



上海交通大学
SHANGHAI JIAO TONG UNIVERSITY



西安交通大学
XI'AN JIAOTONG UNIVERSITY



FreqPDE: Rethinking Positional Depth Embedding for Multi-View 3D Object Detection Transformers

Haisheng Su^{1,3*}, Junjie Zhang^{2,3*}, Feixiang Song³, Sanping Zhou², Wei Wu³, Junchi Yan^{1♡}, Nanning Zheng^{2♡}

(* Equal Contribution ♡ Corresponding Author)

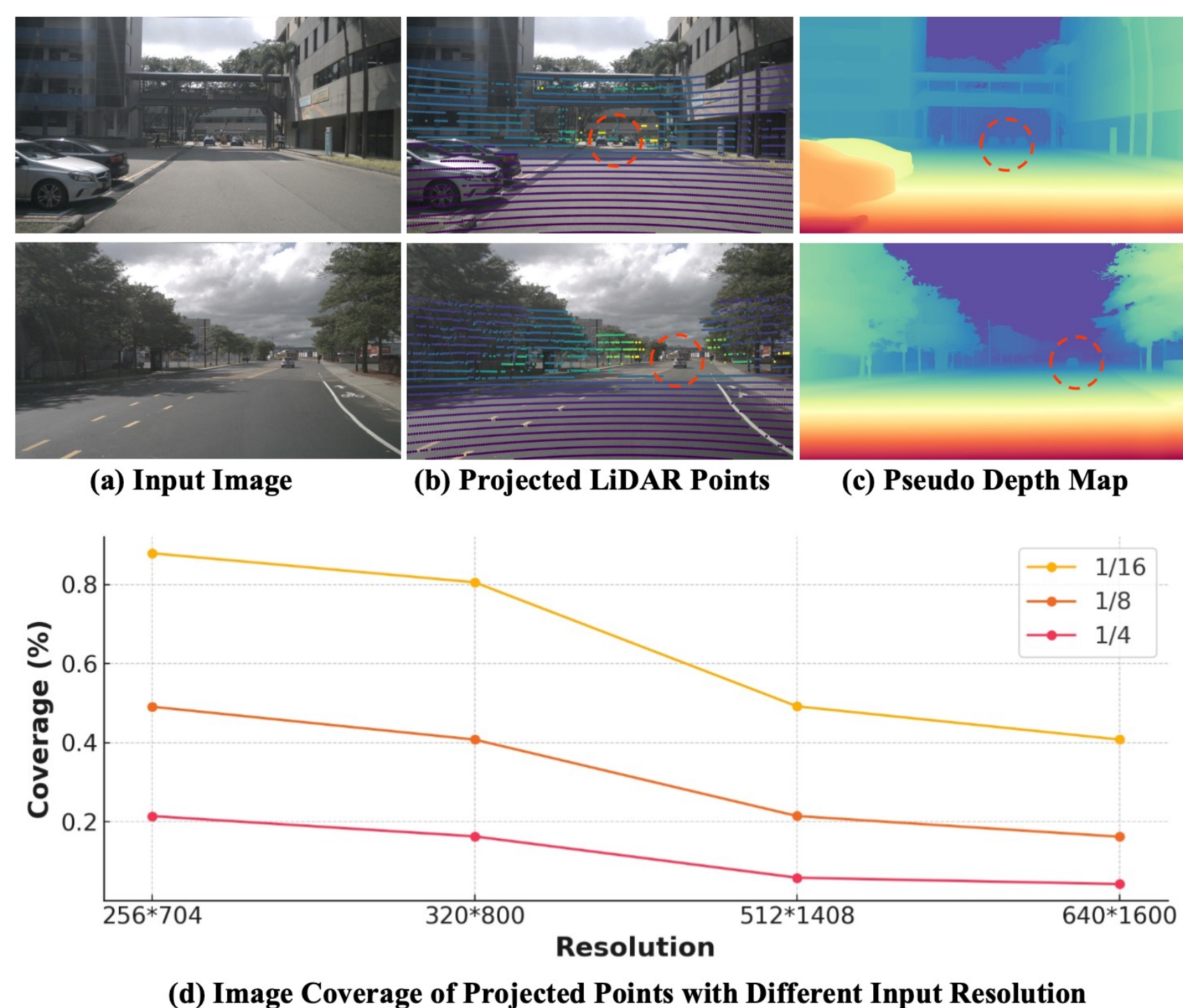
¹Shanghai Jiao Tong University ²Xi'an Jiaotong University ³SenseAuto Research



