V2XScenes: A Multiple Challenging Traffic Conditions Dataset for Large-Range Vehicle-Infrastructure Collaborative Perception

Supplementary Material

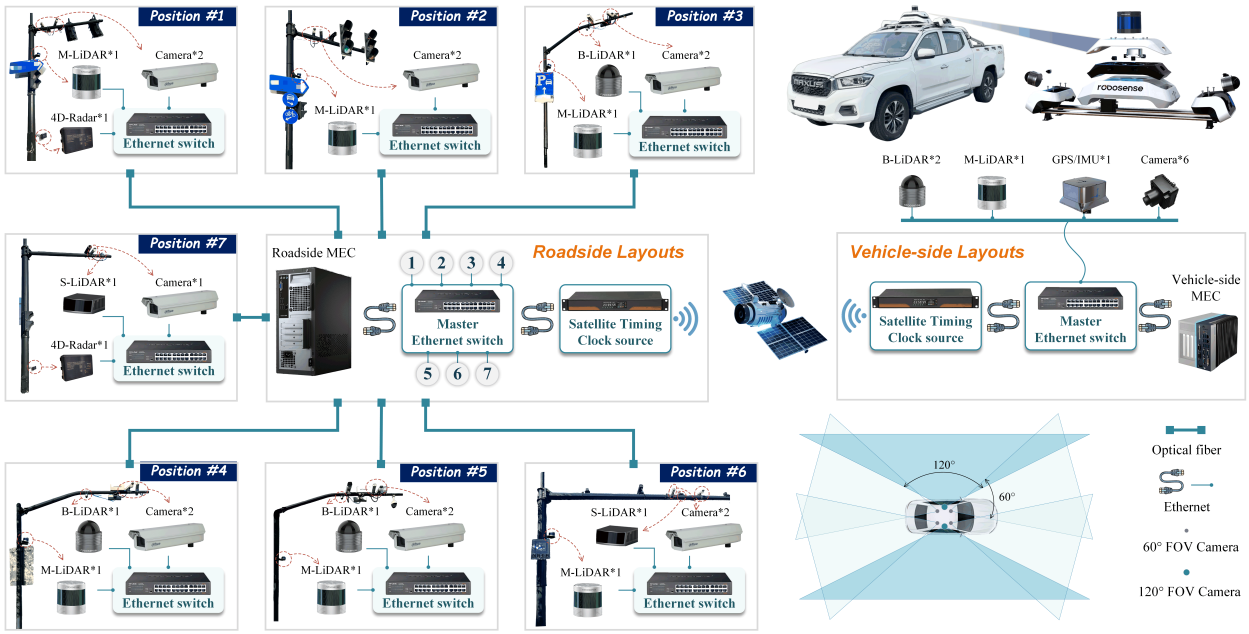


Figure 8. **Detailed sensor layouts for both roadside and vehicle side view in V2XScenes.** We illustrate the roadside and vehicle layouts at left and right respectively. All devices are synchronized via satellite timing. Also, we show the relative mounting position of 4 60°-FOV cameras and 2 120°-FOV cameras in the vehicle-side layouts. M-LiDAR, B-LiDAR and S-LiDAR are the abbreviations of mechanical rotating LiDAR, blind repair LiDAR and solid-state LiDAR respectively. MEC represents the Mobile Edge Computing module.

Contents

A. Outline

In this supplementary material, we add more information and visualization results to demonstrate the data diversity and scenario complexity of V2XScenes. We first present more detailed sensor configurations and layouts for both roadside and vehicle-side view. Then, more analysis and visualization of the calibration and annotation are summarized. Afterwords, additional benchmark qualitative results and visualizations for 3D cooperative perception and tracking under various conditions in V2XScenes are also given in the final part. This supplementary material is organized as shown in the contents. The V2XScenes dataset and benchmark codes will be released.

