PlaneCycle: Training-Free 2D-to-3D Lifting of Foundation Models Without Adapters

Yinghong Yu[1,2⋆][0009−0001−6382−0001], Guangyuan Li[1,2⋆][0009−0005−8063−7873], and Jiancheng Yang[1,2⋆⋆][0000−0003−4455−7145]

[1] ELLIS Institute Finland, Espoo, 02150, Finland
[2] Aalto University, Espoo, 02150, Finland
jiancheng.yang@aalto.fi

Abstract. Large-scale 2D foundation models exhibit strong transferable representations, yet extending them to 3D data typically requires retraining, adapters, or architectural redesign. We introduce *PlaneCycle*, a training-free, adapter-free operator for architecture-agnostic 2D-to-3D lifting of foundation models. *PlaneCycle* reuses the original pretrained 2D backbone by cyclically distributing spatial aggregation across orthogonal HW, DW, and DH planes throughout network depth, enabling progressive 3D fusion while preserving pretrained inductive biases. Using pretrained DINOv3 models, we evaluate *PlaneCycle* on six 3D classification and three 3D segmentation benchmarks. Without any training, the lifted models exhibit intrinsic 3D fusion capability and, under linear probing, outperform slice-wise 2D baselines and strong 3D counterparts, approaching the performance of fully trained models. With full fine-tuning, *PlaneCycle* matches standard 3D architectures, highlighting its potential as a seamless and practical 2D-to-3D lifting operator. These results demonstrate that 3D capability can be unlocked from pretrained 2D foundation models without structural modification or retraining. Code is available at https://github.com/HINTLab/PlaneCycle.

Keywords: Training-Free · Adapter-Free · 2D-to-3D Lifting · Foundation Models · DINOv3.

1 Introduction

Large-scale 2D foundation models [4,15,21] have demonstrated strong robustness and transferable representations across diverse domains, and have shown increasing promise in medical imaging [19,17], where data are limited and heterogeneous [28]. While transferring pretrained 2D models to downstream 2D tasks is straightforward via fine-tuning or low-rank adaptation [11], extending them to volumetric 3D data is considerably less natural, despite many clinical imaging modalities (e.g., CT, MRI, OCT) being inherently 3D.

⋆ Equal contribution.
⋆⋆ Correspondence to: Jiancheng Yang (jiancheng.yang@aalto.fi).

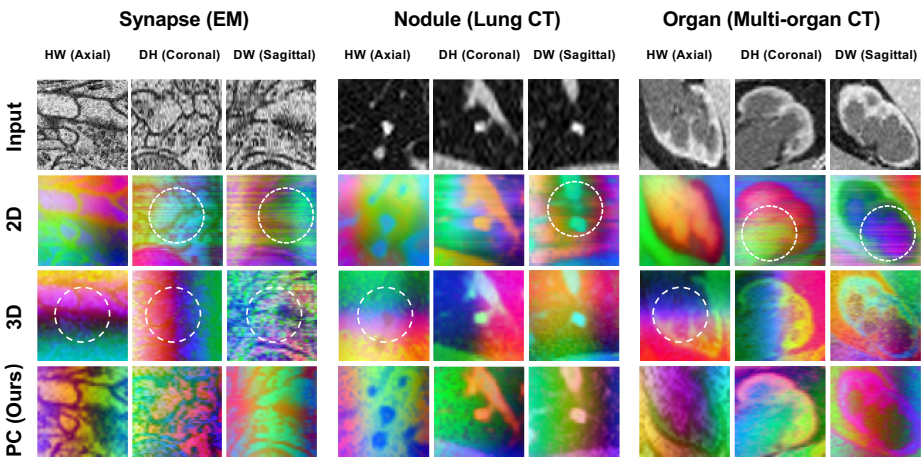

Fig. 1. PCA visualizations of frozen lifted DINOv3 [21] features on three 3D datasets [31] across HW, DW, and DH planes; inconsistencies circled.

A common strategy applies 2D models slice-by-slice and aggregates predictions post hoc [18], which is computationally efficient but neglects cross-slice dependencies. Alternatively, 2D backbones can be converted into full 3D [2] or augmented with adapters [29,26,18]. Notably, such converted models typically exhibit no intrinsic 3D capability prior to 3D retraining. Other efforts rely on large-scale video pretraining [24] or dedicated 3D medical pretraining [9], yet the diversity and accessibility of these data sources remain significantly more limited than natural 2D imagery. In addition, given the enormous computational investment behind modern 2D foundation models, DINOv3 reports 9M H100 GPU hours and 2600 tCO2eq [21], developing mechanisms that more effectively reuse pretrained 2D representations is essential for sustainability.

