Steer the Sampling, Not the Kernel Grid: Geometry-Guided Sampling Operator for Volumetric Segmentation

Sizhe Wang[1(✉)], Himashi Peiris[1], and Zhaolin Chen[1(✉)]

Department of Data Science & AI, Faculty of Information Technology, Monash University, Clayton, VIC, Australia
{sizhe.wang,zhaolin.chen}@monash.edu

Abstract. Accurate 3D segmentation is central to quantitative lesion assessment and anatomy mapping for clinical planning and follow-up. Thin, elongated, and fine anatomical/pathological structures (*e.g.*, vessels) are a particular stress case: small boundary error can disrupt apparent continuity and bias quantitative measurements. In encoder–decoder networks (*e.g.*, U-Net), repeated downsampling and fixed-grid convolution blur/alias fine structures and weaken orientation cues, so early mistakes propagate across scales. We propose a geometry-guided local operator that *steers where features are sampled* (rather than deforming convolutional kernels) under a single formulation for both feature refinement (stride 1) and resolution reduction (stride > 1). At each voxel, it predicts a local orientation and bounded step sizes, samples symmetrically along these directions, and transforms paired samples into compact geometric/boundary cues with lightweight mixing; a cross-scale consensus aligns encoder and decoder features at skip connections to reduce geometric mismatch. Replacing all stride 1 and stride 2 operators in a 3D U-Net yields consistent improvements on BraTS, MSD Hepatic Vessel, and TDSC-ABUS, with notably better boundary metrics (*e.g.*, BraTS Dice 86.1 → 88.9, HD95 7.1 → 6.2; TDSC-ABUS HD95 39.1 → 27.8) while substantially reducing parameters (2.3M→ 0.8M). We further demonstrate that the operator plugs into other backbones (*e.g.* nnU-Net, Swin-UNETR, and MedNeXt) without changing their macro-architectures while providing consistent performance gains. Code available at GeoSample repo.

Keywords: 3D medical image segmentation · geometry-guided sampling · downsampling operator · SO(3) frame field · central differences.

1 Introduction

Accurate volumetric segmentation is a prerequisite for many downstream clinical and scientific workflows, including lesion quantification, preoperative planning, and longitudinal follow-up [21,12]. Encoder–decoder architectures in the U-Net family remain dominant because they combine multi-scale context aggregation with fine localisation through skip connections [23,4]; nnU-Net further

demonstrates strong performance across diverse biomedical tasks using largely standardised backbones and robust training recipes [14]. Despite these successes, thin, elongated, or anisotropic structures remain fragile, and more broadly any task requiring precise boundary localisation can suffer when resolution changes distort local geometry [25]. Such targets often occupy only a few voxels, so small boundary shifts can break continuity and bias quantitative measurements [13]. *Therefore, rather than redesigning the backbone, we fundamentally rethink the convolution/pooling primitives that sample, aggregate, and downsample features.*

A key issue is that existing stride-based refinement and downsampling indiscriminately aggregate voxel information along fixed grids, suppressing or aliasing high-frequency details and distorting fine structures [2,22], an issue exacerbated by extreme foreground sparsity and limited sampling support. Though anti-aliasing strategies partially address instability under downsampling [29], they remain largely orientation-agnostic and fail to provide a unified approach that preserves boundary geometry and directional cues across scales.

Prior work mitigates cross-scale degradation from several directions. Transformer based and hybrid volumetric segmenters strengthen global context modelling [10,9,17,30,27,20], while other studies target cross-scale calibration and alignment within U-Net pipelines [28]. However, these approaches operate on already-sampled features and do not reconsider the geometric bias introduced by fixed-grid refinement and downsampling.

Adaptive sampling operators move beyond fixed-grid convolutional aggregation by learning sampling offsets, *e.g.*, deformable convolutions [6,31] and volumetric variants [16], as well as sparse large-kernel sampling designs such as OBELISK [11]. These methods primarily use learned sampling to gather features from shifted locations; in contrast, we treat sampling as a way to measure local geometric information before feature aggregation.

Differentiable warping modules and equivariant or steerable architectures incorporate geometric transformations or symmetry constraints [15,7,5,26,8]. However, these existing approaches are either anchored in fixed-grid convolutional backbones or impose global equivariance without explicitly modelling boundary-aligned sampling for both refinement and downsampling. Our focus is different: we keep features on the voxel grid and use local geometry to guide how neighbouring information is measured before refinement and downsampling.

