Supplementary Materials Primitive Geometry Segment Pre-training for 3D Medical Image Segmentation

Ryu Tadokoro*[1,2]
tadokororyuryu@gmail.com
Ryosuke Yamada*[1,3]
ryosuke.yamada@aist.go.jp
Kodai Nakashima[1,3]
nakashima.kodai@aist.go.jp
Ryo Nakamura[1,4]
ryo.nakamura@aist.go.jp
Hirokatsu Kataoka[1]
hirokatsu.kataoka@aist.go.jp

[1] National Institute of Advanced Industrial Science and Technology, Japan
[2] Tohoku University, Japan
[3] University of Tsukuba, Japan
[4] Fukuoka University, Japan

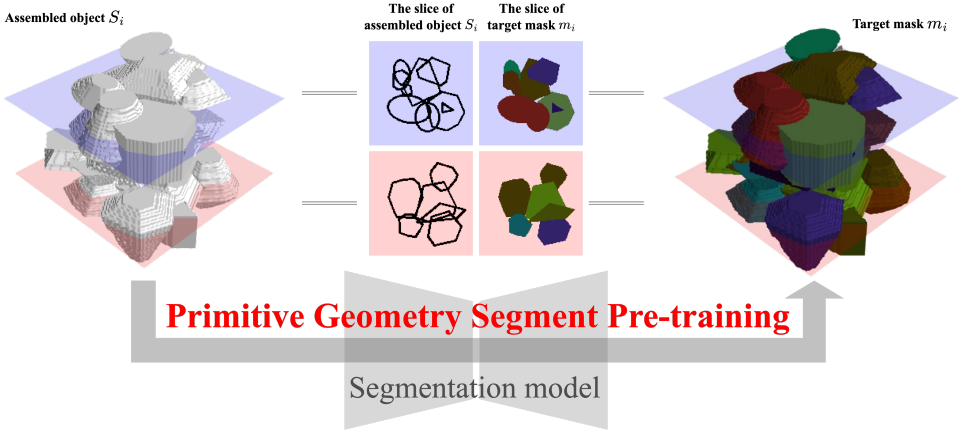

Figure 1: **Visualization of Inputs and Outputs in PrimGeoSeg.** We perform pre-training of the 3D segmentation model with S_i as the input and m_i as the output. Slices of S_i and m_i are shown in the center of the top row in the figure.

The present paper offers an enriched and expanded version of [8], incorporating a more in-depth analysis, essential additional experiments, and comprehensive details. In fundamental experiments, we analyze various elements in PrimGeoSeg and what aspects are effective in 3D medical image segmentation. We also demonstrated the effectiveness and properties of

* These authors contributed equally to this work.

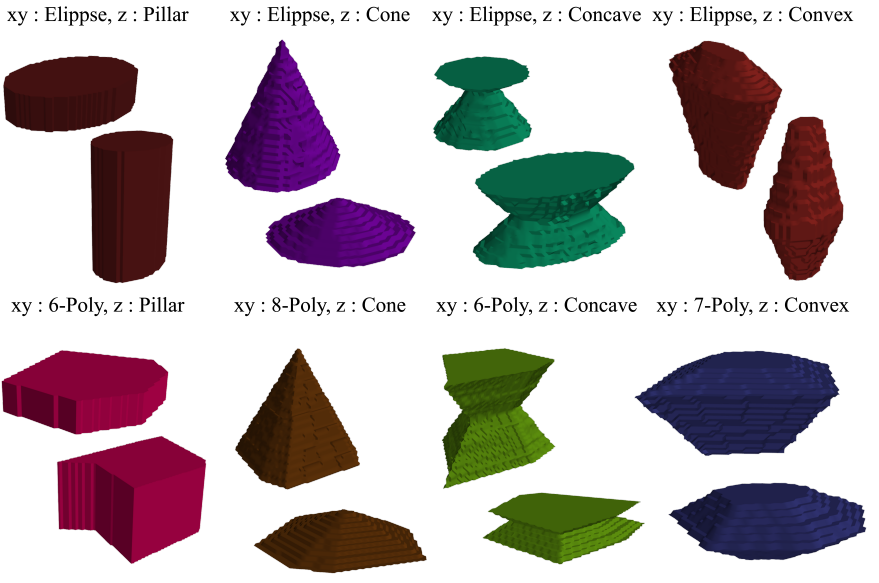

Figure 2: **Visualization of Primitive Objects.** Primitive objects are positioned on the assembled object S_i. The shape class is uniquely determined by the xy-plane and z-axis rules. Even within the same class, diversity in shapes is achieved through instance augmentation.