B. More Details of the Multi-Sensor Layouts

B.1. Details of Sensor Layouts and Configurations

The specific sensors mounting location of each roadside and vehicle-side position are illustrated in Fig. 8. For the roadside devices at each position, all sensor data are connected to the Ethernet switch, and converted into a centralized optical fiber signal based on an Ethernet/Optical-converter module. We employ this data transmission method for all

Table 5. **The sensor configuration in V2XScenes.** The abbreviation of the sensor name can be seen in Fig 8. For the view of infrastructure, we summarize sensors in overall seven positions in the road section.

View	Sensor	Number	Details
Infrastructure	M-LiDAR	6	128beams; 10hz; Distance range 200m; Horizonal FOV [-180°, 180°]; Vertical FOV [-20°, 20°].
	S-LiDAR	2	125beams; 10hz; Distance range 200m; Horizontal FOV [-60°, 60°]; Vertical FOV [-12.5°, 12.5°].
	B-LiDAR	3	32beams; 10hz; Distance range 100m; Horizontal FOV [-180°, 180°]; Vertical FOV [0°, 90°].
	Camera	13	RGB, 60hz, Resolution 2560×1440; Horizontal FOV [-43°, 43°].
	4D-Radar	2	20hz; Detection Radius 300m; Horizontal FOV [-70°, 70°]; Vertical FOV [-20°, 20°]
Vehicle	M-LiDAR	1	128beams; 10hz; Distance range 200m; Horizontal FOV [-180°, 180°]; Vertical FOV [-20°, 20°].
	B-LiDAR	2	32beams; 10hz; Distance range 100m; Horizontal FOV [-180°, 180°]; Vertical FOV [0°, 90°].
	120° Camera	2	RGB, 30hz, Resolution 1920×1080; Horizontal FOV [-60°, 60°].
	60° Camera	4	RGB, 30hz, Resolution 1920×1080; Horizontal FOV [-30°, 30°].

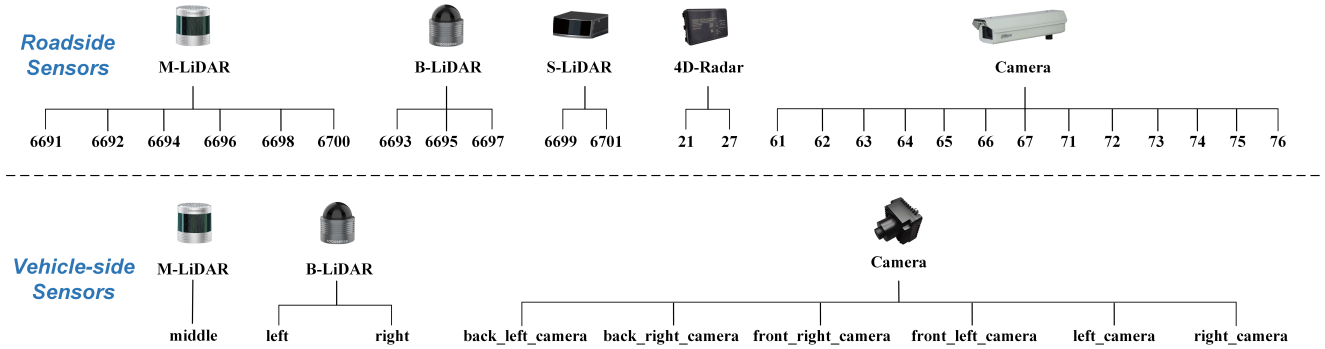


Figure 9. **Specific index for roadside and vehicle-side sensors.** All the given index are used in the benchmark codes to determine the designated sensors' data. The abbreviations are the same to Fig. 8.

Figure 10. **Mounting position of the corresponding roadside sensors with specific index.** The different colors of the numbers represent the various type of the roadside sensors, i.e., five types of the roadside sensors include three types of LiDARs, one type of Radar and one type of camera.

seven roadside position, and the overall collected data are ultimately transmitted to the main switch connected with the roadside MEC. The detailed parameters of all sensors in V2XScenes are listed in Tab 5. Also, we give the following indexes for each sensor to facilitate data organization and analysis in benchmark code as shown in Fig 9, and Fig 10 presents the corresponding mounting position with the index of roadside sensors.