We propose *GeoSample*, a *geometry-guided* local operator as a unified, plug-in module for volumetric encoder–decoder networks. Rather than deforming kernels, GeoSample predicts a voxel-wise local 3D rotation frame (an element of SO(3)) with adaptive step sizes to steer structured symmetric sampling. The resulting average/difference signals are converted into step-size normalised differential cues that compactly summarise gradient- and curvature-like information. To reduce geometric mismatch between decoder and encoder features, we further employ a rotation-consistent cross-scale consensus mechanism.

Our contributions are threefold: **(i)** a novel, geometry-guided local *operator* (*GeoSample*) that replaces fixed-grid convolution/pooling and unifies stride 1 refinement and stride > 1 downsampling under a single formulation; **(ii)** a

rotation-constrained local frame with adaptive step sizes that enables *structured symmetric sampling* and extracts compact geometry-aware signals, improving segmentation fidelity for thin or anisotropic targets; and **(iii)** a rotation-consistent *Consensus Field* module that aligns geometry at the skip connection, improving robustness across backbones. Fig. 1 provides an overview; we next detail the formulation in Sec. 2.

2 Methodology

2.1 Overview and integration

Given an input feature volume $X \in \mathbb{R}^{C\times D\times H\times W}$, the geometry signal used by GeoSample is the learned voxel-wise field $\mathcal{H}(X)$ that guides our operator Ψ. The field defines local sampling directions and distances, and is used to steer symmetric sampling with lightweight token mixing. This yields a drop-in local operator that replaces stride 1 refinement blocks and stride > 1 downsampling in 3D encoder–decoder segmenters (Fig. 1).

Stride 1 ***refinement (replacement of conv blocks).*** For feature refinement without changing spatial resolution, we generate a geometry-conditioned residual update:

$$Y = X + \Psi_1\big(X; \mathcal{H}(X)\big). \tag{1}$$

Stride > 1 ***downsampling (replacement of strided conv/pooling).*** For spatial reduction, we adopt average pooling as a stable low-pass path and augment it with a geometry-conditioned compensation term:

$$Y_\downarrow = \mathrm{AvgPool}_s(Y) + \Psi_{\downarrow,s}\big(X; \mathcal{H}(X)\big), \qquad s \neq (1,1,1). \tag{2}$$

Consensus Field at skip connection. At each decoder stage, we fuse the geometric field from the upsampled decoder feature with the corresponding encoder skip feature, and apply the stride 1 operator using the fused field before concatenation. This reduces mismatch across the skip connection (Sec. 2.6).

2.2 Geometry field prediction

All operations are applied pointwise at voxel locations; we omit the spatial index for readability. A geometry head H_{geo} predicts the field used for sampling:

$$\mathcal{H}(X) = \{R(X), \mathbf{r}(X)\},\ R \in SO(3),\ \mathbf{r} = [r_1, \ldots, r_K]^\top \in [r_{\min}, r_{\max}]^K. \tag{3}$$

Internally, we represent R by a unit quaternion $q \in \mathbb{S}^3$ (with $R = \mathcal{R}(q)$); this is used only for rotation fusion in Sec. 2.6. Let $\{\mathbf{e}_k\}_{k=1}^3$ be the canonical axes of the *image Cartesian* coordinate system. We take $\mathbf{u}_k = R\,\mathbf{e}_k$ as the first $K \leq 3$ unit directions ($K = 3$ in all experiments, yielding a full local 3D frame for voxel-wise measurement) and define sampling offsets

$$\boldsymbol{\delta}_k = r_k\,\mathbf{u}_k \in \mathbb{R}^3. \tag{4}$$

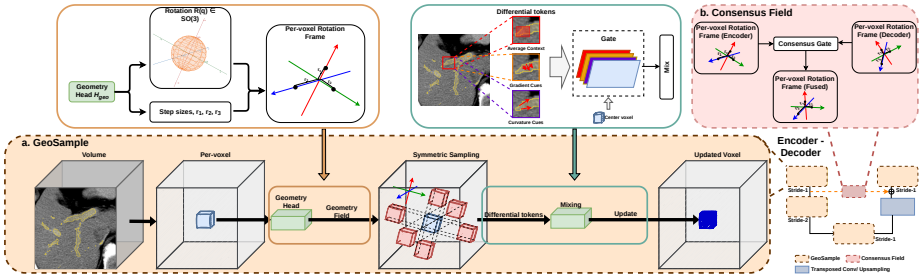

Fig. 1. Overall workflow of the proposed modules. (a) *GeoSample*: per-voxel geometry field, symmetric sampling, and differential-token mixing. (b) *Consensus Field*: rotation-consistent fusion of geometry fields from the encoder skip and upsampled decoder features for pre-skip alignment. Shown on a U-Net, *GeoSample* replaces stride-1 blocks and stride-2 downsampling; *Consensus Field* is applied at skip connection.

