

Supplementary Material: Few-Shot 3D Volumetric Segmentation with Multi-Surrogate Fusion

Meng Zheng¹, Benjamin Planche¹, Zhongpai Gao¹,
Terrence Chen¹, Richard J. Radke², and Ziyang Wu¹

¹ United Imaging Intelligence, Boston, MA, USA
`{first.last}@uii-ai.com`

² Rensselaer Polytechnic Institute, Troy, NY, USA
`rjradke@ecse.rpi.edu`

In this supplementary material, we provide further implementation details and information on the proposed MSFSeg pipeline and framework.

1 Implementation Details

1.1 Model Architecture

CNN Encoder. For visual feature extraction, we adopt a ResNet-101 [4] pre-trained on ImageNet-1K [1] for our proposed MSFSeg, for fair comparison with methods presented in Table 1 in the main paper. We implement our code in PyTorch. We obtain multi-scale (fixing $b = 4$) query/support feature maps from the *conv2*, *conv3*, *conv4*, *conv5* layer of the pretrained ResNet-101 (*c.f.* Methodology – Section 2) from the input query and support images. After generating the initial query masks in b scales from the multi-head attention, we aggregate the b initial query mask features by upsampling and addition to generate multi-scale-informed query mask features $\hat{\mathbf{M}}_{1..n}^q$. Encoding is performed for all visual inputs, *i.e.*, the support image/frame and the n support ones. Our proposed MSFSeg allows flexible n during training/evaluation – we set $n = 3$ during training for all our experiments, and $n = 1$, or 5 for evaluations (as presented in the tables in our main paper).

Multi-Head (Self-) Attention. Our Multi-head layer takes query/support image feature maps and support masks as input and passes a copy to each attention head. Each head flattens the feature maps and adds positional encoding, according to [13], and then applies the attention operation as defined in Section 2 to generate initial query mask features.

Multi-Surrogate Fusion. Each surrogate is implemented according to the equations presented in the main paper. Mask feature maps from the surrogates are concatenated and fused via a 3D convolutional layer, with input channel = 4 (number of surrogates) and output channel = 1, kernel size $1 \times 1 \times 1$, stride = 1, padding = 0, with bias applied. Please note that the only optimizable parameters introduced by our MSF module are the kernel and bias weights from the 3D convolutional layer (*c.f.* Section 2).

Table 1: Model Parameter Analysis of the proposed MSFSeg. “kernel size” indicate the hyperparameter of the 3D convolutional layer in MSF module.

	w/o MSF	w. MSF – kernel size=(1,1,1)	w. MSF – kernel size=(3,1,1)
# of Param.	58,771,354	58,771,359	58,771,367

Mask Decoding. The final decoder architecture follows [9,15], with skip-connection operations same as U-Net [11].

1.2 Training and Hyper-parameters

We use the SGD optimizer with a learning rate of $5e-4$, momentum of 0.9, and weight decay of $1e^{-4}$. Cross-entropy loss calculated between predicted and ground-truth masks is used for MSFSeg network training. We train our MSFSeg first with a pretrained ResNet-101 on COCO [8] images, and then on Abdomen-CT [6] and CHAOS-MRI [5] respectively, following the strategy as [2,10], for all experimental results presented in Table 1-3.

Training of MSFSeg took 12-16 hours on a local server equipped with two NVIDIA A100 graphic cards and a multi-thread AMD EPYC 74F3 24-Core CPU.

2 Model Parameter Analysis

Our proposed MSF module only introduces additional trainable parameters in the 3D convolutional layer (*c.f.* Figure 2), in Table 1 we present the number of model parameters for our proposed MSFSeg network with MSF module using different kernel sizes and without MSF module respectively. In our experiments, we set “kernel size” in the 3D convolutional layers to $1 \times 1 \times 1$ and $3 \times 1 \times 1$ respectively, and empirically found they result in comparable FSS performance. Thus we report results of our MSFSeg with “kernel size= $1 \times 1 \times 1$ ” in all our experiments. From Table 1, we could see that the MSF only introduces $< 0.00001\%$ additional parameters compared to original FSS pipeline based on ResNet-101, further proves its lightweight property.

3 Error Bars for Main Experiments

As mentioned in Experiments section of the main paper, we report the mean values out of five runs for the results presented in the tables there. Here in Table 2, we additionally provide the standard deviation of those results.

Table 2: Mean and standard deviation (SD) of the results presented in Table 1 of the main paper. The values are formatted as “mean \pm (SD)”.

