Self Pre-training with Topology- and Spatiality-aware Masked Autoencoders for 3D Medical Image Segmentation

Pengfei Gu*
University of Texas Rio Grande Valley
Edinburg, TX 78539, USA
pengfei.gu01@utrgv.edu

Huimin Li*
University of Texas Rio Grande Valley
Edinburg, TX 78539, USA
huimin.li01@utrgv.edu

Yejia Zhang
University of Notre Dame
Notre Dame, IN 46556, USA
chazhang0310@gmail.com

Chaoli Wang
University of Notre Dame
Notre Dame, IN 46556, USA
chaoli.wang@nd.edu

Danny Z. Chen
University of Notre Dame
Notre Dame, IN 46556, USA
dchen@nd.edu

***Abstract*—Masked Autoencoders (MAEs) have been shown to be effective in pre-training Vision Transformers (ViTs) for natural and medical image analysis problems. By reconstructing missing pixel/voxel information in visible patches, a ViT encoder can aggregate contextual information for downstream tasks. But, existing MAE pre-training methods, which were specifically developed with the ViT architecture, lack the ability to capture geometric shape and spatial information, which is critical for medical image segmentation tasks. In this paper, we propose a novel extension of known MAEs for self pre-training (i.e., models pre-trained on the same target dataset) for 3D medical image segmentation. (1) We propose a new topological loss to preserve geometric shape information by computing topological signatures of both the input and reconstructed volumes, learning geometric shape information. (2) We introduce a pre-text task that predicts the positions of the centers and eight corners of 3D crops, enabling the MAE to aggregate spatial information. (3) We extend the MAE pre-training strategy to a hybrid state-of-the-art (SOTA) medical image segmentation architecture and co-pretrain it alongside the ViT. (4) We develop a fine-tuned model for downstream segmentation tasks by complementing the pre-trained ViT encoder with our pre-trained SOTA model. Extensive experiments on five public 3D segmentation datasets show the effectiveness of our new approach.**

***Index Terms*—Self-supervised Learning, Masked Autoencoders, Topology, Spatiality, 3D Medical Image Segmentation**

I. Introduction

Accurate segmentation of medical images is critical for medical analysis and applications such as diagnosis, treatment planning, and research. While many deep learning (DL) models (e.g., [1], [25]) have demonstrated impressive performances in medical image segmentation, such methods still face several key challenges. One challenge is the scarcity of high-quality labeled medical images for model training, due to high costs and expertise needed for data collection and annotation. Another challenge is annotation errors, as labeling 3D medical images can be very time-consuming and error-prone.

Self-supervised learning (SSL), a technique that leverages pre-text tasks to derive useful visual representations from unlabeled data, offers a promising avenue to combat the challenge of label scarcity. One representative methodology for SSL is Masked Autoencoders (MAEs) [11]. Specifically, MAE learns to reconstruct the missing pixels after randomly masking a certain fraction (e.g., 75%) of patches of the input images. In the medical image segmentation area, MAE pre-training has also been found to be effective (e.g., UNETR + MAE [28]). Although simple and effective, there are still several limitations. First, geometric shape information (i.e., contextual information on the overall shapes of objects), which is critical for improving segmentation performance, is not captured well (e.g., see Fig. 1). Second, global spatial information is not well explored since the focus has been on reconstructing information from the masked local sub-volumes, possibly neglecting the global context information of the target objects as a whole. Third, the MAE pre-training strategy (i.e., learning representations by reconstructing missing patches from masked image input) is not exploited well with various common medical image segmentation architectures, e.g., those based on convolutional neural networks (CNNs) or hybrid models. This is primarily because MAE was developed using the Vision Transformer (ViT) [6] architecture, potentially restricting its adaptability and effectiveness with other architectures.

To address these limitations, we propose a novel extension of MAEs for self pre-training for 3D medical image segmentation. (I) We extract geometric shape information by exploiting multi-scale topological features (e.g., connected components, cycles/loops, and voids). Our method utilizes cubical complexes [14] to compute topological signatures of both the input and reconstructed volumes, and employs an optimal transport distance (the 2-Wasserstein distance) to

* Equal contribution

derive a new topological loss. Our topology-aware loss is fully differentiable, computationally efficient, can be added to any neural network, and is applicable to 2D/3D images. (II) We propose a pre-text task to predict the positions of multiple key points of crops, enabling the model to aggregate spatial information. Specifically, our method predicts the positions of nine points (the center and eight corners) of a 3D crop in the input volume. By learning where the crops are located in the input volume, the model can capture global spatial information. (III) We extend the MAE pre-training strategy to a hybrid state-of-the-art (SOTA) medical image segmentation architecture, UNETR++ [18], and co-pretrain UNETR++ alongside the ViT. Specifically, masked crops are processed separately by both ViT and UNETR++ to reconstruct the associated missing patches. Reconstruction consistency loss and spatial consistency loss (derived from the pre-text task) are employed to connect the two different types of architectures in pre-training, enhancing their representation learning capability.