Motivation and Contribution of FreqPDE

Motivation

- ✓ Depth discontinuity of obj. boundaries and indistinctions of small objects
- ✓ Sparse supervision of projected points
- ✓ High-level image features used for depth prediction
- ✓ View consistency and scale invariance



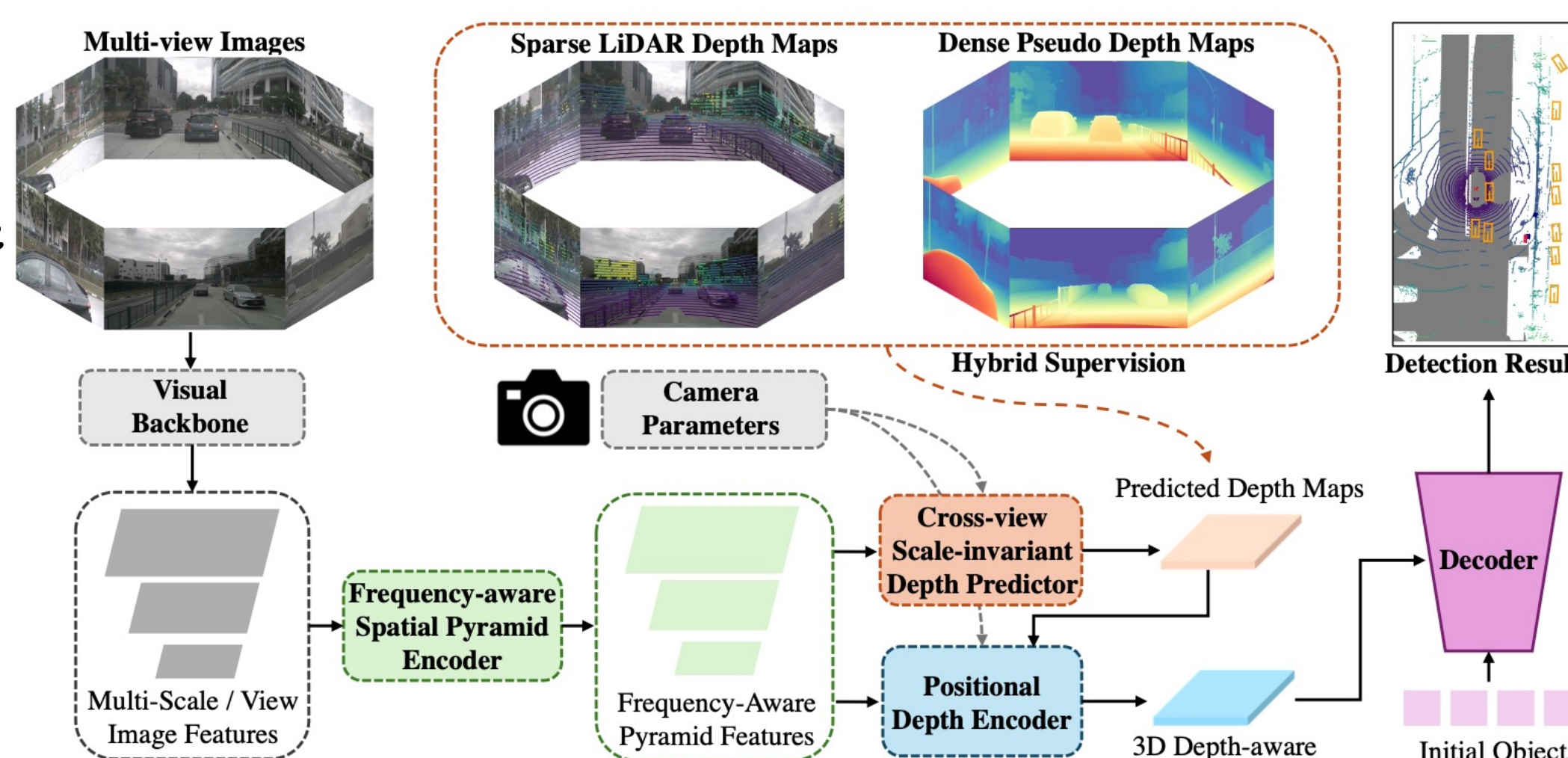
Contribution:

- ✓ FreqPDE (Positional Depth Encoder) for high quality depth prediction, SOTA performance on nuScenes.
- ✓ Frequency-aware Spatial Pyramid Encoder:
 - ✓ Construct a feature pyramid;
 - ✓ Combine high-freq. edge clues and low-freq. semantics from different levels respectively.
- ✓ Cross-view Scale-invariant Depth Predictor:
 - ✓ Estimate pixel-level depth distribution;
 - ✓ Conduct cross-view and channel attention;
 - ✓ Combines 2D image features and 3D position embeddings to generate 3D depth-aware features for query decoding.
- ✓ Hybrid Depth Supervision:
 - ✓ Explicit Metric Supervision: LiDAR points
 - ✓ Implicit Distribution Supervision: DepthAnything

Design of FreqPDE

Framework

1. Multi-view/scale image feature extraction.
2. Frequency-aware feature pyramid construction.
3. Cross-view scale-invariant depth prediction with hybrid supervision from metric and distribution aspects.
4. Positional depth encoder and query decoder for 3D object detection.



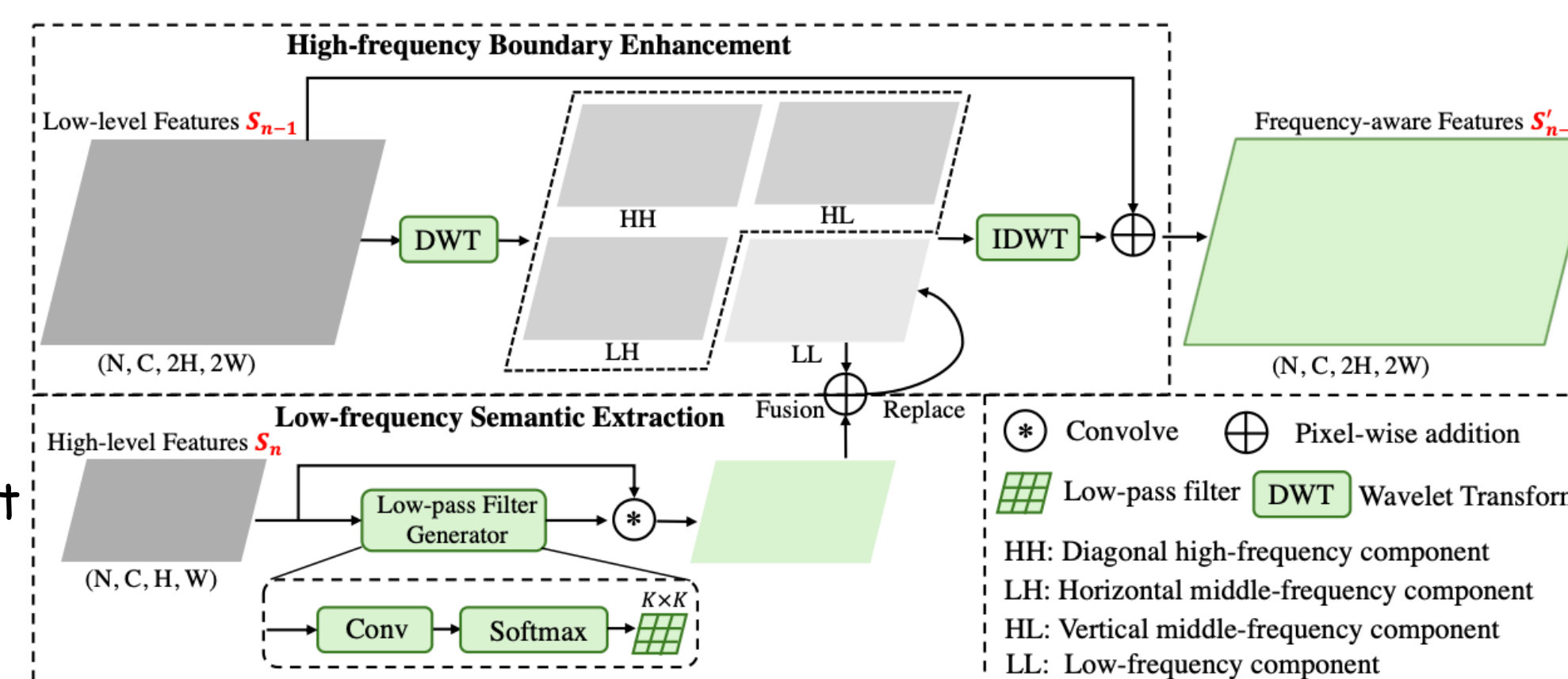
FSPE module

1. Low-freq. Semantic Extraction

$$S'_{n-1} = f_{\text{IDWT}}(f_{\text{lf}}(S_{n-1}), f_{\text{hf}}(S_n)) + S_{n-1},$$

