ExSwin-Unet: An Unbalanced Weighted Unet with Shifted Window and External Attentions for Fetal Brain MRI Image Segmentation

Yufei Wen[1(✉)], Chongxin Liang[2], Jingyin Lin[2], Huisi Wu[2], and Jing Qin[3]

[1] South China University of Technology, Guangzhou, China
201930034695@mail.scut.edu.cn
[2] Shenzhen University, Shenzhen, China
{2060271074,2110276229}@email.szu.edu.cn, hswu@szu.edu.cn
[3] The Hong Kong Polytechnic University, Hung Hom, Hong Kong
harry.qin@polyu.edu.hk

Abstract. Accurate fetal brain MRI image segmentation is essential for fetal disease diagnosis and treatment. While manual segmentation is laborious, time-consuming, and error-prone, automated segmentation is a challenging task owing to (1) the variations in shape and size of brain structures among patients, (2) the subtle changes caused by congenital diseases, and (3) the complicated anatomy of brain. It is critical to effectively capture the long-range dependencies and correlations among training samples to yield satisfactory results. Recently, some transformer-based models have been proposed and achieved good performance in segmentation tasks. However, the self-attention blocks embedded in transformers often neglect the latent relationships among different samples. Model may have biased results due to the unbalanced data distribution in the training dataset. We propose a novel unbalanced weighted Unet equipped with a new ExSwin transformer block to comprehensively address the above concerns by effectively capturing long-range dependencies and correlations among different samples. We design a deeper encoder to facilitate features extracting and preserving more semantic details. In addition, an adaptive weight adjusting method is implemented to dynamically adjust the loss weight of different classes to optimize learning direction and extract more features from under-learning classes. Extensive experiments on a FeTA dataset demonstrate the effectiveness of our model, achieving better results than state-of-the-art approaches.

Keywords: Fetal brain MRI images · Transformer · Medical image segmentation

1 Introduction

Infancy is the origination stage of everyone's life, but some infants, unfortunately, suffer from congenital diseases and severe congenital diseases may lead to the

L. Karlinsky et al. (Eds.): ECCV 2022 Workshops, LNCS 13803, pp. 340–354, 2023.
https://doi.org/10.1007/978-3-031-25066-8_18

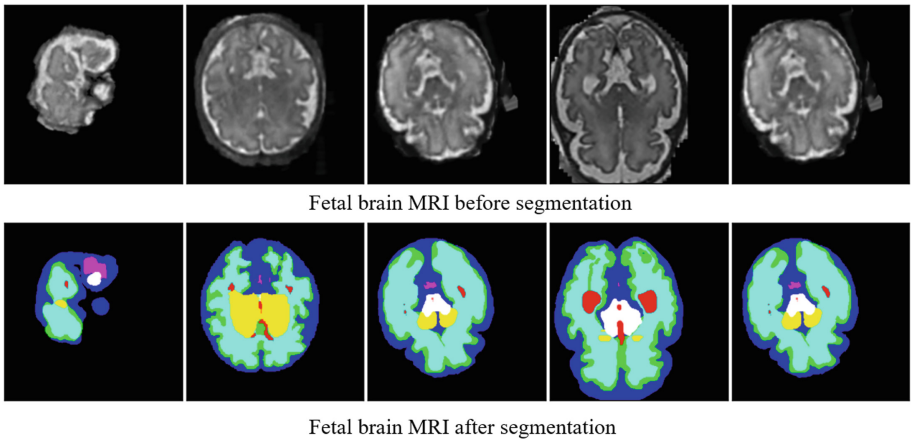

Fig. 1. Samples of fetal brain MRI segmentation dataset

death of infants [15]. In this regard, the timely discovery and treatment of infant congenital diseases are significant. For those unfortunate fetuses with congenital diseases, fetal brain MRI results are especially helpful to study the neuro development of the fetus and aid fetal disease diagnosis and treatment [13,24,25]. When conducting fetal brain analysis, precise segmentation of crucial structures in MRI images is essential. The fetal brain MRI images are complex and many congenital diseases result in subtle changes in brain tissues [4,6,28]. Thus, accurate segmentation of these tissues and structures plays a decisive role in diagnosis and treatment. Manual segmentation is quite laborious, time-consuming, and error-prone. Therefore, automatic segmentation of fetal brain MRI images is highly demanded in practice (Fig. 1).

