Occlusion-aware Text-Image-Point Cloud Pretraining for Open-World 3D Object Recognition

Bridging the Reality Gap in 3D Object Recognition

Khanh Nguyen

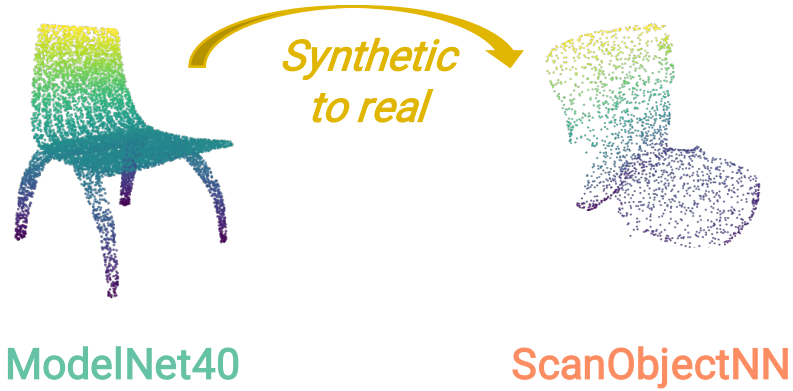
Ghulam Mubashar Hassan

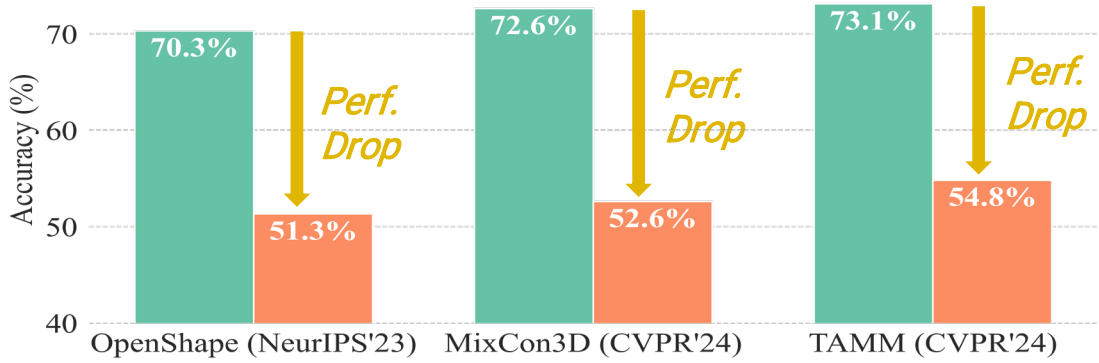
Ajmal Mian

Real-world 3D recognition challenges

Open-world 3D recognition models: must generalize beyond *seen* categories and recognize *novel* objects

1. **Synthetic vs. real:** SOTA models struggle with **real-world, partial 3D scans** due to training on **complete synthetic** data



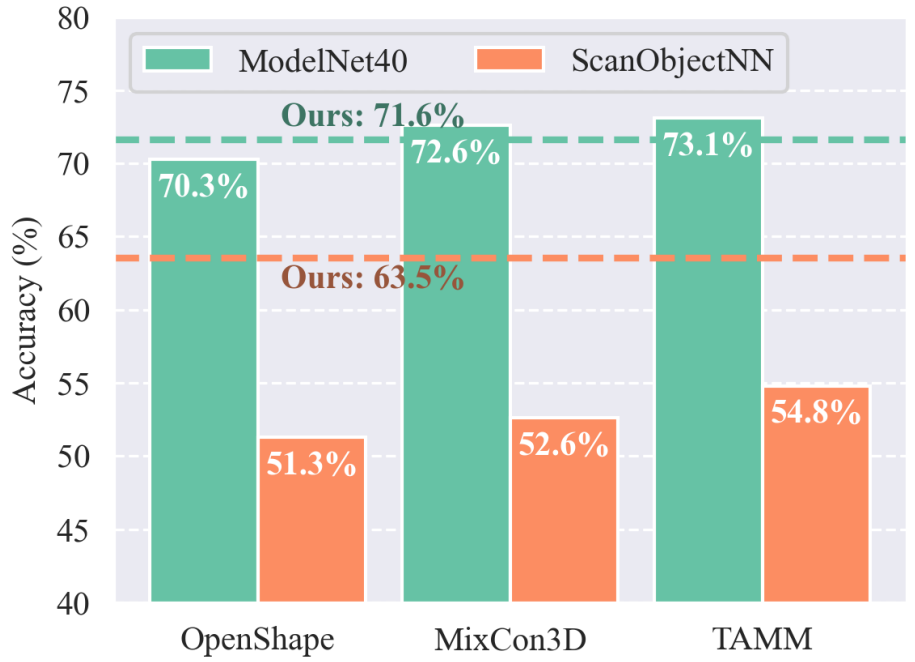


2. **High computation cost:** Transformer-based SOTA models have high inference costs due to **quadratic complexity**

