Pre-training Auto-generated Volumetric Shapes for 3D Medical Image Segmentation

Ryu Tadokoro[1,2]*, Ryosuke Yamada[1,3]*, Hirokatsu Kataoka[1]
[1]National Institute of Advanced Industrial Science and Technology,
[2]Tohoku University, [3]University of Tsukuba

Abstract

In 3D medical image segmentation, data collection and annotation costs require significant human efforts. Moreover, obtaining training data is challenging due to privacy constraints. Consequently, achieving efficient learning with limited data is an urgent 3D medical image segmentation issue. One approach to address this problem is using pre-trained models, which have been widely researched. Recently, self-supervised learning for 3D medical images has gained popularity, but the data available for such learning is also scarce, limiting the number of pre-training datasets. In recent years, formula-driven supervised learning has garnered attention. It can achieve high pre-training effects using only easily accessible synthetic data, making it a promising alternative for pre-training datasets. Inspired by this approach, we propose the Auto-generated Volumetric Shapes Database (AVS-DB) for data-scarce 3D medical image segmentation tasks. AVS-DB is automatically generated from a combination of dozens of 3D models based on polygons and shape similarity ratio variations. Our experiments show that AVS-DB pre-trained models significantly outperform models trained from scratch and achieve comparable or better performance than existing self-supervised learning methods we compared. AVS-DB can potentially enhance 3D medical image segmentation models and address limited data availability challenges.

1. Introduction

3D medical image segmentation is a critical task in medical analysis, with various applications such as surgical planning and measuring treatment effectiveness. However, collecting 3D medical images is more expensive compared to their 2D counterparts. Moreover, the process of annotating segmentation masks is labor-intensive and time-consuming, as it needs to be done for each individual slice. The strict privacy regulations surrounding medical data further complicate data acquisition and sharing. Consequently, the primary challenge in 3D medical image segmentation lies in achieving efficient learning with limited training data.

To address these issues, researchers have mainly focused on advancing model architectures and leveraging pre-trained models. While convolutional neural networks was prevalent in the past, recent studies have demonstrated that transformer models can outperform them [8, 9, 21]. Given the difficulties in obtaining supervised data, self-supervised learning has become increasingly popular [7, 15, 16, 22–24]. For instance, Chen et al. [4] successfully adapted the high-performance Masked Image Modeling technique [10, 18], originally developed for 2D images, to 3D images, leading to significant improvements in accuracy. Tang et al. [17] introduced a self-supervised learning approach that optimizes inpainting, contrastive learning, and rotation tasks concurrently. By employing the SwinUNETR [8] model, they achieved state-of-the-art results on the Multi Atlas Labeling Beyond The Cranial Vault (BTCV) [3] and Medical Segmentation Decathlon (MSD) [2] test leaderboards. In the future, the methodology of pre-training transformer-based models using self-supervised learning strategies is expected to see continued progress. However, due to the challenges in obtaining 3D medical images, self-supervised learning —which typically depends on data volume for performance improvement— faces limitations in capturing effective feature representations.

Meanwhile, in the field of image recognition, Formula-Driven Supervised Learning (FDSL) has gained traction due to its ability to automatically generate training data based on predefined rules [11–13]. FDSL has proven to acquire valuable feature representations for real-world recognition tasks despite being without real-world data. Notably, its effectiveness is most pronounced in Vision Transformers [6], where it surpasses ImageNet [5] pre-training performance in the 2D image domain using only synthetic datasets. Kataoka et al. [11] have shown that during the pre-training process, Vision Transformers focuses on contour shapes, suggesting that Vision Transformers learns fundamental feature representations through shape

*indicates equal contribution.

