GeoSFLoRA: Geometry-Conditioned Spectral Flow Low-Rank Adaptation for 2D-to-3D Transfer in Medical Image Segmentation

Qin Hao[1] , **Bonian Chen**[1] , **Shengwei Tian**[2] and **Long Yu**[3]
[1]School of Computer Science and Technology, Xinjiang University, Urumqi 830046, China
[2]School of Software, Xinjiang University, Urumqi 830046, China
[3]Network and Information Technology Center, Xinjiang University, Urumqi 830046, China
haoqin@stu.xju.edu.cn, chenbn266@163.com, tianshengwei@163.com, yul_xju@163.com

Abstract

Accurate volumetric medical image segmentation is critical for clinical diagnosis, yet adapting 2D vision foundation models (VFMs) to three-dimensional medical imaging remains challenging. While parameter-efficient fine-tuning (PEFT) methods such as LoRA and adapter-based schemes provide efficient alternatives to full fine-tuning, their geometry-agnostic parameter-space adaptations are insufficient to reconcile discrepancies induced by anisotropic voxel spacing and inter-slice discontinuities. We propose Geometry-Conditioned Spectral Flow Low-Rank Adaptation (GeoSFLoRA), a geometry-constrained framework for efficient 2D-to-3D transfer learning. GeoSFLoRA adapts frozen 2D pretrained vision backbones by operating within a fixed low-rank spectral subspace induced by pretrained linear projections. For each adapted layer, a truncated singular value decomposition is computed once and kept frozen, preserving the original pretrained bases. A lightweight Geometry-Conditioned Encoder (GCE) extracts local volumetric descriptors, which are mapped to token-conditional residuals in the singular-value space, enabling bounded and geometry-aware spectral modulation. GeoSFLoRA consistently improves Dice and HD95 over standard PEFT baselines on BraTS20, MSD-Prostate, and MSD-Lung, approaching full fine-tuning performance and demonstrating an effective paradigm for 2D-to-3D medical image segmentation. The code is publicly available at https://github.com/chenbn266/GeoSFLoRA.

1 Introduction

The advent of large scale Vision Foundation Models (VFMs), such as DINOv3 [Siméoni *et al.*, 2025] and the SAM series [Mazurowski *et al.*, 2023; Li *et al.*, 2025], has fundamentally reshaped visual representation learning. Pretrained on massive collections of 2D natural images, these models acquire universal semantic priors that exhibit strong transferability across a wide range of downstream tasks [Awais *et al.*, 2025; Jiao *et al.*, 2025]. Despite this success, extending such models to three-dimensional (3D) medical image segmentation remains challenging. The core difficulty lies in the structural divergence between domains. Natural images are dominated by surface appearance, texture, and color distributions, whereas 3D medical volumes (for example, CT or MRI) are characterized by inter-slice continuity, cross-sectional correlations, and highly anisotropic geometric structures [Litjens *et al.*, 2017; Wu *et al.*, 2025a; Li *et al.*, 2024].

This dimensional mismatch induces a pronounced representation shift that straightforward adaptation strategies struggle to reconcile. Naive approaches, including slice-by-slice 2D inference or the insertion of generic 3D convolutional or inflated adapters, often disrupt volumetric coherence and lead to inter-slice inconsistencies, manifesting as partial-volume artifacts [Wang *et al.*, 2024; Xu *et al.*, 2023; Hung *et al.*, 2024]. Even recent domain-specific adaptations, such as Medical SAM Adapter [Wu *et al.*, 2025b] or 3D SAM Adapter [Gong *et al.*, 2023], face difficulties in fully resolving the semantic gap without incurring significant computational costs or compromising the pretrained manifold structure [VanBerlo *et al.*, 2024].

To alleviate the cost of full retraining [Wang *et al.*, 2025], Parameter-Efficient Fine-Tuning (PEFT) has emerged as a compelling alternative. By introducing low-rank updates to frozen backbones, methods such as LoRA [Hu *et al.*, 2022], DoRA [Liu *et al.*, 2024], GeLoRA [Ed-dib *et al.*, 2024], and their spectral variants attempt to adapt pretrained priors with minimal parameter overhead. However, a critical limitation arises when applying these paradigms to volumetric data. Existing PEFT methods, including recent spectral approaches [Zhang and Pilanci, 2024], typically rely on global or static spectral reweighting, implicitly assuming that feature importance is spatially uniform. This assumption breaks down in 3D anatomy, where local anisotropy varies significantly across voxels. As a result, standard PEFT approaches fail to dynamically redistribute representational energy according to local geometry. Consequently, 2D texture semantics are preserved, while 3D structural awareness remains insufficient. These observations suggest that effective 2D-to-3D transfer requires more than parameter adaptation alone. The model must dynamically redistribute spectral energy according to local volumetric geometry.

To address this challenge, we propose Geometry-Conditioned Spectral Flow Low-Rank Adaptation (GeoSFLoRA), a parameter-efficient framework that injects volumetric priors directly into the spectral subspace of a frozen 2D encoder. GeoSFLoRA consists of two tightly coupled components. First, a Geometry-Conditioned Encoder constructs patch-level descriptors from raw volumetric data to capture local anisotropy and boundary sharpness. Second, a Spectral Flow Low-Rank Adapter leverages these descriptors as control signals to predict token-specific singular-value residuals. In practice, for each adapted linear projection, we compute a truncated singular value decomposition once and keep the left and right singular bases fixed. Geometry-conditioned low-rank residuals are learned and added directly to the singular values, with the truncation rank controlling the adaptation capacity. This design enables position-specific energy redistribution within the pretrained spectral subspace, selectively enhancing or suppressing spectral components in accordance with local anatomical geometry while preserving the integrity of the learned semantic basis.

We empirically evaluated GeoSFLoRA on widely used benchmarks, including BraTS20, MSD-Prostate, and MSD-Lung. We report mean performance and standard deviation across cross-validation folds and compare segmentation accuracy together with parameter counts and inference cost. Experimental results demonstrate that GeoSFLoRA consistently outperforms representative PEFT baselines and achieves performance comparable to full fine-tuning, while requiring only a fraction of the trainable parameters. These results indicate that geometry-conditioned spectral adaptation offers an effective and efficient pathway for migrating 2D foundation models to volumetric medical imaging tasks.

