VisHall3D: Monocular Semantic Scene Completion from Reconstructing the Visible Regions to Hallucinating the Invisible Regions

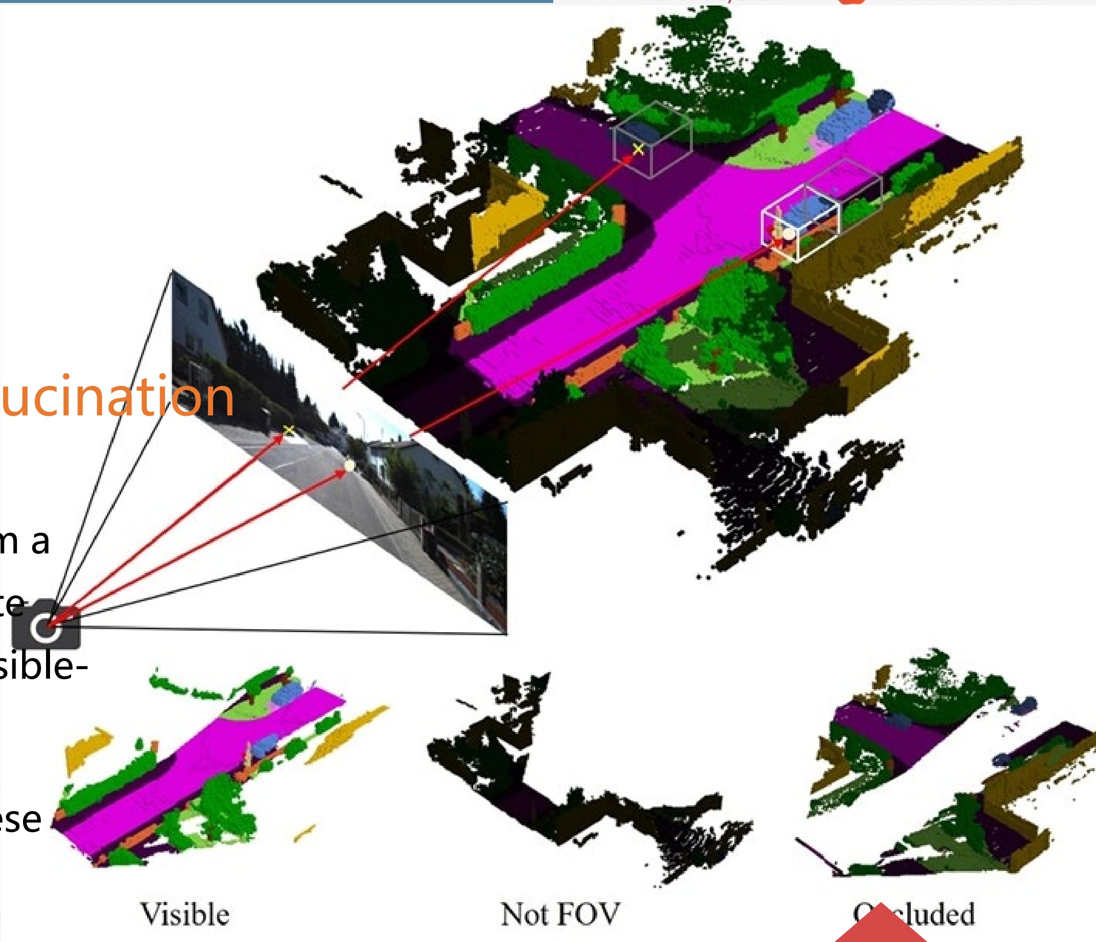
Haoang Lu, Yuanqi Su*, Xiaoning Zhang, Longjun Gao, YuXue, LeWang