Recent years, convolutional neural networks (CNNs) based segmentation models have dominated in medical image computing and achieved remarkable success [3,8,18,21,27]. They apply convolution kernel to perform convolution operations and extract features from local input patches and modularize representations while efficiently utilizing data. However, it is still difficult for these models to precisely segment brain tissues from fetal brain MRI images to capture the subtle changes in different brain tissues. One of the main concerns is the intrinsic locality of convolution operation, which causes CNN-based models difficult to extract long-range semantic information to enhance the segmentation performance in a global view [2]. On the other hand, a newly proposed architecture, namely transformer [23], has achieved great success in the natural language processing domain with its effective self-attention mechanisms. It has been introduced to the vision domain [5] and is widely employed in many computer vision tasks. It performs excellently in the CV area, surpasses the CNN-based model in some areas, and shows that a vigorous model can be constructed with a transformer. Recently, Swin transformer [17] has been proposed and performs well in image classification and detection, while the Swin-Unet [1] has shown its

powerful capability in image segmentation. However, the self-attention mechanism embedded in transformers often ignores the correlations among different samples, while correlations between different samples are essential for image segmentation. And Swin transformer structure brings about training instability when changing the window size of the transformer block. In addition, in medical segmentation tasks, the model often confronts limited and biased labeled data due to the limitation of the dataset, leading to unbalanced training and results.

We propose a novel unbalanced weighted Unet equipped with a new ExSwin transformer block for fetal brain MRI image segmentation in order to effectively capture long-range dependencies and correlations among different samples to enhance the segmentation performance. The ExSwin transformer block is composed of the window attention block [16] and the external memory block based on the external attention scheme [10]. The window attention block is responsible for local and global feature representation learning, while the external memory block combines different intra-samples' features with its two external memory units to reduce the information loss due to dimensional reduction and gain inductive bias information of the dataset. Furthermore, we design a special unbalanced Unet structure where we adopt a larger encoder size to facilitate features extracting and preserving deeper semantic information. In addition, an adaptive weight adjusting method is implemented to dynamically adjust the loss weight of different classes, which contributes to optimizing model learning direction and extracting more features from the under-learning classes. Since our dataset is from FeTA 2021 challenge, we implement comparison with several participators' networks, such as Unet, Res-Unet, and Trans-Unet, where our model has a better performance. Quantitative experiments and ablation studies on the dataset demonstrate the effectiveness of the proposed model, achieving better results than state-of-the-art approaches.

2 Method

The framework of our unbalanced ExSwin-Unet is as shown in Fig. 2. Our ExSwin Unet mainly consists of encoder, bottleneck, decoder and skip connections between encoder and decoder blocks. In our encoder module, input images are divided into non-overlapping patches with patch size 4×4 and the feature dimension of each patch becomes 16 times. Moreover, the feature dimension is projected to a selected dimension C through a linear embedding layer. After that, we continuously apply ExSwin blocks and patch merging layers alternately where ExSwin blocks grasp feature representation and patch merging layers increase feature dimension for down-sampling. Specifically, ExSwin block size is even since it needs to perform window and shift window attention alternately to capture local and global features of the image. Our ExSwin blocks are able to extract high-level features from input images. Then we apply two ExSwin blocks as the bottleneck block to enhance model convergence ability where the input feature dimension and output feature dimension are the same. On the other hand, in the decoder module, we apply patch expanding layers with multiple ExSwin

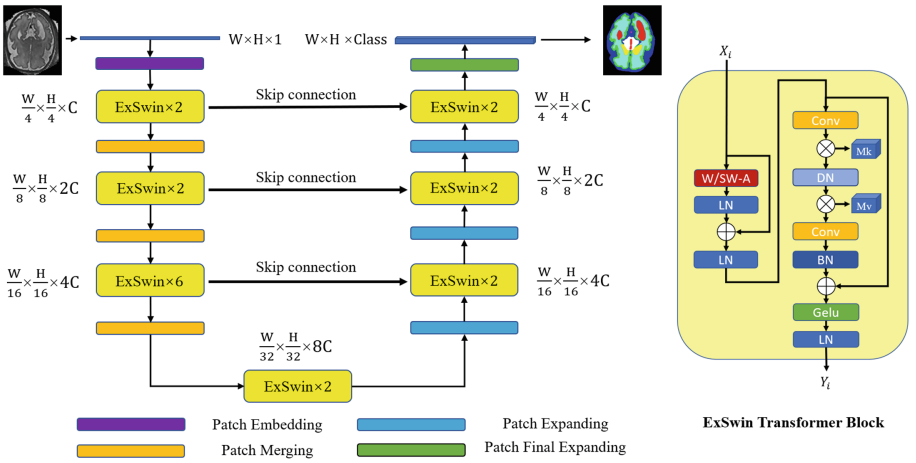

Fig. 2. Overview of our proposed ExSwin-Unet. In ExSwin transformer block, W/SW-A is window and shifted window attention module [17]; LN, DN BN represent layer normlization, double normlization [9] and batch normlization respectively; Mk and Mv are external learnable key and value memory respectively; Gelu is the Gaussian error linear unit.

blocks to perform features up-sampling hierarchically. Skip connections between same-level ExSwin blocks are applied to complement detailed information loss during the down-sampling process and retain more high-resolution details contained in high-level feature maps. At the end of the decoder, a particular patch expanding layer is added to conduct 4× up-sampling where feature resolution is mapped to input resolution. In the end, the up-sampled features will be mapped to segmentation predictions through a linear projection layer.

