FrustumFormer: Adaptive Instance-aware Resampling for Multi-view 3D Detection

Yuqi Wang[1,2] Yuntao Chen[3] Zhaoxiang Zhang[1,2,3]

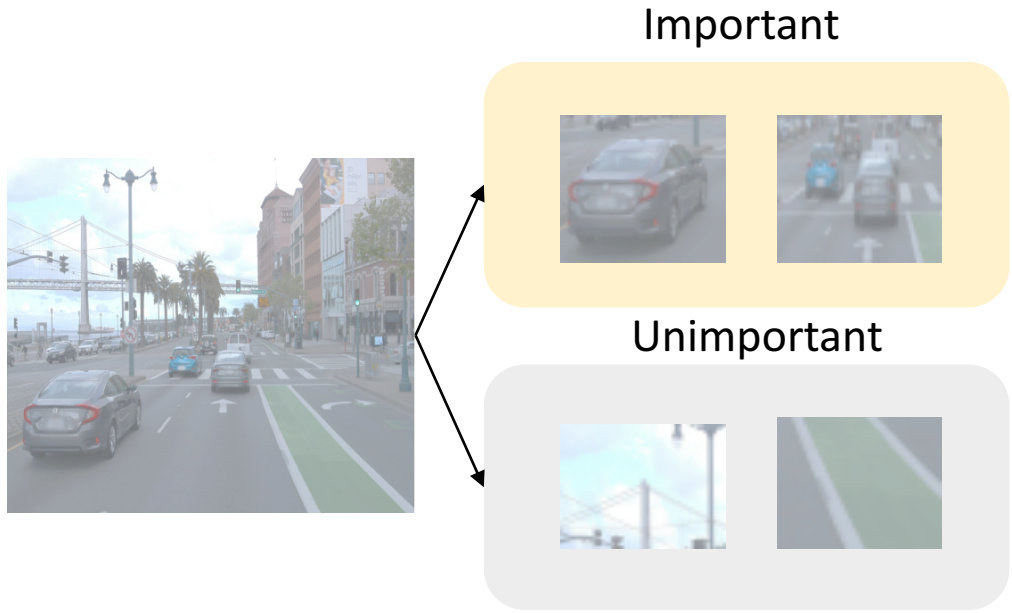
[1] Institute of Automation, Chinese Academy of Sciences (CASIA)

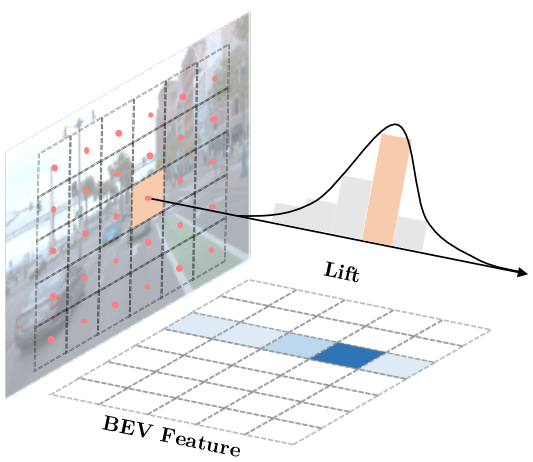
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS)

[3] Centre for Artificial Intelligence and Robotics, HKISI_CAS

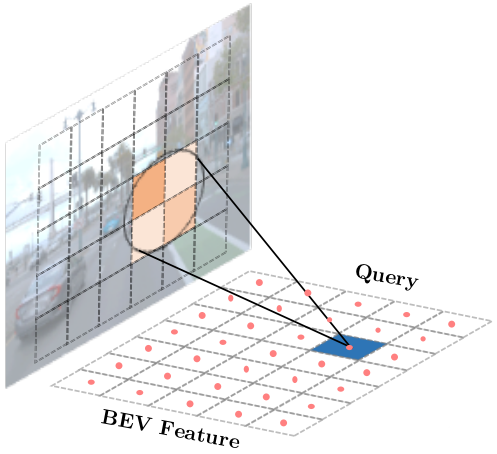
Motivation

- **View Feature Transformation**
  - How to transform view feature? 😋
  - Choosing what to transform? 🤔
- **Adaptive View transformation**
  - Paying more attention to instance regions

