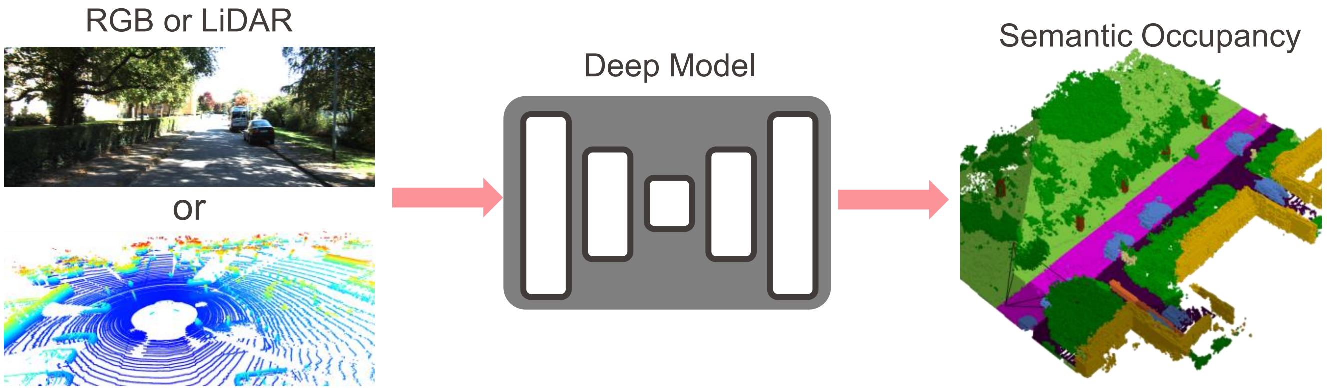
EPFL VITA

VoxDet: Rethinking 3D Semantic Occupancy Prediction as Dense Object Detection

Wuyang Li, Zhu Yu, Alexandre Alahi

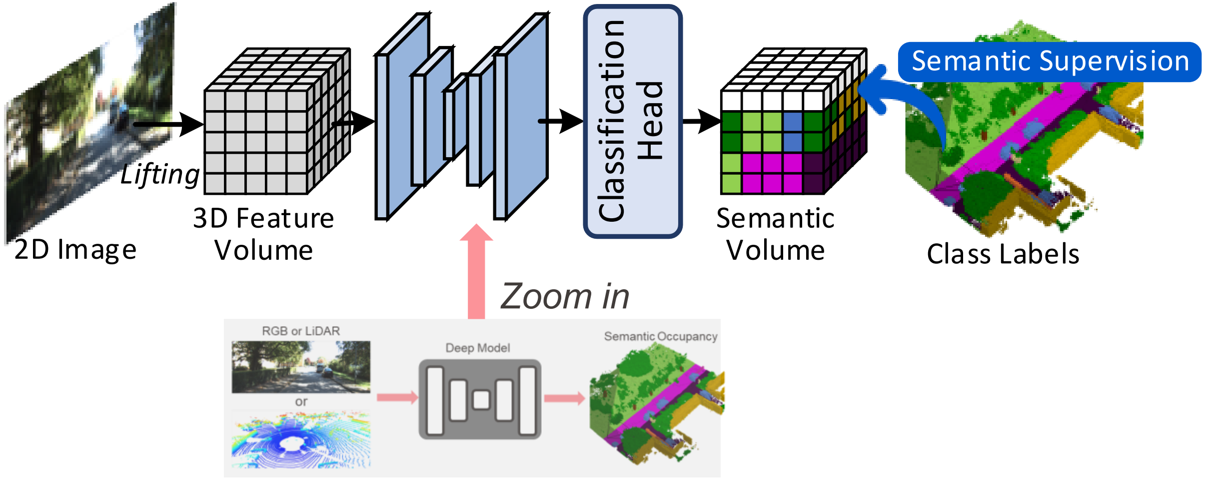
Semantic Occupancy Prediction

- **Objective:** Reconstruct 3D geometry and semantics of surrounding environments from camera or LiDAR inputs



