MAGEFormer: Learning Metric-Consistent Representations for Anisotropic CT Segmentation

Jiaying Li and Paolo Remagnino

Department of Computer Science, Durham University, Durham, UK
jiaying.li2@durham.ac.uk, paolo.remagnino@durham.ac.uk

Abstract. Vision Transformers (ViTs) have shown strong performance in volumetric segmentation, but their effectiveness on clinical CT is limited by an isotropic Euclidean lattice assumption. This conflicts with anisotropic CT acquisition, leading to two key issues: (1) a metric mismatch between voxel indices and physical anatomy, and (2) accuracy degradation from isotropic resampling. To address this, we propose MAGEFormer, a geometry-calibrated framework that embeds physical metric constraints directly into representation learning. Our method introduces Metric-Adaptive Spatial Embedding (MASE) to calibrate positional frequencies using voxel spacing, Geometry-Constrained Attention (GCA) to suppress physically implausible feature correlations, and Geometric View Voting (GVV) to reduce discretization bias during inference. We evaluate MAGEFormer on two multi-organ abdominal CT benchmarks, BTCV and FLARE 22, under a unified protocol against strong CNN and Transformer-based baselines. MAGEFormer achieves the strongest boundary accuracy among the compared methods, with 10.58 mm HD95 on BTCV and 3.40 mm HD95 on FLARE 22, and shows consistent gains in Dice under the same protocol. These results show that geometry-aware internal calibration is more effective than relying on conventional isotropic preprocessing alone for anisotropic CT segmentation.

Keywords: Medical image segmentation · Geometric calibration · Anisotropic CT

1 Introduction

The rapid adoption of Vision Transformers (ViTs) [15, 3] has advanced volumetric medical image segmentation by improving long-range dependency modeling over traditional CNNs [12], with architectures such as nnFormer [20] further integrating convolution and self-attention for 3D volumes. However, most ViT-based architectures implicitly assume an isotropic Euclidean lattice ($\Delta x \approx \Delta y \approx \Delta z$), a geometric inductive bias that is often misaligned with the acquisition physics of clinical Computed Tomography (CT).

In routine abdominal CT, the through-plane spacing (s_z) is frequently much coarser than the in-plane spacing (s_{xy}) due to radiation dose and acquisition speed constraints [19, 9], making the data inherently anisotropic. Without explicit calibration, self-attention may treat a physically distant voxel along the

sparse z-axis as equally local to a neighbor in the dense xy-plane. This metric mismatch can hinder scale-consistent representation learning, especially for small or low-contrast structures (e.g., adrenal glands, pancreas tail) whose boundary quality is highly sensitive to sampling geometry [4].

Existing methods address this issue in two main ways. Implicit adaptation via learnable or input-conditioned positional encodings [2] and relative position biases [17] (e.g., Swin-UNETR [6]) is effective in practice, yet these are parameterized on voxel indices rather than physical distances and do not explicitly enforce metric consistency. Explicit architectural adaptation, such as thickness-aware operator design [18], adapts network behavior to slice thickness but does not embed metric constraints into the learned representations themselves. Our goal is complementary: we treat metric consistency as a representation-level design principle that benefits from being enforced across encoding, feature interaction, and inference, rather than addressed solely through isotropic preprocessing or localized architectural heuristics. Throughout, by isotropic preprocessing we mean not a separately tuned baseline but the conventional resampling/normalization setting in which models operate on voxel-index neighborhoods as if the lattice were geometrically uniform.

We propose MAGEFormer, a geometry-calibrated framework that models anisotropy as an intrinsic physical property within the representation learning process. It integrates three components: Metric-Adaptive Spatial Embedding (MASE), which calibrates positional frequencies using voxel spacing so that positional encoding reflects physical distance; Geometry-Constrained Attention (GCA), which reformulates skip-feature interaction with a metric-aware geometric bias to suppress physically implausible correlations; and Geometric View Voting (GVV), which reduces discretization bias during inference via deterministic sub-voxel view aggregation. Under a unified evaluation protocol on the BTCV and FLARE 22 benchmarks, MAGEFormer achieves the strongest boundary accuracy among compared CNN and Transformer-based baselines, with consistent gains in Dice. Moreover, its advantage increases with anisotropy severity, supporting our central hypothesis that internal geometry calibration is more effective than relying on isotropic preprocessing alone.