B.2. Visualization for Roadside 4D Radar

Compared to other sensors such as LiDAR and camera, 4D high-resolution millimeter-wave Radar has stable detective capabilities, which is not affected by the variation of adverse weather such as fog, heavy rain and air pollution. Due to its robustness under extreme conditions, it has attracted a wide attention from both academia and industry. Currently, 4D high-precision millimeter wave radar is commonly applied and investigated in vehicle-side automated driving. However, there is no autonomous driving-oriented dataset providing 4D millimeter wave radar data under roadside scenarios yet, especially with respect to some challenging scenarios. The potential advantages of roadside 4D millimeter wave Radar for cooperative vehicle-infrastructure cooperative perception still remain to be further explored. To this end, our V2XScenes provides two 4D millimeter wave Radars at both two intersections mounting with a face-to-face opposite direction. Furthermore. Also, we collected data in multiple weather conditions for comparison.

Fig 11 illustrates the visualization of 4D Radar under the weather of sunny and rainy respectively. We can see that in contrast to LiDAR point clouds, 4D millimeter-wave radar point clouds are unable to capture the fine details and contours of detected objects. For each object with 3D bounding boxes, the point clouds from 4D Radar perform a clustered , which can be enhanced through sequential frame-based compensation. From the figure, the roadside images un-

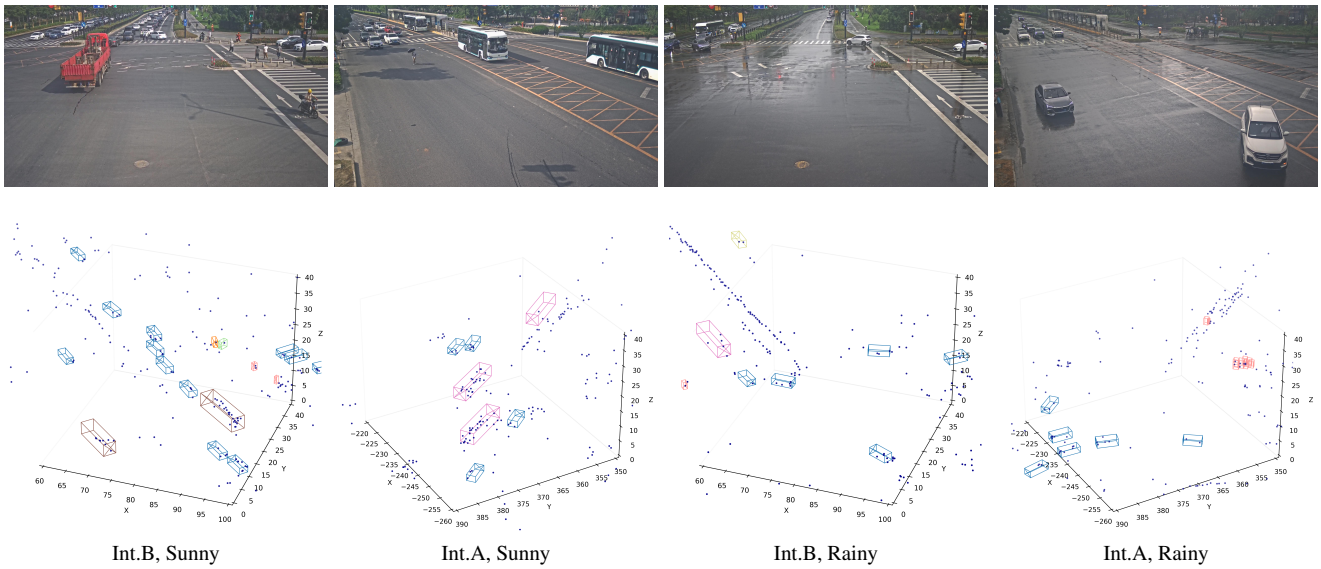


Figure 11. **Visualization of 4D Radar under different weather.** Int.A and Int.B represent the intersection A and B respectively. We use different colors of 3D bounding boxes to illustrate the various label classes, where the 'X' denotes the moving orientation of the bounding boxes. The main categories are colored as follows: **Car**, **Motorcycle**, **Pedestrian**, **Truck**, **Bus**, **Trailer**, **Van**.