Following [28], our method is performed on self pre-training paradigms (i.e., models pre-trained on the same target dataset). In the self pre-training stage, we randomly mask a fraction (e.g., 50%) of patches of the image crops. The masked crops are then processed independently by a ViT model and a UNETR++ model, which are pre-trained with our proposed topological loss, pre-text task that predicts the positions of 9 key points of crops, and spatial and reconstruction consistency losses, learning the geometric shape and global spatial information and enhancing the representation learning capability. In the fine-tuning stage, the pre-trained ViT encoder is complemented with the pre-trained UNETR++ model, which is then fine-tuned for the target segmentation task. A fusion module is utilized to fuse the scale-wise features from both the pre-trained ViT encoder and UNETR++ encoder.

Our main contributions are summarized as follows: (1) We propose a new topological loss and introduce a pre-text task for MAEs to learn geometric shape and spatial information. (2) We extend the MAE pre-training strategy to a hybrid SOTA medical image segmentation architecture and co-pretrain it alongside ViT. (3) We develop a fine-tuned model for downstream segmentation tasks, and demonstrate the effectiveness of our new approach on five public 3D segmentation datasets.

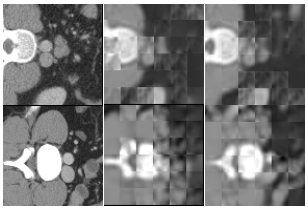

Fig. 1. Illustrating the effect of our proposed topological loss. Left: raw image examples of the Synapse CT dataset; middle: reconstructed images with the mean squared error (MSE) loss [28]; right: reconstructed images with a combination of the MSE and proposed topological losses.

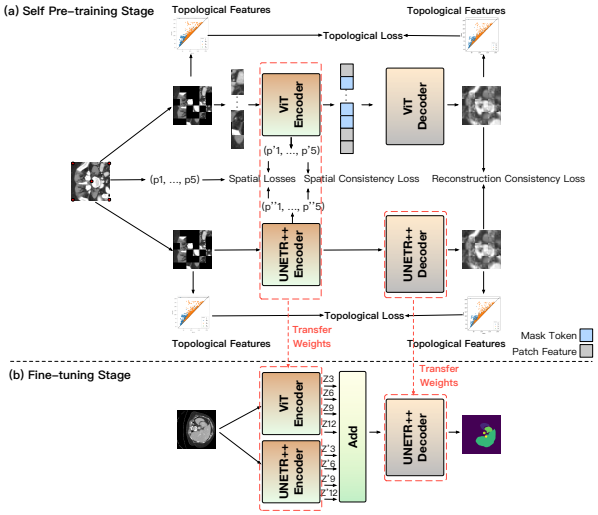

Fig. 2. An overview of our proposed pipeline.

II. Method

Fig. 2 presents an overview of our proposed pipeline, which contains four main components: (1) a topological loss that aims at implicitly extracting geometric shape information by exploiting multi-scale topological features; (2) a pre-text task that captures global spatial information by predicting the positions of 9 key points of 3D crops in the input volume; (3) a spatial consistency loss and a reconstruction consistency loss that enhance the representation learning capability of both the ViT and UNETR++ models by aligning the reconstructed images at both spatial and image levels; (4) a fine-tuned model for improving the downstream segmentation performance.