2 Method

Metric consistency, the requirement that learned representations reflect physical distances rather than voxel indices, must be enforced at multiple pipeline stages. We operationalize this in MAGEFormer, a geometry-calibrated framework with three components addressing metric consistency at complementary levels: Metric-Adaptive Spatial Embedding (MASE) at positional encoding, Geometry-Constrained Attention (GCA) at cross-scale feature interaction, and Geometric View Voting (GVV) at inference-time discretization. An overview is shown in Fig. 1.

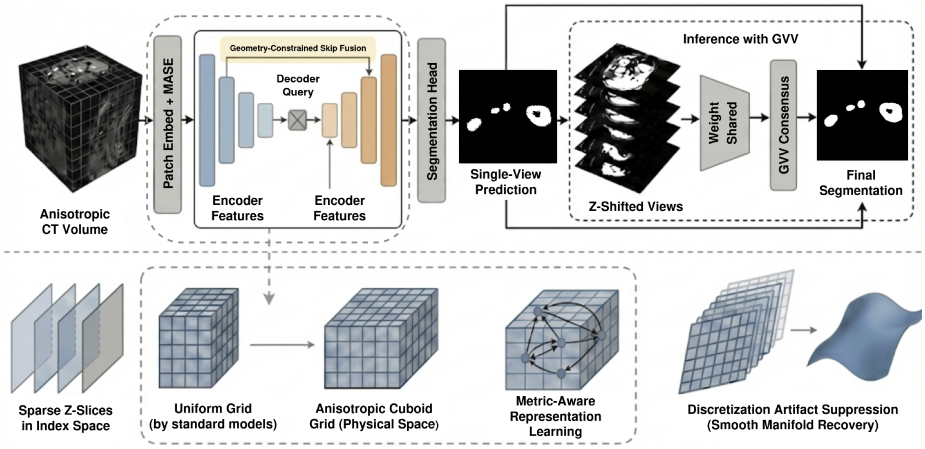

Fig. 1. MAGEFormer: Architecture Overview and Geometric Motivation

2.1 Problem Formulation: Metric Mismatch in Anisotropic Grids

Let a CT volume be sampled on a lattice $\Lambda \subset \mathbb{Z}^3$ with voxel index $\mathbf{p} = (p_x, p_y, p_z)$ and spacing $\mathbf{s}^{(0)} = (s_x^{(0)}, s_y^{(0)}, s_z^{(0)})$, typically anisotropic ($s_z^{(0)} \gg s_x^{(0)}, s_y^{(0)}$), yet most models implicitly assume isotropic geometry. After reorientation to a canonical frame, the physical metric is $G^{(0)} = \mathrm{diag}((s_x^{(0)})^2, (s_y^{(0)})^2, (s_z^{(0)})^2)$ and the physically meaningful distance is

$$d_{\mathcal{S}}(\mathbf{p}, \mathbf{q}) = \sqrt{(\mathbf{p} - \mathbf{q})^\top G^{(0)} (\mathbf{p} - \mathbf{q})}. \tag{1}$$

Equal index offsets thus do not imply equal physical displacements, especially along the sparse through-plane axis.

MAGEFormer adopts a multi-scale U-shaped Transformer backbone. The mismatch propagates through the hierarchy: at stage ℓ, one token spans multiple voxels with effective spacing $\mathbf{s}^{(\ell)} = \mathbf{m}^{(\ell)} \odot \mathbf{s}^{(0)}$, where $\mathbf{m}^{(\ell)}$ is the cumulative stride. This scale-dependent metric forms the shared geometric substrate on which all three components operate.