Semantic Occupancy Prediction

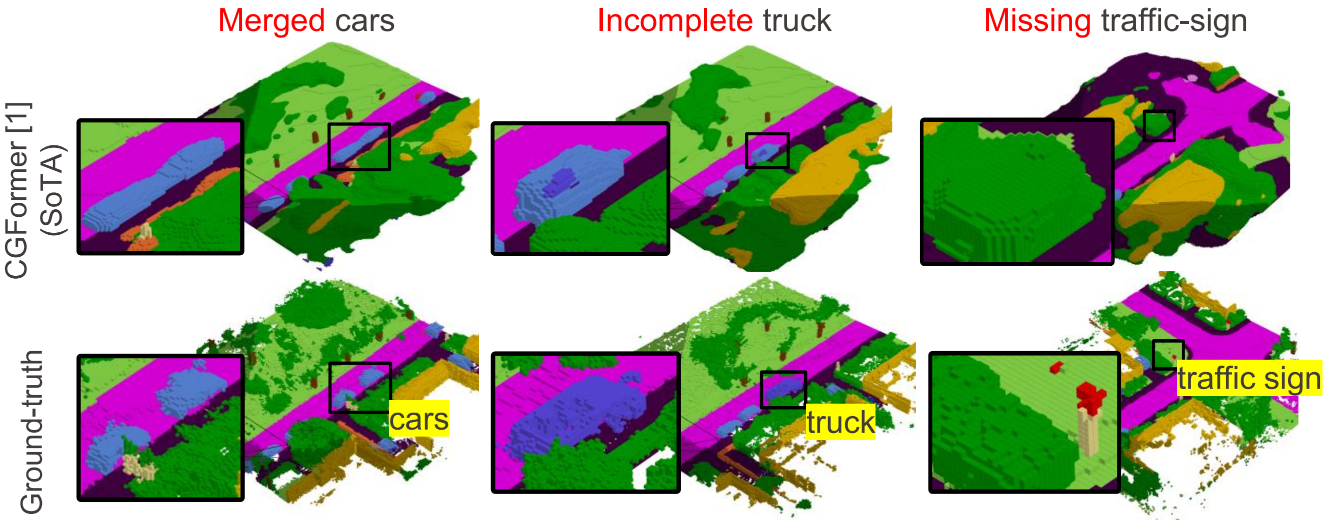
- **Objective:** Reconstruct 3D geometry and semantics of surrounding environments from camera or LiDAR inputs
- **Previous Solutions:** Perform per-voxel recognition (segmentation) on the lifted 3D volume



[1] Context and Geometry Aware Voxel Transformer for Semantic Scene Completion, Yu, Z., et al. NeurIPS, 2024.

Challenge

- **Previous Solutions:** Perform per-voxel recognition (segmentation)
- **Issue:** Segmentation-based formulation **Fails** to perceive object instances well, leading the ambiguity and incompleteness

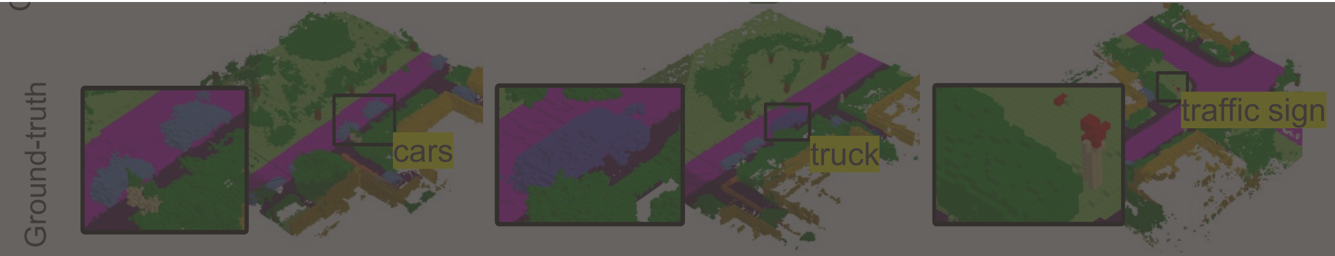


[1] Context and Geometry Aware Voxel Transformer for Semantic Scene Completion, Yu, Z., et al. NeurIPS, 2024.