A. Capturing the Topology of Input Volumes

Given a 3D image I, we represent I with a *cubical complex* C. Typically, the cubical complex C takes each voxel of I as an individual vertex and contains connectivity information on vertex neighbourhoods via edges, squares, and their higher-dimensional counterparts [8], [14]. In this work, we use *persistent homology* (PH) [7] to extract topological features of different dimensions from C, including connected components (0-D), cycles/loops (1-D), and voids (2-D). PH combines the homology of super-level sets by sweeping a threshold function through the entire real numbers. Specifically, for a threshold value $\tau \in \mathbb{R}$, a cubical complex is defined as: $C^{(\tau)} := \{x \in I \mid f(x) \geq \tau\}$, where $f(x)$ is the voxel value of x. When sweeping the threshold, the topology changes only at a finite number of values, $\tau_1 \geq \tau_2 \geq \cdots \geq \tau_{m-1} \geq \tau_m$, and we obtain a sequence of nested cubical complexes, $\emptyset \subseteq C^{(\tau_1)} \subseteq C^{(\tau_2)} \subseteq \ldots \subseteq C^{(\tau_{m-1})} \subseteq C^{(\tau_m)} = I$, which forms the *super-level set filtration*. PH tracks topological features across all the complexes in this filtration, representing each feature as a tuple (τ_i, τ_j) with $\tau_i \geq \tau_j$, indicating the cubical complex in which a feature appears and disappears, respectively. For example, a 0-D tuple (τ_i, τ_j) represents a connected component that appears at threshold τ_i and disappears at threshold τ_j. The tuples of

the k-D ($0 \leq k \leq 2$) features are saved in the *k-th persistence diagram* D_I^k, which is a multi-scale shape descriptor of all topological features of the 3D image I.

Comparing Persistence Diagrams. Given two persistence diagrams D and D', we use the *2-Wasserstein distance* as a metric to measure their similarity or distance, defined as: $W_2(D, D') := \left(\inf_{\eta: D \to D'} \sum_{x \in D} ||x - \eta(x)||_\infty^2\right)^{\frac{1}{2}}$, where $\eta(\cdot)$ denotes a bijection. Note that this equation can be solved by using an optimal transport algorithm, and we use cubical Ripser [14] to compute PHs from volumes.

Constructing Topological Loss. Given an input volume I and a reconstructed volume I', our new topology-aware loss is defined as: $\mathcal{L}_{topo}(I, I') = \left(\sum_{i=0}^{2} (W_2(D_I^i, D_{I'}^i))^2\right)^{\frac{1}{2}}$, where $W_2(\cdot, \cdot)$ denotes the 2-Wasserstein distance, and D_I^i and $D_{I'}^i$ are the i-D persistence diagrams of I and I', respectively. Note that our proposed topological loss differs from that in [19]. First, we extract topological features from 3D images, not from the segmentation. Second, we use the topological loss for self pre-training and medical image segmentation, not for 3D reconstruction tasks.

B. Exploiting Global Spatial Information

MAE [11] and UNETR + MAE [28] lack the ability to learn global spatial information that is vital to 3D medical image segmentation for two reasons: (1) The positional embedding encodes only local position information for each patch, and (2) the methods focus only on low-level patch matching with a local mean squared error (MSE) loss. To address these limitations, we propose a novel pre-text task that complements the known methods with global spatial information. Specifically, the pre-text task aims to predict the positions of 9 key points (the center and eight corners) of 3D crops in the input volume. We attain this by adding two prediction heads to the ViT and UNETR++ encoders. The two prediction heads share the same architecture that consists of a convolutional layer, a two-layer multilayer perceptron (MLP) with 256 hidden dimensions, and a $\tanh$ activation function. This design enables the ViT and UNETR++ encoders to learn global spatial representations.

Constructing Spatial Loss. We denote the 9 key points of a 3D crop as $(p_1, p_2, \ldots, p_9)$, where p_9 is for the crop center, and each $p_i = (x_i, y_i, z_i)$. Given the ground truth (GT) and the prediction of the 9 positions, $P = (p_1, p_2, \ldots, p_9)$ and $P' = (p'_1, p'_2, \ldots, p'_9)$, the spatial loss is defined as: $\mathcal{L}_{spa}(P, P') = \mathcal{L}_{MSE}(P, P')$, where $\mathcal{L}_{MSE}$ is the MSE loss. Our spatial loss definition and implementation are different from those in [23].

C. Co-pretraining the ViT and UNETR++ Models

As illustrated in Fig. 2, the masked crops are processed independently by both the ViT and UNETR++ models. To co-pretrain both the ViT and UNETR++ models to enhance their representation learning capability, we propose to align the reconstructed images in the spatial and image levels.

Constructing Spatial Consistency Loss. Given predictions of 9 key point positions from the ViT and UNETR++ encoders, $P' = (p'_1, p'_2, \ldots, p'_9)$ and $P'' = (p''_1, p''_2, \ldots, p''_9)$, the spatial consistency loss is defined as: $\mathcal{L}_{spa-consis}(P', P'') = \mathcal{L}_{MSE}(P', P'')$. The spatial consistency loss aligns the reconstructed images at the spatial level, enhancing the feature learning capability of both the ViT and UNETR++ models.

