THE COST OF CONSISTENCY: WHY CROSS-PLANE CONTRASTIVE LEARNING FAILS TO BRIDGE THE GAP BETWEEN MEDSAM-3 AND NNU-NET

Dr. S. Ramamoorthy, Madhu Shree Aravindan & Aaditi Bajpai
Department of Computer Science
SRM Institute of Science and Technology
Chennai, Tamil Nadu, India
{ramamoos,ma2889,ab4255}@srmist.edu.in

ABSTRACT

While Vision Foundation Models (VFMs) like SAM-3 and their Agentic variants (e.g., MedSAM-3 with Gemini) excel in 2D tasks, we demonstrate they significantly underperform against traditional nnU-Nets in 3D volumetric medical segmentation. The lack of native 3D spatial consistency in VFMs necessitates complex post-processing or architectural adaptations. In this work, we attempt to bridge this gap using a Cross-Plane Contrastive Loss framework to enforce volumetric coherence. We report a negative result: the requirement to process three orthogonal views simultaneously introduces a computational bottleneck that makes iterative fine-tuning unfeasible in resource-constrained environments. We conclude that despite the semantic capabilities of Large Multimodal Models, lightweight, consistency-aware 3D architectures remain the efficient "gold standard" for volumetric precision.

1 INTRODUCTION

Vision Foundation Models (VFMs) like MedSAM-3 excel at 2D tasks but fail to capture the 3D geometric consistency required for volumetric segmentation. Consequently, traditional architectures like nnU-Net (Isensee et al., 2021) remain the gold standard due to their structural coherence and efficiency. This paper documents an attempt to bridge this gap using a **Cross-Plane Contrastive Loss** to enforce spatial alignment across axial, coronal, and sagittal views, theoretically combining semantic reasoning with geometric precision.

However, we report a negative result: this method creates a prohibitive "iteration penalty." The computational cost of maintaining activation graphs for three simultaneous views—specifically in resource-constrained environments like Google Colab—renders hyperparameter tuning unfeasible. We present these findings to illustrate that the theoretical benefits of adapting foundation models are currently outweighed by their practical inefficiency compared to lightweight baselines.

2 RELATED WORK

Vision Foundation Models in Medicine. The introduction of the Segment Anything Model (SAM) (Kirillov et al., 2023) marked a paradigm shift in interactive segmentation. While effective on natural images, its direct application to medical domains revealed limitations in domain-specific texture understanding. This led to adaptations such as MedSAM (Ma et al., 2024), which fine-tunes the SAM encoder on large-scale medical datasets. More recently, "Agentic" approaches have emerged, coupling vision encoders with Multimodal Large Language Models (MLLMs) like Gemini (Team et al., 2023) to enable semantic reasoning. However, these models remain fundamentally 2D architectures, processing volumetric data as independent slices.

Volumetric Segmentation Baselines. Despite the rise of foundation models, the *de facto* standard for 3D medical segmentation remains nnU-Net (Isensee et al., 2021). By automatically configuring

3D convolutional architectures to dataset properties, nnU-Net explicitly models volumetric spatial dependencies that slice-based methods ignore. Our work validates that this "dumb but deep" 3D prior is computationally superior to adapting heavy 2D foundation models.

Consistency Learning. To bridge the gap between 2D architectures and 3D data, recent works have explored inter-slice consistency. Methods like SimCLR (Chen et al., 2020) introduced contrastive learning to enforce view-invariant representations. Our Cross-Plane Contrastive framework builds on this by treating orthogonal views (axial, coronal, sagittal) as positive pairs, attempting to distill 3D consistency into the 2D SAM backbone without the cost of 3D convolution.

