OCCGen : Generative Multi-modal 3D Occupancy PPrediction for Autonomous Driving

Guoqing Wang, Zhongdao Wang, Pin Tang, Jilai Zheng, Xiangxuan Ren, Bailan Feng, Chao Ma

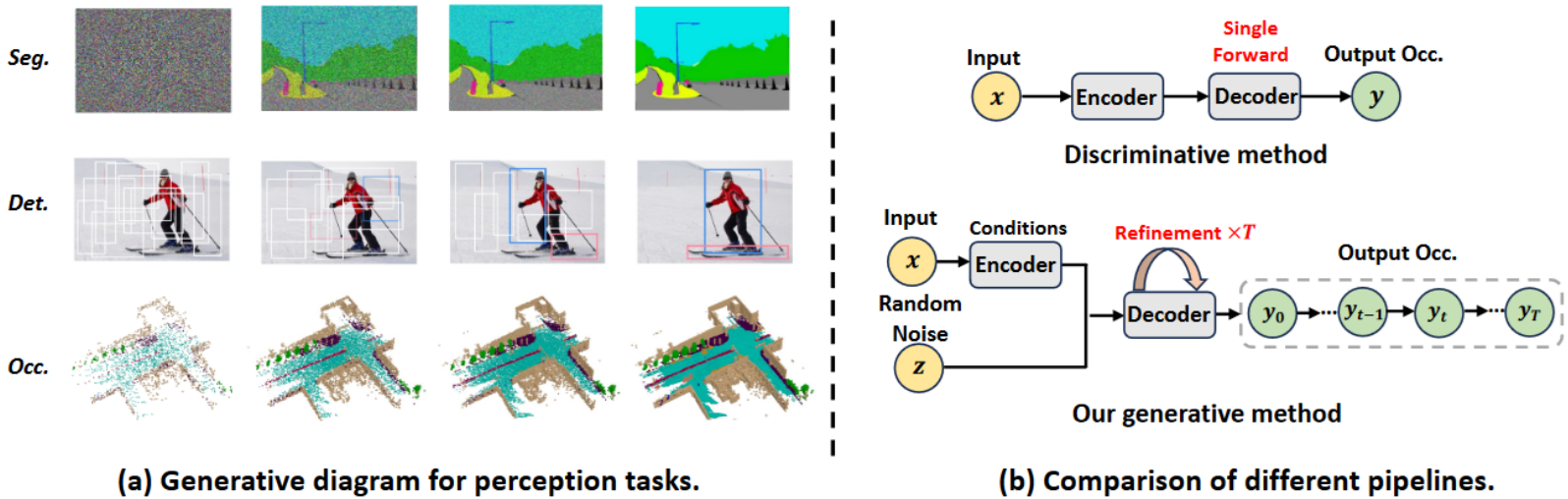
ECCV 2024

- Problem/Objective
  - Semantic Occupancy Prediction

- Contribution/Key Idea
  - Noise-to-occupancy prediction
  - Efficient Encoder-Decoder runner mechanism
  - Improve performance

OCCGen : Generative Multi-modal 3D Occupancy PPrediction for Autonomous Driving

Guoqing Wang, Zhongdao Wang, Pin Tang, Jilai Zheng, Xiangxuan Ren, Bailan Feng, Chao Ma



(a) Generative diagram for perception tasks.

(b) Comparison of different pipelines.

기존 방법 : One-shot voxel wise segmentation problem with a single forward step

본 모델 : Progressive refinement decoder (Coarse-to-fine)