der different weather conditions show significant environmental variations, where the rain on the ground and in the air can also affect the LIDAR reflections. In contrast, the 4D millimeter-wave Radar demonstrates stable perception across both weather scenarios. Additionally, The 4D Radar requires much less data to characterize the same object as the LiDAR although the detailed contours are inadequately, which achieves higher data efficiency and reduces the computational load. In V2XScenes, different types of LiDAR are also provided under the same view, we aim to investigate the advantages of integrating diverse roadside sensors in enhancing single-vehicle autonomous driving.

C. More Details of the Calibration

C.1. Roadside Multi-modal Calibration

Most of the existing calibration boards are usually applied for vehicle-side sensors, thus it is difficult to acquire a large enough of features under roadside perspective based on the commonly used calibration plane. Therefore, we design a customized calibrators with roadside-oriented plane to obtain multi-modal features for aligning the coordinates, and calculating the transformation matrix of $\boldsymbol{T}_{\text{Lidar2Cmaera}}$ between camera and LiDAR. The camera and LiDAR are mounted at a height of approximately 5m, and the customized plane is in the shape of an equilateral rectangular triangle with each length of 2m. We place the calibrators under the same view of LiDAR and camera, and collect several pairs of related features corresponding to the different modal data. Fig. 13 demonstrates the LiDAR-Camera calibration results for all 14 combinations of roadside sensors.

Algorithm 1: Calculation process of $\boldsymbol{T}_{\text{Vehicle2World}}$

Input: Dense map $\boldsymbol{P}_{\text{dense}}$, Vehicle original point cloud $\boldsymbol{P}_v$, Vehicle pose $\boldsymbol{T}^{\text{LLA}}_{\text{veh},i}$ and $\boldsymbol{T}^{\text{LLA}}_{\text{veh},o}$, Sequence length T_{max}, Matrix of $\boldsymbol{T}^{o}_{\text{VehicleLiDAR2IMU}}$ and $\boldsymbol{T}_{\text{LLA2ENU}}$.

Output: $\boldsymbol{T}_{\text{Vehicle2World}}$ for the specific sequence.