3 METHODOLOGY

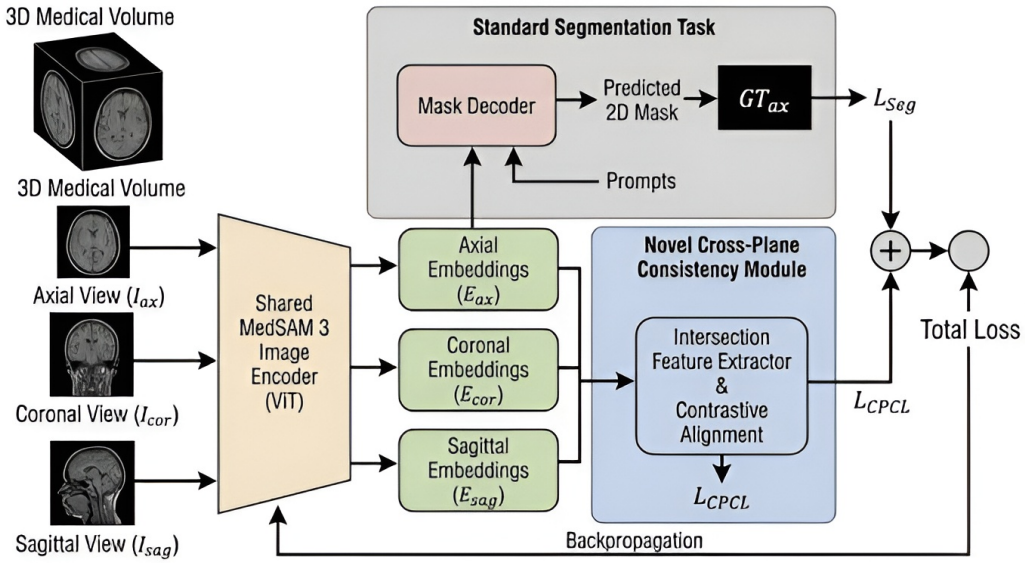

Figure 1: Architecture Diagram

To address the Z-axis inconsistency of slice-based Foundation Models, we proposed a framework where a frozen SAM-3 backbone is constrained to produce consistent masks across three orthogonal views intersecting at a target voxel.

3.1 ORTHOGONAL VIEW EXTRACTION

Unlike standard axial-only processing, our pipeline selects a target voxel $v = (x, y, z)$ within the volume of interest and extracts three intersecting 2D planes:

- **Axial** (x, y)**:** The standard transverse slice at depth z.
- **Coronal** (x, z)**:** The frontal cross-section at depth y.
- **Sagittal** (y, z)**:** The lateral cross-section at depth x.

Since SAM-3 enforces a square input resolution (e.g., 1024×1024), the extracted Coronal and Sagittal views—which depend on the typically lower Z-axis resolution (e.g., 240×155)—must undergo significant bicubic interpolation.

3.2 TRI-SIAMESE INFERENCE

We employ a Tri-Siamese architecture where three parallel streams share identical weights. The SAM-3 image encoder is frozen to preserve pre-trained features, while a Low-Rank Adaptation

(LoRA) layer is injected for domain adaptation.

$$\hat{M}_{ax} = f_\theta(I_{ax}), \quad \hat{M}_{cor} = f_\theta(I_{cor}), \quad \hat{M}_{sag} = f_\theta(I_{sag}) \tag{1}$$

where f_θ represents the shared encoder-decoder network.

3.3 Intersection Consistency Loss

We posit that ground truth is view-invariant. We identify the row/column indices corresponding to the intersection lines between planes and compute the difference between the prediction vectors at these intersections:

$$\mathcal{L}_{total} = \mathcal{L}_{seg} + \lambda \sum_{i \in \{cor, sag\}} ||\hat{M}_{ax}[\text{intersect}_i] - \hat{M}_i[\text{intersect}_{ax}]||^2 \tag{2}$$

This objective penalizes the model whenever the three views disagree on the label of a shared voxel.