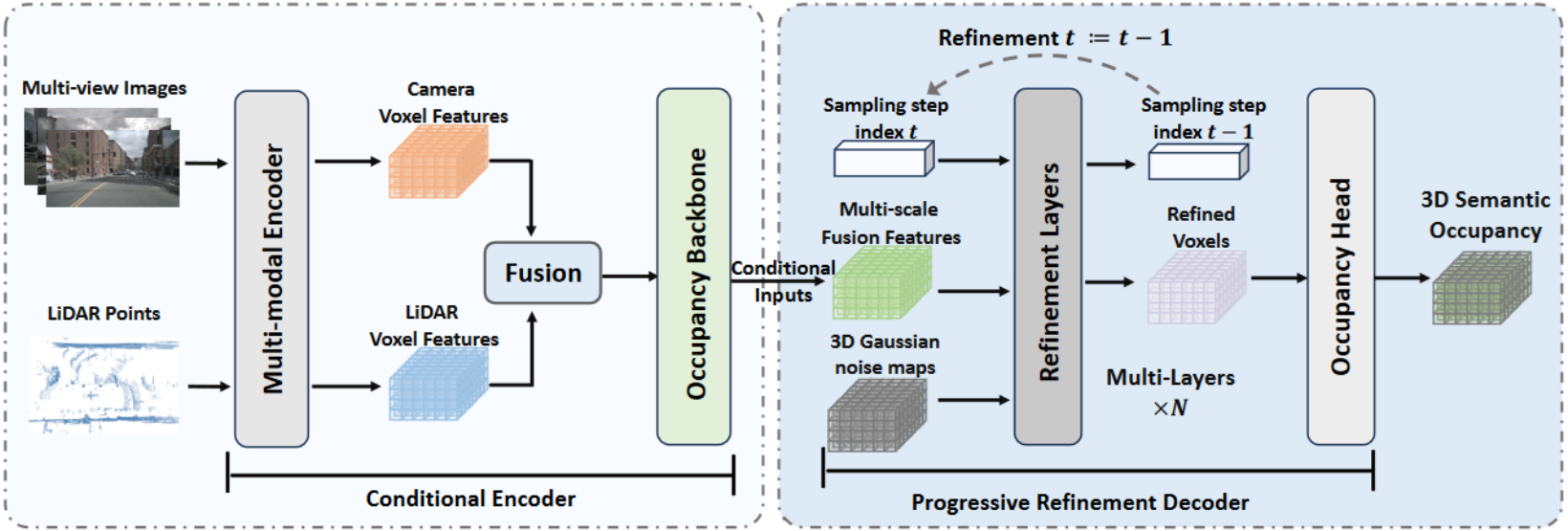
(1) 계산량 - 품질 사이의 trade off 선택가능

(2) occupancy prediction과 동시에 uncertainty estimation 가능

OCCGen : Generative Multi-modal 3D Occupancy PPrediction for Autonomous Driving

Guoqing Wang, Zhongdao Wang, Pin Tang, Jilai Zheng, Xiangxuan Ren, Bailan Feng, Chao Ma

ECCV 2024



$$X_p \in \mathbb{R}^{N_L \times (3+d)}$$

$$X_c \in \mathbb{R}^{N_C \times H_C \times W_C \times 3}$$

$$Y \in \{c_0, c_1, ..., c_N\}^{H \times W \times Z}$$

$$\Delta Y_t = f_\theta\left(x, t, Y_{t+1}\right), \quad Y_t = Y_{t+1} \oplus \Delta Y_t$$

$$Y_T \xrightarrow{f_\theta} Y_{T-1} \xrightarrow{f_\theta} \dots \xrightarrow{f_\theta} Y_0$$

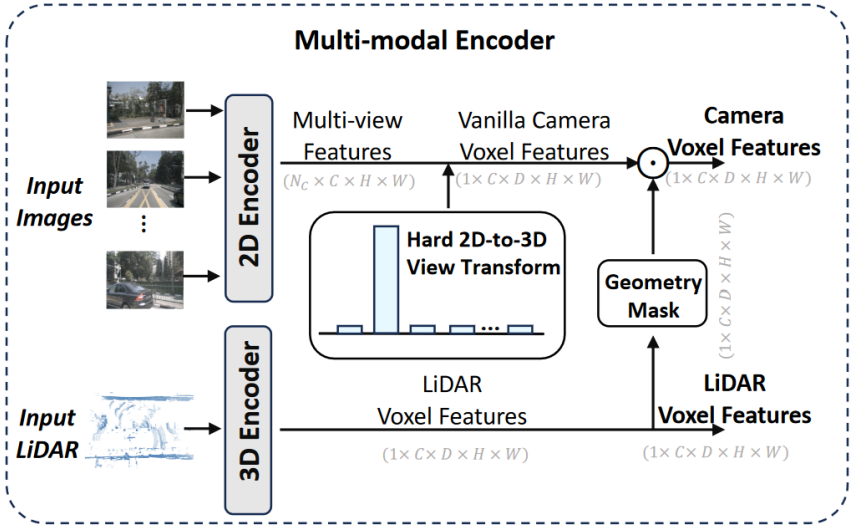
3D Gaussian voxel map

Refined Occupancy

김범준

OCCGen : Generative Multi-modal 3D Occupancy PPrediction for Autonomous Driving

Guoqing Wang, Zhongdao Wang, Pin Tang, Jilai Zheng, Xiangxuan Ren, Bailan Feng, Chao Ma