These observations raise a fundamental question: can 3D capability emerge from pretrained 2D foundation models *without* modifying their architecture or parameters? Ideally, such a mechanism should be training-free while remaining fully compatible with fine-tuning when supervision is available. The closest related perspective is ACS convolution [30], which reorganizes conv kernels across orthogonal planes; however, it is restricted to CNN architectures. In an era where ViT [8] foundation models increasingly dominate, there is a need for architecture-agnostic operators capable of reusing both modern ViT and established CNN backbones [12] within a unified 2D-to-3D lifting framework.

To address this question, we propose *PlaneCycle*, a parameter-free operator for 2D-to-3D lifting. *PlaneCycle* preserves the pretrained 2D backbone, and enables 3D fusion by distributing spatial aggregation across orthogonal planes throughout network depth. Specifically, feature interactions are cyclically performed over the HW, DW, and DH planes across layers, enabling progressive 3D integration in a training-free manner.

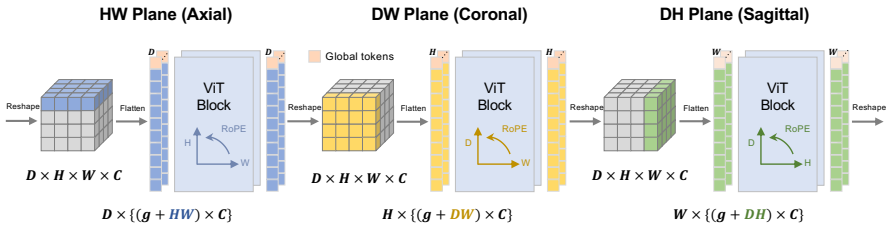

Fig. 2. Overview of *PlaneCycle* across three orthogonal planes: HW, DW, DH. Flattened slice tokens are processed by a shared ViT layer with plane-specific RoPE [22].

We use DINOv3 [21] as a representative. As illustrated in Fig. 1, slice-wise 2D inference produces features that are well-structured within the selected plane (HW) but inconsistent across slices. In contrast, naïvely converted 3D models exhibit weak and poorly aligned 3D representations across all three orthogonal planes prior to retraining. *PlaneCycle*, however, yields well-aligned 3D features across HW, DW, and DH without additional supervision. This intrinsic 3D capability is further reflected in linear probing and zero-training evaluations (Sec. 3.2), where *PlaneCycle* lifting surpasses 2D/3D by considerable margins. After full fine-tuning, performance remains competitive with full 3D models.

Importantly, *PlaneCycle* is not a replacement for adapters [29,26] or 3D pretraining [9,20,6], but a complementary operator; lifted models remain fully compatible with these techniques. We aim to demonstrate that 3D capability can be unlocked directly from pretrained 2D foundation models through a simple, seamless, and practical lifting mechanism.

2 Method

2.1 Problem Formulation

2D-to-3D lifting typically exhibits two extremes: slice-wise 2D processing [16,19] and full-volume 3D modeling [23,25]. The former extracts features independently from individual slices, whereas the latter operates directly on the full volumes. For CNNs, 2D-to-3D lifting [5,30] increases kernel dimensionality while maintaining locally linear complexity; for transformers, it scales self-attention quadratically with sequence length, substantially increasing attention cost.

Formally, given a 3D feature map $\boldsymbol{X}_i \in \mathbb{R}^{D\times H\times W\times C}$, the 2D baseline applies $\mathcal{F}_\theta$ independently to each of the D slices, yielding $\mathcal{O}(D(HW)^2)$ total attention cost. The 3D baseline flattens the volume and applies $\mathcal{F}_\theta$ once, incurring $\mathcal{O}((DHW)^2)$ attention cost, a factor of D higher.

The 2D baseline preserves computational efficiency but restricts attention to intra-slice tokens. The 3D baseline enables global volumetric interaction but incurs quadratic self-attention cost. Together, these limitations reveal a persistent trade-off between efficient 2D weight reuse, scalable volumetric interaction, and architectural simplicity, motivating a unified and efficient lifting strategy.

Algorithm 1: *PlaneCycle*

Input : 3D feature $\boldsymbol{X}_i \in \mathbb{R}^{D\times H\times W\times C}$, global tokens $\boldsymbol{G}_i \in \mathbb{R}^{P'\times g\times C}$.
Parameters : Pretrained 2D layer $\mathcal{F}_\theta$, plane $P \in \{D, H, W\}$.
Output : 3D feature $\boldsymbol{X}_o \in \mathbb{R}^{D\times H\times W\times C}$, global tokens $\boldsymbol{G}_o \in \mathbb{R}^{P\times g\times C}$.