2.2 MASE: Metric-Adaptive Spatial Embedding

Standard positional encodings, including learnable and rotary variants [13, 8], parameterize phase on index offsets and thus cannot distinguish voxels close in index space but distant physically. MASE reformulates positional phase in a spacing-normalized coordinate system so that phase progression tracks physical displacement. Following the principle that Fourier feature mappings resolve high-frequency spatial variations [14], we realize this via an axis-wise rotary form.

At stage ℓ, with in-plane reference scale $s_{\text{ref}}^{(\ell)} = \sqrt{s_x^{(\ell)} s_y^{(\ell)}}$, we define

$$\rho_a^{(\ell)} = \frac{s_a^{(\ell)}}{s_{\text{ref}}^{(\ell)}} \ (a \in \{x, y, z\}), \qquad \gamma^{(\ell)} = \sigma(\hat{\gamma}^{(\ell)}) \ (\gamma^{(\ell)} \in (0, 1)), \tag{2}$$

$$\tilde{\omega}_k^{(\ell,a)} = \omega_k \rho_a^{(\ell)}, \qquad \tilde{\omega}_{z,k}^{(\ell)} = \omega_k \left[1 + \gamma^{(\ell)} \big(\rho_z^{(\ell)} - 1\big)\right]. \tag{3}$$

where ω_k is the base angular frequency and $\sigma(\cdot)$ the sigmoid. We initialize $\hat{\gamma}^{(\ell)}$ so that $\gamma^{(\ell)} \approx 1$, recovering full metric correction. The learnable refinement is restricted to the z-axis—the dominant source of mismatch in abdominal CT—letting the network interpolate between strict physical calibration and a data-driven optimum.

For each head, channels are partitioned into axis-specific subspaces ($x/y/z$), independently transformed, and concatenated. For axis a and token coordinate p_a:

$$\left(u_{2k-1}^{(a)} + i u_{2k}^{(a)}\right) \mapsto \left(u_{2k-1}^{(a)} + i u_{2k}^{(a)}\right) \exp\!\left(i\, p_a\, \tilde{\omega}_k^{(\ell,a)}\right), \tag{4}$$

where $\tilde{\omega}_k^{(\ell,z)}$ uses Eq. (3) and $\tilde{\omega}_k^{(\ell,x)}, \tilde{\omega}_k^{(\ell,y)}$ use Eq. (2)–(3). MASE is applied to query and key streams in all self-attention blocks.

2.3 GCA: Geometry-Constrained Attention

Skip connections in U-shaped architectures assume spatially aligned encoder–decoder tokens correspond to the same physical region, an assumption violated when anisotropic spacing makes index alignment a poor proxy for physical correspondence. GCA formulates skip interaction as geometry-constrained cross-scale alignment, decomposing affinity into a content term and a metric-derived geometric penalty on physically implausible correspondences.

For decoder token at $\mathbf{p}$ and encoder token at $\mathbf{q}$ within window Ω_p:

$$e_{pq}^{(\ell)} = s_{pq}^{\text{cnt}} + B_{\text{geo}}^{(\ell)}(\mathbf{p}, \mathbf{q}), \tag{5}$$

where s_{pq}^{cnt} is content similarity and $B_{\text{geo}}^{(\ell)}$ is the geometric bias.

Let $\Delta_a = p_a - q_a$, $\bar{s}^{(\ell)} = (s_x^{(\ell)} s_y^{(\ell)} s_z^{(\ell)})^{1/3}$, and $\xi_a^{(\ell)} = \frac{s_a^{(\ell)}}{\bar{s}^{(\ell)}} \Delta_a$. The geometric bias is

$$B_{\text{geo}}^{(\ell)}(\mathbf{p}, \mathbf{q}) = -\frac{1}{\tau^{(\ell)}} \sum_{a \in \{x,y,z\}} \big(\xi_a^{(\ell)}\big)^2 + \sum_{a \in \{x,y,z\}} r_a^{(\ell)}\, \xi_a^{(\ell)}, \tag{6}$$

