Boosting Multi-View Indoor 3D Object Detection via Adaptive 3D Volume Construction

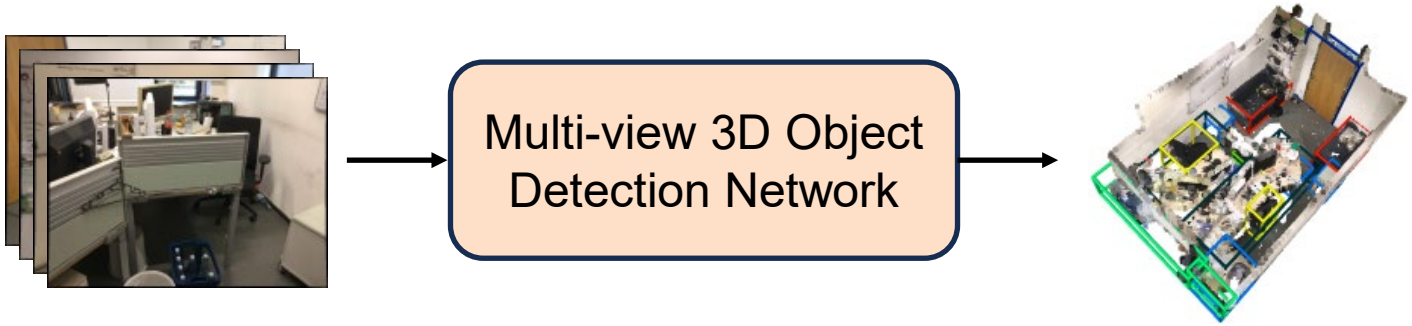
Runmin Zhang, Zhu Yu*, Si-Yuan Cao*, Lingyu Zhu,
Guangyi Zhang, Xiaokai Bai, Hui-Liang Shen

https://github.com/RM-Zhang/SGCDet

Motivation

Problem Setting:

Detect 3D objects from multi-view posed images.



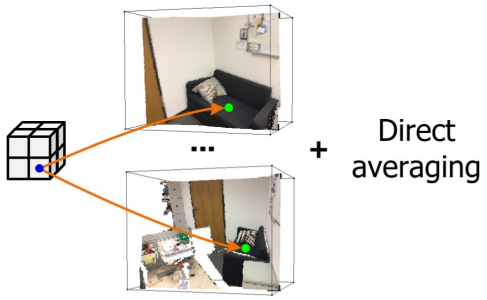
Key:

Bridge the gap between 2D images and 3D representations.

Challenges

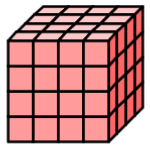
Limitations of previous approaches:

- Restrict the receptive field of voxels to a limited region, overlooking the valuable **contextual information** of images.



Single point sampling

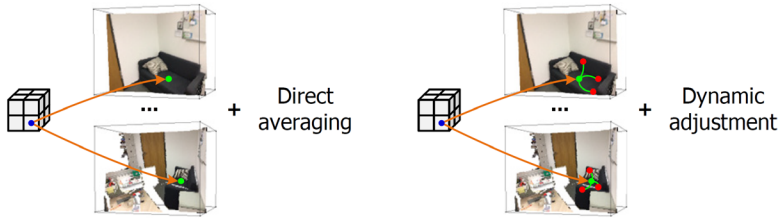
- Construct high-resolution, dense 3D volumes, failing to account for the **inherent sparsity** of 3D scenes.



