresponding ground truth, which is defined as:

$$L_{\text{Seg}}(x) = -\frac{1}{\Omega} \sum_{v\in\Omega} (g_v \log(s(x)_v) + (1 - g_v) \log(1 - s(x)_v)) \tag{9}$$

Finally, the total loss L is a weighted combination of the three loss terms:

$$L = \mathbb{E}_{x\sim\mathscr{D}}[L_{\text{Seg}}(x) + \lambda_1 L_{\text{FOG}}(x) + \lambda_2 L_{\text{Topo}}(x)] \tag{10}$$

where $\mathscr{D}$ represents the empirical distribution of the dataset. λ_1 and λ_2 are the hyper-parameters. (tuned manually and set to $\lambda_1 = 0.1$ and $\lambda_2 = 0.01$ for the hybrid loss function).

IV. Experiment and Results

A. Datasets

1) dHCP: The anatomical MRI data were obtained from a publicly available cohort, the developing Human Connectome Project (dHCP, http://www.developingconnectome.org/), which was approved by the National Research Ethics Committee and informed written consent given by the parents of all participants. The T2w images were acquired on a 3T Philips Scanner with a dedicated neonatal imaging system with resolution $0.8 \times 0.8 \times 1.6mm^3$ [46]. All MRI data was processed with motion correction and resampled to an isotropic voxel size of $0.5mm^3$. Then the dHCP dataset was rigidly registered to a 40 week atlas space by the MIRTK toolbox (https://mirtk.github.io/). Skull-stripping was performed using Draw-EM pipeline for automatic brain MRI segmentation. Nine labels (GM, WM, CSF, background, ventricle, cerebellum, dGM, brainstem, hippocampus) were used for the segmentation. 41 subjects with more than 9,000 MRI slices (images) were included. To train our models, we take 75% of the dataset for training, and 25% for test.

2) iSeg-2017: The MRI (T1-weighted) data were chosen from the pilot study of Baby Connectome Project (http://babyconnectomeproject.org). For each image, 144 sagittal slices were acquired with parameters: TR/TE = 1900/4.38 ms with flip angle = 7°. The spatial resolution is $1\times1\times1mm^3$. A Siemens head-only 3T scanner with a circular polarized head coil is used for data acquisition. Three labels (WM, GM, CSF) were used for the segmentation and 3D volumes of 10 subjects were trained by 4-fold cross validation for a cross-dataset validation.

B. Implementation Details

Our model is implemented in PyTorch and accelerated by 4 NVIDIA 1080Ti GPU and 4 NVIDIA V-100 GPU. Our code is released in (https://github.com/jerryzhang1119). We describe the implemented details as follows:

1) Network Parameter Setting: All the input MRI slices are cropped to 128×128 and the kernel size is set to 3×3 in normal convolution operations. The batch size for training is set to 4 and the maximum iteration number is 300. In addition, we set the learning rate to 5e-3 without any weight decay. The pooling stride is set to 2 by default. We set the scheduler power

to 0.9 and the stride is set to 2. The vision transformer has four heads with 6 layers. The dimension of the linear layer is 1024. λ_1 is set to 0.1 and λ_2 is set to 0.01 for the hybrid loss function. We search λ_1 from $\{0.1, 0.2, 0.3, 0.4, 0.5\}$ and λ_2 from $\{0.01, 0.05, 0.1\}$. We observe the descent speed of Dice coefficient in the initial 50 epochs, and choose the regularization coefficients.

2) Slice Processing: Our pre-training procedure consists of four sections: (1) Randomly shuffle the image slices and divide them into 5 folds, which helps avoid the potential influence of dataset split bias. (2) Randomly extract 2/3 of all slices with 9 labels. (3) Crop the MRI slice to the fixed size 128×128. (4) Keep all slices of each sample volume to be either in the training set or in the testing set.

C. Evaluation Metrics

There are two segmentation tasks in our research (2D segmentation and 2.5D segmentation). We choose the Dice coefficient (DC), the 95th-percentile of the Hausdorff Distance (HD95) to evaluate the segmentation results. The Dice coefficient measures the overlap between the segmented region and the ground truth region for a class. The performance of segmentation is positively correlated with the Dice coefficient.

The Kth-percentile ($K = 95$) of the Hausdorff Distance (HD95) is another metric that evaluates the distance between segmented region and the corresponding ground truth region. Given two regions, it calculates the distance between the second region and each voxel in the first region, and takes the largest 95th percentile distance. Then, the same calculation is repeated with the two regions switched to get another 95th percentile distance. The maximum between these two distances is taken as the final measure between the two regions. Different from the Dice coefficient, the HD95 metric is negatively correlated with the segmentation performance.

TABLE I: Summary of all trained methods with the corresponding parameters, flops, training time and testing time.

Method	Parameters	Flops	Training time	Testing time
U-Net	34.53M	16375.23M	89s/epoch	14ms/slice
U-Net++	36.63M	34656.76M	124s/epoch	26ms/slice
DenseNet	39.44M	31525.33M	298s/epoch	80ms/slice
SegNet	51.95M	41425.55M	354s/epoch	380ms/slice
Deeplabv3+	41.25M	41663.09M	112s/epoch	40ms/slice
FCN	32.96M	26535.69M	138s/epoch	110ms/slice
PSP-Net	35.31M	18760.24M	187s/epoch	140ms/slice
TransUNet	30.08M	5793.23M	180s/epoch	42ms/slice
TW-Net	30.09M	5809.60M	204s/epoch	45ms/slice