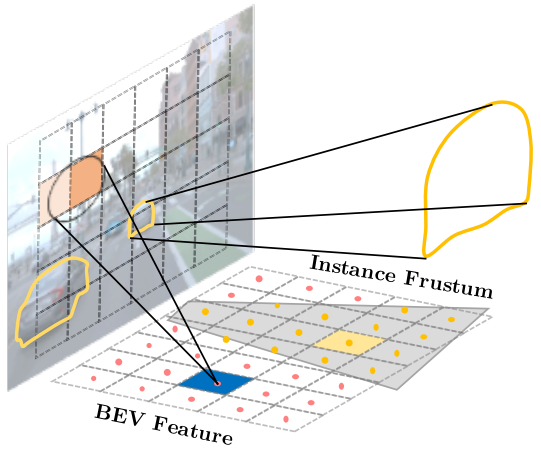




(a) Grid Sampling in Image



(b) Grid Sampling in BEV



(c) Instance-aware Sampling in Frustum

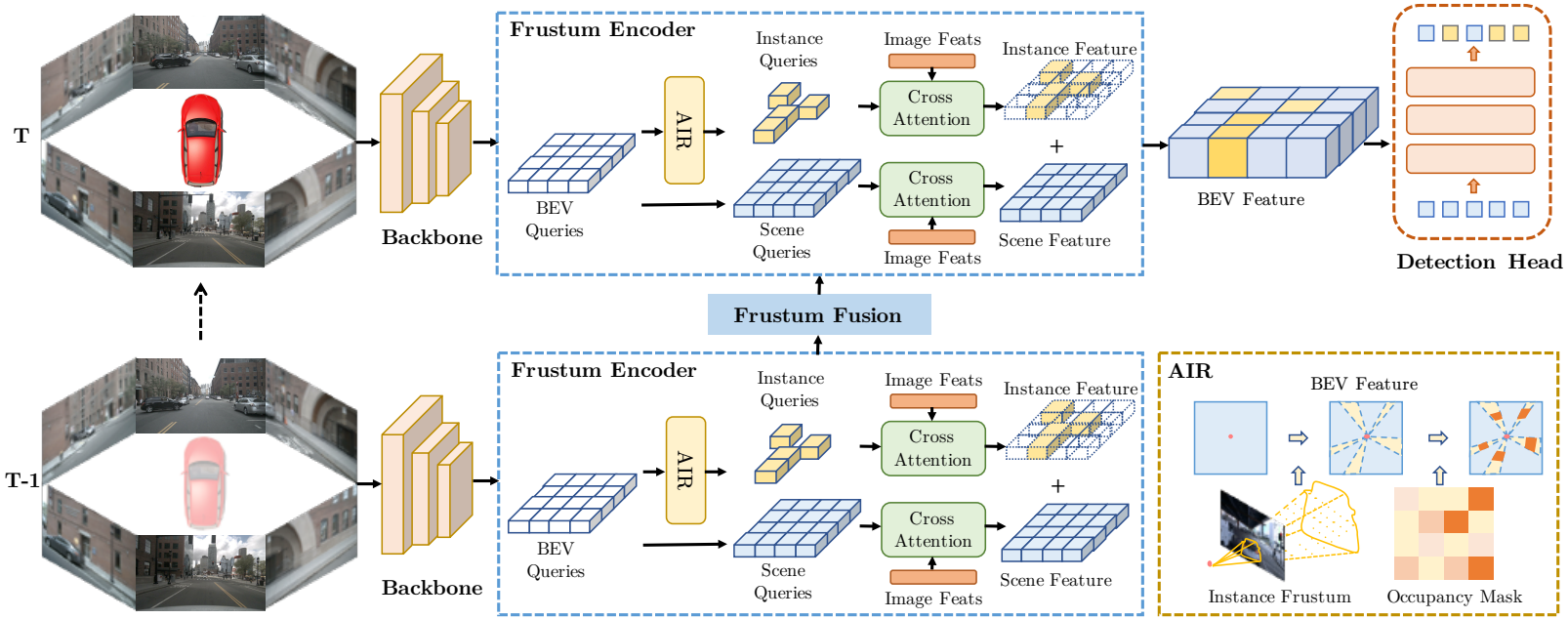
Method Overview

❑ FrustumFormer

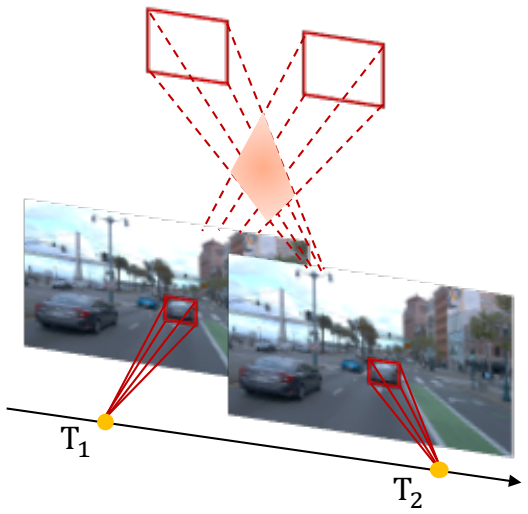
1. Backbone 2. **Frustum Encoder** 3. **Frustum Fusion** 4. Detection Head

❑ Frustum Encoder

- **Instance Queries:** sparse and irregular
- **Scene Queries:** dense and regular



Overview of FrustumFormer



Frustum Fusion

Experiments

- **nuScenes test set:**

Methods	Backbone	CBGS	LiDAR	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
FCOS3D‡ [35]	R101†			0.358	0.428	0.690	0.249	0.452	1.434	**0.124**
PGD [34]	R101†			0.386	0.448	0.626	**0.245**	0.451	1.509	0.127
BEVFormer [19]	R101†			0.445	0.535	0.631	0.257	0.405	0.435	0.143
PolarFormer [13]	R101†			0.457	0.543	0.612	0.257	**0.392**	0.467	0.129
FrustumFormer	R101†			**0.478**	**0.561**	**0.575**	0.257	0.402	**0.411**	0.132
DD3D [28]‡	V2-99*			0.418	0.477	0.572	0.249	0.368	1.014	0.124
DETR3D‡ [36]	V2-99*	✓		0.412	0.479	0.641	0.255	0.394	0.845	0.133
Ego3RT [25]	V2-99*			0.425	0.473	0.549	0.264	0.433	1.014	0.145
M2BEV [40]	X-101			0.429	0.474	0.583	0.254	0.376	1.053	0.190
BEVDet4D‡ [11]	Swin-B	✓		0.451	0.569	**0.511**	**0.241**	0.386	**0.301**	0.121
UVTR [17]	V2-99*			0.472	0.551	0.577	0.253	0.391	0.508	0.123
BEVFormer [19]	V2-99*			0.481	0.569	0.582	0.256	0.375	0.378	0.126