shape pre-training in qualitative results, application to a broader range of datasets, and verification of the effect of pre-training on limited data. In this supplementary material, we also delve into further details that were not feasible to encompass within the main manuscript, including an extensive ablation study. In Section A, we supplementally provide more detailed information on our proposed PrimGeoSeg. In Section B, we introduce the more detailed experimental settings and the the benchmark dataset on 3D medical image segmentation. In Section C, we present experimental results which were omitted from the main study due to space constraints.

A Visualization and Details for PrimGeoSeg

A.1 Input and Output 3D Volumetric Data in PrimGeoSeg

For PrimGeoSeg, we employ the contour components of primitive objects arranged in 3D space as input for the segmentation task. The regions filled with primitive objects are target masks. Figure 2 visualizes the input volumetric image, denoted as S_i, and the corresponding target mask, m_i. Presented in the center top of figure 2 are the slices of S_i and m_i. This visualization confirms that S_i denotes the contour of the primitive objects and m_i functions as a mask, filling the interior of these objects. Here, we arrange primitive objects in 3D space according to their volume, from largest to smallest. The design of the masks allows subsequent Primitive Objects with smaller volumes to overwrite them.

Table 1: Parameters used in the generation of PrimGeoSeg.

Parameter list for the generation of PrimGeoSeg
Parameters for primitive objects generation.
xy-class: {'ellipse', '3-poly', '4-poly', '5-poly', '6-poly', '7-poly', '8-poly', '9-poly'}
z-class: {'concave', 'convex', 'pillar', 'cone'}
$z_{\max} \sim \mathscr{U}(10, 50)$, $z_c \sim \mathscr{U}(3, z_{\max} - 3)$ # For determine the size and the shape along with z-axis.
$o1, o2, o3$ # Determined by the z-class. Due to the space limitation, please see the main text.
$R_{\min} = 15$, $R_{\max} \sim \mathscr{U}(30, 80)$ # For determine the basic size of xy-slice.
Parameters for arrangement of primitive objects.
$(W, H, D) = (96, 96, 96)$ # The size of assembled object S_i and the corresponding mask m_i.
$M = 20$ # The number of shapes intended to be placed on assembled object S_i and the mask m_i.
$Max\_iter = 100$ # The maximum number of times to randomly position search of primitive objects.
$r = 0.7$ # The maximum allowable overlap rate between shapes when placing primitive objects.
$Intensity = 128$ # Each pixel value in S_i is fixed to 128
※ $\sim \mathscr{U}(q, r)$ denotes random selection from a discrete uniform distribution within the range of (q, r)

A.2 Examples of Primitive objects

Primitive objects are formulated by integrating rules applicable to the *xy*-plane and the *z*-axis. With eight rules for the *xy*-plane and four for the *z*-axis, we define a total of 32 shape classes. Examples of these Primitive Objects are illustrated in Figure 1. As can be seen, each unique combination of a *xy*-plane rule and a z-axis rule leads to the specification of a distinct shape class. Even within the same class, objects may have different shapes. This variability results from the random parameters applied during their generation called instance augmentation. It's important to note that internal human anatomy exhibits individual variation, meaning that even the same organ can have different shapes across individuals.

A.3 Parameters for the generation process of PrimGeoSeg

We show the parameters in the whole pipeline of the PrimGeoSeg construction in Table 1. These parameter values were predetermined to fulfill the motivations outlined in Sec. 3.2 of our paper; no parameter tuning was performed.

