

Long-Tailed 3D Detection

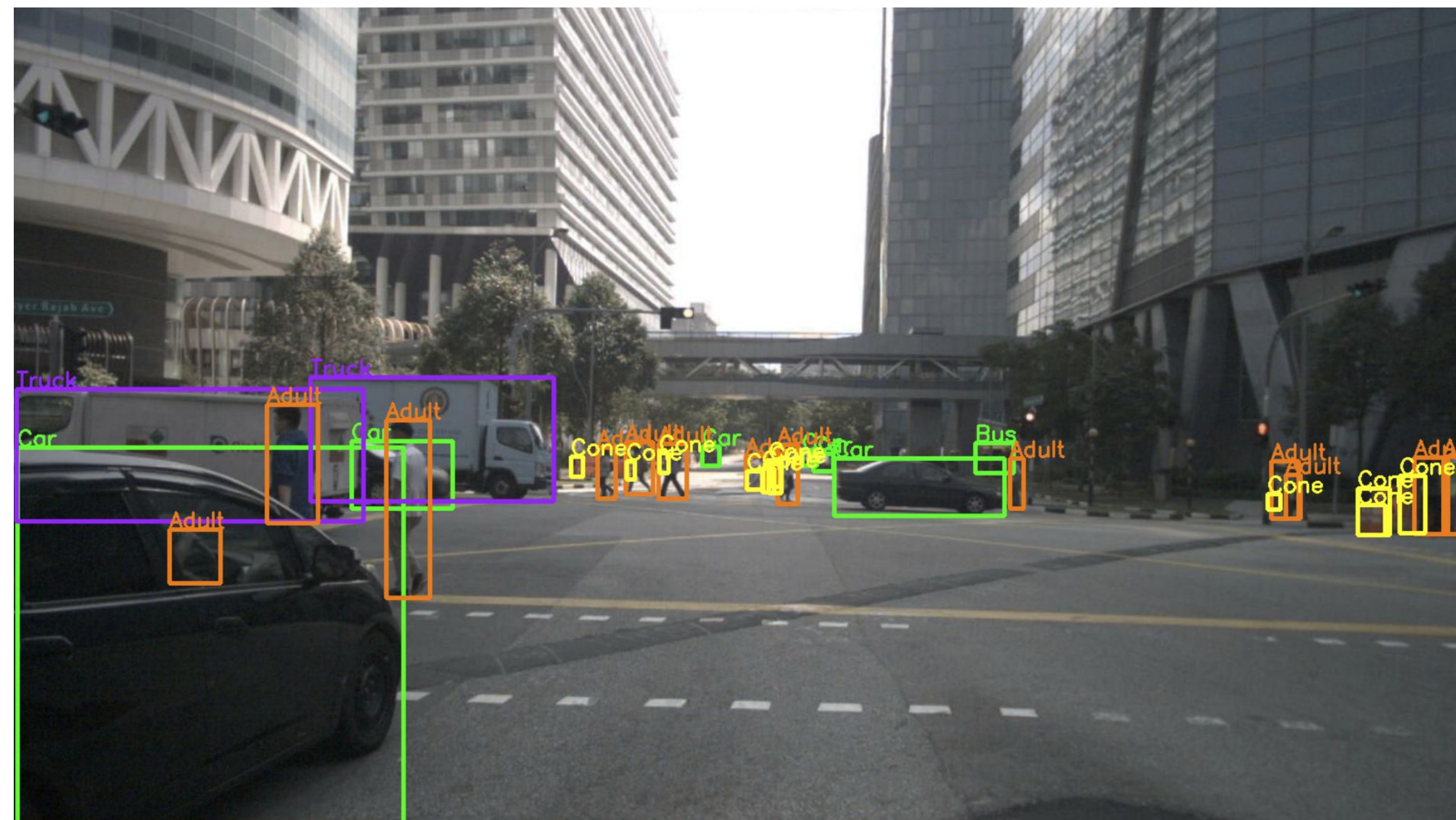


- To deploy a self-driving vehicle safely, it is crucial to detect objects from both common and rare classes, which could be challenging due to limited training data
- Foundation models show promise of bringing useful prior knowledge to tackle long-tailed 3D detection

