2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

3DME: Dual-Branch Encoder with Progressive Masking for 3D Medical Foundation Encoding Model

Hengyi Yuan[1], Zesheng Cheng[1*], Huiru Chen[1], Wang Shixuan[2]

Abstract— **Three-dimensional (3D) medical image analysis faces challenges such as massive data volume, difficulty in integrating cross-slice information, and limited model generalization. This paper proposes 3DME, a foundational model for 3D medical imaging. Its core innovations feature a dual-branch 3D encoder that integrates a Vision Transformer for modeling global long-range dependencies and a 3D graph convolutional network for capturing local voxel structures, enhanced by multi-level deformable attention for cross-planar correlation; a progressive volumetric masking strategy for self-supervised pretraining, which dynamically adjusts masking ratios and block sizes to force the model to learn cross-slice continuity and global semantics; and a unified foundation model framework supporting lightweight adaptation for downstream tasks. Experiments demonstrate that 3DME achieves state-of-the-art (SOTA) performance on 12 segmentation and classification tasks, exhibiting strong zero-shot transfer capabilities, thereby significantly enhancing model generalization and clinical deployment efficiency.**

I. Introduction

Medical image analysis has become a fundamental pillar of modern healthcare systems, playing a central role in disease diagnosis, treatment planning, medical education, and clinical research. High-precision imaging technologies enable accurate lesion localization, surgical navigation, and therapeutic efficacy assessment, thereby accelerating the development of precision medicine, particularly in oncology, neuroscience, and cardiovascular diseases. Compared with conventional two-dimensional imaging, three-dimensional (3D) volumetric modalities such as computed tomography (CT) and magnetic resonance imaging (MRI) provide richer anatomical continuity and spatial context. This volumetric information is essential for complex clinical tasks including organ reconstruction, tumor burden quantification, vascular analysis, and preoperative simulation. For example, 3D vascular reconstruction assists neurosurgeons in avoiding eloquent functional regions during surgical intervention [1], while precise coronary plaque localization directly influences interventional strategies such as stent placement and procedural optimization [2]. Crucially, in the rapidly advancing field of robotic-assisted surgery, high-fidelity 3D anatomical modeling is the absolute cornerstone for autonomous navigation, robotic trajectory planning, and precise intraoperative tissue manipulation. These advances highlight the indispensable value of volumetric imaging not only in conventional clinical workflows but also as the primary spatial perception modality for intelligent medical robots.

Despite these advantages, the massive voxel-level data inherent in 3D imaging introduces substantial challenges. High spatial resolution results in significant computational cost, storage demand, and memory consumption during model training and inference. Manual inspection of volumetric data is time-consuming and cognitively demanding, limiting scalability in routine clinical practice. Furthermore, multi-modality integration—such as combining CT, MRI, PET, and clinical metadata—requires models capable of learning coherent representations across heterogeneous data distributions. These issues hinder the full exploitation of volumetric imaging data. For surgical robots operating in complex, dynamic intraoperative environments, the need for efficient, scalable, and generalizable analytical frameworks is even more critical, as delayed or inaccurate spatial perception can directly compromise surgical safety and robotic control.

Traditional deep learning approaches, particularly 3D convolutional neural networks (3D-CNNs) [3], have demonstrated effectiveness in extracting volumetric features. However, such models are typically trained for narrowly defined tasks and rely heavily on large amounts of manually annotated data. Their task-specific nature limits generalization across institutions and modalities, and their representations often lack transferability. Recently, large-scale vision foundation models such as the Segment Anything Model (SAM) [4] have demonstrated impressive cross-task and cross-modal generalization ability in natural image domains. Nevertheless, directly extending 2D foundation paradigms to 3D medical data leads to prohibitive computational overhead due to cubic growth in spatial complexity. Moreover, medical images differ fundamentally from natural images: they often exhibit low contrast, subtle pathological cues, domain-specific semantics, inter-scanner variability, and inconsistent annotation standards. These intrinsic characteristics complicate direct adaptation of general vision foundation models and motivate the development of specialized foundation architectures tailored for 3D medical imaging and downstream robotic applications.

To address these limitations and empower robotic-assisted clinical applications, we propose **3DME**, a unified foundation model specifically designed for 3D medical

*Corresponding Author: czs_110@hotmail.com.
Hengyi Yuan[1], Zesheng Cheng[1], and Huiru Chen[1] are with College of Computer Science & Technology, Qingdao University, China. Wang Shixuan[2] is with the School of No.1 Middle School of Weifang, China.

This work is supported in part by Shandong Province Natural Science Foundation under Grant No.ZR2024MG034, No.ZR2024MF144 and No.ZR2024MF142, and Key Technology Research and Development Program of Shandong under Grant No.2025CXGC010108.

image analysis and surgical robotic perception. The proposed framework balances representational richness, computational efficiency, and cross-task generalization, making it highly suitable for the stringent perception demands of medical robotics. Our contributions are threefold:

- **Dual-Branch Hybrid Encoder.** We design a novel dual-branch architecture that integrates a Vision Transformer backbone for global contextual modeling with a 3D graph convolution module to capture structural and topological relationships. A deformable attention mechanism adaptively aggregates multi-scale spatial information, enabling efficient modeling of long-range dependencies while preserving fine-grained anatomical details required for precise robotic manipulation.
- **Progressive Volumetric Masking for Self-Supervised Pretraining.** We introduce a progressive masking strategy tailored to volumetric redundancy. Instead of naive random masking, our method gradually increases reconstruction difficulty across spatial hierarchies, encouraging the model to learn semantically meaningful anatomical representations and improving data efficiency under limited annotations.
- **Unified Cross-Task Validation and Strong Generalization.** We conduct comprehensive evaluation across ten heterogeneous 3D medical tasks, including segmentation, classification, and reconstruction benchmarks. 3DME consistently achieves state-of-the-art performance while demonstrating strong zero-shot and few-shot transfer capability. These capabilities provide highly reliable upstream spatial representations for downstream robotic surgical planning, autonomous navigation, and intelligent clinical decision-making.

