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

Abstract

The construction of 3D medical image datasets presents several issues, including requiring significant financial costs in data collection and specialized expertise for annotation, as well as strict privacy concerns for patient confidentiality compared to natural image datasets. Therefore, it has become a pressing issue in 3D medical image segmentation to enable data-efficient learning with limited 3D medical data and supervision. A promising approach is pre-training, but improving its performance in 3D medical image segmentation is difficult due to the small size of existing 3D medical image datasets. We thus present the Primitive Geometry Segment Pre-training (PrimGeoSeg) method to enable the learning of 3D semantic features by pre-training segmentation tasks using only primitive geometric objects for 3D medical image segmentation. PrimGeoSeg performs more accurate and efficient 3D medical image segmentation without manual data collection and annotation. Further, experimental results show that PrimGeoSeg on SwinUNETR improves performance over learning from scratch on BTCV, MSD (Task06), and BraTS datasets by 3.7%, 4.4%, and 0.3%, respectively. Remarkably, the performance was equal to or better than state-of-the-art self-supervised learning despite the equal number of pre-training data. From experimental results, we conclude that effective pre-training can be achieved by looking at primitive geometric objects only. Code and dataset are available at https://github.com/SUPER-TADORY/PrimGeoSeg.

1 Introduction

3D medical image analysis using deep learning is expected to enhance diagnostics and improve patient outcomes through the more accurate detection and visualization of geometric structures inside the human body. For example, 3D medical image segmentation estimates

* These authors contributed equally to this work.

the location and category of human organs from computed tomography (CT) and magnetic resonance imaging (MRI) images. More accurate segmentation of 3D medical images requires a large amount of training data and rich semantic annotation. However, training data collection is difficult due to the high imaging costs and stringent privacy protections. In addition, the annotation process requires expert knowledge of medical science.

In order to solve the above problems, there have been many studies in terms of pre-training methods toward more data-efficient learning under limited training data conditions. In particular, self-supervised learning (SSL) has emerged as a promising approach for pre-training in 3D medical image segmentation [4, 6, 7, 9, 13, 21, 24, 25, 26, 28, 29, 30, 31, 32], as it learns 3D structural features and reduces manual annotation costs by designing and learning a pre-text task on unsupervised data. Chen *et al.* [6] achieved state-of-the-art performance on the Multi-Atlas Labeling Beyond the Cranial Vault (BTCV) dataset [2] and Medical Segmentation Decathlon (MSD) [1] dataset by merging existing 3D medical image datasets and pre-training three pseudo tasks. Nevertheless, pre-training methods for 3D medical image segmentation have lagged compared with other 3D object recognition tasks because of the small scale of pre-training datasets. Therefore, an alternative pre-training approach is needed to address dataset construction issues in 3D medical image segmentation.

Formula-driven supervised learning (FDSL) [14, 15] has been proposed as a synthetic pre-training method without real data and human annotations, which automatically generates synthetic data and supervised labels based on a specific principle rule of the real world. Therefore, a significant advantage of FDSL is that the properties of the synthetic data can be designed considering fine-tuning tasks different from real data. Furthermore, as much as possible, FDSL can reduce dataset issues related to real data, such as social bias and personal information protection. Recently, Nakashima *et al.* [20] reported that Vision Transformer (ViT) tends to focus on the outlines of an object on images in pre-training. Inspired by the above insight, Kataoka *et al.* [16] proposed a Radial Counter DataBase (RCDB) that has improved the complexity of outlines and pre-training performance. Furthermore, Yamada *et al.* [27] proposed a Point Cloud Fractal DataBase (PC-FractalDB) based on fractal geometry to improve performance by designing 3D object detection pre-training. From these insights, we hypothesize that we can design segmentation tasks using only primitive geometric objects to achieve an effective pre-training method for 3D medical segmentation.

The present study proposes a primitive geometry segmentation (PrimGeoSeg) for 3D medical image segmentation by automatically generating pre-training data and expressing semantically supervised labels as an assembly of primitive geometric objects in 3D space, as shown in Figure 1. We generate a primitive geometric object from independent laws in the xy-plane and z-axis directions. We also construct a pre-training dataset by arranging multiple primitive geometric objects in 3D space, overlapping each object. We designed this generation process to consider two aspects of the internal structure of the human body; (i) the variability among individuals and (ii) the com-

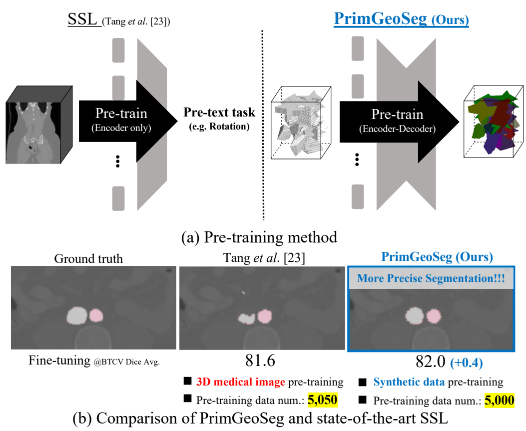

Figure 1: The overview of PrimGeoSeg.

plexity with ambiguous boundaries between organs. The experimental results found that primitive geometric objects only are sufficient to learn the necessary 3D structural representations for achieving a superior pre-training effect for 3D medical image segmentation.

