RoScene: A Large Scale Multi-view 3D Dataset For Roadside Perception

Xiaosu Zhu[1*], Huanlian Sheng[1*], Sijia Cai[1†], Bing Deng[1], Shaopeng Yang[1], Qiao Liang[1], Ken Chen[2], Lianli Gao[1], Jingkuan Song[1‡], and Jieping Ye[1‡]

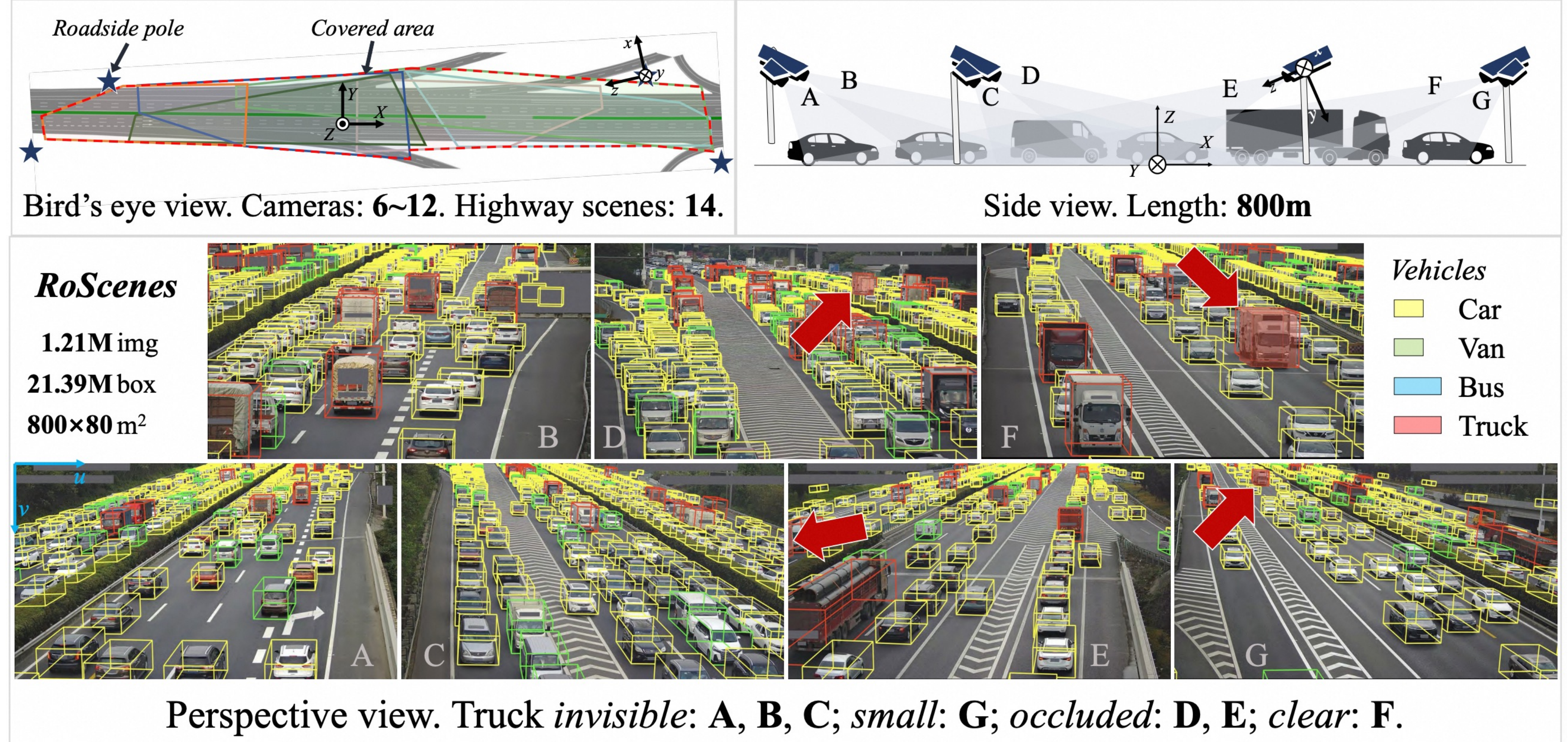
[1]Alibaba Cloud [2]Sichuan Digital Transportation Technology Co., Ltd, [3]Independent Researcher [4]Tongji University