Vision Foundation Models

OWLv2

- Open-vocabulary 2D object detector
- Provides image embeddings
- Provides 2D detections from text prompts

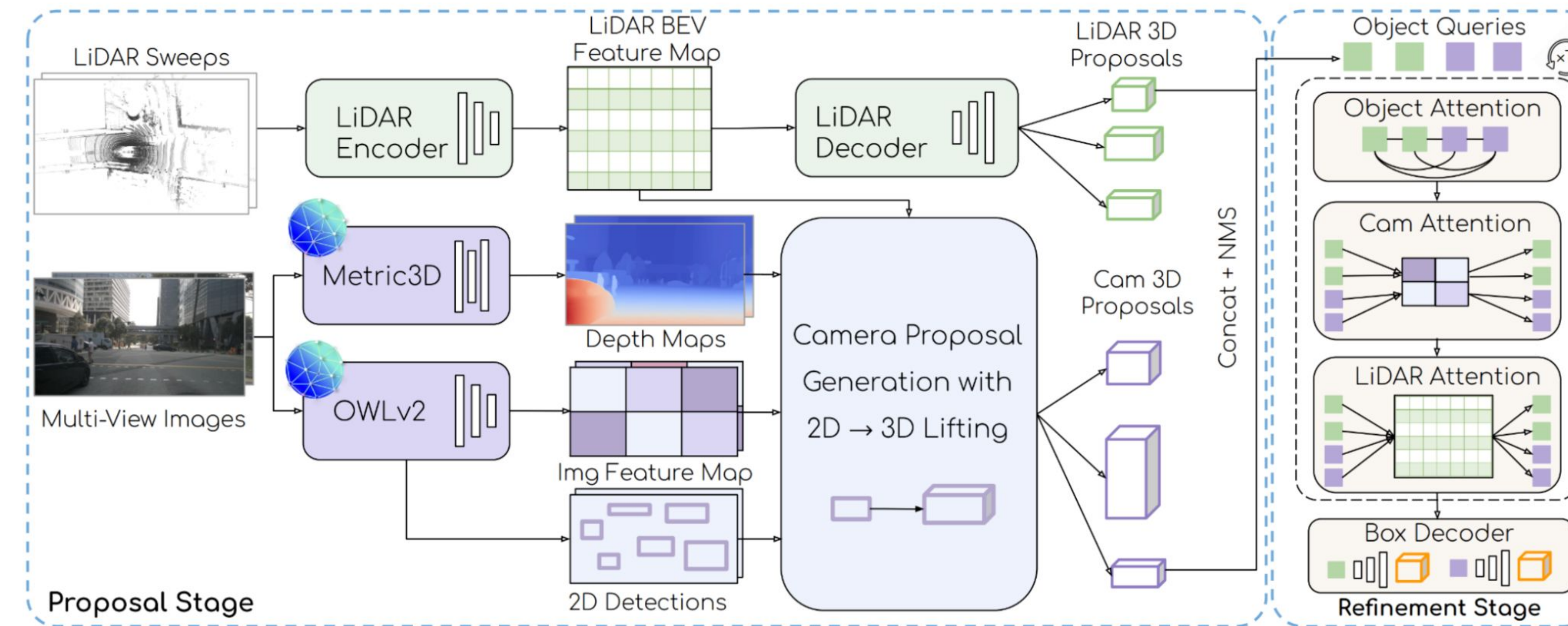


Metric3Dv2 (M3D)