The contributions of this work are as follows: (i) We propose PrimGeoSeg as a pre-training method that enables pre-training by segmentation without real data collection and manual annotation. (ii) We show that pre-training both UNETR and SwinUNETR with PrimGroSeg outperform state-of-the-art SSL accuracy on BTCV and MSD in 3D medical image segmentation. Notably, the number of synthetic pre-training data was almost equal (see Figure 1). In addition, PrimGeoSeg also demonstrates remarkable data efficiency, performing as well with only 30% of the BTCV data as it does when learning from scratch with 100% of training data. (iii) Our proposed method for pre-training synthetic data can reduce problems such as the privacy of 3D medical images.

2 Related Works

Pre-training for 3D medical image segmentation. In 3D medical image segmentation, SSL has been attracting attention for its ability to achieve highly accurate segmentation results by pre-training unsupervised 3D medical images [4, 5, 6, 7, 9, 13, 24, 25, 26, 28, 29, 30, 31, 32]. Even in transformer-based models that achieve higher accuracy than conventional CNN-based models for 3D medical images [10, 11], SSL has shown substantial accuracy improvements. Chen *et al.* [6] improved performance on UNETR through pre-training via masked image modeling, which masks a portion of 3D medical images. In addition, Tang *et al.* [25] achieved state-of-the-art results using the SwinUNETR [10] on BTCV [2] and MSD [1] datasets by pre-training three pre-text tasks including image inpainting, 3D rotation prediction, and contrastive learning. As shown above, SSL can improve the performance of 3D image medical segmentation. However, SSL improvements may be limited by training data available, as SSL performance is often dependent on the amount of training data. We thus believe that the performance of the pre-training of 3D medical image segmentation will be further improved by solving the dataset construction issues.

Formula-driven supervised learning (FDSL). Recently, large-scale pre-training has made tremendous developments in computer vision [3, 8, 18], and among its methods, FDSL can perform large-scale pre-training without real data and manual annotation [12, 14, 15, 16, 17, 20, 23]. Specifically, pre-training data and its label are automatically generated from mathematical formulations based on real-world principles, such as fractal geometry and Perlin noise. Kataoka *et al.* [16], proposed RCDB inspired that ViT pays attention to the outer contours of the fractal region when pre-training with the Fractal Database (FractalDB). RCDB pre-trained model surpasses the ImageNet pre-trained model on ViT despite not learning natural images. More recently, Yamada *et al.* [27] proposed the PC-FractalDB for pre-training in 3D object detection using 3D point clouds. They concluded that one factor for success in pre-training is initializing not only the backbone network but also the entire model.

Based on these findings, we hypothesize that synthetic pre-training through the same segmentation task, similar to the fine-tuning task, will have a greater effectiveness in 3D medical image segmentation using the transformer-based model. Furthermore, we think that learning from synthetic pre-training data, rather than from 3D medical images, can effectively address several issues commonly associated with 3D medical data usage. These issues include societal bias, privacy concerns, and copyright infringement.

This paper notably extends the experiments in [22] and provides new contributions of our proposed method.

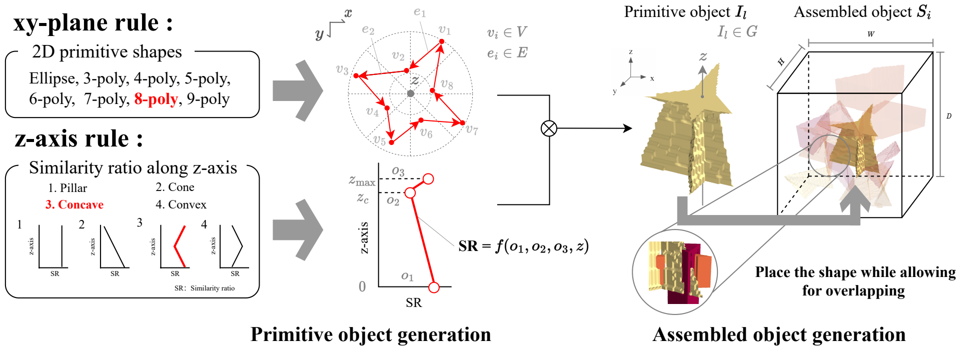

Figure 2: **The generation process of assembled objects as pre-training data for PrimGeoSeg.** We generate an assembled object by arranging randomly multiple primitive objects generated from the individual xy-plane and z-axis rules.

3 PrimGeoSeg: Primitive Geometry Segment Pre-training

In this section, we introduce PrimGeoSeg method, which is the pre-training strategy of generating primitive geometric objects and performing segment pre-training for downstream tasks in 3D medical image segmentation. We generate an assembled object as pre-training data for PrimGeoSeg based on the design concept of the property of 3D medical images by (i) the variability among individuals and (ii) the complexity with ambiguous boundaries between organs. The pre-training dataset consisting of assembled objects and supervised labels denoted by $\mathcal{D} = \{(S_i, m_i)\}_{i=1}^{N}$, where $S_i \in \mathbb{R}^{W \times H \times D}$ is an assembled object, $m_i \in \mathbb{L}^{W \times H \times D}$ is a corresponding segmentation mask, and N is the number of pre-training data for PrimGeoSeg. $\mathbb{L}$ is a set of integers denoting the segmentation label.