Our main contributions are summarized as follows:

- We propose GeoSFLoRA, a parameter-efficient framework that explicitly injects local volumetric geometric information into the spectral subspace of frozen 2D pretrained encoders.
- We design a Geometry-Conditioned Encoder that constructs stable patch-level descriptors using normalized gradient statistics and structure-tensor-inspired eigenspectrum analysis, capturing local anisotropy and directional sharpness to provide modality-agnostic geometric priors.
- We introduce a Spectral Flow Low-Rank Adapter that performs token-specific singular-value modulation based on geometric cues, enabling principled 2D-to-3D transfer while preserving the pretrained orthogonal bases of the encoder.

2 Related Work

2.1 Volumetric Medical Image Segmentation

Early volumetric methods established that 3D convolutional encoder-decoder architectures, such as V-Net [Milletari *et al.*, 2016] and U-Net [Ronneberger *et al.*, 2015], effectively capture local details. System-level frameworks like nnU-Net [Isensee *et al.*, 2021] further optimized these priors. Recent Transformer-based backbones, including TransUNet [Chen *et al.*, 2021] and Swin-UNETR [Hatamizadeh *et al.*, 2022], have pushed the boundaries of long-range context modeling, albeit at the cost of high computational and memory overhead. To mitigate these costs, 2.5D methods employ cross-slice attention [Hung *et al.*, 2024] or bidirectional propagation [Xu *et al.*, 2023]. While these approaches alleviate memory constraints, they often struggle to preserve global 3D geometric consistency, particularly under highly anisotropic sampling. More recently, adapting 2D foundation models such as SAM to volumetric medical data has emerged as a promising direction (e.g., 3DSAM-Adapter [Gong *et al.*, 2023], Medical SAM Adapter [Wu *et al.*, 2025b]). These methods leverage strong 2D semantic priors but typically rely on spatial or channel-wise prompting mechanisms that do not explicitly model volumetric geometry. In contrast, GeoSFLoRA addresses the 2D-to-3D gap at the level of representation geometry by enabling location-specific adaptation driven by local volumetric structure, ensuring anatomical coherence without the burden of full 3D convolutions.

2.2 Parameter-Efficient Fine-Tuning

Parameter-Efficient Fine-Tuning (PEFT) methods adapt large pretrained models by injecting a small number of trainable parameters while freezing the backbone. Representative approaches include LoRA [Hu *et al.*, 2022], DoRA [Liu *et al.*, 2024], and AdaLoRA [Zhang *et al.*, 2023], which introduce low-rank updates to linear projections to reduce optimization cost. More recent work explores adaptation in the spectral domain. Spectral adapters manipulate singular values or truncated SVD representations of pretrained weights, offering a principled way to reshape pretrained subspaces without altering their orthonormal semantic bases [Zhang and Pilanci, 2024]. Geometry-aware variants such as GeLoRA [Ed-dib *et al.*, 2024] further attempt to allocate adaptation capacity based on representation geometry. However, most existing PEFT methods apply global or layer-level adaptations that are shared across all tokens, implicitly assuming spatial homogeneity. This assumption is ill-suited to volumetric medical images, where anatomical geometry and anisotropy vary substantially across space. GeoSFLoRA departs from prior PEFT designs by enabling token-specific spectral modulation conditioned on local volumetric geometry. This allows representational energy to be redistributed within the pretrained subspace in a spatially adaptive and modality-agnostic manner, effectively tailoring the 2D foundation model to the local 3D structural context.

3 Methodology

3.1 Overview

GeoSFLoRA adapts 2D pretrained Vision Transformers (ViTs) to 3D medical segmentation with minimal additional parameters. The overall architecture consists of a frozen 2D-pretrained ViT encoder, GeoSFLoRA adapters inserted into all linear layers, and a lightweight 3D UNet-style decoder. We selected the powerful DINOv3 [Siméoni *et al.*, 2025] as the pretrained backbone. To process volumetric

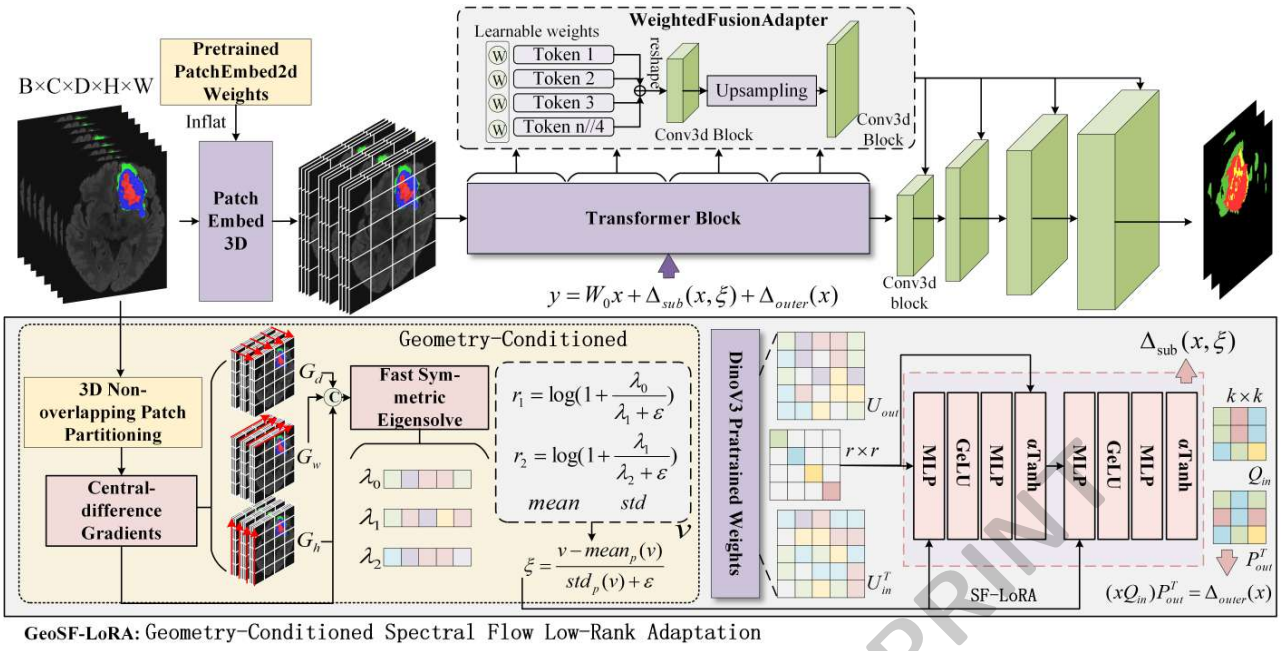

Figure 1: Overview of GeoSFLoRA. A frozen DINOv3-based ViT encoder is adapted for volumetric segmentation using two key modules: a Geometry-Conditioned Encoder (GCE) that extracts patch-wise geometric descriptors, and a Spectral Flow LoRA (SFLoRA) module that injects these cues into pretrained spectral subspaces. A lightweight WeightedFusionAdapter further fuses multi-scale ViT features before decoding.