*Equal contribution †Project lead ‡Corresponding authors

Overview

Dataset	Year	Type V	Type I	Cam	BEV area (m^2)	Duration $(hour)$	Diversity Night	Diversity Rain	Image	Box	Class
KITTI [10]	2012	✓	-	2	70 × 80	1.5	✓	-	0.02M	0.08M	8
ApolloScape [16]	2019	✓	-	6	-	2.5	✓	-	0.14M	0.07M	8-35
nuScenes [1]	2019	✓	-	6	100 × 100	5.5	✓	✓	1.40M	1.40M	23
Argoverse [4]	2020	✓	-	7	205 × 155	320.0	✓	✓	0.02M	0.99M	15
Waymo Open [9]	2020	✓	-	5	150 × 150	6.4	✓	✓	0.23M	12.00M	4
ONCE [25]	2021	✓	-	7	200 × 200	144.0	✓	✓	7.00M	0.42M	5
Rope3D [38]	2022	-	✓	1	104 × 102	-	✓	✓	0.05M	1.50M	12
V2X-Seq [40]	2023	✓	✓	1+1	104 × 102	0.4	-	✓	0.07M	1.20M	10
A9 [5]	2023	-	✓	4	-	0.1	-	✓	0.01M	0.21M	9
RoScenes (Ours)	-	-	✓	**6∼12**	**800 × 80**	23.9	✓	-	**1.30M**	**21.13M**	4

Quantitative comparison with the published vehicle-side and infrastructureside 3D datasets. Our dataset achieves the largest BEV perception area and the largest number of annotations. Type: **V:** Vehicle-side sensors. **I:** Infrastructure-side sensors. "Cam" is the number of synchronized cameras adopted per scene



Bird's eye view. Cameras: **6~12**. Highway scenes: **14**.

Side view. Length: **800m**

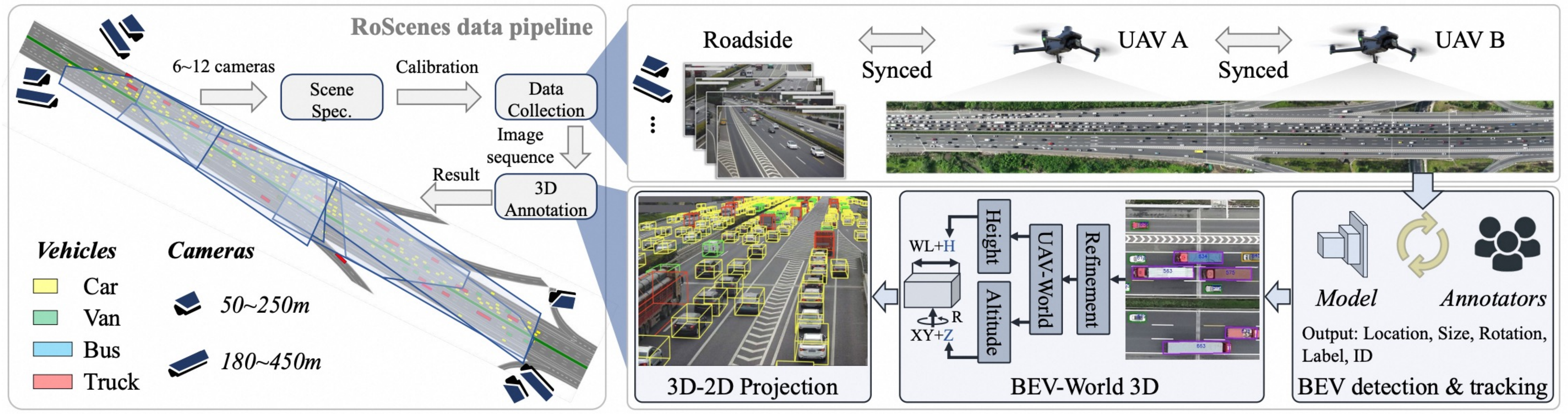
Perspective view. Truck *invisible*: **A, B, C**; *small*: **G**; *occluded*: **D, E**; *clear*: **F**.