Challenge

- **Previous Solutions:** Perform per-voxel recognition like segmentation
- **Issue:** Segmentation-based formulation **Fails** to perceive object instances well, leading the ambiguity and incompleteness

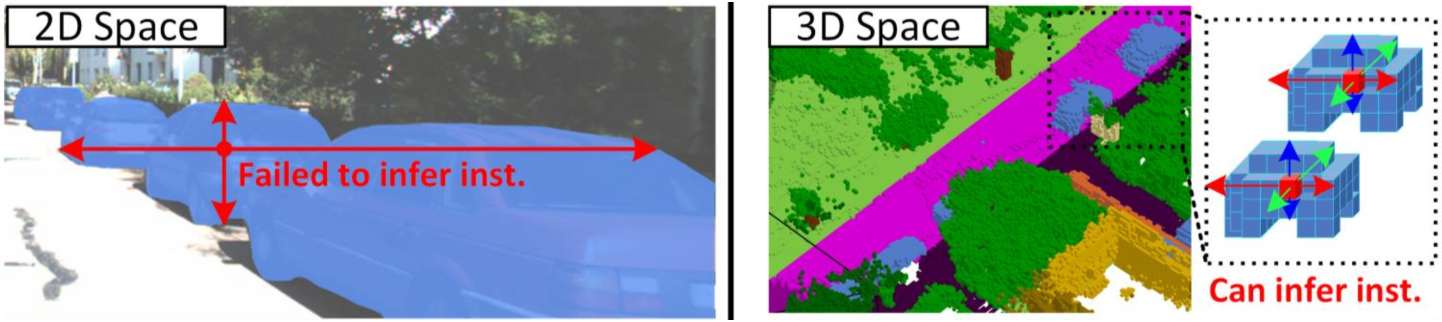
Can We Achieve Instance-Centric Perception Without Additional Labels?



[1] Context and Geometry Aware Voxel Transformer for Semantic Scene Completion, Yu, Z., et al. NeurIPS, 2024.