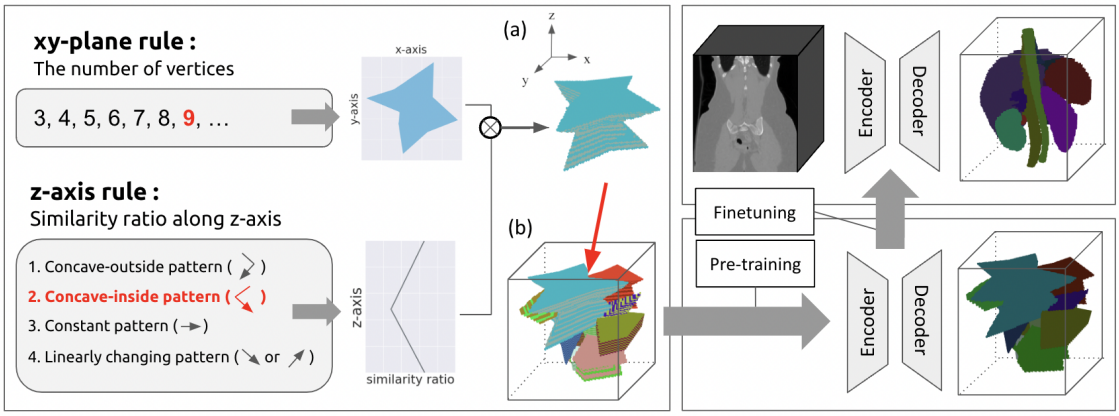

Figure 1. **Overview of the automatic generation method for AVS-DB and its pre-training.** The left side of the figure shows the process of constructing AVS-DB. By combining rules defined in the xy-plane and the z-axis direction, we create 3D shapes as shown in (a), and then arrange the 3D shapes as illustrated in (b). As indicated on the right side of the figure, we attempt to improve the accuracy in downstream tasks by using AVS-DB for pre-training the segmentation model.

pre-training. We propose that incorporating 3D contextual shapes in pre-training could further enhance the capabilities of transformer-based models in 3D medical image analysis, where data scarcity remains a significant challenge.

In this study, we construct a supervised dataset with 3D shapes for the segmentation task, a major task in 3D medical imaging. In the 3D medical image segmentation task, challenges arise due to (i) the internal structures exhibiting individual differences, and (ii) the human body having a highly complex anatomical structure characterized by ambiguous boundaries between different tissues and organs, as well as occasional overlapping regions. Based on these observations, we hypothesize that the following elements are crucial in the 3D medical image segmentation task: (i) 3D shapes with intra-class diversity to represent the diverse shape variations among individuals, and (ii) 3D models in which 3D shapes are spatially arranged with overlapping configurations.

To address these hypotheses, we propose the Auto-generated Volumetric Shapes DataBase (AVS-DB). Building AVS-DB involves a two-step process: synthesizing 3D shapes and arranging them spatially. For shape synthesis, we independently establish rules for the xy plane and z-axis direction and combine the two. During this stage, we introduce various instance augmentations to ensure diversity within the same class of shapes. We then intentionally arrange the synthesized shapes with overlaps in 3D space. Using AVS-DB, we pre-train a transformer-based model and evaluate its performance on the benchmark dataset for 3D medical image segmentation, BTCV, as well as MSD (Task06 and Task09). Our results show that AVS-DB improves performance by +0.4%, +2.01%, and +2.04% over conventional self-supervised learning we compared for BTCV and MSD (Task06 and Task09), respectively.

2. AVS-DB

3D medical images, primarily derived from Computed Tomography (CT) or Magnetic Resonance Imaging (MRI), consist of stacked 2D cross-sectional slices. These images exhibit different properties in the cross-sectional planes (xy-plane) and the perpendicular axis direction (z-axis). Therefore, we propose AVS-DB, automatically constructed by defining and combining rules for both the xy-plane and z-axis. The AVS-DB creation involves generating 3D shapes and arranging them as illustrated in Figure 1.