D. Comparison with 2D state-of-the-art approaches

We train other 2D state-of-the-art methods to segment the aforementioned tissues simultaneously by introducing cross-entrophy as the loss function. The algorithms used in our experiment are listed in Table I with the corresponding number of parameters, Flops, training time per epoch and testing time per slice. Parameter of algorithms could influence the segmentation performance as low-resolution images (the size of input slices) may benefit from the shadow segmentation networks. However, the number of labels in our segmentation task exceeds traditional segmentation tasks and some labels only occur in a few slices. Thus a deeper network with more parameters is suitable for complex tasks (*e.g.* multiple and imbalanced labels). It is difficult to find the optimal parameter of the specific algorithm. To assure a fair comparison, the close parameter of all methods without changing original backbones too much is suggested. Due to the fact that cerebellum, brainstem and hippocampus take a small proportion of labels, some algorithms fail to segment them. We replace the result with '−'. Table II reports the results with 95% confidence intervals of nine different methods with respect to two evaluation metrics. We could observe that TW-Net achieves better segmentation in all brain tissues except deep gray matter, indicating that the proposed architecture in general outperforms other models in this 2D segmentation task. We select three slices uniformly from top to bottom in the axial plane to show the advantages of our model. The visualization of the segmentation results and the corresponding error maps are shown in Fig. 4, in which over-segmentation and under-segmentation are highlighted in red. We select the #55, #75, and #95 slices for error maps visualization. We observe that the proposed method is capable of accurately segmenting small brain regions such as hippocampus and brainstem. In addition, the error of our method is the lowest among all the nine approaches. As shown in Fig. S2, our method outperforms competing methods in segmentation accuracy for most brain structures in 2D experiments. However, TW-Net fails to outperform some methods (*e.g.* U-Net) when segmenting dGM. The focus of TW-Net has been allocated to each label uniformly, such as cerebellum, and brainstem, while U-Net may allocate more focus to dGM, ignoring the segmentation performance of cerebellum, brainstem, and hippocampus. This may account for the inferior performance of TW-Net when segmenting dGM.

1) Ablation Study for SE and Deep supervision module: We perform ablation studies to validate our deep supervision and squeeze-and-excitation (SE) module. In this section, we remove the deep supervision module and the SE module and compare the performances.

To explore the contribution of the deep supervision module, we derive two baselines: backbone only (TW-Net w/o SEDS) and backbone without the deep supervision module incorporated (TW-Net w/o DS : our method without the deep supervision module but with SE blocks). The results in Table II clearly show that deep supervision block is necessary for boosting the segmentation performance of deep gray matter, white matter, CSF, and hippocampus. The Dice score increases 9.6%, 3.8%, 2.6%, 9.9%, respectively. Moreover, as shown in Fig. S3, we present feature maps of four slices from one subject to illustrate the effectiveness of DS module, which benefits the extraction of boundary information.

We also investigate the importance of SE block. From Table II, we observe that backbone with SE block increases the backbone performance in terms of Dice and HD95 in all tissues except ventricle. Specifically, the Dice score increases 3.0% and 9.1% for white matter and deep gray matter, respectively. Moreover, the cerebellum and hippocampus cannot be segmented from backbone while SE block could address

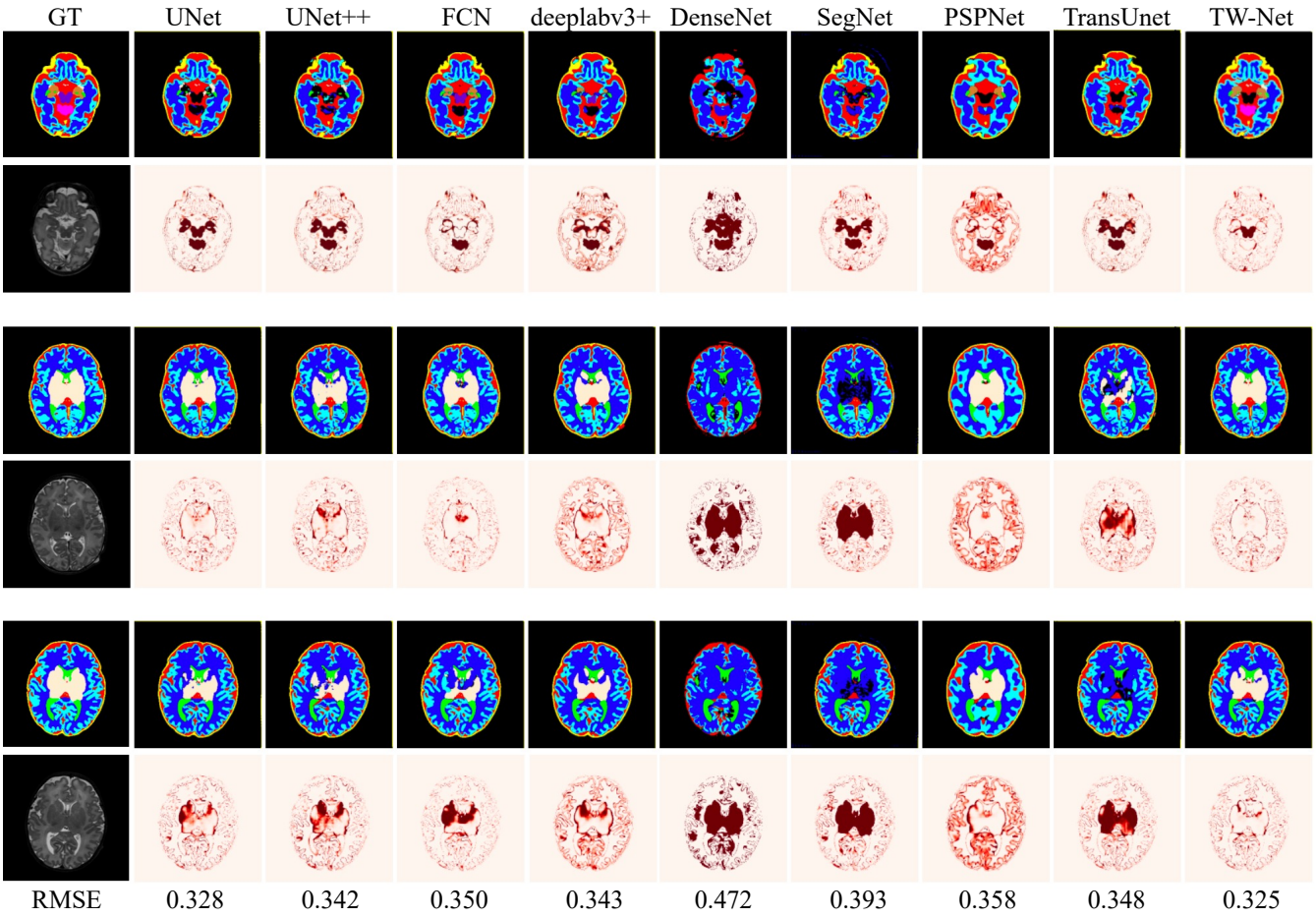

Fig. 4: Comparisons between the automated segmentation of the state-of-the-art approaches and our proposed TW-Net in 2D configuration. Odd rows and even rows represent segmentation results and the corresponding error maps, respectively. We replace the error maps of the ground truth with the original slice.

this problem. This suggests that introducing the SE block can enable our model to accurately segment the minor brain tissues and boost the segmentation performance comprehensively. By this ablation study, we have demonstrated the significance of the combination of Deep Supervision and SE block.