As shown in Figure 2, the generation procedure of an assembled object is composed of two steps: (i) primitive object generation and (ii) arrangement of primitive objects. (i) First, we set a class of each primitive object based on xy-plane and z-axis rules. Moreover, we generate a primitive object based on randomly determined parameters regarding the number of vertices in the xy-plane and the similarity ratio along the z-axis. (ii) Second, in the arrangement of primitive objects, we generate an assembled object S_i and its corresponding segmentation mask m_i by arranging multiple primitive objects in 3D space. Finally, we repeat (i) – (ii) steps N times to automatically construct pre-training dataset $\mathcal{D}$ for PrimGeoSeg.

3.1 Pre-training Data Generation

Primitive object generation. Each primitive object is generated by stacking xy-plane slices, with the similarity ratio of each slice varying along the z-axis. The generation process of a primitive object is based on two rules: the xy-plane rule, which dictates the shape of the slices, and the z-axis rule, which controls the changing rate of the similarity ratio along the z-axis. We define a class of a primitive object considering combing the xy-plane and z-axis rules. We set the maximum 32 classes consisting of eight classes in the xy-plane rule and four classes in the z-axis rule (see Figure 2). The size of the primitive object along the z-axis, denoted as z_{max}, is randomly selected from a uniform distribution, $z_{max} \sim \mathcal{U}(10, 50)$. For each t in the range $0 \leq t \leq z_{max}$, a slice P_t is generated. The similarity ratio of the slice at $z = t$ is determined by a function $f(z = t)$ according to the z-axis rule. The z-axis rule consists of four classes: {'concave', 'convex', 'pillar', 'cone'} in this paper. We define q_z by randomly selecting from the above four classes. Here, the function $f(z)$ represents the similarity ratio of the slices along with the z-axis direction as shown below;

$$f(o_1,o_2,o_3,z) = \begin{cases} o_1 + (o_2 - o_1)\frac{z}{z_c} & (0 \leq z \leq z_c) \\ o_2 + (o_3 - o_2)\frac{z - z_c}{z_{\max} - z_c} & (z_c < z \leq z_{\max}) \end{cases} \tag{1}$$

where $z_c \sim \mathcal{U}(3, z_{max} - 3)$ and $o1$, $o2$, and $o3$ are defined as certain values when q_z was selected. For instance, in the 'pillar' class, all parameters are set to 1: $o1 = o2 = o3 = 1$. In the 'cone' class, we use $o1 = 0$, $o2 = \frac{z_0}{z_{max}}$, and $o3 = 1$. For 'concave', $o1, o3 \sim \mathcal{U}(0.8, 1)$, and $o2 \sim \mathcal{U}(0.2, 0.5)$. For 'convex', $o1, o3 \sim \mathcal{U}(0.2, 0.5)$, and $o2 \sim \mathcal{U}(0.8, 1)$. Choosing a z-axis rule determines the values of $o1$, $o2$, and $o3$, defining the unique function $f(z)$.

We generate the slice P_t using both $f(z)$ and the xy-plane rule. The xy-plane rule is defined by a set of shape definitions: {'ellipse', '3-poly', '4-poly', '5-poly', '6-poly', '7-poly', '8-poly', '9-poly'}, where w-poly represents a w-sided polygon. One shape definition q_{xy} is selected from the above eight rules. To define the size of the slice P_t, we set parameters $R_{min} = 15$ and $R_{max} \sim \mathcal{U}(30, 80)$. If q_{xy} is defined as a polygon, the slice P_t forms a closed shape bounded by edges in the set $E(t)$:

$$V = \left\{ (r_k \cos\theta_k, r_k \sin\theta_k) \mid 1 \leq k \leq C_{xy} \right\} \tag{2}$$

$$E(z=t) = \left\{ f(t)\left(v_k + s(v_{(k+1) \bmod C_{xy}} - v_k) \right) \mid v_k \in V, 1 \leq k \leq C_{xy}, 0 \leq s \leq 1 \right\} \tag{3}$$

where $r_k \sim U(R_{\min}, R_{\max}), \theta_k \sim U\left(\frac{2k}{C_{xy}}\pi, \frac{2(k+1)}{C_{xy}}\pi\right)$ is polar coordinates and C_{xy} is the number of vertices. We have tolerated the alignment of three adjacent vertices in a straight line as acceptable noise. If K_{xy} is defined as 'ellipse", P_t is a closed shape that satisfies the equation $\frac{x^2}{(af(t))^2} + \frac{y^2}{(bf(t))^2} = 1$, where $a, b \sim \mathcal{U}(R_{\min}, R_{\max})$ is the major and minor axes of the ellipse. By integrating the slices P_t from $z = 0$ to z_{max} along the z-axis, we create a single primitive object I_l. By repeating M times, we generate a set of primitive objects $G = \{I_l\}_{l=1}^{M}$. Here, M is the number of primitive objects to be positioned in assembled object S_i.