data, the 2D patch embedding is converted into a 3D patch embedding by inflating the 2D convolutional kernels along the depth dimension. The inflated weights are initialized from the original 2D pretrained parameters, preserving semantic priors while enabling 3D tokenization. All transformer blocks, including self-attention layers and MLP layers, remain unchanged. This design allows direct loading of 2D pretrained weights. GeoSFLoRA introduces two modules: (1) a Geometry-Conditioned Encoder (GCE) that extracts per-patch descriptors of local anisotropy and directional sharpness from volumetric gradients; (2) a Spectral Flow Low-Rank Adapter (SFLoRA) that modulates pretrained singular values within each linear layer based on these descriptors. A lightweight WeightedFusionAdapter is further employed to fuse the geometry-conditioned residuals with backbone token representations through learnable scalar gating before decoder aggregation. During the forward pass, the backbone outputs are corrected by token-wise geometry-aware spectral adjustments, while all original weights remain frozen. This design preserves pretrained semantics and enables geometry-sensitive 3D adaptation with minimal additional parameters.

3.2 Geometry-Conditioned Encoder (GCE)

While 2D-pretrained ViT encoders capture rich semantic priors, they lack explicit knowledge of volumetric geometry. Effective 3D segmentation requires distinguishing line-like, surface-like, and isotropic structures. To this end, the Geometry-Conditioned Encoder (GCE) extracts compact patch-level descriptors encoding local anisotropy, directional sharpness, and intensity statistics. We prioritize compact, interpretable descriptors in GCE because they are robust across modalities and sampling rates, inexpensive to compute, and provide explicit geometric signals that directly guide token-level spectral modulation.

Given a volumetric input $x \in \mathbb{R}^{B\times C\times D\times H\times W}$, we partition it into non-overlapping patches of size (p_d, p_h, p_w), producing N_p patches with $V = p_d p_h p_w$ voxels each. Each patch is flattened into a vector form $\mathbf{x}_p \in \mathbb{R}^V$. To capture local directional variations, we compute normalized central-difference gradients along depth, height, and width:

$$\widetilde{G}_u = \frac{\nabla_u x}{\max(\mathrm{std}_{\mathrm{spatial}}(\nabla_u x), \epsilon)}, \quad u \in \{d, h, w\},$$

where $\epsilon = 10^{-6}$ ensures numerical stability. Normalizing the gradients removes scale variations, emphasizing geometric patterns rather than absolute intensity magnitudes.

To distill local geometric context, we construct the *patch-wise Structure Tensor* using the gradient vectors. Let $\mathbf{J} = [\widetilde{G}_d, \widetilde{G}_h, \widetilde{G}_w] \in \mathbb{R}^{V\times 3}$ denote the gradient matrix for a patch. The structural information is encoded in the covariance matrix $\mathbf{S} = \mathbf{J}^\top \mathbf{J} \in \mathbb{R}^{3\times 3}$. We perform an eigenspectrum analysis on $\mathbf{S}$, extracting the eigenvalues $\lambda_0 \geq \lambda_1 \geq \lambda_2$. These eigenvalues characterize the dominant spatial modes independent of intensity shift. We then define two scale-invariant geometric descriptors:

$$r_1 = \log\left(1 + \frac{\lambda_0}{\lambda_1 + \epsilon}\right), \quad r_2 = \log\left(1 + \frac{\lambda_1}{\lambda_2 + \epsilon}\right). \quad (1)$$

The motivation for employing these specific ratios is twofold. First, they provide a continuous mapping from raw voxels

to classical anatomical primitives: a high r_1 identifies *line-like* structures (e.g., vascular segments); a high r_2 and low r_1 captures *planar* structures (e.g., organ membranes); while low values for both indicate *isotropic* regions (e.g., homogeneous parenchyma). Second, the logarithmic transformation ensures numerical stability across heterogeneous imaging modalities. Finally, we concatenate these geometric priors with the intensity statistics (mean μ and standard deviation σ_{std} of the patch intensity $\mathbf{x}_p$) to form the 4D descriptor:

$$v = [\mu, \sigma_{\text{std}}, r_1, r_2].$$

Sample-wise normalization produces the geometry prior:

$$\xi = \frac{v - \text{mean}_p(v)}{\text{std}_p(v) + \epsilon} \in \mathbb{R}^{B \times N_p \times 4}.$$

To align with the ViT token sequence, the mean descriptor across all patches is assigned to the CLS token, ensuring a one-to-one correspondence between ξ and the full token sequence used in SFLoRA.

3.3 Spectral Flow Adaptation

The geometry descriptors produced by the GCE provide token-wise volumetric context absent in 2D-pretrained ViTs. To effectively incorporate these descriptors without disrupting pretrained semantics or incurring the computational overhead of full fine-tuning, we introduce Spectral Flow LoRA (SFLoRA). SFLoRA adapts each linear layer by modulating its frozen spectral components rather than updating parameters in the ambient weight space. This design injects 3D geometric sensitivity directly into the pretrained singular directions while ensuring that feature representations remain anchored to the robust 2D semantic manifold.

For a frozen linear layer $W_0 \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ and a token vector $x^{(i)} \in \mathbb{R}^{d_{\text{in}}}$ from the input sequence $X \in \mathbb{R}^{N \times d_{\text{in}}}$, the adapted output is formulated as:

$$y^{(i)} = W_0 x^{(i)} + \Delta_{\text{sub}}(x^{(i)}, \xi^{(i)}),$$

where Δ_{sub} represents the geometry-conditioned subspace correction computed per token using the geometry prior ξ from GCE.