Constructing Reconstruction Consistency Loss. Given reconstructed volumes I' and I'' by the ViT and UNETR++ models, our reconstruction consistency loss function computes the MSE between the reconstructed volumes I' and I'', as: $\mathcal{L}_{rec-consis}(I', I'') = \mathcal{L}_{MSE}(I', I'')$. We compute this loss only on masked patches, similar to MAE in [11]. The reconstruction consistency loss aligns the reconstructed images at the image level, further enhancing the representation learning capability of both ViT and UNETR++.

The Overall Loss of Self Pre-training. The overall loss for a volume crop is:

$$\begin{aligned}\mathcal{L} = (1-\lambda_1)(1-2\lambda_2)\mathcal{L}_{MSE-ViT} + (1-\lambda_1)\lambda_2\mathcal{L}_{topo-ViT} \\ + (1-\lambda_1)\lambda_2\mathcal{L}_{spa-Vit} + \lambda_1(1-2\lambda_2)\mathcal{L}_{MSE-UNETR++} \\ + \lambda_1\lambda_2\mathcal{L}_{topo-UNETR++} + \lambda_1\lambda_2\mathcal{L}_{spa-UNETR++} \\ + \lambda_3\mathcal{L}_{spa-consis} + \lambda_3\mathcal{L}_{rec-consis},\end{aligned}$$

where $\mathcal{L}_{MSE-X}$, $\mathcal{L}_{topo-X}$, and $\mathcal{L}_{spa-X}$ are the reconstruction, topological, and spatial losses, respectively, for the X (either ViT or UNETR++) model, and λ_i is a balancing weight.

D. Constructing the Fine-tuned Architecture

In [28], the pre-trained ViT encoder weights were transferred to initialize the segmentation encoder, i.e., the UNETR [10] encoder, achieving impressive performance. Following [28], we utilize the pre-trained ViT encoder weights and propose to complement the pre-trained ViT encoder with the pre-trained UNETR++ [18] to enhance the performance of downstream segmentation tasks.

As shown in Fig. 2, our fine-tuned model consists of four key components: two pre-trained encoders (the pre-trained ViT and UNETR++ encoders), an add fusion module, and a pre-trained UNETR++ decoder. Specifically, the two encoders are employed to capture complementary features, since the ViT encoder is a Transformer-based architecture and the UNETR++ encoder is a convolution-based architecture. Then a scale-wise fusion module, which is addition, is used to fuse the scale-wise features from the two different types of encoders. Finally, a pre-trained UNETR++ decoder is appended to generate the final segmentation. Our fine-tuned model is called **MAE + UNETR++**, which can effectively leverage the pre-trained ViT encoder to capture high-level semantic information and the pre-trained UNETR++ to better capture fine details and edges, resulting in improved segmentation performance.

III. Experiments and Analysis

A. Datasets and Experimental Setup

We conduct experiments on five segmentation datasets: Synapse multi-organ CT segmentation (Synapse CT Dataset) [15], BTCV multi-organ CT segmentation (BTCV CT Dataset) [15], ACDC automated cardiac diagnosis (ACDC) [3], and Medical Segmentation Decathlon (MSD)

TABLE I

SEGMENTATION RESULTS ON THE SYNAPSE CT DATASET. THE BEST RESULTS ARE MARKED IN **bold**, AND THE SECOND-BEST RESULTS ARE UNDERLINED.

Method	Params.	FLOPs	Spl	RKid	LKid	Gal	Liv	Sto	Aor	Pan	Average	
											Dice (↑)	HD95 (↓)
U-Net [17]	—	—	86.67	68.60	77.77	69.72	93.43	75.58	89.07	53.98	76.85	—
TransUNet [5]	96.07M	88.91	85.08	77.02	81.87	63.16	94.08	75.62	87.23	55.86	77.49	31.69
UNETR [10]	92.49M	75.76	85.00	84.52	85.60	56.30	94.57	70.46	89.80	60.47	78.35	18.59
Swin-UNet [4]	—	—	90.66	79.61	83.28	66.53	94.29	76.60	85.47	56.58	79.13	21.55
MISSFormer [12]	—	—	91.92	82.00	85.21	68.65	94.41	80.81	86.99	65.67	81.96	18.20
Swin UNETR [9]	62.83M	384.2	95.37	86.26	86.99	66.54	95.72	77.01	91.12	68.80	83.48	10.55
UNETR + MAE [28]	—	—	90.56	84.00	86.37	75.25	95.95	80.89	88.92	65.02	83.52	10.24
nnFormer [27]	150.5M	213.4	90.51	86.25	86.57	70.17	96.84	86.83	92.04	**83.35**	86.57	10.63
UNETR++ [18]	42.96M	47.98	**95.77**	87.18	87.54	71.25	96.42	86.01	92.52	81.10	87.22	7.53
MAE + UNETR++ (ours)	85.96M	82.49	95.68	**89.30**	**87.64**	**79.60**	**96.98**	**88.47**	**92.58**	81.27	**88.94** (1.72% ↑)	**5.89** (1.64 ↑)
p-values	$< 5e-2$ (Dice), $< 1e-2$ (HD95)											

