Volume Transformer: Revisiting Vanilla Transformers for 3D Scene Understanding

SceneFUN3D Functionality Segmentation Challenge Winner
ScanNet++ Semantic and Instance Segmentation Challenge Winner

Kadir Yilmaz*

Adrian Kruse*

Tristan Höfer

Daan De Geus

Bastian Leibe

* Equal Contribution

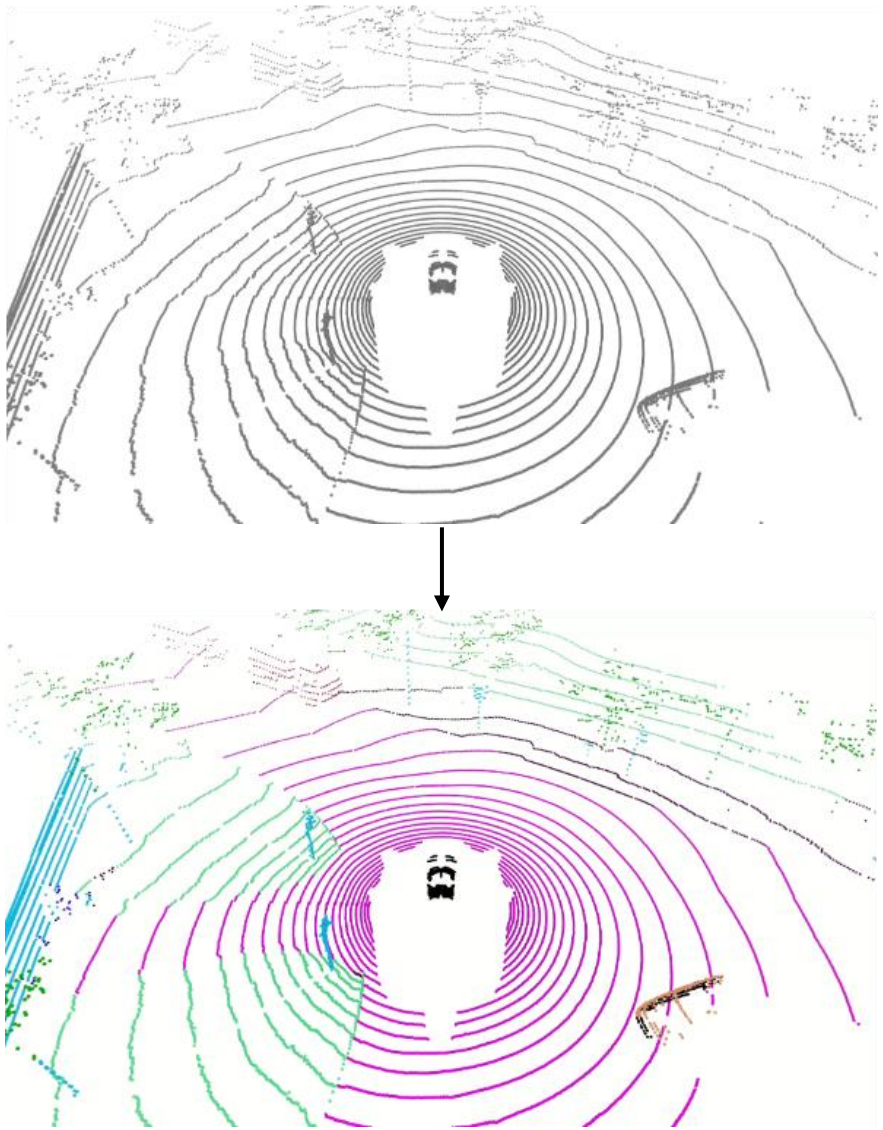
Visual Computing Institute | RWTH AACHEN UNIVERSITY

CVPR JUNE 3-7, 2026 DENVER COLORADO

Task: 3D Scene Understanding

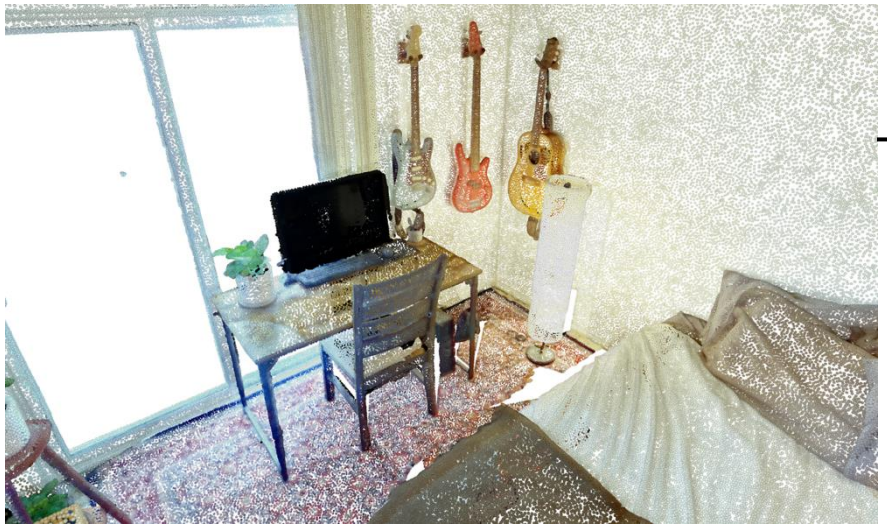
RGB-D Point Clouds (ScanNet++)

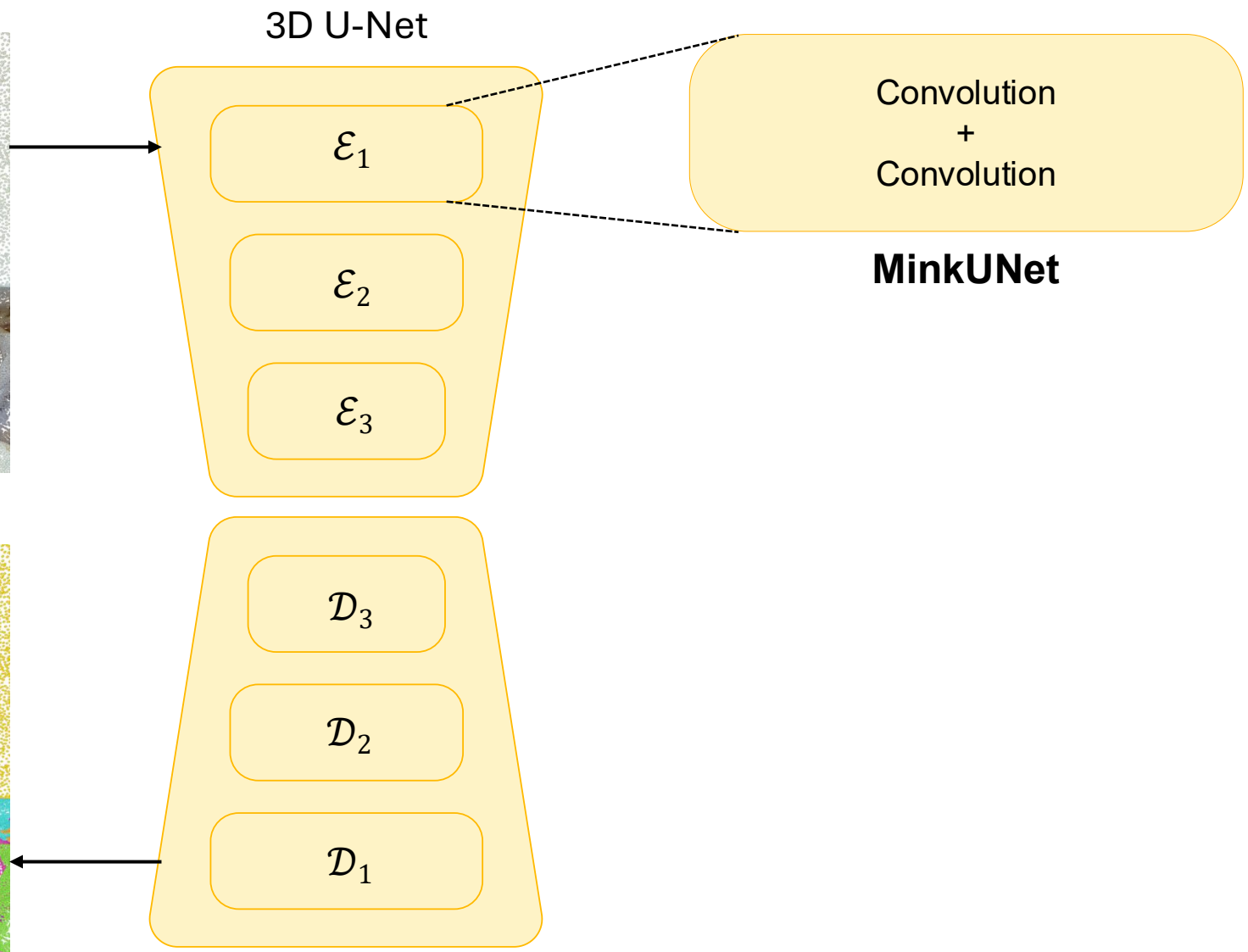
LiDAR Point Clouds (nuScenes)



Current Paradigm: 3D U-Nets

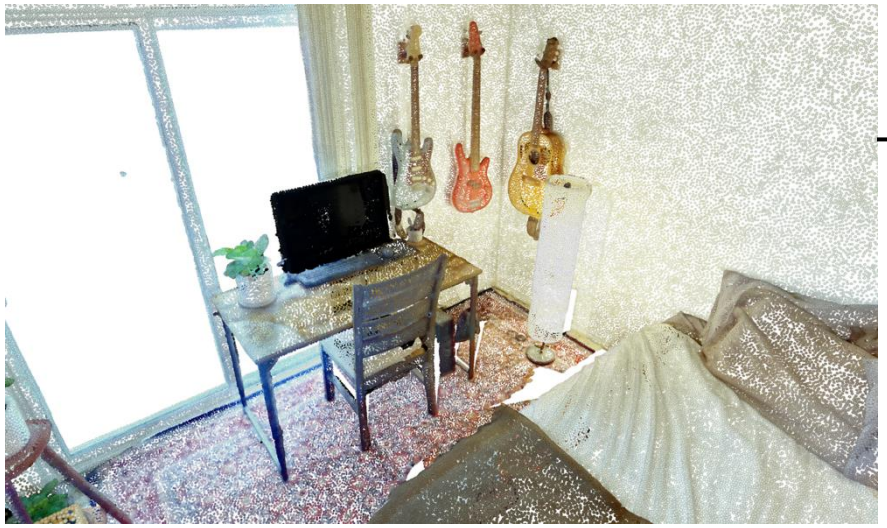
RGB-D Point Clouds (ScanNet++)

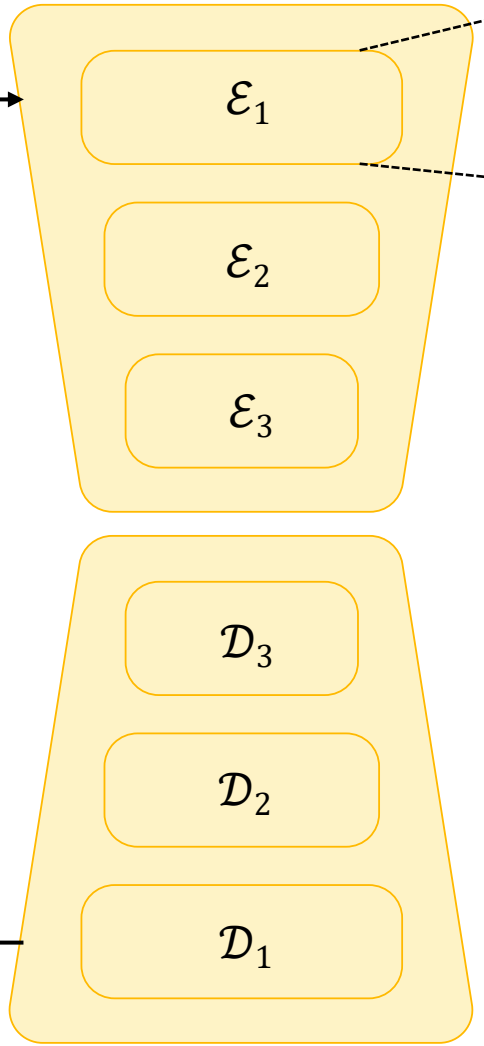




Current Paradigm: 3D U-Nets

RGB-D Point Clouds (ScanNet++)



