Autoregressive Sequence Modeling for 3D Medical Image Representation

Siwen Wang[*3], Churan Wang[*2], Fei Gao[2], Lixian Su[3], Fandong Zhang[3], Yizhou Wang[2 4], Yizhou Yu[†1]

[1]School of Computing and Data Science, The University of Hong Kong
[2]Center on Frontiers of Computing Studies, School of Computer Science, Nat'l Eng. Research Center of Visual Technology, Peking University
[3]Deepwise AI Lab
[4]State Key Lab of General Artificial Intelligence, Inst. for Artificial Intelligence, Peking University
wangsiwendut@gmail.com, churanwang@pku.edu.cn, yizhouy@acm.org

Abstract

Three-dimensional (3D) medical images, such as Computed Tomography (CT) and Magnetic Resonance Imaging (MRI), are essential for clinical applications. However, the need for diverse and comprehensive representations is particularly pronounced when considering the variability across different organs, diagnostic tasks, and imaging modalities. How to effectively interpret the intricate contextual information and extract meaningful insights from these images remains an open challenge to the community. While current self-supervised learning methods have shown potential, they often consider an image as a whole thereby overlooking the extensive, complex relationships among local regions from one or multiple images. In this work, we introduce a pioneering method for learning 3D medical image representations through an autoregressive pre-training framework. Our approach sequences various 3D medical images based on spatial, contrast, and semantic correlations, treating them as interconnected visual tokens within a token sequence. By employing an autoregressive sequence modeling task, we predict the next visual token in the sequence, which allows our model to deeply understand and integrate the contextual information inherent in 3D medical images. Additionally, we implement a random startup strategy to avoid overestimating token relationships and to enhance the robustness of learning. The effectiveness of our approach is demonstrated by the superior performance over others on nine downstream tasks in public datasets.

Introduction

The realm of medical imaging has witnessed significant evolution with the advent of advanced modalities such as Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) (Najjar 2023). These three-dimensional (3D) medical images serve as a cornerstone for clinical diagnosis and treatment planning, providing physicians with a detailed glimpse into the inner workings of the human body (Panayides et al. 2020). Despite their widespread impact, the complexity and richness of 3D medical images present unique challenges in interpretation and analysis. The heterogeneity across different organs, the variability in diagnostic tasks, and the diversity of imaging modalities further complicate the comprehensive representation of these images (Zhou et al. 2023a).

In the data-driven paradigm of deep learning, leveraging large-scale well-annotated data can lead to effective representation learning (Ye et al. 2023; Tian et al. 2024). However, for medical image analysis, obtaining labeled data is particularly challenging, due to the intrusive nature of certain imaging modalities and the laborious process of annotation (Tajbakhsh et al. 2020; Jin et al. 2023). Therefore, we consider learning a general and effective representation from large-scale unlabeled data first. This allows our models to adapt to various downstream tasks with only a small amount of labeled data, significantly reducing the dependency on large-scale annotated data.

Moreover, recent advances in self-supervised learning (SSL) have shown promising results in visual representation learning for certain tasks (Oquab et al. 2023). These methods capitalize on the idea of reconstructing masked input data or contrastive learning to learn robust feature representations without the need for explicit labels. However, 3D medical images offer depth and volume, and are often sparse, posing new challenges when applying these methods to 3D medical image representations. As most existing self-supervised methods consider each image as a whole (Zhou et al. 2023b; El-Nouby et al. 2024; Gao et al. 2024), often overlooking inner and inter-correlations of 3D medical images, e.g., complex relationships among patches, modalities, and semantics. This limitation highlights the need for a more holistic and interconnected representation learning framework.

In this work, we address this challenge by introducing a novel self-supervised method to learn generalizable 3D medical image representations. We design a set of rules to transform diverse 3D medical images into coherent patch sequences. Every patch sequence is composed of several patches cropped from one or multiple original images according to spatial, semantic, or contrast correlation within 3D medical data. Every patch in a patch sequence is further divided into several visual tokens. All tokens from the patches in a patch sequence are concatenated to form a

[*]These authors contributed equally.
[†]Corresponding author.

longer token sequence. We further introduce an autoregressive sequence modeling task to guide the network in learning the spatial, contrast, and semantic correlations among tokens. Thereby our model fosters a deep understanding and integration of the contextual information encapsulated within 3D medical images. To avoid overestimating the relationships among tokens and for better downstream adaptation, we further employ a random startup strategy. This extends the concept of prefix causal attention from individual 2D natural images (El-Nouby et al. 2024) to 3D medical data. This strategy randomizes the starting point of a token sequence, preventing the model from relying on consistent sequence lengths and encouraging more robust learning of the intrinsic correlations within 3D medical data.

To summarize, our contributions are mainly three-fold:

1) We introduce a unified perspective by serializing diverse 3D medical images into token sequences. We propose an autoregressive sequence modeling task that guides the network to effectively learn generalizable 3D medical representations in a self-supervised manner.

2) We also extend prefix causal attention with random startup from 2D images to 3D medical representation to avoid overestimation of correlations among tokens and for better downstream adaptation.

3) We evaluate our method through nine downstream tasks in public datasets, such as segmentation of organs and tumours in CT or MRI and classification of COVID-19 and lung nodules. Our method achieves around 2.1% performance improvements on segmentation and 4%-6% performance improvements on classification, highlighting the potential of our method to advance the field of 3D medical image analysis.

Related Work

3D Medical Images Analysis. Compared to 2D images, 3D medical images possess a significantly higher spatial complexity, hence conventional 2D methods are inadequate for learning 3D representation. Numerous studies have conducted meaningful explorations into 3D medical image analysis, primarily divided into methods based on 2D and 3D models (Ni et al. 2019; Yang et al. 2021). The advantage of 2D models lies in their ability to leverage vast amounts of natural images to obtain powerful pre-trained models. They can take 3D image planes as input channels, blending multiplanar data into 2D models (Moeskops et al. 2016; Prasoon et al. 2013), or use three adjacent slices as channels (Ding et al. 2017; Yu et al. 2018). However, these methods are incapable of learning the context of the 3D space, which is a critical aspect for accurate representation and analysis in 3D medical images.