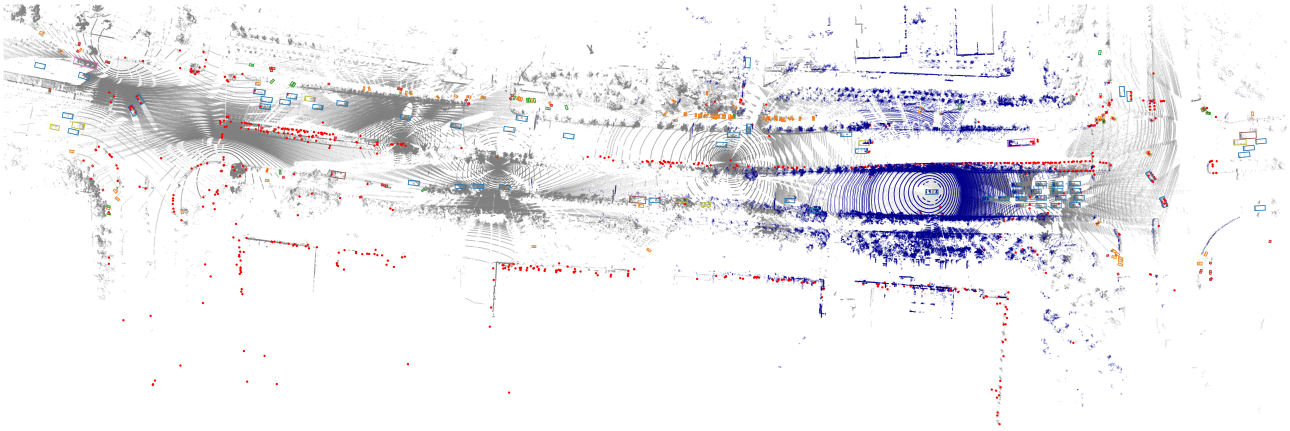
while *i is less than the sequence length T_{max}* **do**
- $\boldsymbol{T}^{\text{ENU}}_{\text{veh},i} \leftarrow \boldsymbol{T}_{\text{LLA2ENU}} \times \boldsymbol{T}^{\text{LLA}}_{\text{veh},i}$
- $\boldsymbol{T}^{\text{ENU}}_{\text{veh},o} \leftarrow \boldsymbol{T}_{\text{LLA2ENU}} \times \boldsymbol{T}^{\text{LLA}}_{\text{veh},o}$
- $\delta\boldsymbol{T}^{\text{ENU}}_{\text{veh},i} \leftarrow \|\boldsymbol{T}^{\text{ENU}}_{\text{veh},i} - \boldsymbol{T}^{\text{ENU}}_{\text{veh},o}\|$
- $\boldsymbol{T}^{i}_{\text{VehLiDAR2IMU}} \leftarrow \delta\boldsymbol{T}^{\text{ENU}}_{\text{veh}} \times \boldsymbol{T}^{o}_{\text{VehLiDAR2IMU}}$
- $\{\boldsymbol{T}^{i,0}_{\text{VehLiDAR2IMU}}, \boldsymbol{P}^{i,0}_{v}\} \leftarrow \{\boldsymbol{T}^{i}_{\text{VehLiDAR2IMU}}, \boldsymbol{P}^{i}_{v}\}$
- **for** *Optimized iteration* $k \in \{0, 1, \cdots N\}$ **do**
  - $\bar{\boldsymbol{P}}^{i,k}_{v} \leftarrow \boldsymbol{T}^{i,k}_{\text{VehLiDAR2IMU}} \times \boldsymbol{P}^{i,k}_{v}$
  - Solve $\min^{\text{ICP}}_{\boldsymbol{T}^{i,k}_{\text{VehLiDAR2IMU}}} \|\boldsymbol{P}_{\text{dense}} - \bar{\boldsymbol{P}}^{i,k}_{v}\|$
  - Obtain the updated matrix of $\boldsymbol{T}^{k+1}_{\text{VehLiDAR2IMU}}$
  - $\boldsymbol{P}^{i,k+1}_{v} \leftarrow \boldsymbol{T}^{i,k+1}_{\text{VehLiDAR2IMU}} \times \boldsymbol{P}_{v,i}$
- **end**
- $\boldsymbol{T}_{\text{Vehicle2World}} \leftarrow \boldsymbol{T}_{\text{Vehicle2World}} \cup \boldsymbol{T}^{i,N}_{\text{VehLiDAR2IMU}}$