Overall, 3DME bridges the gap between task-specific 3D models and emerging foundation paradigms, paving the way toward unified and scalable representation learning for volumetric medical image analysis and next-generation robotic surgery systems.

II. Related Work

Large-scale pre-trained models have recently revolutionized general vision and language tasks, spurring interest in applying them to medical imaging. Three-dimensional modalities (e.g., CT, MRI) provide rich cross-sectional context that is critical for diagnosis and planning. However, proposing large-scale models for volumetric data is challenging, and efforts in 3D medical foundation models are still nascent. Existing work spans diverse directions—model design, pretraining strategy, adaption. But no unified framework yet exists due to architectural fragmentation, limited pretraining objectives, and weak cross-task generalization.

Regarding model architectures, several recent studies propose 3D or hybrid designs. For example, Chen et al. [5]introduce MedBLIP, which uses a MedQFormer module to bridge 3D slice sequences with a frozen 2D image encoder and a language model. He et al.[6] present VISTA3D, a volumetric segmentation foundation model trained on $>$11K CT volumes; it achieves state-of-the-art automatic segmentation and also supports interactive correction in 3D. Similarly, BrainSegFounder[7] propose a two-stage 3D Transformer pretraining: the first stage encodes healthy anatomical structures, and the second stage incorporates spatial relations for lesion localization. While these works advance 3D processing, they generally either rely on processing 2D slices. As a result, they do not fully exploit multi-planar context across entire volumes, limiting their ability to integrate cross-slice information and adapt to multiple tasks.

In parallel, researchers have developed sophisticated self-supervised pretraining strategies for 3D data. Ye et al.[8] propose DeSD, a deep self-distillation scheme that trains multiple nested encoders: a student network's each layer is supervised by a momentum teacher, improving feature quality at both shallow and deep levels. Xie et al.[9] introduce ReFs, which incorporates a reference segmentation task during pretraining: by matching gradients to those of a downstream segmentation loss, ReFs steers representation learning toward downstream-friendly features. Qi et al.[10] present GMIM, which uses adaptive hierarchical masking and multi-scale reconstruction tasks: their framework learns tissue boundaries via dynamic masking networks and enforces consistency across low and high-level representations. These methods enhance 3D image representations, but they still lack explicit mechanisms to model 3D spatial continuity or dynamics across slices. In particular, most focus on improving feature locality or semantics within slices, rather than jointly capturing volumetric coherence over time or through space.

Furthermore, researchers have proposed specialized adaptation schemes for 3D medical imaging. MA-SAM[11] embeds lightweight 3D convolutional adapters into the 2D Segment Anything Model (SAM) so it can process slices in a coordinated way. Make-A-Volume[12] treats a 2D latent diffusion model as a backbone: it inserts volumetric layers into the 2D slice-by-slice generator and fine-tunes on paired 3D data, effectively extending 2D diffusion to consistent 3D volume synthesis. Notably, current solutions lack generalizable 3D foundational architectures capable of simultaneously supporting segmentation, classification, reconstruction, and generation tasks. Their modular designs impede cross-task knowledge transfer, failing to satisfy clinical demands for multifunctional collaborative analysis.

In summary, existing literature has made progress in architecture design, self-supervision, and tasks adaptation, but key gaps remain. Most approaches either treat 3D data as collections of 2D slices or are limited to one task, failing to capture holistic volumetric context or support task diversity. They often overlook temporal or spatial continuity across planes and rely on handcrafted pipelines for each function. 3DME is specifically designed to address these limitations. By using a true 3D dual-branch encoder (combining a Vision Transformer with a 3D graph module) and volumetric

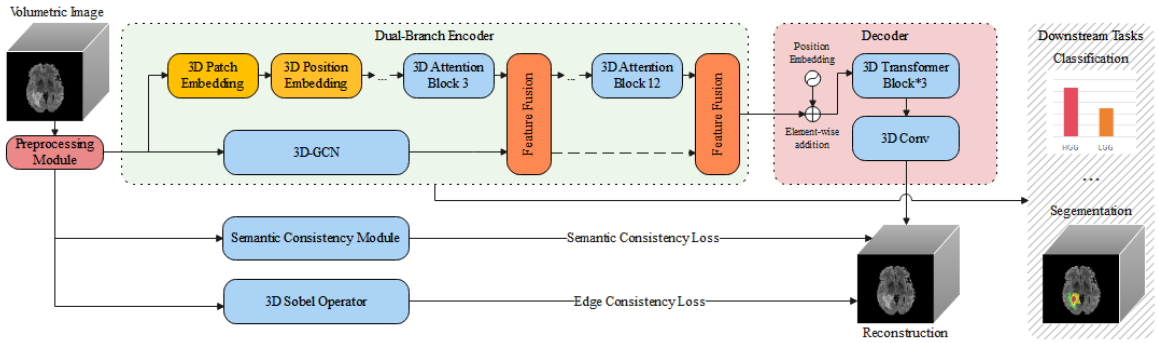

Fig. 1. Overall structure of the proposed 3DME model, including pre-training and downstream stages.

masked modeling, 3DME learns global 3D structure and cross-slice relationships. Its self-supervised objectives (masked reconstruction with semantic and edge consistency losses) enforce coherence in both appearance and topology across the volume. Consequently, 3DME serves as a unified 3D foundation model: it can be fine-tuned for different tasks while leveraging shared volumetric representations, overcoming the fragmentation and single-task focus of prior methods.