Demonstration of our RoScenes dataset. The annotated truck is difficult to recognize in A, B, C, E, F, G, but is clear in D.

Contribution

1) A large-scale multi-view 3D dataset for roadside perception, with a novel, cost-effective annotation pipeline is introduced to obtain 3D annotations in challenging traffic scenarios.
2) A method RoBEV that effectively aggregates 2D image feature to 3D detection queries via feature-guided position embedding.
3) The extensive experimental evaluation also indicates the RoScenes dataset can serve as a benchmark for BEV architectures in the future.

RoScene Data Pipeline



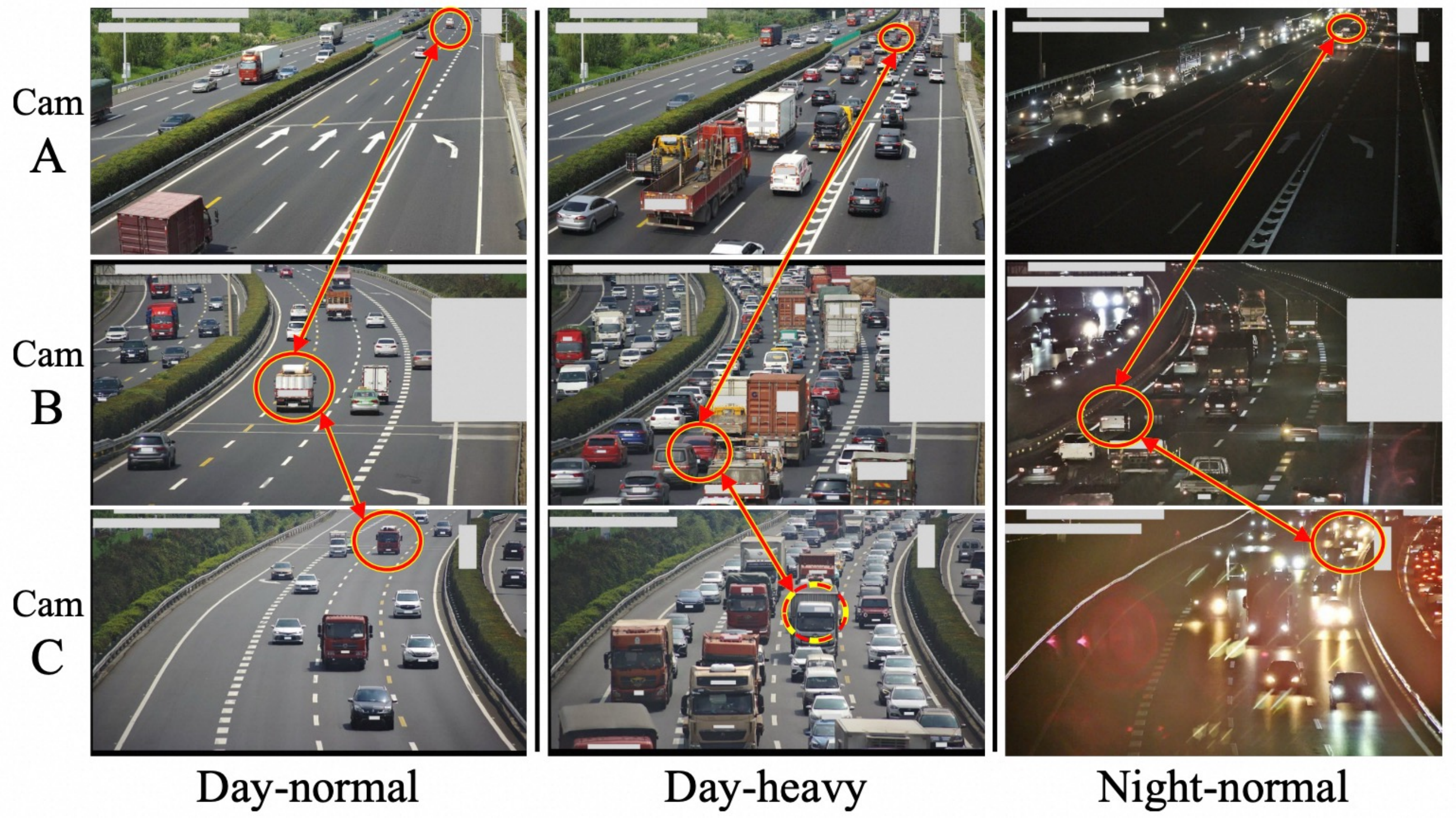
Overall data collection and annotation pipeline. We propose BEV-to-3D joint annotation for efficiency.

