Not All Voxels Are Equal: Hardness-Aware Semantic Scene Completion with Self-Distillation

Song Wang[1], Jiawei Yu[1], Wentong Li[1], Wenyu Liu[1], Xiaolu Liu[1], Junbo Chen[2]* and Jianke Zhu[1]*

[1]Zhejiang University [2]Udeer.ai

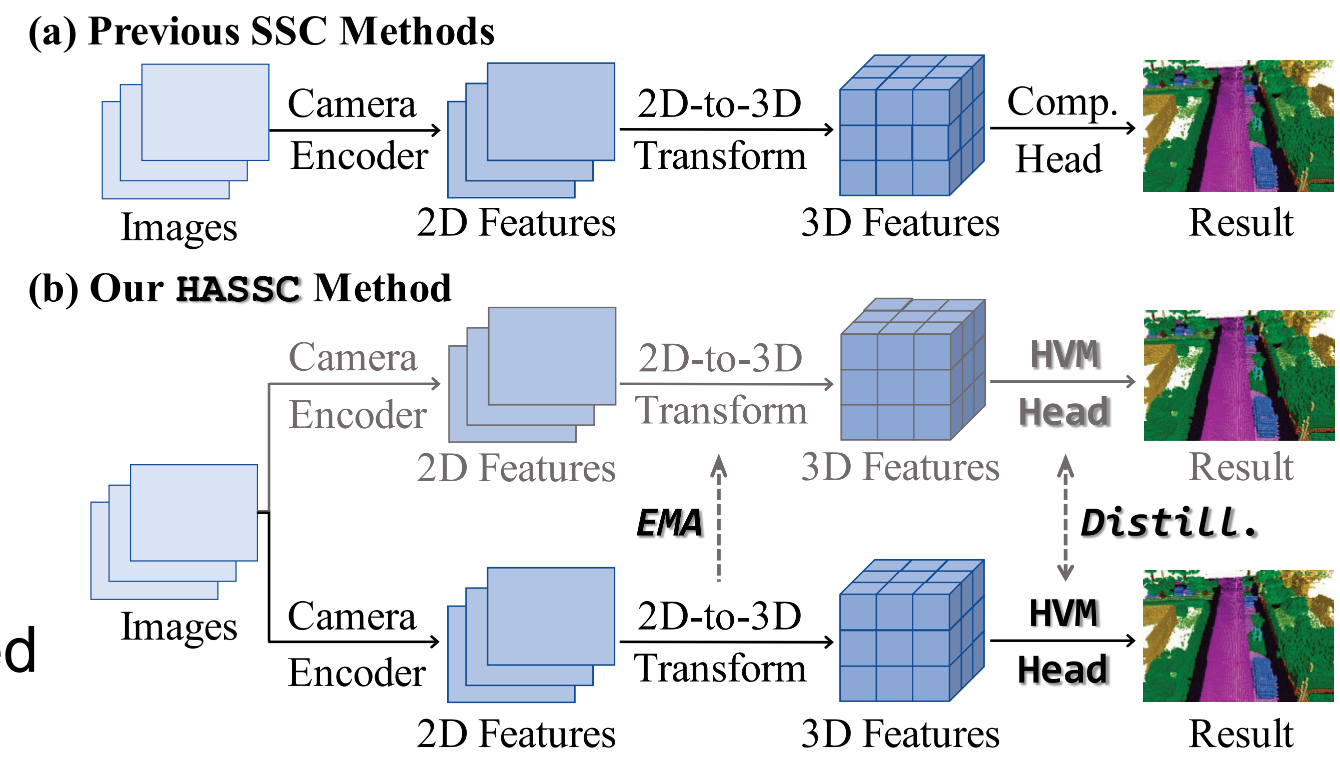
https://github.com/songw-zju/HASSC

Poster: Arch 4A-E-17

Contact: {songw, jkzhu}@zju.edu.cn, junbo@udeer.ai

Overview

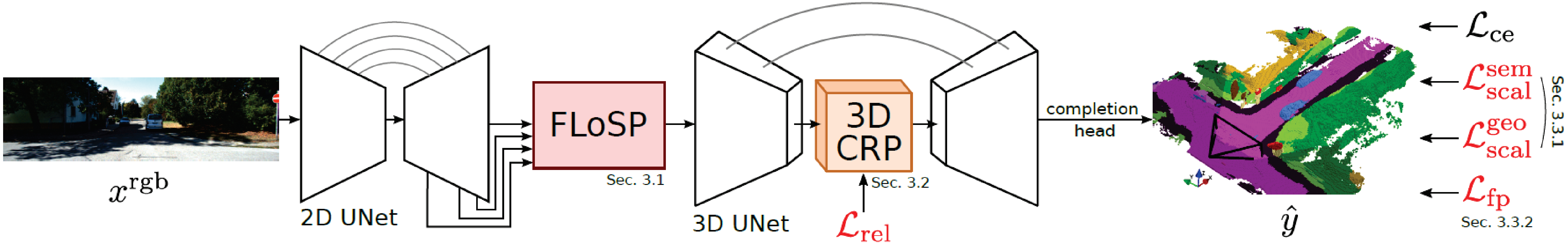
- ➢ The **global hardness** from the network optimization process is defined for dynamical hard voxel selection.
- ➢ The **local hardness** with geometric anisotropy is adopted for voxel-wise refinement.
- ➢ **Self-distillation strategy** is introduced to make training process stable and consistent.