where the quadratic term acts as a metric-aware receptive field contracting attention along sparse axes and dilating along dense ones, and the linear term allows bounded directional preference. Intuitively, when slice spacing is large, $\xi_z^{(\ell)}$ grows rapidly with index offset, causing the quadratic penalty to suppress cross-slice attention to only the nearest physical neighbors. The fused output is

$$\alpha_{pq}^{(\ell)} = \frac{\exp\!\left(e_{pq}^{(\ell)}\right)}{\sum_{j \in \Omega_p} \exp\!\left(e_{pj}^{(\ell)}\right)}, \qquad \mathbf{y}_p = \sum_{q \in \Omega_p} \alpha_{pq}^{(\ell)}\, \mathbf{v}_q, \tag{7}$$

where $\mathbf{v}_q$ is the encoder value and $s_{pq}^{\text{cnt}} = \langle \mathbf{q}_p, \mathbf{k}_q \rangle / \sqrt{d_k}$, yielding efficient windowed cross-attention.

We constrain parameters via $\tau^{(\ell)} = \text{softplus}(\hat{\tau}^{(\ell)}) + \epsilon$ ($\epsilon = 10^{-6}$) and $r_a^{(\ell)} = r_{\max} \tanh(\hat{r}_a^{(\ell)})$, initialized with $\tau^{(\ell)} \approx 1$ and $\hat{r}_a^{(\ell)} = 0$. Parameters $\hat{\tau}^{(\ell)}$ and $\hat{r}_a^{(\ell)}$ are shared across heads within each stage.

2.4 GVV: Geometric View Voting

Training enforces metric consistency in representation space, but a residual geometric error persists at inference: deterministic registration of continuous anatomy onto a fixed lattice introduces systematic boundary bias beyond the reach of representation learning. GVV targets this complementary source through deterministic sub-voxel view aggregation along the anisotropic axis: $\mathcal{D} = \{(0, 0, \delta_z) \mid \delta_z \in \{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}\}\}$, yielding $K_z = 4$ views (fixed across experiments). Each shift displaces by $\delta_z s_z^{(0)}$ mm.

Let $f_\theta(\cdot)$ denote network logits. For $\boldsymbol{\delta} \in \mathcal{D}$, $\mathcal{T}_{\boldsymbol{\delta}}$ applies trilinear sub-voxel translation and $\mathcal{T}_{-\boldsymbol{\delta}}$ maps predictions back. We aggregate:

$$\bar{\mathbf{Z}}(\mathbf{x}) = \frac{1}{|\mathcal{D}|} \sum_{\boldsymbol{\delta} \in \mathcal{D}} \mathcal{T}_{-\boldsymbol{\delta}} \left(f_\theta \big(\mathcal{T}_{\boldsymbol{\delta}}(\mathbf{x}) \big) \right), \tag{8}$$

with $\hat{S}(\mathbf{x}) = \arg\max_c \ \bar{\mathbf{Z}}_c(\mathbf{x})$. Unlike random test-time augmentation [16], GVV is deterministic and metric-derived: shifts and aggregation axis are prescribed by acquisition geometry.

By construction, the three components are orthogonal: MASE calibrates within-scale self-attention, GCA calibrates cross-scale skip interaction, and GVV calibrates inference-time discretization—together enforcing metric consistency wherever geometric information enters the pipeline, while remaining fully compatible with standard U-shaped Transformer backbones.

3 Experiments and Results

3.1 Experimental Setup

Dataset. We evaluate MAGEFormer on two widely used multi-organ abdominal CT benchmarks: FLARE 22 [11] (50 labeled scans) and BTCV [5] (30 scans). Both datasets annotate 13 abdominal organs sharing 12 common structures; dataset-specific organs are noted in Table 1.

Unified Evaluation Protocol. All methods are evaluated under the same fixed 5-fold cross-validation with a shared fold split, using official model/parameter configurations and a unified metric pipeline.