2.1 Window-Based Attention Block

Based on the shifted window mechanism and hierarchical structure, the Swin transformer is able to extract both local and global features of the input images. Since our Feta dataset samples are 2D images generated from 3D images, spatial information can be easily lost. In order to make up for the loss of spatial information and improve the feature fusion among different samples, we propose a new transformer block named as ExSwin transformer block. The ExSwin block is constructed with window-based attention and external attention block. The structure of our ExSwin block is shown in Fig. 2. The operation of the ExSwin block can be formulated as follows:

$$\begin{aligned} X_{i1} &= \text{W/SW-A}\left(\text{LN}\left(X_i\right)\right) + X_i, \\ \hat{X}_i &= \text{LN}\left(X_{i1}\right), \\ \hat{X_{i1}} &= \text{EA}\left(\text{Conv}\left(\hat{X}_i\right)\right) \\ \hat{X_{i2}} &= \text{BN}\left(\text{EA}\left(\text{Conv}\left(\hat{X_{i1}}\right)\right)\right) + \hat{X_{i1}}, \\ Y_i &= \text{LN}\left(\text{Gelu}\left(\hat{X_{i2}}\right)\right) \end{aligned} \tag{1}$$

where $\mathbf{X_i}, \mathbf{Y_i} \in \mathbb{R}^{C \times H \times W}$ represent the input features and output features of the $\mathbf{i^{th}}$ ExSwin transformer block; W/SW-A represents window and shifted window attention module; EA is the external attention module.

Adjustive Window Attention Block. In the window attention block, we apply window-based multi heads self-attention (W-A) module and shift window-based multi-head self-attention (SW-A) [17]. The window-based and shifted window-based multi-head self-attention module are applied in the two successive transformer blocks. Window-based multi-head self attention calculates attention in each window to capture local window features. On the other hand, shifted window-based multi-head self-attention, with its shifting mechanism, calculates attention to mix cross-window features and capture global features. The local self-attention can be formulated as:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}} + B\right)V \tag{2}$$

where $Q, K, V \in \mathbb{R}^{M^2 \times d}$ represents the query, key and value matrices; M^2 denotes the number of patches in a window and d denotes the dimension of the query or key; B is the relative position bias and its values are taken from the bias matrix $\hat{B} \in \mathbb{R}^{(2M-1)\times(2M+1)}$ since the relative position along each axis lies in the range of $[-M+1, M-1]$.

Post Normalization. When training the window attention-based model, we may probably encounter training instability since activation values in the network deep layers are quite low [16]. To ease the unstable situation, post normalization, shown in Fig. 3, is applied in attention blocks and adds an additional layer normalization unit before the external attention block.

Scaled Cosine Attention. While calculating the self-attention in window attention and shifted window attention module, the attention map in some blocks or heads dominated other features, which leads to biased feature extraction.

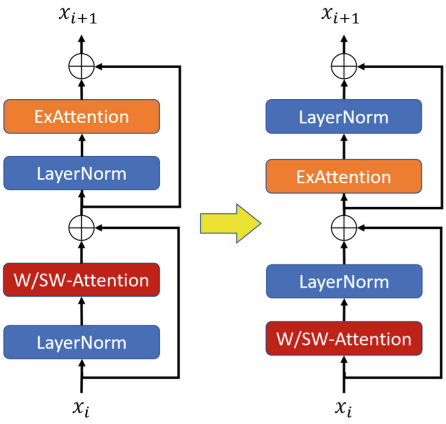

Fig. 3. The Pre-norm is transformed to Post-norm

We can replace the inner product similarity with cosine similarity to improve the problem:

$$\text{Similarity}\left(\mathbf{q}_i, \mathbf{k}_j\right) = \cos\left(\mathbf{q}_i, \mathbf{k}_j\right)/\tau + B_{ij} \tag{3}$$

where B_{ij} is the relative position bias between pixel i and j; τ is a learnable scalar, non-shared across heads and layers. Since the cosine function is equivalently normalized, the substitution can alleviate some inner product domination situation.