Xi'an Jiaotong University, ＊Corresponding Author

Contents Title

VisHall3D

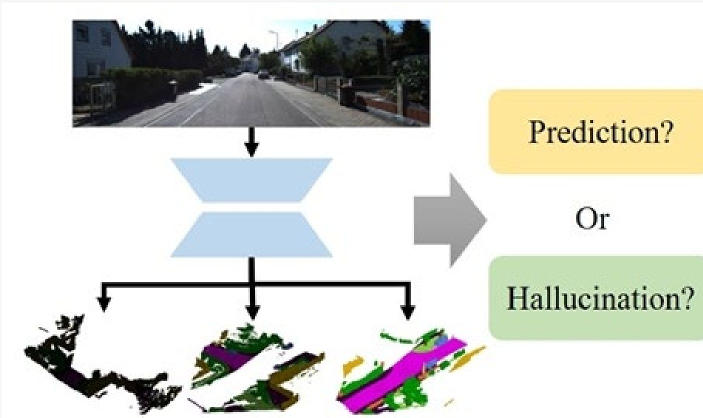
Monocular SSC: Vision vs. Hallucination

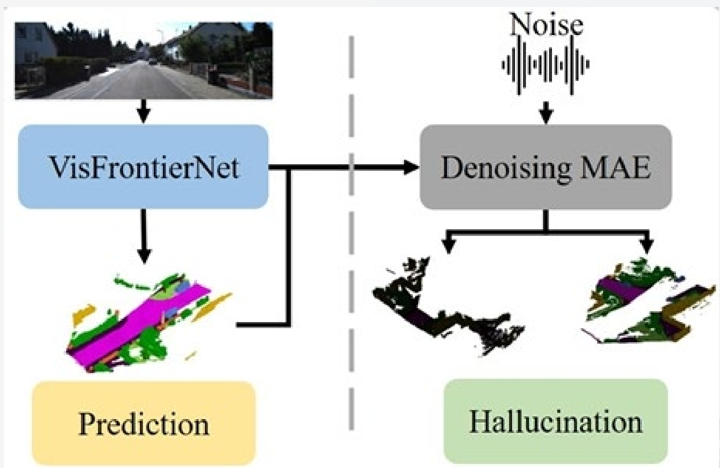


Monocular SSC reconstructs 3D scenes from a single RGB image, demanding both accurate visible-surface modeling and plausible invisible-region inference.

Existing single-stage methods entangle these tasks, causing feature entanglement and geometric inconsistency.

Motivation





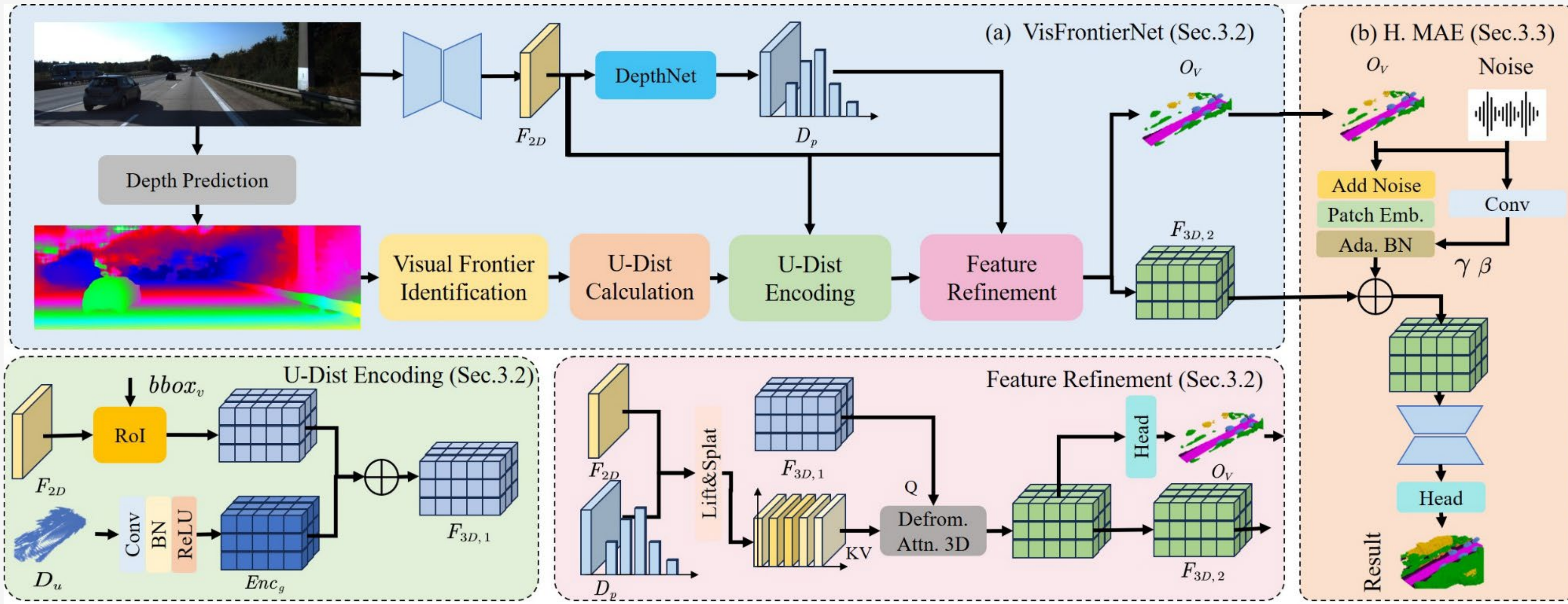
Our Solution:

Decoupling

We propose a two-stage framework: VisFrontierNet for visible regions, followed by OcclusionMAE for invisible ones.

Pipeline

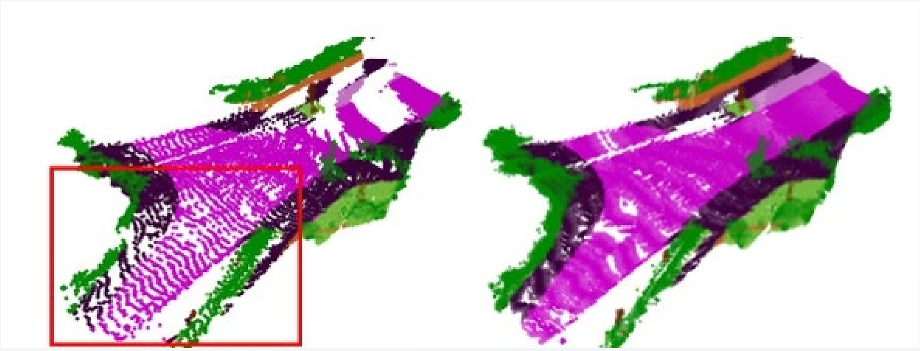
ICCV HONOLULU HAWAII
OCT 19-23, 2025



We propose **VisHall3D**, a two-stage monocular SSC framework that decouples vision and hallucination to reduce feature entanglement and geometric inconsistency.

- **VisFrontierNet**, a visibility-aware projection module that accurately traces the visual frontier by modeling the boundary between visible and invisible regions.
- **OcclusionMAE**, a hallucination network that generates plausible geometries for invisible regions using a noise injection mechanism.

Pipeline

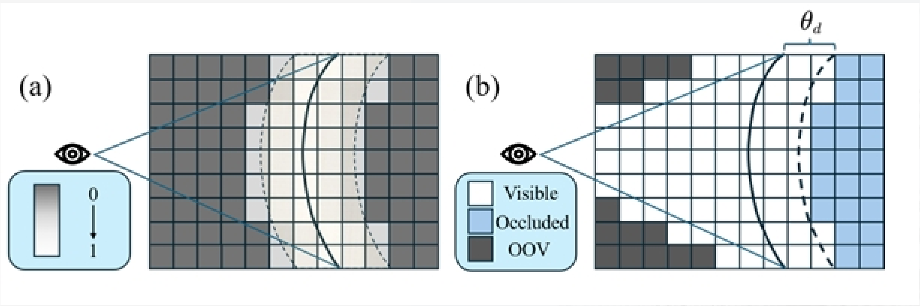


Hard Lifting

Soft Lifting

- Problem: Far distant depth prediction exhibits significant errors, and hard-lifting methods lead to inconsistencies between near and far feature

We propose **Visual Frontier (Soft Lifting)**, a method that models uncertainty in depth prediction while preserving the network' s ability to hallucinate missing regions.