As shown in Table II, our TW-Net performs better than any other settings in terms of Dice score and HD95. These improvements demonstrate that Deep Supervision block together with SE block are the two central components responsible for the good performance of our proposed TW-Net.

2) Sensitivity Analysis on Transformer Layers: To identify the influence of Transformer layers, we conduct experiments (training networks from scratch) when given 4, 6, 8, 10 Transformer layers for the TW-Net, respectively. This involves evaluating every experiment on the validation set with the Dice score and HD95. As shown in Fig. 5 and Fig. S4, we take U-Net as baseline and observe that the Dice score and HD95 are over the baseline steadily, which shows the reliability of our method. The model with 4 layers is more efficient and the model with 6 layers is almost the worst. To prove the robustness of our architecture, we use 6 in our experiments. Fig. 5 highlights that the Transformer block greatly improves the performance compared to the baseline. Except for deep gray matter, our architecture performs the best in all other tissues, suggesting the significance of incorporating long range dependency information in the Transformer.

3) Ablation Study for RF-Layer: The RF-layer conducts linear transformation of feature maps. To investigate the effect of the RF-layer, we remove the RF-layer, and refer to it as 0 conv-layer. In addition, we replace the RF-layer with several scale-invariant convolutional layers (1×1 convolutional layer without changing the size of feature map). We set the number of convolutional layers from 1 to 5. The results are shown in Fig. S5. It is observed that, with the number of convolutional layers increasing, the general performance has been improved until the number reaches 5, where the Dice score declines significantly, which indicates the occurrence of over-fitting. The TW-Net with the RF-layer and the TW-Net with four convolutional layers achieve similar performance and are noticeably better than the others.

4) Ablation Study for Transformer block: Transformer block conducts token-similarity calculation to enhance the representation capability. To investigate the effectiveness of the whole block, we replace the Transformer block with several 1×1 convolutional layer. The number of convolutional layers is set from 0 to 3. The results are presented in Fig. S6.

TABLE II: Segmentation results (mean ± standard) by 5-fold cross-validation achieved using nine 2D methods, in terms of Dice coefficient and HD95. The best results are marked in red. Some methods fail to predict the segmentation, which is marked by '-'. The mean performance is shown in the final column where $*$ represents the mean performance of segmented brain regions instead of all brain regions. The results indicate that our proposed TW-Net is better than all other compared methods.

Methods	Metric	GM	WM	CSF	Background	Ventricle	Cerebellum	dGM	Brainstem	Hippocampus	mean
U-Net	Dice	0.924±0.013	0.909±0.038	0.940±0.023	0.777±0.021	0.890±0.028	-	0.715±0.104	-	-	0.859*
U-Net++		0.919±0.014	0.888±0.041	0.923±0.024	0.819±0.022	0.875±0.030	-	0.656±0.113	-	0.740±0.113	0.831*
Deeplabv3+		0.916±0.011	0.868±0.060	0.909±0.037	0.828±0.023	0.900±0.026	-	0.662±0.115	0.790±0.110	0.786±0.142	0.832*
SegNet		0.836±0.021	0.589±0.137	0.556±0.148	0.813±0.021	0.751±0.104	0.846±0.141	0.190±0.102	0.670±0.139	0.741±0.140	0.666*
FCN		0.937±0.013	0.912±0.053	0.927±0.033	0.775±0.019	0.922±0.017	-	0.773±0.105	0.592±0.064	0.765±0.112	0.825*
PSP-Net		0.878±0.013	0.837±0.031	0.900±0.018	0.820±0.019	0.888±0.024	0.846±0.130	0.756±0.103	-	0.757±0.084	0.836*
Dense-Net		0.807±0.013	0.716±0.023	0.824±0.007	0.595±0.153	0.539±0.096	0.846±0.130	0.273±0.079	0.670±0.220	0.740±0.192	0.668*
TransUNet		0.935±0.012	0.904±0.033	0.932±0.020	0.798±0.022	0.922±0.017	-	0.566±0.113	0.680±0.122	0.773±0.115	0.814*
TW-Net w/o SE		0.938±0.012	0.912±0.013	0.940±0.014	0.796±0.032	0.908±0.015	-	0.678±0.116	0.670±0.220	-	0.835
TW-Net w/o DS		0.939±0.012	0.910±0.015	0.935±0.014	0.794±0.043	0.912±0.014	0.831±0.111	0.673±0.118	0.715±0.120	0.788±0.130	0.833
TW-Net		0.940±0.012	0.917±0.011	0.942±0.014	0.862±0.015	0.930±0.013	0.856±0.082	0.679±0.115	0.798±0.107	0.819±0.105	0.860
U-Net	HD95	3.602±0.425	4.170±0.312	4.603±0.530	1.479±0.026	1.895±0.952	-	2.737±0.633	-	-	3.697*
U-Net++		3.608±0.412	4.364±0.327	4.815±0.508	1.477±0.042	1.944±0.853	-	3.077±1.304	-	1.196±0.535	2.925*
Deeplabv3+		3.758±0.438	4.662±0.492	5.158±1.034	1.456±0.571	1.840±0.848	-	2.996±1.945	1.273±0.311	1.066±0.622	2.776*
SegNet		3.898±0.425	4.662±0.294	6.970±1.301	2.003±0.057	2.073±0.704	3.620±0.132	4.996±1.835	1.630±0.691	1.466±0.506	3.478*
FCN		3.375±0.474	4.225±0.562	4.600±0.865	1.462±0.057	1.850±0.903	-	2.639±0.435	1.880±0.299	0.998±0.295	2.629*
PSP-Net		4.286±0.337	4.978±0.212	4.972±0.470	1.587±0.391	1.909±0.942	3.027±0.686	2.695±0.743	-	2.480±0.875	3.242*
Dense-Net		5.691±0.808	6.835±1.028	7.047±1.023	1.489±0.038	4.598±1.602	3.027±0.686	5.333±1.559	1.625±0.670	1.196±0.515	4.093*
TransUNet		3.371±0.428	4.407±0.289	4.769±0.524	1.472±0.031	1.807±1.036	-	3.601±0.710	1.620±0.625	1.093±0.520	2.768*
TW-Net w/o SE		3.375±0.512	4.332±0.295	4.710±0.526	1.463±0.044	1.820±0.360	-	3.357±1.080	1.445±0.623	-	2.929
TW-Net w/o DS		3.372±0.422	4.347±0.303	4.712±0.526	1.477±0.035	1.804±0.351	3.552±0.710	3.361±1.104	1.414±0.522	1.056±0.509	2.788
TW-Net		3.369±0.312	4.135±0.196	4.550±0.422	1.453±0.034	1.750±0.322	2.836±0.646	3.116±1.108	1.183±0.116	0.972±0.313	2.596