$\boldsymbol{X} = \text{reshape}(\boldsymbol{X}_i) \in \mathbb{R}^{P\times(DHW/P)\times C}$, $\boldsymbol{G} = \text{adaptivePool}(\boldsymbol{G}_i) \in \mathbb{R}^{P\times g\times C}$;
$\boldsymbol{T} = \mathcal{F}_\theta([\boldsymbol{G}, \boldsymbol{X}]) \in \mathbb{R}^{P\times(g+DHW/P)\times C}$; // apply 2D layer P times
$\boldsymbol{X}_o = \text{reshape}(\boldsymbol{T}[:, g:]) \in \mathbb{R}^{D\times H\times W\times C}$, $\boldsymbol{G}_o = \boldsymbol{T}[:, :g] \in \mathbb{R}^{P\times g\times C}$.

2.2 *PlaneCycle*: Training-Free 2D-to-3D Lifting Operator

We propose *PlaneCycle*, an adapter-free operator for 2D-to-3D lifting. As detailed in Algorithm 1, the procedure consists of three steps: plane-wise reshaping, intra-plane aggregation, and 3D restoration. Given a 3D feature map $\boldsymbol{X}_i$ and global tokens $\boldsymbol{G}_i \in \mathbb{R}^{P'\times g\times C}$, where g denotes the number of global tokens per slice ($g = 5$ following DINOv3, consisting of one CLS token and four register tokens), we reuse the pretrained 2D weights $\mathcal{F}_\theta$ and select an axis $P \in \{D, H, W\}$, corresponding to the HW, DW, and DH planes, respectively.

Step 1: Plane-wise reshaping. The volumetric feature map is reshaped along the selected plane axis into P slices, each flattened into a token sequence of length DHW/P; the global tokens are correspondingly resampled from P' to P planes via adaptive average pooling:

$$\boldsymbol{X} = \text{reshape}(\boldsymbol{X}_i) \in \mathbb{R}^{P\times(DHW/P)\times C}, \boldsymbol{G} = \text{adaptivePool}(\boldsymbol{G}_i) \in \mathbb{R}^{P\times g\times C}. \quad (1)$$

Step 2: Intra-plane aggregation. The concatenated global and patch tokens are independently passed P times through the frozen 2D layer $\mathcal{F}_\theta$:

$$\boldsymbol{T} = \mathcal{F}_\theta\big([\boldsymbol{G}, \boldsymbol{X}]\big) \in \mathbb{R}^{P\times(g+DHW/P)\times C}. \quad (2)$$

Step 3: 3D restoration. The output tokens $\boldsymbol{T}$ are split back into patch and global tokens, with the patch tokens reshaped into the volumetric feature map:

$$\boldsymbol{X}_o = \text{reshape}(\boldsymbol{T}[:, g:]) \in \mathbb{R}^{D\times H\times W\times C}, \boldsymbol{G}_o = \boldsymbol{T}[:, :g] \in \mathbb{R}^{P\times g\times C}. \quad (3)$$

PlaneCycle preserves input–output structural consistency and can be inserted into arbitrary layers of both convolutional and transformer-based architectures without modifying pretrained parameters.

2.3 Network-Level Lifting

Cyclic cross-plane feature interaction. In principle, the aggregation plane can be scheduled arbitrarily across network depth. In practice, we adopt a four-operator cycle, HW (axial) → DW (coronal) → DH (sagittal) → HW (axial), repeating HW within each cycle to allocate more capacity to the axial plane, which typically exhibits higher resolution and stronger anatomical continuity in 3D medical data. Figure 2 illustrates this cycle (the final HW operator omitted for brevity). Since DINOv3 incorporates RoPE [22], each plane operator recomputes token coordinates for its plane, introducing no learnable parameters.

Table 1. Linear probing results (averaged over 5 runs) on six 3D classification datasets [31]. R denotes fully fine-tuned ResNet-50 [10]. Best in **bold**, and those achieved by ours in red.

