SQS: Enhancing Sparse Perception Models via Query-based Splatting in Autonomous Driving

Haiming Zhang [1,2], Yiyao Zhu [3*], Wending Zhou [1,2], Xu Yan [4†],*
Yingjie Cai [4], Bingbing Liu [4], Shuguang Cui [2,1], Zhen Li [2,1†]

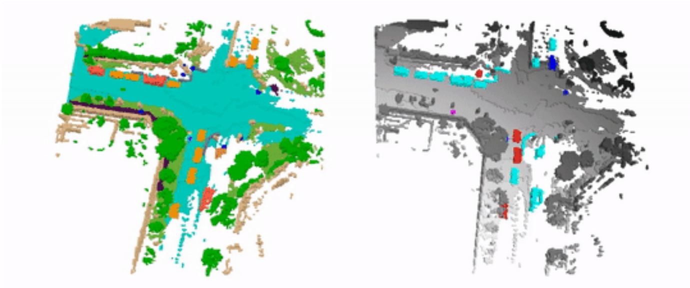
[1] The Future Network of Intelligence Institute, The Chinese University of Hong Kong (Shenzhen),
[2] School of Science and Engineering, The Chinese University of Hong Kong (Shenzhen),
[3] HKUST,
[4] Huawei Noah's Ark Lab

Background

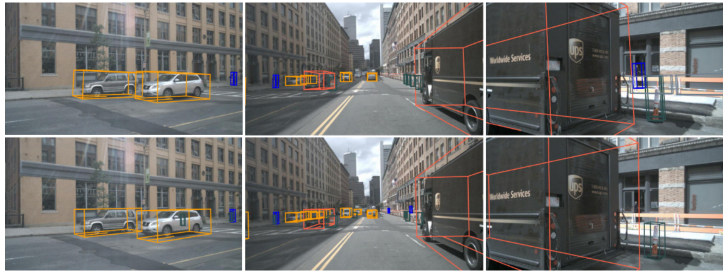
Vision-centric 3D Perception Tasks:

- **Inputs**: Multi-view camera images
- **Outputs**: 3D bounding boxes (3D object detection), 3D semantic occupancy, map segmentation

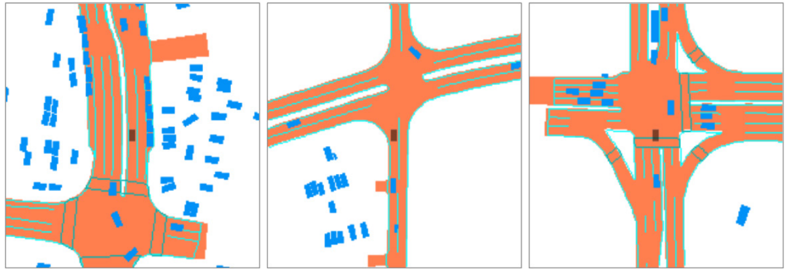
Multi-view images



3D semantic occupancy prediction



3D object detection



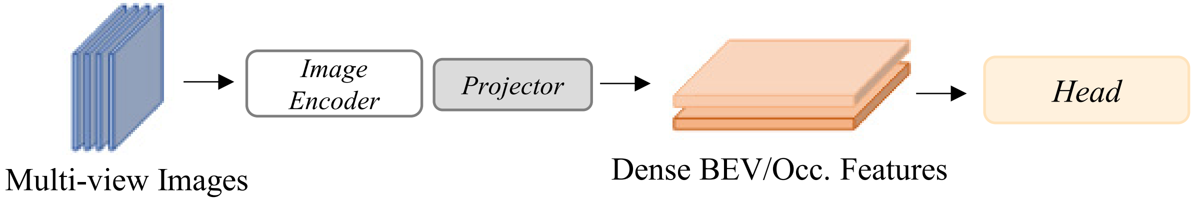
Map segmenation

Background

Existing Vision-centric 3D Perception Paradigms:

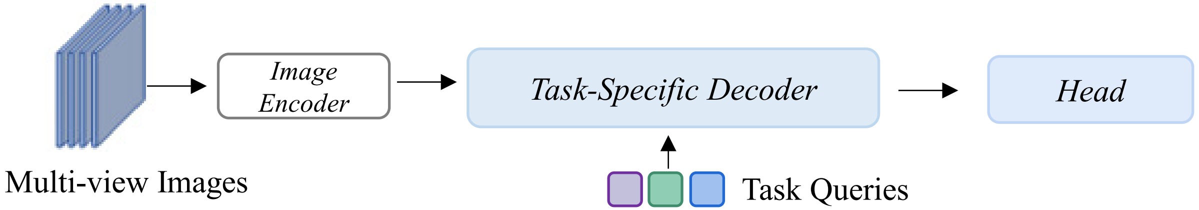
Paradigms

- Dense BEV/Volume-centric



BEVFormer, BEVDet, OccFormer etc.

- Sparse query-centric



Sparse4D, SparseBEV, SparseOcc, OPUS, etc.

Challenges

Pre-training is an effective method to enhance model performance.