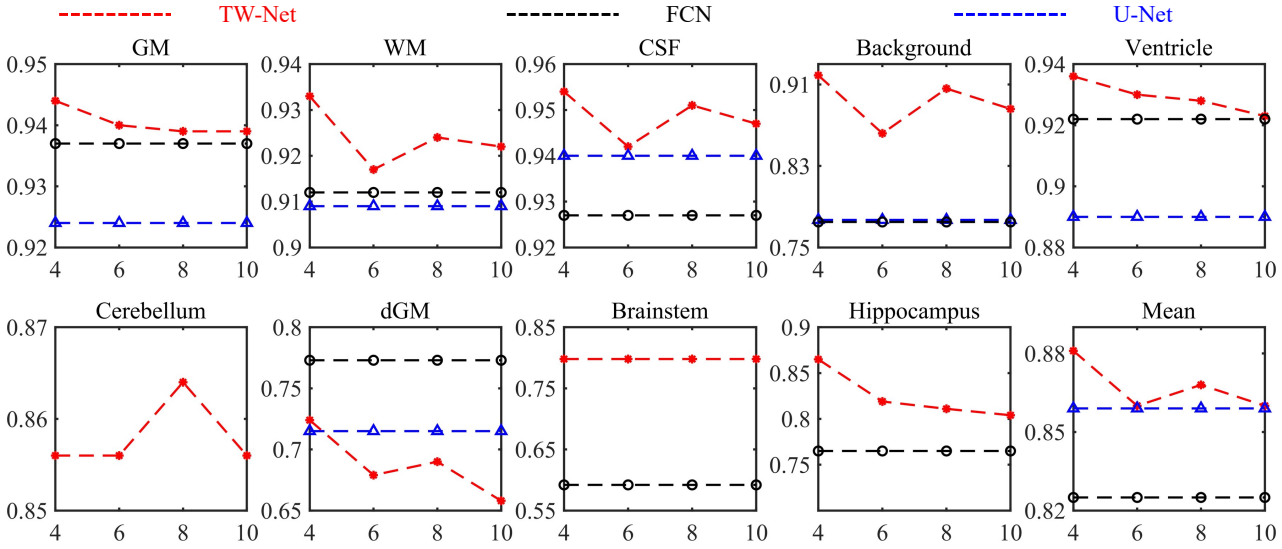

Fig. 5: Sensitivity analysis on Transformer layers using Dice illustrates that Transformer block enhances the network performance, which is independent of the number of Transformer layers. The x-axis represent the number of Transformer layers. We compare our methods with two baselines (FCN and U-Net). The labels not segmented are not shown in the figure. (FCN fails to segment cerebellum, UNet fails to segment cerebellum, brainstem and hippocampus).

We could observe that, with the number of convolutional layers increasing, the segmentation performance declines significantly. The conclusion could be drawn that the Transformer block takes the important position that couldn't be replaced by traditional convolutional layers. The worse performance of convolutional layers indicates that deeper convolutional layers aren't applicable in complex tasks, which requires improvement of the representation of the model.

E. Segmentation Results in 2.5D Configuration

As two-dimensional slices lack spatial information along the z-axis, we incorporate this information into TW-Net by concatenating three adjacent image slices and then using a 3D convolutional layer followed by several 2D layers. We implemented this strategy for all the methods and reran the experiment. The results are shown in Table III.

From Table II and Table III, we observe that with the spatial information from adjacent slices, most of the 2.5D approaches boost the segmentation performance significantly when compared with the corresponding 2D approaches, which demonstrates the significance of using this information.

As shown in Fig. 6 and Table III, comparing with state-of-the-art approaches, our proposed method consistently outperforms other methods in 2.5D segmentation for all brain structures except for the region of dGM in terms of error maps, Dice coefficient, and HD95. For instance, for white matter and background segmentation, our method achieves an improvement of 4.2% and 4.1% respectively (in terms of Dice