Importantly, $\mathbf{u}_k$ are expressed in the same coordinate system as the voxel grid; the frame is used only to steer sampling directions (we do not reparameterise features into a new coordinate system).

2.3 Symmetric sampling and oriented differentials

We view X as a continuous field $x(\cdot)$ via differentiable trilinear interpolation. For each direction $\mathbf{u}_k$ and step size r_k (both predicted at voxel location $\mathbf{p}$), we sample symmetrically:

$$x_k^+ = x(\mathbf{p} + \boldsymbol{\delta}_k), \qquad x_k^- = x(\mathbf{p} - \boldsymbol{\delta}_k), \tag{5}$$

where $\mathbf{p} \in \mathbb{R}^3$ is a voxel coordinate. In the continuous domain, the first- and second-order directional differentials along $\mathbf{u}_k$ are

$$\partial_{\mathbf{u}_k} x = \mathbf{u}_k^\top \nabla x, \qquad \partial^2_{\mathbf{u}_k} x = \mathbf{u}_k^\top H_x \, \mathbf{u}_k, \tag{6}$$

and we approximate them by symmetric finite differences with step-size normalisation:

$$a_k = \tfrac{1}{2}(x_k^+ + x_k^-), \qquad d_k = \frac{x_k^+ - x_k^-}{2(r_k + \epsilon)}, \qquad s_k = \frac{x_k^+ - 2x(\mathbf{p}) + x_k^-}{(r_k + \epsilon)^2}, \tag{7}$$

where $\epsilon > 0$ is a small constant for numerical stability. Here, a_k is an even (context) response, d_k is an odd (directional-change) response, and s_k captures even (2nd-order) curvature-like change. The symmetry makes the odd/even components explicit, while step-size normalisation makes the cues comparable across scales.

2.4 Compact differential tokens and mixing

From $\{d_k\}$ we form a gradient-like cue G expressed in the *image Cartesian* axes, and from $\{s_k\}$ we form a curvature/Laplacian-like cue L:

$$G = \sum_{k=1}^{K} \mathbf{u}_k \otimes d_k \in \mathbb{R}^{3\times C}, \qquad L = \tfrac{1}{K}\sum_{k=1}^{K} s_k \in \mathbb{R}^{C}. \tag{8}$$

Sign stability. One key design is symmetric paired sampling, which makes the directional cue *sign-invariant by construction*: $\mathbf{u}_k \to -\mathbf{u}_k \Rightarrow d_k \to -d_k$ and $(-\mathbf{u}_k)\otimes(-d_k) = \mathbf{u}_k \otimes d_k$. This yields stable directional tokens without frame-sign jitter.

We denote the centre sample by x_0 and collect tokens

$$\mathcal{T} = \Big[\underbrace{x_0}_{\text{centre}}, \underbrace{a_1, \ldots, a_K}_{\text{even / context}}, \underbrace{G_x, G_y, G_z}_{\text{odd / gradient-like}}, \underbrace{L}_{\text{even / curvature-like}} \Big]. \tag{9}$$

We apply a lightweight token-wise gating followed by $1 \times 1 \times 1$ mixing to obtain an update Δ:

$$\Delta = \mathrm{Mix}\big(\mathrm{Gate}(\mathcal{T}) \odot \mathcal{T}\big), \qquad \Psi_1(X; \mathcal{H}) = \Delta. \tag{10}$$

The specific realisation of $\mathrm{Gate}(\cdot)$ and $\mathrm{Mix}(\cdot)$ is summarised in Sec. 2.7.