Contributions

1. **Synthetic vs. real:**

   → **OccTIP:** an occlusion-aware pretraining framework that **simulates incomplete 3D data from synthetic data**



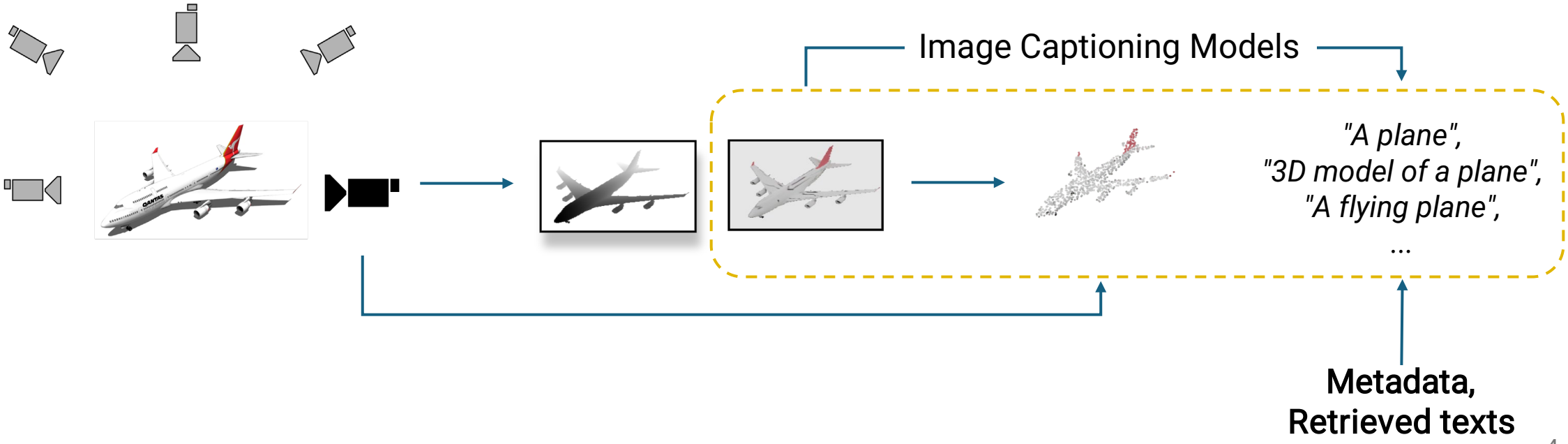
2. **High computation cost:**

   → **DuoMamba:** an efficient model tailored for point clouds with **linear-time complexity**

Contribution 1: OccTIP

Occlusion-aware Text-Image-Point Cloud Pretraining Framework

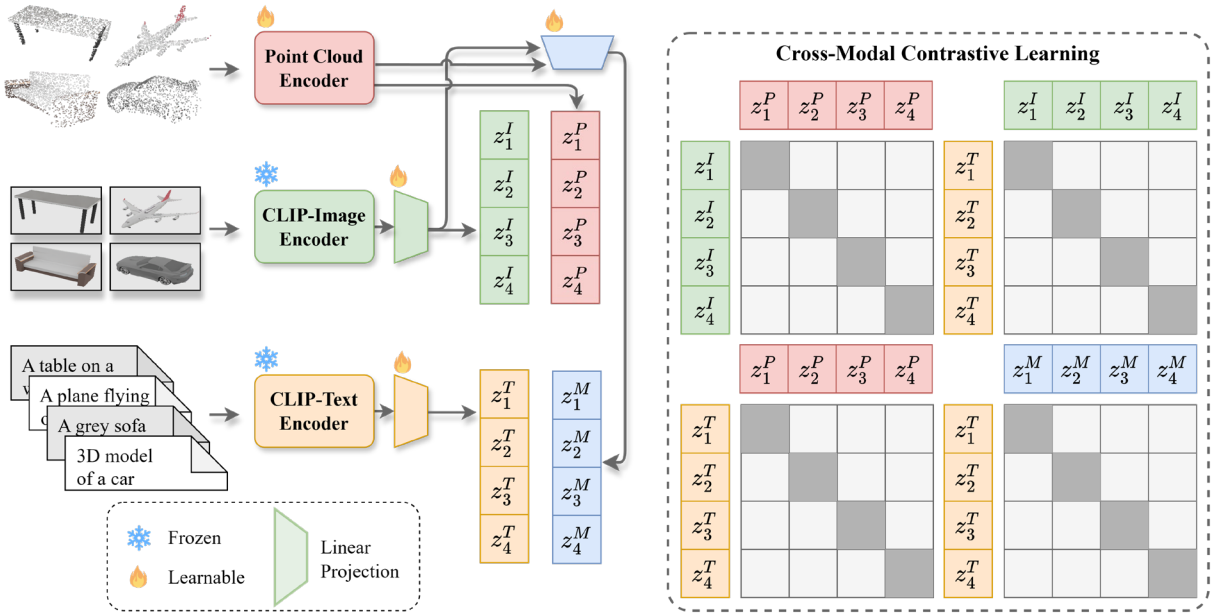
- ✓ Generates **realistic partial 3D scans** by simulating occlusions through rendering