Arrangement of primitive objects. As shown in Figure 3, we place a collection of M primitive objects from set G into a $3D$ volume $F \in \mathbb{R}^{W \times H \times D}$ in descending order of their volume. Initially, the elements of the primitive objects set G are sorted in descending order according to the volumes of the primitive objects and are re-indexed as l' to reflect the sorted order. We set a condition regarding overlaps for the arrangement of primitive objects. When placing the object I_j in 3D volume F, we define the area already occupied by the objects $\{I_{l'}\}_{l'=1}^{j-1}$ as A_j, and the area that I_j occupies as B_j. The condition for overlap stipulates that the volume overlap ratio, represented as $\frac{O(A_j \cap B_j)}{O(B_j)}$, should be less than a threshold r. If this condition is met, we proceed with the placement. In this context, r denotes the maximum overlap ratio of the shapes, and O is a function representing the volume of the occupied region. The placement procedure for each primitive object comprises two main steps: (1) position selection and (2) arrangement. (1) position selection: randomly select a position of the center of the primitive object in the 3D volume F. (2) arrangement: placing the primitive object based on the overlap condition shown in Figure 3. If the overlap condition is met at the position from (1), we

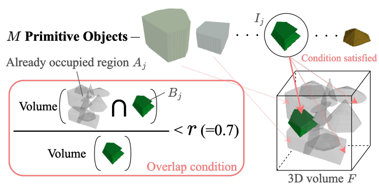

Figure 3: The details of the arrangement of primitive objects.

place the primitive object. If not, we revert to (1). This (1) and (2) process is repeated up to a maximum of Max_iter (=100). If the (1) and (2) process fails after 100 iterations, the primitive object is rejected. We repeat the placement process for each object $I_{l'}$ a total of M times, proceeding sequentially from $l' = 1$ to $l' = M$. The outcome of arranging the contours of the primitive objects is denoted as the assembled object S_i. The outcome of placing the primitive objects filled in the interior is denoted as m_i. In this study, we set the intensity values of the contours for the primitive objects within the assembled object to a fixed value, $Intensity$. We iteratively generate pairs (S_i, m_i) for N times, thereby automatically creating the pre-training dataset $\mathcal{D}$ for PrimGeoSeg. Combining various random parameters increases the diversity of geometric shapes within each class. We call this intra-class diversity of shapes instance augmentation. Please refer to the Supplementary Materials for the parameters and values required to generate the pre-training data for PrimGeoSeg.

3.2 Hypothesis and Motivation of PrimGeoSeg

Reasons for independent rules in the xy-plane and z-axis: 3D medical images are created by constructing of xy slices and reconstructing them along the z-axis. We thus generate primitive objects by stacking slice images on the xy-plane according to the z-axis rule.

Why is each xy-plane and z-axis rule defined as described above?: 3D general object recognition recognizes diverse and complex 3D objects in the real world. On the other hand, 3D medical image segmentation recognizes only a limited number of 3D objects within the human body's internal anatomy. Therefore, in PrimGeoSeg, we considered that a certain number of primitive objects should be sufficient for 3D image segmentation.

Why introduce the overlap when arranging primitive objects? The internal structure of a human being is such that blood vessels can penetrate the interior of organs. In this case, 3D medical images are represented as if the blood vessels overlap a part of the organ region. Therefore, we introduce the overlap when arranging multiple primitive objects to create an assembled object in which the part region that overlaps more closely resembles the internal structure of the human body.

4 Experiments

4.1 Experimental Settings

Datasets. In this experiment, we evaluate the effectiveness of PrimGeoSeg using several datasets: BTCV [2], MSD [1], and the 2021 edition of the Multi-modal Brain Tumor Segmentation Challenge (BraTS). BTCV has 30 samples for organ segmentation, and we split the BTCV training data in an 8:2 ratio for offline evaluation as in [6]. MSD has the lung (Task06) with 63 samples, and the spleen (Task09) has 41 samples for organ segmentation. In the MSD, we focused on the lung (Task06) and the spleen (Task09) due to computational resource constraints, and we split the training data in an 8:2 ratio for offline evaluation as in [28]. BraTS has 1,251 samples for brain tumor segmentation, we performed segmentation of three types of tumors: whole tumor (WT), tumor core (TC), and enhancing tumor (ET), splitting the training data in an 8:2 ratio for offline evaluation, as follows [10].

Architectures. We utilized the prominent transformer-based models, UNETR [11] and SwinUNETR [10], as architectures for 3D medical image segmentation. Both UNETR and

SwinUNETR have demonstrated their state-of-the-art performance on test leaderboards for the BTCV and MSD, surpassing the capabilities of conventional CNN-based models.

