The Cost of Consistency: 3D Volumetric Medical Segmentation

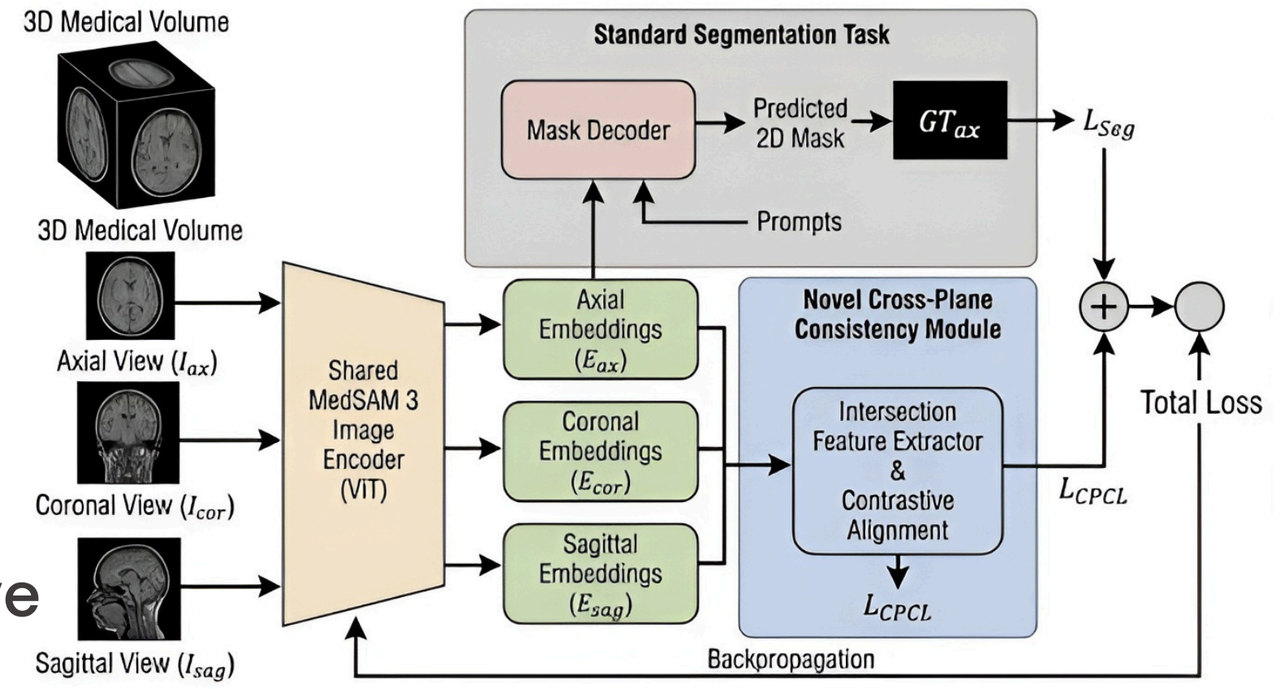
Bridging the Gap Between MedSAM-3 and nnU-Net

Dr. Ramamoorthy S, Madhu Shree Aravindan, Aaditi V Bajpai
SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamil Nadu, India

1 The Problem: 2D Limitation

Vision Foundation Models (VFMs) like MedSAM-3 excel at 2D tasks but fail to capture the 3D geometric consistency required for volumetric medical segmentation.

- **Status Quo:** Traditional architectures like nnU-Net remain the "gold standard" due to their native structural coherence and efficiency.
- **The Challenge:** VFMs process volumetric data as independent 2D slices, ignoring vital spatial dependencies.



2 Our Approach: Cross-Plane Contrastive Learning

Tri-Siamese Architecture: Three parallel streams share weights to process orthogonal views (Axial, Coronal, Sagittal) intersecting at a target voxel.

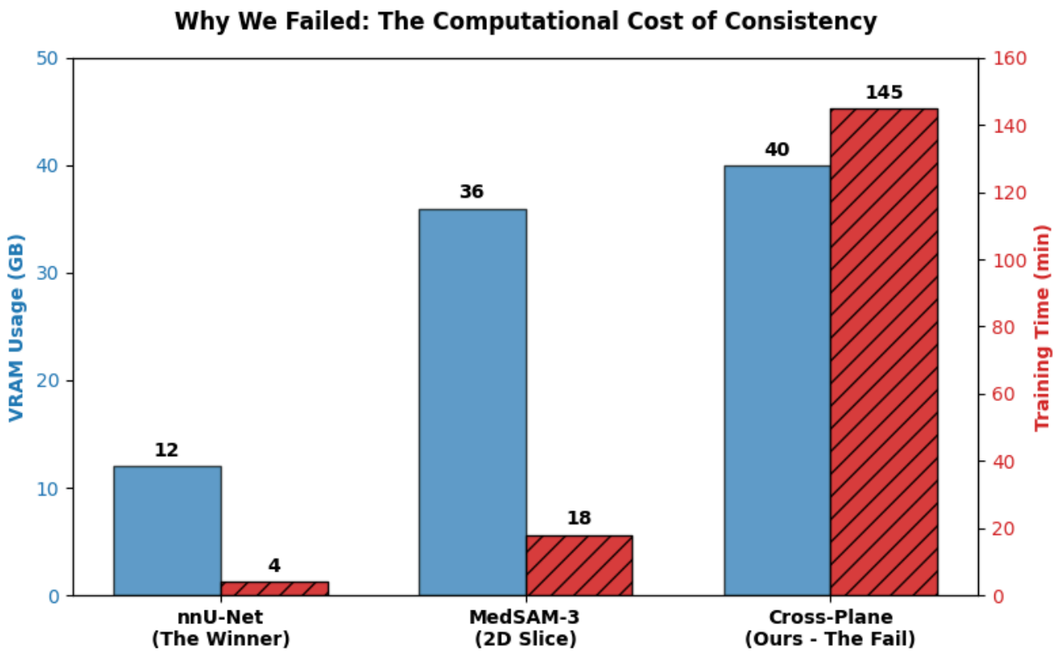
Intersection Consistency Loss: The model is penalized whenever the three views disagree on the label of a shared voxel.

Mechanism: Uses a frozen SAM-3 image encoder with LoRA for domain adaptation.

3 Negative Result

Despite being mathematically sound, this framework introduces a prohibitive "iteration penalty" in resource-constrained environments.

Feature	MedSAM-3 (Agent)	3D nnU-Net (Traditional)	Cross Plane (Ours)
Hardware	NVIDIA A100 (80 GB)	Standard GPUs (T4/RTX)	NVIDIA A100 (40 GB)
Model Size	~850M	~3M	~650M
Compute Demand	Heavy (Inference Bound)	Efficient (Training Bound)	Heavy (Memory Bound)
Inference Processing	Iterative Agent looped 2D slices	Single Pass Volumetric Native 3D Patches	Simultaneous 3-View Orthogonal 2D Planes



4 Three Bottlenecks

- **Activation Stacking:** Simultaneous ViT-H backbones exceed VRAM limits, causing OOM errors on 16GB GPUs.
- **Interpolation Overhead:** Resizing anisotropic data to 1024 *X* 1024 creates "phantom pixels" and blurring artifacts.
- **Memory Bandwidth:** Strided access for Coronal/Sagittal views causes high cache misses, making the process IO-bound.

5 Takeaway

Current Vision Foundation Models are too heavy for multi-view consistency adaptation and not naturally 3D-consistent enough to replace U-Nets.

Future research for resource-constrained labs should pivot toward simpler architectural priors on lightweight networks for a superior return on compute investment.

Related literature

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