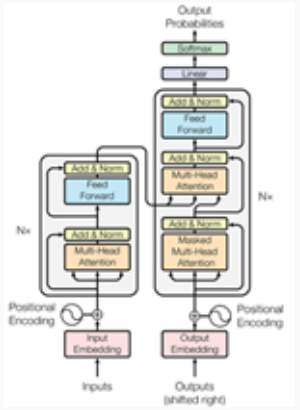


Convolution
+
Local Attention

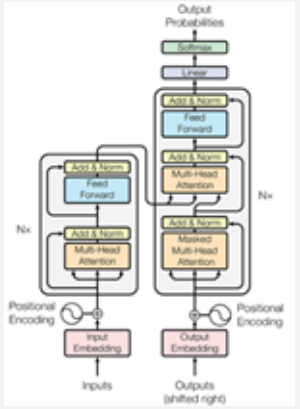
Point“Transformer”v3

Everyone Else: Transformers

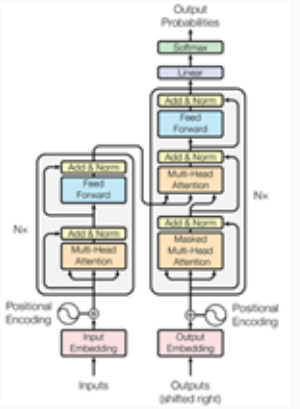
Computer Vision



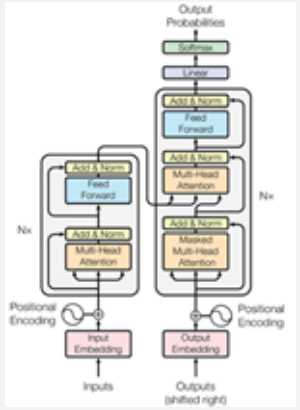
Natural Lang. Proc.



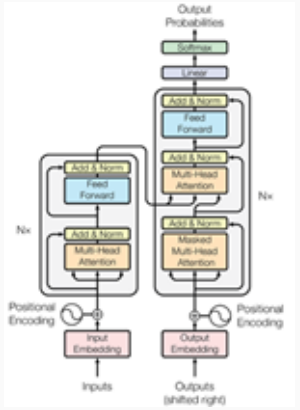
Reinf. Learning



Speech



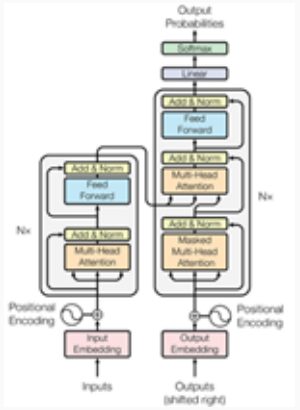
Translation



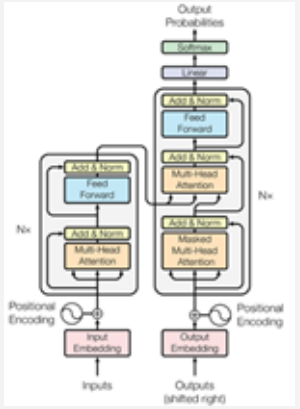
Slide is taken by “Transformer Tutorial” of Lucas Beyer
Transformer image is taken by "Attention Is All You Need" paper

Everyone Else: Transformers

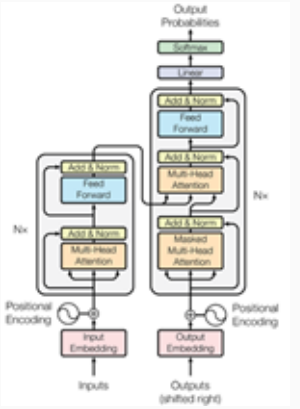
Computer Vision



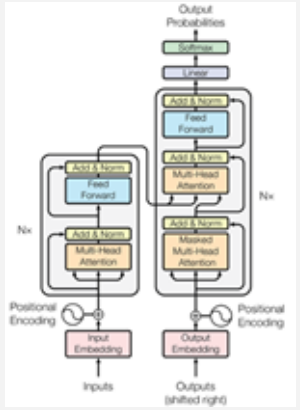
Natural Lang. Proc.



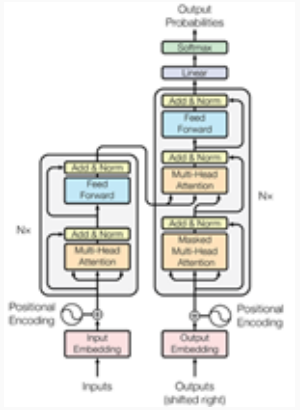
Reinf. Learning



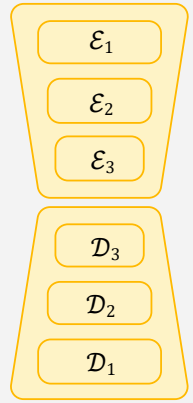
Speech



Translation



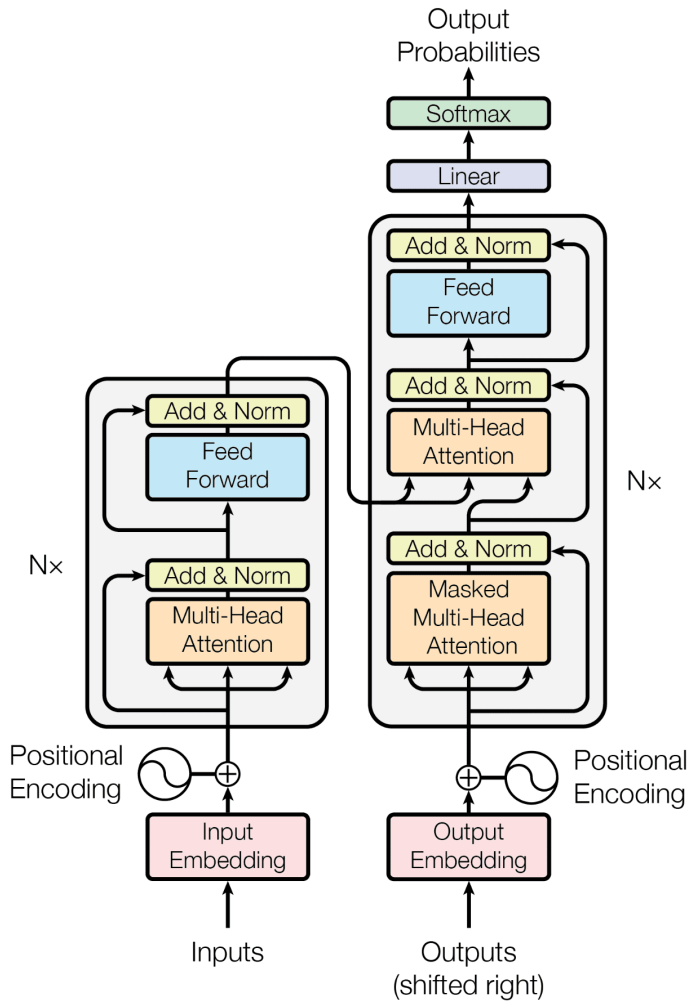
3D vision



Slide is taken by “Transformer Tutorial” of Lucas Beyer
Transformer image is taken by "Attention Is All You Need" paper

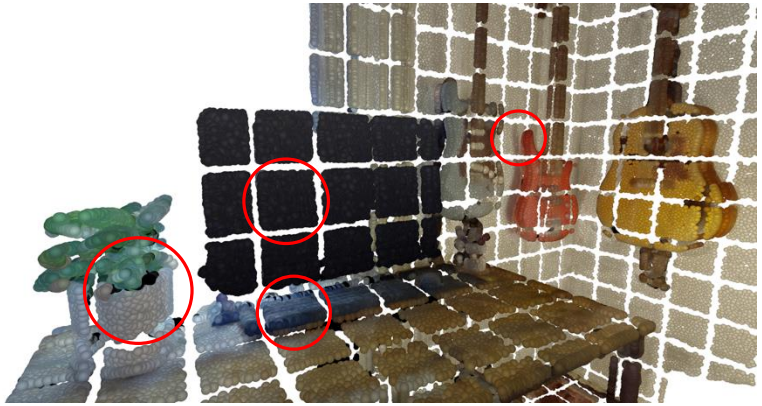
Downsides of 3D-Specific Architectures

- **Inductive biases can be limiting**
  - Convolutions and hand-designed local attention layers are useful at small scale
  - But they can limit the model's ability to learn complex patterns or long-range dependencies
- **Isolates 3D research**
  - Most research is on Transformers (Muon Optimizer, SSL techniques: MAE, DINO, MoE, QKNorm)
  - Hardware/software are focused on the Transformer workloads (H100, GB200, FlashAttention4)

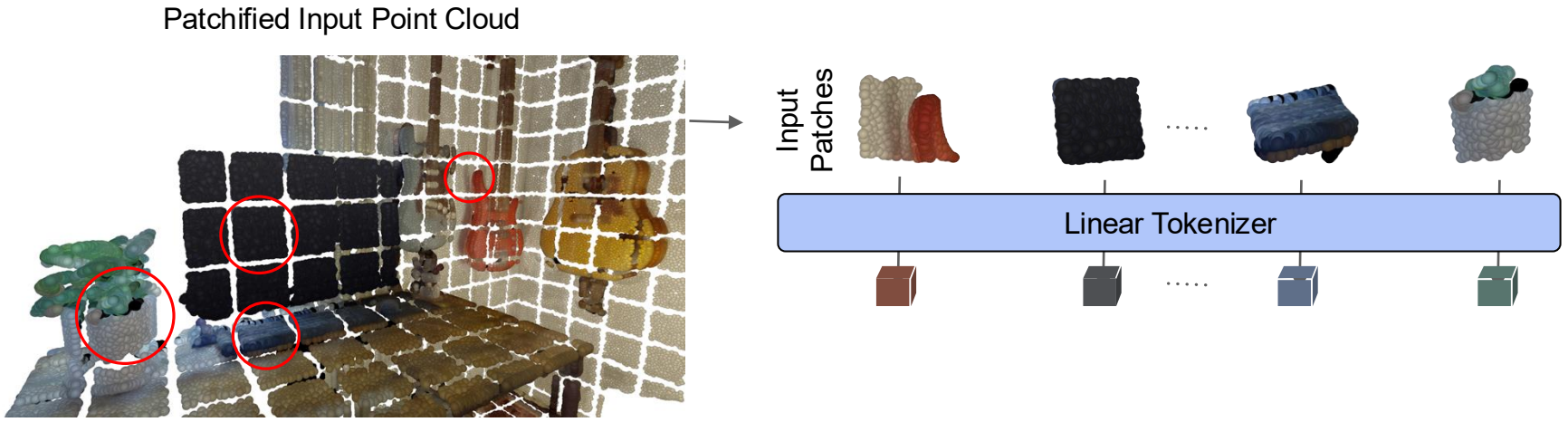


Bridging the Gap: Volume Transformer (Volt)