PolarFormer [13]	V2-99*			0.493	0.572	0.556	0.256	0.364	0.440	0.127
PETRv2 [23]	V2-99*			0.490	0.582	0.561	0.243	**0.361**	0.343	**0.120**
BEVDepth‡ [18]	V2-99*	✓	✓	0.503	0.600	0.445	0.245	0.378	0.320	0.126
BEVStereo [16]	V2-99*	✓	✓	0.525	0.610	0.431	0.246	0.358	0.357	0.138
FrustumFormer	V2-99*			**0.516**	**0.589**	0.555	0.249	0.372	0.389	0.126

- **nuScenes validation set:**

Methods	Backbone	CBGS	LiDAR	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
FCOS3D [35]	R101†			0.295	0.372	0.806	0.268	0.511	1.315	0.170
DETR3D [36]	R101†	✓		0.349	0.434	0.716	0.268	0.379	0.842	0.200
PGD [34]	R101†			0.358	0.425	0.667	**0.264**	0.435	1.276	**0.177**
PETR [22]	R101†	✓		0.370	0.442	0.711	0.267	0.383	0.865	0.201
UVTR [17]	R101†			0.379	0.483	0.731	0.267	0.350	0.510	0.200
BEVFormer [19]	R101†			0.416	0.517	0.673	0.274	0.372	0.394	0.198
PolarFormer [13]	R101†			0.432	0.528	0.648	0.270	**0.348**	0.409	0.201
BEVDepth [18]	R101	✓	✓	0.412	0.535	0.565	0.266	0.358	0.331	0.190
STS [38]	R101	✓	✓	0.431	0.542	0.525	0.262	0.380	0.369	0.204
FrustumFormer	R101†			**0.457**	**0.546**	**0.624**	0.265	0.362	**0.380**	0.191

We achieve **SOTA performance** on nuScenes test/val set without extra LiDAR supervision

FrustumFormer: Adaptive Instance-aware Resampling for Multi-view 3D Detection

Yuqi Wang[1,2] Yuntao Chen[3] Zhaoxiang Zhang[1,2,3]

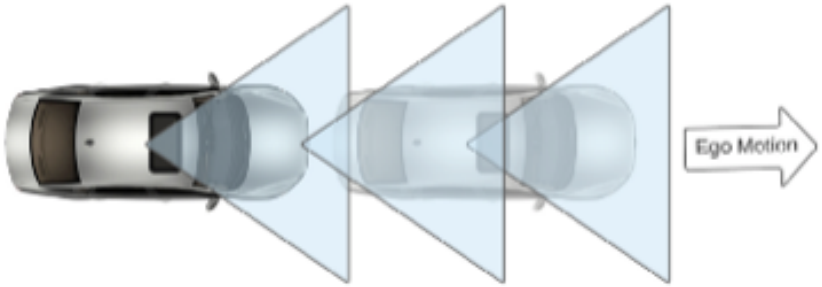
[1] Institute of Automation, Chinese Academy of Sciences (CASIA)