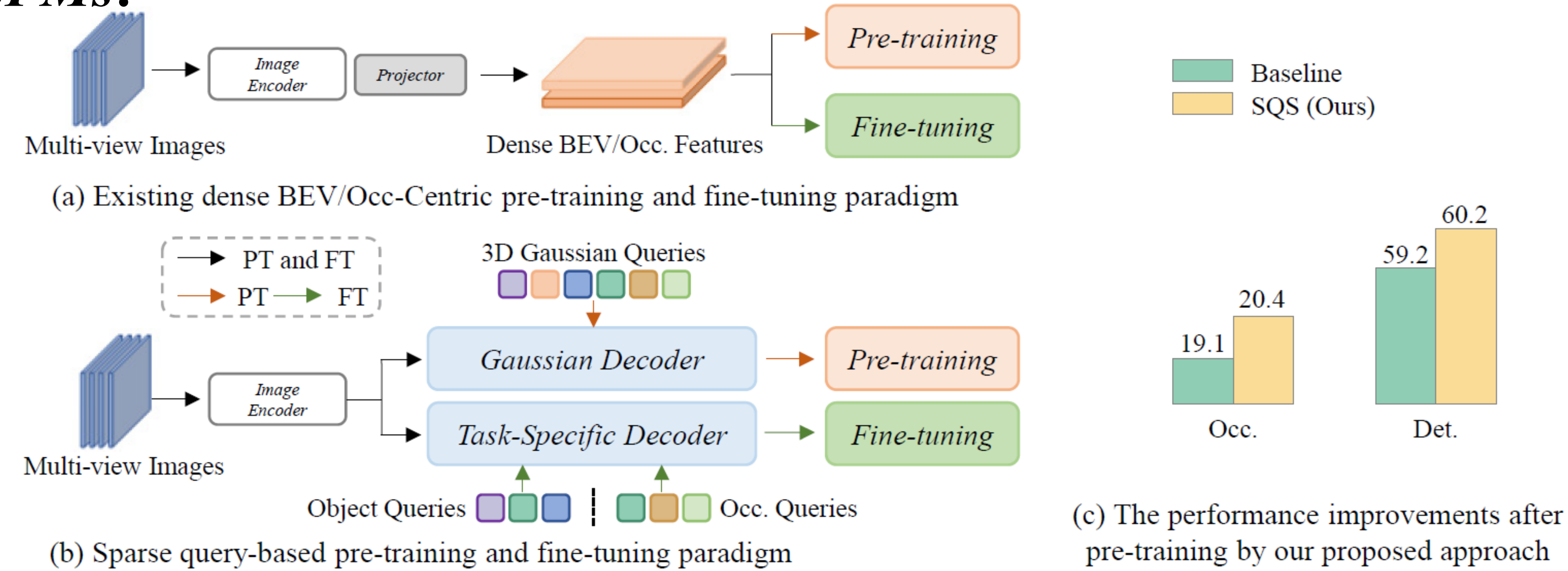
Pre-training in AD
- Contrastive-based
- MAE-based
- Rendering-based

However:

- All existing pre-training approaches for AD operate on **dense BEV or Occupancy** representations (UniPAD, GaussianPretrain, VisionPAD, etc.);
- The queries in **Sparse Perception Models** (**SPMs**) for different tasks play **various roles,** causing difficult to find a **unified** pre-training paradigm for them;

Motivation

Could we design a unified self-supervised pre-training paradigm to enhance different SPMs?