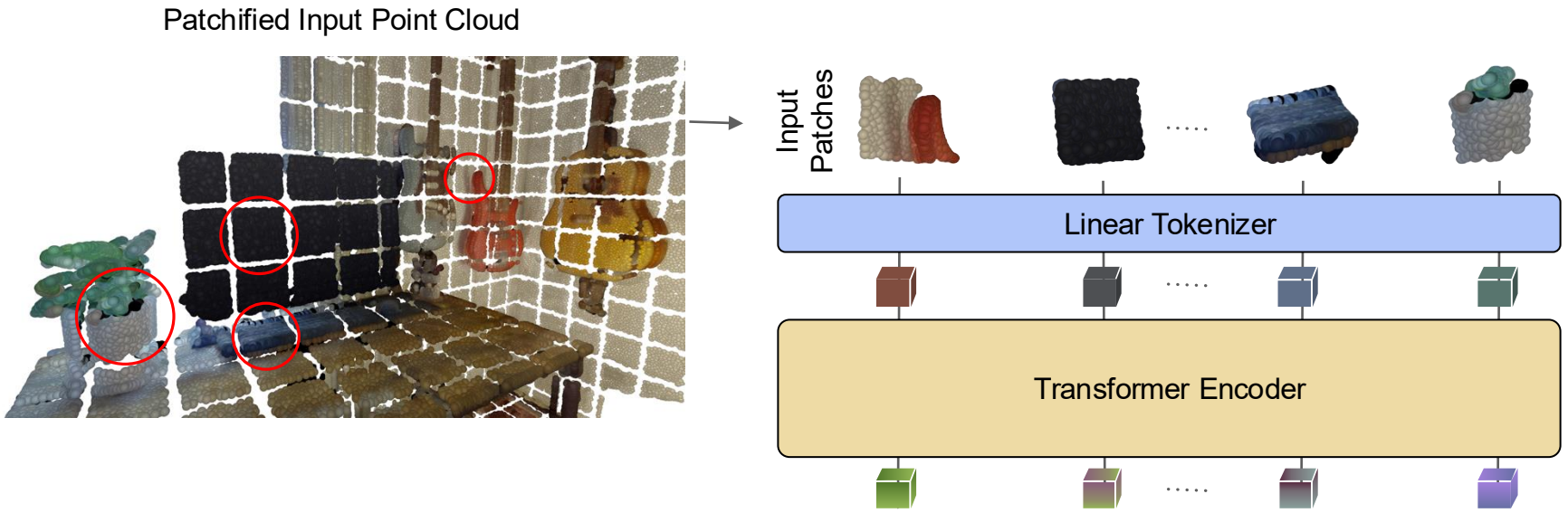
Patchified Input Point Cloud



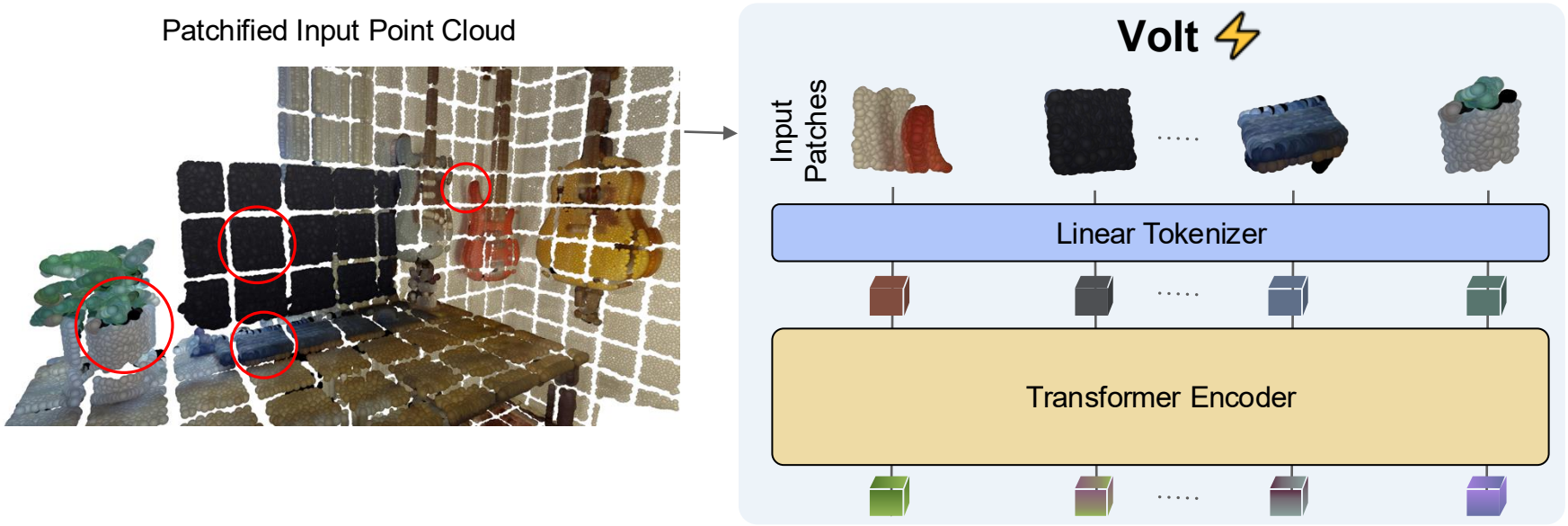
Bridging the Gap: Volume Transformer (Volt)



Bridging the Gap: Volume Transformer (Volt)

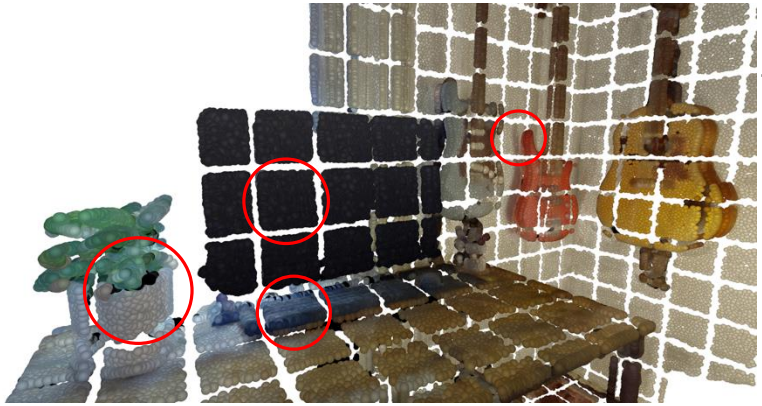


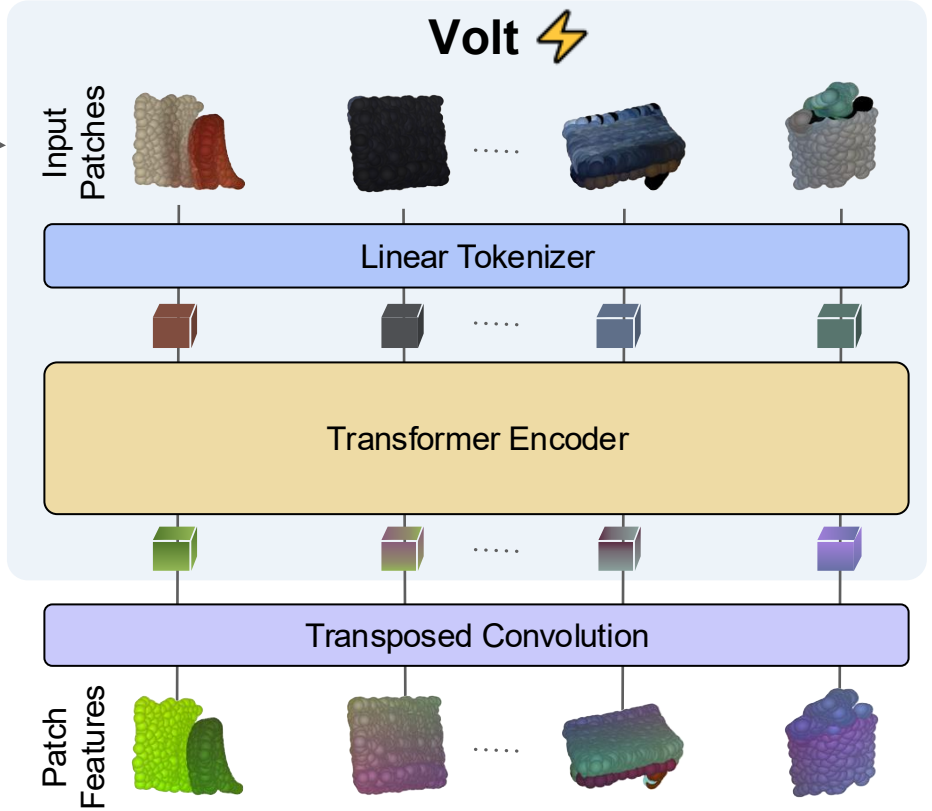
Bridging the Gap: Volume Transformer (Volt)



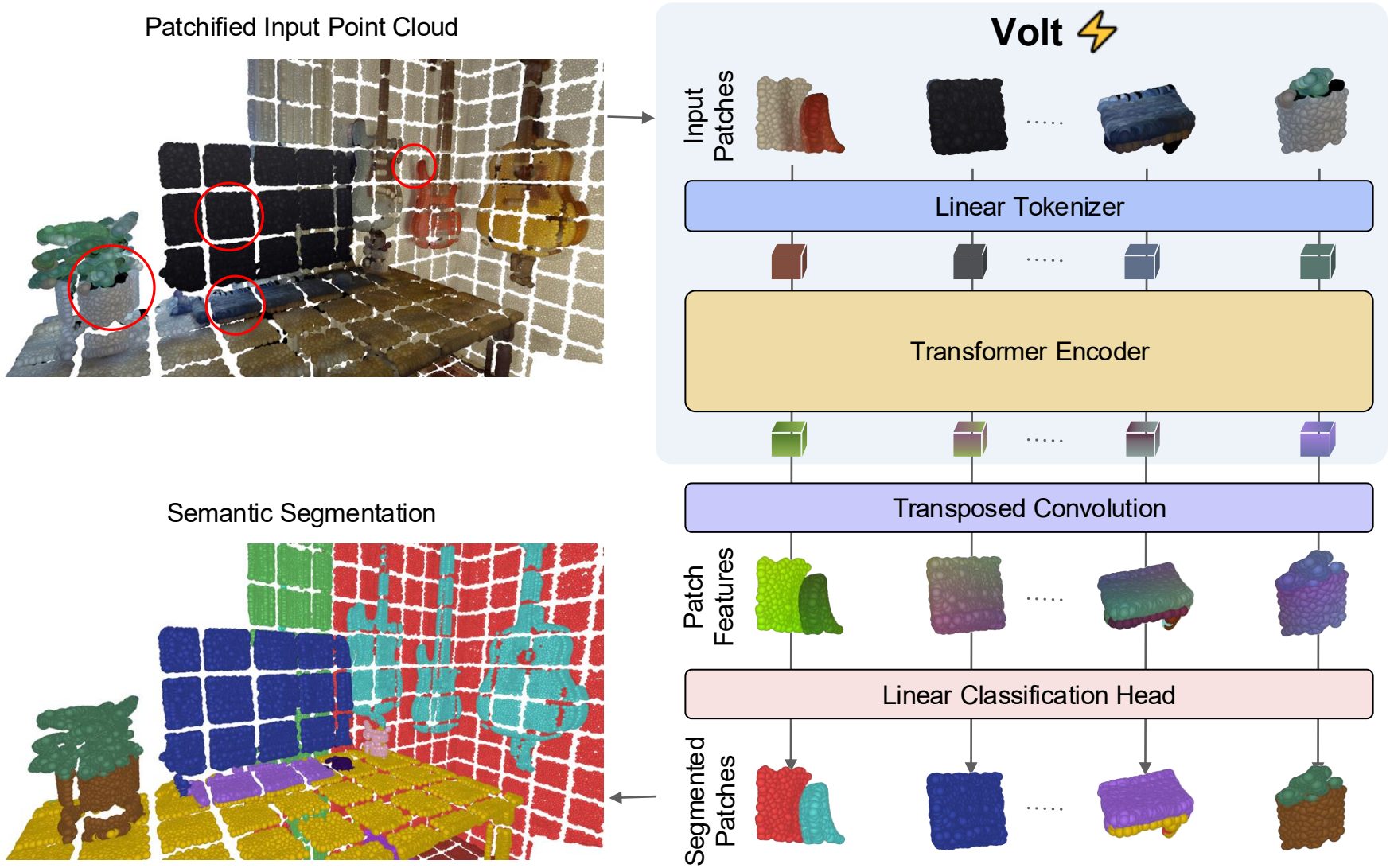
Bridging the Gap: Volume Transformer (Volt)