end

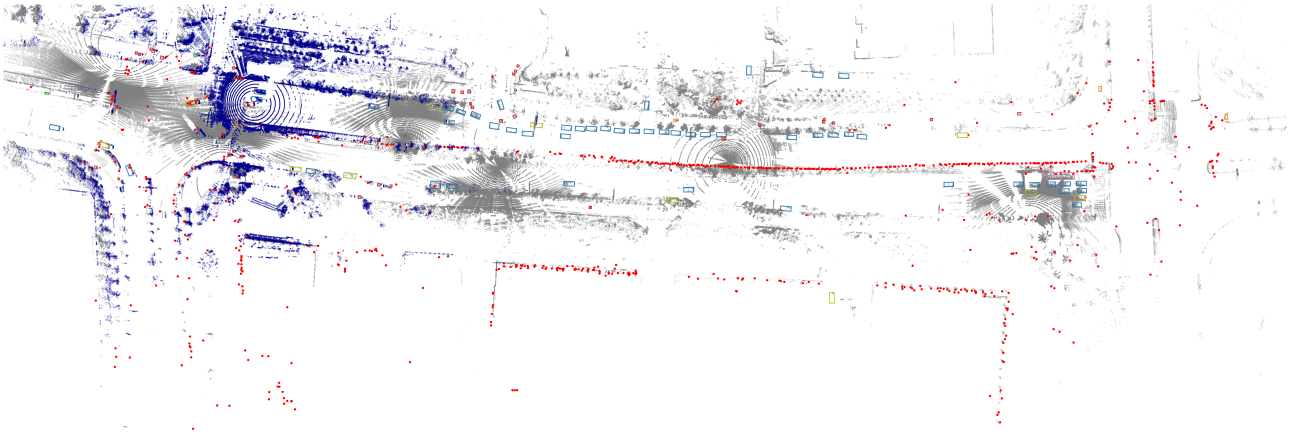
Return the optimized $\boldsymbol{T}_{\text{Vehicle2World}}$

C.2. Detailed Process of LiDAR Fusion

In this part, we give the detailed calculation process of $\boldsymbol{T}_{\text{Vehicle2World}}$ for each data sequence. To start with, the vehicle real-time poses are obtained based on the infor-



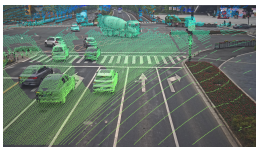
(a) Scene: *facing congestion on a sunny day*



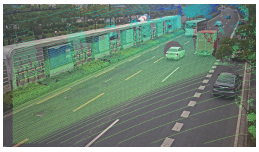
(b) Scene: *Motorbike cut-through during the vehicle U-turn on a rainy day*

Figure 12. **Visualization of data fusion from all point-based sensors under BEV view.** Two key moment of 1720517415.500074 and 1720754211.700128 with different scenes are presented. The **gray points** and **blue points** represent the roadside and vehicle-side LiDAR respectively. The **red points** are denoted as the roadside 4D Radar. We also plot the ground truth labels of 3D bounding boxes with different colors, where each color is denoted as a specific category as the same as Fig. 11.

(a) LiDAR6699 to Camera66



(b) LiDAR6700 to Camera66



(c) LiDAR6692 to Camera62

(d) LiDAR6691 to Camera61

Figure 13. **Roadside LiDAR/Camera calibration results**. The various point color represent the relative distance to the origin.

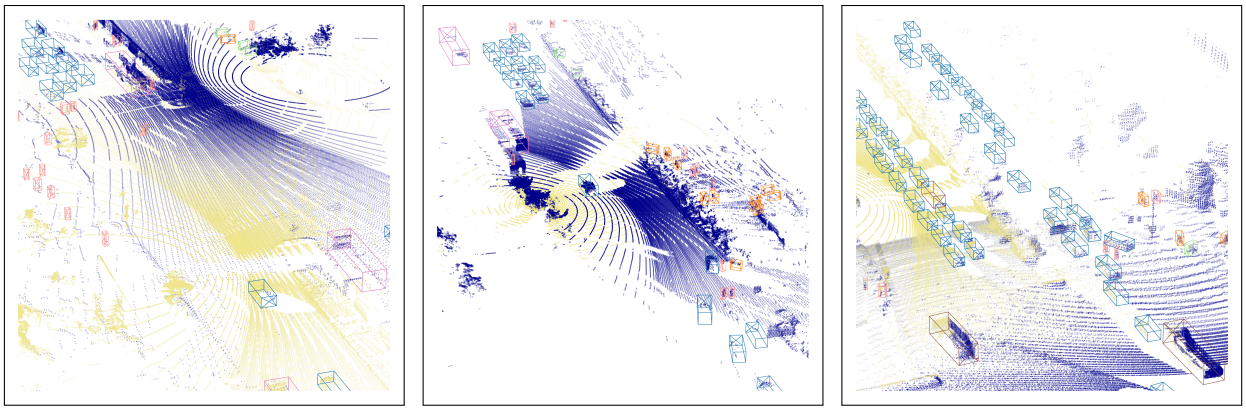
mation of GPS/IMU, we denote the location as $\boldsymbol{T}_{\mathsf{GPS}} = [T_{\mathsf{Long}}, T_{\mathsf{Lat}}, T_{\mathsf{Alt}}]$ in Longitude-Latitude-Altitude (LLA) coordinate and the orientation as $\boldsymbol{R}_{\mathsf{IMU}}^{1\times4} = [q_x, q_y, q_z, q_w]$ based on a quaternion form. In the following, we use $\boldsymbol{T}_{\mathsf{veh},i}^{\mathsf{LLA}}$ to depict the vehicle pose at i^{th} frame in LLA coordinate, which can be written as:

$$\boldsymbol{T}_{\mathsf{veh},i} = \begin{bmatrix} \boldsymbol{R}_{\mathsf{IMU}}^{3\times3} & \boldsymbol{T}_{\mathsf{GPS}} \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} R_{11} & R_{12} & R_{13} & T_{\mathsf{Long}} \\ R_{21} & R_{22} & R_{23} & T_{\mathsf{Lat}} \\ R_{31} & R_{32} & R_{33} & T_{\mathsf{Alt}} \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{1}$$

Where the $\boldsymbol{R}_{\mathsf{IMU}}^{3\times3}$ can be calculated by:

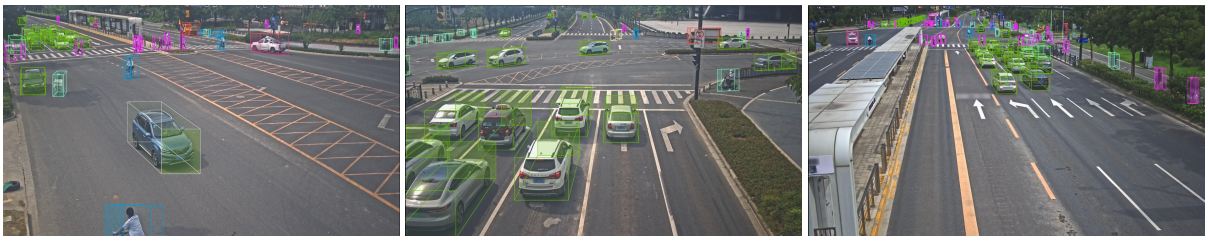
$$\begin{bmatrix} 1-2(q_y^2+q_z^2) & 2(q_xq_y-q_zq_w) & 2(q_xq_z+q_yq_w) \\ 2(q_xq_y+q_zq_w) & 1-2(q_x^2+q_z^2) & 2(q_yq_z-q_xq_w) \\ 2(q_xq_z-q_yq_w) & 2(q_yq_z+q_xq_w) & 1-2(q_x^2+q_y^2) \end{bmatrix} \tag{2}$$

To obtain the transformation matrix of $\boldsymbol{T}_{\mathsf{Vehicle2World}}$ in time-varying, we firstly need to determine an original vehicle pose in the road section. We random collect the vehicle GPS/IMU information of $\boldsymbol{T}_{\mathsf{veh},o}^{\mathsf{LLA}}$ under a static state



(a) Intersection A (6691, **6692**) (b) Corridor (6693, **6694**) (c) Intersection B (6699, 6700, **6701**)

Figure 14. **Visualization of the annotations in roadside LiDARs.** We illustrate three combinations of different types of LiDARs, where the different point clouds under each layout are colored by Khaki, **Dark Blue** and Gray.