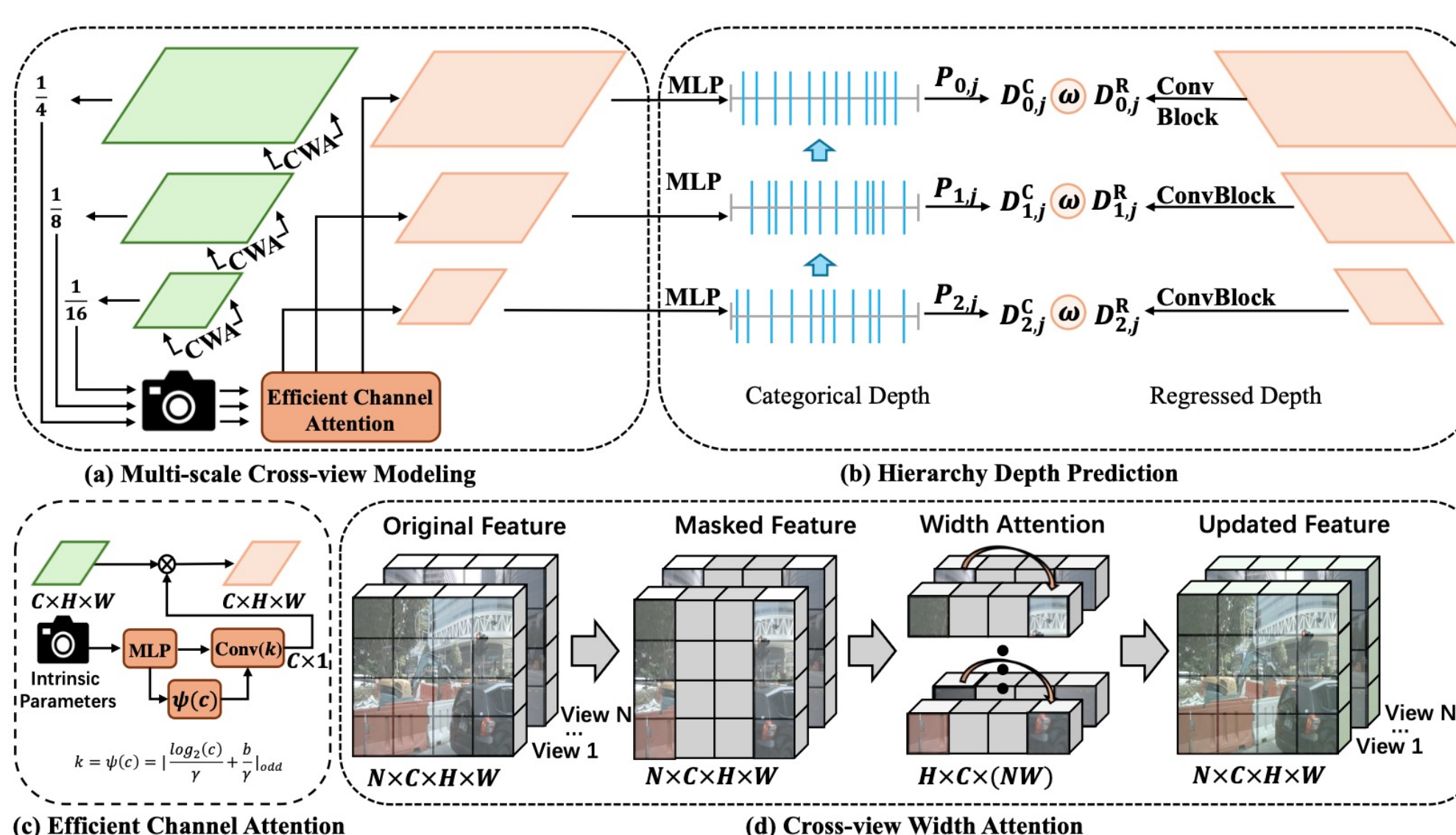
2. High-freq. Boundary Enhancement

$$S'_{n-1} = f_{\text{IDWT}}(LL', (LH, HL, HH)) + S_{n-1}$$



CSDP module

1. Cross-view modeling through cross-view width attention block.
2. Scale-invariant learning through ECANet.
3. Hierarchy depth prediction with hybrid depth supervision