Implementation Details. All models are implemented in PyTorch and trained on a single NVIDIA A100 GPU (80GB). We use AdamW with an initial learning rate of 4×10^{-4}, cosine annealing, and a 50-epoch warmup for 3000 epochs. Input volumes are uniformly cropped to $96 \times 96 \times 96$ with a batch size of 4.

Table 1. Per-organ Dice (%, ↑) on BTCV and FLARE 22 multi-organ CT segmentation. Spl: spleen, RKi: right kidney, LKi: left kidney, Gal: gallbladder, Eso: esophagus, Liv: liver, Sto: stomach, Aor: aorta, IVC: inferior vena cava, Pan: pancreas.

	Spl	RKi	LKi	Gal	Eso	Liv	Sto	Aor	IVC	Pan	AG[§]	V/D[¶]
	BTCV											
TransUNet[†]	87.03	78.60	81.45	57.96	66.95	92.97	75.80	86.42	71.61	57.62	47.95	60.92
nnU-Net[†]	90.91	90.17	87.83	**70.59**	78.33	**95.97**	88.30	91.67	85.84	81.47	71.01	76.46
UNETR[†]	83.70	82.41	79.51	56.59	64.89	91.70	75.06	85.01	75.35	54.43	53.14	56.86
Swin-UNETR[†]	83.23	78.03	81.81	56.35	69.90	93.07	78.22	86.97	81.41	72.58	61.36	69.05
U-Mamba[†]	90.24	**91.20**	89.33	68.63	**79.13**	95.25	87.54	**92.85**	86.36	81.95	71.64	77.05
Ours	**91.21**	90.72	**91.51**	70.46	79.09	95.87	**90.12**	92.37	**86.79**	**82.59**	**72.76**	**77.17**
	FLARE 22											
TransUNet[†]	95.39	91.42	91.42	83.02	80.25	97.29	90.42	92.80	87.01	74.36	73.72	67.87
nnU-Net[†]	98.02	**96.84**	97.01	93.45	89.93	98.93	95.91	97.28	95.16	86.91	89.57	87.54
UNETR[†]	94.93	92.52	92.12	81.19	80.39	97.64	90.34	93.72	88.65	77.76	77.29	72.25
Swin-UNETR[†]	96.01	94.34	94.61	88.75	84.71	97.68	92.23	94.23	91.13	82.53	81.14	81.40
U-Mamba[†]	98.08	95.84	96.52	**93.53**	**89.96**	**98.97**	96.14	**97.40**	95.13	86.14	89.72	88.07
Ours	**98.87**	96.80	**97.65**	93.40	89.58	98.95	**96.73**	97.33	**95.17**	**87.49**	**89.76**	**89.32**

[†] All baselines re-implemented under unified 5-fold protocol with official configurations.
[§] AG: mean Dice of right and left adrenal glands.
[¶] V/D: portal and splenic veins (BTCV) or duodenum (FLARE 22).

Table 2. Comparison of segmentation performance on FLARE 22 and BTCV datasets. Results are mean±std over five folds. Best mean values are in **bold**; green indicates the lowest standard deviation.

	FLARE 22		BTCV	
Method	**Dice**	**HD95**	**Dice**	**HD95**
TransUNet[†]	84.51±0.65	15.90±1.87	70.25±3.81	28.16±5.15
nnU-Net[†]	93.55±0.34	4.16±1.23	83.04±2.30	13.79±5.44
UNETR[†]	85.85±1.09	13.39±2.54	70.14±2.30	64.13±29.37
Swin-UNETR[†]	89.22±0.74	31.30±14.97	74.87±2.72	75.21±24.51
U-Mamba[†]	93.48±0.41	4.57±1.90	83.29±2.45	11.91±**4.16**
Ours	**93.91**±**0.20**	**3.40**±**0.48**	**84.11**±**1.73**	**10.58**±5.10

Evaluation Metrics. We report Dice Similarity Coefficient (DSC, in %) as the overlap metric, and the 95th percentile Hausdorff Distance (HD95, in mm) to assess boundary accuracy.