Truncated SVD Decomposition

Prior to adaptation, we factorize each frozen linear layer via truncated SVD:

$$W_0 \approx U_{\text{out}} \, \text{diag}(\sigma_0) \, U_{\text{in}}^{\top},$$

where $U_{\text{in}} \in \mathbb{R}^{d_{\text{in}} \times r}$, $U_{\text{out}} \in \mathbb{R}^{d_{\text{out}} \times r}$, and $\sigma_0 \in \mathbb{R}^r$ denote the frozen input basis, output basis, and singular values, respectively. The truncation rank r acts as a bottleneck for adaptation capacity. Freezing U_{in} and U_{out} prevents arbitrary rotation of the semantic space, preserving the orthogonal feature axes learned from large-scale 2D pretraining. Consequently, adaptation is constrained to reweighting and mixing these existing axes, which empirically minimizes semantic drift.

Geometry-Conditioned Singular-Value Modulation

While σ_0 defines the global importance of semantic directions, the optimal spectral gain varies locally depending on anatomical geometry. Regions with different geometric anisotropy often require different spectral emphases. For example, boundary-dominant structures benefit from high-frequency directional components, whereas homogeneous regions rely more on low-frequency semantic priors. SFLoRA predicts token-wise bounded residuals for the singular values:

$$\delta_\sigma^{(i)} = \alpha \cdot \tanh\Big(f_\sigma([\sigma_0, \xi^{(i)}])\Big) \in \mathbb{R}^r,$$

$$\sigma^{(i)} = \sigma_0 + \delta_\sigma^{(i)}, \quad i = 1, \ldots, N.$$

The MLP f_σ maps the joint embedding of the static spectrum σ_0 and dynamic geometry $\xi^{(i)}$ to a residual vector. This achieves diagonal spectral modulation, selectively amplifying or suppressing specific semantic bases based on local structure. The usage of $\tanh$ with a global scaler α ensures the spectral updates remain bounded and perturbative, preventing the adapter from overriding the pretrained knowledge base.

Conditional Residual Mixing in Subspace

Diagonal modulation scales existing bases but cannot model interactions between them. To capture complex geometric adaptations, we introduce a dense residual mixing matrix:

$$R_{\text{cond}}^{(i)} = \alpha \cdot \tanh\Big(f_R([\sigma^{(i)}, \xi^{(i)}])\Big) \in \mathbb{R}^{r \times r}.$$

Here, f_R predicts a small, geometry-dependent mixing term. This allows for fine-grained cross-channel communication within the spectral subspace. We include residual mixing to capture off-diagonal interactions that scalar reweighting cannot express. By predicting residuals within the low-rank subspace $r \ll \min(d_{\text{in}}, d_{\text{out}})$, we enable directional specificity with minimal parameter overhead.

Efficient Spectral Flow Computation

The diagonal modulation and residual mixing jointly define the token-specific spectral residual kernel:

$$S^{(i)}(\xi) = \text{diag}(\sigma^{(i)} - \sigma_0) + R_{\text{cond}}^{(i)} \in \mathbb{R}^{r \times r}. \tag{2}$$

To maintain high computational and memory efficiency, we avoid explicitly instantiating the full spectral kernel tensor $S \in \mathbb{R}^{N \times r \times r}$. Instead, we implement a streamlined pipeline to derive the subspace correction for each token in a batch-wise manner. First, the input tokens are projected into the low-rank spectral subspace via $U_{\text{in}}^{\top}$. Second, the geometry-conditioned spectral adjustments are applied by multiplying the kernel $S^{(i)}(\xi)$ with the projected vectors to selectively modulate the spectral energy based on local anatomical priors. Finally, these modulated latent representations are back-projected into the original feature space via U_{out} to yield the final subspace correction. This sequential process is mathematically summarized as:

$$\Delta_{\text{sub}}(x, \xi)^{(i)} = U_{\text{out}} \left(S^{(i)}(\xi) \left(U_{\text{in}}^{\top} x^{(i)} \right) \right). \tag{3}$$

This implementation strategy effectively reduces the computational complexity from $O(Ndr^2)$ to $O(Ndr + Nr^2)$, where $d = \max(d_{\text{in}}, d_{\text{out}})$. Since the truncation rank r is significantly smaller than the embedding dimension d (i.e., $r \ll d$), GeoSFLoRA maintains a memory footprint and inference latency comparable to standard LoRA while providing significantly higher representational flexibility, particularly for high-resolution volumetric medical data.

Methods	MSD-Lung		MSD-Prostate				BraTS20						Training/
	Dice(%)↑	HD95(mm)↓	Dice(%)↑		HD95(mm)↓		Dice(%)↑			HD95(mm)↓			LoRA
	Cancer	Cancer	PZ	TZ	PZ	TZ	WT	ET	TC	WT	ET	TC	Params
nnU-Net	75.34±0.45	15.14±2.98	65.77±1.78	85.73±1.02	15.11±4.22	10.1±2.01	90.61±0.09	77.88±0.12	80.32±0.33	8.21±2.45	18.84±2.54	11.92±2.78	59/-
UNETR++	74.2±0.78	17.14±2.34	65.7±1.97	84.97±1.24	16.78±4.31	11.99±2.06	90.29±0.1	77.15±0.24	81.46±0.34	6.18±1.78	20.58±2.43	17.07±2.94	127.8/-
SegMamba-V2	75.35±0.48	16.14±3.34	64.79±0.78	85.04±0.78	12.41±3.97	12.87±2.34	90.67±0.08	78.15±0.15	**83.46±0.28**	**4.18±2.4**	26.58±2.71	**5.07±1.67**	30/-
DinoU-Net	73.91±0.58	15.99±1.99	66.13±0.91	84.98±0.94	13.56±3.87	12.21±1.83	90.09±0.13	78.11±0.29	81.73±0.29	4.96±2.11	17.95±2.11	9.77±2.36	18.14/-
3DSAM-Adapter	76.01±0.34	17.44±2.41	65.94±0.79	85.41±0.97	11.84±3.16	13.49±2.49	90.46±0.11	77.94±0.17	82.31±0.24	5.21±1.83	16.84±2.41	6.48±1.98	25.5/-
AutoProSAM	75.94±0.21	15.81±2.26	66.31±0.81	85.99±1.64	10.97±2.94	10.87±1.92	90.81±0.14	78.48±0.15	83.14±0.21	4.77±1.92	16.32±2.14	7.74±2.54	26.5/-
MedSAM2	76.38±0.26	14.03±1.94	66.48±1.31	86.14±1.47	11.32±4.01	10.29±1.77	**90.9±0.17**	**78.73±0.21**	82.67±0.38	4.32±2.03	16.52±2.05	5.97±2.34	63/-
2.5D	73.75±0.31	18.94±3.14	55.78±0.78	76.94±1.62	17.54±4.42	15.82±1.88	86.49±0.11	73.94±0.24	76.39±0.28	13.77±1.84	24.81±2.68	15.66±2.03	23/-
Full Train	**76.45±0.34**	**12.76±3.01**	**66.54±1.71**	**86.78±1.56**	**10.12±4.32**	10.09±2.02	90.44±0.14	78.15±0.18	82.19±0.31	7.04±1.78	15.91±2.01	9.37±2.65	342/-
Frozen ViT	73.11±0.12	19.36±1.79	50.11±0.56	74.11±1.64	20.1±4.01	17.14±1.79	80.74±0.09	68.44±0.19	70.11±0.36	15.44±1.97	30.98±2.78	20.11±2.36	23/-
LoRA	74.11±0.23	17.11±2.01	63.89±1.24	84.3±1.45	13.41±4.02	13.44±1.54	87.71±0.11	75.99±0.2	80.99±0.28	11.01±1.57	22.11±2.34	13.78±2.32	40.1/18.1
DoRA	73.94±0.15	15.32±1.85	64.15±1.97	84.61±1.35	15.64±4.06	12.65±1.89	87.99±0.1	75.05±0.21	79.41±0.26	10.93±1.93	24.72±3.15	12.11±1.82	40/18
AdaLoRA	74.9±0.17	14.54±1.83	64.79±1.67	84.99±1.01	12.47±3.99	14.14±2.14	87.48±0.09	75.11±0.19	81.54±0.33	8.37±1.96	19.63±1.87	12.65±1.85	40.1/18.1
Spectral Adapters	74.45±0.16	16.11±1.47	64.06±1.45	85.4±1.7	11.03±4.17	13.51±2.03	87.48±0.12	76.11±0.23	80.94±0.31	9.44±1.89	18.56±1.45	12.97±1.61	40.1/18.1
Our	76.12±0.34	14.01±1.29	66.01±2.01	86.41±0.94	10.95±4.11	**9.44±2.04**	90.23±0.12	77.99±0.2	82.41±0.32	7.44±1.84	**15.77±2.06**	9.91±1.87	39/17