1) A couple of UAVs hover the target scene to capture aerial image sequence along with roadside camera sequence synchronously;
2) We train the UAV model consists of BEV detector and tracker on UAV images for generating image-level BEV annotations, which are then transformed to the XY plane of World Coordinate via the UAV-to-World homography matrix;
3) To further access the altitude and height of each annotated vehicle for converting BEV 2D boxes to world 3D boxes, we choose the center of BEV 2D boxes to query the altitude in pre-built 3D reconstruction model and attach vehicle's height to the average height of its class label;
4) We perform the perspective projection of 8 corners of 3D boxes onto 2D image planes for all roadside cameras using the camera parameters.

Scene	Train	Easy	Hard	Test (Unseen)
#001	329(138k)	51(21k)	51(21k)	#005 ∼ #014
#002	534(256k)	83(40k)	83(40k)	
#003	306(129k)	48(20k)	48(20k)	
#004	534(257k)	80(40k)	80(40k)	
Sum	1,703(779k)	265(121k)	265(121k)	637(274k)

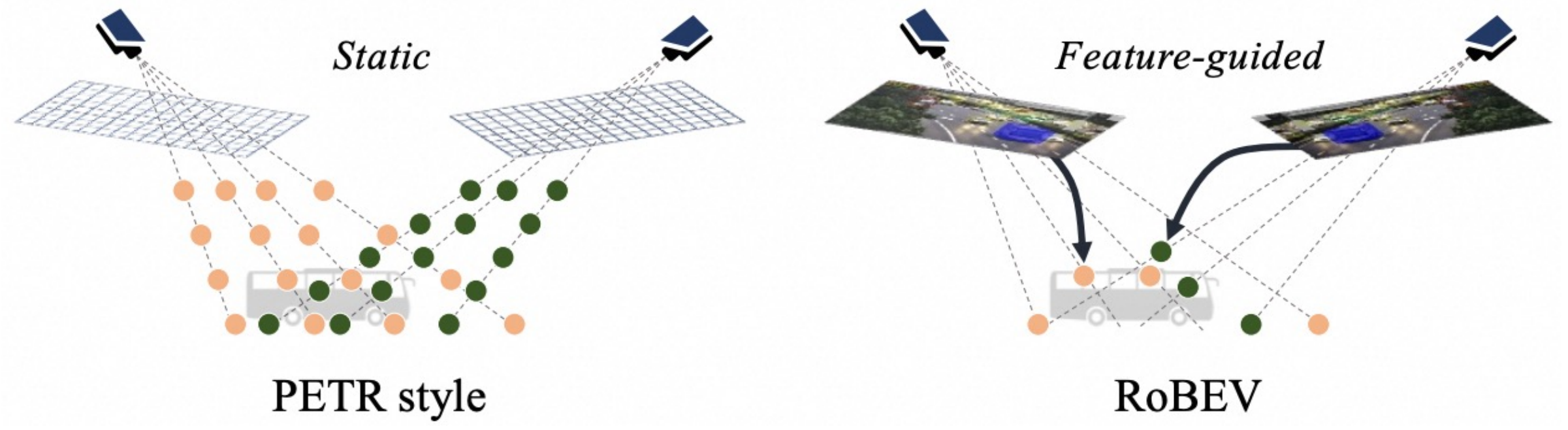
Train, validation (easy, hard) and unseen test splits for benchmark. (Clips and images)