3.2 Results

Table 2 summarizes results against nnU-Net, TransUNet, UNETR, Swin-UNETR, and U-Mamba under the unified protocol. MAGEFormer achieves the best overall performance on both benchmarks (93.91% Dice / 3.40 mm HD95 on FLARE 22; 84.11% / 10.58 mm on BTCV), with the strongest boundary accuracy (lowest HD95), consistent with its geometry-aware design.

Table 3. Ablation study of proposed components (BTCV). **MASE**: Metric-Adaptive Spatial Embedding; **GCA**: Geometry-Constrained Attention; **GVV**: Geometric View Voting.

Method	MASE	GCA	GVV	Avg DSC (↑)	HD95 (mm, ↓)
Baseline	✗	✗	✗	0.7487	75.21
+ MASE	✓	✗	✗	0.7978 (+4.91%)	37.14 (−38.07)
+ GCA	✗	✓	✗	0.7993 (+5.06%)	34.27 (−40.94)
+ GVV	✗	✗	✓	0.7871 (+3.84%)	62.95 (−12.26)
+ MASE + GCA	✓	✓	✗	0.8295 (+8.08%)	18.76 (−56.45)
Ours (Full)	✓	✓	✓	**0.8411 (+9.24%)**	**10.58 (−64.63)**

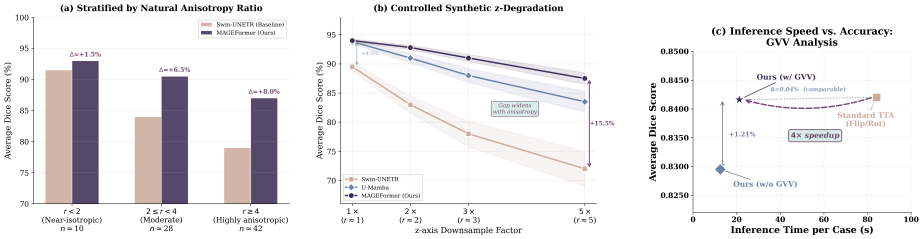

Fig. 2. Ablation and robustness analysis. (a) Performance stratified by natural anisotropy ratio, showing increasing advantage of MAGEFormer over the baseline as anisotropy grows. (b) Controlled synthetic z-degradation experiment demonstrating the robustness gap widens with increasing downsampling factor. (c) Inference speed vs. accuracy trade-off of Geometric View Voting (GVV), achieving comparable accuracy to standard TTA with 4× speedup.

Moreover, MAGEFormer exhibits the lowest cross-fold variance on both benchmarks (Dice std of 0.20 on FLARE 22 and 1.73 on BTCV), with notably stable boundary predictions (HD95 std of 0.48 on FLARE 22), suggesting that geometry-aware calibration yields representations that generalize more consistently across data splits.

The per-organ comparison in Table 1 further shows that the gains are especially evident on small or topology-sensitive organs that are more vulnerable to anisotropic sampling artifacts. On BTCV, MAGEFormer achieves the best Dice among the compared methods on the adrenal glands (AG 72.76%) and improves pancreas segmentation to 82.59%. These structures are highly sensitive to through-plane resolution and partial-volume effects, and the results support our claim that metric-consistent modeling improves fine-grained anatomical delineation.

The ablation on BTCV (Table 3, baseline: Swin-UNETR) shows that MASE and GCA each yield substantial gains individually, with marked HD95 reductions indicating improved geometric consistency. The incorporation of GVV further enhances both metrics by mitigating discretization bias. The full model achieves +9.24 Dice points (0.7487→0.8411) and reduces HD95 from 75.21 to 10.58 mm.