Table 1: Performance comparison of GeoSFLoRA vs. other methods on datasets. (± denotes standard deviation, results are averages of 5-fold cross-validation; DINOv3-Large is used as the pretrained ViT encoder, with $r = 32$ for PEFT and SVD-based adaptation methods)

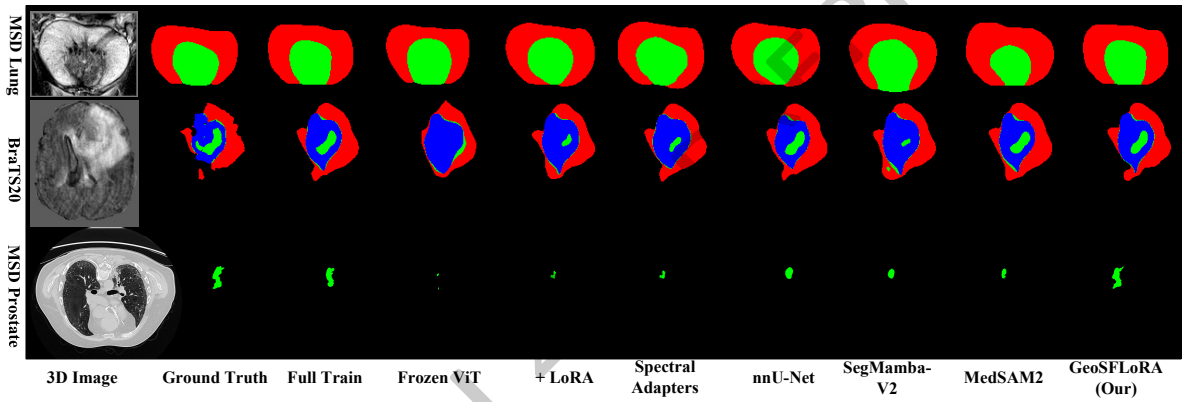

Figure 2: Visual Comparison of 3D Medical Image Segmentation Results Across MSD-Lung, BraTS20, and MSD-Prostate Datasets for Different Methods.

3.4 Loss Function

All adapter parameters and the 3D decoder are optimized end-to-end while the ViT backbone remains frozen. The segmentation objective is a balanced combination of Dice and focal losses:

$$\mathcal{L} = \lambda_{\text{Dice}}\mathcal{L}_{\text{Dice}} + \lambda_{\text{Focal}}\mathcal{L}_{\text{Focal}},$$

with $\lambda_{\text{Dice}} = \lambda_{\text{Focal}} = 0.5$ by default.

4 Experiments

4.1 Datasets and Evaluation Protocol

To comprehensively validate the cross-domain generalization and geometry-conditioned adaptation of GeoSFLoRA, we evaluate our model on three diverse 3D medical segmentation benchmarks: **BraTS20**, which contains 369 multi-modal MRI scans (T1, T1ce, T2, FLAIR) for enhancing tumor (ET), tumor core (TC), and whole tumor (WT) segmentation; **MSD-Prostate**, consisting of 48 dual-modal MRI volumes (T2, ADC) resampled to $1\,\text{mm}^3$ isotropic spacing for peripheral zone (PZ) and transition zone (TZ) segmentation; **MSD-Lung**, featuring 64 thoracic CT scans for lung tumor segmentation. A unified preprocessing pipeline is applied across all datasets, including intensity clipping to the 0.5th–99.5th percentiles, z-score normalization, and random cropping to target input dimensions: $[4, 128, 128, 128]$ for BraTS20, $[2, 20, 320, 256]$ for Prostate, and $[1, 80, 192, 160]$ for Lung. To mitigate overfitting, we employ standard data augmentations including random flipping across three axes ($p = 0.25$), random rotations ($\pm 15°$), and intensity scaling (0.8–$1.2\times$). All datasets use patient-level 5-fold cross-validation, and Dice and HD95 are reported as mean ± standard deviation across folds.