Generation of 3D shapes. We construct 3D shapes as shown in Figure 1 by defining and combining rules for both the cross-sectional direction (xy-plane) and the perpendicular axis direction (z-axis). For the xy-plane, we determine the number of vertices of the polygon by selecting a class. For the z-axis, we set multiple classes based on the patterns of similarity ratio changes along the z-axis direction. The number of vertices p_{xy} in the xy-plane is randomly sampled from the set of vertices $XY = \{p_i \in \mathbb{Z} | 1 \leq i \leq n\}$, where n represents the total number of vertex classes. We construct a shape with p_{xy} vertices as a cross-section. After defining the cross-section parallel to the xy-plane, we stack these shapes along the z-axis direction while changing the similarity ratio according to a specific rule. As shown in Figure 1, we set four rules: Concave-outside (similarity ratio: increases→decrease), 2. Concave-inside (similarity ratio: decrease→increase), 3. Constant, and 4. Linearly changing along the z-axis (similarity ratio :constantly increase / decrease).

By randomly selecting one of these rules, we construct 3D shapes by stacking cross-sectional shapes along the z-axis. To represent the diversity within each class, we perform instance augmentation as follows: In the xy plane,

any shape with the same number of vertices are considered to be part of the same class. Along the z axis, we capture the intra-class variability by specifying values, such as extrema or endpoints, within the range of the predefined rules. We denote the set of generated shape instances as $S = \{s_1, s_2, ..., s_N\}$, where N is the total number of shape instances created.

Arrangement of shapes. We arrange the generated shapes S in 3D space, as shown in Figure 1(b). For AVS-DB, we intentionally arrange the shapes to overlap. Let v_i be the volume of the shape s_i, and v_i' be the volume of the common part between s_i and the regions already filled by other shapes in the space. For each shape s_i, we randomly choose a position where the ratio of v_i' to v_i is less than a constant r, and place the shape s_i at that position.

Regarding the assignment of ground truth labels, we set the pixel values filled by the arranged shapes to the class ID assigned to that shape. In areas where shapes overlap, the masks of smaller shapes overwrite those of larger shapes during random placement.

3. Experiments

In this section, we first describe the detailed experimental settings in Sec 3.1. In Sec. 3.2, we conduct exploratory experiments to investigate the important elements in constructing the AVS-DB. In Sec. 3.3 and Sec. 3.4, we compare the pre-training effects of our models with self-supervised learning on 3D medical images in downstream tasks. In Sec. 3.5, we compare the accuracy of our method with existing FDSL adapted to 3D medical images.

3.1. Implementation Detail

Fine-tuning Datasets. To evaluate the effectiveness of pre-training with AVS-DB, we use the Multi Atlas Labeling Beyond The Cranial Vault (BTCV) [3] and the Medical Segmentation Decathlon (MSD) [2] as a downstream task. The BTCV consists of abdominal CT scans from 30 subjects annotated for 13 internal organs. Following a previous study [4], we divided the BTCV training data into train and test sets (8:2 ratio) for offline evaluation. The MSD focuses on segmentation tasks for ten tumors and internal organs. Due to constraints of machine resources, we chose to evaluate the segmentation performance on the lung (Task06) and spleen (Task09), which have smaller amounts of data, instead of evaluating all tasks. We divided the MSD (Task06 and Task09) similarly to BTCV and used the test set for offline performance evaluation.

Architectures and Hyperparameters. We employed standard Transformer-based models for the architecture, specifically UNETR [9] and SwinUNETR [8]. Unless otherwise specified, we utilized SwinUNETR for all experiments. For pre-training on the AVS-DB, we use $96 \times 96 \times 96$ patches, a batch size of 8, a learning rate of 0.0001, and a weight decay of 0.00001, optimizing the Dice Loss. The number of AVS-DB samples is set to 5,000, equivalent to the number of 3D medical images used in existing research [17]. We employ AdamW [14] with a Warmup Cosine Scheduler for training. For fine-tuning, we adhere to the experimental settings of [1] for BTCV. When pre-training involves only the encoder, we use the encoder's weights; otherwise, we use both the encoder and decoder weights. All evaluations for downstream tasks are conducted using the Dice Score.

Table 1. Results of exploration experiments in AVS-DB.

(a) Effect of shape classes.

	BTCV (Dice↑)
xy:4, z:2	80.29
xy:8, z:2	**81.16**
xy:8, z:4	80.80

(b) Effect of shape overlap.