[2] School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS)

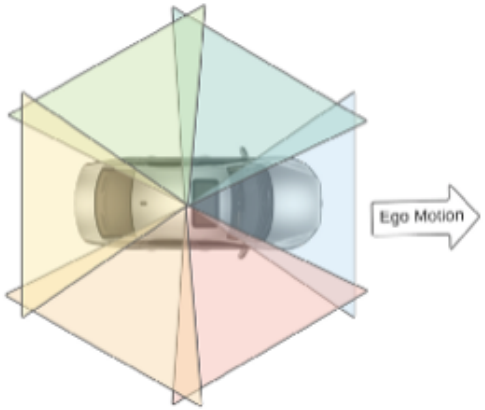
[3] Centre for Artificial Intelligence and Robotics, HKISI_CAS

Camera-based 3D Object Detection

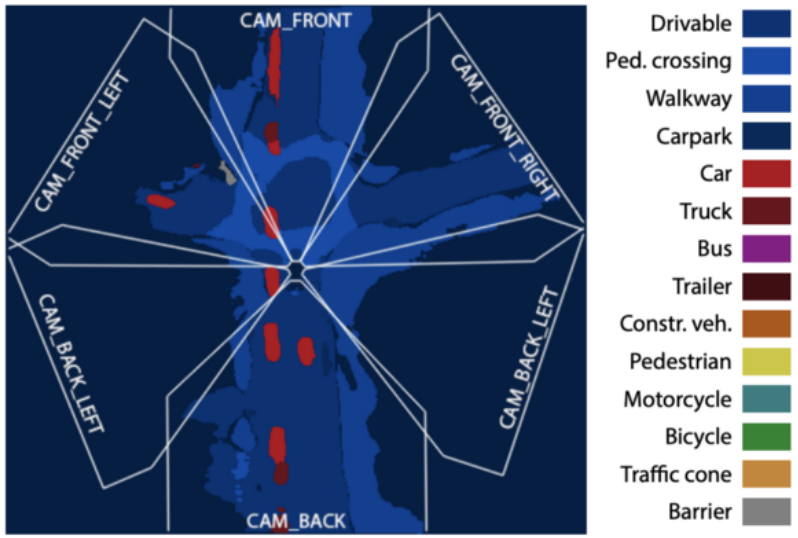
- **Mono Camera**



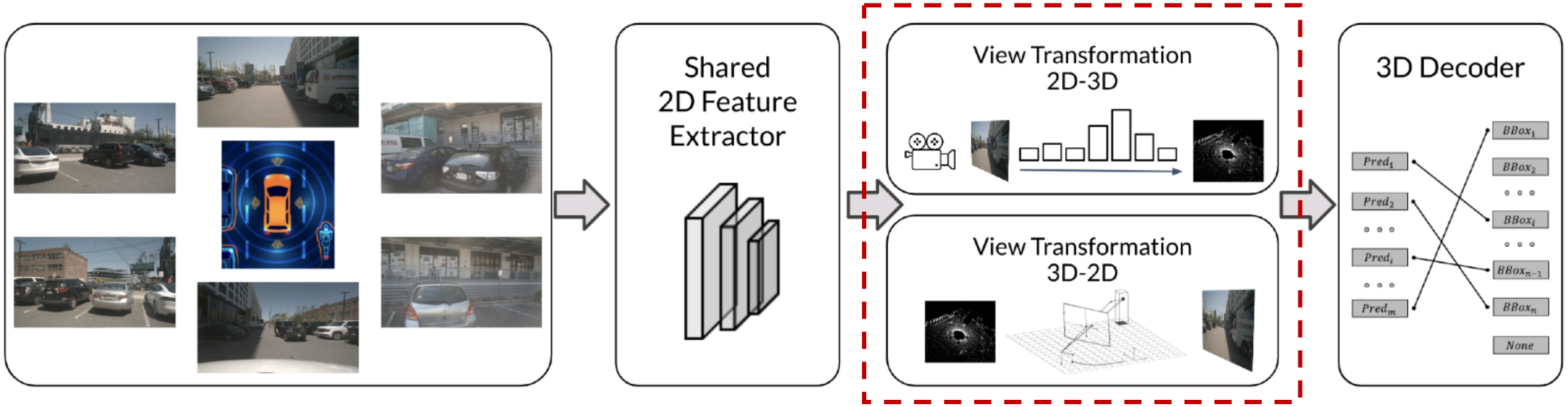
- **Multi-view Camera**



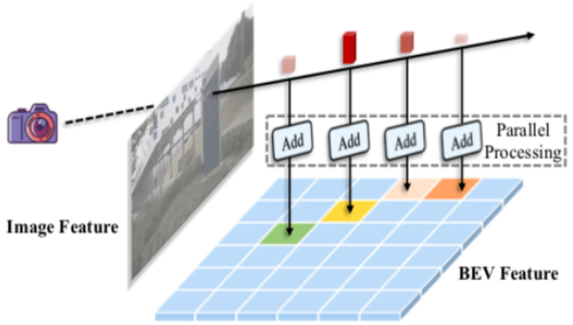
- **3D Object Detection**



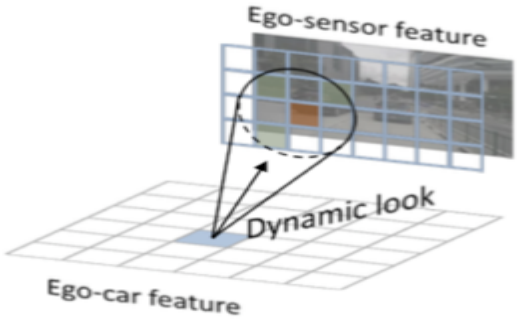
Multi-view 3D Object Detection



➢ **How to transform view feature?**



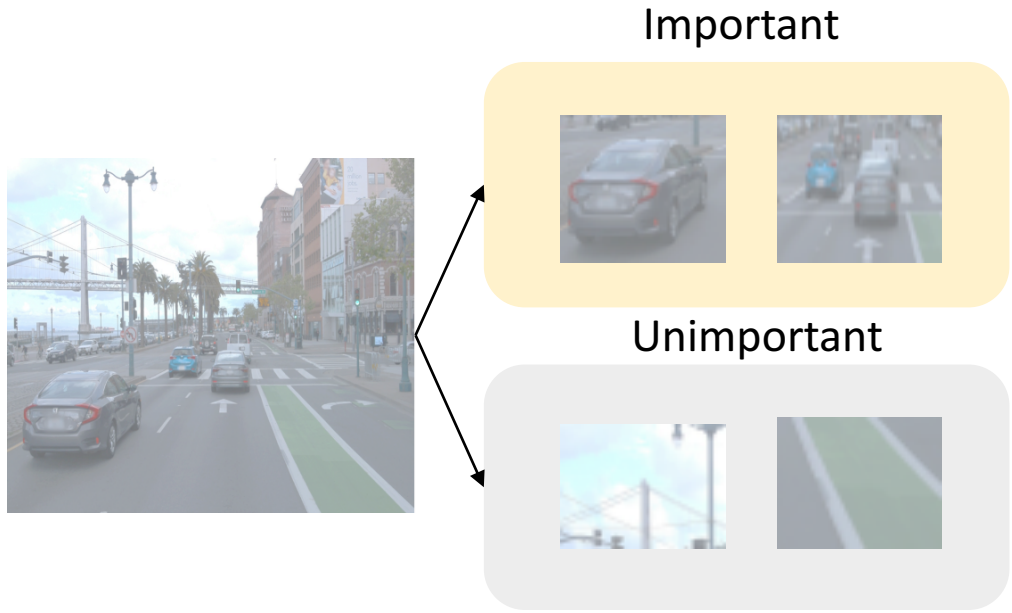
2D **Lift** 3D

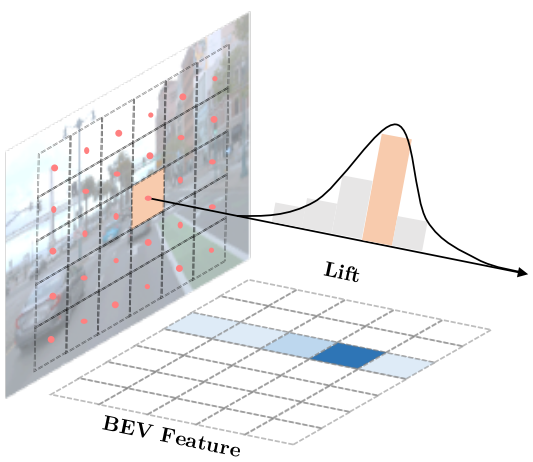


3D **Query** 2D

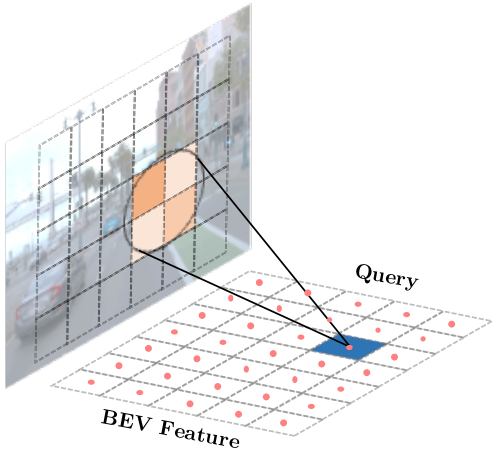
Motivation

- **View Feature Transformation**
  - How to transform view feature? 😋
  - Choosing what to transform? 🤔
- **Adaptive View transformation**
  - Paying more attention to instance regions