III. Methodologies and Structure of the Models

The 3DME model architecture comprises three core modules: the input module, the encoder module, and the decoder module. During pre-training, the input module standardizes image resolutions and performs masking operations, while during fine-tuning and inference, it solely standardizes resolutions. The encoder employs a dual-branch parallel architecture: the ViT branch extracts local spatial features while The 3D-GCN branch captures voxel-level spatial structural information. The outputs from both branches are fused multiple times during the training process after feature alignment. The decoder, based on a ViT architecture, reconstructs the fused features. Pre-training involves self-supervised learning by minimizing the masked region reconstruction loss, semantic consistency loss, and edge consistency loss. During fine-tuning and inference for downstream tasks, the encoder remains fixed for feature extraction, and the decoder is replaced with task-specific head modules such as segmentation or classification. Fig. 1 below illustrates the overall structure of 3DME. The following sections will provide a detailed elaboration on these three modules.

A. Input Module

The input module standardizes and applies masking to 3D medical images. During pre-training, it consists of two sequential submodules. First, all volumes are uniformly resampled to $128\times128\times128$ to ensure spatial consistency. Then, a random block masking strategy is applied to the voxel grid to construct prediction targets for self-supervised learning (see Section IV). During fine-tuning and inference, only the resampling operation is retained to maintain distribution alignment with pre-training, while masking is removed to support supervised or zero-shot downstream tasks.

B. Encoder Module

The encoding module comprises two branches and takes $128\times128\times128\times1$ grayscale images as input.

The ViT branch computes patch and position embeddings. The input is divided into $8\times8\times8$ patches and flattened into 4096×512. After projection by a learnable matrix $\mathbf{W}_E \in \mathbb{R}^{512\times768}$, patch embeddings $E\in\mathbb{R}^{4096\times768}$ are obtained.

Position embeddings adopt separable 3D encoding:

$$\mathrm{PE3D}(x,y,z)_k = \sum_{d\in\{x,y,z\}} \begin{cases} \sin(d\,w_k), & k \text{ even}, \\ \cos(d\,w_k), & k \text{ odd}. \end{cases} \tag{1}$$

where $w_k = 10000^{-2i/768}$, $i\in[0,383]$.

The Transformer input is $z_0 = E + PE_{3D}$. 3DME contains 12 encoder layers:

$$z'_L = \mathrm{MSA}\left(\mathrm{LayerNorm}(z_{L-1})\right) + z_{L-1} \tag{2}$$

$$z_L = \mathrm{MLP}\left(\mathrm{LayerNorm}(z'_L)\right) + z'_L \tag{3}$$

where MSA denotes multi-head self-attention.

The 3D convolutional branch includes three $3\times3\times3$ Conv3D layers:

$$\begin{aligned} \tilde{F}^L &= \mathrm{Conv3D}_{3\times3\times3}(F^L;\,\mathrm{stride}=2), \\ F^{L+1} &= \mathrm{ReLU}\left(\mathrm{BatchNorm3D}(\tilde{F}^L)\right) \end{aligned} \tag{4}$$

with $F^0 = I_{\text{input}}$. The output is a $16\times16\times16\times256$ feature map. A fourth Conv3D layer aligns channels to 768 for fusion:

$$F^4 = \mathrm{Conv3D}_{3\times3\times3}(F^3) \tag{5}$$

The convolutional features are fused with outputs from ViT layers 3, 6, 9, and 12, and then fed to layers 4, 7, 10 and the decoder.

Reshaping the ViT output yields $Z\in\mathbb{R}^{16\times16\times16\times768}$. Two deformable cross-attention operations are performed. First:

$$\hat{F} = F + \mathrm{DeformAttn}(\mathrm{norm}(F),\mathrm{norm}(Z)) \tag{6}$$

Then:

$$\hat{Z} = Z + \mathrm{DeformAttn}(\mathrm{norm}(Z),\mathrm{norm}(\hat{F})) \tag{7}$$

Deformable attention is defined as:

$$\text{DeformAttn}(Q,K,V) = W\left[\sum_{k=1}^{K} A_k \cdot V(p+\Delta p_k)\right] \quad (8)$$

The fused features pass through a residual feed-forward network and a fully connected layer, then are reshaped to $Z_{\text{enc}} \in \mathbb{R}^{4096\times768}$ for subsequent ViT layers.

The dual-branch architecture enables complementary modeling of long-range dependencies (ViT) and voxel-level spatial structures (3D convolution), while multi-level deformable attention enhances cross-dimensional feature interaction in volumetric medical images.