01

Motivation

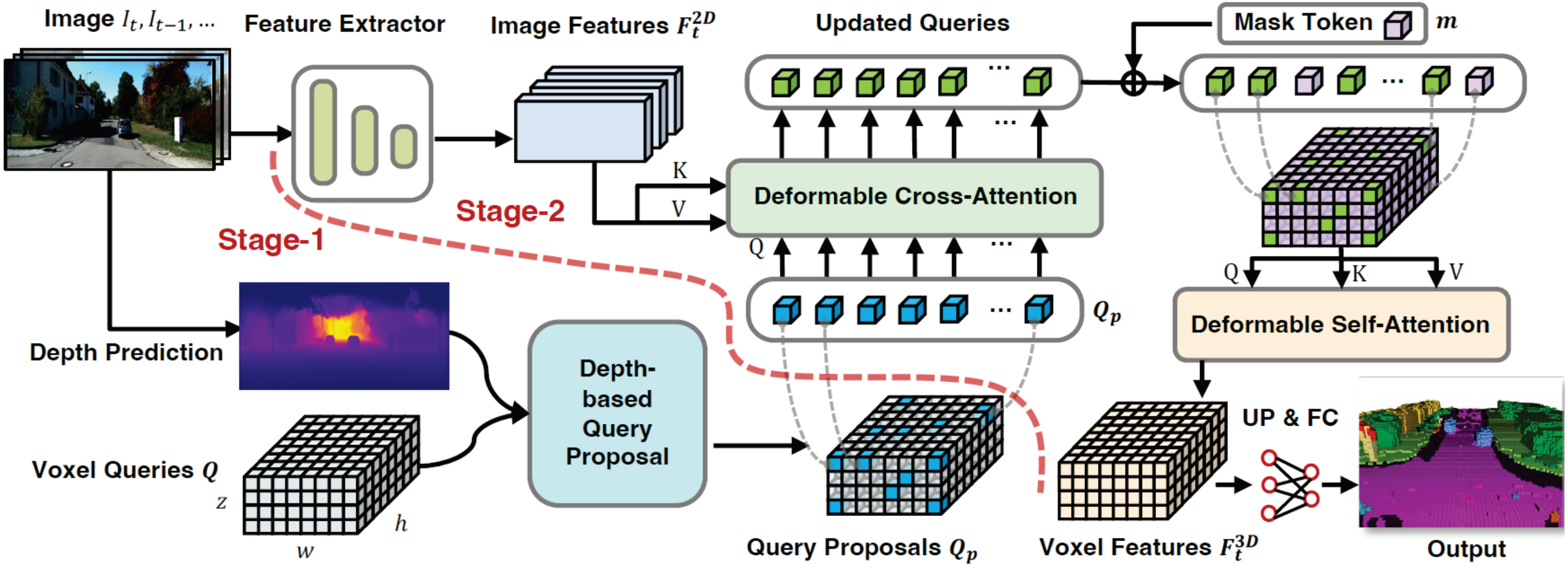
Motivation



MonoScene, CVPR 2022

Cao A Q, De Charette R. Monoscene: Monocular 3d semantic scene completion. CVPR 2022.

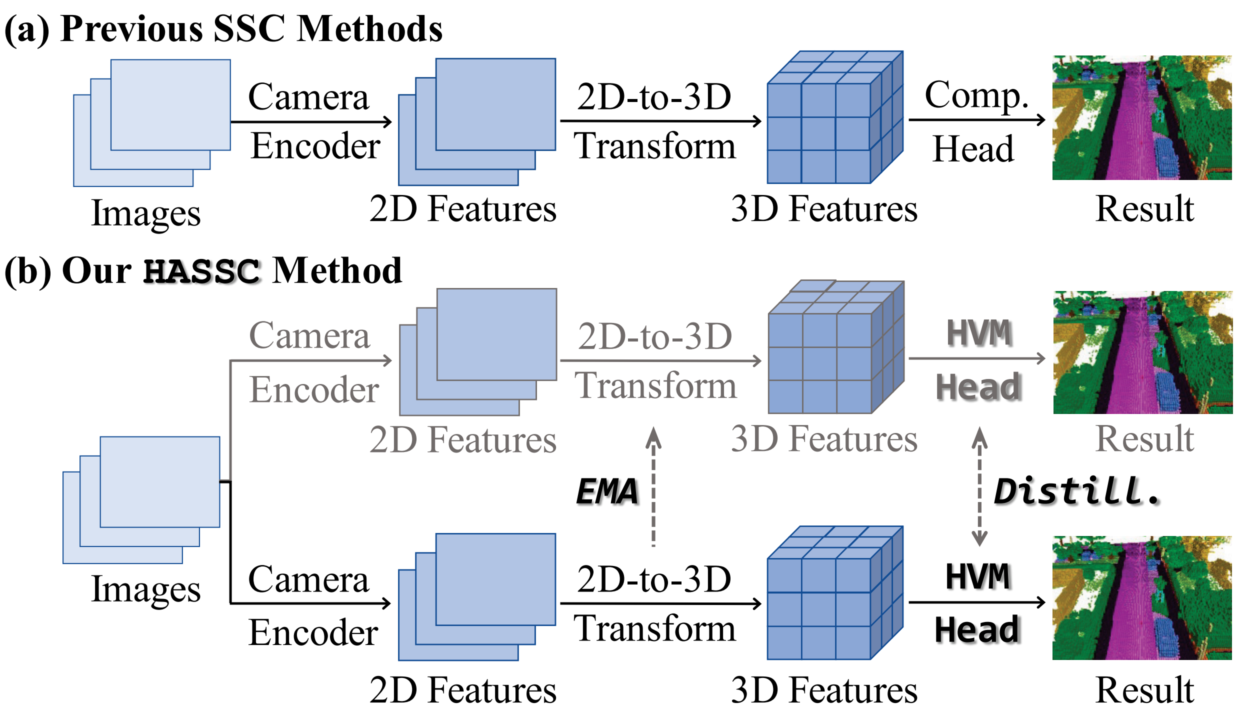
Motivation



VoxFormer, CVPR 2023

Li Y, Yu Z, Choy C, et al. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. CVPR 2023.