- Monocular depth estimation model
- Provides dense depth images

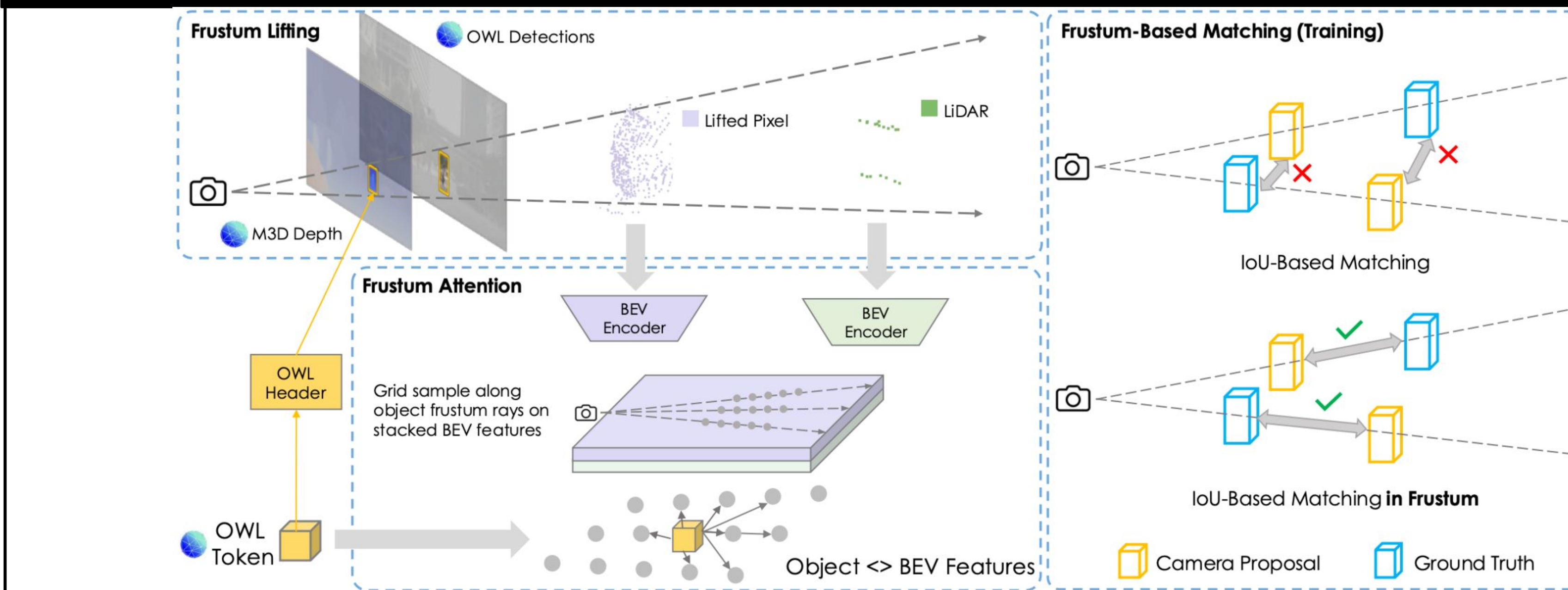


FOMO-3D



- Two-stage DETR-style detection paradigm
- First proposal stage - LiDAR proposal branch (CenterPoint), novel camera proposal branch lifting 2D OWL boxes into 3D detections, using M3D depths and fusing with LiDAR features
- Second refinement stage - object self-attention, object to LiDAR attention, object to OWL features attention