2.5 Downsampling with directional-signal preservation

For downsampling, we first compute the stride 1 refined feature Y in (1). Since odd directional signals can undesirably cancel under naive averaging, we deliberately preserve their magnitude by pooling $|G|$:

$$\bar{\mathcal{T}}_{\downarrow} = \mathrm{AvgPool}_s\Big(\big[X, \underbrace{a_1, \ldots, a_K}_{\text{even terms}}, \underbrace{|G_x|, |G_y|, |G_z|}_{\text{odd energy}}, \underbrace{L}_{\text{even / curvature-like}} \big] \Big). \tag{11}$$

and predict a coarse compensation term

$$\Psi_{\downarrow,s}(X; \mathcal{H}) = \mathrm{Mix}_{\downarrow,s}\big(\mathrm{Gate}_{\downarrow,s}(\bar{\mathcal{T}}_{\downarrow}) \odot \bar{\mathcal{T}}_{\downarrow}\big). \tag{12}$$

2.6 Rotation-consistent Consensus Field and integration

To reduce geometric mismatch between upsampled decoder features and encoder skip features, we fuse two local *rotation fields*: a current field $\mathcal{H} = \{R, \mathbf{r}\}$ from the upsampled decoder features with a reference field $\mathcal{H}^* = \{R^*, \mathbf{r}^*\}$ predicted from encoder skip features, where $R, R^* \in \mathrm{SO}(3)$ and $\mathbf{r}, \mathbf{r}^* \in [r_{\min}, r_{\max}]^K$. For rotation fusion we use their unit-quaternion representations $q, q^* \in \mathbb{S}^3$ (with $R = \mathcal{R}(q)$ and $R^* = \mathcal{R}(q^*)$).

We compute a consensus gate $\omega \in (0,1)$ based on step-size statistics and quaternion similarity:

$$c = |\langle q, q^* \rangle|,\ \bar{r} = \frac{1}{K}\sum_{k=1}^{K} r_k,\ \bar{r}^* = \frac{1}{K}\sum_{k=1}^{K} r_k^*,\ \omega = \sigma(\mathrm{Conv}_{1\times1\times1}([\bar{r}, \bar{r}^*, c])). \quad (13)$$

Rotations are fused by quaternion spherical interpolation and step sizes are blended with the same gate:

$$\bar{q} = \mathrm{SLERP}(q, q^*; \omega), \qquad \bar{\mathbf{r}} = \omega \mathbf{r} + (1-\omega)\mathbf{r}^*, \qquad \bar{R} = \mathcal{R}(\bar{q}). \quad (14)$$

The fused field $\bar{\mathcal{H}} = \{\bar{R}, \bar{\mathbf{r}}\}$ defines the consensus directions and step sizes (via Sec. 2.2) and is used to steer the stride 1 refinement on decoder features at skip connection.

2.7 Implementation notes

We parameterise R with a unit quaternion (from a $1 \times 1 \times 1$ head H_{geo}) and bound r_k to $[r_{\min}, r_{\max}]$ via a sigmoid. Gate is a grouped $1 \times 1 \times 1$ conv (one group per token) followed by a $1 \times 1 \times 1$ Mix conv; stride 1 and downsampling use separate heads. For stability, we stop gradients through r_k in the $1/(r_k + \epsilon)$ factors.

3 Experiments

Datasets. We evaluate our method on three public datasets spanning three different modalities: **BraTS** (MRI) [3,19], **MSD HepaticVessel** (CT)[1], and **TDSC-ABUS** (Ultrasound) [18]. HepaticVessel serves as a thin-structure continuity stress test, while BraTS and TDSC-ABUS assess general lesion/tumour segmentation under distinct appearance statistics. We use a fixed-seed 75/10/15 train/val/test split for all methods.

Training and evaluation. All runs use the same training pipeline including preprocessing, postprocessing, optimisation, loss, patch-based training and sliding-window inference; patch sizes for Params (model size) / FLOPs (compute cost; $\approx$ number of arithmetic operations) are $128 \times 192 \times 128$ (BraTS) and 128^3 (HepaticVessel / TDSC-ABUS). We report mean Dice and boundary metrics (HD95 and ASSD).