Motivation



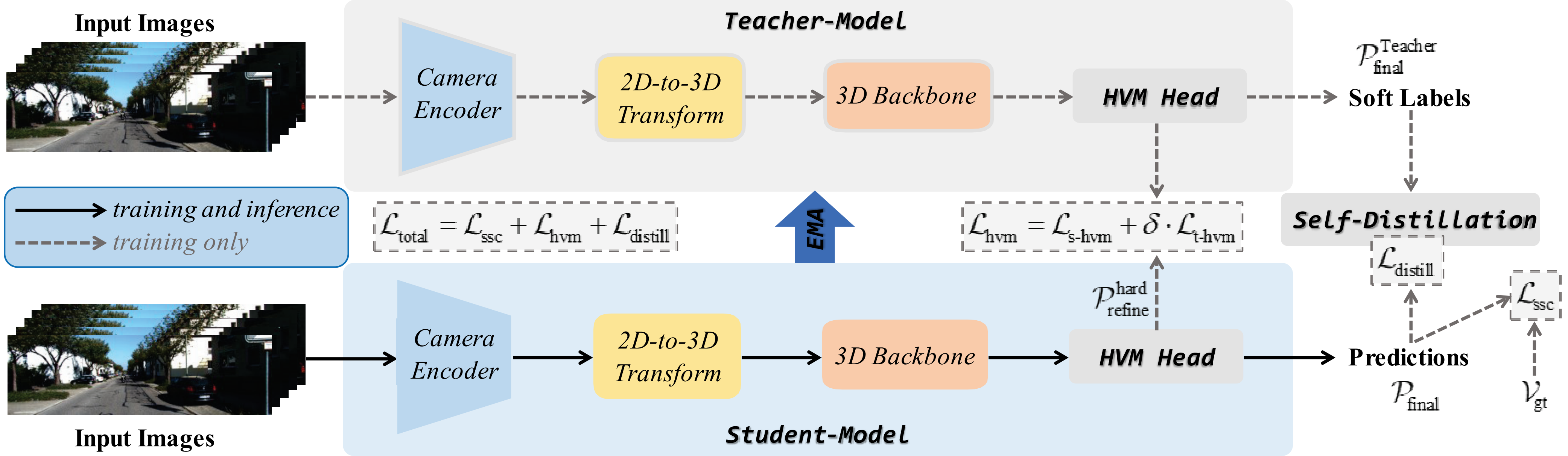
- The 3D dense space typically contains **a large number of empty voxels**, which are easy to learn but require amounts of computation due to **handling all the voxels uniformly** for the existing models.
- The voxels in the **boundary region** are more challenging to differentiate than those in the **interior**.

02

Method

Our Framework

- Propose a **hardness-aware semantic scene completion** (HASSC) scheme that can be easily integrated into existing models without incurring extra cost for inference.
- Take advantage of both the **global and local hardness** to find the hard voxels so that their predictions can be refined by weighted voxel-wise losses during training.



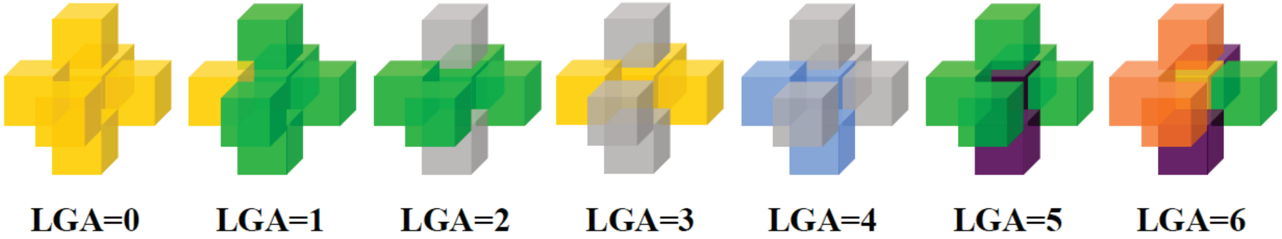
Global Hardness

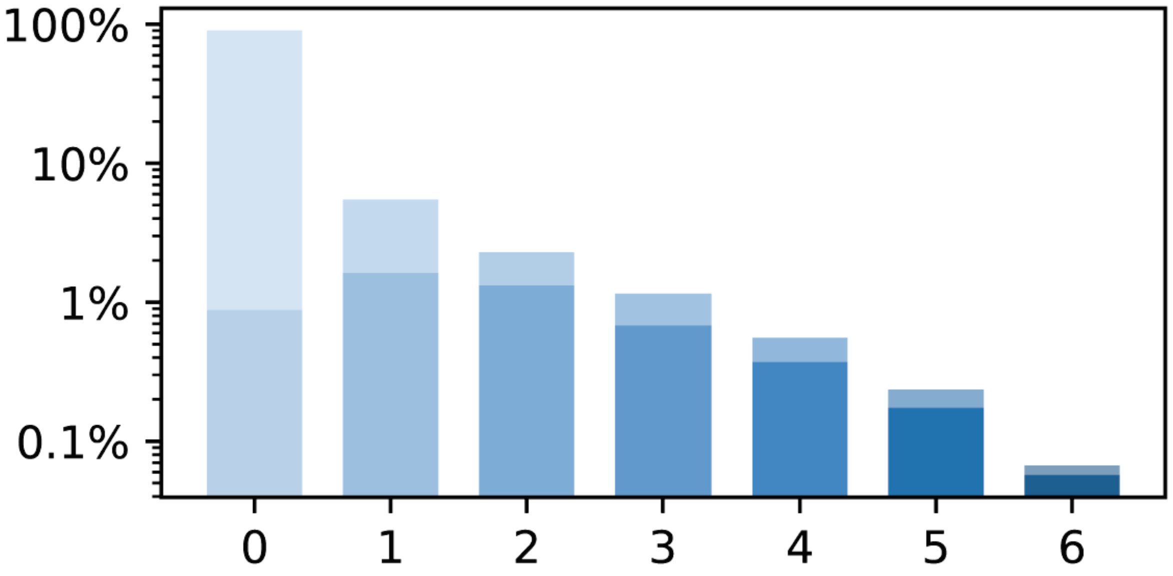
- For the prediction of each voxel, we rank the probabilities of each class in decreasing order. The largest probability in *C* classes is represented as p^a, and the second largest one is denoted as p^b. Then, the global hardness of this voxel is defined as follows

$$\mathcal{H}_{i,j,k}^{\text{global}} = \frac{1}{p^a - p^b}$$

- The global hardness measures the uncertainty of the semantic scene completion prediction between the class a and b, which varies with the optimization of the network. We mainly employ the global hardness to select hard voxels and refine their predictions.