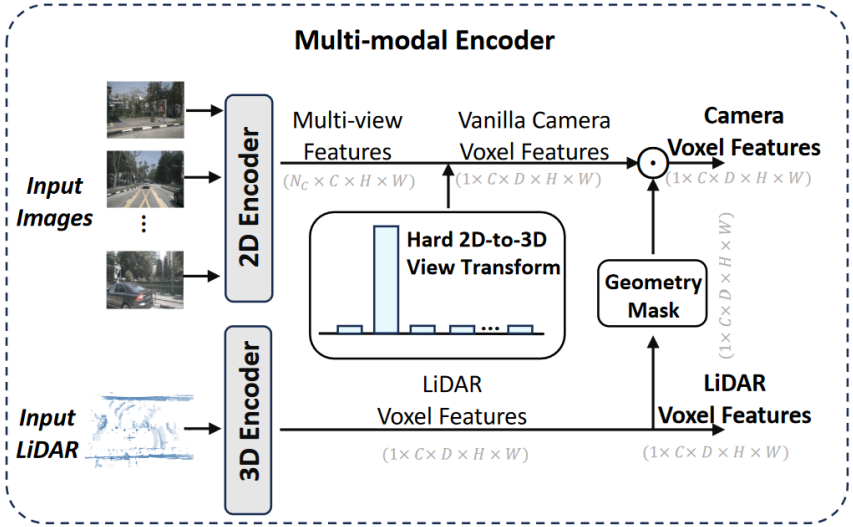
(a) The architecture of multi-modal encoder.

$$W = \mathcal{G}_{\mathrm{C}}\left(\left[\mathcal{G}_{\mathrm{C}}\left(F_p\right), \mathcal{G}_{\mathrm{C}}\left(F_c\right)\right]\right),$$
$$F_m = \sigma(W) \odot F_p + (1 - \sigma(W)) \odot F_c$$

Different from the previous 2D-to-3D view transformation [30, 33, 36, 41] methods that estimate the probabilistic of a set of discrete depths, OccGen proposes a hard 2D-to-3D view transformation to guarantee more accurate depth. We opt for predicting a one-hot vector for depth, as opposed to utilizing softmax on discrete depth values when lifting each image individually into a frustum of features for each camera. However, obtaining one-hot encoding directly through $argmax$ operation is non-differentiable. To address this issue, we propose using Gumbel-Softmax [20] to convert the predicted depth into one-hot encoding.

OCCGen : Generative Multi-modal 3D Occupancy PPrediction for Autonomous Driving

Guoqing Wang, Zhongdao Wang, Pin Tang, Jilai Zheng, Xiangxuan Ren, Bailan Feng, Chao Ma



(a) The architecture of multi-modal encoder.

Different from the previous 2D-to-3D view transformation [30, 33, 36, 41] methods that estimate the probabilistic of a set of discrete depths, OccGen proposes a hard 2D-to-3D view transformation to guarantee more accurate depth. We opt for predicting a one-hot vector for depth, as opposed to utilizing softmax on discrete depth values when lifting each image individually into a frustum of features for each camera. However, obtaining one-hot encoding directly through $argmax$ operation is non-differentiable. To address this issue, we propose using Gumbel-Softmax [20] to convert the predicted depth into one-hot encoding.