Camera Proposal

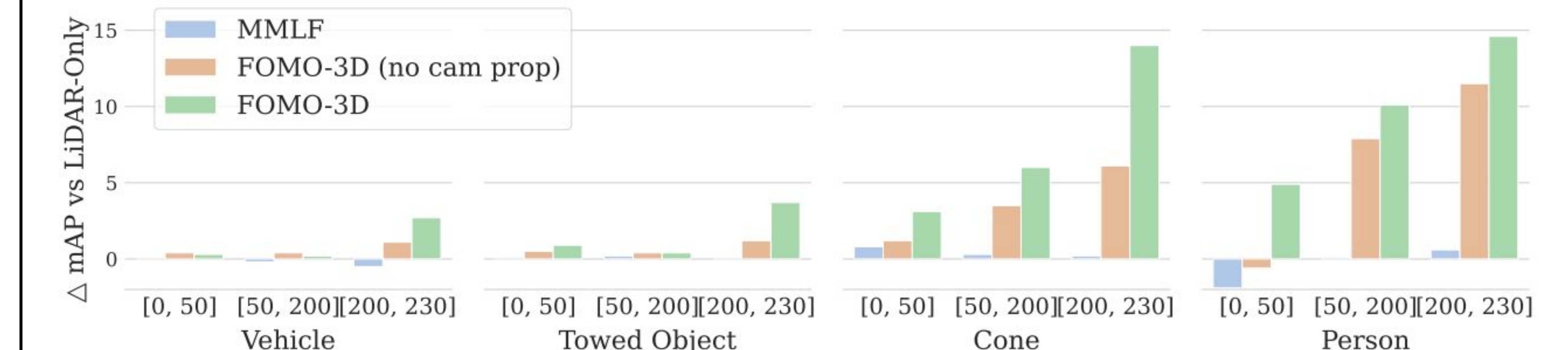


- Obtain 2D detections and image features from OWL
 - Lift each 2D detection into 3D using M3D depths
 - Lift image features to 3D as a feature point cloud, and concat with LiDAR features as a fused BEV feature map
 - Refine 3D detection via attention to the BEV LiDAR and image features in the object frustum
- During training – frustum-aligned matching between proposals and ground truth for supervision

Quantitative Results

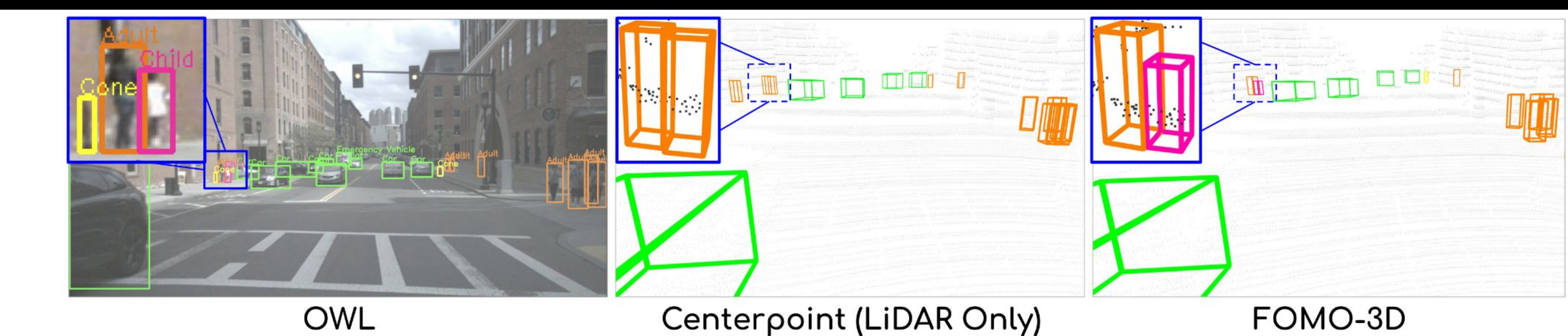
Method	Modality	All	Many	Medium	Few
BEVFormer [38]	C	27.3	52.3	31.6	1.4
CenterPoint (Group-Free) [1, 5]	L	39.2	76.4	43.1	3.5
BEVFusion-L [2]	L	42.5	72.5	48.0	10.6
TransFusion [25]	L+C	39.8	73.9	41.2	9.8
BEVFusion [2]	L+C	45.5	75.5	52.0	12.8
CMT [2]	L+C	44.4	79.9	53.0	4.8
MMF [5]	L+C	43.6	77.1	49.0	9.4
MMF* (w/ OWL + M3D) [5]	L+C	44.3	77.7	46.9	13.4
MMLF [24]	L+C	51.4	77.9	59.4	20.0
FOMO-3D	L+C	54.6	79.9	59.6	27.6

nuScenes:
FOMO-3D outperforms
SoTA detectors in all
aggregated object
groups, esp. on Few



Highway:
FOMO-3D
shines for rare
classes and
long-range
detection.

Qualitative Results



- OWL detects cone but has false positive cone
- LiDAR-only misclassified child as adult
- FOMO-3D combines multimodal information to successfully detect the child and reject the false positive cone

- OWL detects cone but duplicates person and cyclist
- LiDAR-only performs poorly on cone and person
- FOMO-3D performs well on cone, person, without duplicating person/cyclist
- All share a failure mode - confusing trailer and truck