	Methods	Abdomen-CT [6]				CHAOS-MRI [5]			
		LK	RK	Spleen	Liver	LK	RK	Spleen	Liver
Setting 1	Ours – 1-shot	81.11 \pm (0.14)	78.41 \pm (0.15)	73.64 \pm (0.21)	78.91 \pm (0.05)	84.18 \pm (0.33)	88.10 \pm (0.13)	77.12 \pm (0.15)	76.11 \pm (0.12)
	Ours – 5-shot	87.22 \pm (0.03)	85.62 \pm (0.05)	82.71 \pm (0.27)	82.57 \pm (0.01)	88.63 \pm (0.17)	90.94 \pm (0.04)	82.73 \pm (0.28)	82.10 \pm (0.11)
Setting 2	Ours – 1-shot	79.24 \pm (0.25)	77.36 \pm (0.16)	75.21 \pm (0.21)	76.73 \pm (0.05)	82.83 \pm (0.26)	86.98 \pm (0.10)	78.07 \pm (0.07)	76.14 \pm (0.23)
	Ours – 5-shot	85.73 \pm (0.11)	84.51 \pm (0.04)	81.60 \pm (0.20)	81.22 \pm (0.04)	87.70 \pm (0.12)	90.62 \pm (0.09)	81.97 \pm (0.12)	82.52 \pm (0.04)

4 More Experimental Results

4.1 Compared with Other Weakly-Supervised 3D Segmentation Methods

In Table 3, we present more quantitative evaluations of our proposed MSFSeg with other state-of-the-art few-shot segmentation (FSS) and weakly-supervised 3D segmentation methods (experimental setting same as Table 1 in our main paper), on small organs including esophagus, and left adrenal gland from Abdomen-CT dataset [6]. Compared to other FSS methods (in teal color) with 2D segmentation network training, our proposed MSFSeg achieves $>+8.7\%$ mIoU in 1-shot setting. Our MSFSeg – 5-shot achieves competitive results compared to state-of-the-art weakly-supervised 3D segmentation method, PRNet [7] requiring heavy 3D segmentation network training and full 3D supervision, with much cheaper support labels (5 2D slices vs. 3D scribble on 1 data volume which may contain hundreds of slices) and lighter network architecture.

Table 3: Comparison to other weakly-supervised few-shot segmentation methods on Abdomen-CT [6]. All methods do not need 3D annotation input/supervision and 3D network training are in teal color.

Methods	Esophagus	Left AG	mIoU
DataAug – 1-shot [14]	11.9	0.9	6.4
SE-Net – 1-shot [12]	8.7	1.2	5.0
SSL-ALPNet – 1-shot [10]	14.1	5.8	10.0
vSSL-ADNet – 1-shot [3]	12.6	3.6	8.1
PRNet* [7] (scribble on 1 vol.)	38.1	24.6	31.4
Ours – 1-shot (1 slice)	27.6	9.8	18.7
Ours – 5-shot (5 slices)	32.9	24.6	28.8

References

1. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database
2. Ding, H., Sun, C., Tang, H., Cai, D., Yan, Y.: Few-shot medical image segmentation with cycle-resemblance attention. In: WACV. pp. 2487–2496 (2023)
3. Hansen, S., Gautam, S., Jenssen, R., Kampffmeyer, M.: Anomaly detection-inspired few-shot medical image segmentation through self-supervision with supervoxels. Medical Image Analysis **78** (2022)

4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
5. Kavur, A.E., Gezer, N.S., Barış, M., et al.: CHAOS Challenge - combined (CT-MR) healthy abdominal organ segmentation. *Medical Image Analysis* (2021), <http://www.sciencedirect.com/science/article/pii/S1361841520303145>
6. Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A.: Miccai multi-atlas labeling beyond the cranial vault-workshop and challenge (2015). DOI: <https://doi.org/10.7303/syn3193805> (2015)
7. Lei, W., et al.: One-shot weakly-supervised segmentation in 3d medical images. *IEEE Transactions on Medical Imaging* (2023)
8. Lin, T., Maire, M., Belongie, S.J., et al.: Microsoft COCO: common objects in context. *CoRR* (2014)
9. Min, J., Kang, D., Cho, M.: Hypercorrelation squeeze for few-shot segmentation. In: ICCV (2021)
10. Ouyang, C., Biffi, C., Chen, C., Kart, T., Qiu, H., Rueckert, D.: Self-supervision with superpixels: Training few-shot medical image segmentation without annotation. In: ECCV (2020)
11. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015)
12. Roy, A.G., Siddiqui, S., Pölsterl, S., Navab, N., Wachinger, C.: ‘squeeze & excite’ guided few-shot segmentation of volumetric images. *Medical image analysis* **59**, 101587 (2020)
13. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. In: NeurIPS (2017)
14. Zhao, A., Balakrishnan, G., Durand, F., Guttag, J.V., Dalca, A.V.: Data augmentation using learned transformations for one-shot medical image segmentation. In: CVPR (2019)
15. Zhao, H., Zhang, Y., Liu, S., Shi, J., Loy, C.C., Lin, D., Jia, J.: PSANet: Point-wise spatial attention network for scene parsing. In: ECCV (2018)