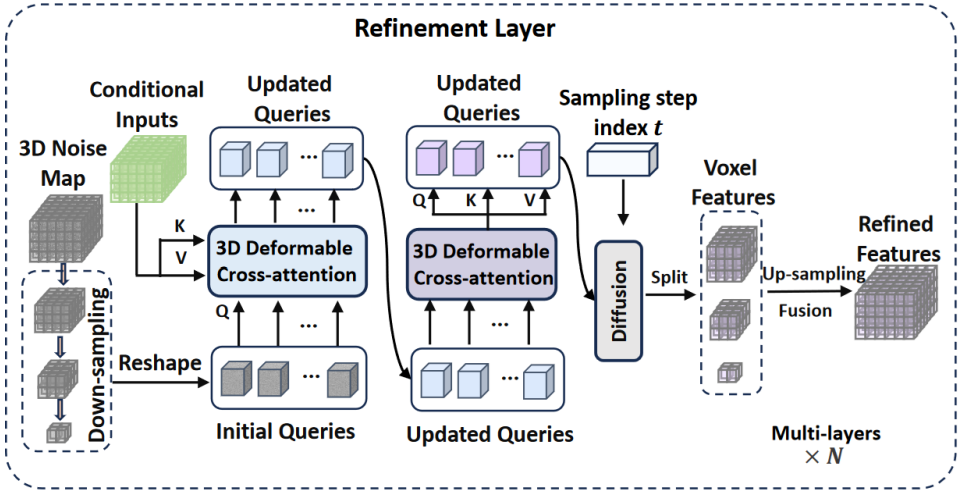
$$W = \mathcal{G}_{\mathrm{C}}\left(\left[\mathcal{G}_{\mathrm{C}}\left(F_p\right), \mathcal{G}_{\mathrm{C}}\left(F_c\right)\right]\right),$$

$$F_m = \sigma(W) \odot F_p + (1 - \sigma(W)) \odot F_c$$

OCCGen : Generative Multi-modal 3D Occupancy PPrediction for Autonomous Driving

Guoqing Wang, Zhongdao Wang, Pin Tang, Jilai Zheng, Xiangxuan Ren, Bailan Feng, Chao Ma

ECCV 2024



(b) The architecture of refinement layer.

Directly operating on the original 3D Gaussian noise map Y_t with high resolution is computationally intensive. Therefore, we first downsample it to obtain smaller multi-scale noise maps $Y_t^i \in \mathbb{R}^{\frac{D}{2^i} \times \frac{H}{2^i} \times \frac{W}{2^i} \times C_i} (i = 1, 2, 3)$. Then, we re-

$$\mathrm{DCA_{3D}}\left(Y_t^i, F_m\right) = \sum_{n \in F_m} \mathrm{DA_{3D}}\left(q, proj(q, n), F_m\right)$$

$$\mathrm{DSA_{3D}}\left(Y_t^i, Y_t^i\right) = \sum_{n \in Y_t^i} \mathrm{DA_{3D}}\left(q, p, \mathbf{Y}_t^i\right)$$

$$Y_t^i := \mathrm{Diff}(Y_t^i, \mathrm{ToEmbed}(t))$$

x N times

김범준

OCCGen : Generative Multi-modal 3D Occupancy PPrediction for Autonomous Driving

Guoqing Wang, Zhongdao Wang, Pin Tang, Jilai Zheng, Xiangxuan Ren, Bailan Feng, Chao Ma

Table 1: Semantic occupancy prediction results on nuScenes-Occupancy validation set. The C, D, L, M denotes **camera, depth, LiDAR** and **multi-modal**. For **Surround**=✓, the method directly predicts surrounding semantic occupancy with 360-degree inputs. Best camera-only, LiDAR-only, and multi-modal results are marked **red**, **blue**, and **black**, respectively.