C. Decoder Module

The decoding module comprises a three-layer Vision Transformer (ViT), forming a symmetric inverse architecture to the encoder. The encoded tokens are combined with 3D positional embeddings, $Z'_{\text{enc}} = Z_{\text{enc}} + PE_{3\text{D}}$, and processed by three Transformer decoder layers (Eq. (2)). The outputs are linearly projected via $W_D \in \mathbb{R}^{512\times768}$ and reshaped into $V_{\text{resh}} \in \mathbb{R}^{16\times16\times16\times512}$, then rearranged and reconstructed to produce $V_{\text{patch}} \in \mathbb{R}^{128\times128\times128}$. A final $1\times1\times1$ 3D convolution yields $J_{\text{out}} = \text{Conv3D}_{1\times1\times1}(V_{\text{patch}})$, which is optimized with the reconstruction loss.

Details of loss function will be described in Section 4. Pre-training Pipeline.

IV. Pre-training Pipeline

The objective of the pre-training phase is to enable the 3DME model to learn robust and generalizable representations for 3D medical images. The pre-training process utilizes approximately 20,000 3D medical images. Training environment contains 8×NVIDIA 4090 GPUs with a single-node training strategy and a batch size of 16. The pipeline comprises two core components: Data Preprocessing with Masking Strategy and Loss Function Design.

A. Data Preprocessing and Masking Strategy

To facilitate self-supervised learning (Fig. 2), raw 3D medical images are first spatially standardized. All volumes are resampled via trilinear interpolation to isotropic voxel spacing of 1mm × 1mm × 1mm, eliminating anisotropy and scale variations introduced by different imaging protocols while preserving physically consistent spatial relationships. The volumes are then adjusted to a fixed resolution of $128\times128\times128$: if any dimension is smaller than 128, upsampling is applied; otherwise, a contiguous $128\times128\times128$ region is randomly cropped. Finally, a validity check is performed to filter out noise-dominated samples, where image blocks containing more than 90% background voxels are discarded.

To encourage structural and semantic understanding of volumetric data, 3DME adopts a self-supervised pre-training strategy based on progressive random block masking tailored to 3D characteristics. Masking is performed on standardized $128\times128\times128$ volumes using non-overlapping cubic blocks with dynamically increasing size and ratio. Specifically, pre-training begins with small blocks ($4\times4\times4$) and a low masking ratio (30%) to emphasize local detail reconstruction. As training proceeds, block sizes are expanded to $8\times8\times8$ and $12\times12\times12$, while the masking ratio increases to 45% and 60%. In the final stage, large blocks ($16\times16\times16$) with a high masking ratio (75%) are applied, forcing the model to leverage long-range contextual dependencies and 3D spatial continuity for voxel prediction. This easy-to-hard curriculum mitigates cold-start instability, progressively enhances global representation learning, and enables efficient extraction of semantically rich features from high-dimensional volumetric data.

B. Loss Function Design

The pre-training stage adopts a triple-loss joint optimization strategy including masked reconstruction loss, semantic consistency loss, and edge consistency loss:

$$L_{\text{total}} = \lambda_{\text{rec}}L_{\text{rec}} + \lambda_{\text{sem}}L_{\text{sem}} + \lambda_{\text{edge}}L_{\text{edge}} \quad (9)$$

where $\lambda_{\text{rec}} = 0.7$, $\lambda_{\text{sem}} = 0.2$, and $\lambda_{\text{edge}} = 0.1$.

Masked reconstruction loss L_{rec} is applied only to masked voxels $\mathscr{M}$, computing the MSE between decoder output J_{out} and original image J, enforcing accurate intensity recovery:

$$L_{\text{rec}} = \frac{1}{2|\mathscr{M}|}\sum_{i\in\mathscr{M}}\left\|J_{\text{out}}^{(i)} - J^{(i)}\right\|^2 \quad (10)$$

Semantic consistency loss L_{sem} aligns masked and unmasked regions in high-level feature space using a pretrained segmentation model φ:

$$L_{\text{sem}} = 1 - \frac{1}{|\mathscr{M}|}\sum_{i\in\mathscr{M}}\frac{\varphi(J_{\text{out}}^{(i)})\cdot\varphi(J^{(i)})}{\|\varphi(J_{\text{out}}^{(i)})\|\,\|\varphi(J^{(i)})\|} \quad (11)$$

Edge consistency loss L_{edge} extracts edge maps via a 3D Sobel operator S and computes the Dice coefficient to preserve structural continuity:

$$L_{\text{edge}} = 1 - \frac{2\sum S(J_{\text{out}})\circ S(J)}{\sum S(J_{\text{out}}) + \sum S(J)} \quad (12)$$

This joint optimization enhances voxel-level reconstruction, semantic coherence, and structural integrity, yielding robust representations for downstream tasks.

V. Experiment and Analysis

To verify the effectiveness and generalizability of the proposed model, extensive experiments were conducted on multiple widely-used medical image segmentation and classification datasets. Same as pre-training process, the experiments were carried out in an environment consisting of a single-node training setup with 8 * NVIDIA 4090 GPUs and a batch size of 16. The following sections will provide a detailed description of the experimental setup, evaluation metrics, and results from fine-tuning the model on various task-specific datasets. The advantages and limitations of the proposed approach will be analyzed and compared with existing state-of-the-art methods.