Implementation details. PrimGeoSeg executes the segmentation task with assembled objects $S_i \in \mathbb{R}^{96\times96\times96}$ as input data and the mask $m_i \in \mathbb{R}^{96\times96\times96}$ as ground truth in pre-training. During pre-training of PrimGeoSeg, we used $96 \times 96 \times 96$ patches, a batch size of 8, a learning rate of 0.0001, and a weight decay of 0.00001, optimizing the dice loss. We employ AdamW [19] with a warmup cosine scheduler for training. For the number of iterations, when the pre-training data is $\{5K, 50K\}$, the iterations are set to $\{100K, 375K\}$. Concerning the pre-training data for PrimGeoSeg, constructing a dataset of 5,000 objects requires less than two hours on a 400GiB CPU memory system, and the storage used is under 3GB. The pre-training process on NVIDIA A100 GPUs takes up to five GPU days for 100,000 iterations. For fine-tuning on BTCV, MSD, and BraTS, we follow the conditions of the hyperparameters on each fine-tuning dataset. For specific hyperparameter settings, we refer readers to the respective conventional research. Specifically, for SwinUNETR in BTCV and BraTS, refer to [25]; and for UNETR in BTCV and MSD, consult [11]. In addition, please see [11] in terms of MSD. All experiments for downstream tasks are conducted using a dice similarity coefficient (Dice) as an evaluation metric. For more detailed settings, please refer to the supplementary materials.

4.2 Fundamental Experiments (see Table 1)

Fundamental experiments aim to clarify the effectiveness of PrimGeoSeg pre-trained model. We focus on five aspects: (a) effects of volumetric shapes, (b) effects of the number of classes, (c) effects of instance augmentation (IA), (d) effects of overlap, and (e) effects of dataset size. We pre-trained UNETR using 2,500 data of PrimGeoSeg for all experiments.

Effects of volumetric shapes: This fundamental experiment (a) aims to compare the effectiveness of pre-training between planar shapes and volumetric shapes. We compare the effects of pre-training when arranging planar shapes and volumetric shapes in 3D space as shown in Figure 4. Table 1a shows that volumetric shapes improve the Dice metrics by 7.78 points compared to planar shapes. This result demonstrates that incorporating volumetric information leads to an improved performance of the pre-training for 3D medical image segmentation.

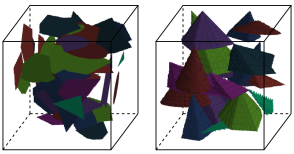

Figure 4: Expetiment (a).

Effects of the number of classes: This fundamental experiment (b) aims to investigate the validity of each xy-plane and z-axis rule of our generation method. Table 1b shows that both xy and z rules contribute to the effective pre-training. Moreover, the pre-training effect improves as the number of classes increases. For example, a maximum performance difference of +2.4 points was observed at 1 class and 32 classes. This result shows that the pre-training effect is enhanced even for primitive shapes when increasing diversity in the shape of the class. We clarify that the diversity of shapes in the xy-plane and z-axis directions in the 3D structure are both important factors in improving the pre-training effect.

Effects of instance augmentation (IA): This fundamental experiment (c) aims to examine the pre-training effect of our proposed IA method, as it considers individual primitive object variations similar to 3D medical images. As demonstrated in Table 1c, IA improves +3.31 points compared to the performance when not using IA pre-training performance. Also, while IA improves accuracy by 3.31 points, class diversity enhances 2.4 points (Table 1b).

Table 1: **Fundamental experiments for BTCV.** Each table show: (a) Effects of volumetric shapes, (b) Effects of the number of classes, (c) Effects of instance augmentation (IA), (d) Effects of overlap, and (e) Effects of the number of pre-training data.

(a) Shapes

	Dice
Planar	69.11
Volumetric	**76.89**

(b) Classes

	Dice
xy:1, *z*:1	75.12
xy:1, *z*:4	76.00
xy:8, *z*:1	76.89
xy:8, *z*:4	**77.52**

(c) IA

	Dice
w/o IA	74.21
w IA	**77.52**

(d) Overlap

	Dice
w/o overlap	77.52
w overlap	**78.14**

(e) Dataset size

	Dice
0.8K	77.39
2.5K	78.14
50K	**80.86**

This result suggests that intra-class shape diversity holds equal or greater importance than inter-class shape diversity in pre-training for 3D medical image segmentation.

Effects of overlap: This fundamental experiment (d) aims to investigate the pre-training effect of overlapping among 3D primitive objects. Because we hypothesize that overlaps between primitive objects could assist pre-training performance by considering, for example, fuzzy boundaries and overlapping regions within the human body. Table 1d shows that overlapping 3D volumetric shapes led to a higher accuracy of +0.62 points. This result suggests that the overlap between primitive objects contributes toward improving the pre-training effect of 3D medical image segmentation.

Effects of dataset size: One of the key advantages of PrimGeoSeg is its ability to generate primitive geometric objects automatically, which enables easily scaling of the pre-training dataset. As demonstrated in Table 1e, there is a positive correlation between the amount of pre-training data and the effectiveness of pre-training, where more data leads to better pre-training outcomes. Note that this experiment is limited to a certain amount of data size due to computational resource constraints.

The above experimental results revealed that volumetric shape extensibility, the number of classes, IA, overlap between primitive objects, and data scalability in PrimGeoSeg contribute to pre-training effects. The above fundamental experiments offered valuable insights into the key elements essential for pre-training in 3D medical image segmentation.