Method	Input	Surround	IoU	mIoU	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. suf.	other flat	sidewalk	terrain	manmade	vegetation
MonoScene [5]	C	✗	18.4	6.9	7.1	3.9	9.3	7.2	5.6	3.0	5.9	4.4	4.9	4.2	14.9	6.3	7.9	7.4	10.0	7.6
TPVFormer [19]	C	✓	15.3	7.8	9.3	4.1	11.3	10.1	5.2	4.3	5.9	5.3	6.8	6.5	13.6	9.0	8.3	8.0	9.2	8.2
3DSketch [8]	C&D	✗	25.6	10.7	12.0	5.1	10.7	12.4	6.5	4.0	5.0	6.3	8.0	7.2	21.8	14.8	13.0	11.8	12.0	21.2
AICNet [28]	C&D	✗	23.8	10.6	11.5	4.0	11.8	12.3	5.1	3.8	6.2	6.0	8.2	7.5	24.1	13.0	12.8	11.5	11.6	20.2
LMSCNet [42]	L	✓	27.3	11.5	12.4	4.2	12.8	12.1	6.2	4.7	6.2	6.3	8.8	7.2	24.2	12.3	16.6	14.1	13.9	22.2
JS3C-Net [61]	L	✓	30.2	12.5	14.2	3.4	13.6	12.0	7.2	4.3	7.3	6.8	9.2	9.1	27.9	15.3	14.9	16.2	14.0	**24.9**
C-OpenOccupancy [57]	C	✓	19.3	10.3	9.9	6.8	11.2	11.5	6.3	8.4	8.6	4.3	4.2	9.9	22.0	15.8	14.1	13.5	7.3	10.2
L-OpenOccupancy [57]	L	✓	30.8	11.7	12.2	4.2	11.0	12.2	8.3	4.4	8.7	4.0	8.4	10.3	23.5	16.0	14.9	15.7	15.0	17.9
OpenOccupancy [57]	M	✓	29.1	15.1	14.3	12.0	15.2	14.9	13.7	15.0	13.1	9.0	10.0	14.5	23.2	17.5	16.1	17.2	15.3	19.5
C-CONet [57]	C	✓	20.1	12.8	13.2	8.1	**15.4**	17.2	6.3	11.2	10.0	8.3	4.7	12.1	31.4	18.8	18.7	16.3	4.8	8.2
L-CONet [57]	L	✓	30.9	15.8	17.5	**5.2**	13.3	18.1	7.8	5.4	9.6	5.6	13.2	13.6	34.9	21.5	22.4	21.7	19.2	23.5
CONet [57]	M	✓	29.5	20.1	23.3	13.3	21.2	24.3	**15.3**	15.9	18.0	13.3	15.3	20.7	33.2	21.0	22.5	21.5	19.6	23.2
C-OccGen	C	✓	**23.4**	**14.5**	**15.5**	**9.1**	15.3	**19.2**	**7.3**	**11.3**	**11.8**	**8.9**	**5.9**	**13.7**	**34.8**	**22.0**	**21.8**	**19.5**	**6.0**	**9.9**
L-OccGen	L	✓	**31.6**	**16.8**	**18.8**	5.1	**14.8**	**19.6**	**7.0**	**7.7**	**11.5**	**6.7**	**13.9**	**14.6**	**36.4**	**22.1**	**22.8**	**22.3**	**20.6**	24.5
OccGen	M	✓	**30.3**	**22.0**	**24.9**	**16.4**	**22.5**	**26.1**	14.0	**20.1**	**21.6**	**14.6**	**17.4**	**21.9**	**35.8**	**24.5**	**24.7**	**24.0**	**20.5**	**23.5**

OCCGen : Generative Multi-modal 3D Occupancy PPrediction for Autonomous Driving

Guoqing Wang, Zhongdao Wang, Pin Tang, Jilai Zheng, Xiangxuan Ren, Bailan Feng, Chao Ma

Table 2: Semantic Scene Completion results on SemanticKITTI [2] validation set. † denotes the results provided by MonoScene [5].