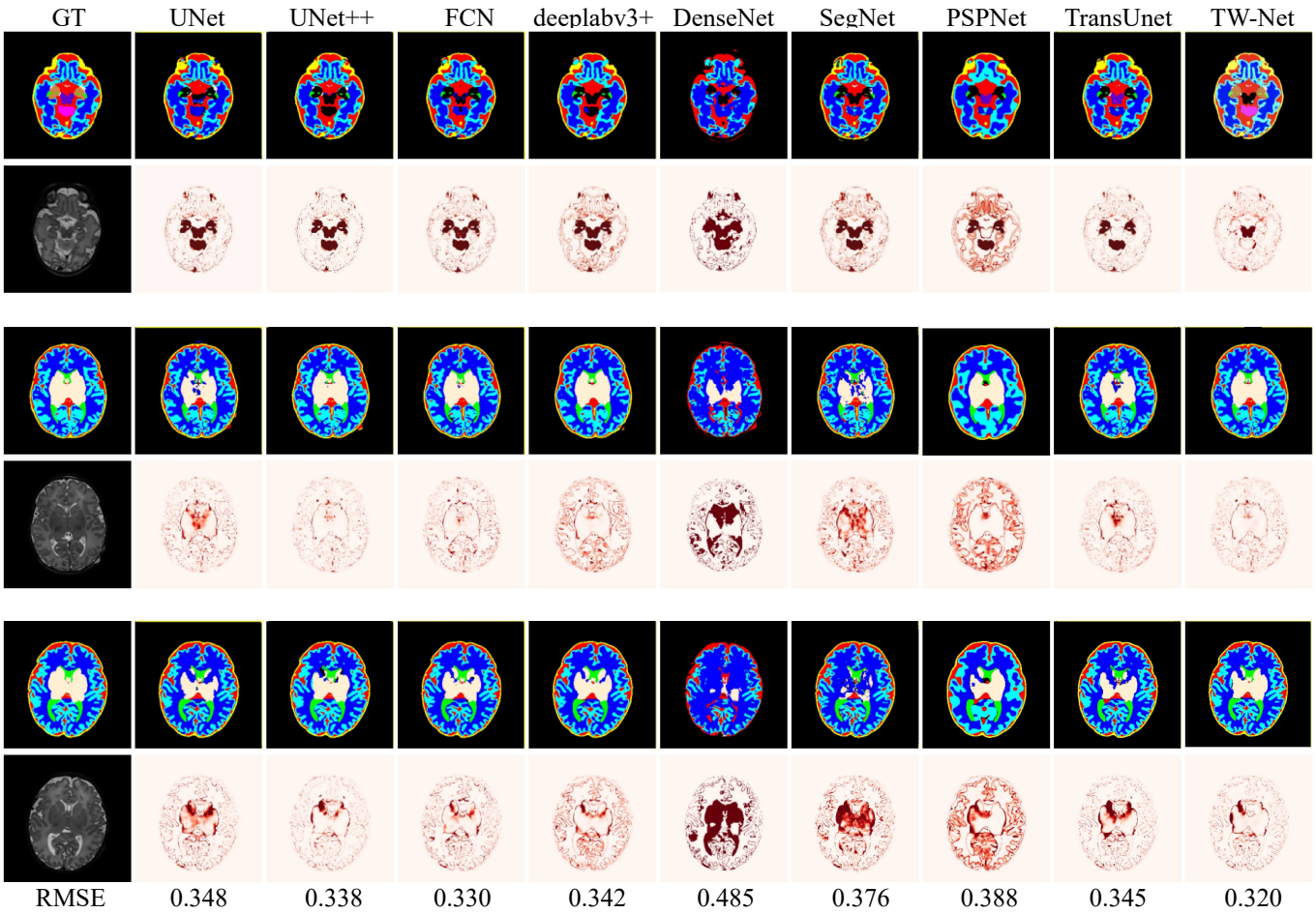

Fig. 6: Comparisons between the automated segmentation of the state-of-the-art approaches and our proposed TW-Net in 2.5D configuration. Odd rows and even rows represent segmentation results and the corresponding error maps, respectively. We replace the error maps of ground truth with the original slice.

coefficient) compared with the second-best result achieved by U-Net++. In addition, our method shows better segmentation accuracy for most brain structures in 2D experiments (Fig. 4). These results further prove the effectiveness of the TW-Net for neonatal brain segmentation.

F. Segmentation Results of our methods compared with 3D baselines

To investigate the effectiveness of our proposed TW-Net, we compare our 2D and 2.5D method with 3D segmentation algorithms 3D U-Net, 3D U-Net++, and TransBTS. As shown in Fig. 7 , our methods in 2D and 2.5D configurations both outperformed 3D U-Net and 3D U-Net++ steadily in terms of the Dice coefficient, even though our methods do not have the full slice information as the 3D algorithms. We could observe that TransBTS share a similar performance with 2D TW-Net and could segment all the labels as well. We could draw the conclusion that convolutional neural networks implemented with a Transformer block could outperform other convolutional networks in 2D and 3D configurations.

G. Segmentation results on iSeg-2017 dataset

To further investigate the generalizability of the proposed TW-Net, we compare our models on the iSeg-2017 dataset. Experimental results on iSeg-2017 are summarized in Table IV. We could observe that TW-Net yields the best segmentation performance in terms of Dice coefficient and HD95 when compared with the other competing methods in 2D and 2.5D configurations. For example, TW-Net significantly improves the previous state-of-the-art method (TransUNet 2D) by 2.3% and 0.54 in terms of Dice and HD95, respectively. Also, TW-Net dramatically improves the baseline (U-Net) from 77.82% to 86.45% in a 2D configuration. Such improvements demonstrate the effectiveness of our model in learning the intrinsic features independent of data sources, as well as accurately identifying the brain regions in low-contrast MRI slices. In addition, it is observed that the TW-Net in a 2.5D configuration outperforms the TW-Net in a 2D configuration in three classes.

H. Ablation Study for Hybrid Loss Function

To explore the effectiveness of the proposed hybrid loss function, we conduct ablation study on both 2D and 2.5D

TABLE III: Segmentation results (mean ± standard) by 5-fold cross-validation achieved using nine 2.5D methods, in terms of Dice coefficient and HD95. The best results are marked in red. Some methods fail to predict the segmentation, which is marked by '-'. The mean performance is shown in the final column where $*$ represents the mean performance of segmented brain regions instead of all brain regions. The results indicate that our proposed TW-Net is better than all other methods.

Method	Metric	GM	WM	CSF	Background	Ventricle	Cerebellum	dGM	Brainstem	Hippocampus	mean
U-Net	Dice	0.868±0.021	0.896±0.028	0.858±0.097	0.695±0.125	0.924±0.013	-	0.707±0.115	-	-	0.779*
U-Net++		0.939±0.012	0.905±0.031	0.904±0.024	0.817±0.032	0.908±0.032	-	0.786±0.108	-	0.741±0.103	0.857*
Deeplabv3+		0.927±0.007	0.906±0.034	0.936±0.026	0.847±0.016	0.926±0.012	0.846±0.135	0.775±0.093	0.671±0.022	-	0.854*
SegNet		0.922±0.019	0.892±0.091	0.918±0.044	0.878±0.054	0.731±0.126	0.846±0.130	0.514±0.143	0.670±0.220	-	0.796*
FCN		0.934±0.014	0.924±0.054	0.945±0.041	0.789±0.015	0.915±0.021	0.845±0.131	0.768±0.108	0.666±0.108	-	0.848*
PSP-Net		0.874±0.014	0.845±0.025	0.901±0.022	0.886±0.030	0.881±0.030	0.846±0.130	0.734±0.108	0.656±0.092	-	0.828*
Dense-Net		0.809±0.057	0.740±0.265	0.832±0.078	0.594±0.014	0.156±0.013	0.846±0.130	0.410±0.141	0.670±0.221	0.741±0.192	0.644*
TransUNet		0.873±0.016	0.896±0.025	0.884±0.041	0.800±0.032	0.918±0.017	-	0.786±0.111	-	0.741±0.125	0.843*
TW-Net w/o SE		0.938±0.010	0.918±0.010	0.944±0.012	0.815±0.027	0.906±0.015	-	0.678±0.115	0.670±0.218	-	0.838
TW-Net w/o DS		0.939±0.012	0.910±0.015	0.935±0.014	0.794±0.043	0.912±0.014	0.831±0.111	0.673±0.118	0.665±0.120	0.788±0.130	0.827
TW-Net		0.956±0.012	0.948±0.023	0.961±0.014	0.888±0.015	0.945±0.014	0.846±0.108	0.769±0.015	0.675±0.019	0.887±0.081	0.875
U-Net	HD95	3.740±0.328	4.260±0.450	4.836±1.183	2.335±0.055	1.875±0.528	-	2.851±1.403	-	-	3.316*
U-Net++		3.523±0.319	4.307±0.412	4.936±0.641	1.364±0.032	1.892±0.408	-	2.902±1.020	-	1.191±0.310	2.874*
Deeplabv3+		3.577±0.331	4.299±0.346	4.527±0.868	1.716±0.057	1.715±0.767	3.627±0.127	2.679±0.520	1.625±0.567	-	2.971*
SegNet		3.664±0.552	4.449±0.573	5.067±1.226	1.605±0.026	1.974±0.961	3.270±0.686	3.146±0.617	1.645±0.537	-	3.103*
FCN		3.521±0.592	4.394±0.559	4.266±1.063	1.790±0.057	1.887±1.015	3.688±0.687	2.637±2.552	1.502±0.781	-	2.961*
PSP-Net		4.321±0.317	4.958±0.214	4.948±0.514	2.113±0.375	1.933±1.012	3.276±0.686	2.778±1.964	2.406±0.976	-	3.342*
Dense-Net		5.444±0.892	6.380±0.871	6.887±0.945	2.886±0.038	4.324±0.671	3.027±0.686	4.293±2.529	1.645±0.669	1.195±0.515	4.009*
TransUNet		3.420±0.337	4.431±0.472	4.877±0.638	1.588±0.026	1.896±1.035	-	3.112±1.390	-	1.090±0.204	2.916*
TW-Net w/o SE		3.371±0.500	4.330±0.283	4.706±0.513	1.460±0.042	1.815±0.350	-	3.352±1.004	1.423±0.620	-	2.922
TW-Net w/o DS		3.373±0.420	4.341±0.298	4.575±0.418	1.475±0.035	1.801±0.342	3.550±0.710	3.358±1.105	1.407±0.512	1.058±0.488	2.778
TW-Net		3.320±0.312	4.200±0.384	3.710±0.506	1.353±0.025	1.623±0.304	3.022±0.556	3.346±1.240	1.257±0.209	1.050±0.201	2.542