Methods	Organ		Nodule		Fracture		Adrenal		Vessel		Synapse		AVG	
	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC
R-3D [31]	99.4	88.3	87.5	84.7	72.5	49.4	82.8	74.5	90.7	91.8	85.1	79.5	86.3	78.0
R-ACS [31]	99.4	88.9	88.6	84.1	75.0	51.7	82.8	75.8	91.2	85.8	71.9	70.9	84.8	76.2
CT-FM [20]	99.0	86.9	83.6	80.6	57.1	37.4	81.7	79.1	81.5	82.4	72.1	75.5	79.2	73.7
SPECTRE [6]	99.0	86.6	85.6	83.9	66.8	50.0	90.0	83.9	90.2	91.1	80.7	78.6	85.4	79.0
ViT-S/16														
2D [16,19]	98.6	84.5	89.2	**88.4**	64.1	48.2	87.0	83.9	84.7	88.7	85.7	83.8	84.9	79.6
2.5D [7]	99.2	88.0	88.3	85.2	62.9	45.4	**89.3**	**84.3**	86.4	88.6	86.7	83.6	85.5	79.2
3D [23,25]	97.4	75.0	77.7	81.5	**70.5**	50.8	77.7	77.2	73.9	88.7	74.5	75.6	78.6	74.8
PCm	**99.4**	**90.2**	88.9	86.4	62.9	45.8	85.0	83.4	**87.6**	88.2	85.3	83.0	84.8	79.5
PCg	99.2	88.2	**90.7**	87.0	69.6	**52.0**	85.9	83.3	86.2	**88.8**	**88.6**	**85.3**	**86.7**	**80.8**
ViT-B/16														
2D [16,19]	98.9	88.3	88.9	85.8	63.6	46.7	85.0	82.2	88.9	89.5	86.7	85.2	85.1	79.0
2.5D [7]	99.4	89.0	87.1	86.1	65.9	48.8	**89.9**	**84.8**	89.3	89.8	86.7	84.9	86.4	80.6
3D [23,25]	97.8	77.6	81.7	83.1	**70.5**	**54.6**	80.5	79.0	77.5	88.7	82.3	80.5	81.7	77.2
PCm	**99.8**	**94.2**	91.0	87.4	65.5	47.1	85.6	82.6	90.3	90.6	87.4	84.9	86.6	81.2
PCg	99.7	93.4	**92.7**	**88.6**	69.4	51.4	86.9	83.4	**90.5**	**90.9**	**88.9**	**85.7**	**87.8**	**82.0**
ViT-L/16														
2D [16,19]	98.2	81.5	90.1	**88.1**	63.3	44.3	86.2	82.4	81.5	88.8	86.1	81.6	84.2	77.8
2.5D [7]	99.3	88.4	89.0	**88.1**	61.4	44.2	**89.2**	**84.2**	89.7	89.4	87.3	82.6	86.0	79.5
3D [23,25]	98.6	83.3	87.5	85.6	**69.9**	47.0	77.6	77.9	82.4	88.9	79.5	79.1	82.6	77.0
PCm	**99.5**	**91.5**	**90.9**	86.8	65.6	47.4	84.0	83.3	85.2	89.2	86.7	83.4	85.3	80.2
PCg	99.2	87.7	**90.9**	85.5	69.3	**51.7**	85.2	82.6	**90.1**	**90.0**	**88.9**	**86.5**	**87.3**	**80.7**

Handling global tokens. Switching across orthogonal planes changes the number of slices, creating a token-length mismatch in the global tokens. We address this with two variants. The global mean variant (PCm) averages global tokens across slices and replicates them for the next plane, but this may distort layer-wise token statistics with a frozen backbone. The grouping variant (PCg) instead applies adaptive average pooling, computing group-wise means for downsampling and interleave-repeating tokens for upsampling, better preserving consistency.

Complexity analysis. Following the attention derivation in Sec. 2.1, *PlaneCycle*'s per-layer token sequence length depends on the active plane. Assuming a cubic volume with $D = H = W$, each layer's self-attention complexity matches that of the 2D baseline, yielding a D-fold reduction over full 3D attention.

3 Experiments

3.1 Setup

Datasets. For classification, we use six 3D datasets from MedMNIST+ [31]: Organ [3,27], Nodule [1], Fracture [14], Adrenal, Vessel [32], and Synapse, covering CT, MRI, and EM across whole-body, lung, abdomen, chest, brain, and cellular structures. We follow the official preprocessing and split protocol.

Table 2. Full fine-tuning (averaged over 5 runs) on six 3D classification datasets [31]. Best in **bold**, and those achieved by ours in red.