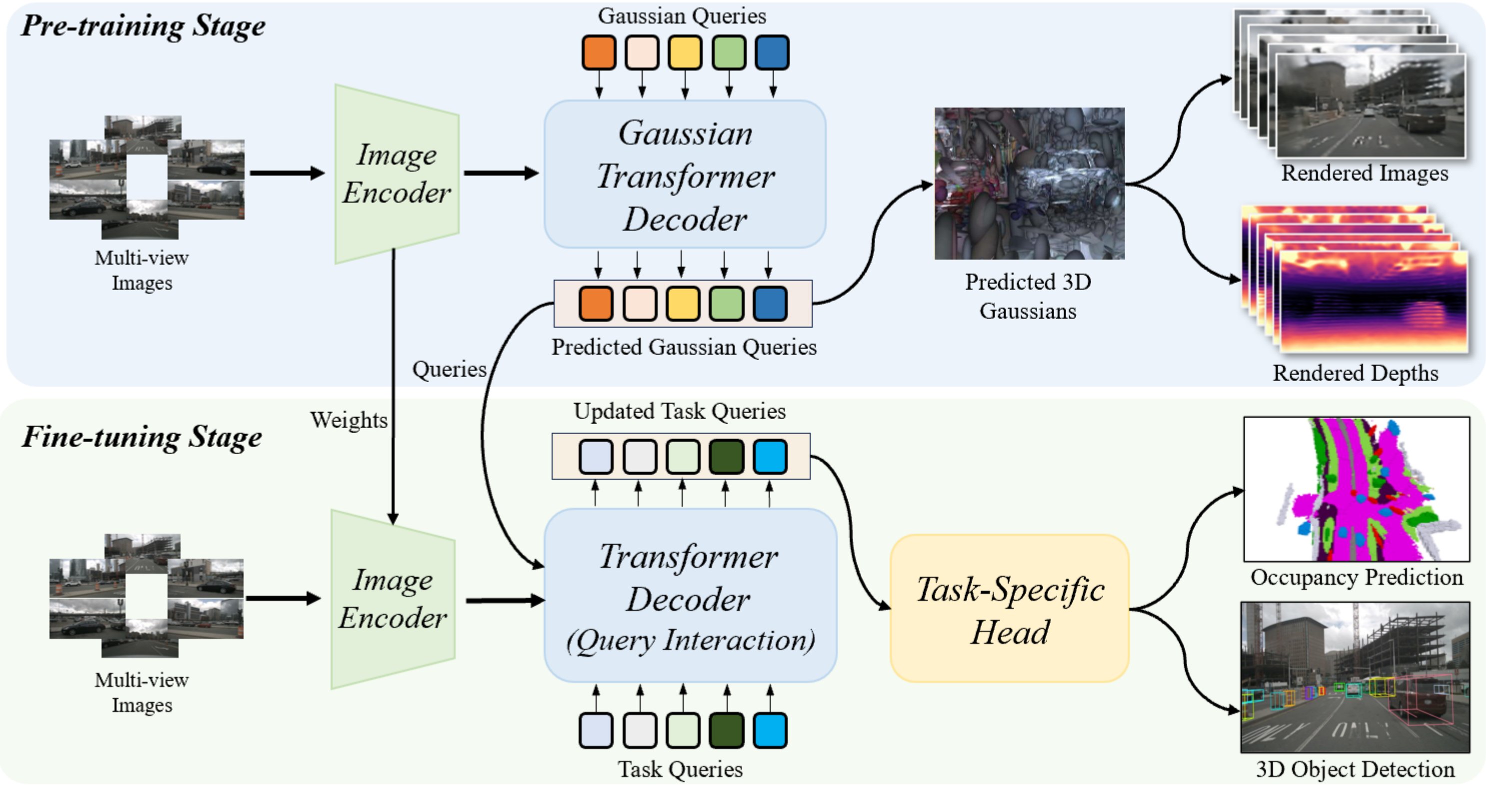


- The proposed **SQS (Sparse Query-based Splatting)** can be integrated into **any sparse query-based perception model**, accepting Gaussian queries for pre-training and utilizing them for prediction;
- We demonstrate the **effectiveness** of SQS on **query-based** 3D semantic occupancy prediction (Occ.) and 3D object detection (Det.) tasks.

SQS

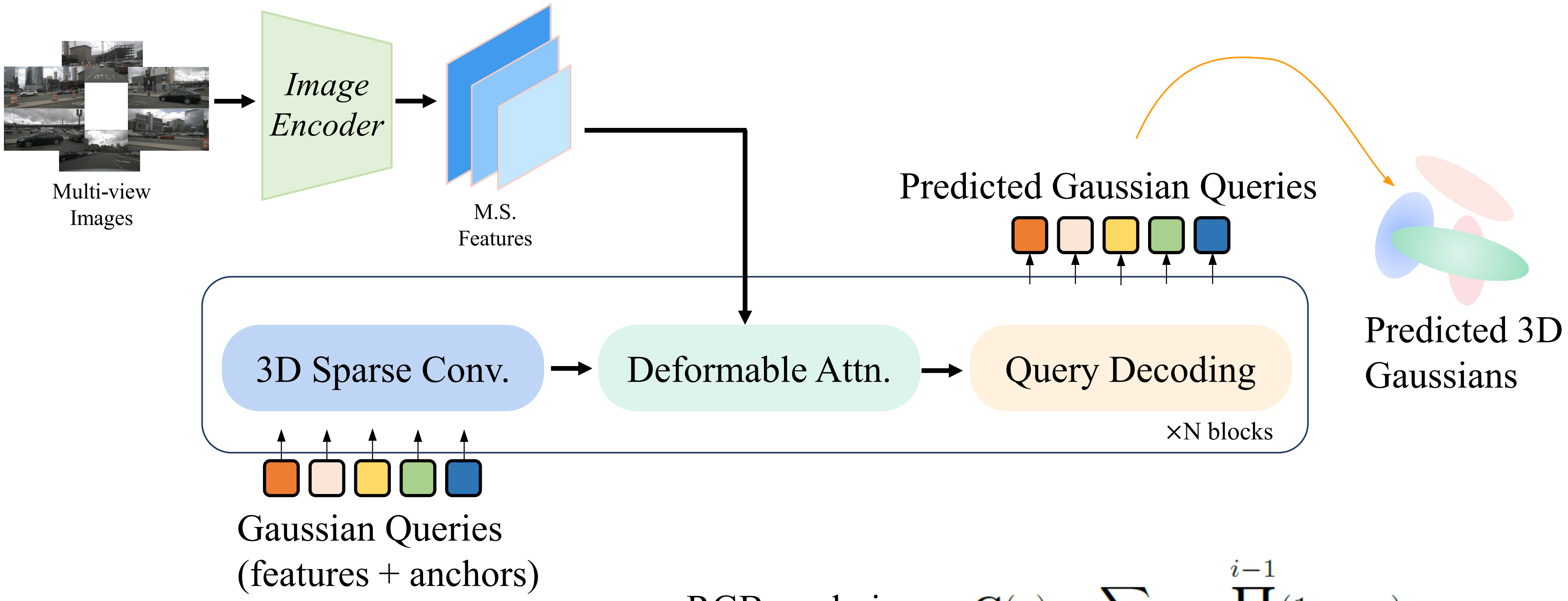
NEURAL INFORMATION PROCESSING SYSTEMS

Framework



- A **sparse query-based 3D Gaussian Splatting pre-training** paradigm with RGB image and depth as supervision;
- A **query interaction module** to fully exploit the knowledge encapsulated in the pre-trained queries;
- The light-weight pre-training paradigm can be **plugged** into **any** sparse query-based downstream tasks to enhance their performance.