Local Hardness

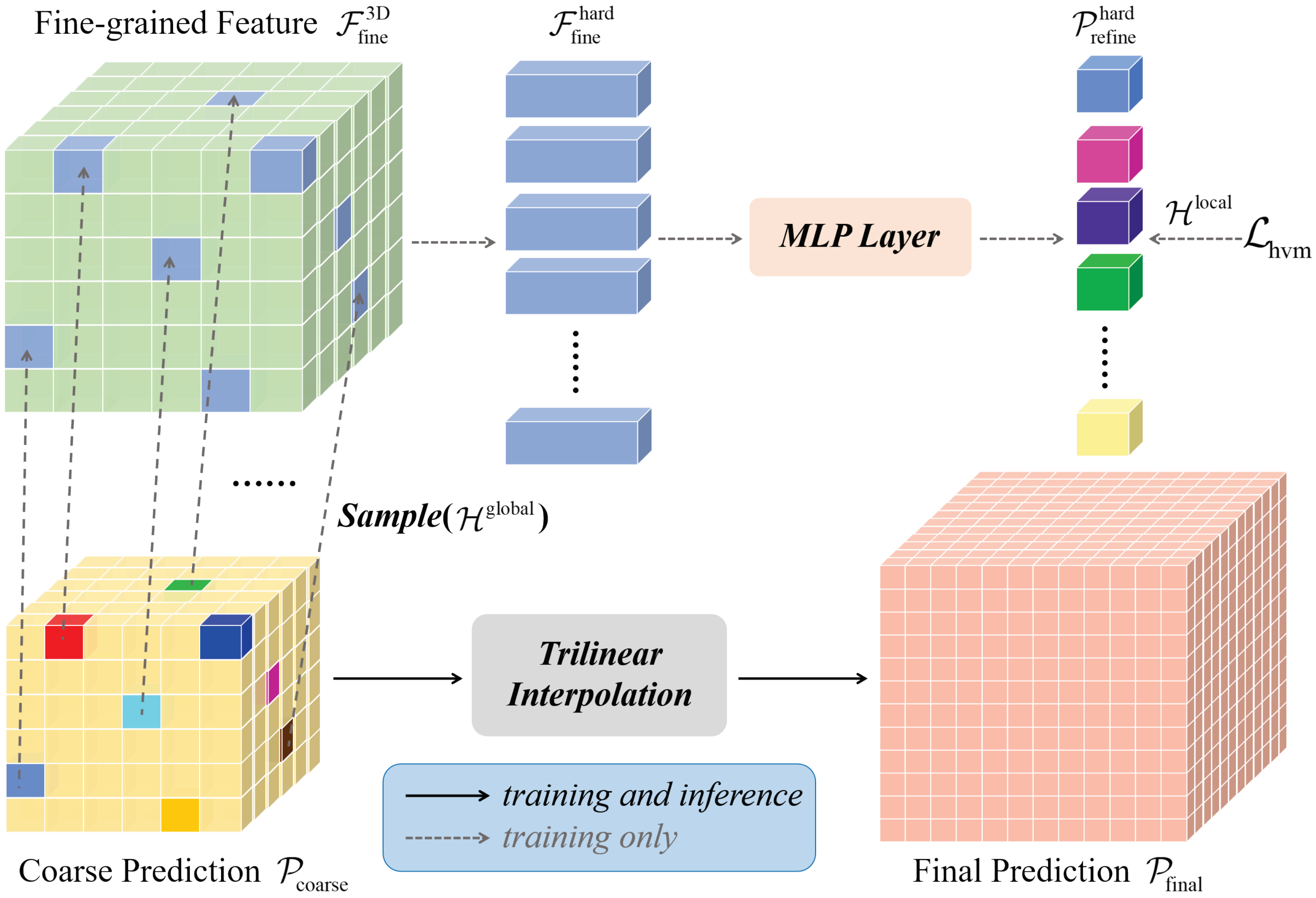




Local Geometric Anisotropy (LGA) Distribution on SemanticKITTI

$$\mathcal{A}_{i,j,k} = \sum_{m=1}^{M} \left(v_{\mathrm{gt}} \oplus v_{\mathrm{gt}}^{m} \right) \qquad \mathcal{H}_{i,j,k}^{\mathrm{local}} = \alpha + \beta \mathcal{A}_{i,j,k}$$

Hard Voxel Mining (HVM) Head



Training Loss

- For Hard Voxel Mining

$$\mathcal{P}_{\text{refine}}^{\text{hard}} = \text{MLP}(\mathcal{F}_{\text{fine}}^{\text{hard}})$$

$$\mathcal{L}_{\text{s-hvm}} = \frac{1}{N} \sum_{n=1}^{N} \mathcal{H}_{n}^{\text{local}} \cdot \text{CE}(v_{\text{refine}}^{n}, v_{\text{gt}}^{n})$$