4.2 Implementation Details

We implemented the proposed SFLoRA in PyTorch and conducted training on 4 NVIDIA RTX 4090 GPUs (24GB each). The Adam optimizer was adopted with an initial learning rate of 1×10^{-4}, complemented by a cosine annealing scheduling

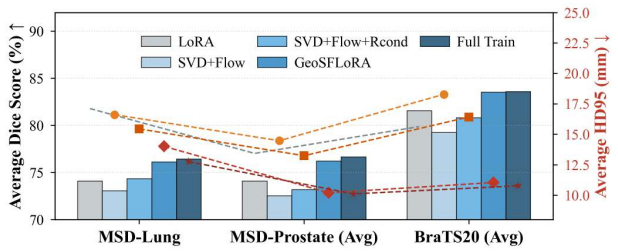

Figure 3: Ablation study of GeoSFLoRA components averaged across MSD-Lung, MSD-Prostate, and BraTS20.

Dataset	Metrics	SVD r=4	SVD r=8	SVD r=16	SVD r=32
MSD Lung	Cancer-Dice	70.45±1.11	73.11±0.94	75.47±0.2	**76.12±0.34**
	Cancer-HD95	25.01±2.14	20.01±2.03	15.99±1.07	**14.01±1.29**
MSD Prostate	PZ-Dice	63.01±1.98	64.40±2.3	65.5±2.02	**66.01±2.01**
	TZ-Dice	84.01±1.21	84.15±0.5	86.01±0.44	**86.41±0.94**
	PZ-HD95	15.15±3.01	13.15±2.54	**10.10±3.02**	10.95±4.11
	TZ-HD95	12.58±2.08	11.97±2.37	10.32±2.03	**9.44±2.04**
BraTS20	WT-Dice	85.23±0.11	88.23±0.11	89.97±0.07	**90.23±0.12**
	ET-Dice	74.91±0.17	75.15±0.19	77.18±0.14	**77.99±0.2**
	TC-Dice	79.14±0.55	80.78±0.45	81.45±0.33	**82.41±0.32**
	WT-HD95	10.69±2.11	8.01±1.87	7.98±1.54	**7.44±1.84**
	ET-HD95	24.77±2.11	19.47±2.6	15.97±2.1	**15.77±2.06**
	TC-HD95	12.41±0.87	10.92±1.42	10.4±1.77	**9.91±1.87**

Table 2: Influence of different truncated SVD ranks r on GeoSFLoRA performance.

strategy [Loshchilov and Hutter, 2016] and a weight decay of 1×10^{-4}. The ViT encoder operates with a patch size of $4 \times 16 \times 16$. The model was trained for 1000 epochs with a batch size of 8, and automatic mixed precision (AMP) [Lam *et al.*, 2013] was employed to reduce memory consumption and accelerate training.

4.3 Comparison with State-of-the-Art Methods

We evaluate GeoSFLoRA against four categories: (1) 3D architectures (nnU-Net [Isensee *et al.*, 2021], UNETR++ [Shaker *et al.*, 2024], SegMamba-V2 [Xing *et al.*, 2025]); (2) a 2D slice-by-slice variant of DinoU-Net [Gao *et al.*, 2025] with a DINOv3-Large backbone ($r = 32$); (3) foundation adapters (MedSAM2 [Ma *et al.*, 2025]); and (4) PEFT baselines (LoRA [Hu *et al.*, 2022], DoRA [Liu *et al.*, 2024], Spectral Adapters [Zhang and Pilanci, 2024]). As shown in Table 1, GeoSFLoRA achieves the best HD95 among PEFT baselines on MSD-Lung (14.01 mm) and the best overall HD95 on Prostate TZ (9.44 mm). On BraTS20-ET, GeoSFLoRA improves the Dice score by ∼2.0% and reduces the HD95 error by over 6 mm compared to LoRA. These results validate that GeoSFLoRA efficiently bridges the 2D-to-3D representational gap while ensuring structural integrity across diverse volumetric modalities. Visual comparisons in Fig. 2 mirror the quantitative gains. GeoSFLoRA produces sharper boundaries and fewer artifacts in small prostate zones and complex lung tumor edges.

4.4 Component-wise Contribution of GeoSFLoRA

We evaluate component contributions on three benchmarks using frozen DINOv3-large [Siméoni *et al.*, 2025]. Starting from baseline spectral modulation (SVD+Flow, $r = 32$), we progressively integrate the conditional residual matrix (R_{cond}) and GCE geometric priors, comparing against standard LoRA [Hu *et al.*, 2022] (Fig. 3).

Although LoRA offers competitive benchmark solutions, its uniform weight update mechanism struggles to adapt to 3D medical images, resulting in suboptimal boundary accuracy. In contrast, the vanilla SVD+Flow lacks cross-dimensional awareness, but adding R_{cond} enables geometry-conditioned subspace mixing, improving Dice by 1.27% on MSD-Lung. The full GeoSFLoRA yields the most significant leap; by using GCE descriptors to guide local 3D adaptation based on structural sharpness, it effectively bridges the 2D-to-3D representational gap.

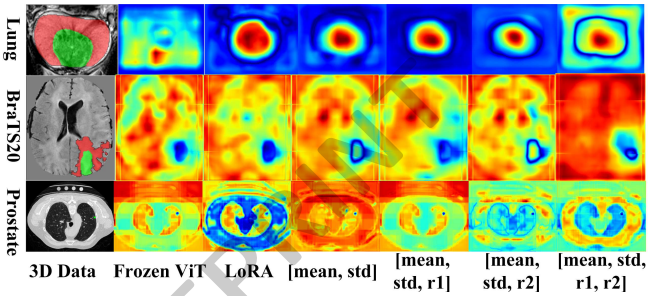

Figure 4: Deepest-layer attention maps generated by DINOv3 under different geometric descriptor configurations. The attention maps are restored to the input resolution via trilinear interpolation.

4.5 Ablation on the Truncated SVD Rank r

We investigate the effect of the truncated SVD rank r on GeoSFLoRA's representational capacity and segmentation performance, as summarized in Table 2. Experiments are conducted on BraTS20, MSD-Prostate, and MSD-Lung datasets using a frozen DINOv3-Large encoder. As r increases from 4 to 32, the Dice coefficient consistently improves across all datasets, while the HD95 decreases, indicating more precise boundary delineation. Mechanistically, the SVD rank r defines the degrees of freedom available for spectral flow modulation. Lower ranks constrain the adaptation of pretrained singular vectors to token-specific geometric priors ξ, limiting the model's ability to capture complex local structures. Increasing r expands the spectral subspace, allowing richer geometry-aware modulation and enhanced cross-dimensional interactions, thereby bridging the 2D-to-3D domain gap more effectively.