Contribution 1: OccTIP

Occlusion-aware Text-Image-Point Cloud Pretraining Framework

- ✓ **Aligns point cloud representations** with rich CLIP-based **image and text** embeddings through **cross-modal contrastive learning**

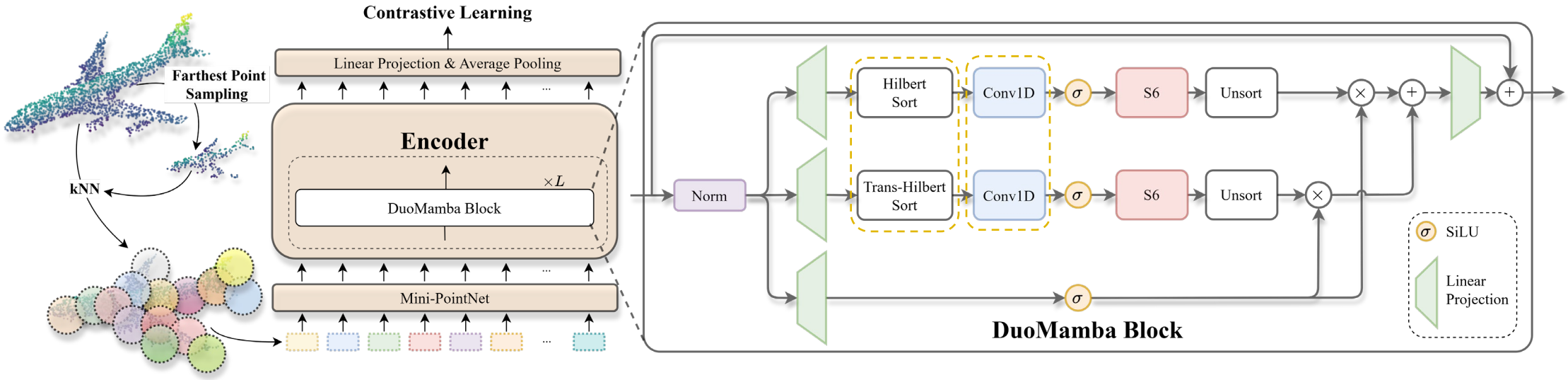


Training objective minimizes the total **InfoNCE** losses:

$$L = L^{P\leftrightarrow T} + L^{P\leftrightarrow I} + L^{I\leftrightarrow T} + L^{M\leftrightarrow T}$$

where T, I, P, M denote text, image, point cloud, and mixed (3D and 2D) modalities

Contribution 2: DuoMamba



To build our **linear-time** model, we adapt Mamba's selective SSMs (S6) into our **two-stream** blocks, incorporating key changes for point clouds:

- ✓ **Two Hilbert curves**: transform unordered point clouds into **spatially meaningful sequences**, preserving local geometry
- ✓ **Standard Conv1D**: captures **richer local geometry** via **bidirectional** information flow among nearby 3D point patches

Experiments: zero-shot classification

Datasets:

- **ModelNet40-P:** synthetic partial point clouds created from ModelNet40 using OccTIP
- **ScanObjectNN:** real-world partial point cloud dataset

Remarks:

- OccTIP framework consistenty **improves the accuracy** of SparseConv and PointBERT
- OccTIP + DuoMamba achieves **the best performance**