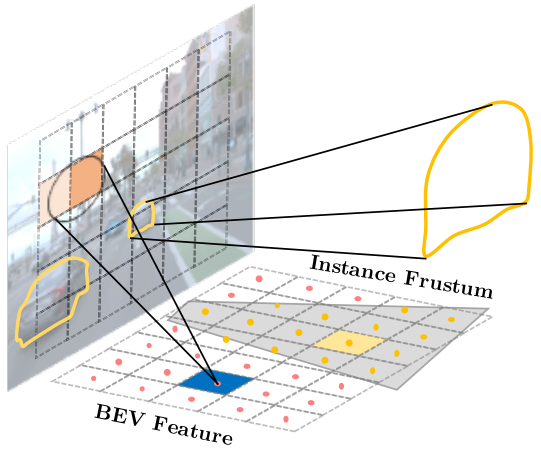




(a) Grid Sampling in Image



(b) Grid Sampling in BEV



(c) Instance-aware Sampling in Frustum

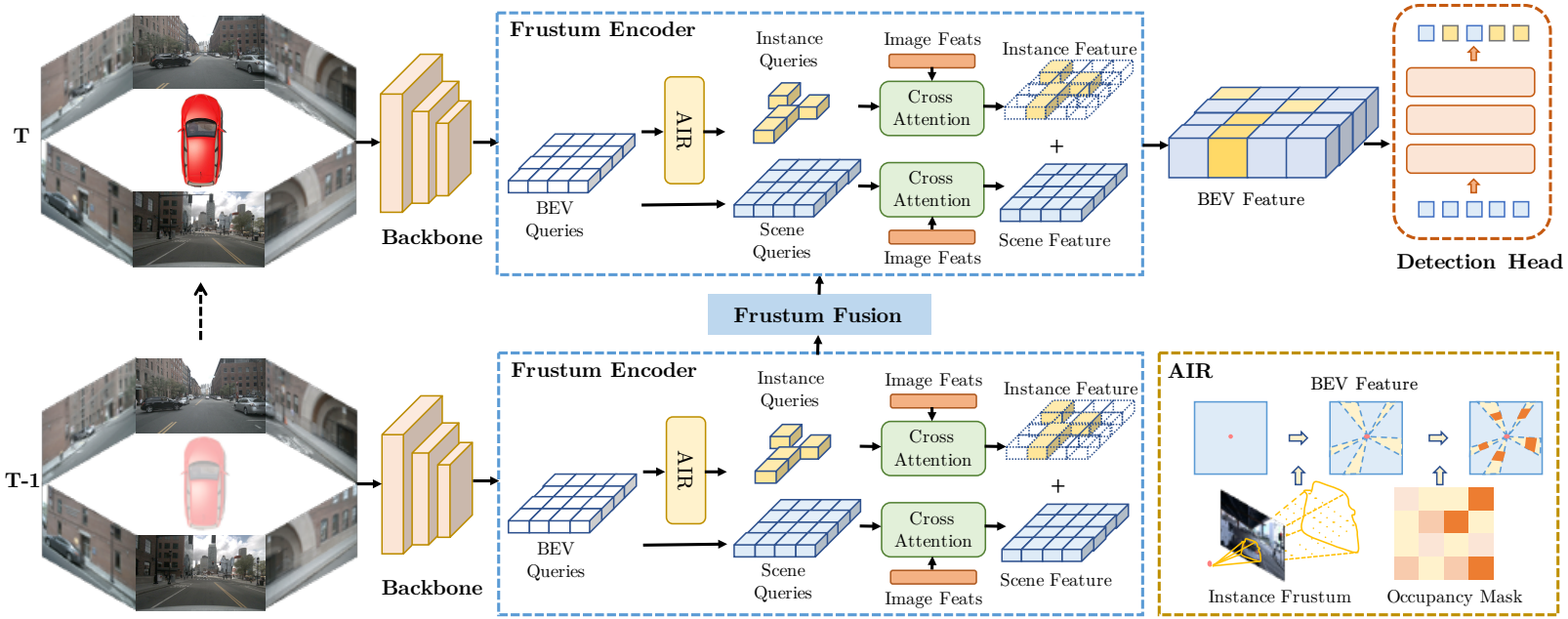
Method Overview

❑ FrustumFormer

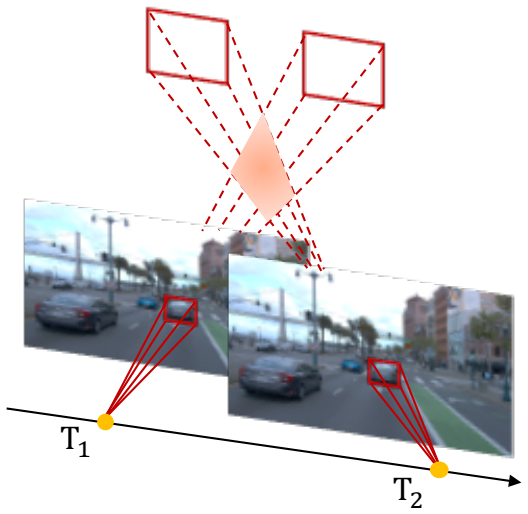
1. Backbone 2. **Frustum Encoder** 3. **Frustum Fusion** 4. Detection Head

❑ Frustum Encoder

- **Instance Queries:** sparse and irregular
- **Scene Queries:** dense and regular



Overview of FrustumFormer

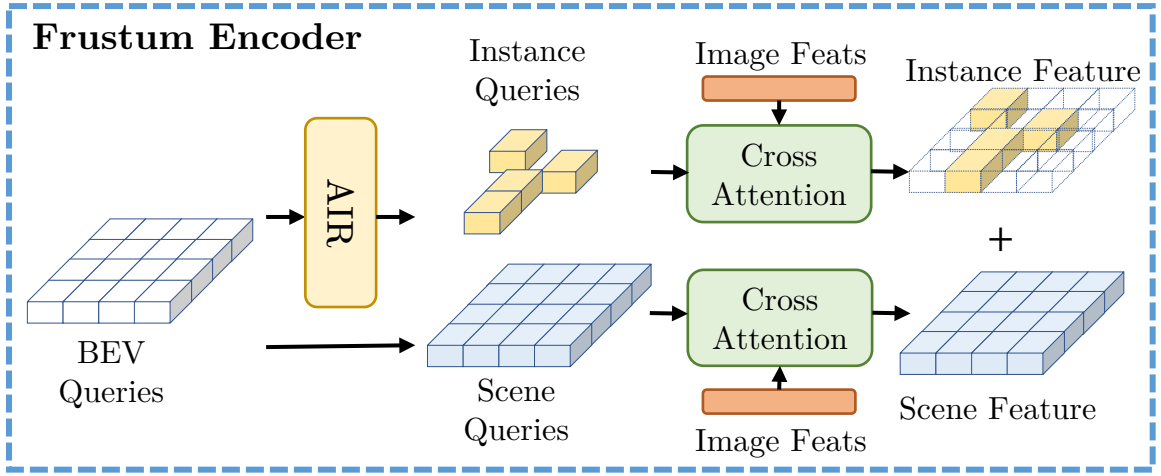


Frustum Fusion

Method Details

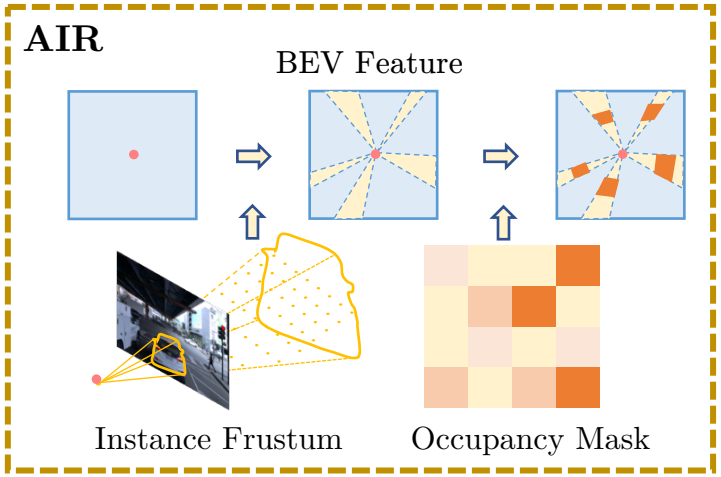
❑ Frustum Encoder

- **Instance queries(yellow):** instance queries are sparse and generated from irregular instance frustum.
- **Scene queries(blue):** scene queries are dense and generated from regular BEV grids.



❑ Adaptive Instance-aware Resampling

- **Instance frustum query generation**: we take advantage of object detection on the image plane and leverage its instance frustum on the BEV plane to select the instance frustum queries.
- **Frustum occupancy mask prediction**: in order to reduce the localization uncertainty, we propose to predict an occupancy mask for all frustums.



Method Details

❑ Instance Frustum Cross-Attention (IFCA)

✓ Deformable Attention

$$IFCA(\mathbf{Q}_i^{p_i}, \mathbf{F}_j) = \frac{1}{|v|} \sum_{j \in v} \sum_{m=1}^{M} DA(\mathbf{Q}_i^{p_i}, \pi_j(\mathbf{p}_i^m), \mathbf{F}_j)$$

❑ Temporal Frustum Cross-Attention (TFCA)

✓ A sequential RNN way

$$TFCA(\mathbf{Q}_f^{p_i}, \mathbf{H}_f) = \sum_{m=1}^{M} DA(\mathbf{Q}_f^{p_i}, \mathbf{p'}_i^m, \mathbf{H}_f)$$