- For Self-Distilltion

$$\mathcal{L}_{\text{distill}} = \lambda e^{\mu} \cdot \boldsymbol{D}_{\text{KL}} \left(\mathcal{P}_{\text{final}}^{\text{Teacher}} \| \mathcal{P}_{\text{final}} \right)$$

- Totally,

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ssc}} + \mathcal{L}_{\text{hvm}} + \mathcal{L}_{\text{distill}}$$

03

Experiments

Performance Comparison

—— SemanticKITTI Validation

Methods	VoxFormer-S [30]			HASSC VoxFormer-S			VoxFormer-T [30]			HASSC VoxFormer-T			StereoScene† [22]			HASSC StereoScene		
Modality	Camera			Camera			Camera			Camera			Camera			Camera		
Range	S	M	L	S	M	L	S	M	L	S	M	L	S	M	L	S	M	L
IoU (%)↑	65.35	57.54	44.02	**65.54**	**57.99**	**44.82**	65.38	57.69	44.15	**66.05**	**58.01**	**44.58**	65.70	56.84	43.66	65.52	**57.01**	**44.55**
mIoU (%)↑	17.66	16.48	12.35	**18.98**	**17.95**	**13.48**	21.55	18.42	13.35	**24.10**	**20.27**	**14.74**	23.27	21.15	15.24	**24.43**	**22.17**	**15.88**
car (3.92%)	39.78	35.24	25.79	**42.37**	**36.78**	**27.23**	44.90	37.46	26.54	**45.79**	**37.70**	**27.33**	47.05	43.52	31.15	46.47	43.02	30.64
bicycle (0.03%)	3.04	1.48	0.59	2.72	**2.26**	**0.92**	5.22	2.87	1.28	4.23	2.11	1.07	2.38	2.15	1.05	**4.20**	**2.63**	**1.20**
motorcycle (0.03%)	2.84	1.10	0.51	**4.49**	**1.63**	**0.86**	2.98	1.24	0.56	**5.64**	**2.03**	**1.14**	4.78	2.84	1.55	**5.26**	**3.34**	0.91
truck (0.16%)	7.50	7.47	5.63	6.25	**11.00**	**9.91**	9.80	10.38	7.26	**22.89**	**21.90**	**17.06**	18.72	22.48	17.55	**24.94**	**34.73**	**23.72**
other-veh. (0.20%)	8.71	4.98	3.77	**14.77**	**8.85**	**5.61**	17.21	10.61	7.81	**22.71**	**13.52**	**8.83**	17.33	13.79	9.26	**20.61**	**14.24**	7.77
person (0.07%)	4.10	3.31	1.78	**5.11**	**4.89**	**2.80**	4.44	3.50	1.93	**5.12**	**4.18**	**2.25**	6.31	4.37	2.17	6.06	3.58	1.79
bicyclist (0.07%)	6.82	7.14	3.32	**6.87**	**8.57**	**4.71**	2.65	3.92	1.97	**4.09**	**6.58**	**4.09**	7.70	4.75	2.30	**8.22**	**5.65**	**2.47**
motorcyclist (0.05%)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
road (15.30%)	72.40	65.74	54.76	**74.49**	**68.04**	**57.05**	75.45	66.15	53.57	**78.51**	**70.02**	**57.23**	79.24	74.16	61.86	**80.61**	**75.53**	**62.75**
parking (1.12%)	10.79	18.49	15.50	**15.49**	**21.23**	**15.90**	21.01	23.96	19.69	**29.43**	**26.69**	**19.89**	21.33	21.19	17.02	**25.21**	**25.95**	**20.20**
sidewalk (11.13%)	39.35	33.20	26.35	**42.69**	**36.32**	**28.25**	45.39	34.53	26.52	**51.69**	**38.83**	**29.08**	50.71	41.86	30.58	**52.68**	**43.61**	**32.40**
other-grnd(0.56%)	0.00	1.54	0.70	**0.02**	**2.38**	**1.04**	0.00	0.76	0.42	0.00	**1.55**	**1.26**	0.00	1.12	0.85	0.00	0.18	0.51
building (14.10%)	17.91	24.09	17.65	**22.78**	**27.30**	**19.05**	25.13	29.45	19.54	**27.99**	**30.81**	**20.19**	26.98	32.52	22.71	**29.09**	31.68	**22.90**
fence (3.90%)	12.98	10.63	7.64	9.81	8.70	6.58	16.17	11.15	7.31	**17.09**	**11.65**	**7.95**	22.50	14.26	8.73	20.88	13.32	8.67
vegetation (39.3%)	40.50	34.68	24.39	40.49	**35.53**	**25.48**	43.55	38.07	26.10	**44.68**	**38.93**	**27.01**	40.20	36.10	24.81	**40.29**	**36.44**	**26.27**
trunk (0.51%)	15.81	10.64	5.08	14.93	**11.25**	**6.15**	21.39	12.75	6.10	**22.22**	**14.11**	**7.71**	21.45	15.28	7.17	**21.65**	14.92	7.14
terrain (9.17%)	32.25	35.08	29.96	**36.66**	**38.28**	**32.94**	42.82	39.61	33.06	**47.04**	**41.37**	**33.95**	45.75	43.67	34.87	**48.50**	**46.95**	**38.10**
pole (0.29%)	14.47	11.95	7.11	**15.25**	**12.48**	**7.68**	20.66	15.56	9.15	18.95	14.76	**9.20**	20.43	18.95	10.66	18.67	16.34	9.00
traf.-sign (0.08%)	6.19	6.29	4.18	5.52	5.61	4.05	10.63	8.09	4.94	9.89	**8.44**	4.81	9.21	8.91	5.19	**10.88**	**9.08**	**5.23**