In contrast, 3D models are capable of learning more complex spatial features (Çiçek et al. 2016; Roth et al. 2018). However, it is often challenging to obtain 3D models pre-trained on large-scale data. Our proposed approach obtains powerful pre-trained 3D models by capturing inner-correlations and inter-correlations of 3D medical images.

Self-Supervised Learning. Self-supervised learning (SSL) has emerged as a promising approach in recent years to harness unlabeled data. Initially, SSL has achieved remarkable results in the domain of natural images. The current methods are divided into two categories approximately: contrastive-based and reconstruction-based (Chen et al. 2020; He et al. 2022). The core idea of contrastive learning is to minimize the feature distance between different views of the same image while maximizing the distance between different images, thereby forcing the model to learn discriminative features for instances (Chen et al. 2020, 2021; Caron et al. 2021). Contrastive methods primarily focus on global representations and lack local focus, which leads to suboptimal performance in dense prediction tasks (Zhou et al. 2021a; Zhang et al. 2022). Reconstruction-based methods primarily focus on masked image modeling, where a significant portion of the image content is masked and then reconstructed (He et al. 2022; Xie et al. 2022). However, they only explore representation in 2D individual nature images without considering the unique correlation within 3D medical images.

Due to the scarcity of annotated data, SSL has also garnered significant attention in the field of medical image analysis. Firstly, the researchers have designed many proxy tasks that focus on the characteristics of medical images, particularly for 3D images (Zhou et al. 2021b; Tang et al. 2022). Additionally, based on the contrastive approach, researchers have proposed improvement strategies, as well as methods that integrate multiple proxy tasks (Zhou et al. 2021a, 2023b). Although many designs have been proposed specifically for 3D medical images, current methods are often focused on individual images. There is a relative scarcity of approaches that utilize cross-sequence characteristics in 3D medical images for self-supervised representation learning. Taking the unique sequential correlation inherent in 3D medical images into full consideration, we propose to transform 3D images into various types of sequences. We design the autoregressive sequence modeling approach to learn inner and inter-correlation within 3D medical images.

Methodology

Our goal is to learn generalizable 3D medical image representations that can be applied across various downstream clinical tasks. As shown in Figure 1, we start by transforming diverse 3D medical images into a set of patch sequences, capturing the inherent spatial, contrast, and semantic correlations during pre-training. We then describe our training mechanism, which involves tokenizing these sequences and employing autoregressive sequence modeling to deeply integrate contextual information. After pre-training, we introduce how the learned representations are adapted to specific clinical tasks through a fine-tuning process, showcasing the versatility and applicability of our method.

Transform the 3D Inputs into Patch Sequences

Our approach leverages the transformation of 3D medical images into patch sequences. This strategy provides a solution to the challenge of capturing the rich interdependencies and contextual information inherent in volumetric data. By strategically transforming these images into different types of patch sequences, our model can effectively learn from the