B Experimental Details

B.1 Datasets for 3D Medical Image Segmentation

B.1.1 BTCV

In our experiments, we use the Multi-Atlas Labeling Beyond the Cranial Vault (BTCV) dataset [3], designed for 3D medical image segmentation focusing on human visceral organs. The BTCV dataset consists of abdominal CT images from 30 subjects, each meticulously annotated by experts to pinpoint 13 major internal organs. Notably, annotations include spleen (Spl), right kidney (RKid), left kidney (LKid), gallbladder (Gall), esophagus (Eso), liver (Liv), stomach (Sto), aorta (Aor), inferior vena cava (IVC), portal vein (Veins), splenic vein (Veins), pancreas (Pan), right adrenal gland (rad), and left adrenal gland (lad). Every

Table 2: **Hyperparameters for Each Benchmark Dataset.**

Dataset	BTCV		MSD		BraTS	
Architecture	UNETR	SwinUNETR	UNETR	SwinUNETR	UNETR	SwinUNETR
Optimizer	AdamW					
Scheduler	Warmup cosine scheduler					
Input size	$96 \times 96 \times 96$				$128 \times 128 \times 128$	
Batch size	6	8	6	8	8	8
Learning rate	0.0001				0.0008	
Iteration	$20K$	$15K$	$20K$	$15K$	$125K$	

3D volumetric image, with slice sizes of 512×512 and slice direction, sizes approximately from 100 to 200, was resampled into voxels of dimensions $1.5mm \times 1.5mm \times 2.0mm$. We then adjusted each pixel value within the soft tissue window and normalized them to fall within a $[0,1]$ range. We split the BTCV dataset into train and test set following an 80:20 ratio, using the test set for offline evaluation. Note that the train and test splits we used are the same as those in the SSL methods we compared with [2, 9].

B.1.2 MSD

The Medical Segmentation Decathlon (MSD) dataset [2] is a comprehensive dataset designed for segmentation tasks across ten different types of tumors and internal organs. However, our study focuses on Task06 (Lung) and Task09 (Spleen), given the significant computational cost demand of analyzing all tasks and the smaller data size of these two tasks for testing data-efficient learning approaches. Task06 involves performing lung tumor segmentation from 3D volumetric images captured via CT scans, while Task09 requires spleen segmentation from similar 3D CT scan images. We resampled all 3D volumetric images into isotropic voxels of $1.0mm$. We split the MSD dataset into train and test set following an 80:20 ratio, using the test set for offline evaluation. The results presented in [8] are given in terms of average dice score for both background and target object. Note that this study reports the Dice score for the target object only.

B.1.3 BraTS

The Multi-modal Brain Tumor Segmentation Challenge (BraTS) dataset [1] targets identifying Glioblastoma tumor areas captured through MRI images. Four distinct types of multimodal MRI images—T1-weighted imaging, T1-weighted imaging with contrast enhancement, T2-weighted imaging, and FLAIR imaging—are merged along the channel direction for input to increase the precision in tumor detection. The BraTS dataset has annotations for three specific regions: WholeTumor (WT), TumorCore (TC), and EnhancingTumor (ET). During training and evaluation, segmentation should be performed for the above three tumor regions. We resampled all 3D volumetric images to isotropic voxels of $1.0mm$. Furthermore, we normalize pixel values to achieve a distribution with a mean of 0 and a standard deviation of 1 using non-zero pixel values. We split the BraTS dataset into train and test set following an 80:20 ratio, using the test set for offline evaluation.

B.2 Implementation Details

Table 3: Comparison of performance in BTCV.

Pre-training	PT Num	Type	Avg.	Spl	RKid	LKid	Gall	Eso	Liv	Sto	Aor	IVC	Veins	Pan	rad	lad
Swin-based model																
Scratch	0	–	79.5	**95.2**	94.3	94.1	43.3	74.2	**96.7**	78.9	**90.2**	83.5	73.1	77.7	67.5	**65.5**
Jiang *et al.* [5]	3.6K	SSL	**81.1**	93.4	94.1	94.0	**58.5**	73.7	96.3	**81.6**	89.3	85.9	74.7	78.5	**68.8**	65.0
PrimGeoSeg	5K	FDSL	80.4	95.0	**94.5**	**94.4**	50.1	**74.4**	96.6	81.1	89.2	**86.1**	**75.9**	**82.1**	67.8	58.2
MiT																
Scratch	0	–	78.8	93.3	93.9	93.7	62.2	70.7	96.3	77.3	86.5	80.5	72.0	73.1	64.2	61.1
Xie *et al.* [10]	5K+α	SSL	79.7	94.9	93.8	94.0	61.6	69.7	96.3	82.1	87.8	81.8	72.4	75.9	**66.0**	60.3
PrimGeoSeg	5K	FDSL	**82.0**	**95.4**	**94.2**	**94.2**	**63.6**	**75.5**	**96.5**	**85.7**	**88.9**	**85.4**	**74.7**	**80.3**	**66.9**	**64.9**

Table 4: Comparison of performance in MSD.