Operator-level results. Under a fixed 3D U-Net macro-architecture with base channel 16, we compare the baseline convolution, Deformable Convolution (DCN) v1/v2 [6,31], Dynamic Downsampling (DCD) [28], and the proposed operator by replacing all stride 1 blocks and stride 2 reductions with the respective operator (Table 1). For BraTS, metrics are averaged over enhanced tumour, tumour core and whole tumour; for Hepatic Vessel, we calculated mean metrics on vessel and tumour; for TDSC-ABUS, we report metrics on tumour. Our operator achieves the best mean Dice/HD95/ASSD (71.5/22.8/4.72). It is best on BraTS

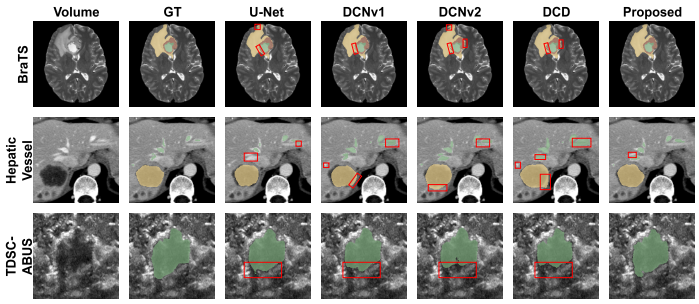

Fig. 2. Qualitative comparison on representative cases from BraTS (top), MSD Hepatic Vessel (middle), and TDSC-ABUS (bottom). Shown are input, ground truth, and predictions from the baseline U-Net, DCNv1/v2, Dynamic Downsampling (DCD), and the proposed operator; red boxes mark noticeable segmentation errors reduced by our method.

Table 1. Operator-level comparison across datasets (top: Dice/HD95/ASSD). Bottom: model complexity (Params/FLOPs) under evaluation patch sizes.

	MSD Hepatic Vessel			BraTS			TDSC-ABUS			Mean		
Model	Dice↑	HD95↓	ASSD↓	Dice↑	HD95↓	ASSD↓	Dice↑	HD95↓	ASSD↓	Dice↑	HD95↓	ASSD↓
Baseline U-Net	53.5	42.8	8.87	86.1	7.1	1.15	64.4	39.1	7.31	68.0	29.7	5.78
Deformable Conv v1	57.2	37.3	7.93	88.3	6.4	1.07	**67.5**	33.1	5.83	71.0	25.6	4.94
Deformable Conv v2	56.7	36.3	8.21	88.3	6.5	1.33	66.3	35.4	6.06	70.4	26.1	5.20
Dynamic Downsampling	56.4	34.9	**7.40**	88.4	6.8	1.14	67.2	33.9	6.87	70.6	25.2	5.13
Proposed	**58.2**	**34.3**	7.43	**88.9**	**6.2**	**0.95**	67.4	**27.8**	**5.78**	**71.5**	**22.8**	**4.72**

Input size	Baseline U-Net		Deformable Conv v1		Deformable Conv v2		Dynamic Downsampling		Proposed	
	Params (M)↓	FLOPs (G)↓	Params (M)↓	FLOPs (G)↓	Params (M)↓	FLOPs (G)↓	Params (M)↓	FLOPs (G)↓	Params (M)↓	FLOPs (G)↓
128^3	2.3	131.2	2.5	137.9	2.6	146.9	2.7	256.7	0.8	69.9
$128 \times 192 \times 128$	2.3	194.8	2.5	203.3	2.6	209.3	2.7	384.5	0.8	108.9

(88.9/6.2/0.95), attains the best Dice/HD95 on Hepatic Vessel (58.2/34.3) with near-best ASSD (7.43 vs. 7.40), and markedly improves boundary metrics on TDSC-ABUS (HD95/ASSD 27.8/5.78) while keeping Dice comparable (67.4), indicating the most stable cross-modality improvements overall.

Computational Complexity. Table 1 (bottom) reports Params and FLOPs in the U-Net setting. Our operator reduces Params from 2.3M to 0.8M and FLOPs from 194.8G to 108.9G. By contrast, DCNv1/v2 slightly increases both, while Dynamic Downsampling raises FLOPs to 256.7G/384.5G. This efficiency comes from using a small number of symmetric samples (*e.g.*, $K=3$ paired directions) and compact token mixing with $1\times1\times1$ gating, rather than dense grids or heavy dynamic branches.

Architecture-level plug-in results. On TDSC-ABUS we plug the proposed operator into nnU-Net, Swin-UNETR, and MedNeXt (base version) [14,9,24] (Table 2). For nnU-Net we replace all stride 1 and stride > 1 blocks, whereas for Swin-UNETR and MedNeXt we preserve backbone-specific refinement (self-attention in Swin-UNETR and ConvNeXt-style blocks in MedNeXt) and re-

Table 2. Architecture-level plug-in validation on TDSC-ABUS dataset.