Visual Frontier

Experiment

Method	Date	IoU↑	mIoU↑	road	sidewalk	parking	other-grnd.	building	car	truck	bicycle	motorcycle	other-veh.	vegetation	trunk	terrain	person	bicyclist	motorcyclist	fence	pole	traf.-sign
Stereo camera-based methods																						
StereoScene[14]	IJCAI2024	43.34	15.36	61.90	31.20	30.70	10.70	24.20	22.80	2.80	3.40	2.40	6.10	23.80	8.40	27.00	2.90	2.20	0.50	16.50	7.00	7.20
Monocular temporal methods																						
VoxFormer-T[20]	CVPR2023	43.21	13.41	54.10	26.90	25.10	7.30	23.50	21.70	3.60	1.90	1.60	4.10	24.40	8.10	24.20	1.60	1.10	0.00	13.10	6.60	5.70
HASSC-T[36]	CVPR2024	42.87	14.38	55.30	29.60	25.90	11.30	23.10	23.00	2.90	1.90	1.50	4.90	24.80	9.80	26.50	1.40	3.00	0.00	14.30	7.00	7.10
HTCL[15]	ECCV2024	44.23	17.09	64.40	34.80	33.80	12.40	25.90	27.30	5.70	1.80	2.20	5.40	25.30	10.80	31.20	1.10	3.10	0.90	21.10	9.00	8.30
H2GFormer-T[37]	AAAI2024	43.52	14.60	57.90	30.40	30.00	6.90	24.00	23.70	5.20	0.60	1.20	5.00	25.20	10.70	25.80	1.10	0.10	0.00	14.60	7.50	9.30
Monocular single-frame methods																						
MonoScene[3]	CVPR2023	34.16	11.08	54.70	27.10	24.80	5.70	14.40	18.80	3.30	0.50	0.70	4.40	14.90	2.40	19.50	1.00	1.40	0.40	11.10	3.30	2.10
VoxFormer-S[20]	CVPR2023	42.95	12.20	53.90	25.30	21.10	5.60	19.80	20.80	3.50	2.60	0.70	3.70	22.40	7.50	21.30	1.40	2.60	0.20	11.10	5.10	4.90
TPVFormer[11]	CVPR2023	34.25	11.26	55.10	27.20	27.40	6.50	14.80	19.20	3.70	1.00	0.50	2.30	13.90	2.60	20.40	1.10	2.40	0.30	11.00	2.90	1.50
SurroundOcc[38]	ICCV2023	34.72	11.86	56.90	28.30	30.20	6.80	15.20	20.60	1.40	1.60	1.20	4.40	14.90	3.40	19.30	1.40	2.00	0.10	11.30	3.90	2.40
OccFormer[46]	ICCV2023	34.53	12.32	55.90	30.30	31.50	6.50	15.70	21.60	1.20	1.50	1.70	3.20	16.80	3.90	21.30	2.20	1.10	0.20	11.90	3.80	3.70
IAMSSC[39]	T-ITS2024	43.74	12.37	54.00	25.50	24.70	6.90	19.20	21.30	3.80	1.10	0.60	3.90	22.70	5.80	19.40	1.50	2.90	0.50	11.90	5.30	4.10
DepthSSC[41]	arXiV2024	44.58	13.11	55.64	27.25	25.72	5.78	20.46	21.94	3.74	1.35	0.98	4.17	23.37	7.64	21.56	1.34	2.79	0.28	12.94	5.87	6.23
HASSC-S[36]	CVPR2024	43.40	13.34	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Symphonize[12]	CVPR2024	42.19	15.04	58.40	29.30	26.90	11.70	24.70	23.60	3.20	3.60	2.60	5.60	24.20	10.00	23.10	**3.20**	1.90	**2.00**	16.10	7.70	8.00
H2GFormer-S[37]	AAAI2024	44.20	13.72	56.40	28.60	26.50	4.90	22.80	23.40	4.80	0.80	0.90	4.10	24.60	9.10	23.80	1.20	2.50	0.10	13.30	6.40	6.30
MonoOcc-L[47]	ICRA2024	-	15.63	59.10	30.90	27.10	9.80	22.90	23.90	7.20	**4.50**	2.40	**7.70**	25.00	9.80	26.10	2.80	4.70	0.60	16.90	7.30	8.40
CGFormer[43]	NIPS2024	44.41	16.63	64.30	**34.20**	**34.10**	12.10	25.80	26.10	4.30	3.70	1.30	2.70	24.50	11.20	**29.30**	1.70	3.60	0.40	18.70	8.70	**9.30**
Ours	ICCV2025	**46.50**	**17.46**	**64.60**	34.10	32.00	**12.50**	**26.90**	**26.70**	**7.50**	2.90	**3.30**	6.20	**27.30**	**12.50**	28.00	2.30	**5.10**	1.90	**19.50**	**9.20**	9.20