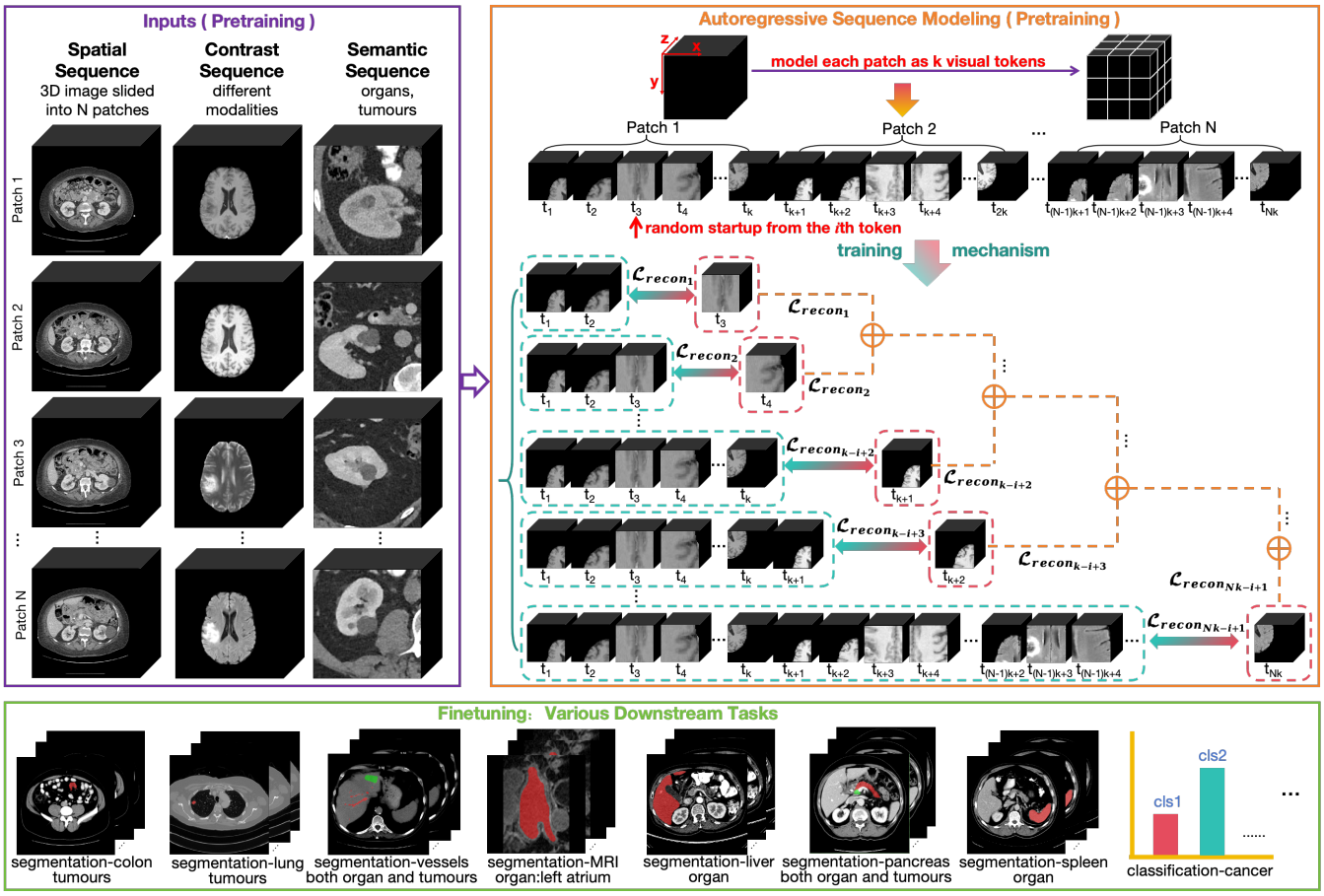

Figure 1: Overview of our Autoregressive Sequence Modeling approach for 3D Medical Images. The left purple box shows the transformation of one or more 3D medical images into a patch sequence with N patches, highlighting spatial, contrast, and semantic relationships within 3D data. In the orange box on the right, patches within the sequence are divided into visual tokens, which are then concatenated to form an ordered token sequence. During pre-training, the start of the token sequence t_i is selected randomly to enhance learning robustness. At the bottom of the orange box, the schematic diagrams of the training mechanism demonstrate how our method leverages autoregressive modeling to predict subsequent tokens and integrate contextual information. The green box shows our method can be generalized to various downstream tasks in the fine-tuning stage.

complex patterns and relationships within 3D medical data. To achieve this, we use diverse sources of medical image data, categorizing them into three distinct types of patch sequences as below:

Spatial Sequences. Capitalizing on the inherent spatial relationships encoded in 3D medical images, we employ a sliding window technique (*e.g.*, stride of 8). This allows creating a set of sequences that implicitly encode spatial contexts by extracting overlapping patches from 3D medical images.

Contrast Sequences. To leverage multi-modal images with varying contrasts (e.g. different image modalities within an MRI scan), we construct sequences that utilize these differences. By aligning and concatenating patches from the same anatomical location across different image modalities, we create sequences that capture the unique information imparted by each contrast.

Semantic Sequences. Given the sparsity and specialized characteristics of medical images, we use categories from the public DeepLesion (Yan et al. 2018) dataset to form semantic sequences. By grouping patches of similar semantic content, such as those from the same lesion or organ category, we enable our model to develop a robust understanding of semantic features.

Learning by Autoregressive Sequence Modeling

The core idea of our methodology lies in autoregressive sequence modeling, which leverages the sequential structure of our preprocessed 3D medical inputs. This section details how we use this model to learn rich, contextual representations of 3D medical data.

Token Sequences. As mentioned above, we first transform 3D images into patch sequences. Suppose each patch sequence has N patches. The next step involves tokenizing each patch into k visual tokens. All $N \times k$ tokens from a patch sequence are concatenated into a token sequence S following the order mentioned above as $\mathrm{T} = \{\boldsymbol{t}_1, ..., \boldsymbol{t}_{Nk}\}$.

Autoregressive Prediction. With token sequences, we employ an autoregressive sequence model to predict the probability distribution of each token in the sequence conditioned on its predecessors. The sequence is decomposed into a product of conditional probabilities for each token,

$$P(t) = \prod_{m=1}^{Nk} P(t_m \mid t_{<m}), \tag{1}$$

where the $P(t_m|t_{<m})$ represents the probability of the mth token given all previous tokens in the sequence.

Training Mechanism. For training our model, we use a normalized pixel-level regression loss, inspired by He *et al.* (He et al. 2022). The loss function is designed to minimize squared distances between predicted tokens and their ground-truth values, encouraging the model to accurately predict the next token in a sequence. To align the autoregressive nature of our pre-training with bidirectional self-attention required for downstream tasks and avoid overestimating correlations among tokens, we implement a prefix self-attention mechanism. This incorporation of **a random starting token** t_i ensures that all tokens preceding it are processed with bidirectional attention, capturing the comprehensive context available up to that point. This allows us to extend attention techniques used for 2D natural images (El-Nouby et al. 2024) to the more complex domain of 3D medical image analysis. The tokens that follow t_i in the sequence are then subject to autoregressive attention, where each token's prediction is conditioned on the preceding tokens. Importantly, only the tokens subsequent to t_i are included in the autoregressive prediction loss calculation, aligning with the formula:

$$\mathcal{L}_{recon} = \min \frac{1}{Nk - i + 1} \sum_{m=i}^{Nk} \|P(t_m) - t_m\|_2. \tag{2}$$

By training our model in this manner, our method can not only learn the intricate patterns within 3D medical images but also adapt flexibly to the diverse needs of various downstream clinical applications. Thereby the generalizability and effectiveness of our approach can be enhanced.

Fine-tuning on Downstream Tasks

After the pre-training stage, our model, enriched with comprehensive representations of 3D medical images, is adept at adapting to a variety of downstream tasks through a fine-tuning process. During fine-tuning, the input can be standard 3D medical images. Using the dataset for a given task, we initialize the model with pre-trained weights. We then optimize the parameters by minimizing the task-specific loss, such as Dice similarity for segmentation tasks or cross-entropy for classification. This fine-tuning approach enables rapid adaptation and enhanced performance on clinical tasks, as the model leverages the robust understanding developed during pre-training rather than starting from scratch. This strategy underscores the versatility and effectiveness of our method in advancing the analysis of 3D medical images for clinical applications.

Experiments

In this section, we first introduce the datasets used for pre-training and downstream tasks. Next, we detail the specific implementation of training and evaluation. Then, we list the methods compared in our study. Finally, we present a comprehensive set of experimental results, including a comparative analysis with existing state-of-the-art (SOTA) methods across multiple tasks, ablation studies highlighting the key components of our approach, and visualization results.

Implementation Details

Pre-training Datasets. The pre-training datasets are divided into several sources: individual images for spatial sequences, multimodal images for contrast sequences, and images belonging to the same semantic category for semantic sequences. For individual images, we collect 23,287 3D CT and MRI volumes from 12 public medical image datasets (RibFrac (Jin et al. 2020), TCIA Covid19 (An et al. 2020), AMOS22 (Ji et al. 2022), ISLES2022 (Hernandez Petzsche et al. 2022), AbdomenCT-1K (Ma et al. 2021), Totalsegmentator (Wasserthal et al. 2023), Verse 2020 (Sekuboyina et al. 2021), RSNA-2022-CSFD (Flanders et al. 2022), RSNA-2020-PED (Colak et al. 2021), STOTIC (Revel et al. 2021), FLARE22 (Ma et al. 2023), and FLARE23 (Ma et al. 2024)). For multimodal images, we collect 2,995 multimodal MRI scans from BraTS 23 (LaBella et al. 2023), which is a series of challenges on brain MRI image analysis. Each scan of this dataset includes four MRI modalities (T1w, T1ce, T2w, and Flair). Images belonging to the same semantic category are obtained from the DeepLesion dataset (Yan et al. 2018), which contains 10,594 CT scans of 4,427 patients.