Dense volume construction

Contributions

- **Geometry and context aware aggregation** that enables each voxel to adaptively aggregate geometric and contextual features within a deformable region **in each view**, and dynamically adjusts feature contributions **across different views.**



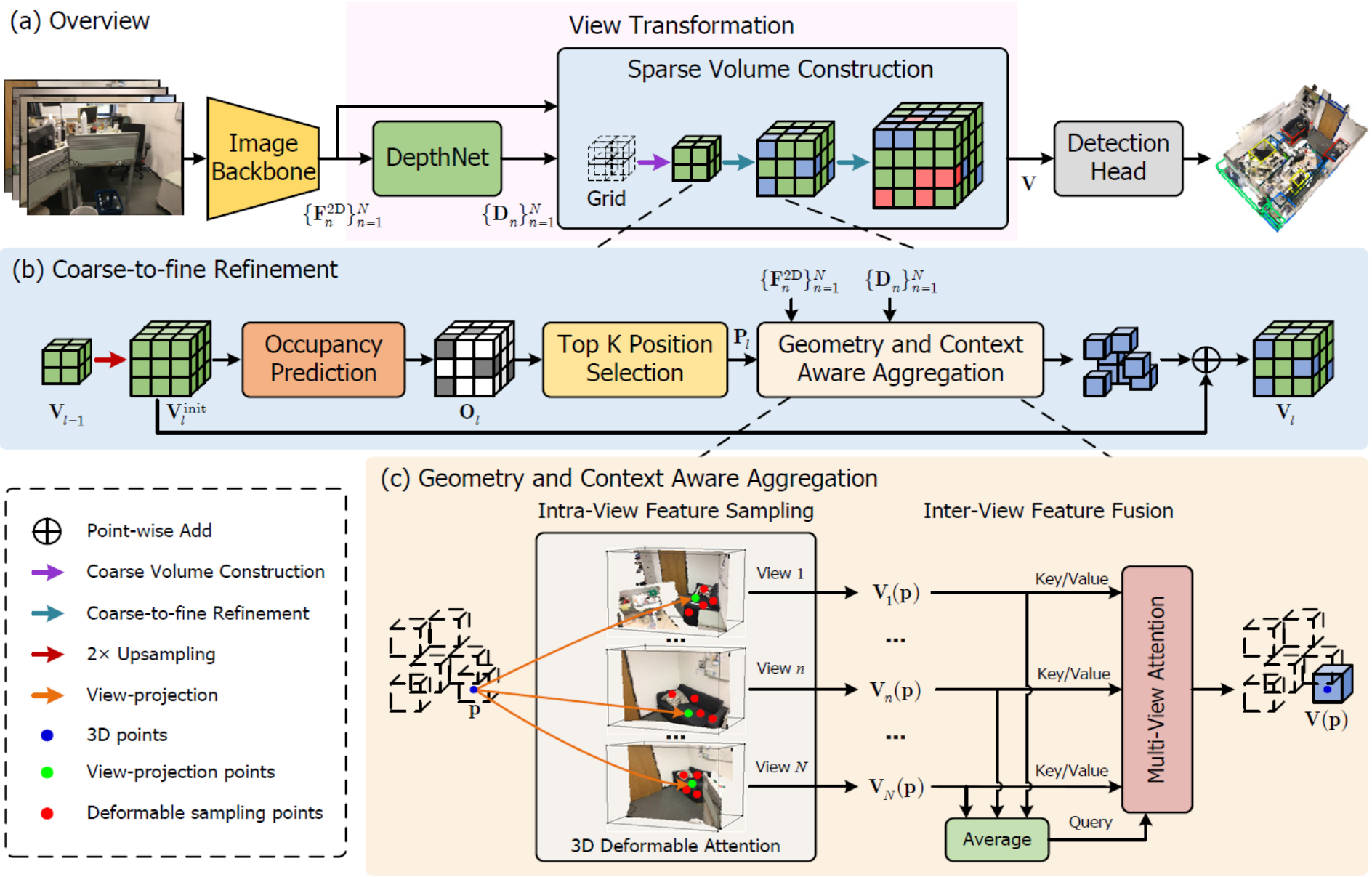
(a) Single point sampling (b) Geometry and context aware aggregation (Ours)

- **Sparse volume construction** that **selectively refines** voxels likely to contain objects, reducing computations in free space.