Gaussian Transformer Decoder and Gaussian Queries



$\Delta\mu$	center offset
s	scale
r	rotation
α	opacity
c	color

Properties for each 3D Gaussian

RGB rendering: $\mathbf{C}(p) = \sum c_i \alpha_i \prod^{i-1} (1 - \alpha_i),$

Depth rendering: $\mathbf{D}(p) = \sum_{i \in K} d_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j),$

SQS

Pre-training Loss

$$\mathcal{L} = \omega_1 \boxed{\mathcal{L}_{\text{rgb}}} + \omega_2 \boxed{\mathcal{L}_{\text{depth}}},$$

L1 loss for reconstructed RGB images

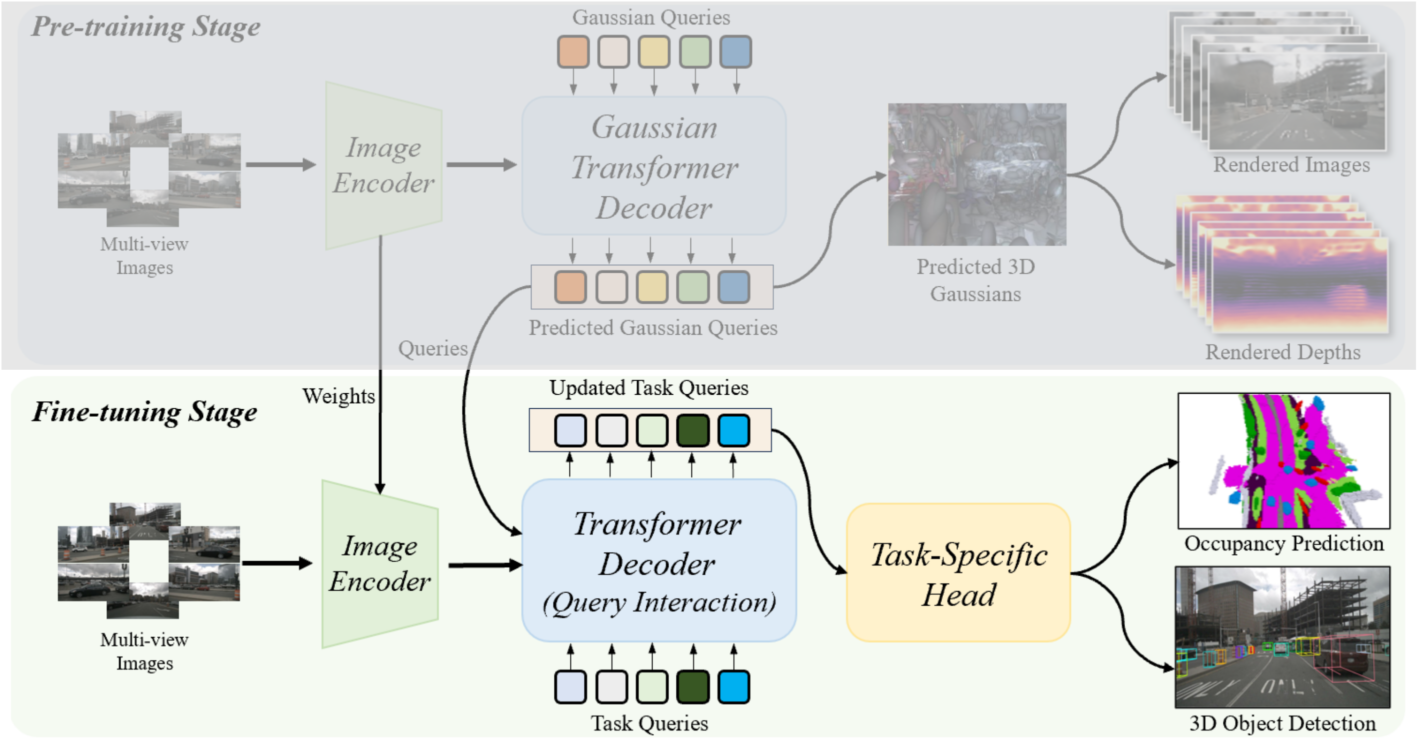
L1 loss for predicted depth maps

where ω_1 and ω_2 are set to 1.0 and 0.05, respectively

- GT depth is obtained from LiDAR projected points with considering valid LiDAR measurements only;
- After pre-training, the image encoder has been enhanced, and the learned queries knowledge can be transferred to downstream models.

