DIV-ATTENTION: A PLUG-AND-PLAY MODULE FOR 3D MEDICAL IMAGE SEGMENTATION

$Zhishan\ Jiang^{1,2}, Ke\ Yan^{2}, Le\ Lu^{2}, Minfeng\ Xu^{2,*}$

[1]Tsinghua University, [2]Alibaba DAMO Academy

ABSTRACT

In medical image segmentation, fusion of multi-scale feature maps using skip-connection is proven to be an effective technique for performance boosting. However, one feature map is sufficient to make a prediction about the segmentation result, and there exists divergence between predictions made by feature maps of different scales. To better utilize feature maps of different scales, in this paper, we propose a plug-and-play module named div-attention. This module can be applied to any deep neural networks with skip-connections. It is placed at the junction between adjacent paths of skip-connections to re-segment the area where divergence exists and to fuse results for segmentation performance improvement, which means that this module is divergence-aware and is proposed for better utilizing this disagreement between predictions made by feature maps of diverse scales within one model. Experiment results show that this plug-and-play module can significantly boost segmentation performance in different deep neural networks and in diverse CT segmentation tasks.

Index Terms**—** *Divergence, attention, plug-and-play*

1. INTRODUCTION AND RELATED WORK

Medical image segmentation plays a pivotal role in intelligent medical treatment. It extracts regions of interests in medical images like CT or MRI and helps to improve efficiency and accuracy of medical decision. However, medical image segmentation remains to be a challenging task for the low resolution of medical images and for the complex and diverse situation in vivo.

The introduction of Unet [1] has brought great improvement in medical image segmentation. Since then, methods based on a Unet-like architecture with skip-connection are proposed. The skip-connection concatenates encoding blocks and decoding blocks on the same level, contributing to fuse feature maps of different receptive fields and resolutions to take advantage of both details and high-level features in feature maps of different scales. Besides, DeepLab [2] uses atrous convolution to extract feature maps from different receptive fields without extra computational expenses.

The prevail of transformer [3] and attention mechanism in natural language processing has attracted much attention since its advent. ViT [4] and SwinTransformer [5] exert attention in computer vision tasks by transforming image patches into sequences. In medical image segmentation, correspondingly, Unetr [6] and SwinUnetr [7] are proposed.

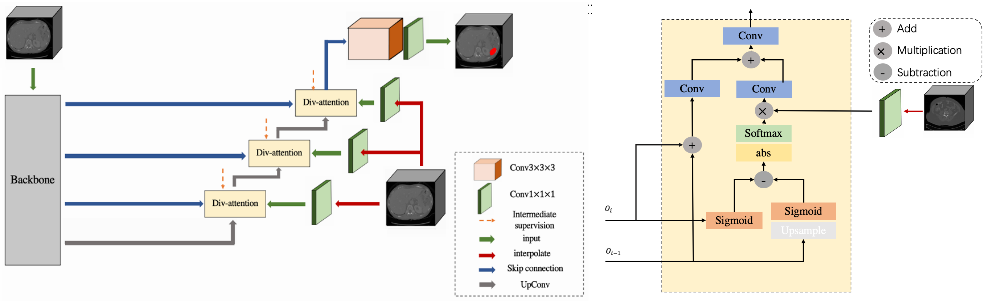
As for segmentation refinement, Golnaz Ghiasi and Charless C. Fowlkes [8] propose a method to trade advantages off between feature maps of different resolutions by suppressing the contribution of higher resolution layers in areas where the segmentation is confident. Chuong Huynh et al. [9] propose a refinement module for segmentation refinement on high-resolution images. This module refines regimentation result based on uncertainty scores acquired by comparing and fusing predictions through multiple stages and different classes.

This paper proposes a novel plug-and-play divergence-aware attention module for emphasizing the area where may brings a boost in segmentation performance. This module is placed at the junction between adjacent skip-connections (will be denoted as path below for simplification) in any encoder-decoder network with skip-connections, making full use of prediction disagreement between feature maps of different scales. The pixel/voxel of divergence will be re-segmented and the re-segmented area will be appended to the previous results for segmentation enhancement. Results on both publicly available and inhouse dataset implies the effectiveness of the proposed module.

2. METHODS

We propose a divergence-aware attention module named div-attention for medical image segmentation enhancement. The implementation of the module is presented in Fig 1-b. To demonstrate where we place our module, a Unet-like architecture with skip-connection between multi-scale feature maps is posed as the example in Fig 1-a. This module aims at giving a refined result at where the output of two adjacent paths disagrees.