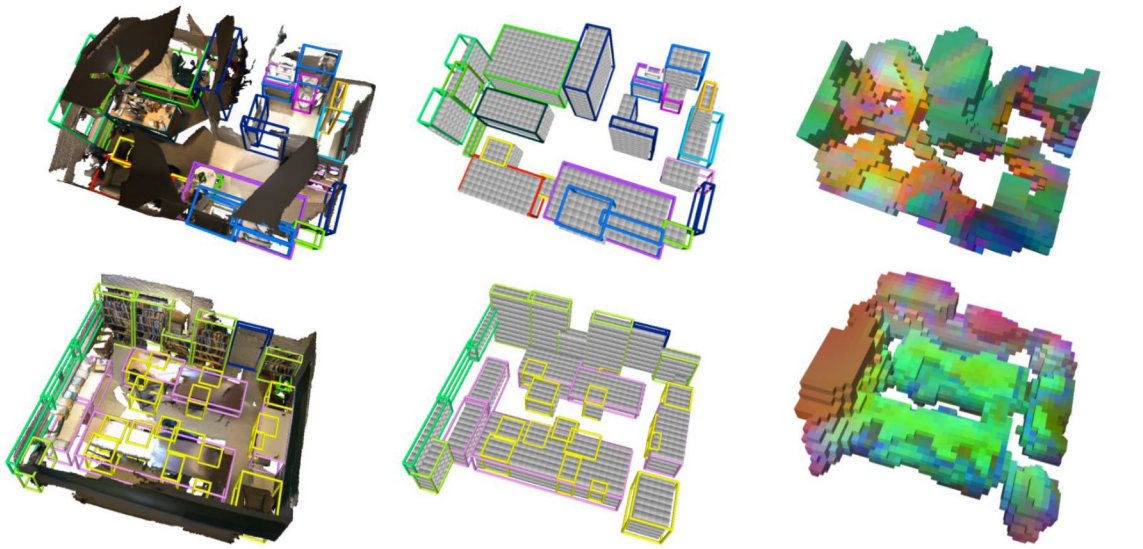
(c) Dense volume construction (d) Sparse volume construction (Ours)

- Training using **only 3D bounding boxes**, without requiring ground-truth geometry.