Methods	Organ		Nodule		Fracture		Adrenal		Vessel		Synapse		AVG	
	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC
ViViT [13]	99.9	96.5	88.3	87.2	69.9	51.0	86.3	82.9	95.1	94.0	93.5	90.5	88.8	83.7
CT-FM [20]	99.9	95.6	90.2	84.8	66.5	45.3	91.3	80.3	95.9	94.7	92.4	87.2	89.4	81.3
SPECTRE [6]	100.0	98.0	92.1	87.2	73.7	54.5	90.0	82.2	95.6	94.7	92.7	86.8	90.7	83.9
ViT-S/16														
2D [16,19]	98.6	85.8	89.1	85.5	62.2	44.1	81.7	80.8	84.2	90.1	93.0	88.2	84.8	79.1
2.5D [7]	99.5	91.3	88.8	86.8	59.9	43.3	86.3	82.9	85.4	88.9	92.0	88.2	85.3	80.2
3D [23,25]	99.8	95.7	**93.8**	**87.8**	**76.0**	55.0	**89.7**	**85.6**	**94.9**	91.9	**96.5**	**92.7**	**91.8**	84.8
PCm	**99.9**	**96.2**	93.4	87.4	73.8	**56.4**	89.2	85.4	**94.9**	93.4	95.1	91.8	91.0	**85.1**
PCg	**99.9**	**96.2**	93.2	**87.8**	72.2	51.9	88.5	84.6	94.1	**93.6**	95.7	92.2	90.6	84.4
ViT-B/16														
2D [16,19]	99.0	87.8	90.8	86.8	61.1	44.9	80.6	80.5	82.0	89.4	94.1	88.5	84.6	79.7
2.5D [7]	99.5	92.5	90.0	86.7	62.4	46.2	82.8	82.0	86.5	90.1	92.6	87.3	85.6	80.8
3D [23,25]	99.8	96.2	92.7	87.7	**74.2**	52.6	89.0	**85.1**	**94.9**	92.0	96.1	91.7	91.1	84.2
PCm	**100.0**	**97.5**	91.8	88.1	72.8	52.7	86.6	81.5	94.6	91.9	96.0	92.7	90.3	84.1
PCg	99.9	97.0	**94.6**	**88.5**	73.1	**53.1**	**89.1**	84.8	93.8	**92.7**	**96.9**	**93.3**	**91.2**	**84.9**
ViT-L/16														
2D [16,19]	98.6	84.6	90.2	86.8	64.6	43.8	81.1	80.1	85.3	89.7	90.5	86.8	85.0	78.6
2.5D [7]	99.6	92.4	90.3	86.3	63.1	46.9	81.9	80.7	87.2	90.9	92.1	86.7	85.7	80.7
3D [23,25]	99.8	97.2	93.6	**88.3**	**76.6**	**58.4**	**88.9**	**84.8**	**96.7**	**94.4**	**97.3**	**93.0**	**92.2**	**86.0**
PCm	**99.9**	97.0	93.2	86.9	73.9	57.1	87.5	83.0	95.1	93.6	96.6	91.9	91.0	84.9
PCg	**99.9**	**97.5**	**94.0**	87.0	73.8	53.7	87.7	83.8	95.8	92.6	97.1	90.1	91.4	84.1

For segmentation, we evaluate on LIDC [1] for CT lung nodule segmentation (2,142 train / 526 test) and MMWHS [33] for multi-structure cardiac segmentation (16 train / 4 test) from CT and MRI, covering 5 cardiac substructures.

Baselines. We compare *PlaneCycle* against the slice-wise 2D and full-volume 3D baselines defined in Sec. 2.1. For 2.5D, we use tri-slice [7], which stacks each slice with its two neighboring slices as three input channels. We also include the 3D medical foundation models CT-FM [20] and SPECTRE [6], pretrained on large-scale CT corpora via self-supervised learning. On classification, we compare against all of these over 5 runs, and additionally cite the fine-tuned results reported on MedMNIST+ for the ResNet-based *R-ACS* and *R-3D* [31], and the ViT-based *ViViT* [13]. On segmentation, we compare against the 2D and 3D baselines only, leaving broader comparisons to future work.

Implementation. All experiments run on a single NVIDIA H200 GPU (141GB). For 2D, 2.5D, 3D, and *PlaneCycle*, we use official pretrained ViT-S/16, ViT-B/16, and ViT-L/16 backbones [21], sharing the identical DINOv3 pretrained weights and tokenizing volumes with the original DINOv3 tokenizer per slice; the 3D baseline additionally adopts 3D RoPE. The 3D medical foundation models use their officially released pretrained weights. In linear probing, the backbone is frozen and only a classification head or 3D segmentation decoder is trained; in full fine-tuning, all parameters are optimized.