2.2 External Attention Block

In the external attention block, we design a multi-head external attention module that applies two convolution layers to grasp feature representation and two external learnable memory units to capture spatial information and sample affinity between different samples. The external attention block applies an external attention mechanism, which adopts two external memory units Mk and Mv to restore the spatial information between adjacent slices and store current global information. The external attention module is designed for capturing intra-sample features and it is capable of learning more representative features from input samples. The external attention block structure is as shown in Fig. 4 and the pseudo-code of our multi-head external attention module is as shown in Algorithm 1.

Since, multi-head attention and convolution mechanism are complementary, we apply two convolution layers in the external attention block. The first convolution layer kernel size is 1×1 in order to aggregate cross-channel features. In order to obtain a useful complement to the attention mechanism, the second convolution layer kernel size is 3×3 with padding size 1. The 3×3 convolution layer captures the local information with a larger receptive field and enhances grasping the feature representation [11].

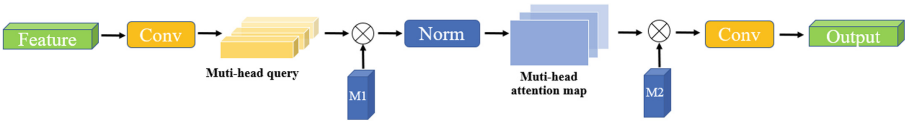

Fig. 4. The framework of external attention block which applies multi-head calculation. M1 and M2 are multiple convolution 1D kernels to store spatial information.

Algorithm 1. The pseudo code of the multi-head attention block.

Input: $\hat{X_{in}}$, a feature vector with shape [B, N, C] # (batch size, pixels, channels)
Parameter: H, the number of heads
Output: $\hat{X_{out}}$, a feature vector with shape [B, N, C]

$Query = Conv(\hat{X_{in}})$ # kernel size $= 1 \times 1$
$Query = Query$.view(B , N, H, C/H) # $shape = [B, N, H, M]$
$Query = Query$.permute(0 ,2 ,1 ,3) # $shape = [B, H, N, M]$
$Attn = M_k$(Query) # $shape = [B, H, N, M]$
Double normalization
$Attn = Softmax$(Attn, dim $= 2$)
$Attn = L1Norm$(Attn, dim $= 3$)
$Out = M_v$(Attn) # $shape = [B, H, N, M]$
$Out = Out$.permute(0 ,2 ,1 ,3) # $shape = [B, N, H, M]$
$Out = Out$.view(B , N, C) # $shape = [B, N, C]$
$\hat{X_{out}} = Conv(Out)$ # kernel size $= 3 \times 3$, stride $= 1$

By utilizing two external memory units to recover and store the spatial information of slices in a 3D sample, our external attention block can be viewed as the dictionary for the whole dataset to calculate attention among 2D slices. The external attention module benefits model learning representative features and alleviates feature loss of dimension reduction process.

2.3 Unbalanced Unet Architecture

In the encoder-decoder unet structure, the sizes of Ex-Swin blocks in the encoder and decoder are different. The Ex-Swin blocks size are $[2, 2, 6]$ and $[2, 2, 2]$ for the encoder and decoder module, respectively. The idea of hyper-parameters setting is inspired by Swin-T model and our experimental results also have proven its effectiveness against balanced Unet structure. The encoder block with deeper size Ex-Swin blocks is able to obtain a better feature extraction and enhance model ability of preserving broad contextual information. The decoder block with a thinner size saves calculation resources and also benefits model convergence.

2.4 Adaptive Weighting Adjustments

In medical segmentation tasks, we may encounter a biased data segmentation training result due to the limited and unbalanced labeled data situation. To

alleviate the biases and increase model performance, we propose an adaptive weighting adjustment strategy on loss function, which conducts model learning on under-learning samples and also prevents the model from overwhelming the well-perform samples during the model training process. In our adaptive weighting adjustment mechanism, the weight value $\mathbf{v_c}$ for the class c is calculated by:

$$v_c = \frac{1}{N_c} \sum_{i=1}^{N_c} \frac{|\mathrm{pre}_c|}{|\mathrm{tru}_c|} \tag{4}$$

where $|\mathrm{pre_c}|$ is prediction pixels that match the ground truth pixels for class **c** and the $|\mathrm{tru_c}|$ is the total number of class **c** in the corresponding ground truth. And the adaptive weight w_c can be calculated by the weight value:

$$w_c = \mathrm{Softmax}(1 - v_c) \tag{5}$$

where the class-wise weight will be updated for every epoch training process. In that case, a suitable weight is generated to enable model adjusting its learning direction and alleviating the biased segmentation result.