	BTCV (Dice↑)
w / o overlap	81.16
w overlap (0.7)	**81.95**

3.2. Exploratory Experiments

We investigate the essential elements of AVS-DB pre-training performance, focusing on (i) the number of shape classes in the xy plane and z-axis direction and (ii) shape overlap. We use 5,000 AVS-DB samples.

Shape Classes. We investigate the relationship between the diversity of shape classes and the effectiveness of pre-training. We vary the number of classes for both the xy plane and z-axis directions. In Table 1, results show the lowest Dice Score (80.29%) when both have few classes. With a total of 16 classes, accuracy is 0.36% higher than with 32 classes (80.8%). An increase in xy plane rules positively contributes to pre-training, while an increase in z-axis direction rules has a negative impact. This suggests that learning complex xy plane shapes improves accuracy, while complex z-axis direction properties are not necessary.

Shape Overlap. Table 1 shows a 0.79% higher accuracy when shapes overlap, indicating that overlap positively contributes to pre-training. Overlapping shapes during pre-training are supposed to increase the effectiveness in downstream tasks, especially in areas where organs are closely packed, in contact, or overlap.

3.3. BTCV

Table 2 compares the accuracy of existing models pre-trained on 3D medical images using BTCV as a downstream task. PT Data denotes the pre-training dataset, with 3D med indicating the use of 3D medical images for pre-training. We selected two state-of-the-art SSL techniques in the context of transformer-based models as our comparison targets. We reference the results from [4] and utilize the pre-trained models from [17]. In order to conduct the fairest possible comparison, we made every effort to unify conditions such as test data, model architecture, and input size as

Table 2. Comparison of accuracy on the BTCV. The table demonstrates the performance of different pre-training methods, including our proposed method (AVS-DB), and their impact on model accuracy.Segmentation is performed on a total of 13 organs, including the spleen (Spl), right and left kidneys (RKid, LKid), gallbladder (Gall), esophagus (Eso), liver (Liv), stomach (Sto), aorta (Aor), inferior vena cava (IVC), portal and splenic veins (Veins), pancreas (Pan), and right and left adrenal glands (rad, lad).

Pre-train	PT Data	Network	Avg.	Spl	RKid	LKid	Gall	Eso	Liv	Sto	Aor	IVC	Veins	Pan	rad	lad
Scratch	-	UNETR	71.18	90.66	89.39	89.44	41.43	51.73	94.97	76.17	85.05	79.22	62.27	63.91	55.46	45.64
SSL [4]	3D med	UNETR	75.75	**95.20**	**95.45**	93.78	51.94	52.32	**98.75**	79.95	87.76	82.67	66.05	68.90	60.76	51.26
FDSL	AVS-DB	UNETR	**79.91**	94.66	94.11	**94.02**	**63.41**	**67.04**	96.32	**85.35**	**89.13**	**84.23**	**73.08**	**81.31**	**60.93**	**55.19**
Scratch	-	SwinUNETR	78.31	92.34	93.19	93.76	55.90	61.25	94.03	77.00	87.52	80.44	74.20	76.07	68.80	63.60
SSL [17]	3D med	SwinUNETR	81.56	95.27	93.17	92.98	**63.63**	73.96	96.21	79.32	**89.99**	83.30	**76.10**	82.26	**69.00**	**65.12**
FDSL	AVS-DB	SwinUNETR	**81.95**	**95.65**	**94.37**	**94.37**	61.03	**75.48**	**96.68**	**83.32**	89.11	**85.58**	75.23	**84.18**	67.91	62.41

Table 3. Comparison of segmentation accuracy on MSD. This table presents accuracy results for lung and spleen segmentation tasks on the MSD using AVS-DB pre-training.

Pre-train	PT Data	Task06:Lung	Task09:Spleen
Scratch	-	82.47	92.79
SSL [1]	3D med	83.30	92.80
FDSL	AVS-DB	**85.31**	**94.84**

much as possible. Our proposed method outperforms not only training from scratch but also SSL-based 3D medical image pre-training. Comparing UNETR and SwinUNETR, the accuracy improvement of FDSL (AVS-DB) over SSL is more pronounced for UNETR. Chen et al. [4] used approximately 800 3D medical images for pre-training UNETR, while Tang et al. [17] used about 5,000 for SwinUNETR. The number of pre-training 3D data may be a factor affecting the results.