Patchified Input Point Cloud

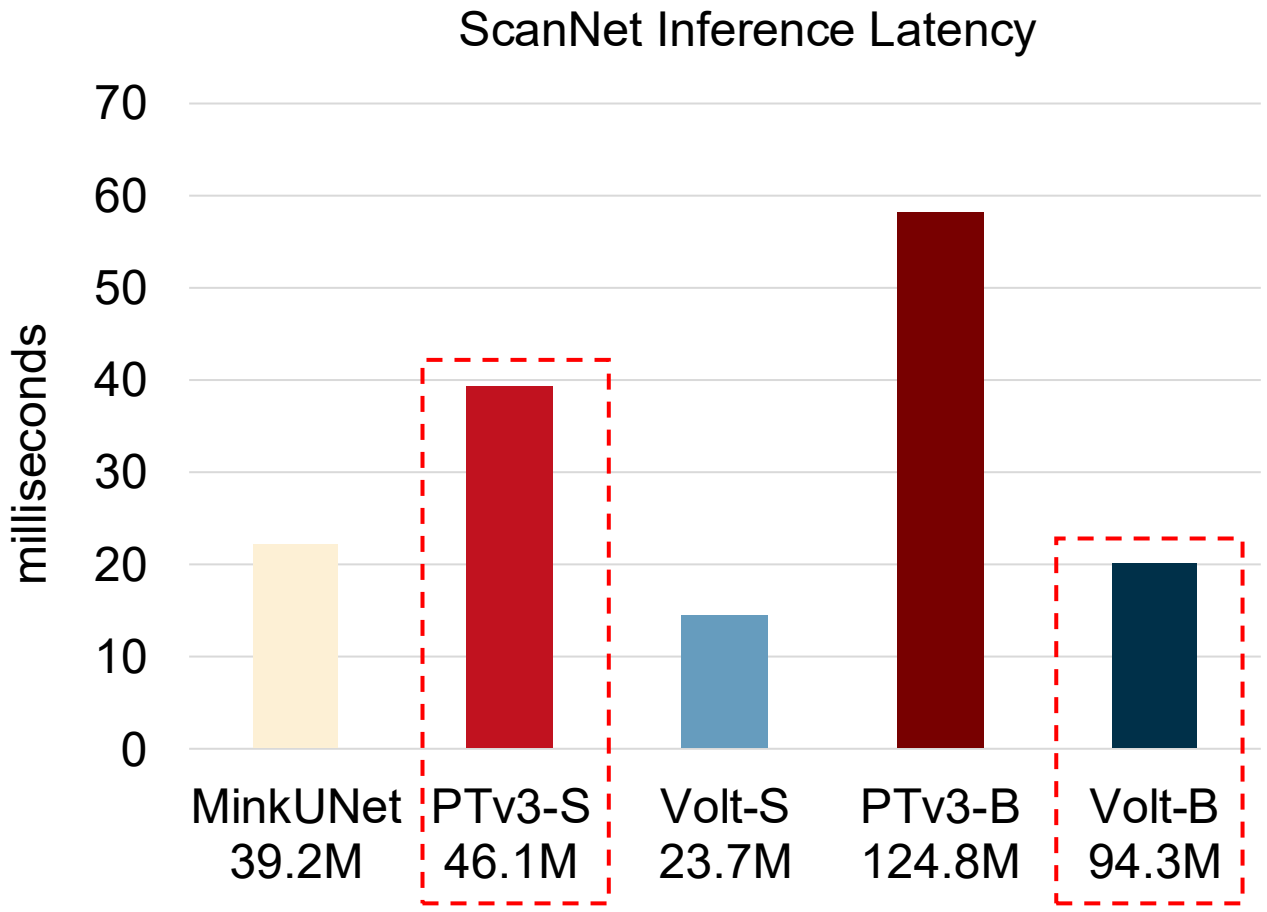


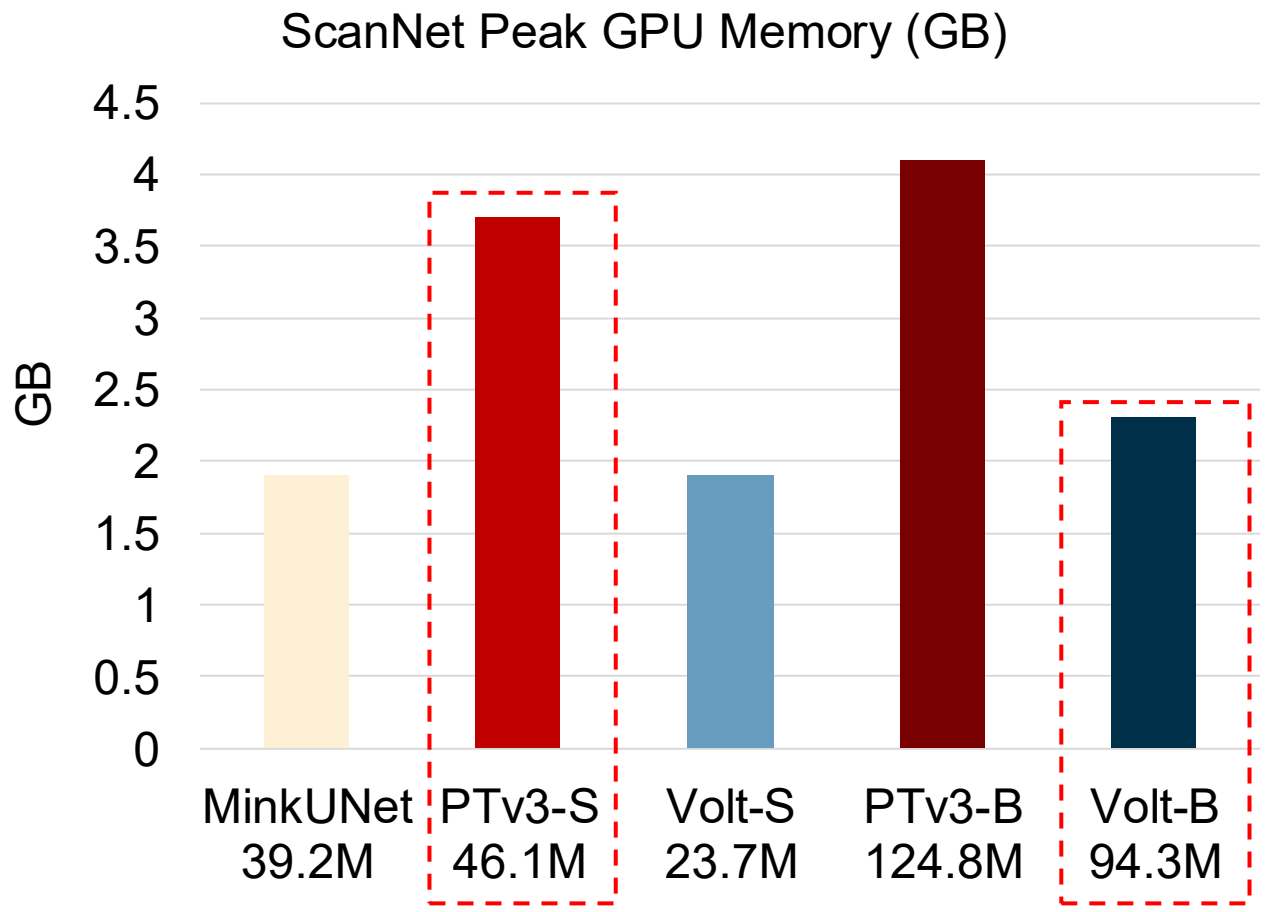


Bridging the Gap: Volume Transformer (Volt)

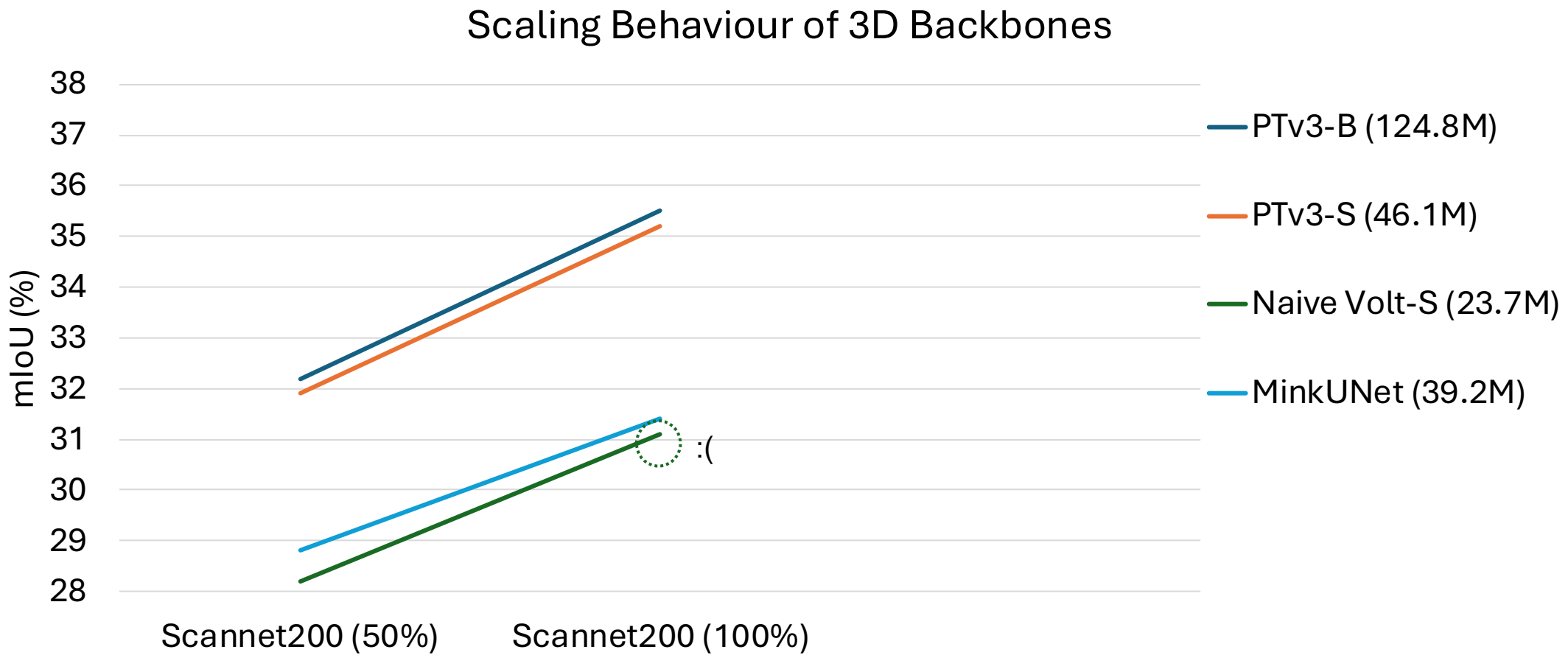


Isn’t Global Attention Expensive?