Backbone	Variant	Overlap			Boundary		Efficiency	
		Dice↑	Sensitivity↑	Specificity↑	HD95↓	ASSD↓	Params (M)↓	FLOPs (G)↓
nnU-Net	Baseline	68.0	72.4	98.8	28.4	5.24	30.8	1250.0
	+ Proposed	**72.8**	75.3	98.5	**25.6**	**5.03**	17.4	750.2
Swin-UNETR	Baseline	69.1	68.0	98.6	30.2	4.97	63.5	751.5
	+ Proposed	**70.9**	74.2	98.7	**27.7**	**4.88**	64.8	802.7
MedNeXt	Baseline	66.8	69.5	99.1	29.4	5.38	2.7	50.1
	+ Proposed	**69.5**	75.5	98.2	**24.2**	**4.89**	2.9	57.0

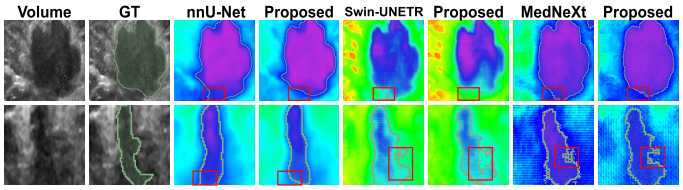

Fig. 3. Qualitative plug-in results on TDSC-ABUS (red boxes: improved regions).

place only downsampling; all variants additionally apply *Consensus Field* at skip connection (Sec. 2.6). Across backbones, the plug-in improves overlap and boundary metrics (Dice +1.8–+4.8, Sens. +2.9–+6.2; HD95 −2.5–−5.2, ASSD −0.09–−0.49), with minor specificity changes (≥98.2%). The largest gain is on nnU-Net (Dice 68.0→72.8) while reducing cost (Params 30.8M→17.4M; FLOPs 1250G→750G); Swin-UNETR and MedNeXt also improve with modest FLOPs changes, consistent with their already-efficient downsampling (patch merging / depthwise separable convolutions). Fig. 3 shows fewer leakage/boundary errors, and Fig. 4 indicates improved volumetry (RVE distributions closer to 0 with fewer outliers and points closer to $y = x$). Overall, the improved boundaries and volumetry suggest more reliable downstream quantitative assessment across backbones.

Ablation Study. We test the effectiveness of the geometry-aware modules. Under the same 3D U-Net setting using TDSC-ABUS data, removing the differentials (G, L; geometry-aware gradient/curvature cues) drops Dice/HD95/ASSD from 67.4/27.8/5.78 to 54.3/51.0/10.71, and removing the *Consensus Field* (skip alignment to reduce geometry mismatch) drops to 58.4/48.2/8.52. This confirms that geometric differential terms drive boundary quality, with consensus alignment providing additional cross-scale robustness.

4 Conclusion

We have introduced *GeoSample*, a geometry-guided operator for volumetric encoder–decoder segmentation that predicts local orientations to steer symmetric sampling and extract geometric signals for refinement and downsampling. Across three public benchmarks and plug-in validation, it consistently improves

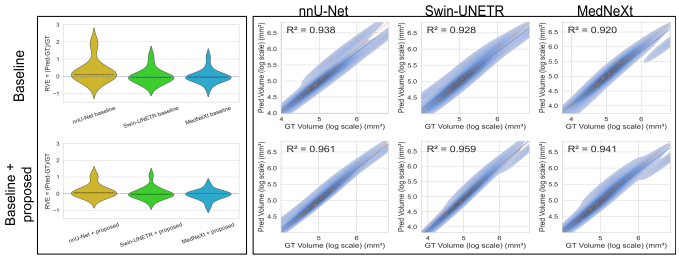

Fig. 4. Volumetric agreement on TDSC-ABUS. Left: relative volume error (RVE) distributions for different backbones. Right: predicted vs. ground-truth tumour volumes (log scale) for baseline and +Proposed; dashed line denotes $y = x$.

segmentation quality, especially boundary metrics, while reducing the number of parameters in the U-Net setting.