Div-attention Module. Denoting feature maps of two adjacent paths as O_i and O_{i-1} with O_i being the upper one. The output of lower path is first up-sampled spatially through transposed convolution and reshaped along channel dimension through convolution with kernel size $1 \times 1 \times 1$ to



(a) (b)

Fig. 1. Proposed div-attention module and diagram of its placement. (a) demonstrates the placement of div-attention module and (b) depicts the detailed architecture of the proposed module.

align it with O_i:

$$O_{i-1} := Conv_{1\times1\times1}(UpConv(\,O_{i-1})) \quad (1)$$

Then Sigmoid is imposed separately to simulate mask prediction and we get M_i and M_{i-1}, which represent the predicted masks of two adjacent paths:

$$M_i = Sigmoid(\,O_i) \quad (2)$$

$$M_{i-1} = Sigmoid(\,O_{i-1}) \quad (3)$$

To get divergence of two adjacent paths, absolute value of residual between M_i and M_{i-1} is used and the result is activated through Softmax function and then rescaled according to the size of M_i as (D, H, W):

$$attn = \frac{|Softmax(M_i - M_{i-1})|}{\sqrt{D \cdot H \cdot W}} \quad (4)$$

The attention map $attn$ is imposed on the feature map F_i extracted from the input by an extra branch aside. F_i is obtained by interpolating and convolving the input. By multiplying F_i by the attention map we get an image with highlighted area where the predictions of the adjacent paths mostly disagree:

$$M_{ref} = attn * F_i \quad (5)$$

This divergence-highlighted output is finally fused with O_{i-1} and O_i by adding and convolutions to retain the final segmentation output of path $i-1$.

An intermediate supervision is imposed at every path except the lowest one. For each path, the intermediate result is predicted by imposing an $1\times 1\times 1$ 3D convolution on the divergence-enhanced feature map obtained from the div-attention module.

Pipeline. As for the segmentation task, we use the model depicted in Fig.1 only in Stage 2. The whole pipeline can be denoted as a coarse-to-fine process. In Stage 1, we use an nnUnet [10] for coarse segmentation. With one-click information fed as cue for the location information of the targeted lesion or organ, In practical application, to better locate the lesion or organ of interest, some interactive information is required to assist medical treatment and analysis. We introduce the one-click information randomly generated within the mask of target lesion or organ in the form of distance transform image to the input in Stage 1, which can simulates the procedure of doctor intervention that enforce the model to focus on the target lesion or organ. According to the prediction of Stage 1, principle component analysis [11] is used to obtain an estimate of the longest axis length of the target tumor or organ. The input image will be cropped into a smaller patch and resized to a uniform size, then fed into the model of Stage 2.

Losses. For intermediate supervision imposed on every path (except for the lowest one), as mentioned, the divergence of adjacent paths will be highlighted by the proposed module and re-segmented. The area where divergence exists, however, will always advent around the contour of target lesion or organ, and there are where false positives prevail. To benefit more from the proposed divergence-aware attention module, we utilize a combination of weighted binary cross entropy loss and Dice loss to supervise both intermediate and final result, in which we impose more penalties on negative instances:

$$wBCE = -\frac{1}{N}\sum\big((1-y)\log(1-p) + \gamma\, y\, \log(p)\big) \quad (6)$$

$$Dice = 1 - \frac{2O_{pred} \cdot M_{gt}}{O_{pred} + M_{gt} - O_{pred} \cdot M_{gt}} \quad (7)$$

$$Loss = \alpha \cdot wBCE + \beta \cdot Dice \quad (8)$$

where p denotes the predicted probability of a pixel belonging to the target lesion and O_{pred} being the output of the model. M_{gt} denotes the ground truth of lesion segmentation task.

Analysis. The div-attention module uses absolute value of residual between predictions of adjacent paths as the

Table. 1. Experiment results on both public and private datasets, baselines with and without the proposed module.

	MSD-Spleen	KITS19	Lung	Lymph
	Dice Recall Precision	Dice Recall Precision	Dice Recall Precision	Dice Recall Precision
3D ESPNet	0.8802 0.8871 0.8883	0.8089 0.8475 0.7874	0.8167 0.8562 0.8003	0.7769 0.8012 0.7646
3D ESPNet+div-attention	0.8852 0.9114 0.8675	0.8357 0.8525 0.8560	0.8339 0.8630 0.8260	0.7880 0.8459 0.7745
Unetr	0.9032 0.9074 0.9072	0.8094 0.8184 0.8150	0.8148 0.8230 0.8283	0.7574 0.7985 0.7166
Unetr+div-attention	0.9194 0.9220 0.9221	0.8225 0.8161 0.8425	0.8340 0.8405 0.8440	0.7751 0.8015 0.7768

attention map. This attention map has voxels valued from 0 to 1 being the extent of the disagreement between their predictions on whether it is foreground or not. Large value implies that the voxel is hard to classify. By introducing an extra branch to extract feature map from the input and imposing the attention map on it, the voxels that are hard to classify will be re- segmented and the result will be fused with predictions made by the two paths through delicate operation. This process boost the segmentation performance by preserving the consistent predictions and examining the contradictory ones.

