Towards Temporal Fusion Beyond the Field of View for Camera-based Semantic Scene Completion

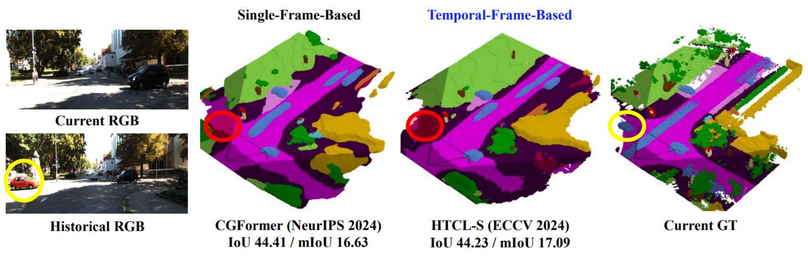
Jongseong Bae, Junwoo Ha, Jinnyeong Heo, Yeongin Lee, and Ha Young Kim

{ js.bae,gkwnsdn0402,wlssud132,zeroin,hayoung.kim }@yonsei.ac.kr

YONSEI UNIVERSITY

MLCF Multimodal Learning & Computational Finance Lab

Motivation



- Temporal fusion is widely used in camera-based 3D SSC, but improvements are mostly **limited to in-camera-view regions.**
- **Out-of-camera-view blind spots** (safety-critical near the ego vehicle) remain poorly completed, even though **past frames may contain the missing context.**
- This motivates **out-of-view completion via temporal cues, by fusing historical and current features in the current frame's 3D metric space.**

Contribution

- **Frist focus on out-of-camera view completion (camera-based SSC):** We are the first (to our knowledge) to primarily address out-of-camera-view completion using temporal cues in camera-based SSC.
- **C3DFusion:** We propose a simple yet effective temporal geometry fusion method, C3DFusion, that performs perspective alignment between explicitly point-mapped historical and current frame features in the 3D metric space of the current frame.
- **2 refinement techniques for robust fusion: historical context blurring** and **current-centric feature densification** that reduce noise caused by geometric inaccuracies in warped historical features and improve geometric fidelity by emphasizing information from the current frame.
- Built on C3DFusion, our model achieves strong SOTA performance on the SemanticKITTI and SSCBench-KITTI-360 benchmarks, while demonstrating robust generalization across diverse existing architectures.

Ablation Studies

✓ **Ablation study of C3DFusion.**

'TPFA', 'HCB', and 'CCFD' denote temporal 3D point feature alignment, historical context blurring, and current-centric feature densification.