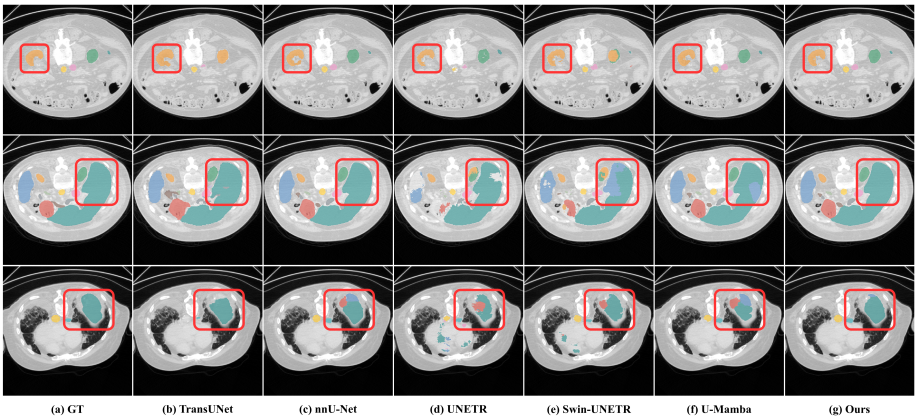

Fig. 3. Qualitative visualization of different architectures. From left to right: (a) Ground Truth, (b) TransUNet [1], (c) nnU-Net [9], (d) UNETR [7], (e) Swin-UNETR [6], (f) U-Mamba [10], (g) Ours. Compared with baselines, MAGEFormer yields more anatomically consistent and physically plausible predictions, with less multi-label fragmentation and fewer intra-organ voids that are anatomically implausible.

The robustness analysis in Fig. 2(a)–(b) further supports our central hypothesis: the advantage of MAGEFormer increases as anisotropy becomes more severe. In Fig. 2(a), pooling test cases from both benchmarks ($n = 80$) and stratifying by the natural anisotropy ratio $r = s_z^{(0)}/\sqrt{s_x^{(0)} s_y^{(0)}}$ show a widening gain over the baseline as r increases. In Fig. 2(b), under controlled synthetic z-degradation (applied at test time only by downsampling along the z-axis and interpolating back to the original grid), MAGEFormer degrades more gracefully than the compared baselines; shaded regions denote the standard deviation across folds.

Fig. 2(c) shows that GVV provides a favorable speed–accuracy trade-off: it achieves accuracy comparable to standard test-time augmentation (TTA) while requiring substantially lower inference time (approximately 4× faster per case in our setup). Visual comparisons in Fig. 3 corroborate the quantitative findings. Compared with baseline methods, MAGEFormer produces smoother and more anatomically consistent boundaries, especially for small organs and thin structures that are prone to block-like artifacts and boundary discontinuities. The improved delineation of low-contrast regions (e.g., vascular boundaries and pancreas tail) is consistent with the role of GCA in geometry-constrained feature interaction and the effect of GVV in mitigating grid-induced discretization artifacts.

4 Conclusion

We presented MAGEFormer, a geometry-calibrated framework that addresses the metric mismatch between isotropic Vision Transformers and anisotropic clinical CT by embedding physical spacing constraints directly into positional encoding,

cross-scale feature interaction, and inference-time view aggregation. Evaluations on the BTCV and FLARE 22 benchmarks show that MAGEFormer achieves the strongest boundary accuracy among the compared baselines, with consistent gains in Dice, validating that internal geometry calibration is more effective than relying on conventional isotropic preprocessing alone. Future work will extend this geometry-aware paradigm to other inherently anisotropic modalities, such as multi-parametric MRI.