Pre-training	Type	Swin-based model		MiT	
		Lung	Spleen	Lung	Spleen
Scratch	–	67.4	96.5	58.6	95.8
Jiang *et al.* [5]	SSL	71.7	96.8	–	–
Xie *et al.* [10]	SSL	–	–	70.8	95.3
PrimGeoSeg	FDSL	**73.6**	**96.8**	**70.8**	**95.9**

In Section 4.2 of the main study, we conducted five foundational experiments, denoted as (a) through (e). In each experiment, we modified the generation of pre-training data for PrimGeoSeg. In experiment (a), as shown in Figure 3 of the main study, we defined planar shapes as shapes where each primitive object was positioned with a thickness of 1 in the z-axis direction. For volumetric shapes, in comparison with planar shapes, the class in the z-axis direction was fixed to "Cone", ensuring that there were no overlaps when the shapes were positioned. In experiment (b), we altered the number of classes for the shapes in PrimGeoSeg. When the class count was set to 1 in the xy plane, we chose "ellipse". In the z-axis direction, "cone" was selected for the experiment. In experiment (c), we compared the pre-training performance based on the presence or absence of instance augmentation for primitive objects. When instance augmentation was disabled, the values for parameters z_{max}, z_c, $o1$, $o2$, $o3$, and R_{max} presented in Table 1 - were fixed for the experiment.

For downstream tasks of all experiments, we fine-tuned our model using hyper-parameters specifically tailored for BTCV, MSD, and BraTS, as detailed in Table 2. Across all experiments, we employ a patch-based approach for both the learning and inference phases. During training, patches of a pre-determined size are randomly cropped from the input images and are then used to train the model. As for the inference stage, we utilize a sliding window technique, with a window overlap of 0.5, to ensure comprehensive coverage.

C Additional Experiments

C.1 Comparison with Other Self-Supervised Learning Methods

The main paper compares SSL techniques using the UNETR and SwinUNETR architectures [4, 9]. In this section, we examine the performance of PrimGeoSeg on architectures other than UNETR and SwinUNETR to confirm its generality and effectiveness. We also compared PrimGeoSeg with other self-supervised learning (SSL) methods, namely SMIT [5] and UniMiSS [10], using the same architectures for each method.

SMIT [5] employs a Teacher-Student Network with Exponential Moving Averaging for

Table 5: Comparison of Normalized Surface Distance.

Pre-training	Type	UNETR			SwinUNETR		
		BTCV	MSD (Lung)	MSD (Spleen)	BTCV	MSD (Lung)	MSD (Spleen)
Scratch	–	0.715	0.511	0.899	0.766	0.673	0.953
Tang *et al.* [9]	SSL	–	–	–	0.829	0.686	0.960
PrimGeoSeg	FDSL	**0.822**	**0.649**	**0.953**	**0.839**	**0.723**	**0.967**

Table 6: The effects of intensity value of PrimGeoSeg.

Intensity value	BTCV
Pixel value = 128	**81.95**
Pixel value range [78, 178]	81.56

Self-Distillation, optimizing a pseudo-task through Masked Image Modeling. For self-supervised pre-training, SMIT uses 3,643 3D CT scans and adopts a 3D segmentation model with the Swin-transformer as its backbone, referred to as the Swin-based model. On the other hand, UniMiSS [10] captures multi-modal representations by self-supervised learning on both 3D CT scans and 2D X-ray images simultaneously. UniMiSS utilizes 5,022 3D CT scans and 108,948 2D X-ray images for self-supervised pre-training, incorporating a uniquely designed pyramid U-like medical Transformer (MiT) for its architecture. For simplicity, we trained the Swin-based model and MiT using settings similar to UNETR's.

Table 3 and Table 4 presents the accuracy comparison when using the Swin-based model and MiT as architectures, with the Dice Score as the evaluation metric. Using the Swin-based model, our proposed method, PrimGeoSeg, showed an improvement of 0.9 points in BTCV, 6.2 points in MSD (Lung), and 0.3 points in MSD (Spleen) compared to training from scratch. With the MiT, the improvements were 3.2 points in BTCV, 12.2 points in MSD (Lung), and 0.1 points in MSD (Spleen).