3. EXPERIMENTS AND RESULTS

To demonstrate the effectiveness of the proposed div-attention module, we plug it into different baselines and we train models on several datasets including both publicly available and inhouse ones.

Datasets and Evaluation Criteria. We use the following publicly available datasets in our experiments: MSD-Spleen [12], KITS19 [13]. As for other lung tumor and lymph tumor, we use our inhouse datasets. Our private lung and lymph tumor segmentation datasets includes 2600 and 1883 cases respectively. All the datasets mentioned above are datasets of CT images, 3D masks are labeled in binary form for every single lesion or organ. For publicly available datasets mentioned above, the training set is split into two parts, with 80% for training and 20% for validation. For the private ones, 30% cases of the dataset act as the test set, while 80% and 20% of the rest 70% will be utilized for training and evaluation respectively. To evaluate segmentation result, we use pixel-wise dice, recall and precision.

Implementation Details. We use the same model in Stage 1 for all experiments. Cropped patches according to prediction are fed into different baselines and different baselines with our proposed models. For Stage 2, we set batch size to be 4 or 8 for different datasets and the optimizer adopted is Adam. We set learning rate to be 1e-4, and this will be decayed by 0.1 after 15, 30, 60, 90 epochs. The epoch is set as 120, early-stopping is also used which means the training process will suspend when over-fitting is met. For Stage 2, all the input are resized to be 96x96x96, lesions with too small sizes (less than 2 in one dimension or less than 5 in pixels) are discarded during dataset preprocessing.

Ablation Study on Div-attention module. We plug our model into different baselines and the ablation study results on the test sets of all the datasets are demonstrated in Table 1. For baseline, we choose 3D ESPNet [14] and Unetr, both of them have skip-connections between feature maps of different scale. In Unetr, we simply plug the proposed module at the connection between two adjacent skip-connections(paths) for div-attention module experiment. For 3D ESPNet, however, we replace the ESP module in decoder with the proposed module for experiment 3D ESPNet + div-attention.

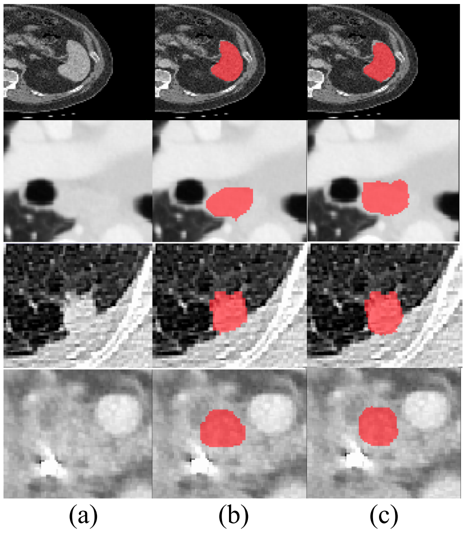
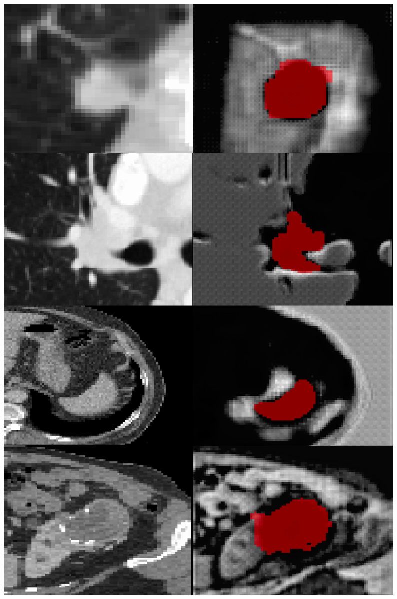


Fig. 2. Visualization of segmentation result in Stage 2. Given a cropped 3D CT image (a), (b) is corresponding segmentation annotation, (c) is prediction of Unetr with proposed module. All of the sub-images presented are obtained from results on axial plane.