4.6 Visualizing Geometric Guidance

Fig. 4 and Fig. 5 illustrate how geometric descriptors guide the 2D ViT toward 3D structural awareness. While Frozen ViT and LoRA exhibit diffuse or coarse attention, incorporating descriptors (r_1, r_2) progressively localizes attention maps onto target boundaries, while the summary plot in Fig. 5 shows consistent improvements in average Dice and HD95 across benchmarks. Overall, the attention evolution in Fig. 4 is consistent with the quantitative trends shown in Fig. 5, demonstrating that geometric descriptors effectively redirect pretrained 2D semantic directions toward volumetric 3D anatomy.

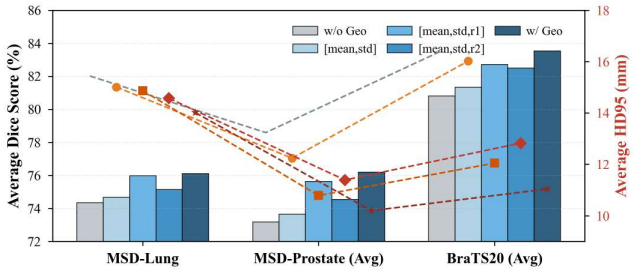

Figure 5: Ablation of geometric descriptor configurations on GeoSFLoRA across MSD-Lung, MSD-Prostate, and BraTS20.

Dataset	Metrics	w/o	MLP only	QKV only	Full
MSD Lung	Cancer-Dice	73.11±0.12	75.18±0.21	74.99±0.25	**76.12±0.34**
	Cancer-HD95	19.36±1.79	16.41±1.57	15.14±1.21	**14.01±1.29**
MSD Prostate	PZ-Dice	50.11±0.56	62.48±2.46	60.79±1.45	**66.01±2.01**
	TZ-Dice	74.11±1.64	82.14±1.44	80.55±1.97	**86.41±0.94**
	PZ-HD95	20.1±4.01	12.48±4.56	14.11±3.46	**10.95±4.11**
	TZ-HD95	17.14±1.79	11.99±1.47	12.44±1.97	**9.44±2.04**
BraTS20	WT-Dice	80.74±0.09	88.14±0.15	86.65±0.44	**90.23±0.12**
	ET-Dice	68.44±0.19	74.35±0.17	71.48±0.41	**77.99±0.2**
	TC-Dice	70.11±0.36	78.21±0.43	75.68±0.24	**82.41±0.32**
	WT-HD95	15.44±1.97	9.63±1.94	10.34±1.77	**7.44±1.84**
	ET-HD95	30.98±2.78	18.67±2.43	20.11±2.13	**15.77±2.06**
	TC-HD95	20.11±2.36	13.57±1.57	15.75±1.72	**9.91±1.87**

Table 3: Impact of adapter insertion position on GeoSFLoRA performance. Results are averages of 5-fold cross-validation with DINOv3-Large and $r = 32$.

4.7 Layer Insertion Strategy

Table 3 evaluates the placement of GeoSFLoRA adapters within the ViT encoder. Frozen weights alone yield the lowest performance, while the joint MLP+QKV configuration consistently achieves superior results.

4.8 Sensitivity of Modulation Strength α

We evaluate the modulation strength $\alpha \in \{0.4, 0.6, 0.8, 1.0\}$ across three benchmarks (Fig. 6). Results show that $\alpha = 0.8$ consistently yields the highest and most stable Dice scores, providing an optimal balance between geometric adaptation and training stability.

Lower values ($\alpha = 0.4$) lead to under-adaptation and slow convergence, while excessive modulation ($\alpha = 1.0$) causes performance oscillations, particularly on BraTS20, by over-perturbing pretrained singular values. Consequently, we adopt $\alpha = 0.8$ as the default setting to ensure robust 3D adaptation without destabilizing the frozen backbone.

4.9 Impact of Backbone Capacity

Across DINOv3-Small, Base, and Large variants, validation performance improves with model capacity in our setting; therefore, DINOv3-Large is used as the default backbone in the main experiments.

4.10 Efficiency Analysis

Table 4 shows that GeoSFLoRA matches or slightly reduces training time compared with LoRA while incurring only marginal inference overhead, indicating a favorable balance between cost and accuracy.

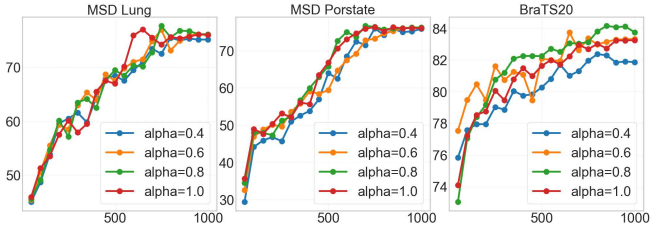

Figure 6: Effect of modulation strength α on mean Dice scores across MSD-Lung, MSD-Prostate, and BraTS20. $\alpha = 0.8$ consistently achieves the highest and most stable performance.

Method	MSD-Lung	MSD-Prostate	BraTS20
Total Training 1000 Epochs Time (h)			
nnU-Net	**3.06**	**1.94**	**8.34**
LoRA	3.88	2.22	11.54
GeoSFLoRA	3.33	2.22	11.45
Avg. Inference Time per Image (s)			
nnU-Net	**61.40**	**1.14**	**4.27**
LoRA	71.40	1.24	5.12
GeoSFLoRA	71.50	1.25	5.15

Table 4: Training cost and sliding-window inference efficiency (overlap=0.25).

5 Conclusion

GeoSFLoRA bridges the 2D-to-3D domain gap via geometry-conditioned spectral modulation in frozen ViTs. By leveraging voxel-wise anisotropy and structural sharpness, it improves volumetric consistency with minimal overhead and outperforms standard PEFT methods across three benchmarks. Future work will explore modality-aware geometric conditioning for varied MRI and CT profiles.

Ethical Statement

There are no ethical issues.

Acknowledgements

This work was supported in part by the Tianshan Talent Training Program 2023TSYCLJ0023. This work was supported by the university-level computing platform of Xinjiang University's Computing and Data Center.

Contribution Statement

Qin Hao and Bonian Chen contributed equally to this work. Shengwei Tian and Long Yu are the corresponding authors. Qin Hao and Bonian Chen developed the method, conducted the experiments, and drafted the manuscript. Shengwei Tian and Long Yu supervised the study, revised the manuscript, and provided project support.

References

[Awais *et al.*, 2025] Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Foundation models defining a new era in vision: a survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[Chen *et al.*, 2021] 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.