NEURAL INFORMATION PROCESSING SYSTEMS

Query Interaction for Fine-tuning



$$q_t = \mathrm{LocalAttn}(q_t + \mathrm{MLP}(\mu_t), q_k + \mathrm{MLP}(g_k)).$$

- The queries in different SPMs play various roles, so it's hard to share the queries in the pre-training stage;
- Therefore, we propose a plug-in framework based on Query Interaction operation;

Experiments

Datasets and Tasks

- nuScenes: 3D object detection task
- Dense occupancy annotations from SurroundOcc: 3D semantic occupancy prediction task

Metrics

- **3D object detection**: nuScenes Detection Score (NDS) and mean Average Precision (mAP)
- **3D occupancy prediction**: mean Intersection-over-Union (mIoU), Intersection-over-Union (IoU);

Finetune Settings

- We strictly follow the official training configurations during fine-tuning without any modifications.

Results

3D semantic occupancy prediction results on SurroundOcc validation set

Table 1: **3D semantic occupancy prediction results on the SurroundOcc `val` set.** While the original TPVFormer [15] is trained with LiDAR segmentation labels, TPVFormer* is supervised by dense occupancy annotations.

Method	SC IoU	SSC mIoU	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. suf.	other flat	sidewalk	terrain	manmade	vegetation
MonoScene [4]	23.96	7.31	4.03	0.35	8.00	8.04	2.90	0.28	1.16	0.67	4.01	4.35	27.72	5.20	15.13	11.29	9.03	14.86
Atlas [40]	28.66	15.00	10.64	5.68	19.66	24.94	8.90	8.84	6.47	3.28	10.42	16.21	34.86	15.46	21.89	20.95	11.21	20.54
BEVFormer [24]	30.50	16.75	14.22	6.58	23.46	28.28	8.66	10.77	6.64	4.05	11.20	17.78	37.28	18.00	22.88	22.17	13.80	22.21
TPVFormer [15]	11.51	11.66	16.14	7.17	22.63	17.13	8.83	11.39	10.46	8.23	9.43	17.02	8.07	13.64	13.85	10.34	4.90	7.37
TPVFormer* [15]	30.86	17.10	15.96	5.31	23.86	27.32	9.79	8.74	7.09	5.20	10.97	19.22	38.87	21.25	24.26	23.15	11.73	20.81
OccFormer [70]	31.39	19.03	18.65	10.41	23.92	30.29	10.31	14.19	13.59	10.13	12.49	20.77	38.78	19.79	24.19	22.21	13.48	21.35
SurroundOcc [57]	31.49	20.30	20.59	11.68	28.06	30.86	10.70	15.14	14.09	12.06	14.38	22.26	37.29	23.70	24.49	22.77	14.89	21.86
GaussianFormer [16]	29.83	19.10	19.52	11.26	26.11	29.78	10.47	13.83	12.58	8.67	12.74	21.57	39.63	23.28	24.46	22.99	9.59	19.12
GaussianFormer + SQS (Ours)	**31.52**	**20.40**	19.98	11.86	28.21	30.68	10.87	15.03	14.28	9.57	14.74	22.98	39.82	23.88	25.46	23.09	14.56	21.31

Results

3D object detection results on nuScenes validation set

Table 2: **3D object detection results on the nuScenes `val` split.** † benefits from perspective pre-training [31]. ‡ indicates methods with CBGS [73] which will elongate 1 epoch into 4.5 epochs.

Method	Backbone	Input Size	Epochs	NDS	mAP	mATE	mASE	mAOE	mAVE	mAAE
PETRv2 [34]	ResNet50	704 × 256	60	45.6	34.9	0.700	0.275	0.580	0.437	0.187
BEVStereo [21]	ResNet50	704 × 256	90 ‡	50.0	37.2	0.598	0.270	0.438	0.367	0.190
BEVPoolv2 [12]	ResNet50	704 × 256	90 ‡	52.6	40.6	0.572	0.275	0.463	0.275	0.188
SOLOFusion [43]	ResNet50	704 × 256	90 ‡	53.4	42.7	0.567	0.274	0.511	0.252	0.181
Sparse4Dv2 [28]	ResNet50	704 × 256	100	53.9	43.9	0.598	0.270	0.475	0.282	0.179
StreamPETR † [53]	ResNet50	704 × 256	60	55.0	45.0	0.613	0.267	0.413	0.265	0.196
SparseBEV [31]	ResNet50	704 × 256	36	54.5	43.2	0.606	0.274	0.387	0.251	0.186
SparseBEV † [31]	ResNet50	704 × 256	36	55.8	44.8	0.581	0.271	0.373	0.247	0.190
SparseBEV † + SQS (Ours)	ResNet50	704 × 256	36	56.6	45.2	0.564	0.263	0.362	0.232	0.182
Sparse4Dv3 † [29]	ResNet50	704 × 256	100	56.1	46.9	0.553	0.274	0.476	0.227	0.200
Sparse4Dv3 † + SQS (Ours)	ResNet50	704 × 256	100	**56.9**	**47.4**	0.542	0.266	0.458	0.218	0.191
DETR3D † [55]	ResNet101-DCN	1600 × 900	24	43.4	34.9	0.716	0.268	0.379	0.842	0.200
BEVFormer † [24]	ResNet101-DCN	1600 × 900	24	51.7	41.6	0.673	0.274	0.372	0.394	0.198
BEVDepth [22]	ResNet101	1408 × 512	90 ‡	53.5	41.2	0.565	0.266	0.358	0.331	0.190
Sparse4D † [26]	ResNet101-DCN	1600 × 900	48	55.0	44.4	0.603	0.276	0.360	0.309	0.178
SOLOFusion [43]	ResNet101	1408 × 512	90 ‡	58.2	48.3	0.503	0.264	0.381	0.246	0.207
SparseBEV † [31]	ResNet101	1408 × 512	24	59.2	50.1	0.562	0.265	0.321	0.243	0.195
SparseBEV † + SQS (Ours)	ResNet101	1408 × 512	24	60.2	50.9	0.531	0.251	0.318	0.241	0.185
Sparse4Dv3 † [29]	ResNet101	1408 × 512	100	62.3	53.7	0.511	0.255	0.306	0.194	0.192
Sparse4Dv3 † + SQS (Ours)	ResNet101	1408 × 512	100	**63.1**	**54.4**	0.498	0.241	0.298	0.187	0.188