4.3 Organ and Tumor Segmentation (see Table 2 and Figure 5)

In this section, we verify the effectiveness of PrimGeoSeg on organ segmentation (BTCV and MSD Task09) and tumor segmentation (MSD Task06 and BraTS). We employed [6, 25] of the state-of-the-art SSL only because of limited computational resources and non-integrality of various conditions such as test data, architecture and input size.

BTCV. In Table 2a, we compare the fine-tuning results of our proposed method, learning from scratch, and the recent state-of-the-art SSL [6, 25], respectively. PrimGeoSeg showed an overall higher recognition performance than Scratch for each class. Even when utilizing an equivalent volume of pre-training data as with the SSL, we observed performance improvements: UNETR increased by 1.6 points and SwinUNETR by 0.4 points in average Dice score. Moreover, as detailed in Section 4.2, the performance of UNETR continued to improve as we increased the volume of pre-training data. It is worth noting that with only synthetic 3D pre-training data, the performance of our proposed method is superior to that of the baseline. In addition, Figure 5 shows several examples of SwinUNETR output results in BTCV. In areas that are over or under-segmented by Scratch and SSL, PrimGeoSeg is able to segment more accurately. Thus, we speculate that the distinct contours of PrimGeoSeg allow for the acquisition of more accurate 3D structual features during pre-training.

Table 2: **3D medical image segmentation benchmark datasets.** The results for PrimGeoSeg on BTCV, MSD, and BraTS in comparison to the previous SSL. The best value for each fine-tuning dataset is in bold.

(a) Comparison of performance in BTCV.

Pre-training	PT Num	Type	Avg.	Spl	RKid	LKid	Gall	Eso	Liv	Sto	Aor	IVC	Veins	Pan	rad	lad
UNETR																
Scratch	0	–	73.0	90.2	91.1	90.7	47.0	63.8	95.3	76.5	85.1	82.1	67.9	72.3	46.1	40.8
Chen *et al.* [6]	0.8K	SSL	75.8	95.2	**95.5**	93.8	51.9	52.3	**98.8**	80.0	87.8	82.7	66.1	68.9	60.8	51.3
PrimGeoSeg	0.8K	FDSL	77.4	88.9	94.0	93.8	59.8	65.7	95.4	79.3	88.3	82.6	69.9	76.8	58.5	53.3
PrimGeoSeg	50K	FDSL	**80.9**	**95.7**	94.2	**94.1**	**61.9**	**69.6**	96.7	**85.5**	**89.5**	**84.4**	**74.7**	**81.9**	**64.3**	**58.7**
SwinUNETR																
Scratch	0	–	78.3	92.3	93.2	93.8	55.9	61.3	94.0	77.0	87.5	80.4	74.2	76.1	68.8	63.6
Tang *et al.* [23]	5K	SSL	81.6	95.3	93.2	93.0	**63.6**	74.0	96.2	79.3	**90.0**	83.3	**76.1**	82.3	**69.0**	**65.1**
PrimGeoSeg	5K	FDSL	**82.0**	**95.7**	**94.4**	**94.4**	61.0	**75.5**	**96.7**	**83.3**	89.1	**85.6**	75.2	**84.3**	67.9	62.4

(b) Comparison of performance in MSD.

Pre-training	Type	UNETR		SwinUNETR	
		Lung	Spleen	Lung	Spleen
Scratch	–	52.5	95.0	63.5	96.3
Tang *et al.* [23]	SSL	–	–	65.2	96.5
PrimGeoSeg	FDSL	**62.2**	**96.3**	**67.9**	**96.6**

(c) Comparison of performance in BraTS.

Pre-training	Type	UNETR				SwinUNETR			
		Avg.	ET	WT	TC	Avg.	ET	WT	TC
Scratch	–	88.1	84.8	91.3	88.1	90.0	86.8	**92.9**	90.3
PrimGeoSeg	FDSL	**88.7**	**85.6**	**91.8**	**88.9**	**90.3**	**87.0**	**92.9**	**91.0**

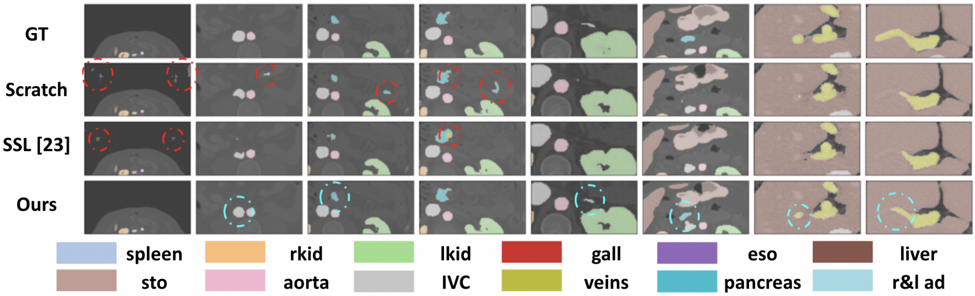

Figure 5: **Qualitative results on BTCV.** Red dashes indicate misidentified areas and blue dashes indicate more accurately identified areas.