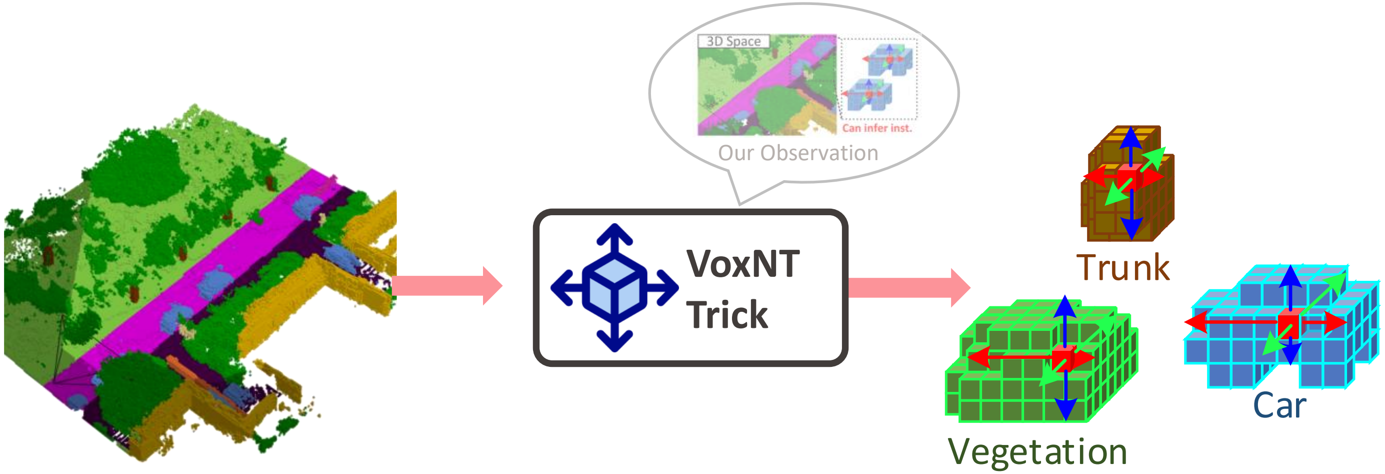
Motivation

- **Observation:** Voxel-level class labels have told instance-level insights
  - Fail to infer instances in 2D pixels due to occlusion
  - Can infer instances in 3D voxels due to occlusion-free nature



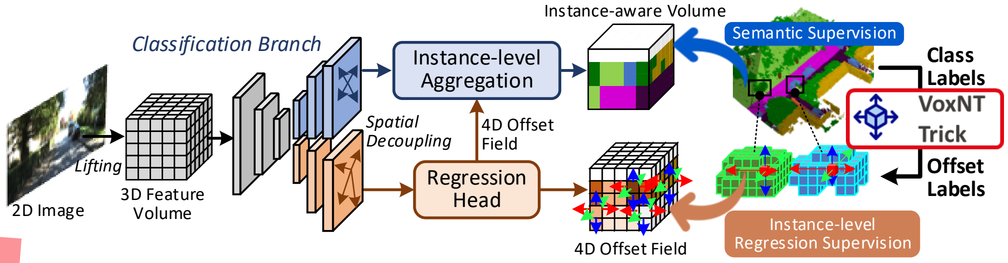
Motivation

- **Observation:** Voxel-level class labels have told instance-level insights
- **Voxel-to-Instance (VoxNT) Trick:** Freely convert voxel-level class labels into instance-level offset labels based on our observation

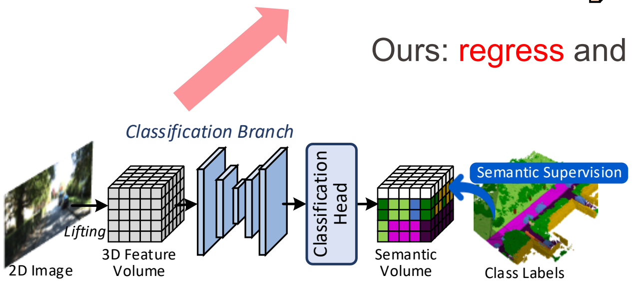


Motivation

- **Voxel-to-Instance (VoxNT) Trick:** Freely convert voxel-level class labels into instance-level offset labels based on our observation
- **VoxDet:** Reformulate occupancy prediction as instance-centric dense object detection based on our free offset labels



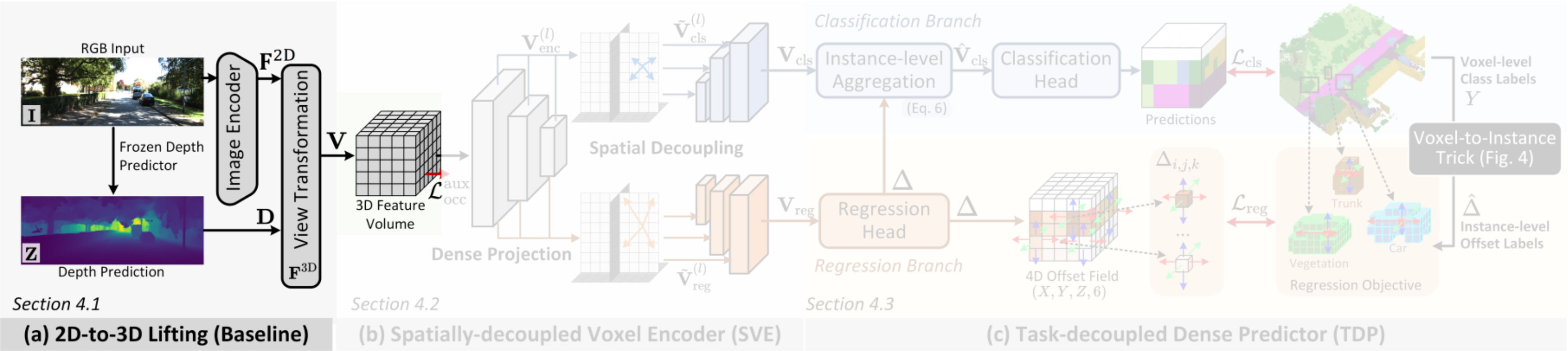
Ours: regress and recognize each instance