Experiments

- **nuScenes test set:**

Methods	Backbone	CBGS	LiDAR	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
FCOS3D‡ [35]	R101†			0.358	0.428	0.690	0.249	0.452	1.434	**0.124**
PGD [34]	R101†			0.386	0.448	0.626	**0.245**	0.451	1.509	0.127
BEVFormer [19]	R101†			0.445	0.535	0.631	0.257	0.405	0.435	0.143
PolarFormer [13]	R101†			0.457	0.543	0.612	0.257	**0.392**	0.467	0.129
FrustumFormer	R101†			**0.478**	**0.561**	**0.575**	0.257	0.402	**0.411**	0.132
DD3D [28]‡	V2-99*			0.418	0.477	0.572	0.249	0.368	1.014	0.124
DETR3D‡ [36]	V2-99*	✓		0.412	0.479	0.641	0.255	0.394	0.845	0.133
Ego3RT [25]	V2-99*			0.425	0.473	0.549	0.264	0.433	1.014	0.145
M2BEV [40]	X-101			0.429	0.474	0.583	0.254	0.376	1.053	0.190
BEVDet4D‡ [11]	Swin-B	✓		0.451	0.569	**0.511**	**0.241**	0.386	**0.301**	0.121
UVTR [17]	V2-99*			0.472	0.551	0.577	0.253	0.391	0.508	0.123
BEVFormer [19]	V2-99*			0.481	0.569	0.582	0.256	0.375	0.378	0.126
PolarFormer [13]	V2-99*			0.493	0.572	0.556	0.256	0.364	0.440	0.127
PETRv2 [23]	V2-99*			0.490	0.582	0.561	0.243	**0.361**	0.343	**0.120**
BEVDepth‡ [18]	V2-99*	✓	✓	0.503	0.600	0.445	0.245	0.378	0.320	0.126
BEVStereo [16]	V2-99*	✓	✓	0.525	0.610	0.431	0.246	0.358	0.357	0.138
FrustumFormer	V2-99*			**0.516**	**0.589**	0.555	0.249	0.372	0.389	0.126

- **nuScenes validation set:**

Methods	Backbone	CBGS	LiDAR	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
FCOS3D [35]	R101†			0.295	0.372	0.806	0.268	0.511	1.315	0.170
DETR3D [36]	R101†	✓		0.349	0.434	0.716	0.268	0.379	0.842	0.200
PGD [34]	R101†			0.358	0.425	0.667	**0.264**	0.435	1.276	**0.177**
PETR [22]	R101†	✓		0.370	0.442	0.711	0.267	0.383	0.865	0.201
UVTR [17]	R101†			0.379	0.483	0.731	0.267	0.350	0.510	0.200
BEVFormer [19]	R101†			0.416	0.517	0.673	0.274	0.372	0.394	0.198
PolarFormer [13]	R101†			0.432	0.528	0.648	0.270	**0.348**	0.409	0.201
BEVDepth [18]	R101	✓	✓	0.412	0.535	0.565	0.266	0.358	0.331	0.190
STS [38]	R101	✓	✓	0.431	0.542	0.525	0.262	0.380	0.369	0.204
FrustumFormer	R101†			**0.457**	**0.546**	**0.624**	0.265	0.362	**0.380**	0.191

We achieve **SOTA performance** on nuScenes test/val set without extra LiDAR supervision

Experiments

- **Ablation of Components in FrustumFormer**

	IF	OM	FF	mAP↑	NDS↑	mATE↓
(a)				0.318	0.366	0.771
(b)	✓			0.326	0.373	0.765
(c)		✓		0.328	0.381	0.759
(d)	✓	✓		0.337	0.383	0.749
(e)	✓	✓	✓	**0.360**	**0.463**	**0.719**

- **Ablation of Instance-aware Sampling**

	Total	Scene	Instance	mAP↑	NDS↑
(a)	1×	1×	-	0.318	0.366
(b)	2×	2×	-	0.318	0.362
(c)	2×	1×	1×	**0.326**	**0.373**

Experiments

- **Ablation of Occupancy Mask Learning**

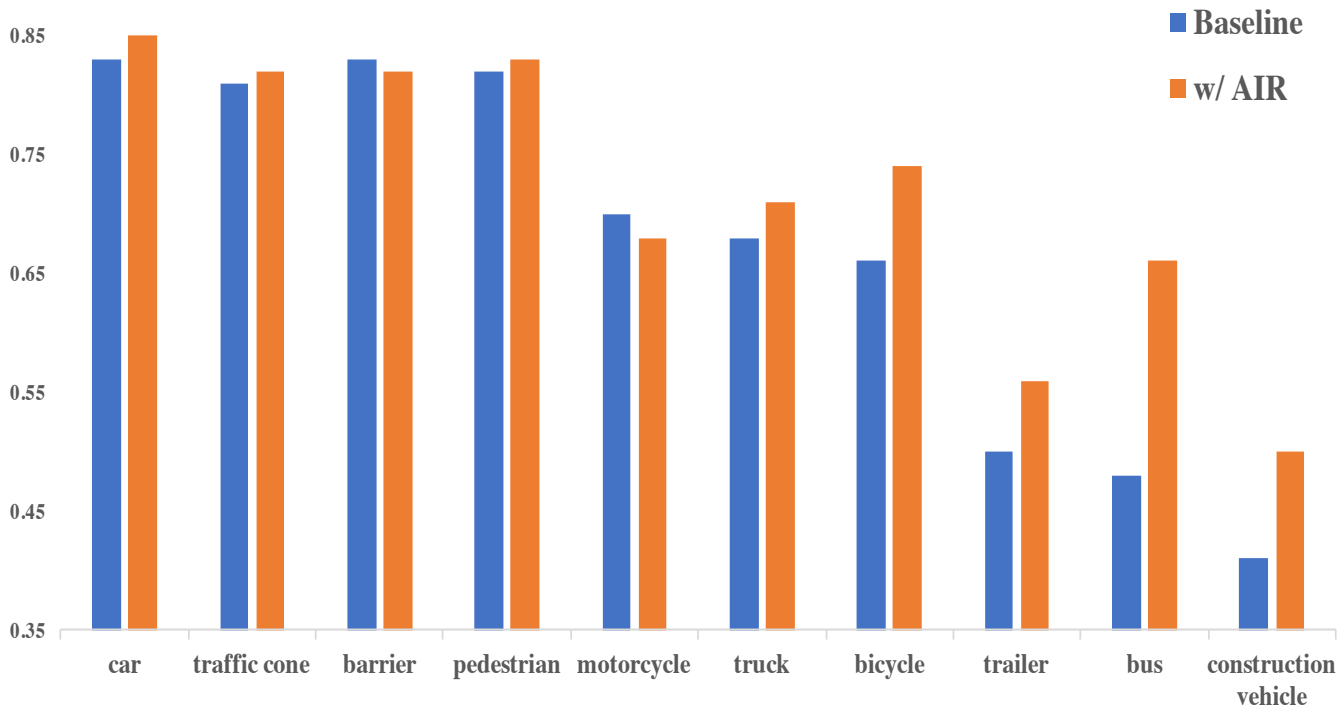
	Supervision	α	mAP↑	NDS↑	mATE↓
(a)	w/o	0.0	0.318	0.366	0.771
(b)	w/ BEV box*	5.0	0.324	0.374	0.756
(c)	w/ BEV box	5.0	**0.328**	**0.381**	0.759
(d)	w/ BEV box	10.0	0.322	0.381	**0.749**
(e)	w/ BEV box	1.0	0.326	0.379	0.751