MSD. Table 2b shows the accuracy when fine-tuning to MSD (Task06 and Task09). For lung segmentation, both UNETR and SwinUNETR, initialized by PrimGeoSeg improved accuracy compared to training from scratch by 9.7 points and 4.4 points, respectively. In addition, PrimGeoSeg on SwinUNETR showed a 2.7 points accuracy improvement compared to SSL. For spleen segmentation, when using PrimGeoSeg on UNETR and SwinUNETR, PrimGeoSeg exhibited accuracy improvements from Scratch of 1.3 points and 0.3 points, respectively. PrimGeoSeg on SwinUNETR had a 0.1 points accuracy improvement compared to SSL. From this result, PrimGeoSeg demonstrates superior pre-training performance without depending on a specific dataset.

BraTS. We verify the effectiveness of PrimGeoSeg for tumor segmentation. Due to the difficulty in conducting a fair comparison with other pre-training methods, we primarily focused on comparing PrimGeoSeg with Scratch. The BraTS results shown in Table 2c, indicate that PrimGeoSeg improves accuracy by approximately 0.5 points for ET, WT, and TC, respectively. Interestingly, although PrimGeoSeg is designed considering the key elements of the human body's internal structure, it has been proven effective for brain tumor segmentation. This suggests that PrimGeoSeg is capable of acquiring a 3D structural representation.

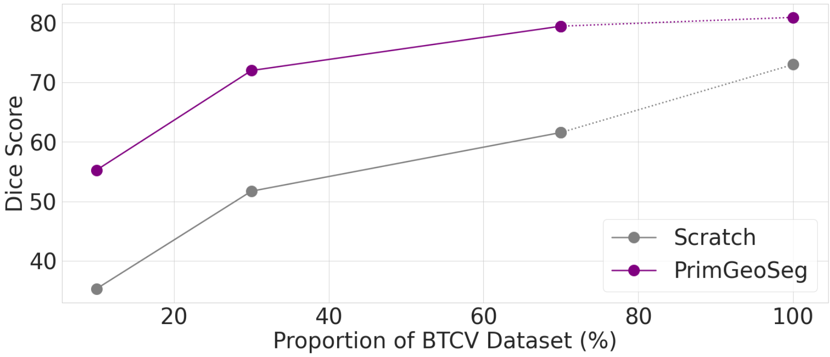

Figure 6: **Performance with limited training data.**

4.4 Pre-training Effect on Limited Training Data (see Figure 6)

In 3D medical image segmentation, achieving accurate recognition using a small number of 3D medical images is considered ideal. Figure 6 illustrates the results of limited training data in BTCV. Specifically, this experiment used training data (10% [2 samples], 30% [6 samples], 70% [15 samples]) to compare the performances of Scratch and PrimGeoSeg. As shown in Figure 6, the accuracy improvements of PrimGeoSeg over Scratch were {19.9, 20.3, 17.9}, respectively. Even more surprising, PrimGeoSeg uses only 30% of the training data, yet it achieves a performance comparable to that of Scratch, which uses 100% of the training data. This result demonstrates that PrimGeoSeg is beneficial, even with limited training data. Therefore, we consider it a promising pre-training approach to handling limited training data for 3D medical image segmentation.

5 Conclusion

This paper demonstrated the effectiveness of pre-training the proposed PrimGeoSeg, resulting in significant accuracy improvements compared to training from scratch for organ and tumor segmentation. Our proposed method also showed equal or superior performance to self-supervised learning. The findings through experimental results are described below;

The effect of the intra-class diversity. We observed that despite organs in the human body being essentially identical, the size and shape of these organs differ from person to person. In light of this observation, we experimented with a variety of 3D object types and shapes while building PrimGeoSeg. The results demonstrated that increasing the diversity of 3D objects contributes to the enhancement of medical image segmentation performance (see Table 1c).

The effect of spatial overlap of 3D objects. We also focused on the fact that the anatomical structure of the human body is complex and the boundaries between different tissues and organs are ambiguous, resulting in overlapping regions. Through our exploratory experiments on overlap (see Table 1d), it became clear that incorporating a certain amount of overlap can help improve performance.

While we empirically confirmed the performance enhancement due to shape pre-training, further analytical justification is required. Although the current focus is on addressing data scarcity in segmentation tasks, we plan to investigate the applicability of our method to other domains, such as 3D medical image classification and registration, in future research.

Acknowledgement. Computational resource of AI Bridging Cloud Infrastructure (ABCI) provided by the National Institute of Advanced Industrial Science and Technology (AIST) was used. We want to thank Hideki Tsunashima, Hiroaki Aizawa, Shinagawa Seitaro, and Takuma Yagi for their helpful research discussions.

References

[1] 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.

[2] 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.

[3] Josh Beal, Hao-Yu Wu, Dong Huk Park, Andrew Zhai, and Dmitry Kislyuk. Billion-scale pretraining with vision transformers for multi-task visual representations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 564–573, 2022.

