IPFormer: Visual 3D Panoptic Scene Completion with Context-Adaptive Instance Proposals

Markus Gross[1,2] Aya Fahmy[1] Danit Niwattananan[2] Dominik Muhle[2] Rui Song[1,2] Daniel Cremers[2] Henri Meeß[1]

[1]Fraunhofer Institute IVI [2]Technical University of Munich

Project Page

Task Description

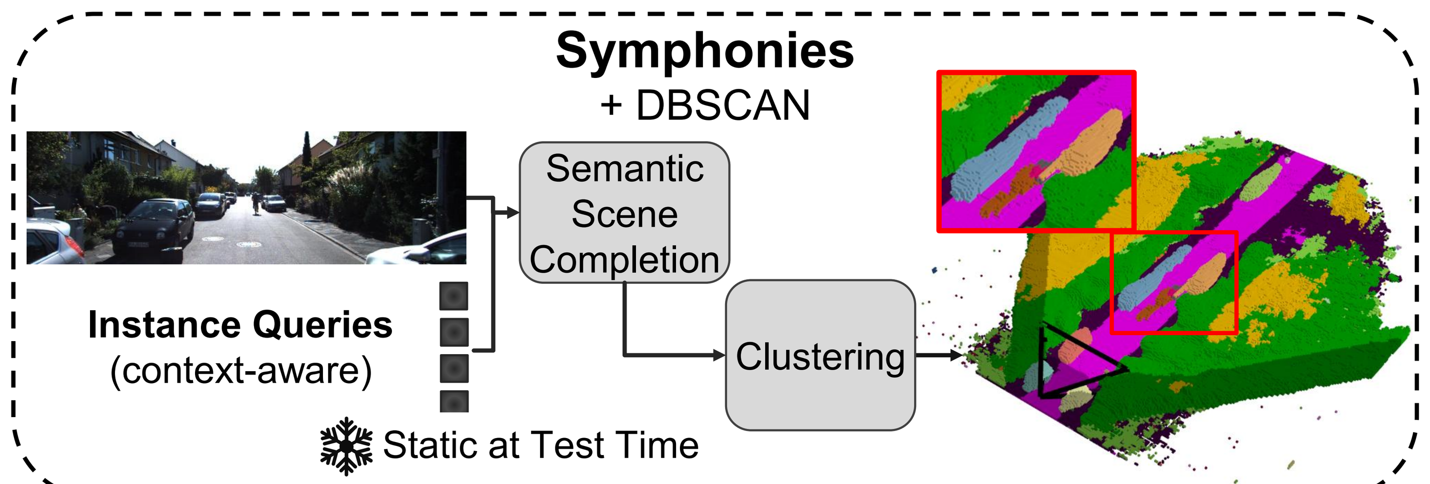
Using camera images, infer the complete 3D structure of a scene as a voxel grid, including both visible and occluded regions. Every voxel carries