Downstream Datasets. We conducted downstream experiments in nine clinical tasks on public medical image datasets to evaluate the effectiveness of our method. These datasets cover a variety of organs, lesions, and modalities, including Task03 Liver (131 cases), Task06 Lung (64 cases), Task07 Pancreas (282 cases), Task08 Hepatic Vessel (303 cases), Task09 Spleen (41 cases), and Task10 Colon (126 cases) from Medical Segmentation Decathlon (MSD) (Antonelli et al. 2022), Left Atrium (LA) (Xiong et al. 2021) (100 cases), RICORD (Tsai et al. 2021) (330 cases) and LIDC-IDRI (Armato III et al. 2011) (1633 cases). These datasets can be categorized into 3D segmentation and 3D classification tasks. Specifically, Task03, Task09, and LA are used for organ segmentation, while Task06 and Task10 focus on tumour segmentation. Task07 and Task08 are designed for segmenting both organs and tumours. RICORD is used for COVID-19 binary classification (being COVID-19 or not). LIDC-IDRI is used for lung nodule binary classification (level 1/2 into negative class and 4/5 into positive class, ignoring the cases with malignancy level 3) similar to other researches that have used this dataset (Wu et al. 2018). We randomly split the whole set into training, validation, and test at a ratio of 7:1:2 for the tasks on the MSD dataset. For the LA, RICORD, and LIDC-IDRI datasets, we follow the data split in (Yu et al. 2019), (Ye et al. 2024), and (Yang et al. 2023), respectively.

Methodology	CT: MSD dataset							MRI: LA dataset
	Task03 Liver	Task06 Lung	Task07 Pancreas	Task08 Hepatic Vessel	Task09 Spleen	Task10 Colon	Avg Dice	Dice Score
(Train From Scratch)								
UNETR (Hatamizadeh et al. 2022)	0.9285	0.4758	0.5384	0.5665	0.9372	0.2446	0.6152	0.8656
3D UNet (Çiçek et al. 2016)	0.9376	0.5222	0.5547	0.5770	0.9375	0.4057	0.6558	0.8755
(with General SSL)								
SimCLR (Chen et al. 2020)	0.9271	0.5631	0.5466	0.5519	0.9472	0.3330	0.6448	0.8988
MoCov3 (Chen et al. 2021)	0.9298	0.5730	0.5563	0.5480	0.9465	0.4200	0.6623	0.9009
DINO (Caron et al. 2021)	0.9392	0.5381	0.5478	0.5772	0.9470	0.4016	0.6585	0.9029
(with Medical SSL)								
PCRLv2 (Zhou et al. 2023b)	0.9451	0.6138	0.5894	0.5887	0.9417	0.4423	0.6868	0.9053
MAE3D (Chen et al. 2023)	0.9435	0.6277	0.5728	0.5878	0.9431	0.4522	0.6879	0.9054
MedCoSS (Ye et al. 2024)	0.9401	0.6292	0.5685	0.5890	0.9475	0.4381	0.6854	0.9062
Ours	**0.9593**	**0.6529**	**0.5910**	**0.6014**	**0.9585**	**0.4896**	**0.7088**	**0.9157**

Table 1: Segmentation results on Task03 Liver, Task06 Lung, Task07 Pancreas, Task08 Hepatic Vessel, Task09 Spleen, and Task10 Colon on MSD dataset (Antonelli et al. 2022) and LA dataset (Xiong et al. 2021).

Methodology	COVID-19		Lung Nodule	
	ACC	AUC	ACC	AUC
(Train From Scratch)				
ResNet (He et al. 2016)	0.7500	0.8133	0.8440	0.8630
ViT (Dosovitskiy et al. 2020)	0.7381	0.7940	0.8290	0.8587
(with General SSL)				
SimCLR (Chen et al. 2020)	0.7976	0.7904	0.8484	0.8605
MoCov3 (Chen et al. 2021)	0.7738	0.8446	0.8452	0.8683
DINO (Caron et al. 2021)	0.7857	0.8297	0.8323	0.8755
(with Medical SSL)				
PCRLv2 (Zhou et al. 2023b)	0.8095	0.8632	0.8516	0.8981
MAE3D (Chen et al. 2023)	0.8095	0.8703	0.8419	0.8906
MedCoSS (Ye et al. 2024)	0.8333	0.8803	0.8323	0.8983
Ours	**0.8929**	**0.9259**	**0.8871**	**0.9361**

Table 2: Classification results of COVID-19 diagnosis on RICORD (Tsai et al. 2021) and of lung nodule malignancy diagnosis on LIDC-IDRI (Armato III et al. 2011).

Training and Evaluation Details We use the AdamW optimizer and cosine learning rate decay scheduler for both pre-training and downstream tasks. In the pre-training stage, the initial learning rate is 1e-4, and we set 100K training steps with a batch size of 288. During the fine-tuning stage, the layer-wise learning rate decay strategy with the ratio of 0.75 is adopted for stabilizing the ViT training. The evaluation metric in classification tasks is the area under the receiver operator curve (AUC), and accuracy (ACC). For segmentation tasks, we use Dice similarity as the evaluation metric. To make a fair comparison, ViT-B (Dosovitskiy et al. 2020) is adopted as the backbone network, and UNETR (Hatamizadeh et al. 2022) is employed for segmentation tasks. More details of our implementation are provided in the supplemental material.

Compared Baselines Our baselines for segmentation include 1)UNETR (Hatamizadeh et al. 2022) and 2) 3D UNet (Çiçek et al. 2016), for classification include 1) ResNet (He et al. 2016) and 2) ViT-B (Dosovitskiy et al. 2020), which are trained from scratch. We further compare with three general SSL methods: 1) SimCLR (Chen et al. 2020), 2) MoCov3 (Chen et al. 2021), 3) DINO (Caron et al. 2021). The first two are based on contrastive learning. And DINO (Caron et al. 2021) introduces a self-distillation framework where a student model learns to predict the output of a teacher model. Then, we compare with three powerful methods specifically tailored for medical images: 1) PCRLv2 (Zhou et al. 2023b), 2) MAE3D (Chen et al. 2023), 3) MedCoSS (Ye et al. 2024). Specifically, MAE3D (Chen et al. 2023) is a generative method, which leverages an autoencoder and mask-based pretext task to learn visual representations from medical images. PCRLv2 (Zhou et al. 2023b) integrates pixel restoration and hybrid feature contrast into a multi-task optimization problem. MedCoSS (Ye et al. 2024) adopts a sequential pre-training paradigm using a continual learning approach.