2.5 Dual Loss Functions

In order to improve segmentation accuracy and learning speed, we define a dual loss function. Since we adopt an adaptive weight adjusting method, loss is obtained by calculating weighted loss for each class by taking the average value. Assume that w_c is the weight for each class and C is the total number of classes.

Multi-class Cross Entropy Loss. Cross entropy loss measures the difference between two probability distributions. It fastens model convergence and reduces model training resource consumption.

$$\mathcal{L}_{ce} = -\frac{1}{C} \times \sum_{i=1}^{C} w_c \times l_c \log(p_c) \tag{6}$$

where p_c is the segmentation probability for class c in the output, l_c is the identification for class c which ranges 0 or 1 and w_c is adaptive weight for class c.

Square Dice Loss. Dice loss measures similarity of two distributions and we apply it to calculate the similarity between output prediction and the ground truth. It conduces to improve model performance and increase accuracy.

$$\mathcal{L}_{dice} = \frac{1}{C} \times \sum_{i=1}^{C} w_c \times \left(1 - \frac{2 \times \sum_{\mathrm{pixels}} y_{\mathrm{pred}}\, y_{\mathrm{true}}}{\sum_{\mathrm{pixels}} \left(y_{\mathrm{pred}}^2 + y_{\mathrm{true}}^2\right)}\right) \tag{7}$$

where $\sum_{\mathrm{pixels}}$ represents the sum of pixel value, y_{pixels} and y_{true} are segmentation prediction and segmentation ground truth respectively and w_c is adaptive weight for class c.

Table 1. Quantitative comparison results of segmentation results on feta2021 dataset. The table shows different methods segmentation prediction with % unit. All methods are evaluated by Dice, Jaccard, Sensitivity and Specificity coefficient.

Method	Year	Dice	Jaccard	SE	SP	Para(M)	Flop(GMac)
Attn-Unet [19]	2018	88.6	80.5	88.9	99.7	**34.9**	66.6
Segmenter [22]	2021	87.5	78.6	87.3	99.7	102.3	25.8
Swin-Unet [1]	2021	88.7	80.7	89.1	99.7	34.2	**9.6**
Utnet [7]	2021	89.1	81.1	88.9	99.8	35.1	49.7
Trans-Unet [2]	2021	89.2	81.3	89.3	99.8	105.3	35.2
Ours	2022	**90.1**	**82.3**	**90.2**	**99.8**	50.4	11.3

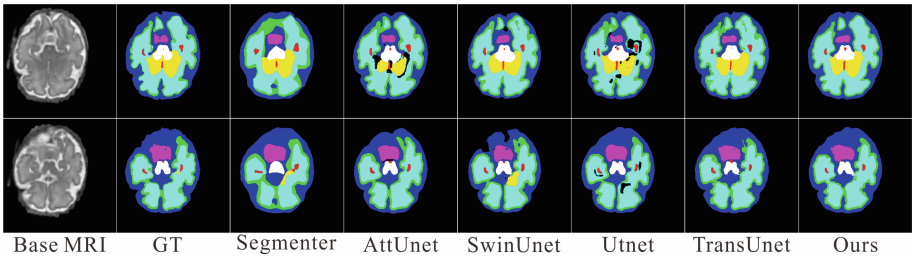

Fig. 5. Segmentation visual results of different methods on FeTA2021 dataset.

Total Loss. The total loss is linear combination of average weighted CE loss and average weighted DICE loss with coefficient α.

$$\mathcal{L}_{total} = \alpha \times \frac{1}{N} \times \sum_{i=1}^{N} \mathcal{L}_{ce} + (1-\alpha) \times \frac{1}{N} \times \sum_{i=1}^{N} \mathcal{L}_{dice} \tag{8}$$

3 Experimental Results

3.1 Datasets

The dataset is Fetal Brain Tissue Annotation and Segmentation Challenge released in 2021 [20]. The fetal brain MRI was manually segmented into 8 different classes with in-plane resolution of 0.5 mm × 0.5 mm. Dataset includes 80 3D T2-weighted fetal brain and reconstruction methods were used to create a super-resolution reconstruction of the fetal brain. We divided the dataset into 60 training set and 20 testing set. In order to save time and energy consumption, we transform dataset to about 2D images with size 256 × 256.

3.2 Implement Details

We train and test our model on a single NVIDIA RTX 2080Ti (11 GB RAM). The ExSwin-Unet model is trained on Python 3.7 and Pytorch 1.7.0. In order to increase data diversity and avoid data overfitting, we applied simple data augmentation flipping and rotation on dataset. We adopt weighted dual loss function and employ lookahead optimizer [26] with Adam optimizer [12] as inner optimizer. Moreover, we experimentally set the coefficient of total loss $\alpha = 0.4$ to obtain a relatively better performance. During the model training period, the initial learning rate is 1e−4 and loss decay for each epoch. We trained the model for 200 epochs with a batch size of 16.