SGCDet: Overall Framework

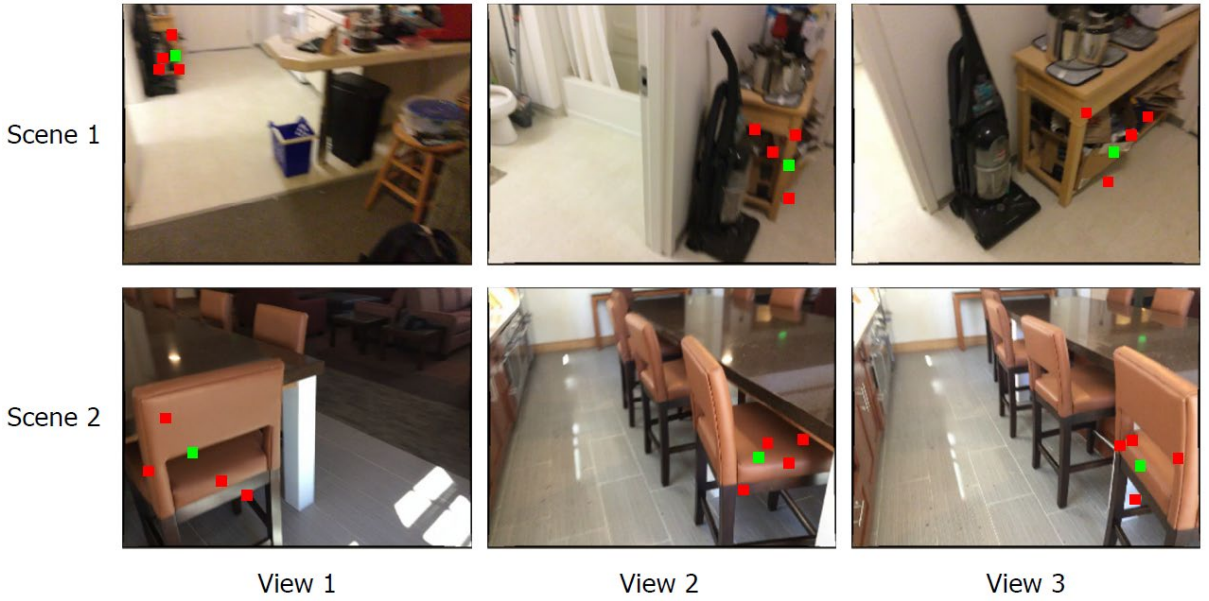


Visualization



(a) Ground-truth (b) Pseudo label for occupancy (c) Refined volume features

Visualization of sparse volume construction



Visualization of sampling locations in intra-view feature sampling

Experiments

Quantitative results

Table 1. Quantitative results and computational cost on the ScanNet dataset. * denotes the results are directly cited from [30, 40].

Method	Voxel Resolution	Performance		Training Cost		Inference Cost	
		mAP@0.25	mAP@0.50	Memory (GB)	Time (Hours)	Memory (GB)	FPS
With ground-truth geometry supervision.							
ImGeoNet* [31]	40×40×16	54.8	28.4	13	16	11	2.50
CN-RMA* [30]	256×256×96	58.6	**36.8**	43	242	12	0.26
Without ground-truth geometry supervision.							
ImVoxelNet* [29]	40×40×16	46.7	23.4	11	13	9	2.60
NeRF-Det* [38]	40×40×16	53.5	27.4	13	14	12	1.30
MVSDet* [40]	40×40×16	56.2	31.3	35	36	28	0.87
SGCDet (Ours)	40×40×16	**61.2**	35.2	20	19	14	1.46

Table 2. Quantitative results on the ScanNet200 dataset. * denotes the results are directly cited from [31]. The voxel resolution of all approaches is 80×80×32.