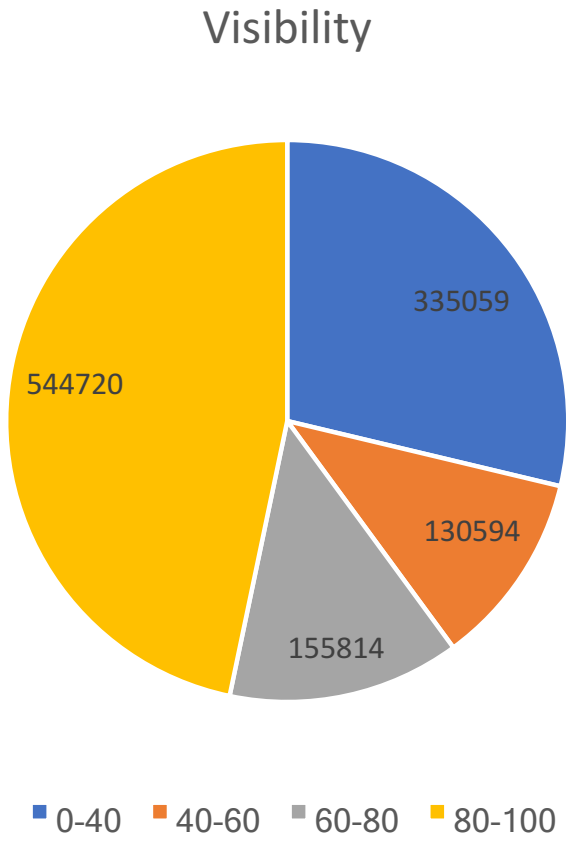
- **Ablation of Temporal Frustum Fusion**

	W	K	Frustum	mAP↑	NDS↑	mAVE↓
(a)	4	2		0.353	0.454	0.497
(b)	4	2	✓	0.355	0.457	0.479
(c)	8	4	✓	0.360	**0.463**	**0.463**
(d)	16	4	✓	**0.364**	0.457	0.568

Experiments

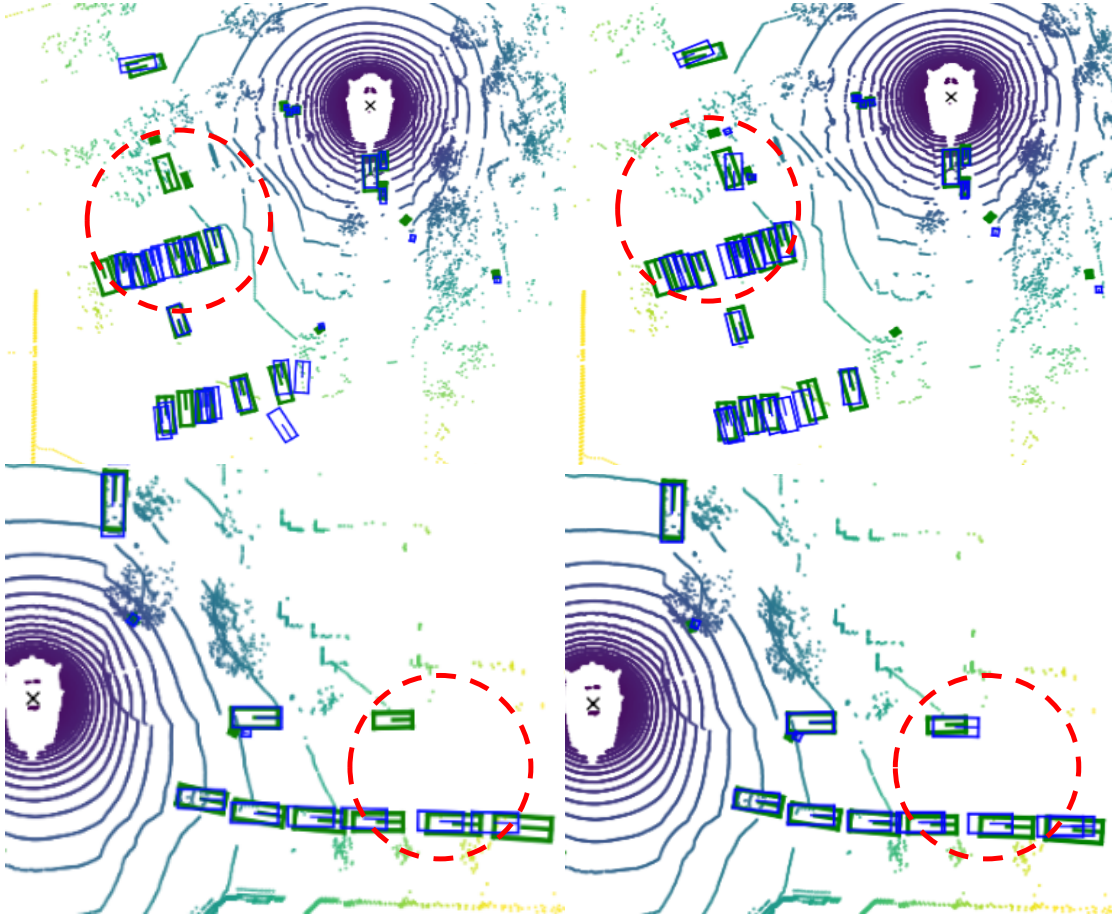
➢ **Recall Improvement Under Low Visibility(0-40%)**