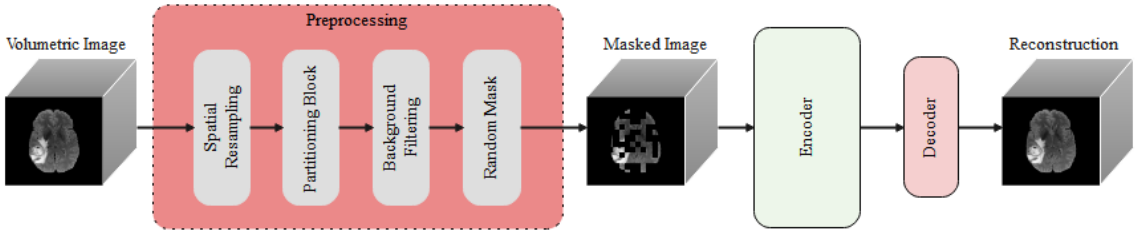

Fig. 2. Data pre-processing in pre-training process

TABLE I

QUANTITATIVE COMPARISON ON 3D MEDICAL IMAGE SEGMENTATION BENCHMARKS. HIGHER DICE AND LOWER HD95 INDICATE BETTER PERFORMANCE.

Dataset	Metric	nnU-Net	TransBTS	UNETR	SwinUNETR	3D UX-Net	MedNeXt	3DME-S0	**3DME-S10**
2*BraTS	Dice ↑	0.659	0.690	0.723	0.892	0.867	0.838	0.815	**0.907**
	HD95 ↓	14.537	9.655	7.986	5.212	5.199	5.220	6.111	**5.033**
2*MSD Liver	Dice ↑	0.649	0.683	0.724	0.778	0.768	0.760	0.703	**0.781**
	HD95 ↓	13.291	13.006	9.204	7.758	7.617	7.354	8.122	**7.293**
2*MSD Hippocampus	Dice ↑	0.696	0.710	0.788	0.872	0.890	0.885	0.798	**0.898**
	HD95 ↓	12.298	11.046	8.008	6.023	6.033	5.987	7.226	**6.012**
2*MSD Pancreas	Dice ↑	0.669	0.646	0.733	0.761	0.766	0.797	0.720	**0.813**
	HD95 ↓	13.690	10.590	7.996	5.443	5.459	5.416	5.971	**5.311**
2*BTCV	Dice ↑	0.624	0.638	0.723	0.796	0.791	0.790	0.729	**0.804**
	HD95 ↓	12.361	9.450	8.694	8.331	6.657	6.053	8.911	**6.028**
2*FeTA 2021	Dice ↑	0.765	0.774	0.852	0.850	0.879	0.872	0.822	**0.888**
	HD95 ↓	14.702	8.301	6.527	5.330	5.390	5.545	7.001	**5.293**

A. 3D Medical Image Segmentation Task

For 3D medical image segmentation task, this study replaces the decoder module applied in pre-training process with the 3D U-Net Decoder[13]. The experiments implement two fine-tuning strategies: zero-shot transfer without fine-tuning (3DME-S0) and 10-epoch fine-tuning (3DME-S10) to evaluate the adaptability of the pre-trained encoder.

Datasets: BraTS[14]: Brain Tumor Segmentation dataset featuring multi-modal MRI images. Targets: segmentation of tumor core, enhancing regions, and whole tumor. MSD Liver[15]: Liver and liver tumor segmentation based on CT imaging. MSD Hippocampus: Hippocampus segmentation using MRI data. MSD Pancreas: Pancreas and pancreatic tumor segmentation from CT images. Challenges: Small target volumes with ambiguous annotations. BTCV[16]: Abdominal multi-organ segmentation CT dataset covering 13 organs. FeTA 2021[17]: Fetal brain segmentation MRI dataset.

Baselines: Baselines: This study employs six methods as baseline models: nnU-Net[18], SwinUNETR[19], 3D UX-Net[20], MedNeXt.M.K3[21], UNETR[22], and TransBTS[23].

Evaluation Matrix: The experiment adopted the mean DICE coefficient and HD95 values commonly used in segmentation tasks to evaluate the segmentation accuracy and boundary error of the method, respectively.

Quantitative Experiment

As shown in TABLE I, in the systematic evaluation of 3D medical image segmentation tasks, the 3DME model demonstrates advantages in performance and generalization capability. Without fine-tuning (3DME-S0), it significantly outperforms traditional models in zero-shot scenarios: On the BraTS dataset, it achieves a DICE coefficient of 0.815, surpassing nnU-Net (0.659) and TransBTS (0.690). In the MSD Hippocampus task, it exceeds UNETR (0.788) which requires full annotation for training. This validates the pre-trained encoder's strong ability to capture anatomical structural features, enabling it to independently solve complex segmentation problems without downstream annotations.

After 10 rounds of fine-tuning (3DME-S10), the model achieves comprehensive performance improvements and surpasses current SOTA methods across multiple tasks: For instance, it reaches 0.907 DICE on the BraTS task, exceeding the previous best model SwinUNETR (0.892). In the BTCV multi-organ segmentation task, the DICE score improves to 0.804. Its HD95 error also decreases significantly, dropping to 5.311mm in the MSD Pancreas task (outperforming MedNeXt's 5.416mm), demonstrating refined segmentation capabilities for fine structures and highlighting the robustness of 3D continuous modeling for small-volume targets like the pancreas. Compared to existing methods, 3DME-S10 achieves optimal results in 5 out of 6 datasets while exhibiting significantly higher training efficiency than models requiring training from scratch (e.g., nnU-Net). Notably, the zero-shot performance of 3DME-S0 already rivals most traditional supervised models,