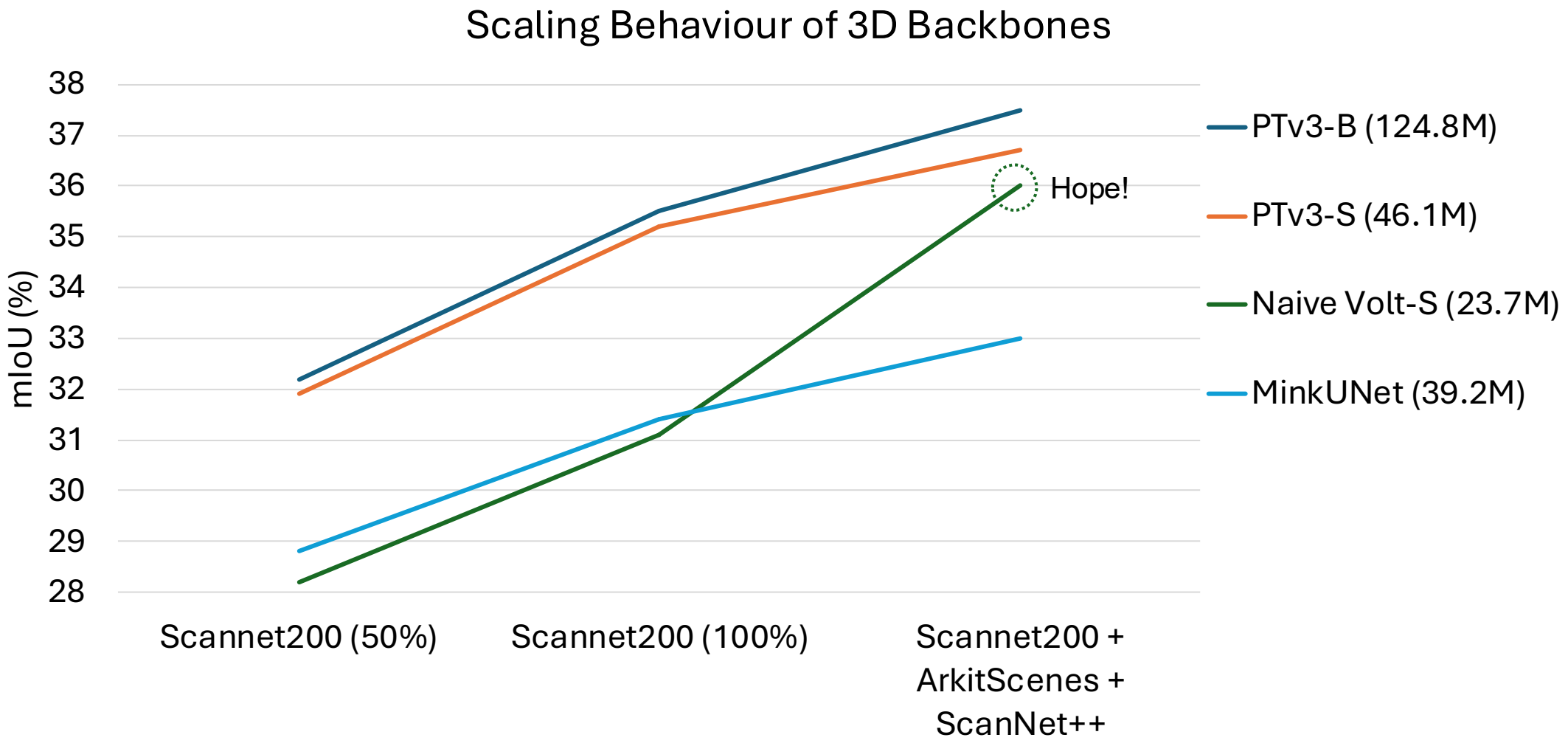




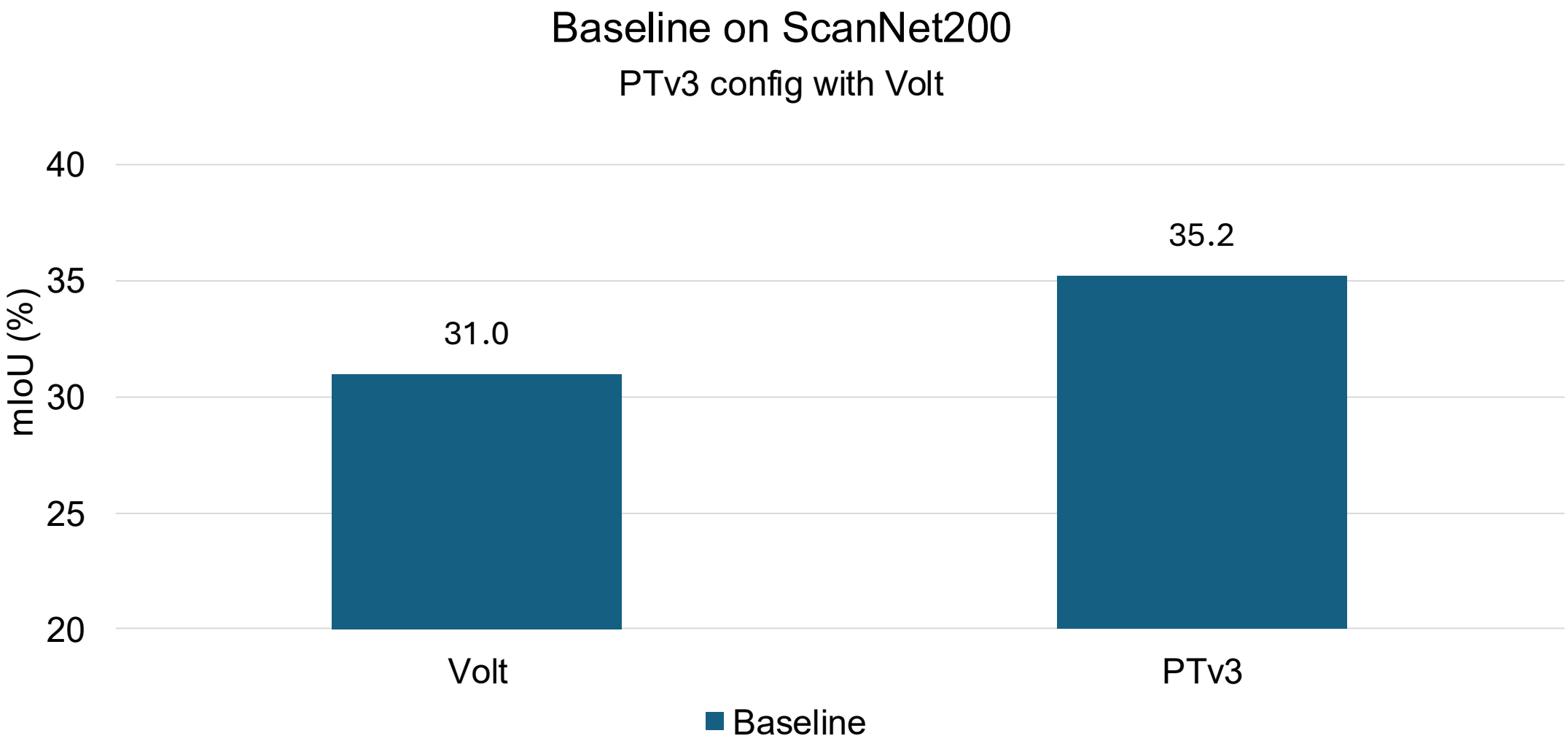
Naive Training of Volt



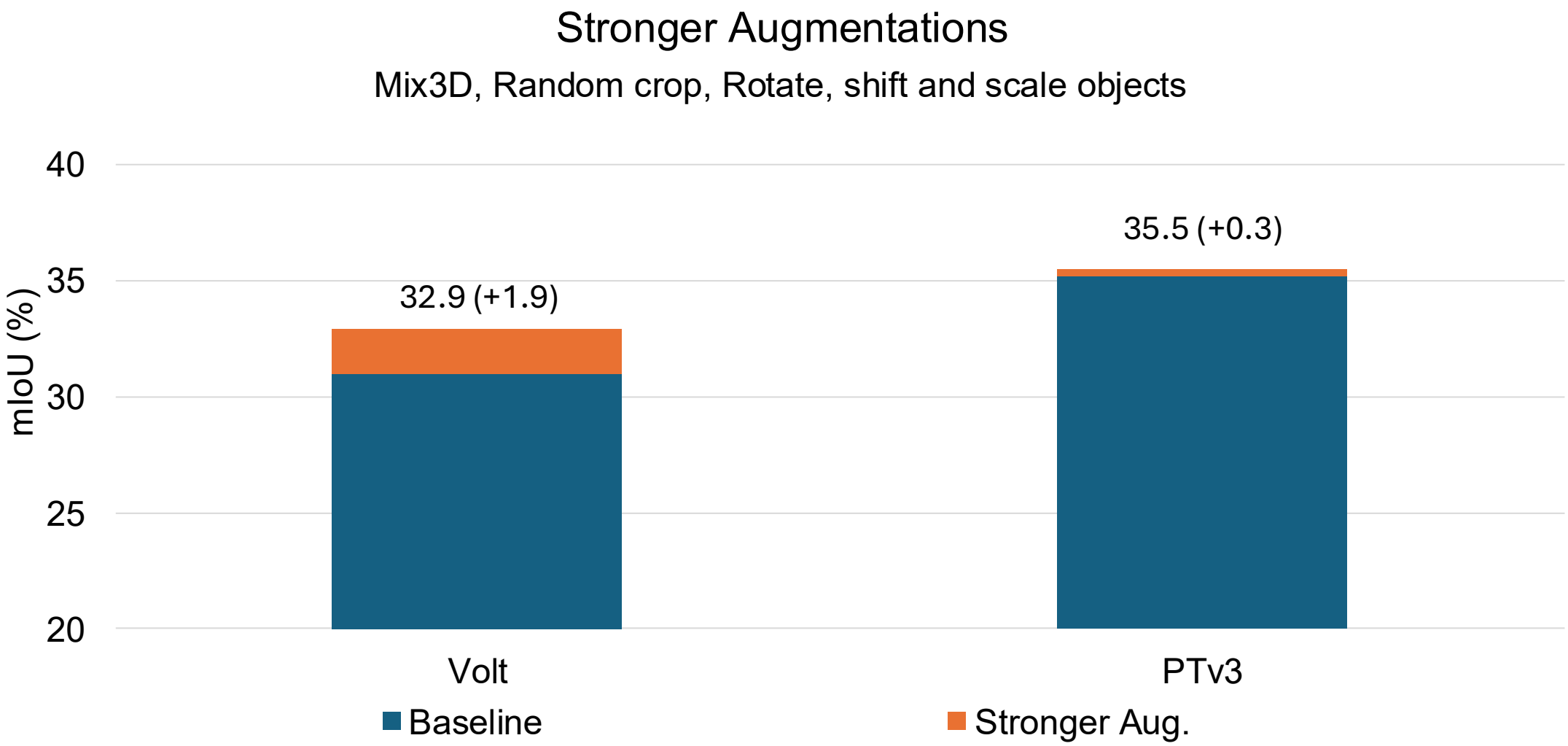
Naive Training of Volt



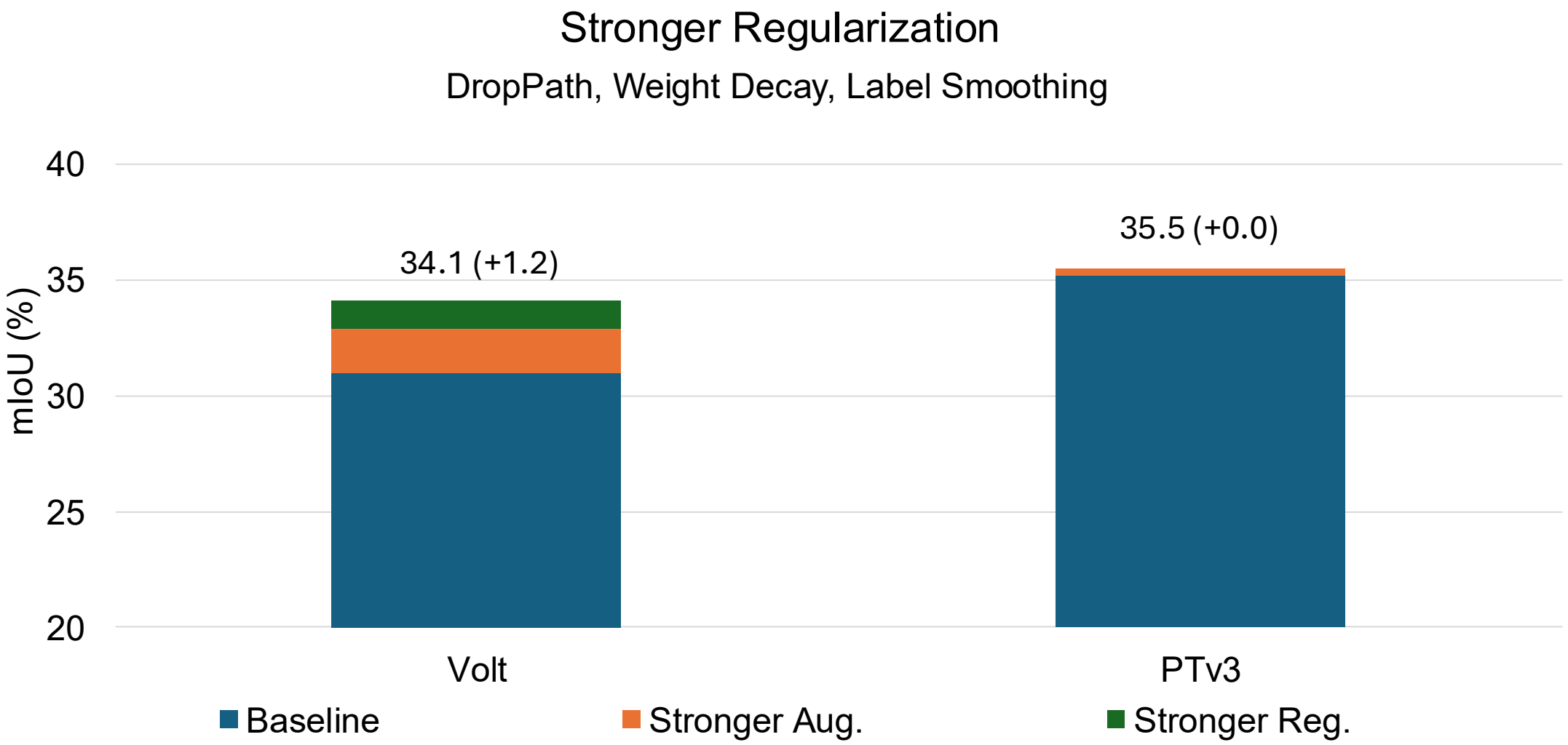
Training Recipe



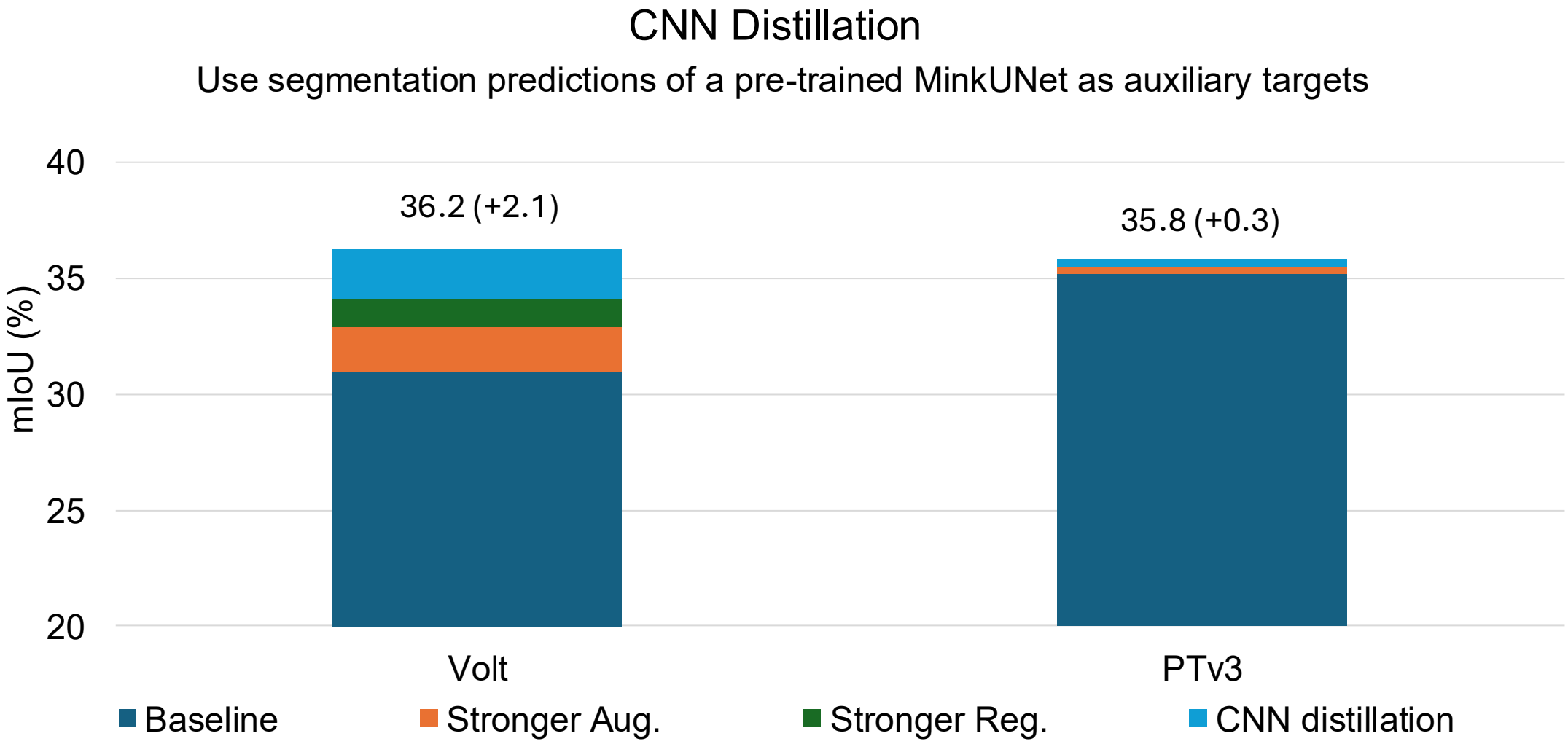
Training Recipe



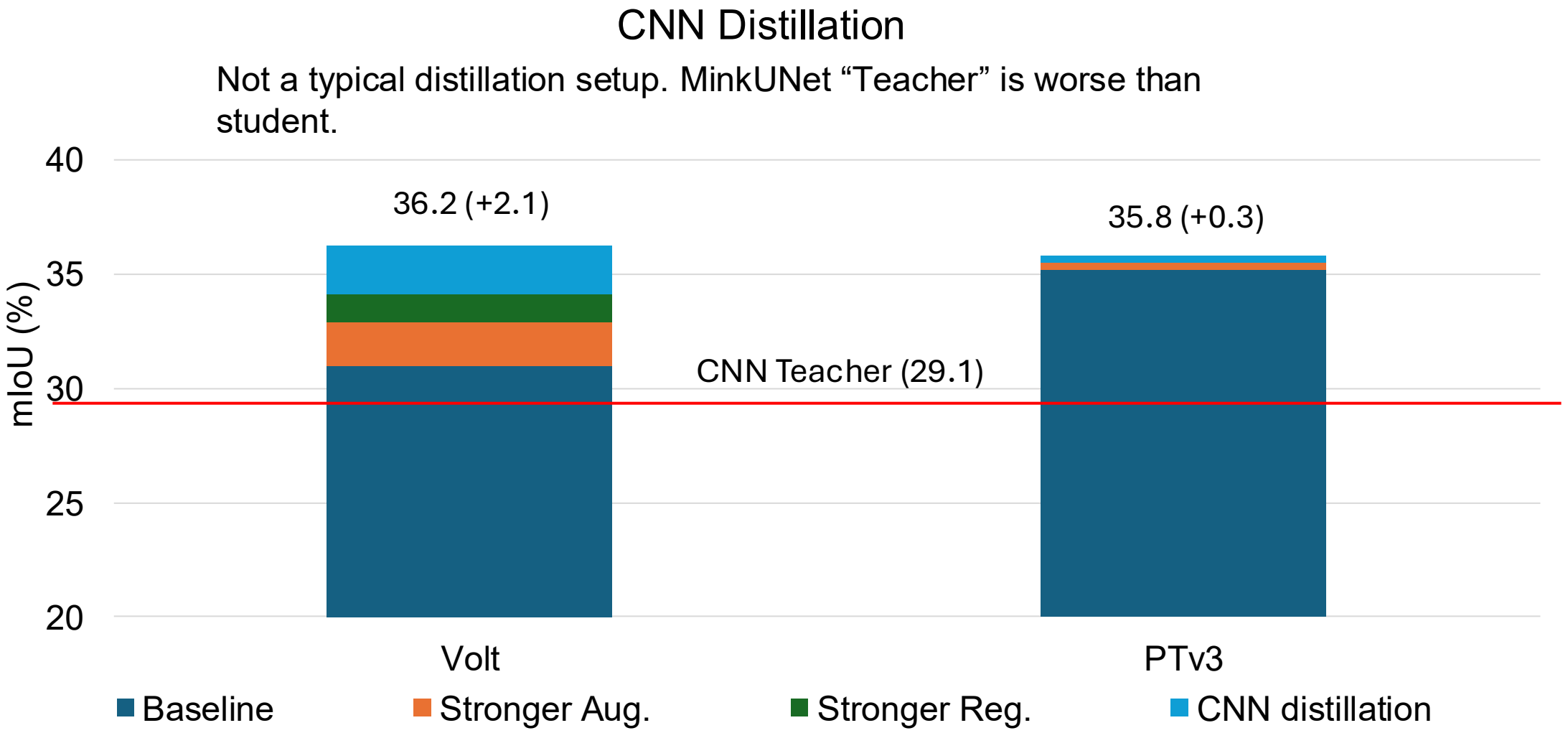
Training Recipe



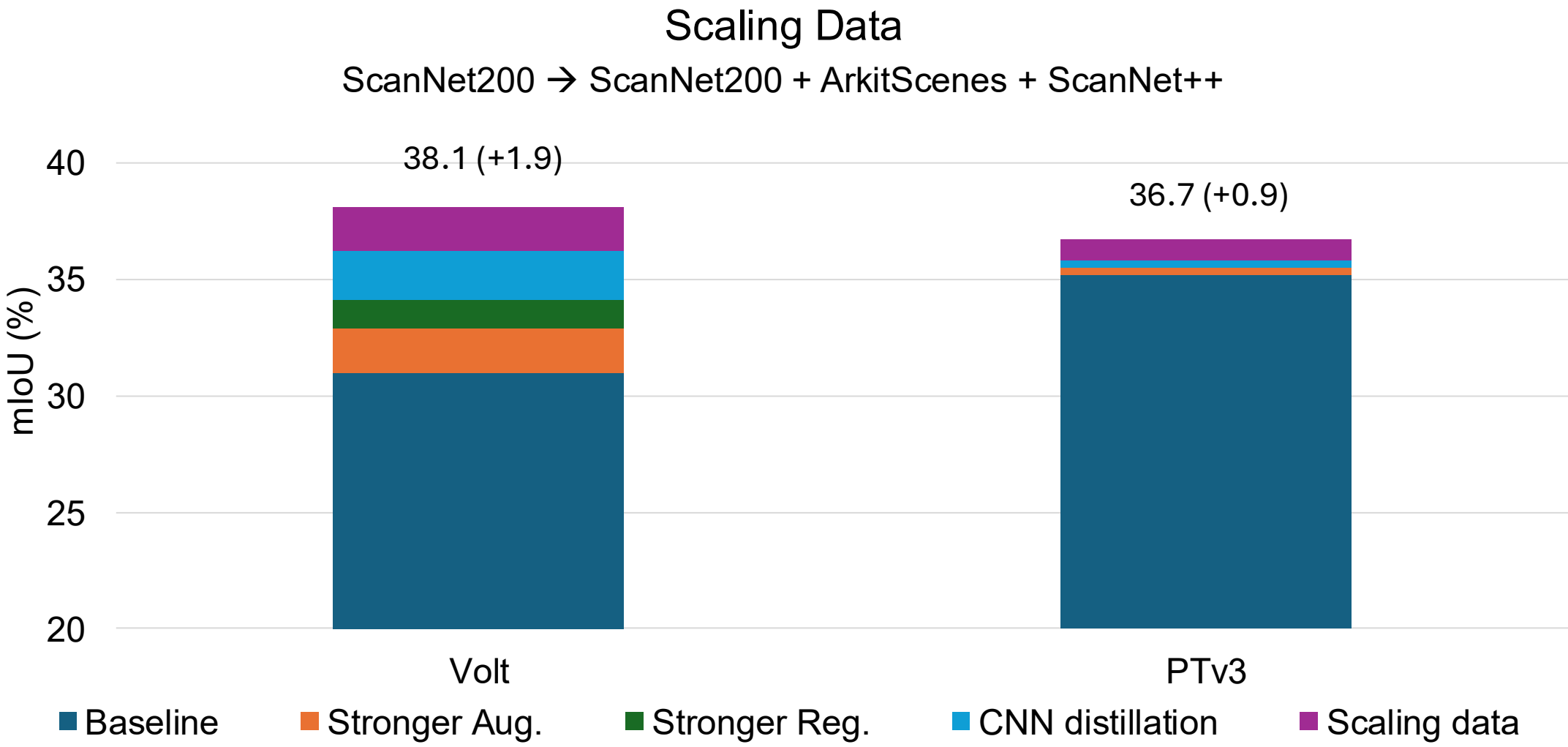
Training Recipe



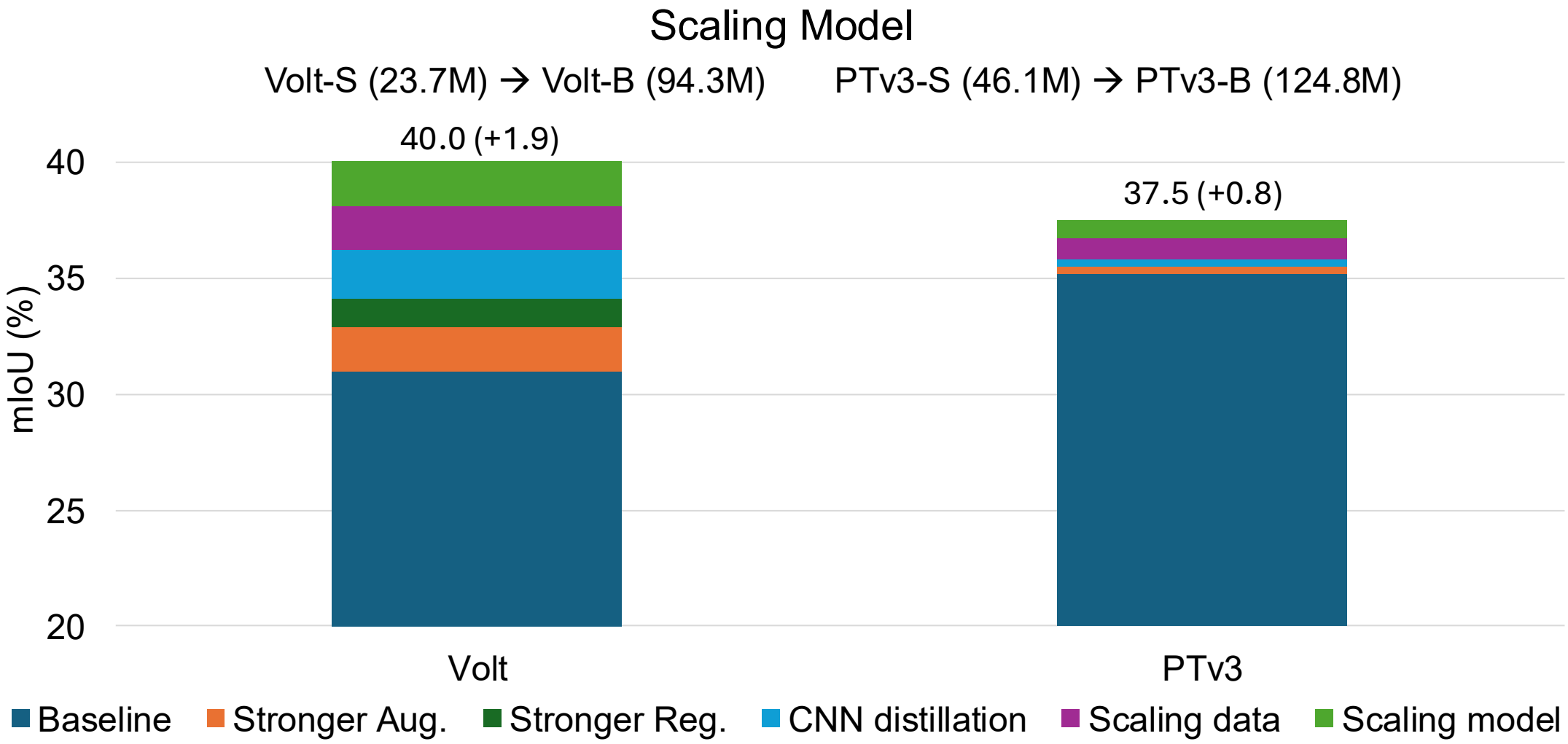