Labeling Ratio	Dice (LA dataset)	AUC (RICORD)
100%	0.9157	0.9259
50%	0.9039	0.8696
25%	0.8741	0.8382
10%	0.8365	0.8054

Table 3: Results on segmentation (LA dataset (Xiong et al. 2021)) and classification (RICORD (Tsai et al. 2021)) tasks under our proposed method by training with different amounts of annotated data in fine-tuning stage.

Results on Downstream Tasks

We evaluate our proposed method in nine downstream tasks (organ segmentation for liver, spleen and left atrium, tumour segmentation for lung and colon, both organ and tumour segmentation for pancreas and hepatic vessel as shown in Table 1 and COVID-19 and lung nodule malignancy binary classification as shown in Table 2). In addition to the different goals of the tasks themselves, there are also two different modalities in them, *i.e.*, CT and MRI.

Transform 3D Images into Patch Sequences	Proposed Training Mechanism	Source Inputs for Pre-training			Stride	MSD Task06	RICORD
		Spatial	Contrast	Semantic		Dice	AUC
×	√	√	√	√	-	0.5880	0.8617
√	×	√	√	√	8	0.5852	0.8389
√	√	×	√	√	8	0.6165	0.8610
√	√	√	×	√	8	0.6076	0.8710
√	√	√	√	×	8	0.6137	0.8632
√	√	√	√	√	4	0.6258	0.8931
√	√	√	√	√	12	0.6435	0.8988
√	√	√	√	√	8	**0.6529**	**0.9259**

Table 4: Ablation study of our proposed method on segmentation (MSD Task06 (Antonelli et al. 2022)) and classification (RICORD (Tsai et al. 2021)) tasks.

Can the learned representation be used for CT analysis? We first conduct comparative evaluations of the proposed approach against the SOTA methods on six CT-based segmentation tasks from the MSD challenge. As illustrated in Column2-8 of Table 1, our method demonstrates superior performance across all evaluated tasks on the MSD challenge, outperforming other advanced methods. The average Dice score of our method is 0.7088, which is notably higher than the scores achieved by SimCLR (Chen et al. 2020) and MoCov3 (Chen et al. 2021), which are 0.6448 and 0.6623, respectively. Compared with DINO (Caron et al. 2021), our method also outperforms it by a large margin. These methods are designed for general SSL, relying heavily on the negative sampling process or data augmentations, which is not well-suited for the nuanced and complex nature of medical images. We further conduct experimental comparisons with PCRLv2 (Zhou et al. 2023b), MAE3D (Chen et al. 2023), and MedCoSS (Ye et al. 2024), which employ medical SSL. However, their performance gains are still limited. Our method gains an average 2.1% improvement in Dice score relative to the medical SSL methods. The improvements suggest that the proposed method can effectively leverage the contextual relationships inherent in 3D medical image data.

To evaluate the performance of our method on classification tasks, we conduct experiments on the COVID-19 (RICORD) dataset and lung nodule (LIDC-IDRI) dataset. As detailed in Table 2, the proposed method significantly surpasses all other methods. Compared with the SOTA methods, we achieve around 4%-6% improvement in performance. These results suggest that our method is effective for classification tasks.

Can the learned representation be used for MRI analysis? To validate the generalizability of our pre-trained model across different modalities, we further proceed with evaluations on the LA dataset which contains MRI data. As shown in the last column of Table 1, our model achieves a Dice coefficient of 0.9157, surpassing existing SOTA methods. This demonstrates the effectiveness of our approach in multimodal generalization.

Can the learned representation be still useful in less annotations? We take segmentation for the left atrium in LA dataset and classification for COVID-19 in RICORD dataset as examples. As shown in Table 3, our model shows competitive outcomes on both tasks, despite varying amounts of annotation availability. This indicates that our model effectively harnesses extensive unlabeled datasets to develop a versatile representation, thereby reducing reliance on fully labeled datasets. Remarkably, even when utilizing as little as 10% of the available labeled data, our model maintains satisfactory performance on both tasks, rivaling the results of compared baselines trained on complete datasets. This capability underscores the practical applicability of our model, particularly in scenarios where annotations are limited.

Visualization Results. We also visualize the segmentation results in Figure 2, providing the qualitative analysis of our performance on 3D medical images. The visualizations demonstrate the effectiveness of our autoregressive sequence modeling in capturing fine details and producing high-quality segmentations.

In summary, our method outperforms other compared methods in above all tasks. These results demonstrate the effectiveness of our method in improving the segmentation and classification of various anatomical structures and pathologies. Additionally, our method shows great potential for seamless integration into diverse clinical workflows.

Ablation Study

We evaluate some variant models to verify the effectiveness of each component, including transforming 3D images into patch sequences, the proposed training mechanism, and the source inputs for pre-training. Take segmentation for lung tumours in MSD dataset and classification for COVID-19 in RICORD dataset as examples. Some explanations of terms in the Table 4 are lists as below:

Transform 3D Images into Patch Sequences: × denotes not forming patch sequences from 3D images (El-Nouby et al. 2024), *i.e.*, inputs of the pre-training stage are individual images. √ denotes using our proposed method, *i.e.*, inputs of the pre-training stage are patch sequences, each of which is constructed from one or multiple 3D images.

Proposed Training Mechanism: × denotes not using our proposed training mechanism, *i.e.*, capturing the correlations among tokens by using the standard autoregressive attention mechanism without random startup during pre-training. √ denotes using our proposed method, *i.e.*, capturing the correlations among tokens by using the proposed training mechanism with a random startup during pre-training.