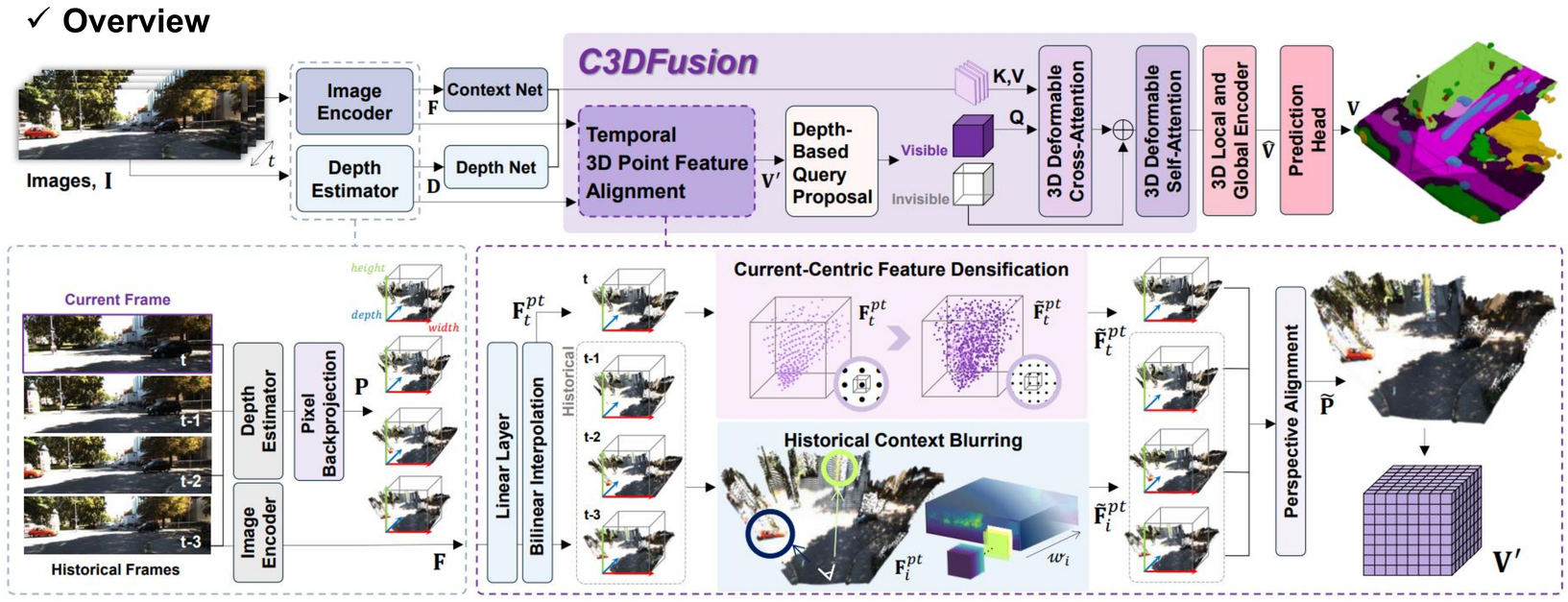
	TPFA	HCB	CCFD	IoU	mIoU
Baseline				48.59	16.58
(a)	✓			49.09	18.45
(b)	✓	✓		48.99	18.88
(c)	✓		✓	48.87	18.86
Ours	✓	✓	✓	**49.53**	**19.31**

✓ **Generalization of C3DFusion**
across other leading camera-based SSC models.

Method	IoU	mIoU
VoxFormer-S	44.02	12.35
+ *Temporal LSS fusion*	44.39 (+0.37)	9.56 (-2.79)
+ ***C3DFusion***	**45.98 (+1.96)**	**15.12 (+2.77)**
OccFormer	36.50	13.46
+ *Temporal LSS fusion*	44.48 (+7.98)	16.91 (+3.45)
+ ***C3DFusion***	**44.83 (+8.36)**	**18.29 (+4.83)**
ScanSSC	45.95	17.12
+ *Temporal LSS fusion*	49.31 (+3.36)	17.67 (+0.55)
+ ***C3DFusion***	**49.89 (+3.94)**	**18.73 (+1.61)**
CGFormer	45.99	16.87
+ *Temporal LSS fusion*	48.59 (+2.60)	16.58 (-0.29)
+ ***C3DFusion (Ours)***	**49.53 (+3.54)**	**19.31 (+2.44)**