Semantic-KITTI Comparison

- 3,834 train samples
- 815 validation samples
- 3,992 test samples
- 20 valid classes
- 1 invalid class

Experiment

ICCV HONOLULU HAWAII OCT 19-23, 2025

Method	Date	IoU↑	mIoU↑	car	bicycle	motorcycle	truck	other-veh.	person	road	parking	sidewalk	other-grnd.	building	fence	vegetation	terrain	pole	traf.-sign	other-struct.	other-obj.
LiDAR-based methods																					
SSCNet [34]	CVPR2017	53.58	16.95	31.95	0.00	0.17	10.29	0.00	0.07	65.70	17.33	41.24	3.22	44.41	6.77	43.72	28.87	0.78	0.75	8.69	0.67
LMSCNet [32]	3DV 2020	47.35	13.65	20.91	0.00	0.00	0.26	0.58	0.00	62.95	13.51	33.51	0.20	43.67	0.33	40.01	26.80	0.00	0.00	3.63	0.00
Monocular camera-based methods																					
MonoScene [3]	CVPR2023	37.87	12.31	19.34	0.43	0.58	8.02	2.03	0.86	48.35	11.38	28.13	3.32	32.89	3.53	26.15	16.75	6.92	5.67	4.20	3.09
TPVFormer [11]	CVPR2023	40.22	13.64	21.56	1.09	1.37	8.06	2.57	2.38	52.99	11.99	31.07	3.78	34.83	4.80	30.08	17.52	7.46	5.86	5.48	2.70
OccFormer [46]	ICCV2023	40.27	13.81	22.58	0.66	0.26	9.89	3.82	2.77	54.30	13.44	31.53	3.55	36.42	4.80	31.00	19.51	7.77	8.51	6.95	4.60
VoxFormer [20]	CVPR2023	38.76	11.91	17.84	1.16	0.89	4.56	2.06	1.63	47.01	9.67	27.21	2.89	31.38	4.97	28.99	14.69	6.51	6.92	3.79	2.43
IAMSSC [39]	T-ITS2024	41.80	12.97	18.53	2.45	1.76	5.12	3.92	3.09	47.55	10.56	28.35	4.12	31.53	6.28	29.17	15.24	8.29	7.01	6.35	4.19
DepthSSC [41]	arXiV2024	40.85	14.28	21.90	2.36	4.30	11.51	4.56	2.92	50.88	12.89	30.27	2.49	37.33	5.22	29.61	21.59	5.97	7.71	5.24	3.51
Symphonies [12]	CVPR2024	44.12	18.58	30.02	1.85	5.90	**25.07**	**12.06**	8.20	54.94	13.83	32.76	**6.93**	35.11	8.58	38.33	11.52	14.01	9.57	**14.44**	**11.28**
CGFormer[43]	NIPS2024	48.07	20.05	29.85	**3.42**	3.96	17.59	6.70	6.63	63.85	17.15	40.72	5.53	42.73	8.22	38.80	24.04	16.24	17.45	10.18	6.77
Ours	ICCV2025	**49.12**	**20.95**	**30.77**	1.91	**6.60**	17.99	8.72	**8.67**	**64.35**	**18.83**	**41.53**	4.48	**43.87**	**9.07**	**39.75**	**24.94**	**16.52**	**20.66**	10.30	7.99

KITTI360-SSCBench Comparison

- 8487 train samples
- 1812 validation samples
- 2566 test samples
- 19 valid classes

Experiment

Method	IoU↑	mIoU↑	Params (M)	Memory (M)
Baseline	42.11	14.56	57.2	15260
VisFrontierNet w/o Feature Refinement	**46.44** (+4.33)	15.11 (+0.55)	74.3	17349
+ OcclusionMAE w/o Denoising	46.38 (-0.06)	15.99 (+0.88)	86.7	18746
+ Feature Refinement	45.88 (-0.50)	16.59 (+0.60)	125.5	21246
+ Denoising	46.14 (+0.26)	**17.06** (+0.47)	127.8	22597