datasets [2] for two different segmentation tasks, spleen segmentation and lung segmentation. For each experiment, we perform 5 runs using different random seeds and report the average results. Additionally, we compute p-values to ascertain the statistical significance of the results.

Synapse CT Dataset: This dataset [15] contains 30 abdominal CT volumes with 8 organs. Following [5], [24], we split the dataset randomly into 18 volumes and 12 volumes for training and testing, and report the average Dice and 95% Hausdorff distance (HD95) on 8 abdominal organs: spleen (Spl), right kidney (RKid), left kidney (LKid), gallbladder (Gal), liver (Liv), stomach (Sto), aorta (Aor), and pancreas (Pan).

BTCV CT Dataset:: This dataset [15] consists of 30 abdominal CT volumes with 13 organs, including 8 organs of the Synapse CT dataset, along with esophagus (Eso), inferior vena cava (IVC), portal and splenic veins (PSV), right adrenal gland (RAG), and left adrenal gland (LAG).

ACDC Dataset: This dataset [3] contains 100 samples, and aims to segment the cavity of the right ventricle, the myocardium of the left ventricle, and the cavity of the left ventricle. Each sample's labels involve left ventricle (LV), right ventricle (RV), and myocardium (MYO). Following [27], we split the dataset into 70 training samples, 10 validation samples, and 20 test samples.

MSD Spleen Dataset: This dataset [2] contains 41 CT volumes for spleen segmentation. Following [10], we split the dataset into training, validation, and test sets (80:15:5).

MSD Lung Dataset:: This dataset [2] comprises 64 CT volumes for lung cancer segmentation. We split the dataset with a 80:20 ratio for training and testing following [18].

B. Implementation Details

Our experiments are implemented with PyTorch and MONAI.The model training is performed on an NVIDIA Tesla V100 Graphics Card with 32GB GPU memory using the AdamW optimizer with a weight decay = 0.005.

For the Synapse CT and BTCV CT datasets, we clip the raw values between -175 and 250, normalize the values into the range of [0, 1], and re-sample the spacing to $[1.5, 1.5, 2.0]$. All the models are trained with input images of size $96 \times 96 \times 96$. For the ACDC dataset, we re-sample the spacing to $[1.52, 1.52, 6.35]$. All the models are trained with input of size $160 \times 160 \times 16$. For the MSD spleen dataset, we clip the raw values between -57 and 164, normalize the values into the range of [0, 1], and re-sample the spacing to $[1.5, 1.5, 2.0]$. All the models are trained with input of size $96 \times 96 \times 96$. For the MSD lung dataset, we clip the raw values between -1000 and 3071, normalize the values into the range of [0, 1], and re-sample the spacing to $[1.0, 1.0, 1.0]$. All the models are trained with input of size $192 \times 192 \times 32$.

We use a learning rate of $6.4e-3$ for self pre-training on all the datasets. We pre-train on the Synapse CT, BTCV CT, and MSD spleen and lung segmentation datasets with $10,000$ epochs, and on the ACDC dataset with $2,000$ epochs.

For all the downstream segmentation tasks, we use a learning rate of $1e-1$, and fine-tune with $5,000$ epochs for the Synapse CT, BTCV CT, and MSD spleen and lung segmentation datasets, and $1,000$ epochs for the ACDC dataset. The batch size for each case is set as the maximum size allowed by the GPU. We set $\lambda_1 = 0.5$, $\lambda_2 = 0.1$, and $\lambda_3 = 0.1$.