EUROPEAN CONFERENCE ON COMPUTER VISION
MILANO 2024

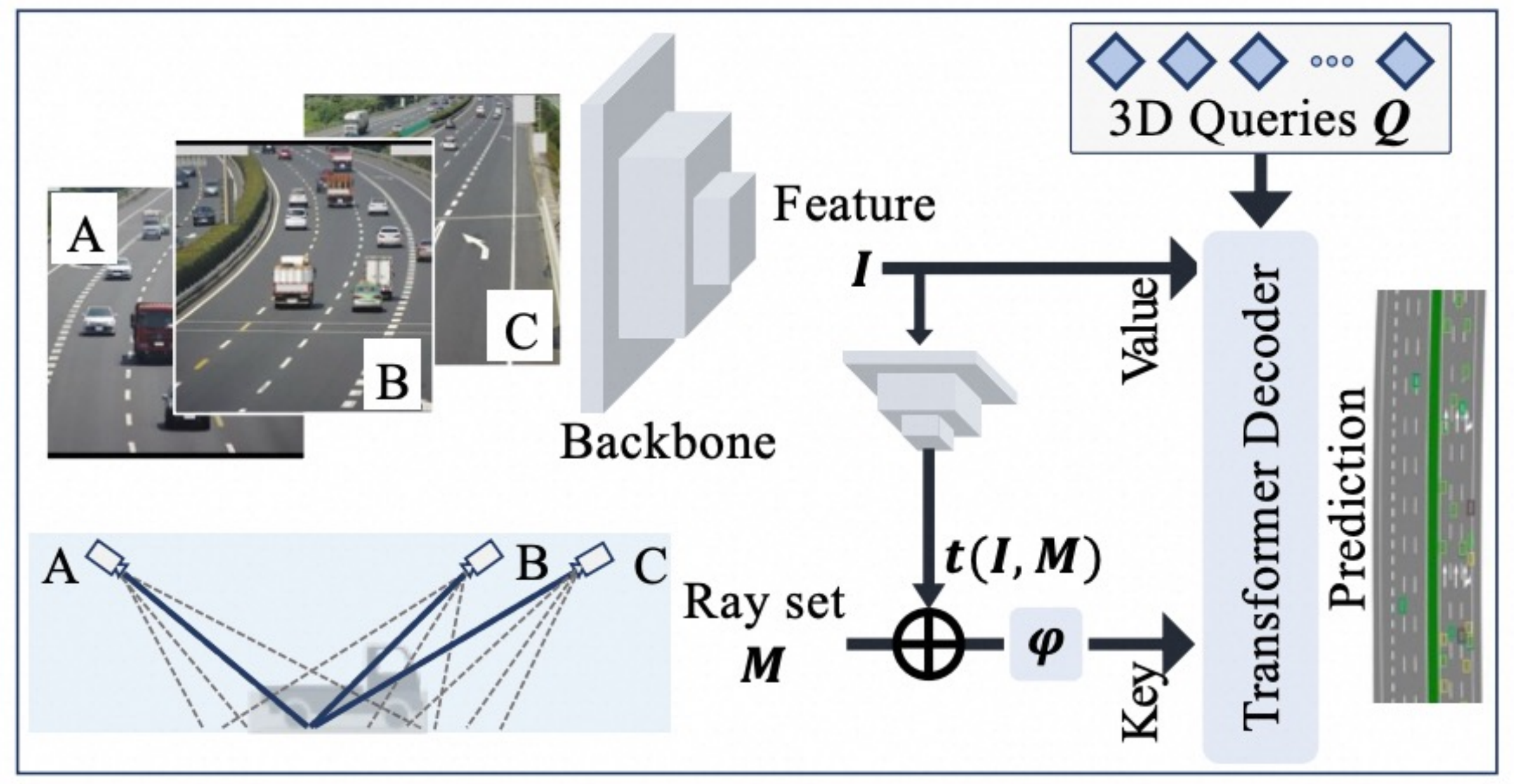


Multi-view images under different conditions. Connected vehicles are identical.