TABLE IV: Segmentation results (mean ± standard) by 4-fold cross-validation achieved using nine 2D and 2.5D methods, in terms of Dice coefficient and HD95. The best results are marked in red. The results indicate that our proposed TW-Net is better than all other methods.

Method	Metric	2D configuration			2.5D configuration		
		WM	GM	CSF	WM	GM	CSF
U-Net	Dice	77.82±1.56	80.49±1.27	83.57±0.89	78.18±1.52	82.33±1.38	84.50±0.88
U-Net++		78.04±1.35	81.25±1.25	84.30±1.03	79.13±1.33	81.85±1.18	84.87±1.02
Deeplabv3+		77.65±1.51	80.53±1.20	83.44±0.97	77.88±1.43	81.34±1.16	84.45±0.86
SegNet		73.19±2.13	76.44±1.98	80.36±1.25	73.15±2.15	76.46±1.93	81.32±1.24
FCN		80.42±1.53	83.92±1.24	87.78±0.83	82.85±1.50	83.95±1.25	88.13±0.97
PSP-Net		80.04±1.42	82.73±1.05	86.63±0.82	81.15±1.40	82.90±1.05	86.92±0.80
Dense-Net		72.21±1.86	75.28±2.29	78.38±0.95	70.15±1.56	75.48±2.32	76.73±0.95
TransUNet		84.14±1.58	86.05±1.18	89.46±0.73	85.06±1.38	86.95±1.28	90.05±0.82
TW-Net		86.45±1.38	88.23±0.89	91.32±0.65	87.53±1.42	89.35±0.79	91.82±0.54
U-Net	HD95	8.52±1.48	7.63±1.40	10.35±1.26	8.44±1.51	7.57±1.64	10.22±1.18
U-Net++		8.40±1.39	7.60±1.33	10.05±1.08	8.29±1.35	7.52±1.30	10.02±1.06
Deeplabv3+		8.55±1.36	7.58±1.25	9.95±1.12	8.52±1.38	7.46±1.28	9.92±1.26
SegNet		9.21±1.28	9.63±1.06	10.25±1.14	9.23±1.24	9.60±1.08	10.14±1.28
FCN		7.48±1.37	7.20±1.13	9.75±1.16	6.82±1.27	7.22±1.14	9.46±1.23
PSP-Net		7.61±1.25	7.35±1.14	9.68±0.92	7.48±1.24	7.30±1.12	9.65±0.95
Dense-Net		10.30±1.35	9.52±1.04	10.52±1.07	10.52±1.54	9.40±1.12	11.28±1.27
TransUNet		6.99±1.12	6.86±1.22	9.57±1.32	6.87±1.06	6.77±1.26	9.36±1.24
TW-Net		6.45±1.27	6.78±1.07	9.35±1.28	6.40±1.24	6.65±1.12	9.32±1.27

networks. As shown in Fig. 8, the FOG loss function and the topology loss function improve the segmentation performance in all brain tissues, and the combination of the two loss functions achieve the best performance. We could also observe that the FOG loss function outperforms the topology loss function when segmenting gray matter, background, cerebellum and deep gray matter on both 2D and 2.5D tasks. And it underperforms the topology loss function when segmenting brainstem and hippocampus. Since brainstem and hippocampus contain few labeled voxels, the topology loss function may be beneficial to few-shot learning in medical image segmentation.

V. Conclusion

We have developed a Transformer-weighted network combining deep supervision and SE-block for neonatal brain tissue segmentation. To make the network more suitable for neonatal brain structure segmentation, we implement the network with a RF-layer. An ablation study demonstrates the utility of the deep supervision module, Transformer layer, SE-block, and RF-layer. Then, we apply a 2.5D strategy that takes advantage of spatial information in adjacent MRI slices. As a result, our algorithm outperforms other models in 2D and 2.5D segmentation tasks when compared with other state-of-the-art segmentation algorithms. Moreover, when compared with several 3D baselines, TW-Net still leads the segmentation performance in terms of Dice coefficient. In addition, the segmentation performance of TW-Net on another dataset (iSeg-2017) present the advantages in neonatal brain region segmentation. Our automatic segmentation model could provide a effective way for segmenting multiple brain tissues simultaneously, which benefits the subsequent tasks such as neurodegenerative disease diagnosis and evaluation. Although the proposed architecture has achieved promising performance on two segmentation tasks, some limitations and future work should still be considered: (1) A 2D network is utilized