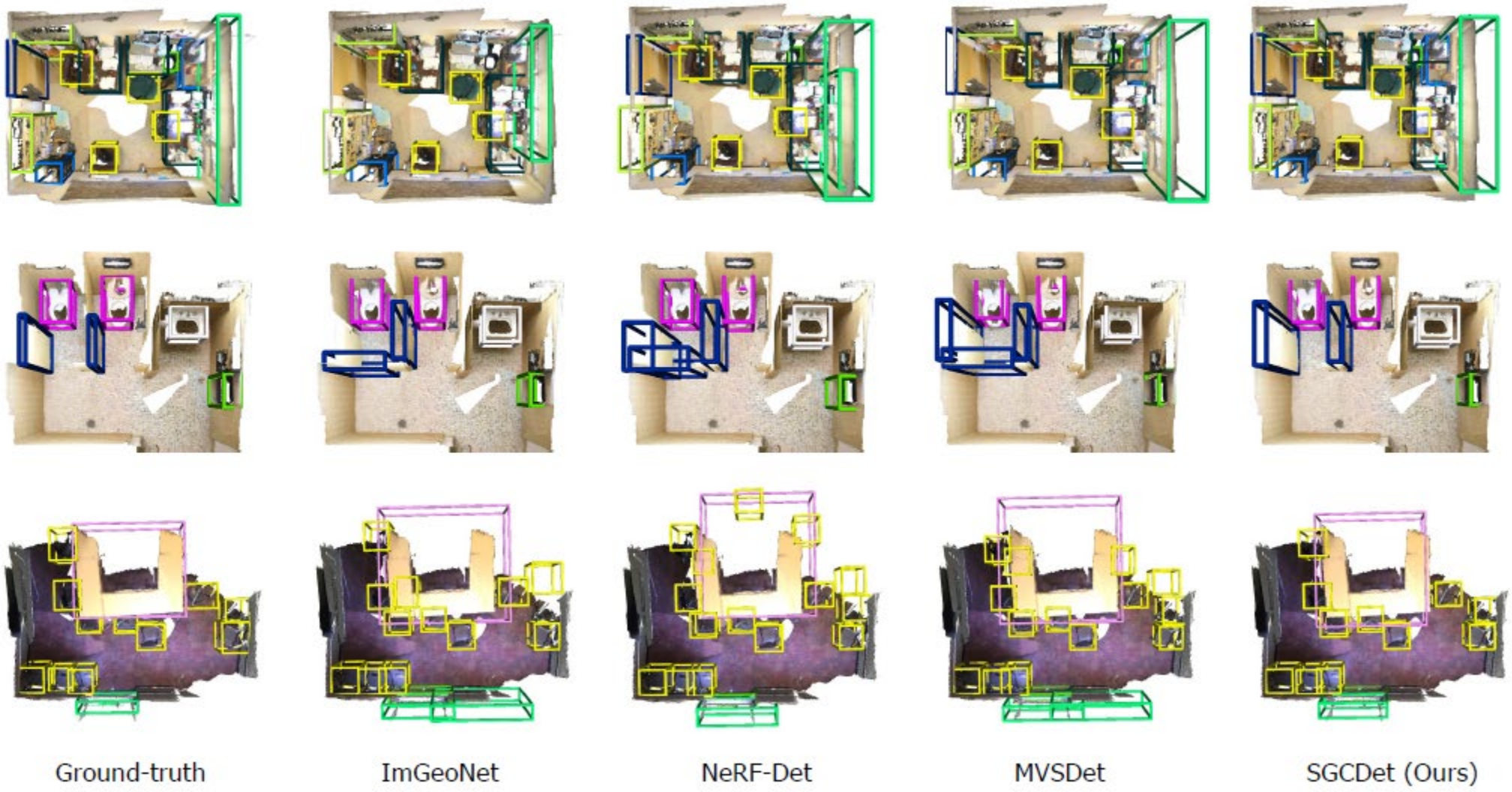
Method	Performance (mAP@0.25)			
	Total	Head	Common	Tail
ImVoxelNet* [29]	19.0	34.1	14.0	7.7
ImGeoNet* [31]	22.3	38.1	17.3	9.7
SGCDet-L (Ours)	**28.9**	**46.0**	**24.0**	**14.9**

Table 3. Quantitative results on the ARKitScenes dataset. * denotes the results are directly cited from [30, 40].

Method	Voxel Resolution	mAP@0.25	mAP@0.50
With ground-truth geometry supervision.			
ImGeoNet* [31]	40×40×16	60.2	43.4
CN-RMA* [30]	192×192×80	67.6	56.5
Without ground-truth geometry supervision.			
ImVoxelNet* [29]	40×40×16	27.3	4.3
NeRF-Det* [38]	40×40×16	39.5	21.9
MVSDet* [40]	40×40×16	42.9	27.0
ImVoxelNet [29]	40×40×16	58.0	33.2
NeRF-Det [38]	40×40×16	60.4	38.3
MVSDet [40]	40×40×16	60.7	40.1
SGCDet (Ours)	40×40×16	62.3	44.7
SGCDet-L (Ours)	80×80×32	**70.4**	**57.0**

Experiments

Qualitative results



Conclusions

- A novel framework named **SGCDet** for multi-view indoor 3D object detection.
- **Geometry and context aware aggregation** for multi-view feature lifting.
- **Sparse volume construction** for adaptive refinement.
- Using **only 3D bounding boxes** for supervision.
- **SOTA** performance on ScanNet, ScanNet200, and ARKitScenes.

浙江大學
ZHEJIANG UNIVERSITY

Thanks for watching!