C. Experimental Results

Synapse CT Dataset Results. In Table I, we compare our method with an array of baseline methods (U-Net [17], TransUNet [5], UNETR [10], Swin-UNet [4], MISSFormer [12], Swin UNETR [9], and nnFormer [27]) and SOTA models (UNETR++ [18], and the MAE-based self pre-training method, i.e., UNETR + MAE [28]). On this dataset, UNETR++ yields superior performance over the other known methods. Our method outperforms UNETR++ by 1.72% and 1.64 mm in average Dice and HD95, respectively, which are quite impressive improvements on the Synapse CT dataset. Specifically, our method achieves the highest Dice scores on six organs (kidney (right), kidney (left), gallbladder, liver, stomach, and aorta). Compared to the known methods, our method is more advantageous in segmenting gallbladder, which is difficult to delineate using known segmentation methods. Our method is able to surpass the UNETR + MAE by large margins in both the evaluation metrics, demonstrating the effectiveness of our method (see Fig. 3).

BTCV CT Dataset Results. Table II showcases the segmentation results of various methods on the BTCV CT dataset. Among the known methods, nnU-Net [13] and UNETR++ [18] achieve average Dice scores of 83.16% and 83.28%, respectively. Our method outperforms the SOTA method UNETR++ by 0.8% in average Dice. This is particularly commendable given the challenging nature of the BTCV CT dataset, which encompasses 13 distinct organs.

TABLE II
SEGMENTATION RESULTS ON THE BTCV CT DATASET.

Method	Spl	RKid	LKid	Gal	Eso	Liv	Sto	Aor	IVC	PSV	Pan	RVG	LAG	Average Dice (↑)
UNETR [10]	90.48	82.51	86.05	58.23	71.21	94.64	72.06	86.57	76.51	70.37	66.06	66.25	63.04	76.00
Swin UNETR [9]	94.59	88.97	92.39	65.37	75.43	95.61	75.57	88.28	81.61	76.30	74.52	68.23	66.02	80.44
TransBTS [20]	94.55	89.20	90.97	68.38	75.61	96.44	83.52	88.55	82.48	74.21	76.02	67.23	67.03	81.31
nnFormer [27]	94.58	88.62	**93.68**	65.29	76.22	96.17	83.59	89.09	80.80	75.97	77.87	70.20	66.05	81.62
nnU-Net [13]	**95.95**	88.35	93.02	70.13	76.72	96.51	86.79	88.93	82.89	**78.51**	79.60	**73.26**	68.35	83.16
UNETR++ [18]	94.94	**91.90**	93.62	70.75	77.18	95.95	85.15	89.28	83.14	76.91	77.42	72.56	68.17	83.28
MAE + UNETR++ (ours)	94.97	87.93	87.37	**78.46**	**78.97**	**96.99**	**88.31**	**92.51**	**89.01**	76.94	**80.18**	69.88	**71.48**	**84.08** (0.8% ↑)
p-value	$< 1e-2$ (Dice)													

TABLE III
SEGMENTATION RESULTS ON THE ACDC DATASET.

Method	RV	Myo	LV	Average
VIT-CUP [6]	81.46	70.71	92.18	81.45
R50-VIT-CUP [6]	86.07	81.88	94.75	87.57
MISSFormer [12]	86.36	85.75	91.59	87.90
UNETR [10]	85.29	86.52	94.02	88.61
TransUNet [5]	88.86	84.54	95.73	89.71
Swin-UNet [4]	88.55	85.62	95.83	90.00
LeViT-UNet-384s [22]	89.55	87.64	93.76	90.32
UNETR + MAE [28]	— (88.44)	— (87.87)	— (94.58)	— (90.30)
nnFormer [27]	90.94	89.58	95.65	92.06
UNETR++ [18]	91.89	90.61	96.00	92.83
MAE + UNETR++ (ours)	**92.59**	**91.38**	**96.37**	**93.45** (0.62% ↑)
p-value	$< 1e-2$ (Dice)			

TABLE IV
SEGMENTATION RESULTS ON THE MSD SPLEEN DATASET.

Method	Dice (↑)	HD95 (↓)
SETR MLA [26]	0.950	4.091
TransUNet [5]	0.950	4.031
AttUNet [16]	0.951	4.091
U-Net [17]	0.953	4.087
CoTr [21]	0.954	3.860
UNETR [10]	0.964	1.333
UNETR + MAE [28]	— (0.966)	— (1.295)
UNETR++ [18]	0.966	1.246
MAE + UNETR++ (ours)	**0.974** (0.8% ↑)	**1.002** (0.244 ↑)
p-values	$< 1e-2$ (Dice), $< 5e-2$ (HD95)	

TABLE V
SEGMENTATION RESULTS ON THE MSD LUNG DATASET.