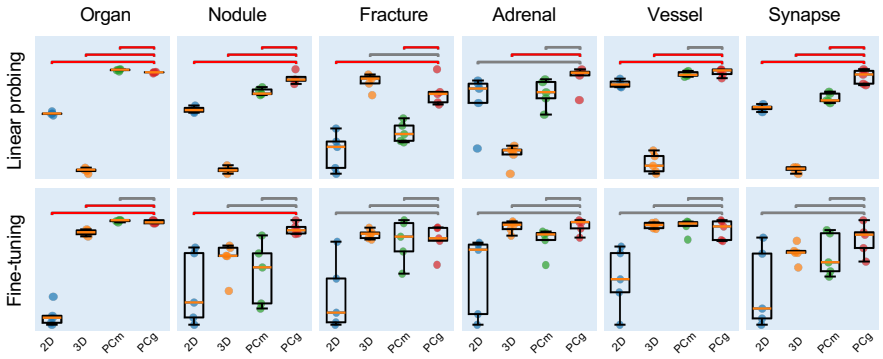

Fig. 3. Paired t-tests of AUC on six 3D classification datasets [31] on ViT-B/16, computed over five runs. Red indicates significance ($p < 0.05$).

Metrics. We report AUC and ACC for classification, and Dice for segmentation. To assess volumetric feature coherence without additional optimization, we define *FeatDice*, as no established metric directly evaluates this. For each volume, the feature at the central lesion voxel serves as a reference to compute similarities to all voxel-wise features, yielding a dense similarity map. The Dice coefficient is calculated between the map and the ground-truth lesion mask.

3.2 Key Insights

Proper lifting unlocks the 3D capability of pretrained 2D foundation models. Under linear probing alone, *PlaneCycle* outperforms 2D, 2.5D, and 3D baselines, even surpassing the fully fine-tuned *R-ACS* and *R-3D* [31] (Tab. 1). After fine-tuning (Tab. 2), it further exceeds *ViViT* [13], which lacks comparable pretraining, while slice-wise 2D still falls short, failing to transfer pretrained knowledge into coherent 3D modeling. These results demonstrate that *PlaneCycle* unlocks 3D capability from 2D foundation models pretrained solely on natural images.

Without training, PlaneCycle yields discriminative 3D representations. We analyze zero-training results (Fig. 1, Tab. 3): *PlaneCycle* produces more coherent feature maps and achieves higher *FeatDice* scores than both 2D and 3D, indicating stronger 3D representations. Under linear probing (Tab. 1, Tab. 3), *PlaneCycle* achieves the best performance in nearly all settings, with paired t-tests (Fig. 3) confirming statistical significance ($p < 0.05$) on 5/6 datasets.

With fine-tuning, PlaneCycle matches full 3D without compromise. Under full fine-tuning (Tab. 2, Tab. 3), *PlaneCycle* consistently outperforms slice-wise 2D and tri-slice 2.5D, matches 3D flattening on classification, and even surpasses it on segmentation. However, 3D flattening comes at substantially higher cost: with ViT-L/16 on LIDC, it requires 36.2 h (3D) vs. 16.3 h (*PlaneCycle*) ($> 2\times$). Moreover, 3D flattening performs poorly under linear probing, indicating it relies

Table 3. Segmentation performance on LIDC [1] and MMWHS [33] under different settings. Best in **bold**, and those achieved by ours in red.

Methods	ViT-S/16				ViT-B/16				ViT-L/16			
	LIDC	MMWHS CT	MMWHS MRI	AVG	LIDC	MMWHS CT	MMWHS MRI	AVG	LIDC	MMWHS CT	MMWHS MRI	AVG
Zero-training Features (FeatDice)												
2D [16,19]	36.9	**31.3**	26.5	31.6	28.3	30.4	26.1	28.3	22.4	29.1	25.1	25.5
3D [23,25]	13.9	22.4	24.4	20.2	22.3	26.2	28.6	25.7	21.5	26.2	27.4	25.0
PCm	42.3	**31.3**	**30.8**	**34.8**	30.8	**33.5**	**29.5**	31.3	27.9	**33.7**	**29.2**	**30.3**
PCg	**42.6**	30.9	29.9	34.5	**32.2**	33.2	29.3	**31.6**	**28.0**	33.5	29.0	30.2
Linear Probing												
2D [16,19]	60.1	67.7	58.4	62.1	62.2	68.2	57.6	62.7	62.6	68.8	59.5	63.6
3D [23,25]	59.4	65.9	60.9	62.1	62.6	66.8	59.2	62.9	61.5	65.6	57.5	61.5
PCm	**64.9**	**69.3**	**61.5**	**65.2**	**65.8**	69.4	**61.4**	**65.5**	**66.1**	**69.2**	59.8	65.0
PCg	64.4	68.7	61.4	64.8	64.8	**69.5**	61.3	65.2	65.7	**69.2**	**61.8**	**65.6**
Full Fine-tuning												
2D [16,19]	71.2	68.4	60.7	66.8	71.3	69.8	61.2	67.4	71.1	70.2	61.2	67.5
3D [23,25]	72.1	71.5	66.2	69.9	73.4	71.0	**66.2**	70.2	74.0	72.6	**65.5**	70.7
PCm	72.6	**74.2**	67.8	71.5	**73.5**	73.8	64.9	70.7	73.8	74.3	64.0	70.7
PCg	**72.8**	74.0	**70.8**	**72.5**	**73.5**	**76.5**	64.6	**71.5**	**74.4**	**76.2**	63.8	**71.5**