Method	Encoder	ModelNet40-P			ScanObjectNN		
		Top 1	Top 3	Top 5	Top 1	Top 3	Top 5
OpenShape [26]	SparseConv [4]	42.1	61.6	69.4	52.7	72.7	83.6
TAMM [59]		45.5	64.8	73.1	57.9	75.3	83.1
MixCon3D [12]		-	-	-	54.4	73.9	83.3
MixCon3D† [12]		42.1	59.3	67.5	56.0	73.2	82.8
OccTIP		**64.5**	**81.0**	**86.7**	**61.7**	**78.4**	**86.9**
OpenShape [26]	PointBERT [55]	46.3	64.2	71.9	51.3	69.4	78.4
TAMM [59]		45.6	66.2	74.7	54.8	74.5	83.3
MixCon3D [12]		-	-	-	52.6	69.9	78.7
MixCon3D† [12]		50.3	69.7	78.6	55.5	72.8	81.1
OccTIP		**67.7**	**82.7**	**87.3**	**60.6**	**78.2**	**86.0**
OpenDlign [29]	ViT-H-14 [10]	-	-	-	59.5	76.8	83.7
OccTIP	DuoMamba	**67.7**	**82.9**	**87.8**	**63.5**	**81.3**	**89.2**

Experiments: 3D object detection

Experimental setting:

We leveraged off-the-shelf **3DETR-m for bounding box prediction** and the **pretrained zero-shot models for classification**

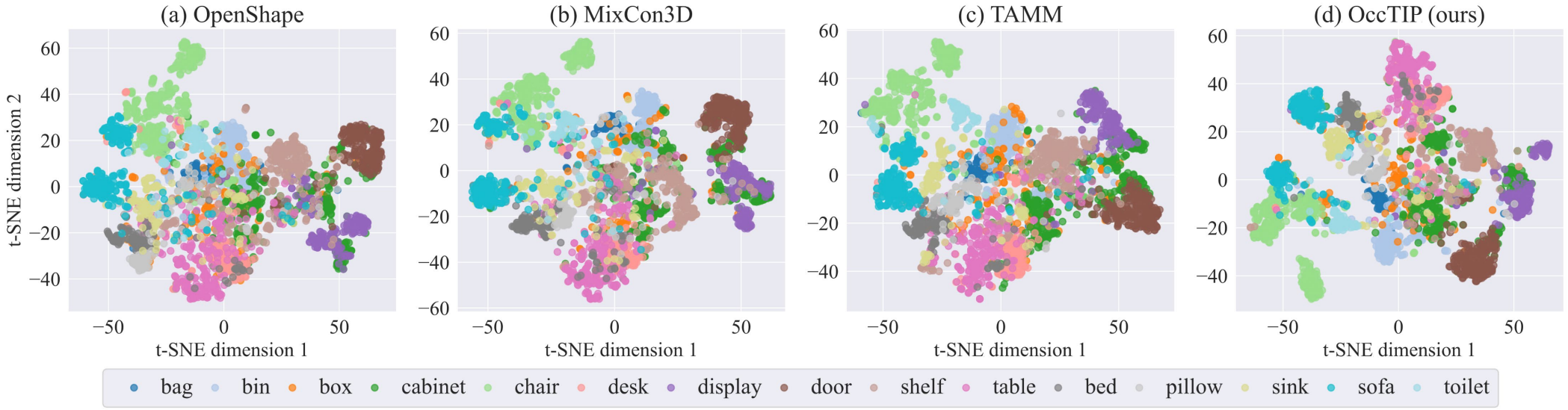
Remarks:

OccTIP (ours) achieves the **best results** on both ScanNetV2 and SUN RGB-D datasets

	Method	ScanNetV2	SUN RGB-D
mAP_{25}	PointCLIP [19]	6.0	-
	PointCLIP V2 [65]	19.0	-
	OpenShape* [26]	20.4	18.6
	MixCon3D† [12]	24.1	18.7
	TAMM* [59]	23.1	18.9
	OccTIP	**28.9**	**24.4**
mAP_{50}	PointCLIP [58]	4.8	-
	PointCLIP V2 [65]	11.5	-
	OpenShape* [26]	16.1	9.8
	MixCon3D† [12]	19.1	9.6
	TAMM* [59]	18.1	10.0
	OccTIP	**22.7**	**13.0**