RoBEV Perception Algorithm



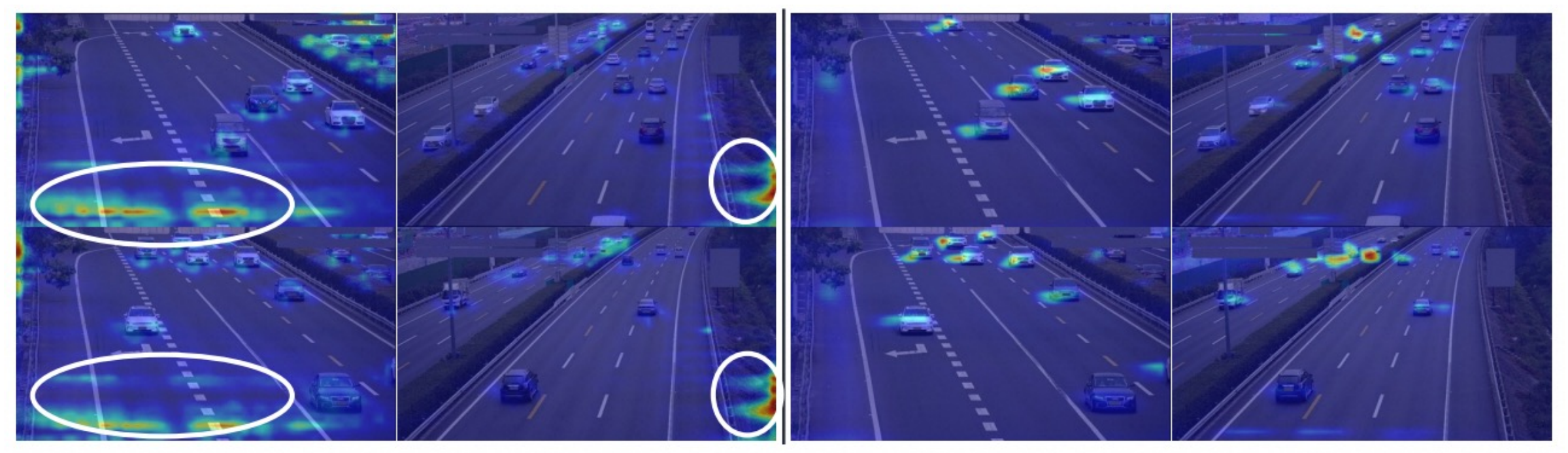
Previous methods learn positional encoding $\phi(\cdot)$ which constrained to the specific camera view, without taking into account the variations of different camera layouts



Our RoBEV has an enhanced feature-guided position embedding that leverages contextual information from images to augment feature assignment process.

Method	Easy NDS	Easy mAP	Easy mATE	Easy mASE	Easy mAOE	Hard NDS	Hard mAP	Hard mATE	Hard mASE	Hard mAOE	*Avg.* NDS
BEVDet [18]	0.506	0.299	0.742	0.079	0.042	0.445	0.184	0.754	0.087	0.043	0.476
BEVDet4D [17]	0.428	0.200	0.896	0.094	0.041	0.393	0.139	0.922	0.099	0.038	0.411
SOLOFusion [32]	0.308	0.129	0.878	0.144	0.517	0.202	0.066	0.844	0.144	1.000	0.255
BEVFormer [23]	0.693	0.609	0.560	0.078	0.030	0.597	0.433	0.600	0.090	0.029	0.645
DETR3D [43]	0.722	0.644	0.501	0.067	0.031	0.633	0.471	0.508	0.080	0.028	0.678
PETRv2 [27]	0.674	0.587	0.590	0.090	0.032	0.580	0.414	0.633	0.100	0.029	0.627
StreamPETR [41]	0.619	0.513	0.690	0.102	0.032	0.496	0.284	0.739	0.107	0.031	0.558
RoBEV (Ours)	**0.753**	**0.684**	0.442	0.058	0.031	**0.672**	**0.524**	0.438	0.077	0.027	**0.713**

Performance comparison of BEV methods on RoScenes dataset. RoBEV achieve state-of-the-art performance.



Attention heatmap visualization. PETRv2 has static artifacts in heatmap across scenes and cameras.