Methodology

✓ **Overview**



Experiments

✓ **Quantitative results on SemanticKITTI hidden test set.**

Method	Input	IoU	mIoU	road	sidewalk	parking	other-grnd.	building	car	truck	bicycle	motorcycle	other-veh.	vegetation	trunk	terrain	person	bicyclist	motorcyclist	fence	pole	traf.-sign	OOV (val.) IoU	OOV (val.) mIoU
Single-Frame-Based																								
MonoScene	Mono	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	31.07	7.02
TPVFormer	Mono	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	30.76	7.87
OccFormer	Mono	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	30.46	8.68
Symphonies	Stereo	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	23.48	6.40
CGFormer	Stereo	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	33.54	9.06
L2COcc-C	Stereo	44.31	17.03	66.00	35.00	33.10	13.50	25.10	27.20	3.00	3.50	3.60	4.30	25.20	11.50	30.10	1.50	2.40	0.20	20.50	9.10	8.90	32.24	8.55
ScanSSC	Stereo	44.54	17.40	66.20	35.90	35.10	12.50	25.30	27.10	3.50	3.50	3.20	6.10	25.20	11.00	30.60	1.80	**5.30**	0.70	20.50	8.40	8.90	33.60	9.50
L2COcc-D	Stereo	45.37	18.18	**68.20**	**36.90**	34.60	16.20	25.80	28.30	4.50	4.90	3.30	**7.20**	26.20	11.90	32.00	2.10	2.40	0.30	21.60	9.60	9.50	31.85	10.05
Temporal-Frame-Based																								
VoxFormer-T	Stereo	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	40.21	11.58
HTCL-S	Stereo	44.23	17.09	64.40	34.80	33.80	12.40	25.90	27.30	**10.80**	1.80	2.20	5.40	25.30	10.80	31.20	1.10	3.10	0.90	21.10	9.00	8.30	33.14	9.04
Hi-SOP	Stereo	44.57	17.49	63.95	34.27	**35.85**	13.77	25.91	27.35	7.18	2.99	2.59	7.19	26.07	10.35	30.77	1.68	4.81	1.06	20.15	8.70	7.90	-	-
FlowScene	Stereo	45.20	17.70	64.10	35.00	33.70	13.00	27.70	26.40	10.00	4.20	3.10	7.00	26.30	10.00	30.20	3.10	5.10	1.10	20.20	8.90	9.10	-	-
CF-SSC	Stereo	46.21	16.40	61.30	33.30	29.20	11.90	30.40	26.30	4.80	2.60	2.70	6.30	28.50	11.40	28.30	1.50	1.40	0.40	17.70	7.20	6.30	-	-
Ours	**Stereo**	**47.62**	**18.98**	67.00	36.30	33.20	**19.30**	**30.60**	**29.00**	3.30	**5.40**	**4.40**	4.70	**29.60**	**14.70**	**33.80**	1.60	2.80	0.30	**22.80**	**11.40**	**10.40**	**44.37**	**17.17**

✓ **Quantitative results on SSCBench-KITTI-360 test set.**

Method	Input	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.	OOV (test) IoU	OOV (test) mIoU
Single-Frame-Based																							
MonoScene	Mono	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	Mono	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	Mono	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	-	-
Symphonies	Stereo	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**	34.39	11.93
CGFormer	Stereo	48.07	20.05	29.85	3.42	3.96	17.59	6.79	6.63	**63.85**	17.15	40.72	5.53	42.73	8.22	38.80	24.94	16.24	17.45	10.18	6.77	44.72	15.61
ScanSSC	Stereo	48.29	20.14	29.91	3.78	4.28	14.34	9.08	6.65	62.21	18.16	40.19	5.16	42.68	8.83	38.84	25.50	16.60	19.14	10.30	6.89	45.09	15.44
Temporal-Frame-Based																							
FlowScene	Stereo	46.98	19.12	29.83	4.44	3.78	16.71	8.71	7.77	60.70	16.99	39.59	6.01	43.17	9.45	37.32	25.14	17.35	18.12	10.63	7.56	-	-
CF-SSC	Stereo	45.79	19.10	28.10	3.39	6.87	16.76	7.75	5.68	59.01	16.80	37.60	4.95	42.16	8.26	36.14	21.89	14.73	17.72	9.73	7.14	-	-
Ours	**Stereo**	**49.28**	**21.74**	**31.16**	**5.39**	**7.01**	18.12	8.25	5.66	63.70	**19.12**	**41.64**	5.09	**43.93**	**10.43**	**40.73**	**27.62**	**19.30**	**23.08**	12.34	8.74	**52.41**	**23.72**

✓ **Qualitative Results**