Method	Dice (↑)
UNETR [10]	73.29
nnU-Net [13]	74.31
Swin UNETR [9]	75.55
nnFormer [27]	77.95
UNETR + MAE [28]	— (78.90)
UNETR++ [18]	80.68
MAE + UNETR++ (ours)	**82.55** (1.87% ↑)
p-value	$< 5e-2$ (Dice)

ACDC Dataset Results. Table III reports the quantitative results on the ACDC dataset. We observe that nnFormer [27] and UNETR++ [18] attain better performances of 92.06% and 92.83% in average Dice, respectively. Remarkably, our method surpasses the SOTA method UNETR++ by 0.62% in the average Dice score. Furthermore, our method outperforms the MAE-based self pre-training method UNETR + MAE [28] by an impressive 3.15% in average Dice, confirming the effectiveness of our new approach.

MSD Spleen Dataset Results. As Table IV shows, on the MSD spleen dataset, both UNETR + MAE [28] and UNETR++ [18] already achieve a very high 0.966 Dice score, giving a limited margin for big improvement. Nonetheless, our method still manages to enhance both the Dice and HD95 scores by 0.8% and 0.244, respectively. These results reinforce the superiority of our method over known SOTA methods.

MSD Lung Dataset Results. Table V presents the experimental results on the MSD lung dataset. One can see that the best known method is UNETR++ [18], whose Dice score is higher than the second-best method, UNETR + MAE [28], by a margin of 1.78%. In comparison, our method outperforms UNETR++ by 1.87% and UNETR + MAE by a notable 3.65% in Dice score. These substantial improvements validate the effectiveness of our method.

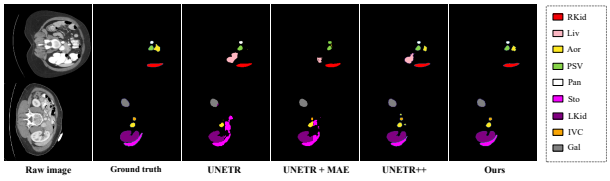

Fig. 3. Visual results of different methods on the Synapse CT dataset.

D. Ablation Study

We conduct ablation study using the Synapse CT dataset to examine the effects of different key components in our method. From the results in Table VI, we observe the following. (1) When applying the pre-trained ViT encoder on top of UNETR++, the Dice is improved by 0.4% (p-value = 0.027, t-test), showing the effect of our fine-tuned model. (2) When adding the spatial pre-text task to the pre-training of ViT only, the Dice is further improved by 0.47% (p-value = 0.009, t-test), demonstrating the effect of our proposed pre-text task in extracting spatial information. (3) When further adding the topological loss to the pre-training of ViT only, the Dice is further improved by 0.25% (p-value = 0.015, t-test), validating the effect of our proposed topological loss in capturing geometric shape information. (4) When transferring the weights of both the ViT encoder and UNETR++ to the fine-tuned model, the Dice is improved by 0.60% (p-value = 0.011, t-test), validating the effects of the MAE pre-training strategy on UNETR++ as well as the spatial and reconstruction consistency losses on co-pretraining ViT and UNETR++.

IV. CONCLUSIONS

In this paper, we proposed a novel extension of masked autoencoders (MAEs) for self pre-training (i.e., models pre-trained on the same target dataset) for 3D medical image segmentation. In particular, we proposed a new topological loss for extracting geometric shape information, introduced a pre-text task to aggregate global spatial information, extended the

TABLE VI
ABLATION STUDY OF THE EFFECTS OF DIFFERENT KEY COMPONENTS IN OUR MAE + UNETR++ APPROACH ON THE SYNAPSE CT DATASET.

Method	Pre-trained ViT Encoder	Spatial Pre-text Task	Topological Loss	Pre-trained UNETR++	Dice (↑)	HD95 (↓)
UNETR++					87.22	7.53
UNETR++ with pre-trained ViT encoder	√				87.62	7.03
UNETR++ with pre-trained ViT encoder & spatial pre-text task	√	√			88.09	6.12
UNETR++ with pre-trained ViT encoder & spatial pre-text task & topological loss	√	√	√		88.34	5.96
MAE + UNETR++	√	√	√	√	**88.94**	**5.89**

MAE pre-training strategy to a hybrid SOTA medical image segmentation architecture, and developed a fine-tuned model to further improve the downstream segmentation performance. Experimental results on five public 3D segmentation datasets demonstrated the effectiveness of our proposed approach.