Method	IoU	road. (%)	sidewalk. (%)	parking. (%)	otherground. (%)	building. (%)	car. (%)	truck. (%)	bicycle. (%)	motorcycle. (%)	othervehicle. (%)	vegetation. (%)	trunk. (%)	terrain. (%)	person. (%)	bicyclist. (%)	motorcyclist. (%)	fence. (%)	pole. (%)	trafficsign. (%)	mIoU
LMSCNet† [42]	28.61	40.68	18.22	4.38	0.00	10.31	18.33	0.00	0.00	0.00	0.00	13.66	0.02	20.54	0.00	0.00	0.00	1.21	0.00	0.00	6.70
AICNet† [28]	29.59	43.55	20.55	11.97	0.07	12.94	14.71	4.53	0.00	0.00	0.00	15.37	2.90	28.71	0.00	0.00	0.00	2.52	0.06	0.00	8.31
JS3C-Net† [61]	38.98	50.49	23.74	11.94	0.07	15.03	24.65	4.41	0.00	0.00	6.15	18.11	4.33	26.86	0.67	0.27	0.00	3.94	3.77	1.45	10.31
MonoScene [5]	37.12	57.47	27.05	15.72	**0.87**	14.24	23.55	7.83	0.20	0.77	3.59	18.12	2.57	30.76	1.79	1.03	0.00	6.39	4.11	2.48	11.50
TPVFormer [19]	35.61	56.50	25.87	20.60	0.85	13.88	23.81	8.08	0.36	0.05	4.35	16.92	2.26	30.38	0.51	0.89	0.00	5.94	3.14	1.52	11.36
VoxFormer [29]	44.02	54.76	26.35	15.50	0.70	17.65	25.79	5.63	0.59	0.51	3.77	24.39	**5.08**	29.96	1.78	3.32	0.00	**7.64**	7.11	4.18	12.35
OccFormer [65]	36.50	58.85	26.88	19.61	0.31	14.40	25.09	**25.53**	0.81	1.19	8.52	19.63	3.93	**32.62**	2.78	2.82	0.00	5.61	4.26	2.86	13.46
Symphonize [23]	41.44	55.78	26.77	14.57	0.19	**18.76**	**27.23**	15.99	**1.44**	2.28	9.52	**24.50**	4.32	28.49	3.19	**8.09**	0.00	6.18	**8.99**	**5.39**	13.44
OccGen (ours)	36.87	**61.28**	**28.30**	20.42	0.43	14.49	26.83	15.49	**1.60**	**2.53**	**12.83**	20.04	3.94	32.44	**3.20**	3.37	0.00	6.94	4.11	2.77	**13.74**

OCCGen : Generative Multi-modal 3D Occupancy PPrediction for Autonomous Driving

Guoqing Wang, Zhongdao Wang, Pin Tang, Jilai Zheng, Xiangxuan Ren, Bailan Feng, Chao Ma

	Encoder	Decoder	IoU	mIoU
(a)	-	-	28.1	20.4
(b)	✓	-	28.6	20.7
(c)	-	✓	30.1	21.6
(d)	✓	✓	30.3	22.0

Hard LSS	Geo. Mask	IoU	mIoU
-	-	29.8	21.4
✓	-	30.2	21.5
-	✓	30.3	21.6
✓	✓	30.3	22.0

Models	Latency(ms)	IoU	mIoU
C-Baseline [57]	172.4	19.3	10.3
C-CONet [57]	285.7	20.1	12.8
C-OccGen(step1)	294.1	23.0	14.2
C-OccGen(step2)	312.5	23.3	14.4
Baseline [57]	243.9	29.1	15.1
CONet [57]	344.8	29.5	20.1
OccGen(step1)	357.1	29.3	21.7
OccGen(step2)	400.0	29.7	21.8

performance is further boosted to 21.8% and 14.4% on the multi-modal and camera-only benchmarks, at a loss of 20 $\sim$ 50 ms. These results show that OccGen can progressively refine the output occupancy multiple times with reasonable time cost.

OCCGen : Generative Multi-modal 3D Occupancy PPrediction for Autonomous Driving

Guoqing Wang, Zhongdao Wang, Pin Tang, Jilai Zheng, Xiangxuan Ren, Bailan Feng, Chao Ma

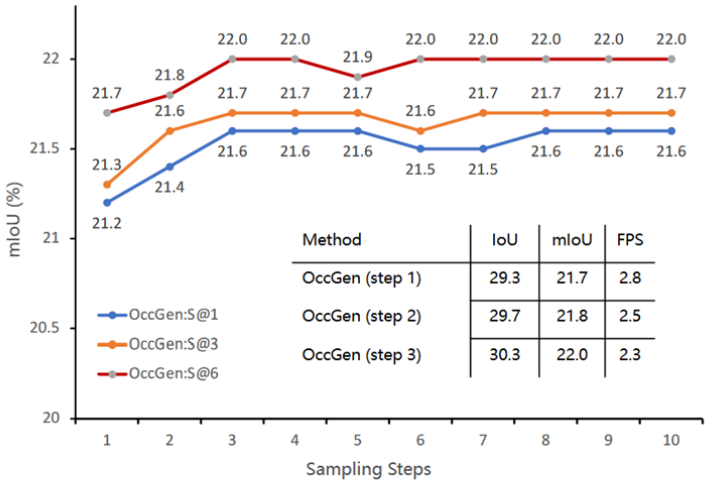


Fig. 5: The results of multiple inferences on nuScenese-Occupancy under multi-modal setting.