3.3 Comparison with SOTA Methods

To evaluate the performances of our method, we compared our network with five state-of-the-art methods including Segmenter [22], Attn-Unet [19], Utnet [7], Swin-Unet [1] and Trans-Unet [2]. The compared models consist of four transformer based model structures and a CNN based models, namely Attention Unet. We implemented the comparison under the same computational environments without using any pre-trained models. Both visual and statistical comparisons are conducted using the same datasets and with same data processing method. The statistical comparison results are shown in Table 1 and visualization results are shown in Fig. 5.

Our model with its unique features generally outperforms other SOTA methods on dice and jaccard score and cost less calculation consumption. We save 50% parameters than Trans-Unet and achieve a better segmentation performance. Visually compared with other segmentation methods in the Fig. 5, our model also outperforms on segmenting fetal tissue with different scales and irregular shapes. Demonstrating that the proposed ExSwin-Unet is capable to improve the segmentation performance.

3.4 Ablation Studies

In order to demonstrate the effectiveness of the proposed components, we conduct ablation studies with different components and unbalanced structure. The component ablation experiment results are as shown in Table 2 and the attention hotspots of different methods are as shown in Fig. 6. To illustrate the effectiveness of our unbalanced structure, we conducted ablation studies on the unbalanced structure as shown in Table 4.

As shown in the Table 2, unbalanced Unet structure benefits feature learning process with its larger encoder size. The model with the external attention unit is able to combine different intra-sample features and mitigate spatial information loss. The adaptive loss alleviates model imbalance class under-learning problems.

As shown in Table 3, we further implement comparison with focal loss [14], a typical class imbalance loss, to verify the effectiveness of our adaptive loss

Table 2. Ablation studies result on FeTA2021 dataset. The table shows different methods segmentation prediction dice level with % unit. ECF, GM, WM, Ve, Ce, DGM and Br are 7 segmented brain tissues representing External Cerebrospinal Fluid, Grey Matter, White Matter, Ventricles, Cerebellum, Deep Grey Matter, Brainstem, respectively. Specially, except for the first Swin-Unet method is balanced, others structure are unbalanced Unet structure version with encoder size $[2, 2, 6]$ and decoder size $[2, 2, 2]$.

Method	Mean	ECF	GM	WM	Ve	Ce	DGM	Br
Swin-Unet(Balanced)	88.7	89.9	79.9	92.7	91.3	88.4	88.3	89.5
Swin-Unet(Unbalanced)	89.2	90.7	80.8	93.3	91.7	89.3	88.9	89.8
Swin-Unet+Adaptive	89.5	90.5	82.5	93.6	91.5	89.0	89.2	89.7
ExSwin-Unet	89.7	**91.3**	81.4	93.8	92.1	89.4	89.6	**90.4**
ExSwin-Unet+Adaptive	**90.1**	91.2	**82.9**	**94.2**	**92.3**	**89.8**	**89.9**	90.2

Table 3. Ablation studies result on FeTA2021 dataset. The table shows different methods segmentation prediction dice level with % unit. ECF, GM, WM, Ve, Ce, DGM and Br are 7 segmented brain tissues representing External Cerebrospinal Fluid, Grey Matter, White Matter, Ventricles, Cerebellum, Deep Grey Matter, Brainstem, respectively.

Method	Mean	ECF	GM	WM	Ve	Ce	DGM	Br
Swin-Unet+Focal	88.6	90.1	81.7	93.3	90.8	87.3	87.9	89.4
Swin-Unet+Adaptive	89.5	90.5	82.5	93.6	91.5	89.0	89.2	89.7
ExSwin-Unet+Focal	89.3	90.6	82.0	93.5	91.7	88.7	89.0	89.9
ExSwin-Unet+Adaptive	**90.1**	**91.2**	**82.9**	**94.2**	**92.3**	**89.8**	**89.9**	**90.2**