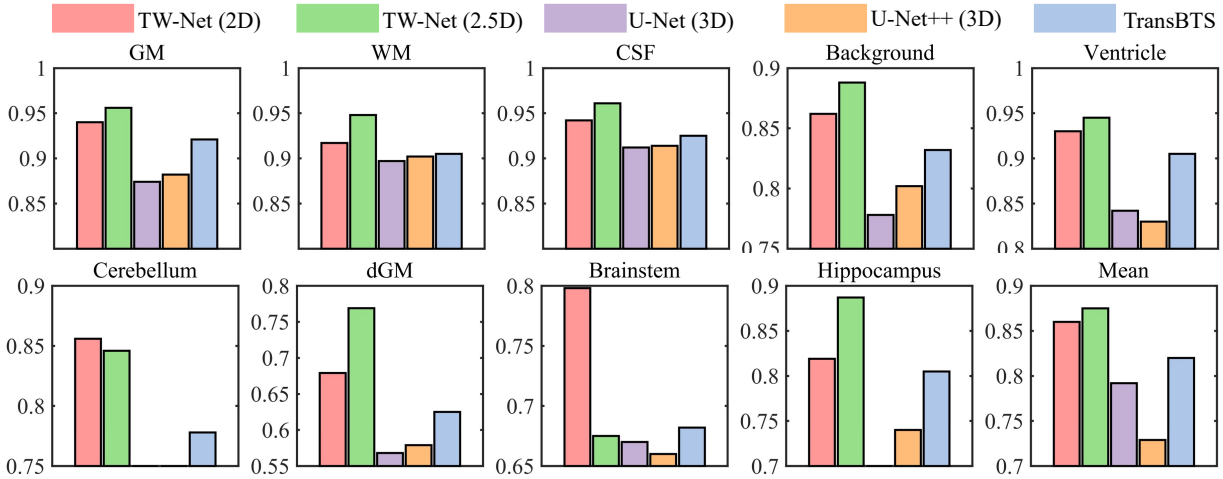

Fig. 7: Comparison with 3D baselines. We compare TW-Net in 2D and 2.5D configurations with several 3D baselines, including 3D U-Net, 3D U-Net++, and TransBTS. Some methods fail to predict the segmentation, which is not shown in this figure.

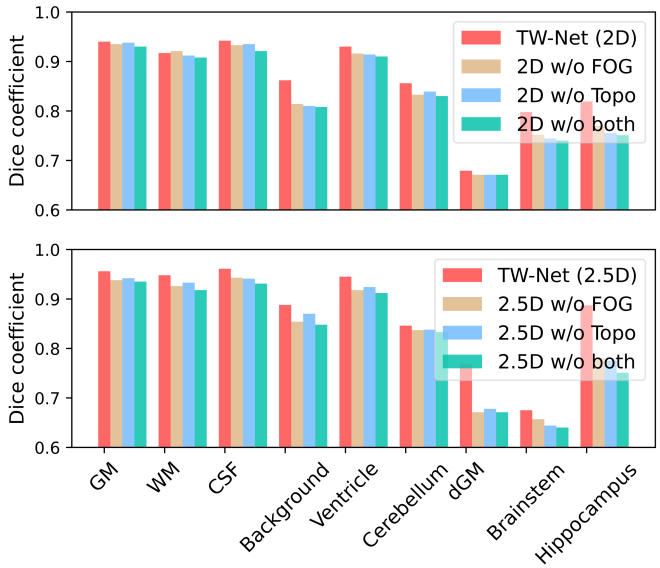

Fig. 8: Ablation study of the proposed hybrid loss function in 2D and 2.5D configurations. The first and second rows represent the TW-Net in 2D and 2.5D configurations, respectively. 'w/o FOG' means that the FOG loss function is not utilized during the training process. Also, 'w/o both' means that there is only the segmentation loss (cross entropy) when training the TW-Net.

in our experiment, while 2D networks lose the 3D spatial information. (2) The sample size used is not sufficient enough for multiple brain structures' segmentation and more samples could improve our performance. In the future, we will explore the 3D architecture with more samples for neonatal brain segmentation.

REFERENCES

[1] G. Li, J. Nie, L. Wang, F. Shi, W. Lin, J. H. Gilmore, and D. Shen, "Mapping region-specific longitudinal cortical surface expansion from birth to 2 years of age," *Cerebral cortex*, vol. 23, no. 11, pp. 2724–2733, 2013.

[2] G. Li, J. Nie, L. Wang, F. Shi, A. E. Lyall, W. Lin, J. H. Gilmore, and D. Shen, "Mapping longitudinal hemispheric structural asymmetries of the human cerebral cortex from birth to 2 years of age," *Cerebral cortex*, vol. 24, no. 5, pp. 1289–1300, 2014.

[3] L. Wang, F. Shi, W. Lin, J. H. Gilmore, and D. Shen, "Automatic segmentation of neonatal images using convex optimization and coupled level sets," *NeuroImage*, vol. 58, no. 3, pp. 805–817, 2011.

[4] G. Li, W. Lin, J. H. Gilmore, and D. Shen, "Spatial patterns, longitudinal development, and hemispheric asymmetries of cortical thickness in infants from birth to 2 years of age," *Journal of neuroscience*, vol. 35, no. 24, pp. 9150–9162, 2015.

[5] F. Shi, P.-T. Yap, W. Gao, W. Lin, J. H. Gilmore, and D. Shen, "Altered structural connectivity in neonates at genetic risk for schizophrenia: a combined study using morphological and white matter networks," *Neuroimage*, vol. 62, no. 3, pp. 1622–1633, 2012.

[6] E. T. Ahrens and J. W. Bulte, "Tracking immune cells in vivo using magnetic resonance imaging," *Nature Reviews Immunology*, vol. 13, no. 10, pp. 755–763, 2013.

[7] K. M. Brown, G. Barrionuevo, A. J. Canty, V. De Paola, J. A. Hirsch, G. S. Jefferis, J. Lu, M. Snippe, I. Sugihara, and G. A. Ascoli, "The diadem data sets: representative light microscopy images of neuronal morphology to advance automation of digital reconstructions," *Neuroinformatics*, vol. 9, no. 2, pp. 143–157, 2011.

[8] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

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

[10] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE transactions on medical imaging*, vol. 39, no. 6, pp. 1856–1867, 2019.