Results

Data efficiency analysis with limited data

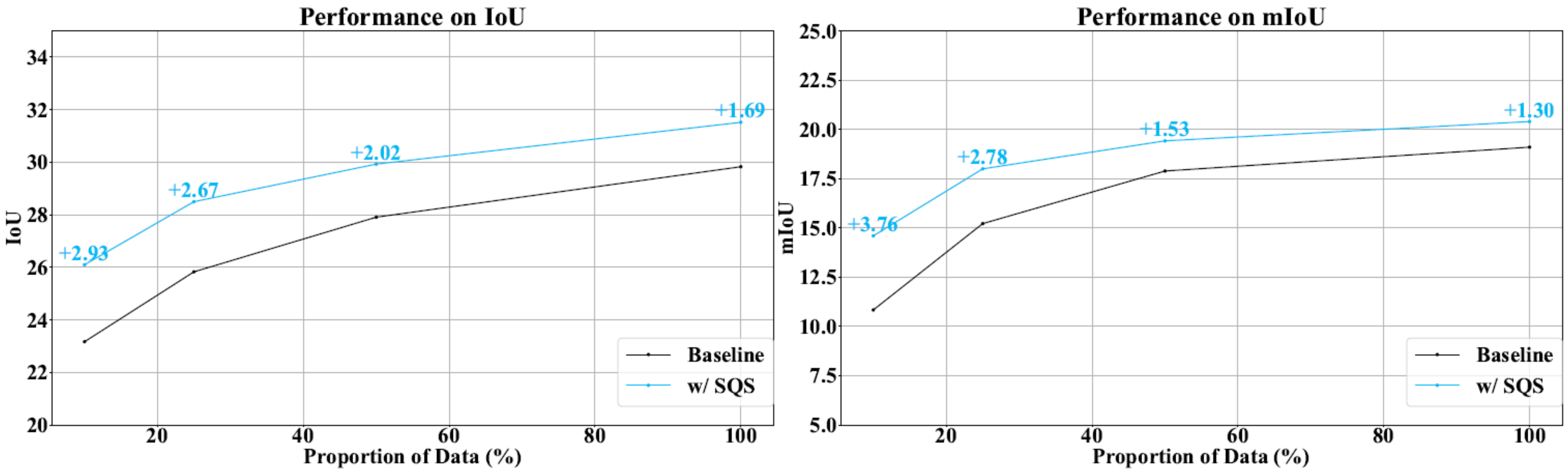


Figure 3: **Data efficiency analysis.** To assess data efficiency under limited annotation scenarios, we reduce the amount of labeled data used for downstream fine-tuning in the 3D semantic occupancy prediction task. The outcomes demonstrate that our pre-training method significantly enhances performance, even when only a small portion of annotations is available.

Ablation

Ablation Study on main designs

Table 3: **Ablation studies.** We report the IoU and mIoU metrics on the SurroundOcc *val* set for the 3D semantic occupancy prediction task. "Rend.", "B.b." and "Inter." denote rendering, image backbone, and query interaction, respectively.

Methods	Rend. RGB	Rend. Depth	Load B.b.	Query Inter.	IoU	mIoU
Baseline [16]					25.8	15.2
Model A	✓		✓		23.8 ↓2.0	12.2 ↓3.0
Model B		✓	✓		27.9 ↑2.1	17.3 ↑2.1
Model C	✓	✓	✓		28.2 ↑2.4	17.5 ↑2.3
Model D	✓	✓		✓	26.3 ↑0.5	15.9 ↑0.7
Model E				✓	25.7 ↓0.1	15.3 ↑0.1
SQS (Ours)	✓	✓	✓	✓	**28.5** ↑2.7	**18.0** ↑2.8