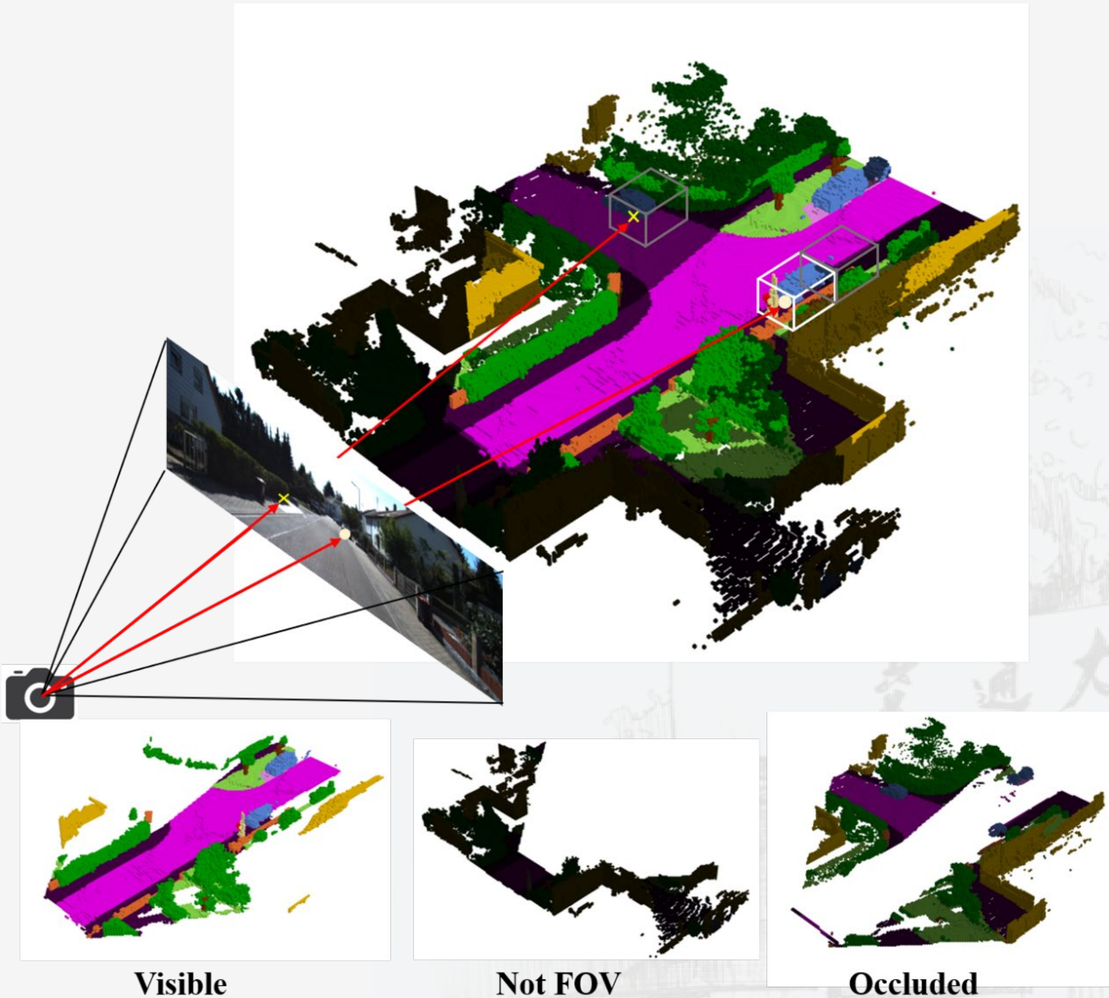
- Ablation on architectural components

R_h (Voxel)	R_d (Voxel)	IoU↑	mIoU↑
0	0	46.03	16.32
0	2	45.92	16.79
0	3	45.83	**16.92**
0	4	45.73	16.66
1	3	**46.19**	16.52
1	4	46.12	16.69

Invisibility	Threshold θ_d (m)	IoU↑	mIoU↑
OOV	-	43.98	15.89
OOV + Occ.	1.5	45.87	16.65
OOV + Occ.	2.5	45.83	16.92
OOV + Occ.	3.5	**46.14**	**17.06**
OOV + Occ.	4.5	45.94	16.95

- Ablation on Visual Frontier
- Ablation on MAE noise

Conclusion



In this paper, we proposed VisHall3D. VisHall3D sets a new standard for Monocular SSC, paving the way for more accurate and reliable scene understanding in various applications

西安交通大學
XI'AN JIAOTONG UNIVERSITY

Thank You