Prior works: recognize each voxel

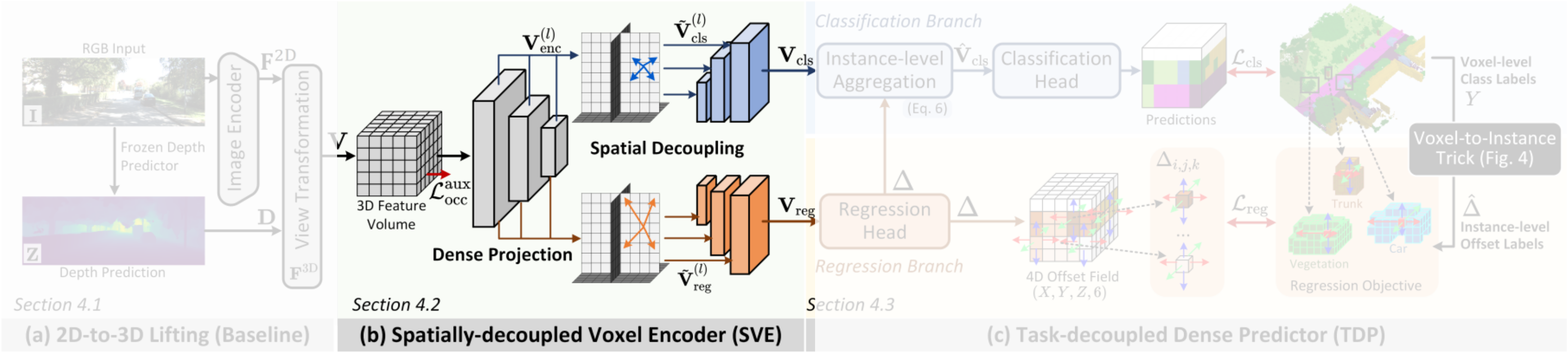
Method

- **2D-to-3D Lifting**
  - Lift 2D image to 3D feature volume



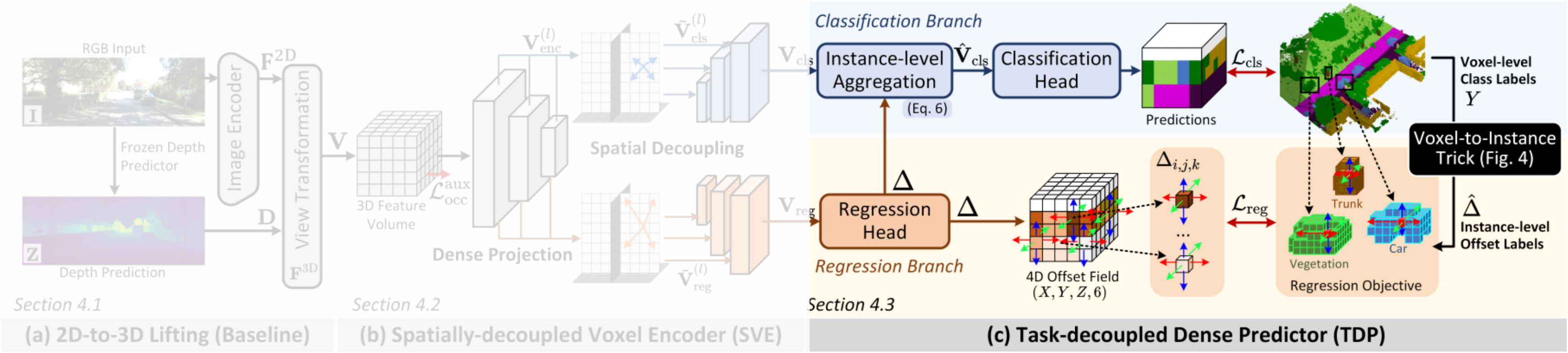
Method

- **Spatially-decoupled Voxel Encoder**
  - Learn task-specific voxel representation with different spatial deformations



Method

- **Task-decoupled Dense Predictor**
  - Regression: densely regress the instance borders with a 4D offset field
  - Classification: aggregate instance-level semantics based on regression



Experiments: VoxDet is Versatile

- VoxDet is state-of-the-art on both Camera and LiDAR benchmarks

T indicates using multi-frame temporal information

Camera-based results on SemanticKITTI test set

Method	Arch.	T	IoU	mIoU	road	sidewalk	parking	other-grnd.	building	car
MonoScene* [6]	Eff-B7		34.16	11.08	54.70	27.10	24.80	5.70	14.40	18.8
TPVFormer [21]	Eff-B7		34.25	11.26	55.10	27.20	27.40	6.50	14.80	19.2
SurroundOcc [64]	Eff-B7		34.72	11.86	56.90	28.30	30.20	6.80	15.20	20.6
OccFormer [80]	Eff-B7		34.53	12.32	55.90	30.30	31.50	6.50	15.70	21.6