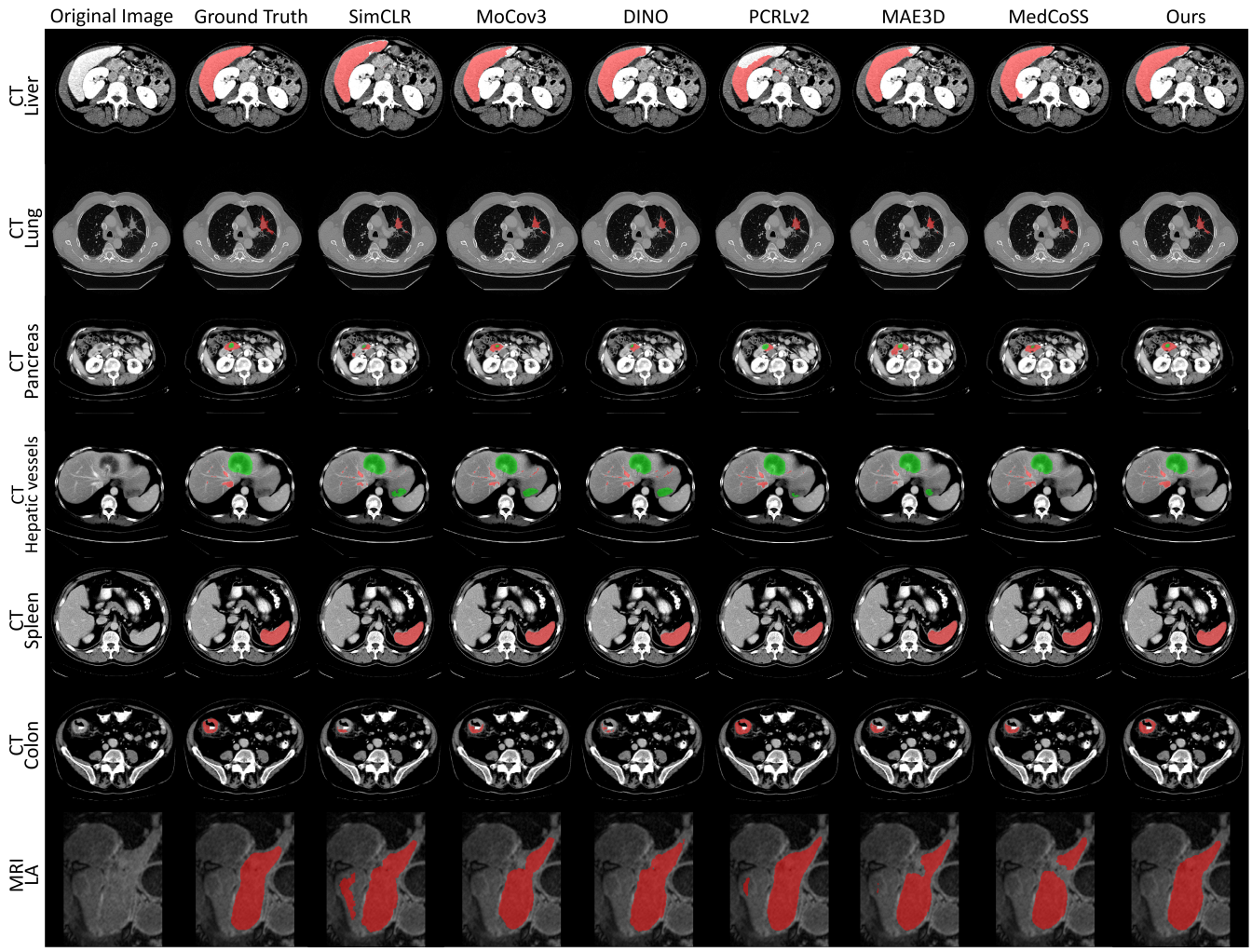

Figure 2: Visualizations of the segmentation results for various organs and pathologies from CT scans and MRI based on our proposed method and compared baselines. Each row denotes a different task. Each column denotes a different method.

Source Inputs for Pre-training: the source data used as inputs during pre-training.

Stride: used stride while transforming into spatial sequences. - denotes no stride operation is required if using individual images (without using our proposed transformation for patch sequences) as inputs of the pre-training stage.

In Column1, comparison between Row1 and Row8 shows a marked improvement when transforming 3D images into patch sequences is applied. This suggests that the patch sequence approach can better capture the inner and inter-correlation of the 3D medical images, e.g., the complex relationships between patches, modalities, and semantics. Row2 and Row8 in Column2 reflect the effectiveness of our proposed training mechanism with a random startup. The source of the inputs used during the pre-training phase is indicated in Column3-5 (Row3-5 and Row8). The results underscore the benefit of using a diverse set of data sources, as it enriches the model's understanding of different anatomical structures and imaging modalities. We further compare the model performance across different sliding stride settings in Column6 (Row6-8). Generally, larger sliding strides yield better performance (stride = 8 and stride = 12 outperform stride = 4). However, excessively large strides do not provide additional improvements. As shown in Table 4, with stride = 8, the proposed method achieves the best performance on segmentation and classification tasks. The ablative results in Table 4 show that removing or changing any of the components would lead to a descent in performance.

Conclusions

In this paper, we introduce an autoregressive sequence modeling approach to 3D medical image analysis, effectively capturing the generalizable representation of 3D medical images. Our proposed model's state-of-the-art performance across various diverse downstream tasks demonstrates its robustness and generalizability. Through a strategic fine-tuning process, our method rapidly adapts to specific clinical applications, significantly enhancing the capabilities of 3D medical image analysis and laying a solid foundation for future innovations in the field.

Acknowledgements

This work was partially supported by Hong Kong Research Grants Council under Collaborative Research Fund (Project No. HKU C7004-22G).

References

An, P.; Xu, S.; Harmon, S. A.; Turkbey, E. B.; Sanford, T. H.; Amalou, A.; Kassin, M.; Varble, N.; Blain, M.; Anderson, V.; Patella, F.; Carrafiello, G.; Turkbey, B. T.; and Wood, B. J. 2020. CT images in COVID-19 [Data set]. *The Cancer Imaging Archive. https://doi.org/10.7937/TCIA.2020.GQRY-NC81*.

Antonelli, M.; Reinke, A.; Bakas, S.; Farahani, K.; Kopp-Schneider, A.; Landman, B. A.; Litjens, G.; Menze, B.; Ronneberger, O.; Summers, R. M.; et al. 2022. The medical segmentation decathlon. *Nature Comm.*, 13(1): 4128.