Ablation

NEURAL INFORMATION PROCESSING SYSTEMS

Ablation Study on main designs

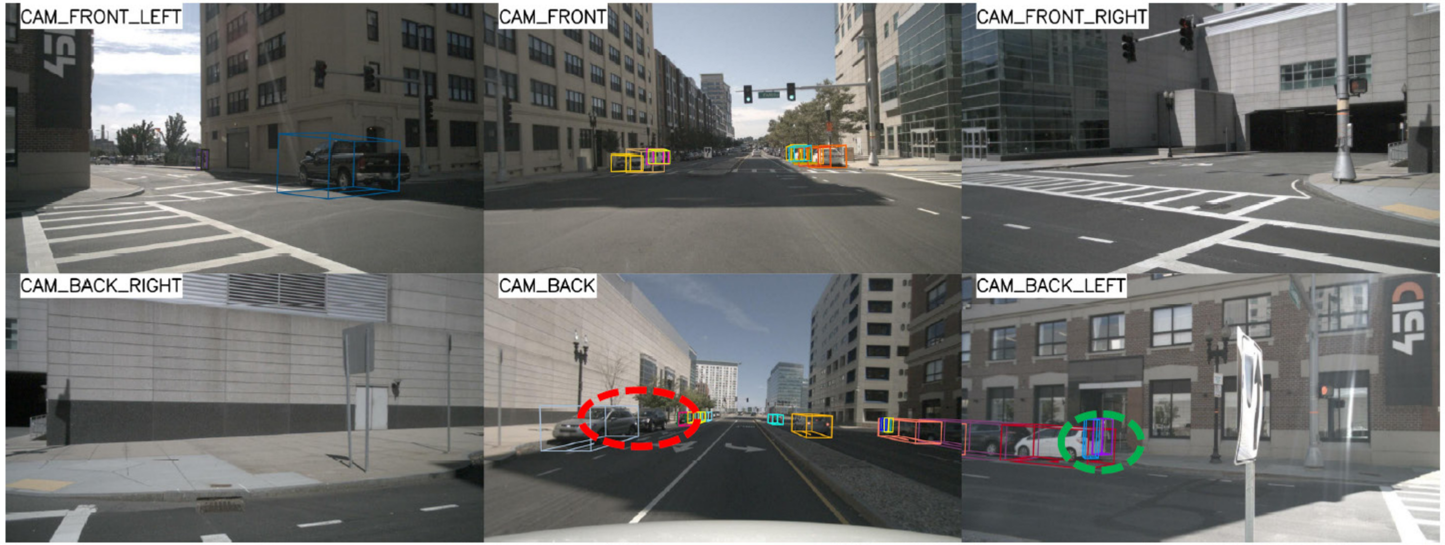
Table 1: **Ablation on the number of Gaussians.** The latency and memory are tested with batch size one during inference.

Number of Gaussians	Latency	Memory	mIoU	IoU
7500	**180** ms	**4615** M	15.2	26.6
12500	198 ms	4650 M	17.8	28.1
25600	210 ms	4680 M	18.0	**28.5**
51200	259 ms	4796 M	**18.4**	28.4
144000	372 ms	5635 M	18.2	28.3