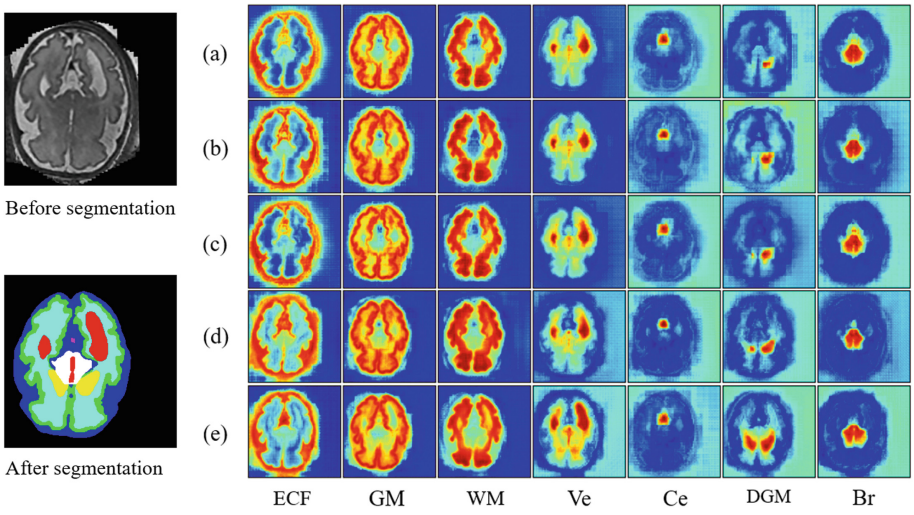

Fig. 6. Multiple classes attention hotspot of different methods. (a)–(e) are the ablation experiment methods balanced Swin-Unet, unbalanced Swin-Unet, unbalanced Swin-Unet+Adaptive, unbalanced ExSwin-Unet and unbalanced ExSwin-Unet+Adaptive correspondingly.

Table 4. Ablation studies on unbalanced Unet architecture. The mean dice is calculated through five-fold cross-validation to verify our method's effectiveness.

Method	Encoder size	Decoder size	Mean Dice ± std
ExSwin-Unet	[2, 2, 6]	[2, 2, 6]	89.5 ± 0.349
ExSwin-Unet	[2, 2, 2]	[2, 2, 2]	89.7 ± 0.193
ExSwin-Unet	[2, 2, 6]	[2, 2, 2]	$\mathbf{90.1 \pm 0.252}$

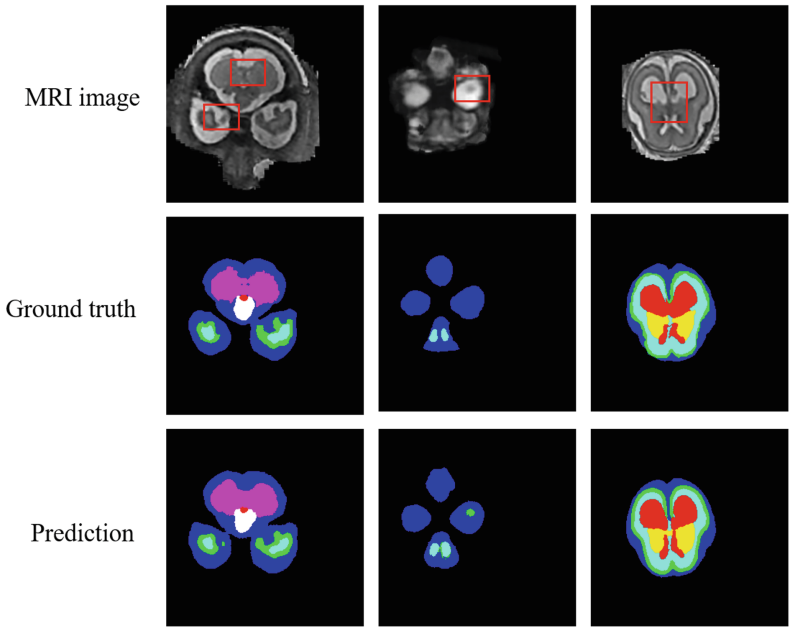

Fig. 7. Visual segmentation results of some failure predicting cases.

method. The results demonstrate the our method effectiveness, where our adaptive weighted loss function benefits model learning ability by grasping information of under-learning classes and improving overall performance.

The ablation studies on unbalanced Unet structure is shown in the Table 4, demonstrating the effectiveness of unbalanced Unet structure. With a lager encoder size, our model can achieve a better performance than two other balanced models. The experiment indicates that the unbalanced structure benefit model feature extraction process and improve segmentation result.

4 Discussions and Limitations

Through the above ablation studies and comparative experiments, we design an effective 2D-based segmentation network with external attention to implement

segmentation tasks on 3D image slices. Our purpose is to discover intra-sample relationships to alleviate spatial information loss and benefit the feature learning process. The external attention module achieves this goal, and experiments demonstrated its effectiveness. Moreover, we discover that balanced Unet structure may not be necessary for Unet framework where unblanced Unet can obtain a better performance than balanced Unet. On the other hand, our method still has some limitations. As shown in Fig. 7, our model fails to achieve correct predictions on some small scales.