Armato III, S. G.; McLennan, G.; Bidaut, L.; McNitt-Gray, M. F.; Meyer, C. R.; Reeves, A. P.; Zhao, B.; Aberle, D. R.; Henschke, C. I.; Hoffman, E. A.; et al. 2011. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. *Medical Physics*, 38(2): 915–931.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *ICCV*, 9650–9660.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *ICML*, 1597–1607. PMLR.

Chen, X.; Xie, S.; and He, K. 2021. An empirical study of training self-supervised vision transformers. In *ICCV*, 9640–9649.

Chen, Z.; Agarwal, D.; Aggarwal, K.; Safta, W.; Balan, M. M.; and Brown, K. 2023. Masked image modeling advances 3d medical image analysis. In *WACV*, 1970–1980.

Çiçek, Ö.; Abdulkadir, A.; Lienkamp, S. S.; Brox, T.; and Ronneberger, O. 2016. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *MICCAI*, 424–432. Springer.

Colak, E.; Kitamura, F. C.; Hobbs, S. B.; Wu, C. C.; Lungren, M. P.; Prevedello, L. M.; Kalpathy-Cramer, J.; Ball, R. L.; Shih, G.; Stein, A.; et al. 2021. The RSNA pulmonary embolism CT dataset. *Radiology: Artificial Intelligence*, 3(2): e200254.

Ding, J.; Li, A.; Hu, Z.; and Wang, L. 2017. Accurate pulmonary nodule detection in computed tomography images using deep convolutional neural networks. In *MICCAI*, 559–567. Berlin, Heidelberg: Springer-Verlag. ISBN 978-3-319-66178-0.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

El-Nouby, A.; Klein, M.; Zhai, S.; Bautista, M. A.; Toshev, A.; Shankar, V.; Susskind, J. M.; and Joulin, A. 2024. Scalable pre-training of large autoregressive image models. *arXiv preprint arXiv:2401.08541*.

Flanders, A.; Carr, C.; Colak, E.; Kitamura, F.; Lin, H.; Rudie, J.; Mongan, J.; Andriole, K.; Prevedello, L.; Riopel, M.; et al. 2022. RSNA 2022 cervical spine fracture detection. *Radiological Society of North America (RSNA)*.

Gao, F.; Wang, S.; Zhang, F.; Zhou, H.-Y.; Wang, Y.; Wang, C.; Yu, G.; and Yu, Y. 2024. Cross-dimensional medical self-supervised representation learning based on a pseudo-3D transformation. In *MICCAI*, 178–188. Springer.

Hatamizadeh, A.; Tang, Y.; Nath, V.; Yang, D.; Myronenko, A.; Landman, B.; Roth, H. R.; and Xu, D. 2022. Unetr: Transformers for 3d medical image segmentation. In *WACV*, 574–584.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *CVPR*, 16000–16009.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Hernandez Petzsche, M. R.; de la Rosa, E.; Hanning, U.; Wiest, R.; Valenzuela, W.; Reyes, M.; Meyer, M.; Liew, S.-L.; Kofler, F.; Ezhov, I.; et al. 2022. ISLES 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. *Scientific Data*, 9(1): 762.

Ji, Y.; Bai, H.; Ge, C.; Yang, J.; Zhu, Y.; Zhang, R.; Li, Z.; Zhanng, L.; Ma, W.; Wan, X.; et al. 2022. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *NeurIPS*, 35: 36722–36732.

Jin, C.; Guo, Z.; Lin, Y.; Luo, L.; and Chen, H. 2023. Label-efficient deep learning in medical image analysis: Challenges and future directions. *arXiv preprint arXiv:2303.12484*.

Jin, L.; Yang, J.; Kuang, K.; Ni, B.; Gao, Y.; Sun, Y.; Gao, P.; Ma, W.; Tan, M.; Kang, H.; et al. 2020. Deep-learning-assisted detection and segmentation of rib fractures from CT scans: Development and validation of FracNet. *EBioMedicine*, 62.

LaBella, D.; Adewole, M.; Alonso-Basanta, M.; Altes, T.; Anwar, S. M.; Baid, U.; Bergquist, T.; Bhalerao, R.; Chen, S.; Chung, V.; et al. 2023. The asnr-miccai brain tumor segmentation (brats) challenge 2023: Intracranial meningioma. *arXiv preprint arXiv:2305.07642*.

Ma, J.; Zhang, Y.; Gu, S.; Ge, C.; Ma, S.; Young, A.; Zhu, C.; Meng, K.; Yang, X.; Huang, Z.; et al. 2023. Unleashing the strengths of unlabeled data in pan-cancer abdominal organ quantification: the flare22 challenge. *arXiv preprint arXiv:2308.05862*.

Ma, J.; Zhang, Y.; Gu, S.; Ge, C.; Wang, E.; Zhou, Q.; Huang, Z.; Lyu, P.; He, J.; and Wang, B. 2024. Automatic organ and pan-cancer segmentation in abdomen CT: the flare 2023 challenge. *arXiv preprint arXiv:2408.12534*.

Ma, J.; Zhang, Y.; Gu, S.; Zhu, C.; Ge, C.; Zhang, Y.; An, X.; Wang, C.; Wang, Q.; Liu, X.; et al. 2021. Abdomenct-1k: Is abdominal organ segmentation a solved problem? *IEEE T-PAMI*, 44(10): 6695–6714.

Moeskops, P.; Wolterink, J. M.; Van Der Velden, B. H.; Gilhuijs, K. G.; Leiner, T.; Viergever, M. A.; and Išgum, I. 2016. Deep learning for multi-task medical image segmentation in multiple modalities. In *MICCAI*, 478–486. Springer.

Najjar, R. 2023. Redefining radiology: a review of artificial intelligence integration in medical imaging. *Diagnostics*, 13(17): 2760.

Ni, T.; Xie, L.; Zheng, H.; Fishman, E. K.; and Yuille, A. L. 2019. Elastic boundary projection for 3D medical image segmentation. In *CVPR*, 2109–2118.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Panayides, A. S.; Amini, A.; Filipovic, N. D.; Sharma, A.; Tsaftaris, S. A.; Young, A.; Foran, D.; Do, N.; Golemati, S.; Kurc, T.; et al. 2020. AI in medical imaging informatics: current challenges and future directions. *IEEE Journal of Biomedical and Health Informatics*, 24(7): 1837–1857.

