Learning Spatial-Temporal Consistency for 3D Semantic Scene Completion

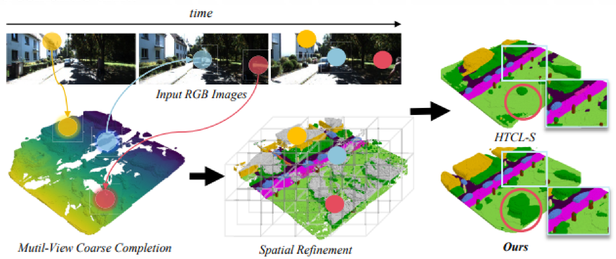
Yujie Xue, Meng Wang, Ruihui Li, Fan Wu, Zhizhong Liu, Zhuo Tang, Kenli Li
Hunan University

CVPR JUNE 3-7, 2026 DENVER COLORADO

Introduction

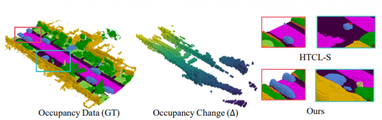
• **Background:** Camera-based Semantic Scene Completion (SSC) is to in fer the full geometry of objects and scenes from only 2D images, which i s cost-effective but suffers from occlusion and perspective ambiguity.
• **Limitation:** Lack of explicit spatial-temporal consistency leads to geom etric fragmentation and semantic drift.
• **Research Question:** How to leverage temporal cues to generate comple te and consistent 3D scenes?

Limitations of Existing Methods:
Single-frame SSC: Insufficient information, rendering the recovery of oc cluded regions impossible.
Temporal SSC: Merely stacks multi-frame features without explicit spati otemporal consistency constraints, leading to geometric discontinuities an d semantic drift.

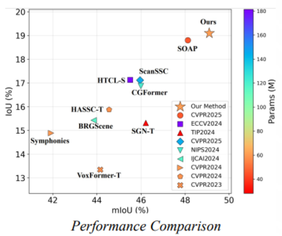


Motivation:

✓ Incomplete scene occupancy leads to a lack of semantic context, resulting in semantic degradation.
✓ By synergistically leveraging historical 2D and 3D cues, the problem of spatiotemporal inconsistency is effectively resolved.



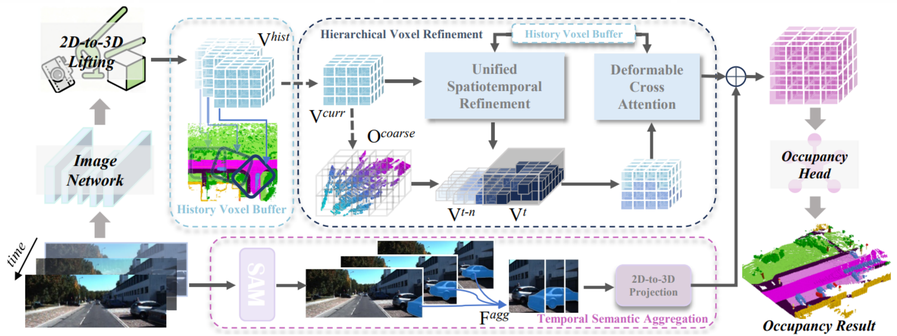
Comparing ground truth occupancy with coa rse occupancy changes. The change map illu strates the differences before and after coars e occupancy generation.



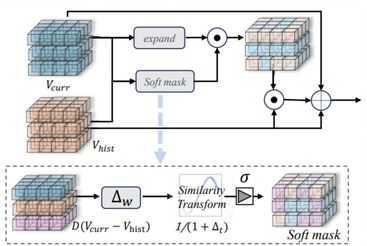
Performance Comparison

Method

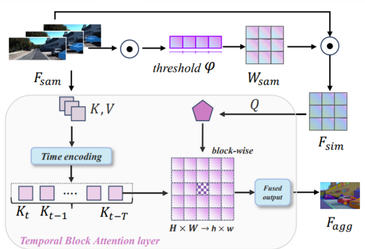
Pipeline:



Module:



Unified Spatiotemporal Refinement



Temporal Semantic Aggregation

Contribution

We summarize our Contribution as follows:

◆ We introduce a Spatial-Temporal Consistency camera-based SSC framework, designed to leverage both **2D and 3D historical information** to update 3D features across space and time**.**
◆ Hierarchical Voxel Refinement, which uses historical 2D features to generate more **coarse occupancy** and leverages historical 3D features to progressively refine the missing structures. Temporal Semantic Aggregation, which integrates **visibility-aware multi-view semantic cues** to correct ambiguous predictions in occluded regions**..**
◆ Our method achieves state-of-the-art performance among all camera-based SSC methods on the SemanticKITTI and KITTI-360 benchmarks.