Performance Comparison

—— SemanticKITTI Test

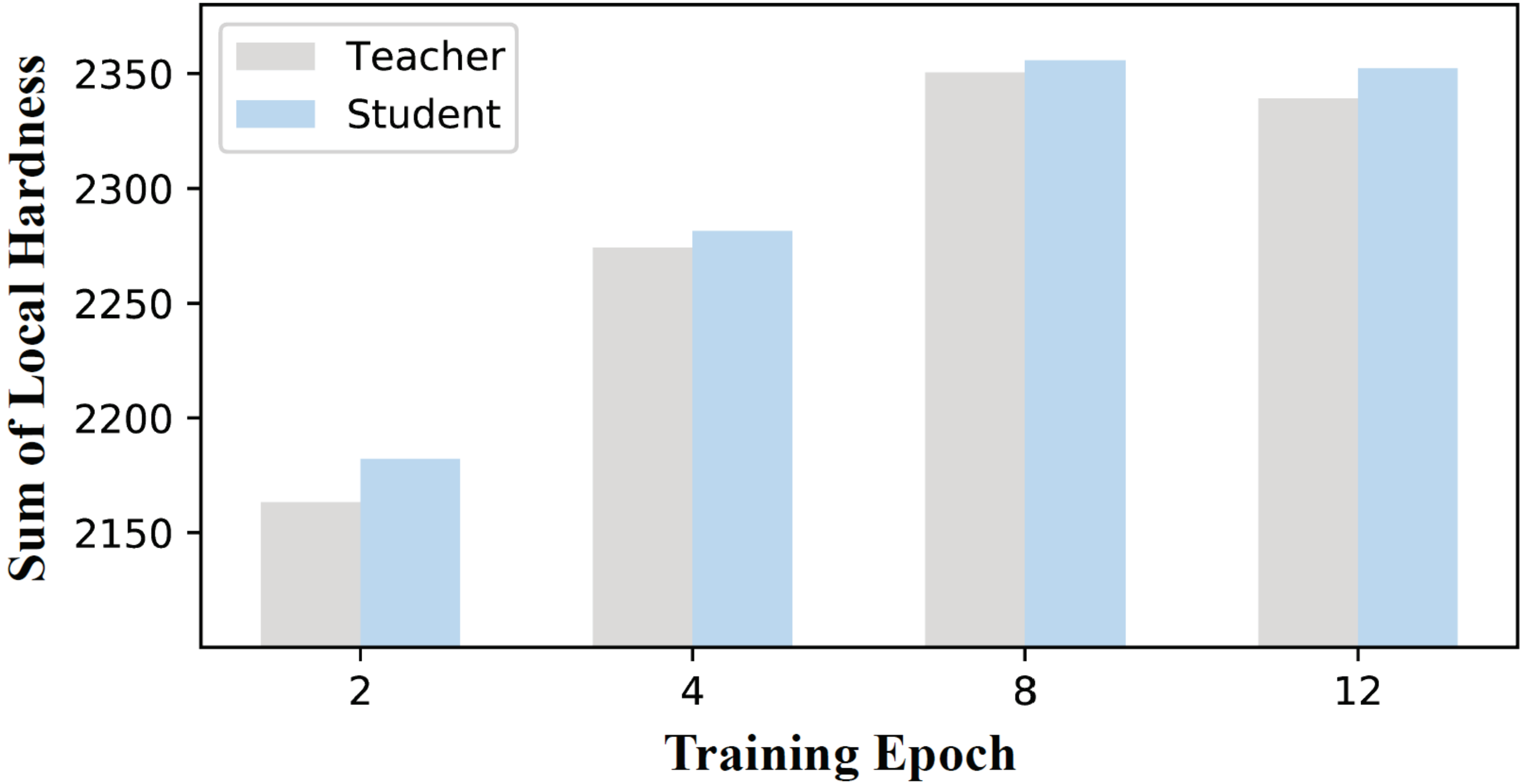
Methods	SSC Input	Pub.	IoU (%)↑	mIoU (%)↑
LMSCNet* [40]	$\hat{x}^{\text{occ}}_{\text{3D}}$	3DV 2020	31.38	7.07
3DSketch* [5]	$x^{\text{rgb}}, \hat{x}^{\text{TSDF}}$	CVPR 2020	26.85	6.23
AICNet* [24]	$x^{\text{rgb}}, \hat{x}^{\text{depth}}$	CVPR 2020	23.93	7.09
JS3C-Net* [53]	$\hat{x}^{\text{pts}}$	AAAI 2021	34.00	8.97
MonoScene [4]	x^{rgb}	CVPR 2022	34.16	11.08
TPVFormer [18]	x^{rgb}	CVPR 2023	34.25	11.26
OccFormer [60]	x^{rgb}	ICCV 2023	34.53	12.32
NDC-Scene [56]	x^{rgb}	ICCV 2023	36.19	12.58
VoxFormer-S [30]	x^{rgb}	CVPR 2023	42.95	12.20
VoxFormer-T [30]	$x^{\text{rgb}} \times 5$	CVPR 2023	43.21	13.41
HASSC-VoxFormer-S	x^{rgb}	-	**43.40**	13.34
HASSC-VoxFormer-T	$x^{\text{rgb}} \times 5$	-	42.87	**14.38**

Ablation Study

Global	Local	T-HVM	T-Distill	IoU (%)↑	mIoU (%)↑
				44.16	13.33
✓				43.89	13.30
	✓			44.00	13.40
✓	✓			43.98	13.91
✓	✓	✓		44.12	14.03
			✓	44.38	13.65
✓	✓	✓	✓	**44.58**	**14.74**

Ablation study on our proposed **HASSC** scheme

Ablation Study



Visualization of the sum of the local hardness change during training on both student and teacher branches

Ablation Study

Methods	VoxFormer-T	HASSC-VoxFormer-T
Params (M)	57.91	58.43
Inference Speed (ms)	724.05	720.84
IoU (%)↑	44.16	**44.58**
mIoU (%)↑	13.33	**14.74**