Experimental Results

Quantative Results on nuScenes dataset

Table 1. Comparison of other methods on the nuScenes val set. [†] The backbone benefits from perspective pretraining. The best is in **bold**.

Method	Backbone	Input Size	NDS [†]	mAP [†]	mATE [↓]	mASE [↓]	mAOE [↓]	mAVE [↓]	mAAE [↓]
BevDet4D [16]	ResNet50	256 × 704	45.7	32.2	0.703	0.278	0.495	0.354	0.206
PETrv2 [33]	ResNet50	256 × 704	45.6	34.9	0.700	0.275	0.580	0.437	0.187
BEVStereo [24]	ResNet50	256 × 704	50.0	37.2	0.598	0.270	0.438	0.367	0.190
SOLOFusion [39]	ResNet50	256 × 704	53.4	42.7	0.567	0.274	0.511	0.252	0.181
FreqPDE (Ours)	ResNet50	256 × 704	54.3	43.5	0.577	0.270	0.442	0.257	0.199
Sparse4Dv2 [†] [29]	ResNet50	256 × 704	53.8	43.9	0.598	0.270	0.475	0.282	0.179
BEVFormerv2 [†] [57]	ResNet50	-	52.9	42.3	0.618	0.273	0.413	0.333	0.188
StreamPETR [†] [50]	ResNet50	256 × 704	55.0	45.0	0.613	0.267	0.413	0.265	0.196
SparseBEV [†] [30]	ResNet50	256 × 704	55.8	44.8	0.581	0.271	0.373	0.247	0.190
FreqPDE (Ours) [†]	ResNet50	256 × 704	56.2	46.3	0.575	0.268	0.405	0.245	0.203
StreamPETR [†] [50]	V2-99	320 × 800	57.2	48.2	0.602	0.256	0.372	0.267	0.192
FreqPDE (Ours) [†]	V2-99	320 × 800	58.6	50.6	0.576	0.261	0.375	0.253	0.200
Far3D [†] [19]	ResNet101	512 × 1408	59.4	51.0	0.551	0.258	0.372	0.238	0.195
Sparse4Dv2 [†] [29]	ResNet101	512 × 1408	59.4	50.5	0.548	0.268	0.348	0.239	0.184
StreamPETR [†] [50]	ResNet101	512 × 1408	59.2	50.4	0.569	0.262	0.315	0.257	0.199
FreqPDE (Ours) [†]	ResNet101	512 × 1408	60.1	51.9	0.562	0.258	0.324	0.242	0.196

Table 2. Comparison of other methods on nuScenes test set. These results are reported without test-time augmentation, model ensemble, and any future information. The best is in **bold**.

Method	Backbone	Input Size	NDS [†]	mAP [†]	mATE [↓]	mASE [↓]	mAOE [↓]	mAVE [↓]	mAAE [↓]
BEVDepth [25]	V2-99	640 × 1600	60.0	50.3	0.445	0.245	0.378	0.320	0.126
CAPE-T [55]	V2-99	640 × 1600	61.0	52.5	0.503	0.242	0.361	0.306	0.114
FB-BEV [30]	V2-99	640 × 1600	62.4	53.7	0.439	0.250	0.358	0.270	0.128
HoP [63]	V2-99	640 × 1600	61.2	52.8	0.491	0.242	0.332	0.343	0.109
StreamPETR [50]	V2-99	640 × 1600	63.6	55.0	0.479	0.239	0.317	0.241	0.119
SparseBEV [30]	V2-99	640 × 1600	63.6	55.6	0.485	0.244	0.332	0.246	0.117
Sparse4Dv2 [29]	V2-99	640 × 1600	63.8	55.6	0.462	0.238	0.328	0.264	0.115
Ours	V2-99	640 × 1600	64.2	56.0	0.468	0.241	0.315	0.242	0.115

Qualitative Results on nuScenes dataset