IAMSSC [66]	R-50		43.74	12.37	54.00	25.50	24.70	6.90	19.20	21.3
VoxFormer [34]	R-50		42.95	12.20	53.90	25.30	21.10	5.60	19.80	20.8
VoxFormer [34]	R-50	✓	43.21	13.41	54.10	26.90	25.10	7.30	23.50	21.7
DepthSSC [74]	R-50		44.58	13.11	55.64	27.25	25.72	5.78	20.46	21.9
Symphonize [22]	R-50		42.19	15.04	58.40	29.30	26.90	11.70	24.70	23.6
HASSC [60]	R-50		43.40	13.34	54.60	27.70	23.80	6.20	21.10	22.8
HASSC [60]	R-50	✓	42.87	14.38	55.30	29.60	25.90	11.30	23.10	23.0
StereoScene [25]	Eff-B7		43.34	15.36	61.90	31.20	30.70	10.70	24.20	22.8
H2GFormer [63]	R-50		44.20	13.72	56.40	28.60	26.50	4.90	22.80	23.4
H2GFormer [63]	R-50	✓	43.52	14.60	57.90	30.40	30.00	6.90	24.00	23.7
MonoOcc [81]	R-50		-	13.80	55.20	27.80	25.10	9.70	21.40	23.2
CGFormer [77]	Eff-B7		44.41	16.63	64.30	34.20	34.10	12.10	25.80	26.1
L2COcc-C [59]	Eff-B7		44.31	17.03	**66.00**	35.00	33.10	13.50	25.10	27.2
HTCL [24]	Eff-B7	✓	44.23	17.09	64.40	34.80	33.80	12.40	25.90	**27.3**
VoxDet (Ours)	R-50		47.27	18.47	64.70	35.50	34.80	**14.40**	28.10	26.9
VoxDet† (Ours)	R-50		**47.81**	**18.67**	65.50	**36.10**	**35.50**	13.20	**28.40**	**27.3**