[Ed-dib *et al.*, 2024] Abdessalam Ed-dib, Zhanibek Datbayev, and Amine Mohamed Aboussalah. Gelora: Geometric adaptive ranks for efficient lora fine-tuning. *arXiv preprint arXiv:2412.09250*, 2024.

[Gao *et al.*, 2025] Yifan Gao, Haoyue Li, Feng Yuan, Xiaosong Wang, and Xin Gao. Dino u-net: Exploiting high-fidelity dense features from foundation models for medical image segmentation. *arXiv preprint arXiv:2508.20909*, 2025.

[Gong *et al.*, 2023] Shizhan Gong, Yuan Zhong, Wenao Ma, Jinpeng Li, Zhao Wang, Jingyang Zhang, Pheng-Ann Heng, and Qi Dou. 3dsam-adapter: Holistic adaptation of sam from 2d to 3d for promptable medical image segmentation. *arXiv e-prints*, pages arXiv–2306, 2023.

[Hatamizadeh *et al.*, 2022] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *BrainLes / MICCAI Workshops*, 2022.

[Hu *et al.*, 2022] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[Hung *et al.*, 2024] Alex Ling Yu Hung, Haoxin Zheng, Kai Zhao, Xiaoxi Du, Kaifeng Pang, Qi Miao, Steven S Raman, Demetri Terzopoulos, and Kyunghyun Sung. Csam: A 2.5 d cross-slice attention module for anisotropic volumetric medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5923–5932, 2024.

[Isensee *et al.*, 2021] Fabian Isensee, Paul F. Jaeger, Simon A. A. 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.

[Jiao *et al.*, 2025] Licheng Jiao, Jingyi Yang, Ruiyang Li, Fang Liu, Xu Liu, Puhua Chen, Yuwei Guo, Lingling Li, Ronghua Shang, and Wenping Ma. Foundation model for medical imaging: A comprehensive review. *IEEE Transactions on Artificial Intelligence*, 2025.

[Lam *et al.*, 2013] Michael O Lam, Jeffrey K Hollingsworth, Bronis R de Supinski, and Matthew P Legendre. Automatically adapting programs for mixed-precision floating-point computation. In *Proceedings of the 27th international ACM conference on International conference on supercomputing*, pages 369–378, 2013.

[Li *et al.*, 2024] Wenxuan Li, Alan Yuille, and Zongwei Zhou. How well do supervised 3d models transfer to medical imaging tasks? In *ICLR 2024 (OpenReview)*, 2024.

[Li *et al.*, 2025] Chengyin Li, Rafi Ibn Sultan, Prashant Khanduri, Yao Qiang, Chetty Indrin, and Dongxiao Zhu. Autoprosam: Automated prompting sam for 3d multi-organ segmentation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3570–3580. IEEE, 2025.

[Litjens *et al.*, 2017] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud A A Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A W M van der Laak, et al. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

[Liu *et al.*, 2024] Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning*, 2024.

[Loshchilov and Hutter, 2016] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[Ma *et al.*, 2025] Jun Ma, Zongxin Yang, Sumin Kim, Bihui Chen, Mohammed Baharoon, Adibvafa Fallahpour, Reza Asakereh, Hongwei Lyu, and Bo Wang. Medsam2: Segment anything in 3d medical images and videos. *arXiv preprint arXiv:2504.03600*, 2025.

[Mazurowski *et al.*, 2023] Maciej A Mazurowski, Haoyu Dong, Hanxue Gu, Jichen Yang, Nicholas Konz, and Yixin Zhang. Segment anything model for medical image analysis: an experimental study. *Medical Image Analysis*, 89:102918, 2023.

[Milletari *et al.*, 2016] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571. IEEE, 2016.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas 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*, pages 234–241. Springer, 2015.

[Shaker *et al.*, 2024] 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. *IEEE Transactions on Medical Imaging*, 43(9):3377–3390, 2024.

[Siméoni *et al.*, 2025] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

[VanBerlo *et al.*, 2024] Blake VanBerlo, Jesse Hoey, and Alexander Wong. A survey of the impact of self-supervised pretraining for diagnostic tasks in medical x-

ray, ct, mri, and ultrasound. *BMC Medical Imaging*, 24(1):79, 2024.

[Wang *et al.*, 2024] Xin Wang, Zhiyun Song, Yitao Zhu, Sheng Wang, Lichi Zhang, Dinggang Shen, and Qian Wang. Inter-slice super-resolution of magnetic resonance images by pre-training and self-supervised fine-tuning. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2024.

[Wang *et al.*, 2025] Luping Wang, Sheng Chen, Linnan Jiang, Shu Pan, Runze Cai, Sen Yang, and Fei Yang. Parameter-efficient fine-tuning in large language models: a survey of methodologies. *Artificial Intelligence Review*, 58(8):227, 2025.

[Wu *et al.*, 2025a] Jing Wu, Yuli Wang, Zhusi Zhong, Weihua Liao, Natalia Trayanova, Zhicheng Jiao, and Harrison X Bai. Vision-language foundation model for 3d medical imaging. *npj Artificial Intelligence*, 1(1):17, 2025.

[Wu *et al.*, 2025b] Junde Wu, Ziyue Wang, Mingxuan Hong, Wei Ji, Huazhu Fu, Yanwu Xu, Min Xu, and Yueming Jin. Medical sam adapter: Adapting segment anything model for medical image segmentation. *Medical image analysis*, 102:103547, 2025.

[Xing *et al.*, 2025] Zhaohu Xing, Tian Ye, Yijun Yang, Du Cai, Baowen Gai, Xiao-Jian Wu, Feng Gao, and Lei Zhu. Segmamba-v2: Long-range sequential modeling mamba for general 3d medical image segmentation. *IEEE Transactions on Medical Imaging*, 2025.

[Xu *et al.*, 2023] Xin Xu, Wenjing Lu, Jiahao Lei, Peng Qiu, Hong-Bin Shen, and Yang Yang. Sliceprop: A slice-wise bidirectional propagation model for interactive 3d medical image segmentation. In *2023 IEEE International Conference on Medical Artificial Intelligence (MedAI)*, pages 414–424. IEEE, 2023.

[Zhang and Pilanci, 2024] Fangzhao Zhang and Mert Pilanci. Spectral adapter: Fine-tuning in spectral space. arXiv preprint arXiv:2405.13952, 2024.

[Zhang *et al.*, 2023] Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.10512*, 2023.