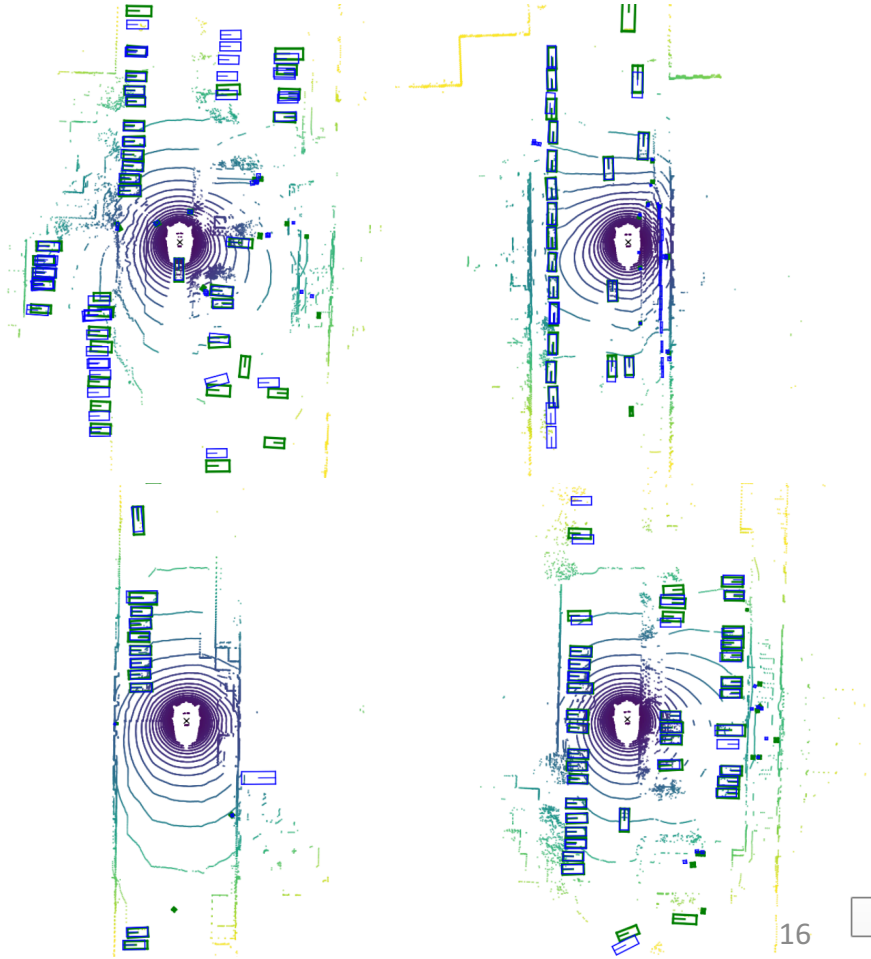




Experiments

- **Recall Improvement Under Low Visibility(0-40%)**



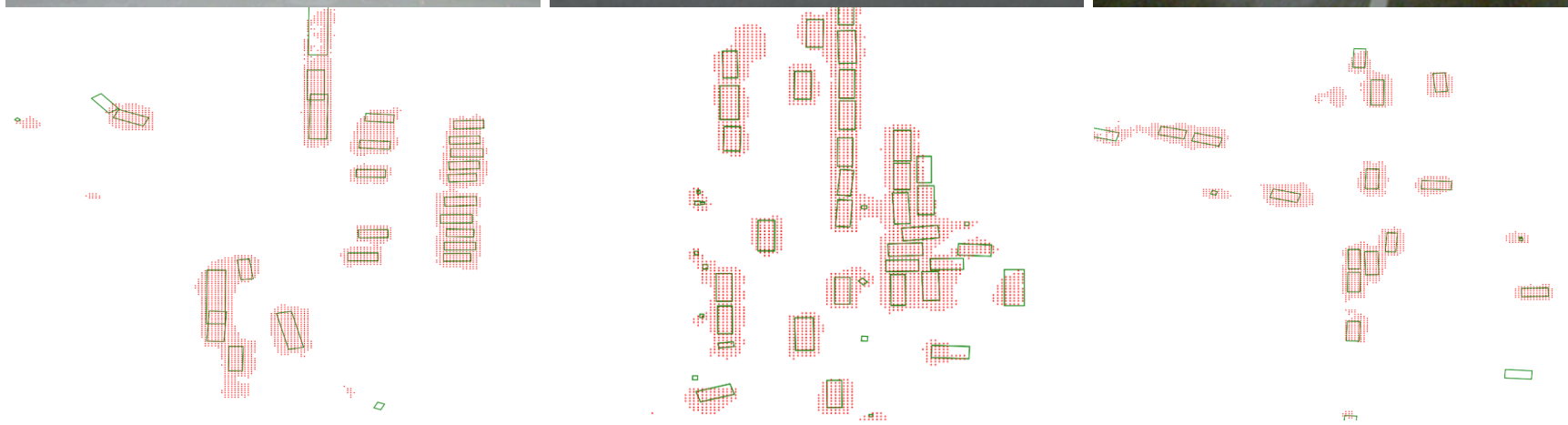


Experiments

➢ **Instance Queries Visualization**

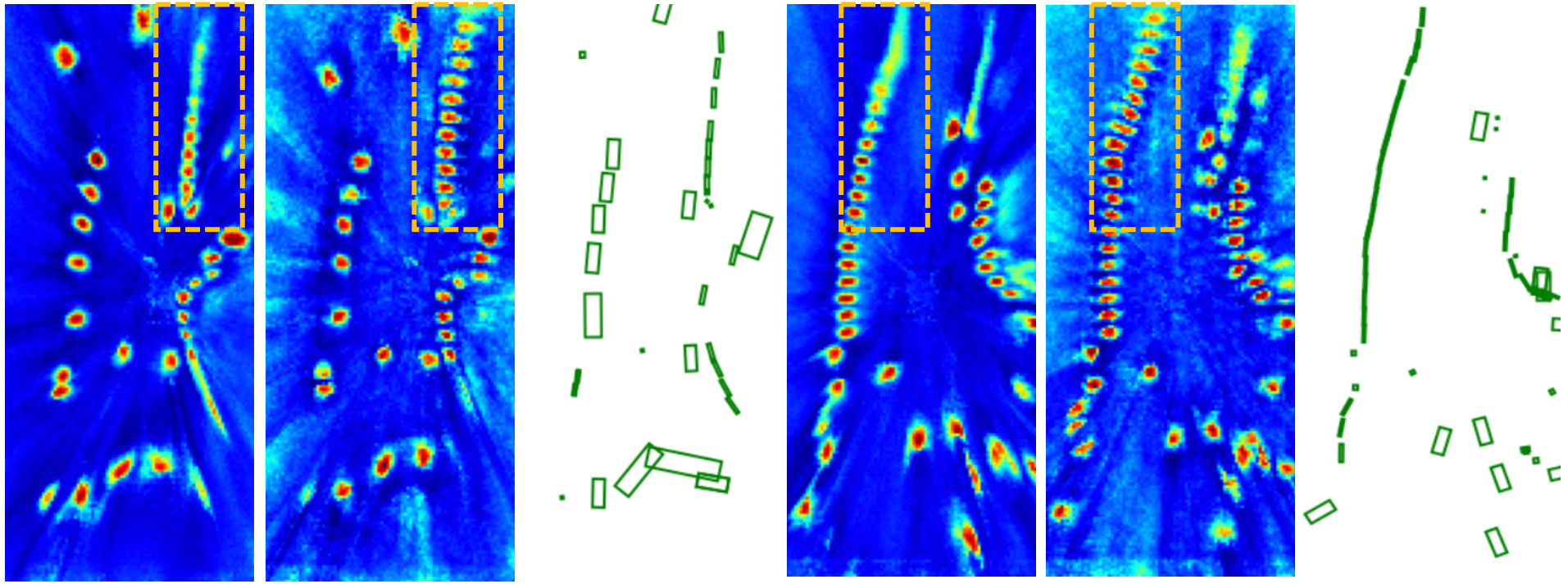
Image

BEV



Experiments

➢ **BEV Feature Visualization**



Baseline w/ AIR GT Baseline w/ AIR GT

Thanks For Your Listening

Paper link: *https://arxiv.org/pdf/2301.04467.pdf*

Code link: *https://github.com/Robertwyq/Frustum*