heavily on fine-tuning to realize its potential. *PlaneCycle*, by contrast, performs well under both settings while retaining 2D computational efficiency.

PlaneCycle also outperforms 3D medical foundation models CT-FM [20] and SPECTRE [6] on classification under both linear probing and fine-tuning. Importantly, *PlaneCycle* is not a replacement but a complementary operator: lifted models remain fully compatible with these 3D pretraining techniques.

As for variant selection, *PCg* is recommended for linear-probing classification due to better CLS-to-patch token consistency, while *PCm* is preferred otherwise for its simplicity and comparable performance.

4 Conclusion and Future Work

We introduced *PlaneCycle*, a parameter-free operator for architecture-agnostic 2D-to-3D lifting that unlocks 3D capability in pretrained 2D foundation models without architectural modification or retraining, while remaining fully compatible with 3D fine-tuning whenever supervision is available.

Several limitations warrant further investigation. Although *PlaneCycle* is inherently compatible with both CNN and ViT backbones, we have not yet conducted systematic comparisons. The *PlaneCycle*-lifted models are also applicable to 3D pretraining [9] and adapters such as LoRA [11,18], which remain to be explored. Besides, while *PlaneCycle* inherits 2D computational complexity and suggests favorable scaling behavior over full 3D, its large-scale potential remains unvalidated; preliminary results show that lifting DINOv3-7B is feasible, a scale not previously achieved in 3D settings. Moreover, our segmentation study intentionally isolates the lifting effect using a shared lightweight decoder, whose

simplicity and the 16× downsampling in ViT may limit spatial recovery. Separately, cyclic design enables 3D fusion but gives each plane fewer opportunities for within-plane attention; this trade-off remains unexplored. We leave these directions, along with broader architectures, datasets, task-specific decoders, and extensions to multimodal or vision–language settings, for future work.

Acknowledgments. This study was supported by ELLIS Institute Finland and Aalto University. We acknowledge CSC-IT Center for Science, Finland, for providing access to the supercomputers Mahti and Puhti, as well as LUMI, owned by the European High Performance Computing Joint Undertaking (EuroHPC JU) and hosted by CSC Finland in collaboration with the LUMI consortium. We also acknowledge the computational resources provided by the Aalto Science-IT project through the Triton cluster.

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

References

1. Armato III, S.G., McLennan, G., Bidaut, L., others.: The lung image database consortium (lidc) and image database resource initiative (idri): A completed reference database of lung nodules on ct scans. Medical Physics **38**(2), 915–931 (2011)
2. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: A video vision transformer. In: International Conference on Computer Vision. pp. 6836–6846 (2021)
3. Bilic, P., Christ, P.F., et al.: The liver tumor segmentation benchmark (lits). arXiv Preprint (2019)
4. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: International Conference on Computer Vision. pp. 9650–9660 (2021)
5. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
6. Claessens, C., Viviers, C., D'Amicantonio, G., Bondarev, E., van der Sommen, F.: Scaling self-supervised and cross-modal pretraining for volumetric ct transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13636–13647 (2026)
7. Ding, J., Li, A., Hu, Z., Wang, L.: Accurate pulmonary nodule detection in computed tomography images using deep convolutional neural networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 559–567. Springer (2017)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
9. Hamamci, I.E., Er, S., Wang, C., Almas, F., Simsek, A.G., Esirgun, S.N., Dogan, I., Durugol, O.F., Hou, B., Shit, S., et al.: Generalist foundation models from a multimodal dataset for 3d computed tomography. Nature Biomedical Engineering pp. 1–19 (2026)