[11] S. Li, J. Zhang, C. Ruan, and Y. Zhang, "Multi-stage attention-unet for wireless capsule endoscopy image bleeding area segmentation," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2019, pp. 818–825.

[12] Z. Gu, J. Cheng, H. Fu, K. Zhou, H. Hao, Y. Zhao, T. Zhang, S. Gao, and J. Liu, "Ce-net: Context encoder network for 2d medical image segmentation," *IEEE transactions on medical imaging*, vol. 38, no. 10, pp. 2281–2292, 2019.

[13] A. Vakanski, M. Xian, and P. E. Freer, "Attention-enriched deep learning model for breast tumor segmentation in ultrasound images," *Ultrasound in medicine & biology*, vol. 46, no. 10, pp. 2819–2833, 2020.

[14] M. Byra, P. Jarosik, A. Szubert, M. Galperin, H. Ojeda-Fournier, L. Olson, M. O'Boyle, C. Comstock, and M. Andre, "Breast mass segmentation in ultrasound with selective kernel u-net convolutional

neural network," *Biomedical signal processing and control*, vol. 61, p. 102027, 2020.

[15] Z. Ning, K. Wang, S. Zhong, Q. Feng, and Y. Zhang, "Cf2-net: Coarse-to-fine fusion convolutional network for breast ultrasound image segmentation," *arXiv preprint arXiv:2003.10144*, 2020.

[16] A. Aquino, M. E. Gegúndez-Arias, and D. Marín, "Detecting the optic disc boundary in digital fundus images using morphological, edge detection, and feature extraction techniques," *IEEE transactions on medical imaging*, vol. 29, no. 11, pp. 1860–1869, 2010.

[17] A. Q. Ye, O. A. Ajilore, G. Conte, J. GadElkarim, G. Thomas-Ramos, L. Zhan, S. Yang, A. Kumar, R. L. Magin, A. G Forbes *et al.*, "The intrinsic geometry of the human brain connectome," *Brain informatics*, vol. 2, no. 4, pp. 197–210, 2015.

[18] H. Zhang, A. M. Valcarcel, R. Bakshi, R. Chu, F. Bagnato, R. T. Shinohara, K. Hett, and I. Oguz, "Multiple sclerosis lesion segmentation with tiramisu and 2.5 d stacked slices," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 338–346.

[19] Y. Gonzalez, C. Shen, H. Jung, D. Nguyen, S. B. Jiang, K. Albuquerque, and X. Jia, "Semi-automatic sigmoid colon segmentation in ct for radiation therapy treatment planning via an iterative 2.5-d deep learning approach," *Medical image analysis*, vol. 68, p. 101896, 2021.

[20] F. Iandola, M. Moskewicz, S. Karayev, R. Girshick, T. Darrell, and K. Keutzer, "Densenet: Implementing efficient convnet descriptor pyramids," *arXiv preprint arXiv:1404.1869*, 2014.

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

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

[23] H. Zheng, F. Lin, X. Feng, and Y. Chen, "A hybrid deep learning model with attention-based conv-lstm networks for short-term traffic flow prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 11, pp. 6910–6920, 2020.

[24] A. Roy Choudhury, R. Vanguri, S. R. Jambawalikar, and P. Kumar, "Segmentation of brain tumors using deeplabv3+," in *International MICCAI Brainlesion Workshop*. Springer, 2018, pp. 154–167.

[25] Z. Zhang, H. Fu, H. Dai, J. Shen, Y. Pang, and L. Shao, "Et-net: A generic edge-attention guidance network for medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 442–450.

[26] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[27] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[28] W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, and J. Li, "Transbts: Multimodal brain tumor segmentation using transformer," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 109–119.

[29] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," *arXiv preprint arXiv:2105.05537*, 2021.

[30] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6881–6890.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[32] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[33] C. Li, M. Z. Zia, Q.-H. Tran, X. Yu, G. D. Hager, and M. Chandraker, "Deep supervision with intermediate concepts," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1828–1843, 2018.

[34] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[35] F. Wu, P. Hu, and D. Kong, "Flip-rotate-pooling convolution and split dropout on convolution neural networks for image classification," *arXiv preprint arXiv:1507.08754*, 2015.

[36] Q. Yu, L. Xie, Y. Wang, Y. Zhou, E. K. Fishman, and A. L. Yuille, "Recurrent saliency transformation network: Incorporating multi-stage visual cues for small organ segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8280–8289.

[37] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[38] H. Shan, Y. Zhang, Q. Yang, U. Kruger, M. K. Kalra, L. Sun, W. Cong, and G. Wang, "3-d convolutional encoder-decoder network for low-dose ct via transfer learning from a 2-d trained network," *IEEE transactions on medical imaging*, vol. 37, no. 6, pp. 1522–1534, 2018.

[39] L. Sifre and S. Mallat, "Rotation, scaling and deformation invariant scattering for texture discrimination," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 1233–1240.

[40] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, and D. Tran, "Image transformer," in *International Conference on Machine Learning*. PMLR, 2018, pp. 4055–4064.

[41] J. Liu, S. Chen, B. Wang, J. Zhang, N. Li, and T. Xu, "Attention as relation: learning supervised multi-head self-attention for relation extraction," in *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, 2021, pp. 3787–3793.

[42] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[43] R. Gu, G. Wang, T. Song, R. Huang, M. Aertsen, J. Deprest, S. Ourselin, T. Vercauteren, and S. Zhang, "Ca-net: Comprehensive attention convolutional neural networks for explainable medical image segmentation," *IEEE transactions on medical imaging*, vol. 40, no. 2, pp. 699–711, 2020.

[44] H. Zhang, J. Zhang, R. Wang, Q. Zhang, S. A. Gauthier, P. Spincemaille, T. D. Nguyen, and Y. Wang, "Geometric loss for deep multiple sclerosis lesion segmentation," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2021, pp. 24–28.

[45] J. R. Clough, N. Byrne, I. Oksuz, V. A. Zimmer, J. A. Schnabel, and A. P. King, "A topological loss function for deep-learning based image segmentation using persistent homology," *arXiv preprint arXiv:1910.01877*, 2019.

[46] A. Makropoulos, E. C. Robinson, A. Schuh, R. Wright, S. Fitzgibbon, J. Bozek, S. J. Counsell, J. Steinweg, K. Vecchiato, J. Passerat-Palmbach *et al.*, "The developing human connectome project: A minimal processing pipeline for neonatal cortical surface reconstruction," *Neuroimage*, vol. 173, pp. 88–112, 2018.