V. ACKNOWLEDGEMENTS

This research was supported in part by NSF Grants CCF-2523787 and OAC-2104158.

REFERENCES

[1] Diego Adame, Jose A Nunez, Fabian Vazquez, Nayeli Gurrola, Huimin Li, Haoteng Tang, Bin Fu, and Pengfei Gu. Topo-VM-UNetV2: Encoding topology into vision Mamba UNet for polyp segmentation. In *CBMS*, pages 258–263, 2025.

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

[3] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Transactions on Medical Imaging*, 37(11):2514–2525, 2018.

[4] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-UNet: UNet-like pure Transformer for medical image segmentation. In *ECCV Workshops*, pages 205–218, 2023.

[5] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. TransUNet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[7] Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. Topological persistence and simplification. *Discrete & Computational Geometry*, 28:511–533, 2002.

[8] Pengfei Gu, Hongxiao Wang, Yejia Zhang, Huimin Li, Chaoli Wang, and Danny Chen. TopoImages: Incorporating local topology encoding into deep learning models for medical image classification. In *ACM MM*, pages 1938–1947, 2025.

[9] 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 *MICCAI Brainlesion Workshop*, pages 272–284, 2022.

[10] 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 *WACV*, pages 1748–1758, 2022.

[11] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022.

[12] Xiaohong Huang, Zhifang Deng, Dandan Li, and Xueguang Yuan. MISSFormer: An effective medical image segmentation Transformer. *arXiv preprint arXiv:2109.07162*, 2021.

[13] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021.

[14] Shizuo Kaji, Takeki Sudo, and Kazushi Ahara. Cubical Ripser: Software for computing persistent homology of image and volume data. *arXiv preprint arXiv:2005.12692*, 2020.

[15] Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, T Langerak, and Arno Klein. MICCAI multi-atlas labeling beyond the cranial vault – workshop and challenge. In *MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, page 12, 2015.

[16] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention U-Net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[17] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015.

[18] Abdelrahman Shaker, Muhammad Maaz, Hanoona Rasheed, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. UNETR++: Delving into efficient and accurate 3D medical image segmentation. *arXiv preprint arXiv:2212.04497*, 2022.

[19] Dominik JE Waibel, Scott Atwell, Matthias Meier, Carsten Marr, and Bastian Rieck. Capturing shape information with multi-scale topological loss terms for 3D reconstruction. In *MICCAI*, pages 150–159, 2022.

[20] Wenxuan Wang, Chen Chen, Meng Ding, Hong Yu, Sen Zha, and Jiangyun Li. TransBTS: Multimodal brain tumor segmentation using Transformer. In *MICCAI*, pages 109–119, 2021.

[21] Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. CoTr: Efficiently bridging CNN and Transformer for 3D medical image segmentation. In *MICCAI*, pages 171–180, 2021.

[22] Guoping Xu, Xingrong Wu, Xuan Zhang, and Xinwei He. LeVit-UNet: Make faster encoders with Transformer for medical image segmentation. *arXiv preprint arXiv:2107.08623*, 2021.

[23] Yejia Zhang, Pengfei Gu, Nishchal Sapkota, Hao Zheng, Peixian Liang, and Danny Z Chen. A point in the right direction: Vector prediction for spatially-aware self-supervised volumetric representation learning. In *ISBI*, 2023.

[24] Yejia Zhang, Nishchal Sapkota, Pengfei Gu, Yaopeng Peng, Hao Zheng, and Danny Z Chen. Keep your friends close & enemies farther: Debiasing contrastive learning with spatial priors in 3D radiology images. In *BIBM*, pages 1824–1829, 2022.

[25] Yizhe Zhang, Tao Zhou, Yuhui Tao, Shuo Wang, Ye Wu, Benyuan Liu, Pengfei Gu, Qiang Chen, and Danny Z Chen. TestFit: A plug-and-play one-pass test time method for medical image segmentation. *Medical Image Analysis*, 92:103069, 2024.

[26] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with Transformers. In *CVPR*, pages 6881–6890, 2021.

[27] 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.

[28] Lei Zhou, Huidong Liu, Joseph Bae, Junjun He, Dimitris Samaras, and Prateek Prasanna. Self pre-training with masked autoencoders for medical image analysis. In *ISBI*, 2023.