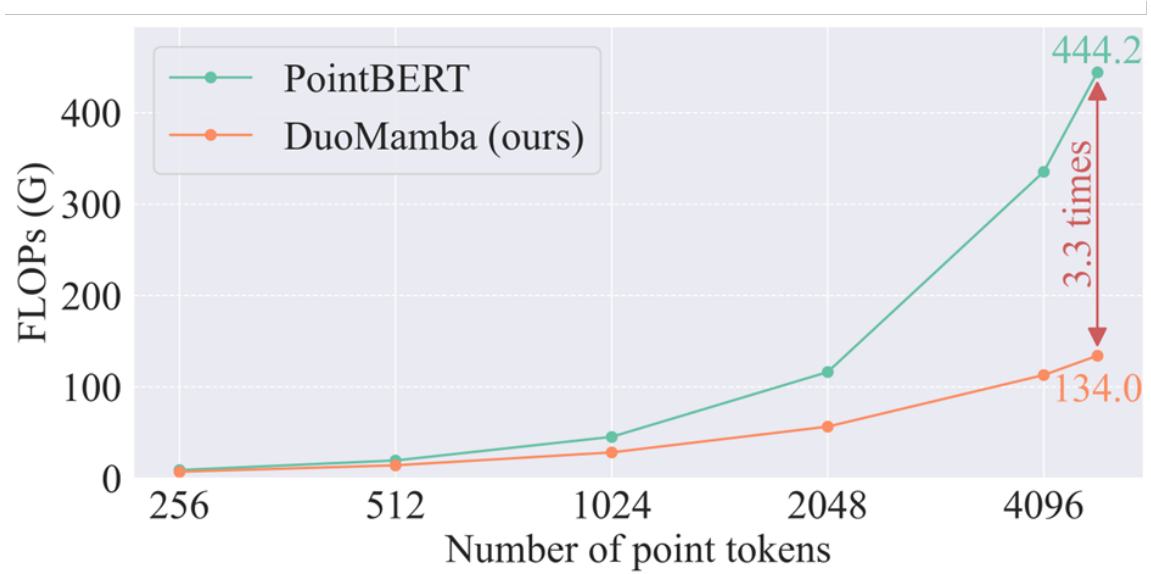
Latent space visualization

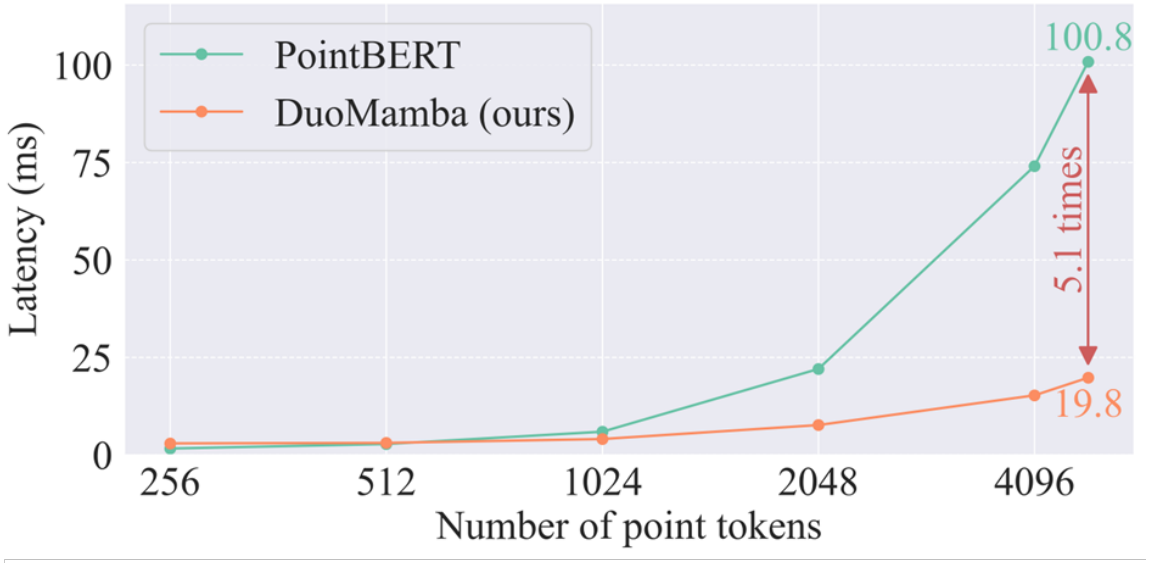
- We extract the features of ScanObjectNN objects using pretrained models
- Our method **OccTIP** achieves **clearer category separation** and significantly **reduces inter-class overlap**



Computation and latency

DuoMamba dramatically reduces **computation** (up to 3.3 times lower FLOPS) and inference **latency** (up to 5.1 times faster) compared to Transformer-based PointBERT, especially at higher point token counts





Summary:

- **OccTIP**: bridging synthetic-to-real 3D gap
- **DuoMamba**: fast, state-of-the-art 3D architecture
- **Impact**: enabling robust real-world 3D AI

Project page: ndkhanh360.github.io/project-occtip

Contact information: duykhanh.nguyen@research.uwa.edu.au

Thank you for listening!