Comparison with baseline model on the training and inference efficiency

Methods	Hardness	IoU (%)↑	mIoU (%)↑
PALNet [23]	Local	44.28	13.28
PointRend [20]	Global	44.29	13.57
Xiao *et al.* [52]	Global	44.10	13.33
Ours	Global & Local	**44.58**	**14.74**

Comparison with other hard sample mining schemes

Voxel Numbers (N)	0	1024	2048	4096	8192
IoU (%)↑	**44.16**	44.01	43.92	44.12	44.09
mIoU (%)↑	13.33	13.52	13.64	**14.03**	13.74

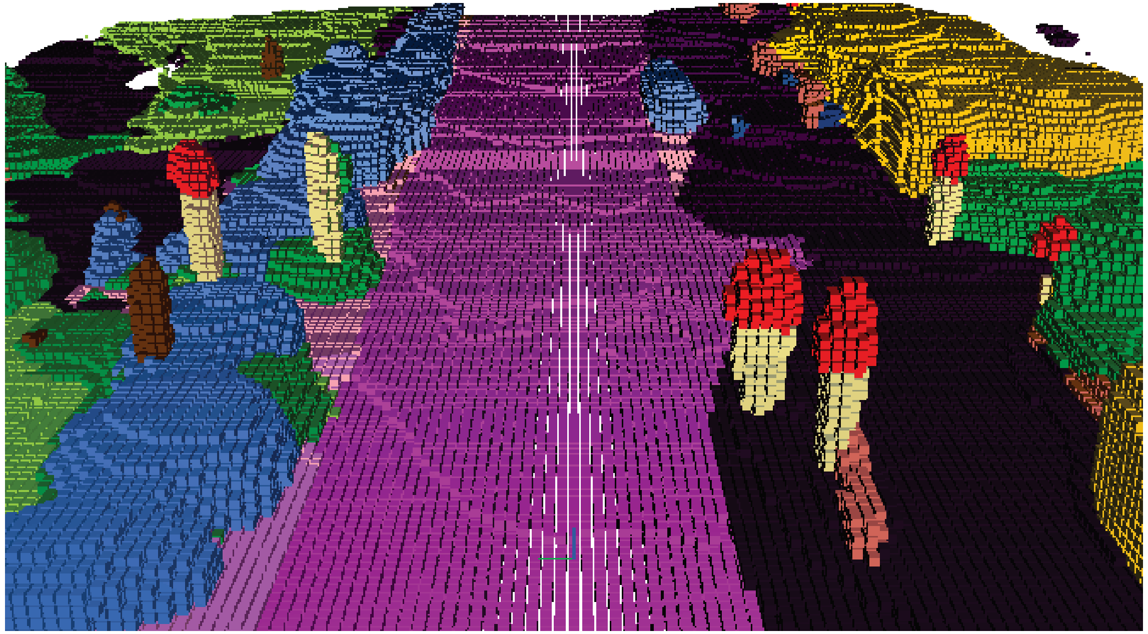
Ablation study on the number of hard voxel selection

Distill Weight (λ)	0	12	24	48	96
IoU (%)↑	44.12	44.06	44.25	**44.58**	44.51
mIoU (%)↑	14.03	13.80	14.23	**14.74**	14.39

Ablation study on the weight of self-distillation from teacher model

Performance on Sequence 08
(*validation set*)

Camera View (Left)



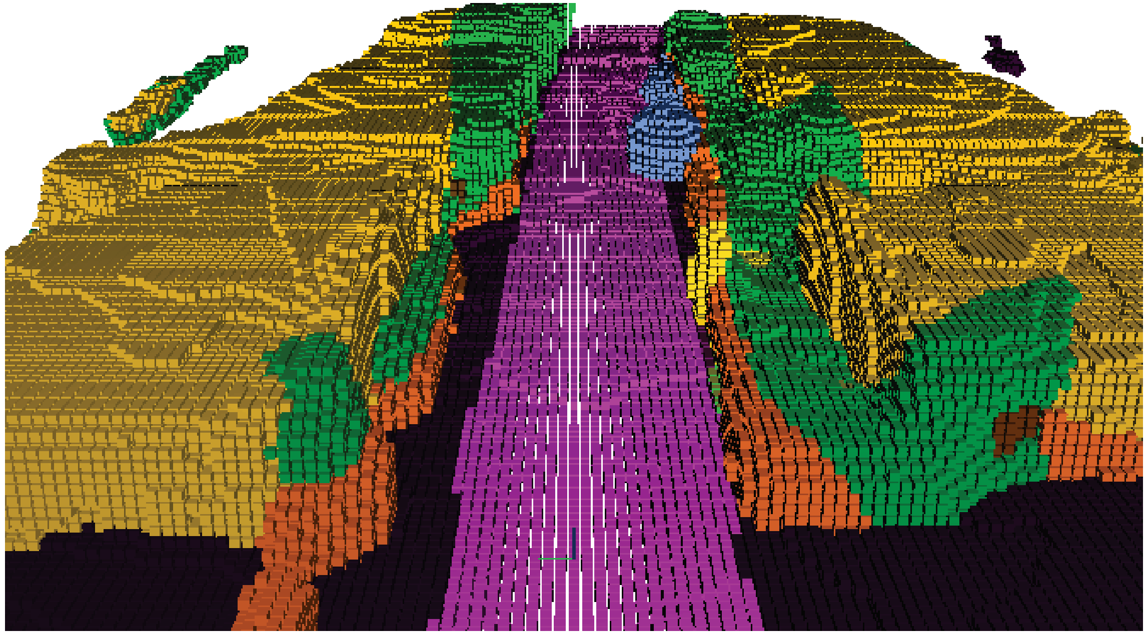
Semantic Scene Completion (Result)

car | pole | fence | person | bicycle | parking | vegetation | other-vehicle | bicyclist | motorcyclist

road | truck | terrain | trunk | building | sidewalk | traffic-sign | other-ground | motocycle