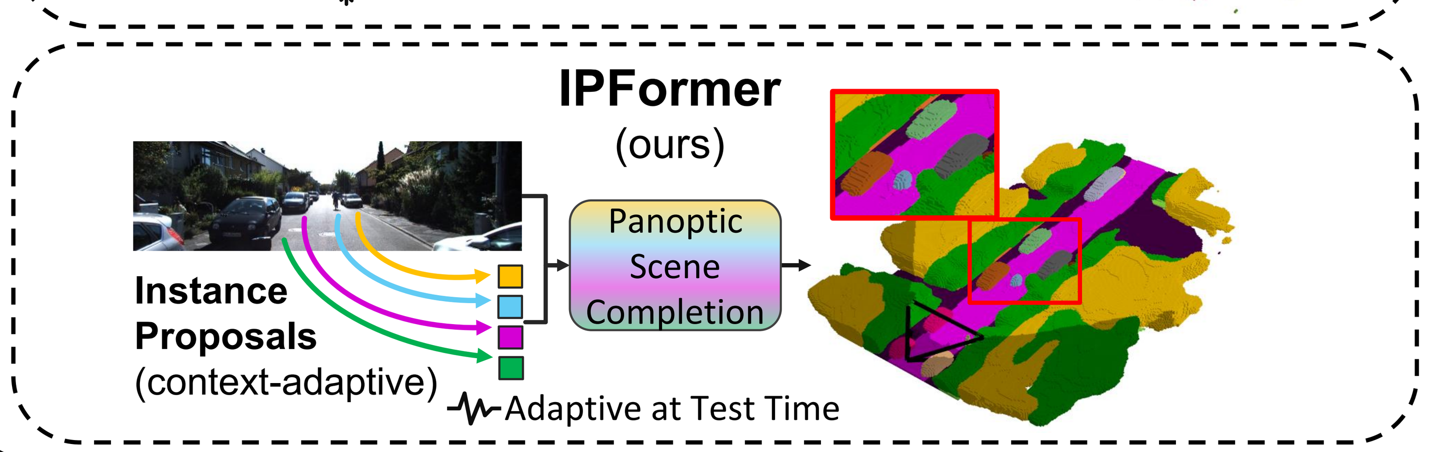
1. binary **occupancy**
2. a **semantic** label
3. an **instance** ID to group countable objects

Challenges

Previous methods

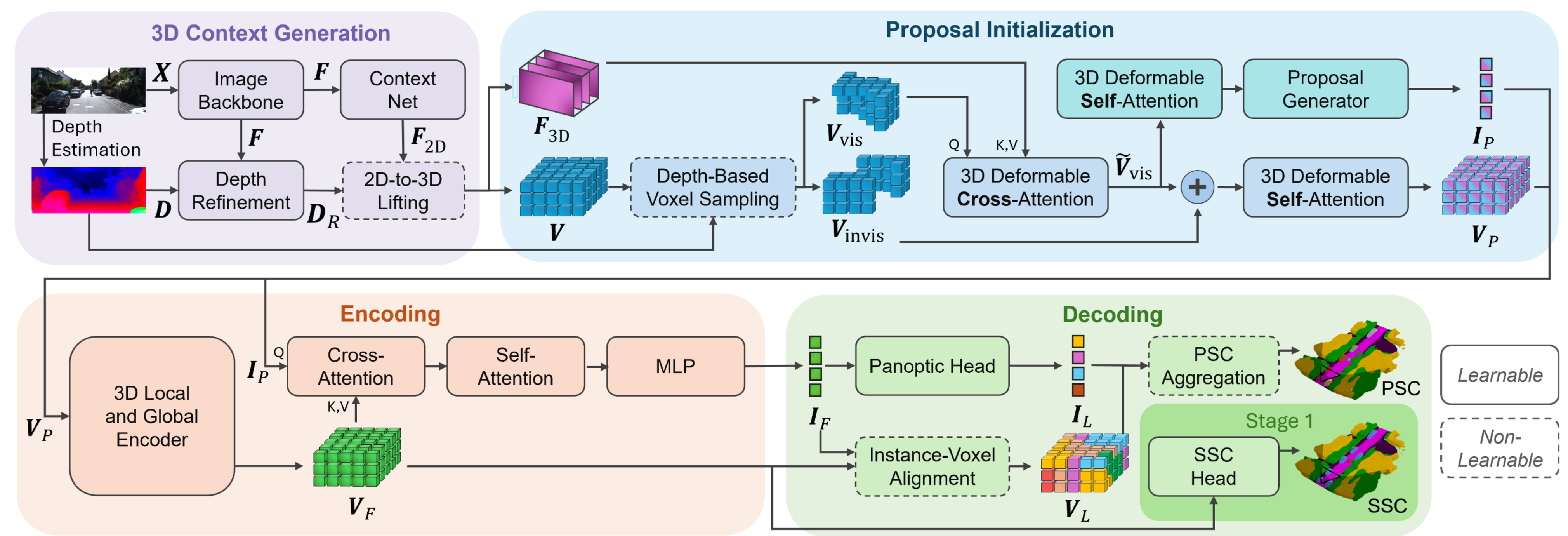
1. only infer **occupancy** and **semantics** in an end-to-end fashion (Semantic Scene Completion), but require subsequent, time-consuming Euclidean clustering to retrieve individual **instances**.
2. reconstruct objects using a fixed set of learned queries that are updated with image context during training, but remain static at test time and thus fail to dynamically adapt specifically to the observed scene.