Prasoon, A.; Petersen, K.; Igel, C.; Lauze, F.; Dam, E.; and Nielsen, M. 2013. Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network. In *MICCAI*, 246–253. Springer.

Revel, M.-P.; Boussouar, S.; de Margerie-Mellon, C.; Saab, I.; Lapotre, T.; Mompoint, D.; Chassagnon, G.; Milon, A.; Lederlin, M.; Bennani, S.; et al. 2021. Study of thoracic CT in COVID-19: the STOIC project. *Radiology*, 301(1): E361–E370.

Roth, H. R.; Oda, H.; Zhou, X.; Shimizu, N.; Yang, Y.; Hayashi, Y.; Oda, M.; Fujiwara, M.; Misawa, K.; and Mori, K. 2018. An application of cascaded 3D fully convolutional networks for medical image segmentation. *Computerized Medical Imaging and Graphics*, 66: 90–99.

Sekuboyina, A.; Husseini, M. E.; Bayat, A.; Löffler, M.; Liebl, H.; Li, H.; Tetteh, G.; Kukačka, J.; Payer, C.; Štern, D.; et al. 2021. VerSe: a vertebrae labelling and segmentation benchmark for multi-detector CT images. *Medical Image Analysis*, 73: 102166.

Tajbakhsh, N.; Jeyaseelan, L.; Li, Q.; Chiang, J. N.; Wu, Z.; and Ding, X. 2020. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical Image Analysis*, 63: 101693.

Tang, Y.; Yang, D.; Li, W.; Roth, H. R.; Landman, B.; Xu, D.; Nath, V.; and Hatamizadeh, A. 2022. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *CVPR*, 20730–20740.

Tian, Y.; Shi, M.; Luo, Y.; Kouhana, A.; Elze, T.; and Wang, M. 2024. FairSeg: A large-scale medical image segmentation dataset for fairness learning using segment anything model with fair error-bound scaling. In *ICLR*.

Tsai, E. B.; Simpson, S.; Lungren, M. P.; Hershman, M.; Roshkovan, L.; Colak, E.; Erickson, B. J.; Shih, G.; Stein, A.; Kalpathy-Cramer, J.; et al. 2021. The RSNA international COVID-19 open radiology database (RICORD). *Radiology*, 299(1): E204–E213.

Wasserthal, J.; Breit, H.-C.; Meyer, M. T.; Pradella, M.; Hinck, D.; Sauter, A. W.; Heye, T.; Boll, D. T.; Cyriac, J.; Yang, S.; et al. 2023. TotalSegmentator: robust segmentation of 104 anatomic structures in CT images. *Radiology: Artificial Intelligence*, 5(5).

Wu, B.; Zhou, Z.; Wang, J.; and Wang, Y. 2018. Joint learning for pulmonary nodule segmentation, attributes and malignancy prediction. In *2018 IEEE 15th ISBI*, 1109–1113. IEEE.

Xie, Z.; Zhang, Z.; Cao, Y.; Lin, Y.; Bao, J.; Yao, Z.; Dai, Q.; and Hu, H. 2022. Simmim: A simple framework for masked image modeling. In *CVPR*, 9653–9663.

Xiong, Z.; Xia, Q.; Hu, Z.; Huang, N.; Bian, C.; Zheng, Y.; Vesal, S.; Ravikumar, N.; Maier, A.; Yang, X.; et al. 2021. A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging. *Medical Image Analysis*, 67: 101832.

Yan, K.; Wang, X.; Lu, L.; and Summers, R. M. 2018. DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *Journal of Medical Imaging*, 5(3): 036501–036501.

Yang, J.; Huang, X.; He, Y.; Xu, J.; Yang, C.; Xu, G.; and Ni, B. 2021. Reinventing 2d convolutions for 3d images. *IEEE Journal of Biomedical and Health Informatics*, 25(8): 3009–3018.

Yang, J.; Shi, R.; Wei, D.; Liu, Z.; Zhao, L.; Ke, B.; Pfister, H.; and Ni, B. 2023. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1): 41.

Ye, Y.; Xie, Y.; Zhang, J.; Chen, Z.; Wu, Q.; and Xia, Y. 2024. Continual self-supervised learning: Towards universal multi-modal medical data representation learning. In *CVPR*, 11114–11124.

Ye, Y.; Xie, Y.; Zhang, J.; Chen, Z.; and Xia, Y. 2023. Uniseg: A prompt-driven universal segmentation model as well as a strong representation learner. In *MICCAI*, 508–518. Springer.

Yu, L.; Wang, S.; Li, X.; Fu, C.-W.; and Heng, P.-A. 2019. Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In *MICCAI*, 605–613. Springer.

Yu, Q.; Xie, L.; Wang, Y.; Zhou, Y.; Fishman, E. K.; and Yuille, A. L. 2018. Recurrent saliency transformation network: Incorporating multi-stage visual cues for small organ segmentation. In *CVPR*, 8280–8289.

Zhang, T.; Qiu, C.; Ke, W.; Süsstrunk, S.; and Salzmann, M. 2022. Leverage your local and global representations: A new self-supervised learning strategy. In *CVPR*, 16580–16589.

Zhou, H.-Y.; Guo, J.; Zhang, Y.; Han, X.; Yu, L.; Wang, L.; and Yu, Y. 2023a. nnformer: Volumetric medical image segmentation via a 3d transformer. *IEEE TIP*.

Zhou, H.-Y.; Lu, C.; Chen, C.; Yang, S.; and Yu, Y. 2023b. A unified visual information preservation framework for self-supervised pre-training in medical image analysis. *IEEE T-PAMI*, 45(7): 8020–8035.

Zhou, H.-Y.; Lu, C.; Yang, S.; Han, X.; and Yu, Y. 2021a. Preservational learning improves self-supervised medical image models by reconstructing diverse contexts. In *ICCV*, 3499–3509.

Zhou, Z.; Sodha, V.; Pang, J.; Gotway, M. B.; and Liang, J. 2021b. Models genesis. *Medical Image Analysis*, 67: 101840.