[4] Sihong Chen, Kai Ma, and Yefeng Zheng. Med3d: Transfer learning for 3d medical image analysis. *arXiv preprint arXiv:1904.00625*, 2019.

[5] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[6] 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.

[7] Eli Gibson, Wenqi Li, Carole Sudre, Lucas Fidon, Dzhoshkun I Shakir, Guotai Wang, Zach Eaton-Rosen, Robert Gray, Tom Doel, Yipeng Hu, et al. Niftynet: a deep-learning platform for medical imaging. *Computer methods and programs in biomedicine*, 158: 113–122, 2018.

[8] Priya Goyal, Mathilde Caron, Benjamin Lefaudeux, Min Xu, Pengchao Wang, Vivek Pai, Mannat Singh, Vitaliy Liptchinsky, Ishan Misra, Armand Joulin, et al. Self-supervised pretraining of visual features in the wild. *arXiv preprint arXiv:2103.01988*, 2021.

[9] Fatemeh Haghighi, Mohammad Reza Hosseinzadeh Taher, Zongwei Zhou, Michael B Gotway, and Jianming Liang. Transferable visual words: Exploiting the semantics of anatomical patterns for self-supervised learning. *IEEE transactions on medical imaging*, 40(10):2857–2868, 2021.

[10] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 7th International Workshop, BrainLes 2021, Held in Conjunction with*

MICCAI 2021, Virtual Event, September 27, 2021, Revised Selected Papers, Part I, pages 272–284. Springer, 2022.

[11] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584, 2022.

[12] Nakamasa Inoue, Eisuke Yamagata, and Hirokatsu Kataoka. Initialization using perlin noise for training networks with a limited amount of data. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 1023–1028. IEEE, 2021.

[13] 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.

[14] 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.

[15] Hirokatsu Kataoka, Asato Matsumoto, Ryosuke Yamada, Yutaka Satoh, Eisuke Yamagata, and Nakamasa Inoue. Formula-driven supervised learning with recursive tiling patterns. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 4098–4105, October 2021.

[16] Hirokatsu Kataoka, Ryo Hayamizu, Ryosuke Yamada, Kodai Nakashima, Sora Takashima, Xinyu Zhang, Edgar Josafat Martinez-Noriega, Nakamasa Inoue, and Rio Yokota. Replacing labeled real-image datasets with auto-generated contours. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21232–21241, 2022.

[17] Hirokatsu Kataoka, Eisuke Yamagata, Kensho Hara, Ryusuke Hayashi, and Nakamasa Inoue. Spatiotemporal initialization for 3d cnns with generated motion patterns. 2022.

[18] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 491–507. Springer, 2020.

[19] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[20] Kodai Nakashima, Hirokatsu Kataoka, Asato Matsumoto, Kenji Iwata, and Nakamasa Inoue. Can vision transformers learn without natural images? *arXiv preprint arXiv:2103.13023*, 2021.

[21] Duy MH Nguyen, Hoang Nguyen, Mai TN Truong, Tri Cao, Binh T Nguyen, Nhat Ho, Paul Swoboda, Shadi Albarqouni, Pengtao Xie, and Daniel Sonntag. Joint self-supervised image-volume representation learning with intra-inter contrastive clustering. *arXiv preprint arXiv:2212.01893*, 2022.

[22] 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.

[23] Sora Takashima, Ryo Hayamizu, Nakamasa Inoue, Hirokatsu Kataoka, and Rio Yokota. Visual atoms: Pre-training vision transformers with sinusoidal waves, 2023.

[24] Aiham Taleb, Winfried Loetzsch, Noel Danz, Julius Severin, Thomas Gaertner, Benjamin Bergner, and Christoph Lippert. 3d self-supervised methods for medical imaging. *Advances in neural information processing systems*, 33:18158–18172, 2020.

[25] 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.

[26] 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.

[27] Ryosuke Yamada, Hirokatsu Kataoka, Naoya Chiba, Yukiyasu Domae, and Tetsuya Ogata. Point cloud pre-training with natural 3d structures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21283–21293, 2022.

[28] Yiwen Ye, Jianpeng Zhang, Ziyang Chen, and Yong Xia. Desd: Self-supervised learning with deep self-distillation for 3d medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part IV*, pages 545–555. Springer, 2022.

[29] Chenyu You, Ruihan Zhao, Lawrence H Staib, and James S Duncan. Momentum contrastive voxel-wise representation learning for semi-supervised volumetric medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part IV*, pages 639–652. Springer, 2022.

[30] Hong-Yu Zhou, Chixiang Lu, Sibei Yang, Xiaoguang Han, and Yizhou Yu. Preservational learning improves self-supervised medical image models by reconstructing diverse contexts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3499–3509, 2021.

[31] Zongwei Zhou, Vatsal Sodha, Jiaxuan Pang, Michael B Gotway, and Jianming Liang. Models genesis. *Medical image analysis*, 67:101840, 2021.

[32] Jiuwen Zhu, Yuexiang Li, Yifan Hu, Kai Ma, S Kevin Zhou, and Yefeng Zheng. Rubik's cube+: A self-supervised feature learning framework for 3d medical image analysis. *Medical image analysis*, 64:101746, 2020.