Future work can include topology-specific evaluations, scalability test on larger backbones, and reducing the memory/wall-clock time overhead of interpolation-based sampling and improve early-training stability.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., et al.: The medical segmentation decathlon. Nature communications **13**(1), 4128 (2022)
2. Azulay, A., Weiss, Y.: Why do deep convolutional networks generalize so poorly to small image transformations? Journal of Machine Learning Research **20**(184), 1–25 (2019)
3. Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F.C., Pati, S., et al.: The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv preprint arXiv:2107.02314 (2021)
4. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: International conference on medical image computing and computer-assisted intervention. pp. 424–432. Springer (2016)
5. Cohen, T., Welling, M.: Group equivariant convolutional networks. In: International conference on machine learning. pp. 2990–2999. PMLR (2016)
6. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 764–773 (2017)
7. Frangi, A.F., Niessen, W.J., Vincken, K.L., Viergever, M.A.: Multiscale vessel enhancement filtering. In: International conference on medical image computing and computer-assisted intervention. pp. 130–137. Springer (1998)

8. Fuchs, F., Worrall, D., Fischer, V., Welling, M.: Se (3)-transformers: 3d roto-translation equivariant attention networks. Advances in neural information processing systems **33**, 1970–1981 (2020)
9. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI brainlesion workshop. pp. 272–284. Springer (2021)
10. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 574–584 (2022)
11. Heinrich, M.P., Oktay, O., Bouteldja, N.: Obelisk-net: Fewer layers to solve 3d multi-organ segmentation with sparse deformable convolutions. Medical image analysis **54**, 1–9 (2019)
12. Hesamian, M.H., Jia, W., He, X., Kennedy, P.: Deep learning techniques for medical image segmentation: achievements and challenges. Journal of digital imaging **32**(4), 582–596 (2019)
13. Hu, X., Li, F., Samaras, D., Chen, C.: Topology-preserving deep image segmentation. Advances in neural information processing systems **32** (2019)
14. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
15. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. Advances in neural information processing systems **28** (2015)
16. Lee, H.H., Liu, Q., Yang, Q., Yu, X., Bao, S., Huo, Y., Landman, B.A.: Deformux-net: Exploring a 3d foundation backbone for medical image segmentation with depthwise deformable convolution. arXiv preprint arXiv:2310.00199 (2023)
17. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
18. Luo, G., Xu, M., Chen, H., Liang, X., Tao, X., Ni, D., Jeong, H., Kim, C., Stock, R., Baumgartner, M., et al.: Tumor detection, segmentation and classification challenge on automated 3d breast ultrasound: The tdsc-abus challenge. arXiv preprint arXiv:2501.15588 (2025)
19. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging **34**(10), 1993–2024 (2014)
20. Peiris, H., Hayat, M., Chen, Z., Egan, G., Harandi, M.: A robust volumetric transformer for accurate 3d tumor segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 162–172. Springer (2022)
21. Pham, D.L., Xu, C., Prince, J.L.: Current methods in medical image segmentation. Annual review of biomedical engineering **2**(1), 315–337 (2000)
22. Ribeiro, A.H., Schön, T.B.: How convolutional neural networks deal with aliasing. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2755–2759. IEEE (2021)
23. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)

24. Roy, S., Koehler, G., Ulrich, C., Baumgartner, M., Petersen, J., Isensee, F., Jaeger, P.F., Maier-Hein, K.H.: Mednext: transformer-driven scaling of convnets for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 405–415. Springer (2023)
25. Sinclair, B., Pham, W., Vivash, L., Moses, J., Lynch, M., Dorfman, K., Marotta, C., Koh, S., Bunyamin, J., Rowsthorn, E., et al.: Perivascular space identification nnunet for generalised usage (pingu). Medical Image Analysis p. 103903 (2025)
26. Weiler, M., Geiger, M., Welling, M., Boomsma, W., Cohen, T.S.: 3d steerable cnns: Learning rotationally equivariant features in volumetric data. Advances in Neural information processing systems **31** (2018)
27. Xie, Y., Zhang, J., Shen, C., Xia, Y.: Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 171–180. Springer (2021)
28. Yang, J., Marcus, D.S., Sotiras, A.: Dynamic u-net: Adaptively calibrate features for abdominal multiorgan segmentation. In: Medical Imaging 2025: Computer-Aided Diagnosis. vol. 13407, pp. 326–334. SPIE (2025)
29. Zhang, R.: Making convolutional networks shift-invariant again. In: International conference on machine learning. pp. 7324–7334. PMLR (2019)
30. Zhou, H.Y., Guo, J., Zhang, Y., Han, X., Yu, L., Wang, L., Yu, Y.: nnformer: Volumetric medical image segmentation via a 3d transformer. IEEE transactions on image processing **32**, 4036–4045 (2023)
31. Zhu, X., Hu, H., Lin, S., Dai, J.: Deformable convnets v2: More deformable, better results. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9308–9316 (2019)