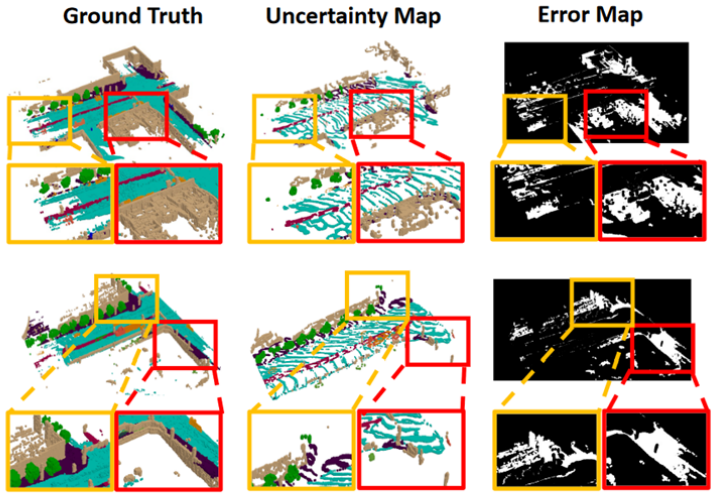


Fig. 6: The visualization of uncertainty map and error map on nuScenese-Occupancy under multi-modal setting.

OCCGen : Generative Multi-modal 3D Occupancy PPrediction for Autonomous Driving

Guoqing Wang, Zhongdao Wang, Pin Tang, Jilai Zheng, Xiangxuan Ren, Bailan Feng, Chao Ma

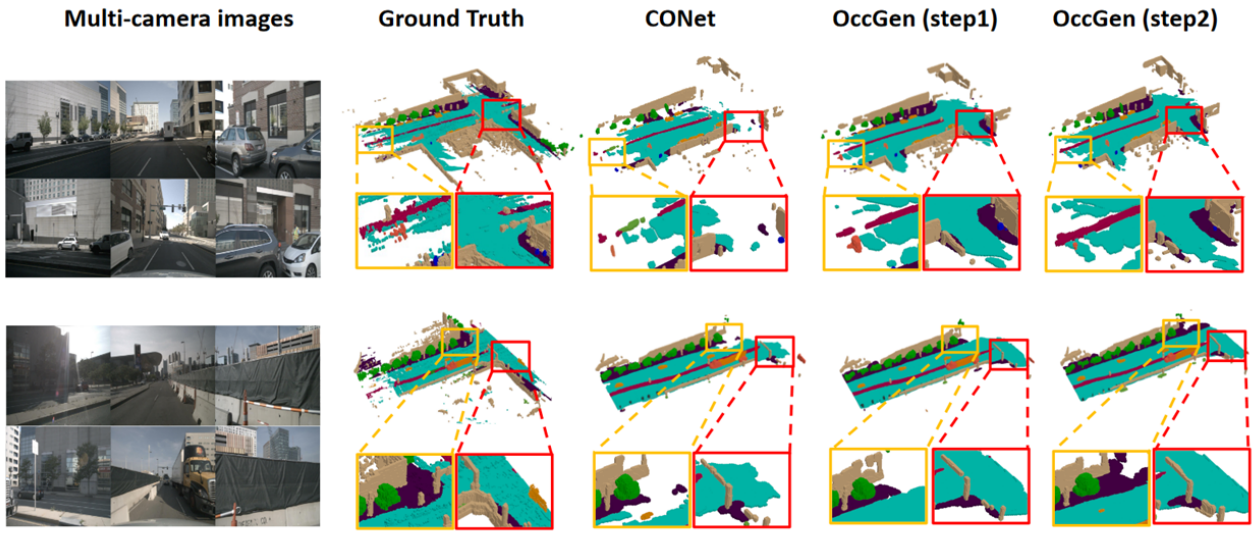


Fig. 4: Qualitative results of the 3D semantic occupancy predictions on nuScenes-Occupancy. The leftmost column shows the input surrounding images, and the following four columns visualize the 3D semantic occupancy results from the ground truth, CONet [57], OccGen(step1), and OccGen(step2). The regions highlighted by rectangles indicate that these areas have obvious differences (better viewed when zoomed in).