IoU +7.9% mIoU +9.2%

LiDAR-based results on SemanticKITTI test set

Method	T	IoU	mIoU	road	sidewalk	parking	other-grnd.	building	car	truck	bicycle
SSCNet [54]		29.8	9.5	27.6	17.0	15.6	6.0	20.9	10.4	1.8	0.0
SSCNet-full [54]		50.0	16.1	51.2	30.8	27.1	6.4	34.5	24.3	1.2	0.5
TS3D [15]		29.8	9.5	28.0	17.0	15.7	4.9	23.2	10.7	2.4	0.0
TS3D/DNet [4]		25.0	10.2	27.5	18.5	18.9	6.6	22.1	8.0	2.2	0.1
LMSCNet [50]		55.3	17.0	64.0	33.1	24.9	3.2	38.7	29.5	2.5	0.0
LMSCNet-SS [50]		56.7	17.6	64.8	34.7	29.0	4.6	38.1	30.9	1.5	0.0
Local-DIFs [49]		57.7	22.7	67.9	42.9	**40.1**	11.4	40.4	34.8	4.4	3.6
JS3C-Net [68]		56.6	23.8	64.7	39.9	34.9	14.1	39.4	33.3	**7.2**	**14.4**
SSA-SC [71]		58.8	23.5	72.2	43.7	37.4	10.9	43.6	36.5	5.7	13.9
L2COcc-D [59]		45.3	18.1	68.2	36.9	34.6	**16.2**	25.8	28.3	4.5	4.9
L2COcc-L [59]		60.3	23.3	68.5	40.6	33.2	6.1	41.5	36.8	5.4	8.7
OccMamba [28]	✓	-	24.6	-	-	-	-	-	-	-	-
VPNet [56]	✓	60.4	25.0	72.4	**44.3**	40.5	14.8	44.0	37.2	4.3	14.0
VoxDet-L (Ours)		**63.0**	**26.0**	**73.0**	43.6	37.5	10.3	**44.5**	**37.7**	6.6	9.9

IoU +4.3% mIoU +4.0%

Experiments: VoxDet is Leaderboard Topper

- VoxDet gives 63.0 IoU, ranking **1st** on SemanticKITTI leaderboard*

Results

#	User	Entries	Date of Last Entry	mIoU ▲	completion ▲	Detailed Results
1	VITA-a	3	05/21/25	26.0 (9)	63.0 (1)	View
2	DPS2CNet	2	03/17/25	26.5 (7)	62.6 (2)	View
3	VITA	10	05/20/25	24.8 (19)	61.8 (3)	View
4	OccFiner_anonymous	3	03/06/24	37.8 (2)	61.7 (4)	View
5	JM	6	10/27/23	24.9 (16)	61.4 (5)	View
6	auto23	10	01/19/25	24.8 (17)	60.9 (6)	View
7	Lubo_Wang	4	03/01/24	25.6 (12)	60.7 (7)	View
8	sixwood	4	12/22/24	26.2 (8)	60.6 (8)	View
9	jdgalviss	8	08/03/23	27.1 (6)	60.6 (9)	View
10	Hailey	2	07/29/23	20.8 (29)	60.2 (10)	View
11	TALoS	1	05/20/24	37.9 (1)	60.2 (11)	View
12	liumu	10	06/17/24	25.2 (13)	60.2 (12)	View
13	luzonghao	9	08/14/23	25.0 (15)	60.2 (13)	View
14	shuminwang	9	07/10/24	24.6 (20)	59.7 (14)	View
15	jmwang	10	06/15/24	24.4 (21)	59.6 (15)	View
16	GSDSY	3	12/22/24	25.6 (10)	58.5 (16)	View
17	yiningma	3	06/03/23	23.7 (23)	58.5 (17)	View