Performance on Sequence 11
(*hidden test set*)

Camera View (Left)

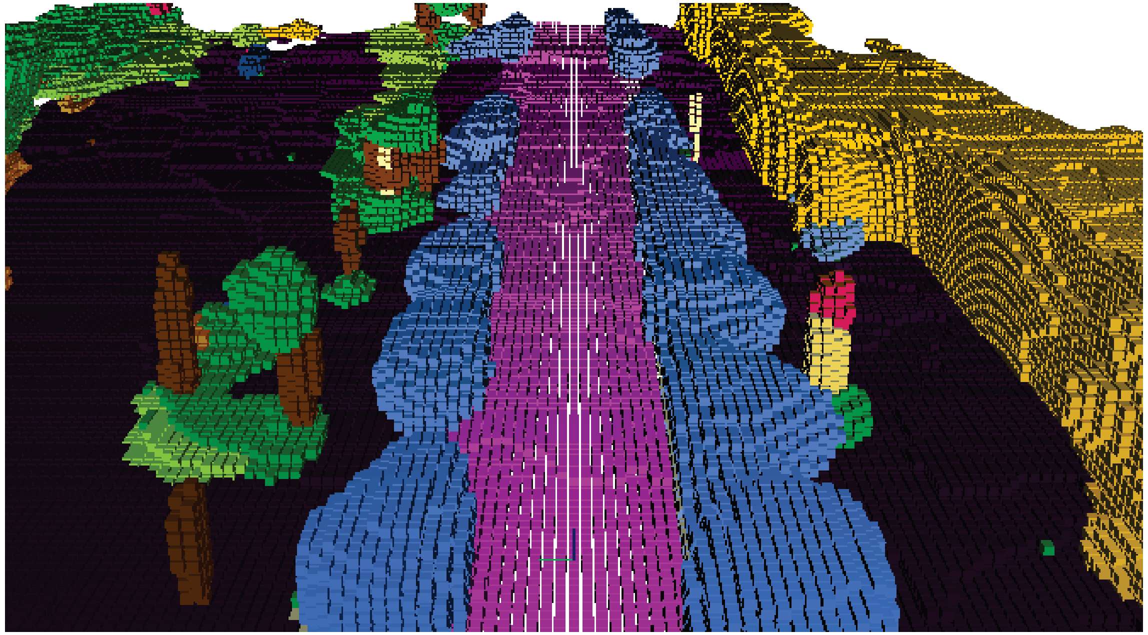


Semantic Scene Completion (Result)

car, pole, fence, person, bicycle, parking, vegetation, other-vehicle, bicyclist, motorcyclist,
road, truck, terrain, trunk, building, sidewalk, traffic-sign, other-ground, motocycle

Performance on Sequence 13
(*hidden test set*)

Camera View (Left)

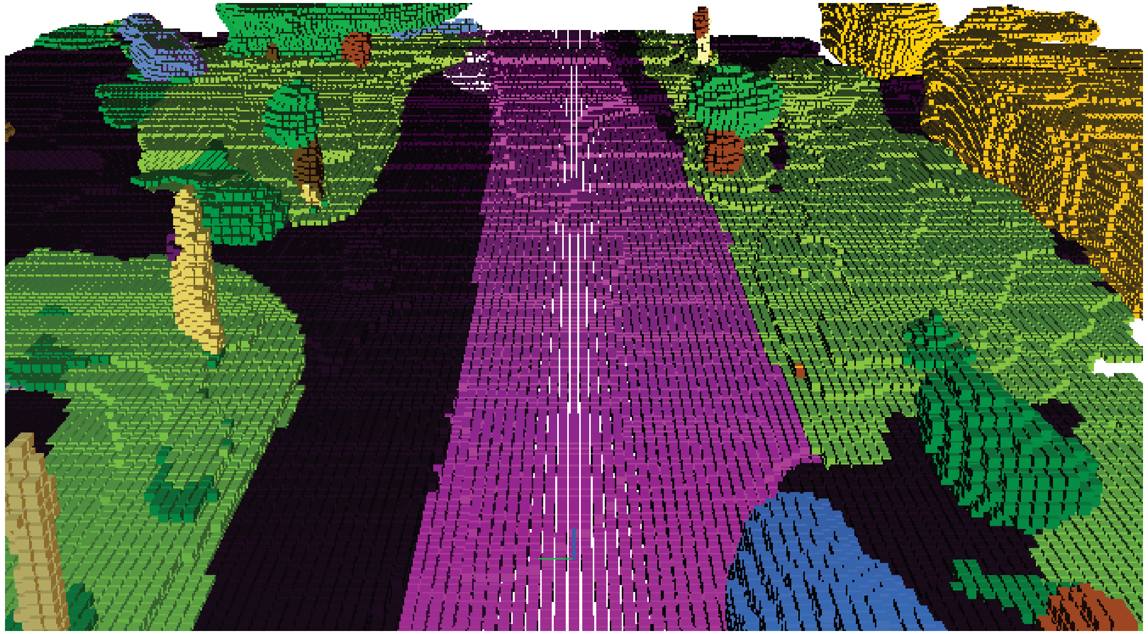


Semantic Scene Completion (Result)

car, pole, fence, person, bicycle, parking, vegetation, other-vehicle, bicyclist, motorcyclist
road, truck, terrain, trunk, building, sidewalk, traffic-sign, other-ground, motocycle

Performance on Sequence 16
(*hidden test set*)

Camera View (Left)



Semantic Scene Completion (Result)

car, pole, fence, person, bicycle, parking, vegetation, other-vehicle, bicyclist, motorcyclist

road, truck, terrain, trunk, building, sidewalk, traffic-sign, other-ground, motocycle

Thanks!