Segmentation Result Visualization. The result of segmentation on the test sets is visualized in Fig. 2, as displayed, the baseline with proposed module has good performance on segmentation tasks on different lesion or organ type.

Attention Map Visualization. The attention maps within div-attention module as the visualization of "attn" demonstrated in (4) are displayed in Fig.3. From Fig.3, we can see that the highlighted areas always advent around the contour of the target lesion or organ, which means these are areas where predictions made by feature maps of different scales most disagree. As we also feed the CT input into the module using an extra branch aside, voxels with high intensity are also visible in the attention map, this is because simply several convolutional layers will not bring significant change to the original input, the voxel with high intensity will be partially preserved.



(a) (b)

Fig. 3. Visualization of attention map with mask annotation and corresponding input(only in Stage 2 as the div-attention module is used). Images in (b) denote attention map in grayscale and target lesion or organ annotation in red. Images in (a) denote the corresponding input.

4. CONCLUSION

This paper proposes a novel plug-and-play attention module based on divergence or disagreement of segmentation results between adjacent branches in multi-branch network. This module boosts segmentation performance of a network by highlighting the area of disagreement and enforcing re-segmentation of these areas. Results on several datasets demonstrates the effectiveness of the proposed module.

5. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access by (Source information). Ethical approval was not required as confirmed by the license attached with the open access data.

6. REFERENCES

[1] O. Ronneberger, P. Fischer, and T.Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Proc.Med. Image Comput. Comput-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234-241.

[2] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuile. “Deeplab: Semantic iamge segmentation with deep convolutional nets, atrous convolution, and fully connected crfs.”, *axXiv: 1606.00915*, 2016.

[3] Vaswani Ashish, Shazeer Noam, Parmar Niki, Uszkoreit Jakob, Jones Llion, Gomez Aidan N., Kaiser Lukasz, Polosukhin Illia, “Attention is all you need”, In *NeurIPS***,** 2017.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, “An iamge is worth 16x16 words: Transformer for image recognition at scale”, In *International Conference on Learning Representations*, 2021

[5] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo, “Swin Transformer: Hierarchical vision transformer using shifted windows”, In *Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV)*, pages 10012-10022, 2021

[6] Hatamizadeh, A., Yang D., Roth H., and Xu D, “Unetr: Transformers for 3D Medical Image Segmentation”, In *IEEE Winter Conference on Application of Computer Vision*, 2020

[7] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger Roth, and Daguang Xu. “Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images”, *arXiv preprint* arXiv:2201.01266, 2022.

[8] Golnaz Ghiasi, Charless C. Fowlkes. “Laplacian Pyramid Reconstruction and Refinemnet for Semantic Segmentation”, In *Conference on Computer Vision and Recognition*, 2016

[9] Chuong Huynh, Anh Tuan Tran, Khoa Luu, Minh Hoai. “Progressive Semantic Segmentation”, In *Conference on Computer Vision and Recognition*, 2021

[10] F.Isensee, P.Jger, J. Wsserthal, D. Zimmerer, J, Petersen, S. Kohl, J. Schock, A.Klein, T. Ro, S. Wikert, P.Neher, S. Dinkelacker, G. Köhler, K. Maier-Hein, “nn-Unet: A self-configuring method for deep learning-based biomedical image segmentation”, *Nature Methdos*, vol. 18, no 2. pp. 203-211, Feb. 2021.

[11] H. Abdi, L.J. Williams, “Principal component analysis”, *Wiley Interdisciplinary Reviews: Computational Statistics*, 2 (4) 433–459, 2010.

[12] M. Antonelli, A. Reinke, S. Bakas, K. Farahani, B. A. Landman, G. Litjens, B. Menze, O. Ronneberger, R. M. Summers, B. van Ginneken et al., “The medical segmentation decathlon,” *arXiv*, 2021.

[13] Heller, N., Sathianathen, N.J., Kalapara, A.A., Walczak, E., Moore, K., Kaluzniak, H., Rosenberg, J.E., Blake, P., Rengel, Z., Oestreich, M., Dean, J., Tradewell, M.B., Shah, A., Tejpaul, R., Edgerton, Z., Peterson, M., Raza, S., Regmi, S.K., Papanikolopoulos, N., & Weight, C.J., “The KiTS19 Challenge Data: 300 Kidney Tumor Cases with Clinical Context, CT Semantic Segmentations, and Surgical Outcomes”, *arXiv: 1904.00445, 2019*

[14] Nuechterlein, N., Mehta, S., “3d-espnet with pyramidal refinement for volumetric brain tumor image segmentation”, In *International MICCAI Brainlesion Workshop*, pp. 245-253, Springer, 2018.