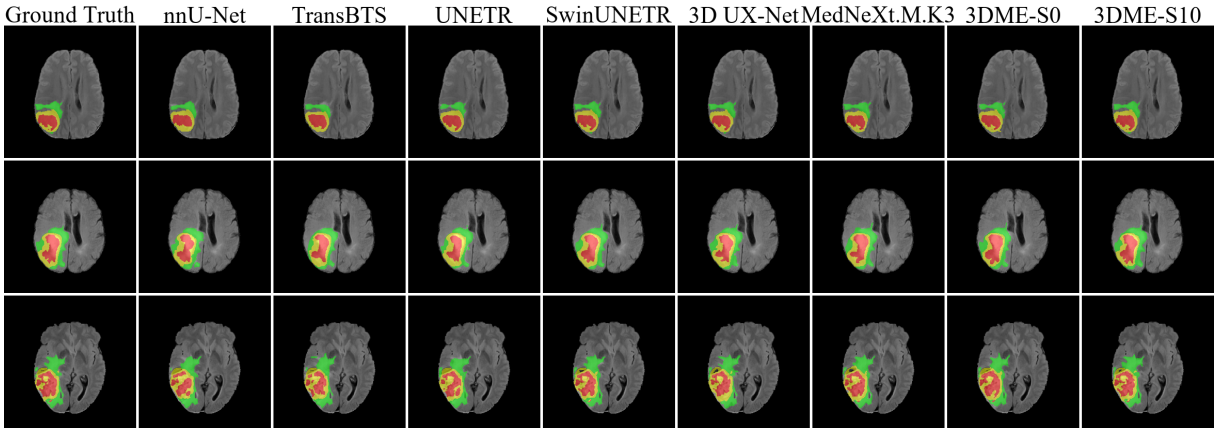

Fig. 3. Qualitative analysis of segmentation effects (BraTs)

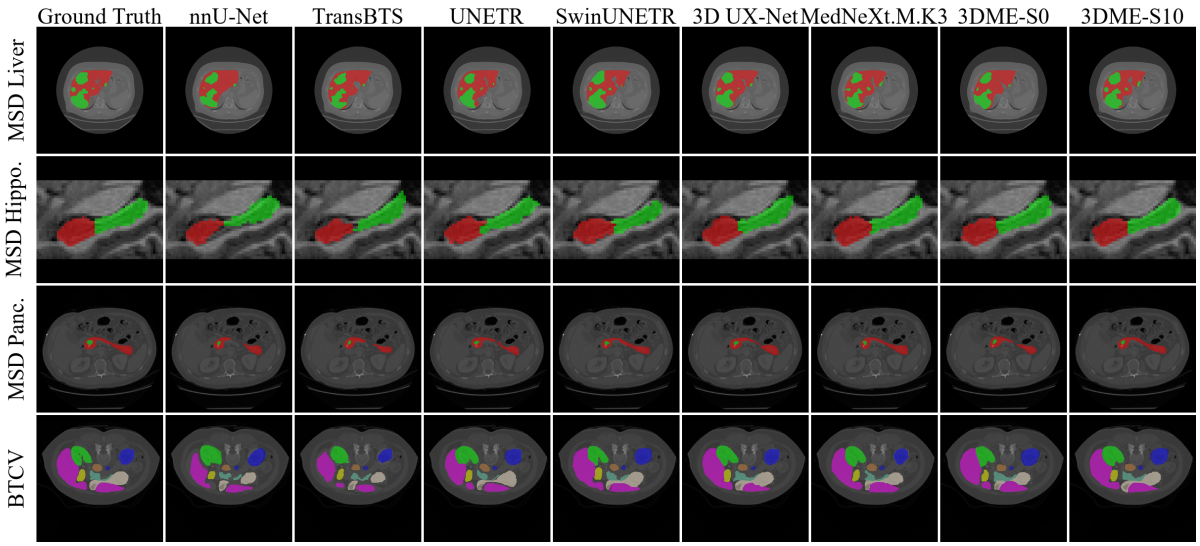

Fig. 4. Qualitative analysis of segmentation effects (MSD Liver, MSD Hippocampus, MSD Pancreas, BTCV)

and lightweight fine-tuning enables substantial performance leaps. This provides significant value for clinical deployment.

Qualitative Analysis

As shown in Fig. 3 and 4, the segmentation results of the 3DME model, particularly 3DME-S10, are visually superior to baseline methods. In the BraTS task, as shown in Fig. 3, 3DME-S10 demonstrates more precise, continuous, and ground-truth-aligned boundary delineations for the brain tumor core (red), enhancing regions (yellow), and edema regions (green) - especially in subtle structures at tumor margins and areas of internal heterogeneity. In contrast, segmentation results from other models (e.g., nnU-Net, TransBTS) often exhibit blurred boundaries, regional discontinuities (e.g., holes), or missed and over segmented subtle lesions.

In other challenging datasets such as that shown in Fig. 4, 3DME-S10 also achieves higher segmentation integrity and morphological fidelity for various organs and lesions, effectively overcoming ambiguities in annotations. These visualizations intuitively confirm the superiority of the quantitative metrics, highlighting the robust capability of 3DME's dual-branch encoder and progressive masking strategy in capturing 3D continuous anatomical structures and fine-grained boundary information. This lends greater reliability to clinical diagnoses.