4 Analysis of Negative Results

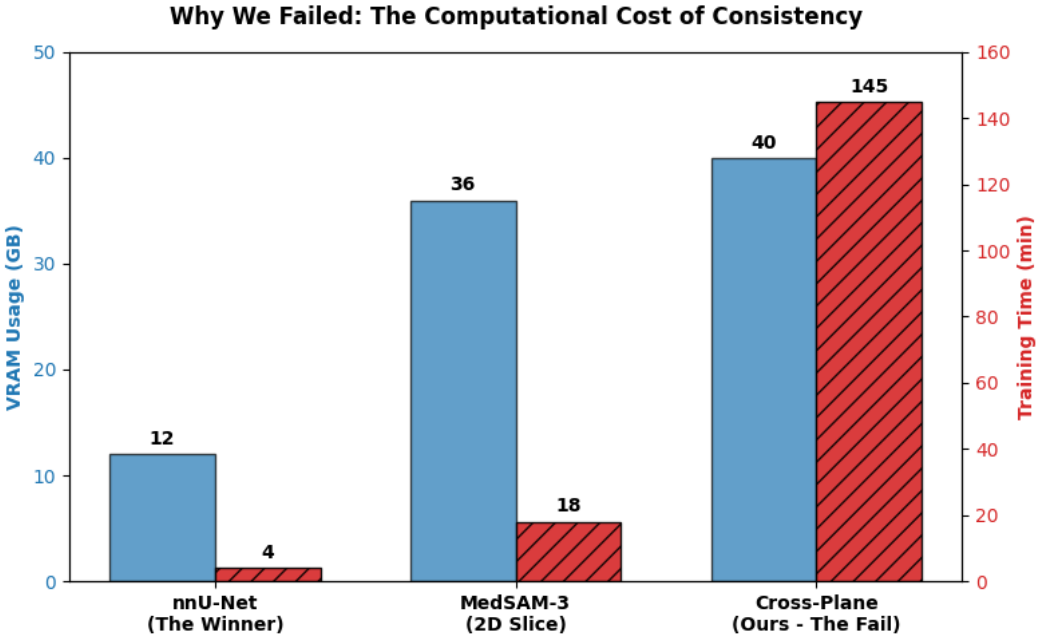

Figure 2: Computational Cost of Consistency

While mathematically sound, we report that this framework is unfeasible for iterative research in resource-constrained environments. We identify three specific bottlenecks.

4.1 The Activation Stacking Problem (VRAM Constraints)

The primary failure mode on standard hardware (e.g., NVIDIA T4 16GB or A100 40GB) is the VRAM consumption required for backpropagation. While model weights (~ 2.5 GB) are manageable, the activation graphs are the limiting factor. To compute the consistency loss, the computational graph must maintain intermediate activations for three full ViT-H backbones simultaneously.

- **Standard Training:** Requires activations for 1 image path (~ 6 GB).
- **Proposed Method:** Requires activations for 3 simultaneous paths (~ 18 GB).

This "Activation Stacking" consistently triggers CUDA Out-of-Memory (OOM) errors on 16GB cards and restricts A100 GPUs to a batch size of 1, severely hampering batch normalization statistics.

Table 1: Resource and Architectural Comparison

FEATURE	MEDSAM-3 AGENT	TRADITIONAL 3D	CROSS-PLANE (OURS)
Hardware	1–2 NVIDIA A100 (80GB)	Standard GPUs (T4/RTX)	NVIDIA A100 (40GB+)
Model Size	$\sim$850M + Billions (MLLM)	$\sim$3M (3D U-Net)	$\sim$650M (Frozen SAM)
Compute Demand	Heavy (Inference Bound)	Efficient (Training Bound)	Heavy (Memory Bound)
Inference	Iterative Agent Loop	Single Pass Volumetric	Simultaneous 3-View
Processing	2D Slices (Stacked)	Native 3D Patches	Orthogonal 2D Planes

4.2 The Resizing & Interpolation Overhead

We observed that the interpolation required to map anisotropic medical data (e.g., 240×155) to the isotropic input required by SAM (1024×1024) introduces significant overhead. The extreme upsampling ($155 \rightarrow 1024$) introduces blurring artifacts that the model struggles to segment, slowing convergence. Furthermore, a significant portion of GPU FLOPs are wasted processing interpolated "phantom pixels."

4.3 Memory Bandwidth and Striding Latency

Even when utilizing high-end hardware, training throughput remained low due to non-contiguous memory access. While Axial data is stored contiguously (row-major), Coronal/Sagittal extraction requires strided access (skipping W pixels), which defeats memory coalescing mechanisms. This results in a high rate of cache misses, shifting the bottleneck from Compute-bound to IO-bound. Consequently, effective training time per epoch increased by a factor of 4.5x compared to the baseline, making hyperparameter tuning (finding optimal λ) computationally prohibitive under a 100-compute-unit budget.

5 Resource and Architectural Comparison

To quantify the "iteration economy" discussed in our results, we compare the architectural and computational requirements of the Agentic Foundation Model approach (MedSAM-3 + MLLM) against the traditional 3D Deep Learning baseline (nnU-Net). Table 1 highlights the disparity in resource intensity that makes the former unfeasible for rapid experimental feedback loops.

As shown in Table 1, the resource overhead for the Foundation Model approach is orders of magnitude higher. While the Agentic workflow allows for open-ended reasoning, the requirement to run an iterative inference loop (where the MLLM "reflects" on the mask) multiplies the inference time per volume. For a dataset of hundreds of MRI scans, this latency is prohibitive compared to the single-pass efficiency of a specialized 3D U-Net.

6 Conclusion

Our findings suggest that for 3D medical segmentation, the current generation of Vision Foundation Models occupies an "unhappy medium." They are too computationally heavy to be adapted via multi-view consistency, yet they are not naturally 3D-consistent enough to replace simple U-Nets. We conclude that future research for resource-constrained labs should pivot away from heavy loss-based adaptations of Foundation Models. Instead, simpler architectural priors—such as consistency-aware training on lightweight networks—offer a far superior return on compute investment.

REFERENCES

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

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.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chengwei You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.