Our Approach

Our method (1) addresses Panoptic Scene Completion in an end-to-end fashion, and (2) initializes object queries as instance proposals that dynamically adapt specifically to the observed scene at train and test time.



Specifically, we

1. propose a dual-head architecture and a two-stage training scheme that effectively guides the latent space toward **occupancy** and **semantics** before **instance** registration.
2. introduce a visibility-based sampling strategy, which utilizes visible voxels and respective image context to adaptively initialize instance proposals.

Quantitative Results

In-Domain Performance:

Method	PSC Metrics: All PQ†↑	PQ↑	SQ↑	RQ↑	Thing PQ↑	SQ↑	RQ↑	Stuff PQ↑	SQ↑	RQ↑	SSC Metrics: IoU↑	mIoU↑	Runtime [s] ↓
MonoScene [4] + DBSCAN	10.12	3.43	15.15	5.33	0.51	7.36	0.87	5.56	20.81	8.57	36.80	11.31	4.51
Symphonies [21] + DBSCAN	11.69	3.75	26.09	5.95	1.07	27.65	1.76	5.70	24.95	8.99	41.92	15.02	4.54
OccFormer [63] + DBSCAN	11.25	4.32	24.19	6.69	0.68	21.47	1.15	6.96	26.16	10.73	36.43	13.51	4.70
CGFormer [59] + DBSCAN	14.39	6.16	**48.14**	9.48	**2.20**	**44.46**	**3.47**	9.03	**50.82**	13.86	**45.98**	**16.89**	4.70
IPFormer (ours)	**14.45**	**6.30**	41.95	**9.75**	2.09	42.67	3.33	**9.35**	41.43	**14.43**	40.90	15.33	**0.33**

Out-of-Domain Zero-Shot Generalization Performance:

	PSC Metrics: All PQ†↑	PQ↑	SQ↑	RQ↑	Thing PQ↑	SQ↑	RQ↑	Stuff PQ↑	SQ↑	RQ↑	SSC Metrics: IoU↑	mIoU↑
SemanticKITTI												
CGFormer [59] + DBSCAN	14.39	6.16	**48.14**	9.48	**2.20**	**44.46**	**3.47**	9.03	**50.82**	13.86	**45.98**	**16.89**
IPFormer (ours)	**14.45**	**6.30**	41.95	**9.75**	2.09	42.67	3.33	**9.35**	41.43	**14.43**	40.90	15.33
KITTI-360												
CGFormer [59] + DBSCAN	8.44	1.08	17.82	1.87	**0.53**	20.06	**0.96**	1.48	16.19	2.54	28.11	9.44
IPFormer (ours)	**9.41**	**1.23**	**24.68**	**2.16**	0.52	**22.76**	0.95	**1.68**	**25.89**	**2.93**	**28.74**	**9.53**
Relative Gap ↓												
CGFormer [59] + DBSCAN	41.37%	82.47%	62.98%	80.28%	75.91%	54.89%	72.34%	83.61%	68.15%	81.67%	38.88%	44.09%
IPFormer (ours)	**34.88%**	**80.48%**	**41.19%**	**77.85%**	**75.12%**	**46.64%**	**71.53%**	**82.03%**	**37.52%**	**79.69%**	**29.73%**	**37.81%**

Qualitative Results