B. 3D Medical Image Recognition Task

For 3D medical image recognition task, this study replaces the decoder module applied in pre-training process by the 3D Vision Transformer with Attention Enhancement. Same as segment task, the experiments implement two fine-tuning strategies: zero-shot transfer without fine-tuning (3DME-R0) and 10-epoch fine-tuning (3DME-R10) to evaluate the adaptability of the pre-trained encoder.

Datasets: BraTS 2018[24]: Brain Tumor dataset featuring multi-modal MRI images. Targets: Classification of tumor grades or survival prediction. ADNI[25]: Alzheimer's Disease Neuroimaging Initiative dataset based on brain MRI. Targets: Classification of Alzheimer's Disease (AD), Mild Cognitive Impairment (MCI), and Cognitively Normal

TABLE II

QUANTITATIVE COMPARISON ON 3D MEDICAL IMAGE RECOGNITION BENCHMARKS. HIGHER VALUES INDICATE BETTER PERFORMANCE.

Dataset	**Metric**	FSU	UNet3D	UniFormer	MAE3D	SimMM	MoCov3	3DME-R0	**3DME-R10**
3*BraTS 2018	ACC ↑	**0.798**	0.732	0.778	0.696	0.743	0.732	0.712	0.751
	AUC ↑	0.784	0.743	0.779	0.674	0.825	0.818	0.755	**0.852**
	F1 ↑	0.655	0.426	0.703	0.558	0.699	0.731	0.538	**0.746**
3*ADNI	ACC ↑	0.573	0.574	0.556	0.582	.629	0.602	0.577	**0.631**
	AUC ↑	0.610	0.490	0.641	0.545	0.605	0.567	0.603	**0.666**
	F1 ↑	0.565	0.365	0.555	0.562	0.536	**0.601**	0.508	0.599
3*ADHD-200	ACC ↑	**0.671**	0.647	0.611	0.615	0.633	0.619	0.633	0.652
	AUC ↑	0.654	0.690	0.647	0.644	0.665	0.622	0.642	**0.682**
	F1 ↑	0.620	0.426	0.581	0.594	0.581	0.601	0.546	**0.644**
3*ABIDE-I	ACC ↑	0.611	0.611	0.588	0.615	0.537	0.591	0.585	**0.617**
	AUC ↑	0.543	0.498	0.447	0.493	0.581	0.633	0.609	**0.635**
	F1 ↑	0.557	0.459	0.426	0.459	0.530	0.594	0.435	**0.616**

(CN) subjects. ADHD-200[26]: Brain resting-state fMRI and structural MRI dataset. Targets: Classification of Attention Deficit Hyperactivity Disorder (ADHD) versus healthy controls. ABIDE-I[27]: Autism Brain Imaging Data Exchange dataset. Targets: Classification of Autism Spectrum Disorder (ASD) versus typical controls.

Baselines: This study employs six methods as baseline models: FromScratch UNETR (FSU), UNET3D[28], UniFormer[29], MAE3D[30], [31], SimMM[32], and MoCov3[33].

Evaluation Matrix: The experiment adopted the accuracy (ACC), area under the ROC curve (AUC) and F1 score (F1) which are commonly used in image recognition tasks to evaluate the performance of each method.

Quantitative Experiment

As presented in TABLE II, consistent with the findings in medical image segmentation, the zero-shot 3DME-R0 model exhibits robust intrinsic task-solving capabilities without any downstream fine-tuning. This underscores the pre-trained encoder's effectiveness in capturing discriminative anatomical features. With merely 10 epochs of lightweight fine-tuning, 3DME-R10 achieves substantial performance gains, surpassing state-of-the-art baselines in 9 out of the 12 evaluated metrics. These results confirm that the proposed dual-branch encoder and progressive masking strategy successfully model continuous 3D semantics, delivering a highly adaptable, high-performance solution well-suited for resource-constrained clinical applications.

VI. CONCLUSION AND FUTURE WORK

This paper presents 3DME, a unified foundation framework for 3D medical image analysis that integrates a dual-branch encoder architecture—leveraging the complementary strengths of Vision Transformers (ViT) and 3D Graph Convolutional Networks (3D-GCN)—with a progressive masked pre-training strategy. This synergy facilitates the joint modeling of global long-range dependencies and local voxel-level spatial topologies, yielding robust representations that significantly alleviate the dependency on large-scale expert annotations. Extensive evaluations across ten heterogeneous benchmarks underscore 3DME's competitive edge over existing state-of-the-art methods. Notably, its remarkable zero-shot transferability highlights the encoder's inherent adaptability, offering substantial promise for streamlining clinical workflows.

Despite these advancements, several avenues for improvement remain. First, the high-resolution nature of volumetric data poses challenges regarding GPU memory efficiency and computational latency, hindering deployment in real-time, resource-limited clinical settings. Future research will explore lightweight architectural paradigms—including token pruning, channel compression, and linear-complexity sequence modeling—to enhance scalability. Second, while 3DME achieves superior visual generalization, its capacity for multimodal synergy remains nascent. Recognizing that clinical decisions derive from diverse data streams, we aim to develop cross-modal fusion mechanisms to integrate 3D imaging with textual diagnostics and real-time surgical telemetry.

Furthermore, the fine-grained segmentation of minuscule or low-contrast anatomical structures remains a non-trivial task. Precise boundary delineation is not only vital for early diagnostics but also fundamental for high-precision interventions, particularly in robotic-assisted surgery. In these scenarios, providing medical robots with real-time, high-fidelity spatial awareness of lesions and the surrounding vasculature is critical for safe autonomous navigation and preoperative planning. To address this, we will investigate anatomy-informed masking and adaptive uncertainty modeling to enforce structural consistency across complex anatomical variations.