Training Recipe



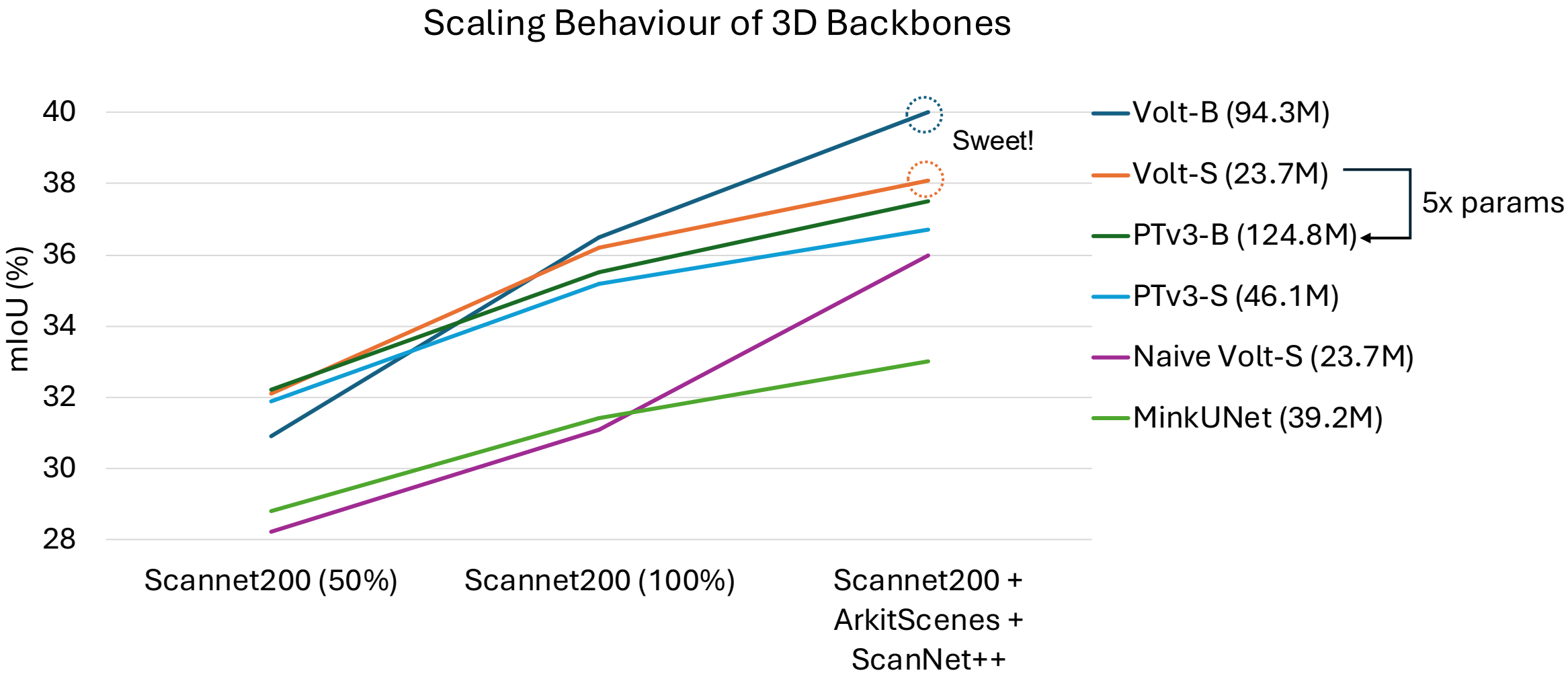
Training Recipe



Training Recipe



Volt Scales Better with Data



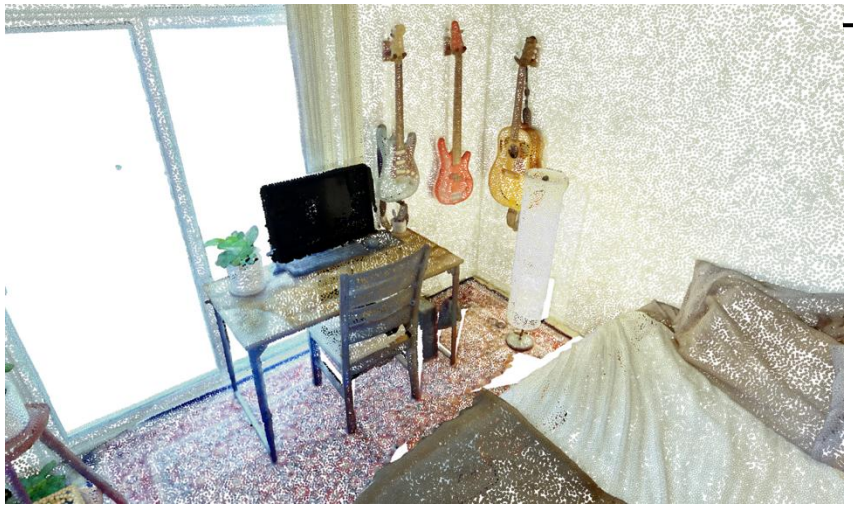
Volt Achieves New Best Scores

Table 1: **Indoor semantic segmentation results (mIoU).**

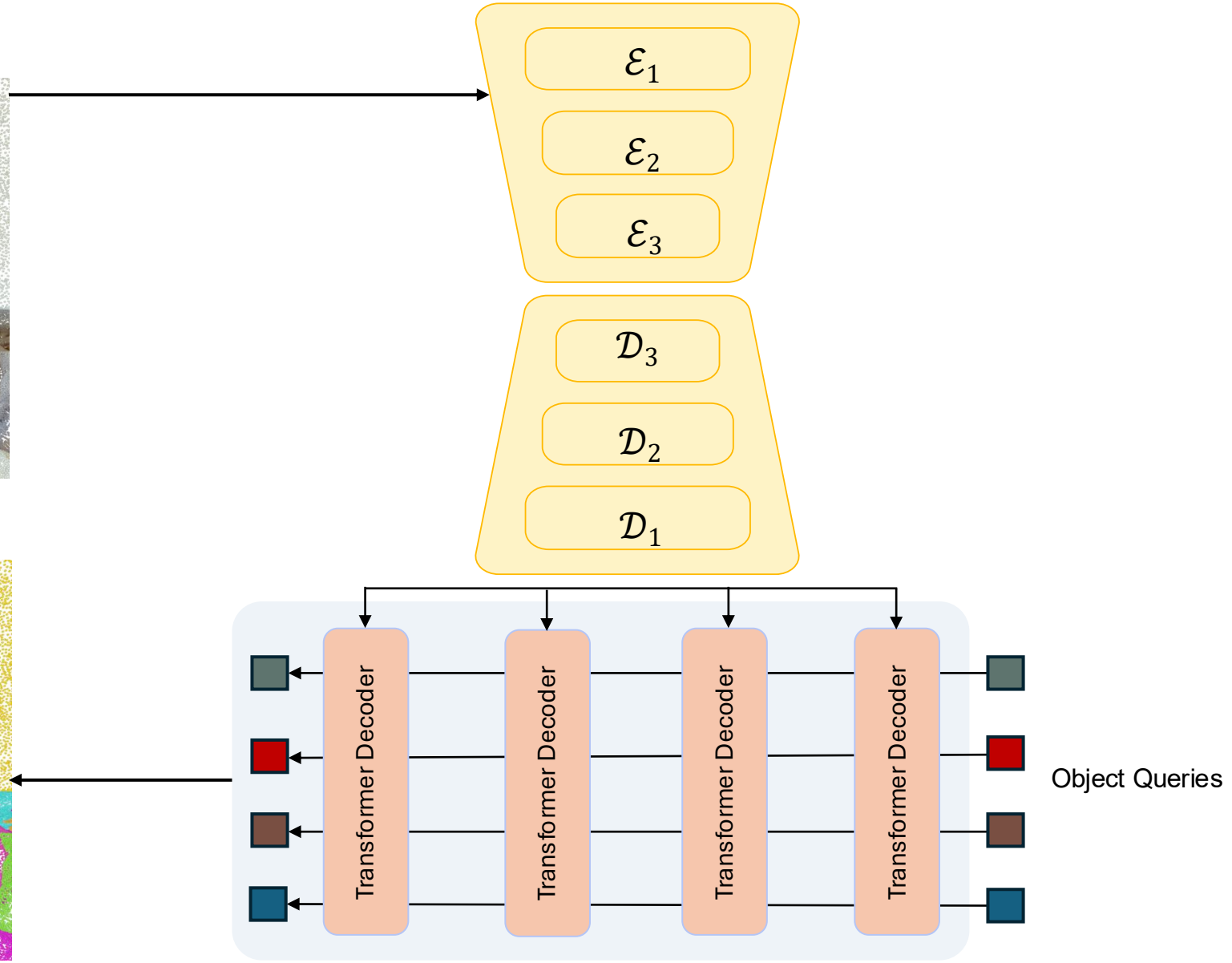
Method	#Params	ScanNet++	ScanNet200		ScanNet	
		Test	Val	Test	Val	Test
ST [22]	18.8M	–	–	–	74.3	73.7
PTv1 [18]	11.4M	–	27.8	–	70.6	–
PointNeXt [74]	41.6M	–	–	–	71.5	71.2
MinkUNet [13]	39.2M	45.6	25.0	25.3	72.2	73.6
OctFormer [54]	44.0M	46.0	32.6	32.6	75.7	76.6
Swin3D [21]	23.6M	–	–	–	76.4	–
PTv2 [19]	12.8M	44.5	30.2	–	75.4	74.2
OA-CNN [75]	51.5M	47.0	32.3	33.3	76.1	75.6
PTv3 [20]	46.1M	48.8	35.2	37.8	77.5	77.9
Volt-S	23.7M	49.3	36.2	–	77.2	–
↓ Jointly trained on multiple datasets						
PTv3/PPT [72, 71]	124.8M	–	37.5	41.4	79.1	79.8
Volt-S	23.7M	–	38.1	–	80.2	–
Volt-B	87.7M	**49.5**	**40.0**	**41.6**	**80.5**	**80.5**