These results suggest that PrimGeoSeg is effective for specific models such as SwinUNETR or UNETR and offers generalizable benefits across other models. Compared to the SSL technique SMIT, there was a decrease of 0.7 points in the Average Dice Score for BTCV; however, it performed equivalently or better in half of the classes. In MSD (Lung), PrimGeoSeg outperformed SMIT by 1.9 points. Against UniMiSS, PrimGeoSeg showed an improvement of 2.3 points. In the MSD metric, PrimGeoSeg achieved performance on par with UniMiSS. These findings further reinforce the efficacy of our proposed PrimGeoSeg.

C.2 Evaluation using Other Metric

In the main paper, we primarily employed the Dice Score, a prevalent metric in 3D medical image segmentation, for our evaluations. In this section, we also evaluate PrimGeoSeg using the Normalized Surface Distance (NSD) [7] to provide a more comprehensive assessment. NSD serves as a metric to evaluate the congruence between predicted and ground truth segmentation boundaries, quantifying deviations and providing insights into the precision of boundary delineation. It assigns scores ranging from 0 to 1, where a score closer to 1 indicates a higher congruence between the predicted and actual boundaries. We factored the tolerance threshold of 1mm for each class for NSD. The results of our NSD evaluations, presented in Table 5, mirrored the trends observed with the Dice Score for both UNETR and SwinUNETR architectures. The enhanced performance evident in both NSD and Dice Score evaluations underscores the effectiveness of our proposed method, PrimGeoSeg.

C.3 The Effects of Intensity value of PrimGeoSeg

In the main study, we set the intensity value of the assembled object S_i in PrimGeoSeg to a fixed value of 128. This setting is based on existing findings from shape pre-training research [8], which indicated that fixing the intensity value led to better pre-training performance. In this section, we examined the pre-training effect when the intensity value was changed from a fixed value to a random value in the range [78,178], using the SwinUNETR architecture and the BTCV dataset. As seen in Table 6, it is clear that fixing the intensity of the assembled object S_i results in better performance. We believe that fixing the intensity makes the model less focused on texture and more attuned to the shape. Such a focus on shape is crucial for effective shape pre-training.

References

[1] Brain tumor segmentation challenge 2021. Available at: https://www.synapse.org/#!Synapse:syn25829067/wiki/612712.

[2] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022.

[3] J Igelsias M Styner T Langerak B Landman, Z Xu and A Klein. Miccai multi-atlas labeling beyond the cranial vault– workshop and challenge. 2015.

[4] Zekai Chen, Devansh Agarwal, Kshitij Aggarwal, Wiem Safta, Mariann Micsinai Balan, and Kevin Brown. Masked image modeling advances 3d medical image analysis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1970–1980, 2023.

[5] J. Jiang, N. Tyagi, K. Tringale, C. Crane, and H. Veeraraghavan. Self-supervised 3d anatomy segmentation using self-distilled masked image transformer (smit). In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 556–566, Cham, September 2022. Springer Nature Switzerland.

[6] Hirokatsu Kataoka, Kazushige Okayasu, Asato Matsumoto, Eisuke Yamagata, Ryosuke Yamada, Nakamasa Inoue, Akio Nakamura, and Yutaka Satoh. Pre-training without natural images. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[7] Stanislav Nikolov, Sam Blackwell, Alexei Zverovitch, Ruheena Mendes, Michelle Livne, Jeffrey De Fauw, Yojan Patel, Clemens Meyer, Harry Askham, Bernardino Romera-Paredes, et al. Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. *arXiv preprint arXiv:1809.04430*, 2018.

[8] Ryu Tadokoro, Ryosuke Yamada, and Hirokatsu Kataoka. Pre-training auto-generated volumetric shapes for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4739–4744, June 2023.

[9] Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20730–20740, 2022.

[10] Y. Xie, J. Zhang, Y. Xia, and Q. Wu. Unimiss: Universal medical self-supervised learning via breaking dimensionality barrier. In *European Conference on Computer Vision*, pages 558–575, Cham, October 2022. Springer Nature Switzerland.