In conclusion, 3DME establishes a scalable and robust foundation for volumetric representation learning. By advancing computational efficiency and multimodal integration, we believe such 3D foundation models will serve as the "perception backbone" for the next generation of medical robots, ultimately elevating surgical automation, reducing clinician cognitive load, and optimizing patient outcomes within the digital medicine landscape.

REFERENCES

[1] Isikay, I., et al., Narrative review of patient-specific 3D visualization and reality technologies in skull base neurosurgery: enhancements in surgical training, planning, and navigation. Frontiers in Surgery, 2024. 11: p. 1427844.

[2] Shlofmitz, E., et al., Intravascular imaging-guided percutaneous coronary intervention: a universal approach for optimization of stent implantation. Circulation: Cardiovascular Interventions, 2020. 13(12): p. e008686.

[3] Tran, D., et al. Learning spatiotemporal features with 3d convolutional networks. in Proceedings of the IEEE international conference on computer vision. 2015.

[4] Kirillov, A., et al. Segment anything. in Proceedings of the IEEE/CVF international conference on computer vision. 2023.

[5] Chen, Q. and Y. Hong. Medblip: Bootstrapping language-image pre-training from 3d medical images and texts. in Proceedings of the Asian Conference on Computer Vision. 2024.

[6] He, Y., et al., Vista3d: Versatile imaging segmentation and annotation model for 3d computed tomography. arXiv preprint arXiv:2406.05285, 2024.

[7] Cox, J., et al., BrainSegFounder: towards 3D foundation models for neuroimage segmentation. Medical Image Analysis, 2024. 97: p. 103301.

[8] Ye, Y., et al. Desd: Self-supervised learning with deep self-distillation for 3d medical image segmentation. in International Conference on Medical Image Computing and Computer-Assisted Intervention. 2022. Springer.

[9] Xie, Y., et al., ReFs: A hybrid pre-training paradigm for 3D medical image segmentation. Medical Image Analysis, 2024. 91: p. 103023.

[10] Qi, L., et al., GMIM: self-supervised pre-training for 3D medical image segmentation with adaptive and hierarchical masked image modeling. Computers in Biology and Medicine, 2024. 176: p. 108547.

[11] Chen, C., et al., Ma-sam: Modality-agnostic sam adaptation for 3d medical image segmentation. Medical Image Analysis, 2024. 98: p. 103310.

[12] Zhu, L., et al. Make-a-volume: Leveraging latent diffusion models for cross-modality 3d brain mri synthesis. in International Conference on Medical Image Computing and Computer-Assisted Intervention. 2023. Springer.

[13] Çiçek, Ö., et al. 3D U-Net: learning dense volumetric segmentation from sparse annotation. in Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19. 2016. Springer.

[14] Menze, B.H., et al., The multimodal brain tumor image segmentation benchmark (BRATS). IEEE transactions on medical imaging, 2014. 34(10): p. 1993–2024.

[15] Antonelli, M., et al., The medical segmentation decathlon. Nature communications, 2022. 13(1): p. 4128.

[16] Landman, B., et al. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. in Proc. MICCAI multi-atlas labeling beyond cranial vault—workshop challenge. 2015. Munich, Germany.

[17] Payette, K., et al., An automatic multi-tissue human fetal brain segmentation benchmark using the fetal tissue annotation dataset. Scientific data, 2021. 8(1): p. 167.

[18] Isensee, F., et al., nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods, 2021. 18(2): p. 203–211.

[19] Hatamizadeh, A., et al. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. in International MICCAI brainlesion workshop. 2021. Springer.

[20] Lee, H.H., et al., 3d ux-net: A large kernel volumetric convnet modernizing hierarchical transformer for medical image segmentation. arXiv preprint arXiv:2209.15076, 2022.

[21] Roy, S., et al. Mednext: transformer-driven scaling of convnets for medical image segmentation. in International Conference on Medical Image Computing and Computer-Assisted Intervention. 2023. Springer.

[22] Hatamizadeh, A., et al. Unetr: Transformers for 3d medical image segmentation. in Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2022.

[23] Wenxuan, W., et al. Transbts: Multimodal brain tumor segmentation using transformer. in International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. 2021.

[24] Bakas, S., et al., Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. arXiv preprint arXiv:1811.02629, 2018.

[25] Jack Jr, C.R., et al., The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine, 2008. 27(4): p. 685–691.

[26] consortium, A.-. The ADHD-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience. Frontiers in systems neuroscience, 2012. 6: p. 62.

[27] Di Martino, A., et al., The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Molecular psychiatry, 2014. 19(6): p. 659–667.

[28] Ronneberger, O., P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. in Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. 2015. Springer.

[29] Li, K., et al., Uniformer: Unifying convolution and self-attention for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023. 45(10): p. 12581–12600.

[30] Chen, Z., et al. Masked image modeling advances 3d medical image analysis. in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023.

[31] He, K., et al. Masked autoencoders are scalable vision learners. in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

[32] Xie, Z., et al. Simmim: A simple framework for masked image modeling. in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

[33] Chen, X., S. Xie, and K. He. An empirical study of training self-supervised vision transformers. in Proceedings of the IEEE/CVF international conference on computer vision. 2021.