5 Conclusion

In this paper, we present a novel unbalanced weighted Unet equipped with a new ExSwin transformer block to improve fetal brain MRI segmentation results. The ExSwin transformer is composed of shift-window attention and external attention module. The ExSwin transformer block not only can grasp essential sample features representation, but it also is able to capture intra-sample correlation and spatial information between different 3D slices. And the Unet is unbalanced where the encoder has a larger size to facilitate the feature extracting process. Furthermore, we introduce an adaptive weight adjustment strategy to improve biased data segmentation situations. The quantitative comparison experiments and ablation studies demonstrate the well performance of our proposed model.

Acknowledgments. This work was supported partly by National Natural Science Foundation of China (No. 61973221), Natural Science Foundation of Guangdong Province, China (No. 2019A1515011165), the Innovation and Technology Fund-Mainland-Hong Kong Joint Funding Scheme (ITF-MHKJFS) (No. MHP/014/20) and the Project of Strategic Importance grant of The Hong Kong Polytechnic University (No. 1-ZE2Q).

References

1. Cao, H., et al.: Swin-Unet: Unet-like pure transformer for medical image segmentation. arXiv preprint arXiv:2105.05537 (2021)
2. Chen, J., et al.: TransUNet: transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
3. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. **40**(4), 834–848 (2017)
4. Clouchoux, C., et al.: Delayed cortical development in fetuses with complex congenital heart disease. Cereb. Cortex **23**(12), 2932–2943 (2013)
5. Dosovitskiy, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
6. Egaña-Ugrinovic, G., Sanz-Cortes, M., Figueras, F., Bargalló, N., Gratacós, E.: Differences in cortical development assessed by fetal MRI in late-onset intrauterine growth restriction. Am. J. Obstet. Gynecol. **209**(2), 126-e1 (2013)

7. Gao, Y., Zhou, M., Metaxas, D.N.: UTNet: a hybrid transformer architecture for medical image segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12903, pp. 61–71. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87199-4_6
8. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
9. Guo, M.H., Cai, J.X., Liu, Z.N., Mu, T.J., Martin, R.R., Hu, S.M.: PCT: point cloud transformer. Comput. Vis. Media **7**(2), 187–199 (2021)
10. Guo, M.H., Liu, Z.N., Mu, T.J., Hu, S.M.: Beyond self-attention: external attention using two linear layers for visual tasks. arXiv preprint arXiv:2105.02358 (2021)
11. Guo, R., Niu, D., Qu, L., Li, Z.: SOTR: segmenting objects with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7157–7166 (2021)
12. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
13. Li, Z., Pan, J., Wu, H., Wen, Z., Qin, J.: Memory-efficient automatic kidney and tumor segmentation based on non-local context guided 3D U-Net. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12264, pp. 197–206. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59719-1_20
14. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)
15. Liu, L., et al.: Global, regional, and national causes of child mortality: an updated systematic analysis for 2010 with time trends since 2000. Lancet **379**(9832), 2151–2161 (2012)
16. Liu, Z., et al.: Swin transformer v2: scaling up capacity and resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12009–12019 (2022)
17. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
18. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
19. Oktay, O., et al.: Attention U-Net: learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
20. Payette, K., et al.: An automatic multi-tissue human fetal brain segmentation benchmark using the fetal tissue annotation dataset. Sci. Data **8**(1), 1–14 (2021)
21. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
22. Strudel, R., Garcia, R., Laptev, I., Schmid, C.: Transformer for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7262–7272 (2021)
23. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
24. Wu, H., Lu, X., Lei, B., Wen, Z.: Automated left ventricular segmentation from cardiac magnetic resonance images via adversarial learning with multi-stage pose estimation network and co-discriminator. Med. Image Anal. **68**, 101891 (2021)

25. Wu, H., Pan, J., Li, Z., Wen, Z., Qin, J.: Automated skin lesion segmentation via an adaptive dual attention module. IEEE Trans. Med. Imaging **40**(1), 357–370 (2020)
26. Zhang, M., Lucas, J., Ba, J., Hinton, G.E.: Lookahead optimizer: k steps forward, 1 step back. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
27. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2881–2890 (2017)
28. Zugazaga Cortazar, A., Martín Martinez, C., Duran Feliubadalo, C., Bella Cueto, M.R., Serra, L.: Magnetic resonance imaging in the prenatal diagnosis of neural tube defects. Insights Imaging **4**(2), 225–237 (2013). https://doi.org/10.1007/s13244-013-0223-2