* https://codalab.lisn.upsaclay.fr/competitions/7170#results

Experiments: VoxDet is Efficient

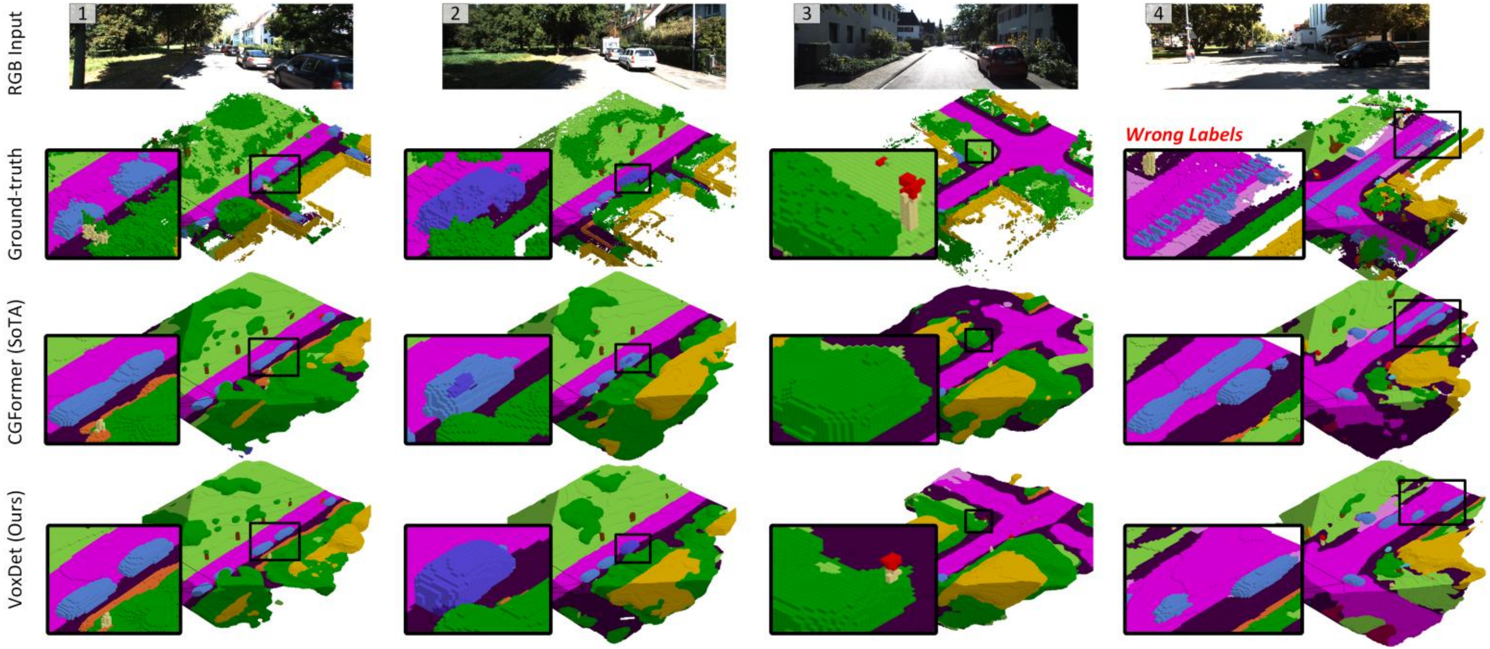
- VoxDet is highly efficient
  - Fewer parameters
  - Faster inference-speed
  - Stronger performance

Camera-based results on SemanticKITTI validation set

Method	N_{param} ↓	T_{inf} ↓	IoU (%) ↑	mIoU (%) ↑
OccFormer [80] [ICCV'23]	214	199	36.42	13.50
StereoScene [25] [IJCAI'24]	117	258	43.85	15.43
CGFormer [77] [NeurIPS'24]	122	205	45.99	16.89
SGFormer [18] [CVPR'25]	126	-	45.01	16.68
ScanSSC [2] [CVPR'25]	145	261	45.95	17.12
VoxDet (Ours)	**53**	**159**	**47.36**	**18.73**
	Reduce 63.4% Para.	1.3 × Faster	IoU +3.1%	mIoU +9.4%

Experiments: VoxDet is Powerful 💪

- VoxDet effectively addresses instance-level challenges



Take-Home Message

- Your occupancy labels are not just class labels

  Try VoxNT Trick! Freely convert voxel-level class labels to instance-level offsets

- Your occupancy predictor should not be just a segmentor

  Try VoxDet! Effectively detect all objects in your 3D voxel space

Project Page