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

References

1. Chen, J., Mei, J., Li, X., Lu, Y., Yu, Q., Wei, Q., Luo, X., Xie, Y., Adeli, E., Wang, Y., et al.: Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. Medical Image Analysis p. 103280 (2024)
2. Chu, X., Tian, Z., Zhang, B., Wang, X., Shen, C.: Conditional positional encodings for vision transformers. arXiv preprint arXiv:2102.10882 (2021)
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=YicbFdNTTy
4. Ge, R., Shi, F., Chen, Y., Tang, S., Zhang, H., Lou, X., Zhao, W., Coatrieux, G., Gao, D., Li, S., Mai, X.: Improving anisotropy resolution of computed tomography and annotation using 3d super-resolution network. Biomedical Signal Processing and Control **82**, 104590 (2023). https://doi.org/https://doi.org/10.1016/j.bspc.2023.104590, https://www.sciencedirect.com/science/article/pii/S174680942300023X
5. Gibson, E., Giganti, F., Hu, Y., Bonmati, E., Bandula, S., Gurusamy, K., Davidson, B., Pereira, S.P., Clarkson, M.J., Barratt, D.C.: Multi-organ abdominal ct reference standard segmentations (Feb 2018). https://doi.org/10.5281/zenodo.1169361, https://doi.org/10.5281/zenodo.1169361
6. 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: Crimi, A., Bakas, S. (eds.) Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. pp. 272–284. Springer International Publishing, Cham (2022)
7. 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 (WACV). pp. 574–584 (January 2022)
8. Heo, B., Park, S., Han, D., Yun, S.: Rotary position embedding for vision transformer. In: European Conference on Computer Vision. pp. 289–305. Springer (2024)
9. Isensee, F., Jaeger, P.F., Kohl, S.A.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 (Feb 2021). https://doi.org/10.1038/s41592-020-01008-z, https://www.nature.com/articles/s41592-020-01008-z

10. Ma, J., Li, F., Wang, B.: U-mamba: Enhancing long-range dependency for biomedical image segmentation (2024), https://arxiv.org/abs/2401.04722
11. Ma, J., Zhang, Y., Gu, S., Ge, C., Mae, S., Young, A., Zhu, C., Yang, X., Meng, K., Huang, Z., Zhang, F., Pan, Y., Huang, S., Wang, J., Sun, M., Zhang, R., Jia, D., Choi, J.W., Alves, N., de Wilde, B., Koehler, G., Lai, H., Wang, E., Wiesenfarth, M., Zhu, Q., Dong, G., He, J., He, J., Yang, H., Huang, B., Lyu, M., Ma, Y., Guo, H., Xu, W., Maier-Hein, K., Wu, Y., Wang, B.: Unleashing the strengths of unlabelled data in deep learning-assisted pan-cancer abdominal organ quantification: the flare22 challenge. The Lancet Digital Health **6**(11), e815–e826 (2024). https://doi.org/https://doi.org/10.1016/S2589-7500(24)00154-7, https://www.sciencedirect.com/science/article/pii/S2589750024001547
12. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. pp. 234–241. Springer International Publishing, Cham (2015)
13. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024). https://doi.org/https://doi.org/10.1016/j.neucom.2023.127063, https://www.sciencedirect.com/science/article/pii/S0925231223011864
14. Tancik, M., Srinivasan, P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J., Ng, R.: Fourier features let networks learn high frequency functions in low dimensional domains. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 7537–7547. Curran Associates, Inc. (2020), https://proceedings.neurips.cc/paper_files/paper/2020/file/55053683268957697aa39fba6f231c68-Paper.pdf
15. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
16. Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T.: Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. Neurocomputing **338**, 34–45 (2019). https://doi.org/https://doi.org/10.1016/j.neucom.2019.01.103, https://www.sciencedirect.com/science/article/pii/S0925231219301961
17. Wu, K., Peng, H., Chen, M., Fu, J., Chao, H.: Rethinking and improving relative position encoding for vision transformer. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10013–10021 (2021). https://doi.org/10.1109/ICCV48922.2021.00988
18. Yang, J., He, Y., Huang, X., Xu, J., Ye, X., Tao, G., Ni, B.: Alignshift: Bridging the gap of imaging thickness in 3d anisotropic volumes. In: Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, S.K., Racoceanu, D., Joskowicz, L. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. pp. 562–572. Springer International Publishing, Cham (2020)
19. Yu, P., Zhang, H., Kang, H., Tang, W., Arnold, C.W., Zhang, R.: Rplhr-ct dataset and transformer baseline for volumetric super-resolution from ct scans. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. pp. 344–353. Springer Nature Switzerland, Cham (2022)

20. 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). https://doi.org/10.1109/TIP.2023.3293771