10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
11. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. International Conference on Learning Representations **1**(2), 3 (2022)
12. Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K., Jaeger, P.F.: nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. In: Conference on Medical Image Computing and Computer Assisted Intervention. pp. 488–498. Springer (2024)
13. Jain, S., Li, X., Xu, M.: Knowledge transfer from macro-world to micro-world: enhancing 3d cryo-et classification through fine-tuning video-based deep models. Bioinformatics **40**(7), btae368 (2024)
14. Jin, L., Yang, J., Kuang, K., Ni, B., Gao, Y., Sun, Y., Gao, P., Ma, W., Tan, M., Kang, H., Chen, J., Li, M.: Deep-learning-assisted detection and segmentation of rib fractures from ct scans: Development and validation of fracnet. EBioMedicine **62**, 103106 (2020)
15. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: International Conference on Computer Vision. pp. 4015–4026 (2023)
16. Li, Y., Wu, Y., Lai, Y., Hu, M., Yang, X.: Meddinov3: How to adapt vision foundation models for medical image segmentation? arXiv Preprint (2025)
17. Liu, C., Chen, Y., Shi, H., Lu, J., Jian, B., Pan, J., Cai, L., Wang, J., Zhang, Y., Li, J., et al.: Does dinov3 set a new medical vision standard? arXiv Preprint (2025)
18. Liu, H., Georgescu, B., Zhang, Y., Yoo, Y., Baumgartner, M., Gao, R., Wang, J., Zhao, G., Gibson, E., Comaniciu, D., et al.: Revisiting 2d foundation models for scalable 3d medical image classification. arXiv Preprint (2025)
19. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature communications **15**(1), 654 (2024)
20. Pai, S., Hadzic, I., Bontempi, D., Bressem, K., Kann, B.H., Fedorov, A., Mak, R.H., Aerts, H.J.: Vision foundation models for computed tomography. arXiv preprint arXiv:2501.09001 (2025)
21. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv Preprint (2025)
22. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024)
23. Wald, T., Roy, S., Isensee, F., Ulrich, C., Ziegler, S., Trofimova, D., Stock, R., Baumgartner, M., Köhler, G., Maier-Hein, K.: Primus: enforcing attention usage for 3d medical image segmentation (2025)
24. Wang, Y., Li, K., Li, X., Yu, J., He, Y., Chen, G., Pei, B., Zheng, R., Wang, Z., Shi, Y., et al.: Internvideo2: Scaling foundation models for multimodal video understanding. In: European Conference on Computer Vision. pp. 396–416. Springer (2024)
25. Wei, X., Liu, X., Zang, Y., Dong, X., Zhang, P., Cao, Y., Tong, J., Duan, H., Guo, Q., Wang, J., et al.: Videorope: What makes for good video rotary position embedding? arXiv Preprint (2025)
26. Wu, J., Wang, Z., Hong, M., Ji, W., Fu, H., Xu, Y., Xu, M., Jin, Y.: Medical sam adapter: Adapting segment anything model for medical image segmentation. Medical Image Analysis **102**, 103547 (2025)
27. Xu, X., Zhou, F., et al.: Efficient multiple organ localization in ct image using 3d region proposal network. IEEE Transactions on Medical Imaging **38**(8), 1885–1898 (2019)

28. Yang, J.: Multi-task learning for medical foundation models. Nature Computational Science **4**(7), 473–474 (2024)
29. Yang, J., He, Y., Kuang, K., Lin, Z., Pfister, H., Ni, B.: Asymmetric 3d context fusion for universal lesion detection. In: Conference on Medical Image Computing and Computer Assisted Intervention. pp. 571–580. Springer (2021)
30. Yang, J., Huang, X., He, Y., Xu, J., Yang, C., Xu, G., Ni, B.: Reinventing 2d convolutions for 3d images. IEEE Journal of Biomedical and Health Informatics **25**(8), 3009–3018 (2021)
31. Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., Ni, B.: Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. Scientific Data **10**(1), 41 (2023)
32. Yang, X., Xia, D., Kin, T., Igarashi, T.: Intra: 3d intracranial aneurysm dataset for deep learning. In: Conference on Computer Vision and Pattern Recognition (June 2020)
33. Zhuang, X., Li, L., Payer, C., Štern, D., Urschler, M., Heinrich, M.P., Oster, J., Wang, C., Smedby, Ö., Bian, C., et al.: Evaluation of algorithms for multi-modality whole heart segmentation: an open-access grand challenge. Medical Image Analysis **58**, 101537 (2019)