Experiments

Semantic KITTI:

Methods	Pub.	Input	IoU	road (15.30%)	sidewalk (11.13%)	parking (1.12%)	other-grnd (0.56%)	building (14.10%)	car (3.92%)	truck (0.16%)	bicycle (0.03%)	motocycle (0.03%)	other-vehicle (0.20%)	vegetation (39.3%)	trunk (0.51%)	terrain (9.17%)	person (0.07%)	bicyclist (0.07%)	motorcyclist (0.05%)	fence (3.90%)	pole (0.29%)	traf.-sign (0.08%)	mIoU
MonoScene †	CVPR22	S	34.16	54.7	27.1	24.8	5.7	14.4	18.8	3.3	0.5	0.7	4.4	14.9	2.4	19.5	1.0	1.4	0.4	11.1	3.3	2.1	11.08
TPVFormer†	CVPR23	S	34.25	55.1	27.2	27.4	6.5	14.8	19.2	3.7	1.0	0.5	2.3	13.9	2.6	20.4	1.1	2.4	0.3	11.0	2.9	1.5	11.26
OccFormer†	CVPR23	S	34.53	55.9	30.3	31.5	6.5	15.7	21.6	1.2	1.5	1.7	3.2	16.8	3.9	21.3	2.2	1.1	0.2	11.9	3.8	3.7	12.32
VoxFormer-T	CVPR23	T	43.21	54.1	26.9	25.1	7.3	23.5	21.7	3.6	1.9	1.6	4.1	24.4	8.1	24.2	1.6	1.1	0.0	13.1	6.6	5.7	13.41
HASSC-T	CVPR24	T	42.87	55.3	29.6	25.9	11.3	23.1	23.0	2.9	1.9	1.5	4.9	24.8	9.8	26.5	1.4	3.3	0.0	14.3	7.0	7.1	14.38
H2GFormer	AAAI24	T	43.52	57.9	30.4	30.0	6.9	24.0	23.7	5.2	0.6	1.2	5.0	25.2	10.7	25.8	1.1	0.1	0.0	14.6	7.5	7.1	14.60
Symphonies	CVPR24	S	42.19	58.4	29.3	26.9	11.7	24.7	23.6	3.2	3.6	2.6	5.6	24.2	10.0	23.1	**3.2**	1.9	**2.0**	16.1	7.7	8.0	15.04
BRGScene †	IJCAI24	S	43.34	61.9	31.2	30.7	10.7	24.2	22.8	2.8	3.4	2.4	6.1	23.8	8.4	27.0	2.9	2.2	0.5	16.5	7.0	7.2	15.36
Bi-SSC†	CVPR24	S	45.10	63.4	33.3	31.7	11.2	26.6	25.0	6.8	1.8	2.9	**6.8**	26.1	10.5	28.9	1.7	3.3	1.0	19.4	9.3	8.4	16.73
CGFormer †	NIPS24	S	44.41	64.3	34.2	34.1	12.1	25.8	26.1	4.3	3.7	1.3	2.7	24.5	11.2	29.3	1.7	3.6	0.4	18.7	8.7	9.3	16.63
ScanSSC†	CVPR25	S	44.54	**66.2**	35.9	35.1	12.5	25.3	27.1	3.5	3.5	**3.2**	6.1	25.2	11.0	30.6	1.8	**5.3**	0.7	20.5	8.4	8.9	17.40
SGN	TIP24	T	43.71	57.9	29.7	25.6	5.5	27.0	25.0	1.5	0.9	0.7	3.6	26.9	12.0	26.4	0.6	0.3	0.0	14.7	9.0	6.4	14.49
HTCL-S †	ECCV24	T	44.23	64.4	34.8	33.8	12.4	25.9	27.3	**10.8**	1.8	2.2	5.4	25.3	10.8	31.2	1.1	3.1	0.9	21.1	9.0	8.3	17.09
SOAP	CVPR25	T	47.54	63.2	36.6	35.8	**16.3**	30.3	28.4	4.3	4.6	2.9	4.9	31.6	15.0	33.2	2.3	0.9	0.1	20.6	11.5	13.3	18.72
Ours		T	**48.17**	64.5	**37.9**	**37.0**	15.5	**31.8**	**29.5**	0.2	**5.7**	2.7	6.6	**31.7**	**15.4**	**33.9**	2.0	0.4	0.0	**22.5**	**12.5**	**13.3**	**19.20**

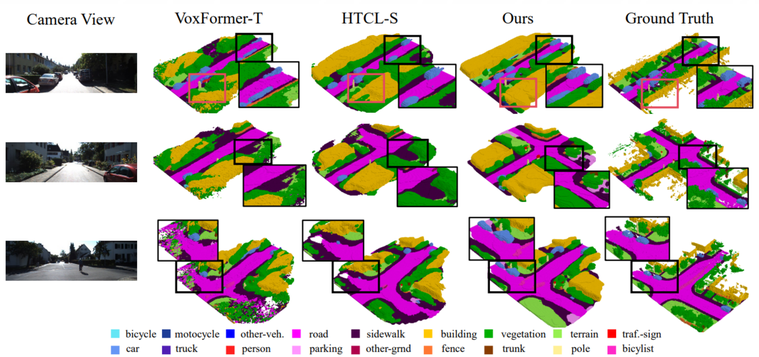
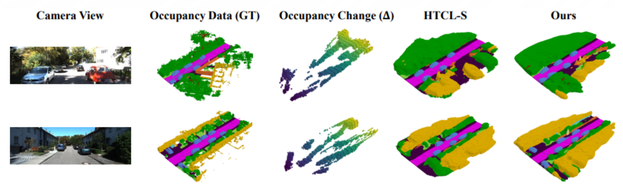


Figure 6. The Visualization results of our ConSSC and the state-of-the-art methods, include VoxFormer-T and HTCL-S.



Comparison between the ground truth occupancy and coarse occupancy changes