Instance Segmentation (SPFormer)

Input Point Cloud

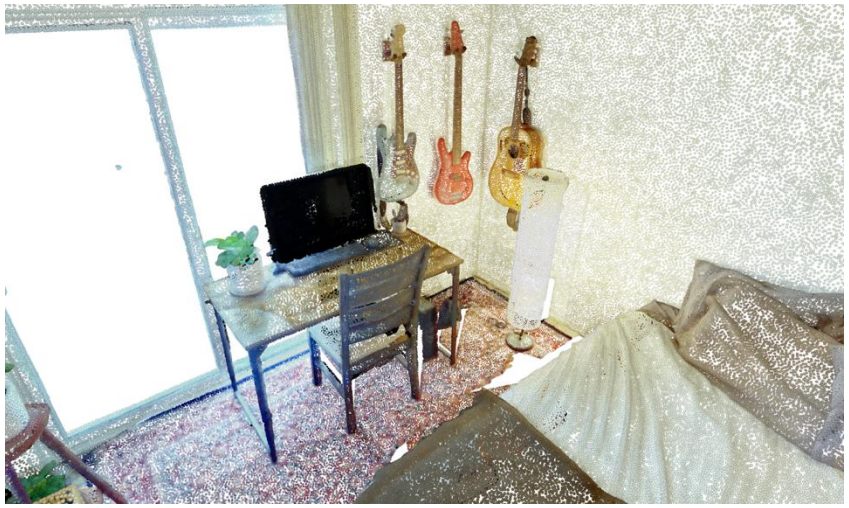


Instance Segmentation

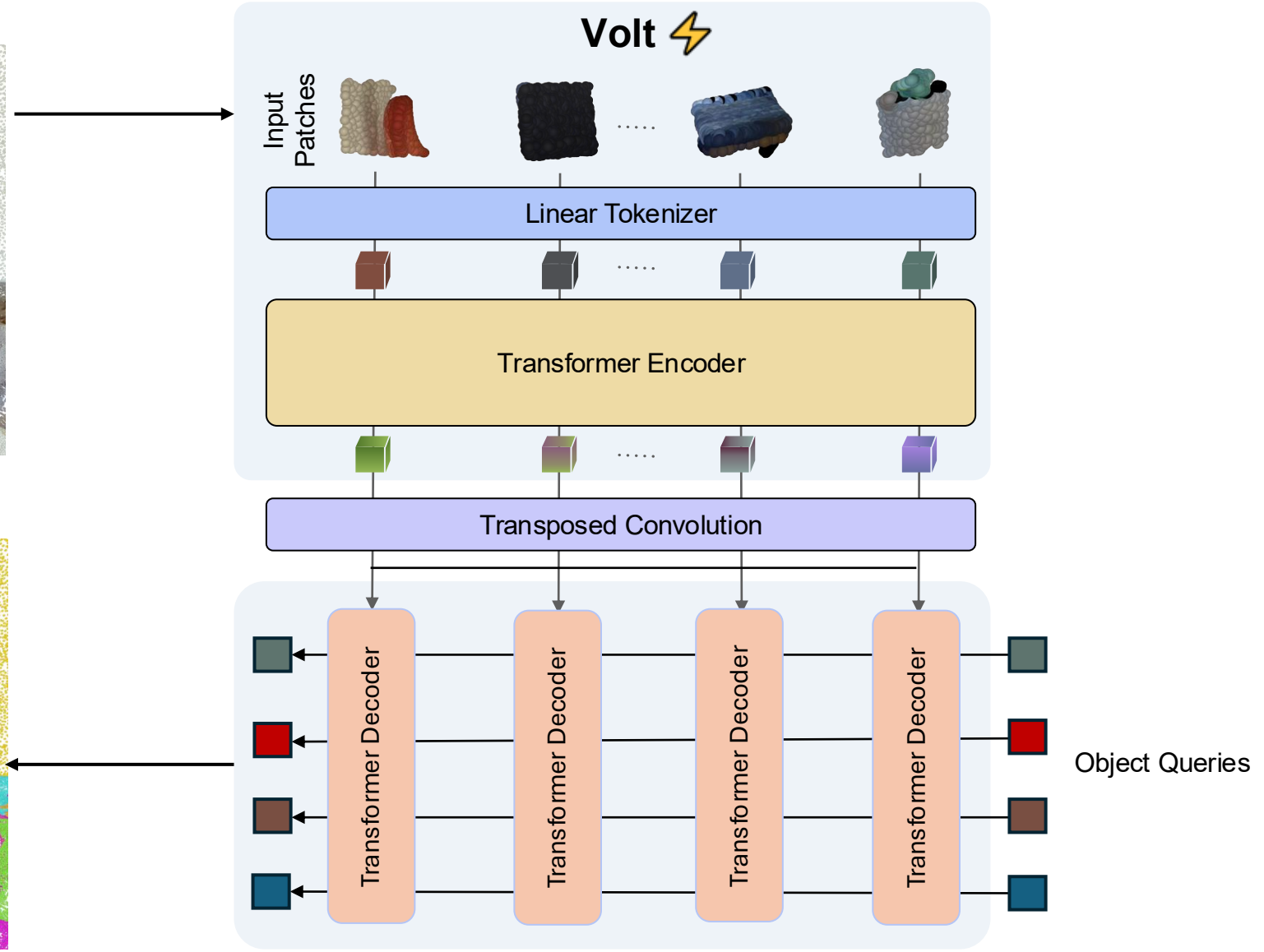


Volt as Backbone for Instance Segmentation

Input Point Cloud



Instance Segmentation



Instance Segmentation Results

Table 3: **Indoor instance segmentation results (mAP50).**

Method	Backbone	ScanNet++	ScanNet200		ScanNet	
		Test	Val	Test	Val	Test
SPFormer [65]	MinkUNet	43.5	33.8	–	73.9	77.0
Mask3D [15]	MinkUNet	–	37.0	38.8	73.7	78.0
OneFormer3D [16]	MinkUNet	43.3	40.8	–	78.1	80.1
MAFT [85]	MinkUNet	–	38.2	–	76.5	78.6
QueryFormer [86]	MinkUNet	–	37.1	–	74.2	78.7
SphericalMask [87]	MinkUNet	–	–	–	79.9	81.2
Relation3D [88]	MinkUNet	–	41.2	–	80.2	81.6
SGIFormer [89]	MinkUNet	45.7	39.4	–	**81.2**	79.9
SPFormer [65]	**Volt-B**	**54.9**	**48.4**	**47.5**	78.3	**82.7**

SceneFUN3D Challenge

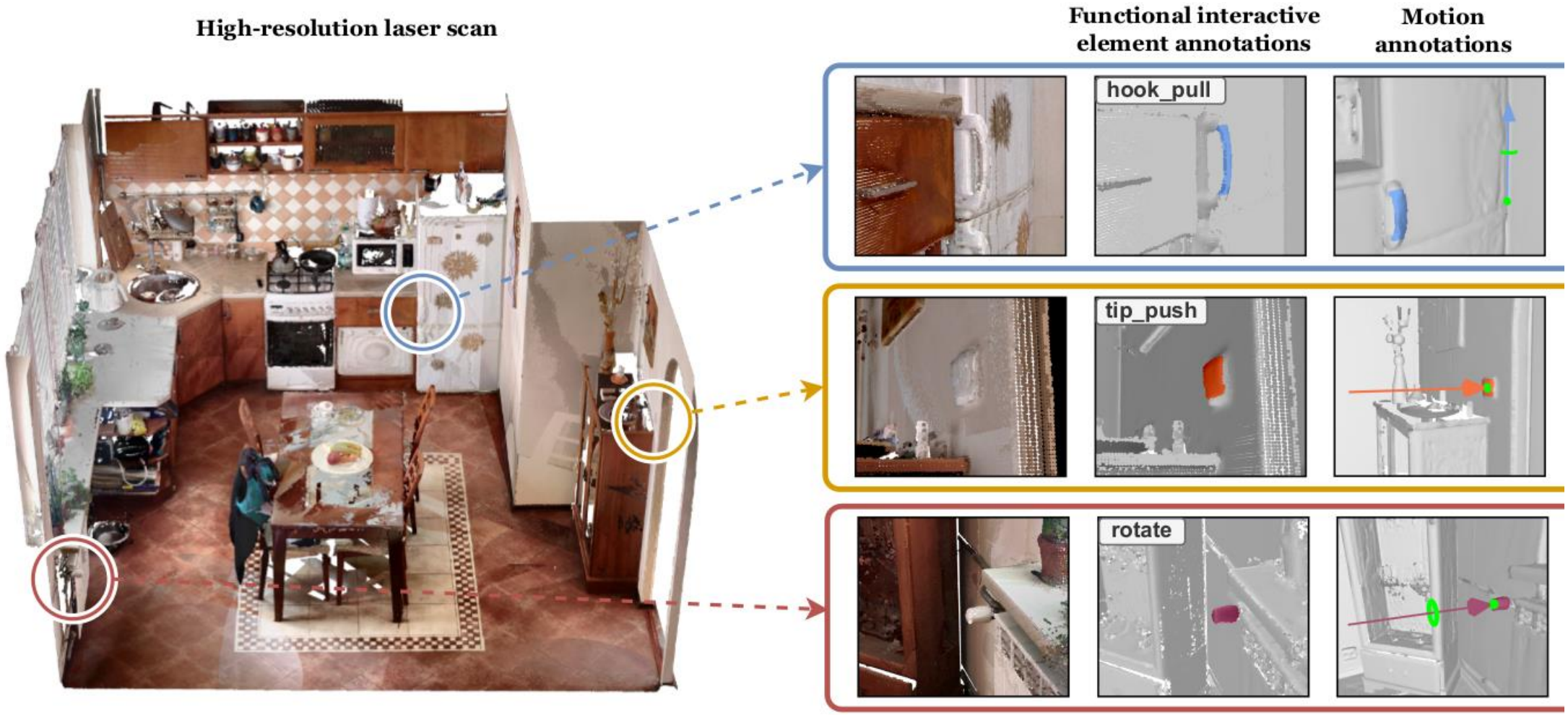


Image taken from SceneFUN3D paper

SceneFUN3D Results

Functionality Segmentation (test split)

Team/Method	AP ↑	AP_50 ↑	AP_25 ↑
Volt	**22.72**	**34.47**	**42.63**
servette	7.75	13.19	19.30
pico-mr	6.54	13.97	27.82
Z. Zhou, S. Wei, Z. Wang, C. Wang, X. Yan, X. Liu. "OpenTrack3D: Towards Accurate and Generalizable Open-Vocabulary 3D Instance Segmentation". CVPR 2026 Findings			
curiosAI	4.83	9.67	16.21
RPL	2.91	6.46	16.91
Night's Watch	2.39	4.76	7.88
MaNET	0.76	3.76	13.92
SegFunCT	0.00	0.00	10.69

ViT moment for 3D Scene Understanding

- New ViT-style backbone for 3D scenes
- Fast and memory efficient even with global attention
- Use the training recipe for training from scratch
- Better scaling with more data
- Better scaling with new hardware/software
- Enables transfer of research from the general scientific community

Volt Project Page