3.4. MSD

Table 3 shows the accuracy when using MSD (Task06 and Task09) as the downstream task. For SSL [1], we report the results obtained by fine-tuning the publicly available pre-trained weights. In Task06, we observed a 2.84% accuracy improvement compared to training from scratch and a 2.01% accuracy improvement compared to SSL. In Task09, we observed a 2.05% accuracy improvement compared to training from scratch and a 2.04% accuracy improvement compared to SSL. In addition to the BTCV dataset, the superior pre-training performance of AVS-DB demonstrates its usefulness without depending on a specific dataset.

3.5. Comparison of AVS-DB and V-PCF

Table 4 compares pre-training performance between the existing Point Cloud Fractal DataBase [20] method for FDSL in 3D point cloud tasks and our proposed AVS-DB, using BTCV and MSD as downstream tasks. Since Point Cloud Fractal DataBase has a different data format than 3D images, we converted Point Cloud Fractal DataBase to a Voxelized Point Cloud Fractal DataBase (V-PCF) using the voxelization method as described in [19] and used it for pre-training classification tasks. Pre-training with AVS-DB yielded higher accuracy in BTCV and MSD (Task06, Task09) by 1.21%, 2.59%, and 1.25% compared to using V-PCF. We believe that the sparse data structure of V-PCF and the impact of the tasks imposed during pre-training are the causes of the difference in pre-training effects between AVS-DB and V-PCF.

Table 4. Comparison of pre-training performance between AVS-DB and V-PCF. Comparison of pre-training performance using BTCV and MSD as downstream tasks.

PT Data	BTCV	MSD (Task06)	MSD(Task09)
V-PCF	80.74	82.72	93.59
AVS-DB	**81.95**	**85.31**	**94.84**

4. Conclusion

In this paper, we demonstrated that by using our proposed AVS-DB for pre-training, a improvement in accuracy was achieved compared to training from scratch. Furthermore, the accuracy improvement was found to be on par with or better than self-supervised learning methods we compared. AVS-DB enables more data-efficient 3D medical image segmentation without relying on real-world data, which is burdened with data collection costs and privacy concerns. As future prospects, a deeper analysis of AVS-DB and exploration of its applicability to a broader range of tasks are necessary.

Acknowledgement. This paper is based on results obtained from a project, JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO). 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 and Seitaro Shinagawa for their helpful research discussions.

References

[1] Yamen Ali, Aiham Taleb, Marina M-C Höhne, and Christoph Lippert. Self-supervised learning for 3d medical image analysis using 3d simclr and monte carlo dropout. *arXiv preprint arXiv:2109.14288*, 2021. 3, 4

[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. 1, 3

[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. 1, 3

[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. 1, 3, 4

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[7] 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. 1

[8] 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. 1, 3

[9] 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. 1, 3

[10] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 1

[11] 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. 1

[12] 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. 1

[13] 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. 1

[14] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 3

[15] 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. 1

[16] 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. 1

[17] 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. 1, 3, 4

[18] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022. 1

[19] Yusheng Xu, Xiaohua Tong, and Uwe Stilla. Voxel-based representation of 3d point clouds: Methods, applications, and its potential use in the construction industry. *Automation in Construction*, 126:103675, 2021. 4

[20] 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. 4

[21] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnformer: Interleaved transformer for volumetric segmentation. *arXiv preprint arXiv:2109.03201*, 2021. 1

[22] 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. 1

[23] Zongwei Zhou, Vatsal Sodha, Jiaxuan Pang, Michael B Gotway, and Jianming Liang. Models genesis. *Medical image analysis*, 67:101840, 2021. 1

[24] Xinrui Zhuang, Yuexiang Li, Yifan Hu, Kai Ma, Yujiu Yang, and Yefeng Zheng. Self-supervised feature learning for 3d medical images by playing a rubik's cube. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*, pages 420–428. Springer, 2019. 1