Visualization

3D Object Detection

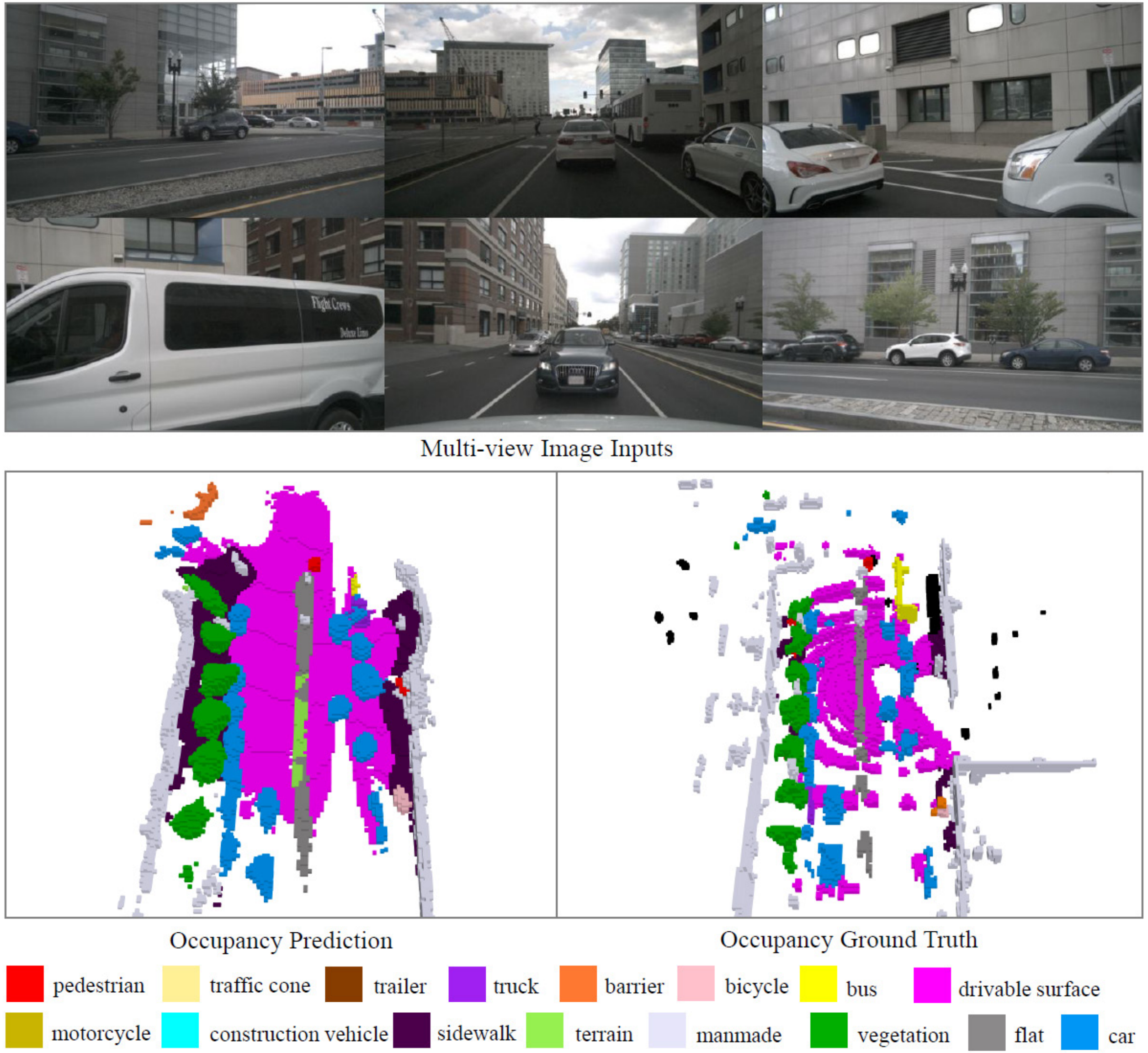
SparseBEV



Ours

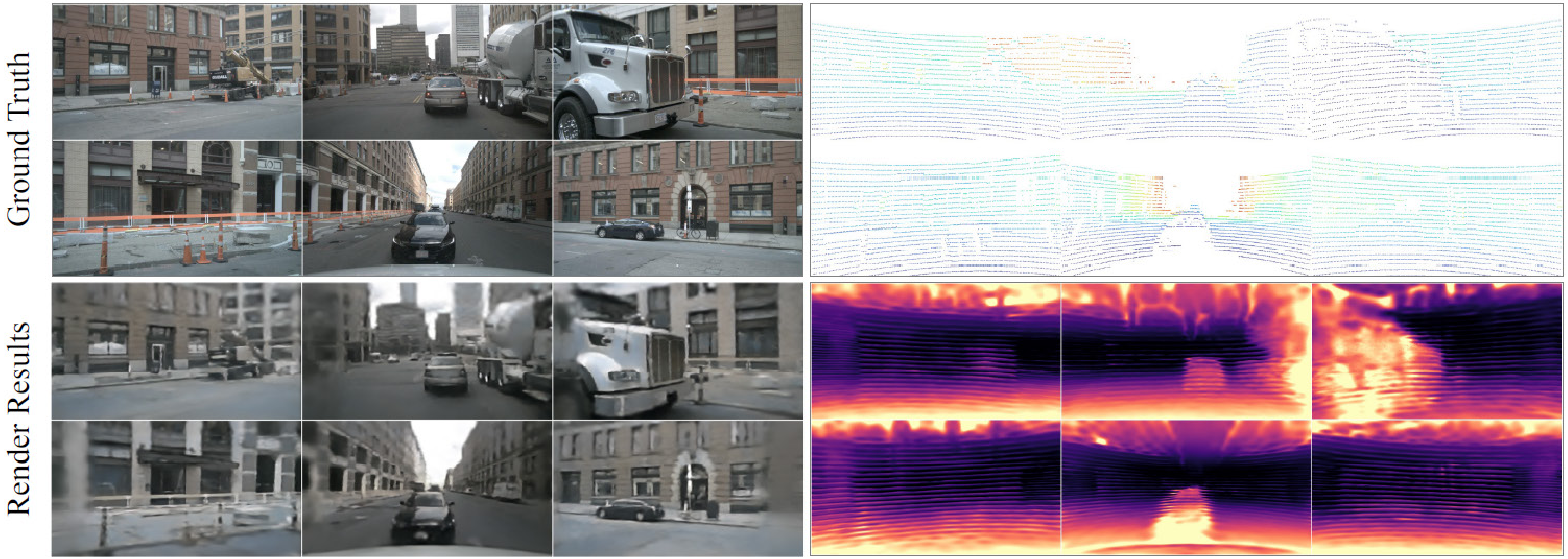
Visualization

3D Occupancy Prediction



Visualization

Rendered Results



Conclusion

NEURAL INFORMATION PROCESSING SYSTEMS

SQS: Enhancing Sparse Perception Models via Query-based Splatting in Autonomous Driving:

- We propose SQS, the **first query-based splatting pre-training** technique specifically designed to advance Sparse Perception Models (**SPMs**);
- We introduce **plug-and-play Gaussian queries**, which learns fine-grained features in a self-supervised manner during pre-training, and further enhances downstream tasks via **interactive feature** fusion during fine-tuning;
- SQS significantly enhances performance in both **occupancy prediction and 3D object detection**, surpassing previous state-of-the-art results on multiple autonomous driving benchmarks.

SQS: Enhancing Sparse Perception Models via Query-based Splatting in Autonomous Driving

Thanks for watching!

Haiming Zhang [1,2], Yiyao Zhu [3*], Wending Zhou [1,2], Xu Yan [4†], Yingjie Cai [4], Bingbing Liu [4], Shuguang Cui [2,1], Zhen Li [2,1†]*

[1] The Future Network of Intelligence Institute, The Chinese University of Hong Kong (Shenzhen),
[2] School of Science and Engineering, The Chinese University of Hong Kong (Shenzhen),
[3] HKUST
[4] Huawei Noah's Ark Lab