(a) Camera 61 (b) Camera 66 (c) Camera 73

Figure 15. **Visualization of the annotations in roadside cameras.** Three typical views from roadside cameras with labeled 3D bounding boxes are selected for illustration, where the dark transparent side represents the moving orientation.

as the initial pose matrix. Then we use the calibrator to calculate the transformation matrix of $\boldsymbol{T}^{o}_{\text{VehLiDAR2IMU}}$ which can project the LiDAR points to the IMU coordinate system at this specific moment. The following calculation process is demonstrated in Alg. 1. Directly using the matrix may cause large projecting error due to inaccurate GPS/IMU information, hence we apply the optimization method of ICP to reduce the calibration errors as shown in the line 9 of Alg. 1. Particularly, we build a road map with dense point cloud, which is regarded as the high precision reference for aligning both the roadside and vehicle-side point cloud data.

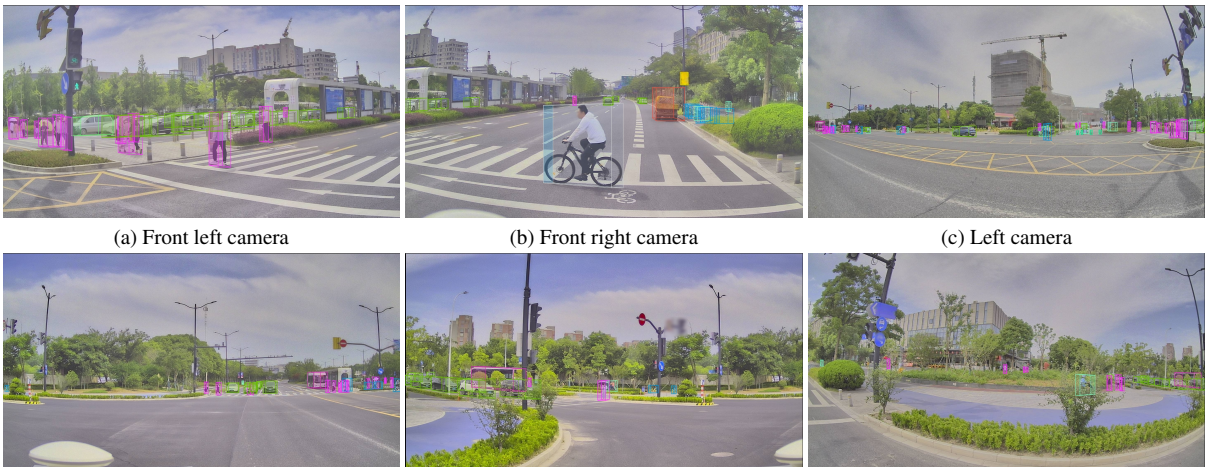
We provide more visualization of the finial point cloud fused results as shown in Fig. 12 by projecting the vehicle-side LiDAR to the roadside combined point cloud using the refined $\boldsymbol{T}_{\text{Vehicle2World}}$ and $\boldsymbol{T}_{\text{Lidar2World}}$.

D. More Details of the Annotation

V2XScenes have 332596 labeled 3D bounding boxes in total based on human expert experiences with various categories. Fig 14 illustrates three typical layouts of sensors overlapping with labeled data. In general, the ground truth of 3D bounding boxes in camera view can be obtained based on the calibration results. Considering the calibrating errors and the different field of view overlapping rates, we also annotate 315439 labels for roadside and vehicle-side camera by employing the expert annotators. Fig 15 and Fig 16 demonstrate the annotation results of 3D bounding boxes under some typical camera views.

Detailed statistics of the labeled ground truth with different categories are demonstrated in Fig. 17, we summarize the 8 categories into four main class of "*pedestrian*", "*car*", "*bus*" and "*bicycle*". We can see the object distribution of the overall road section from the estimation of distance density, where the defined original point of the zero distance is near to the intersection B. For instance, the category of "*car*" has a larger density under the range of 0-300m, which shows the vehicle flow is usually greater than intersection B. In addition, the distance distribution of "*pedestrian*" and "*bicycle*" are more focused in the middle, which is consistent of the common fact.

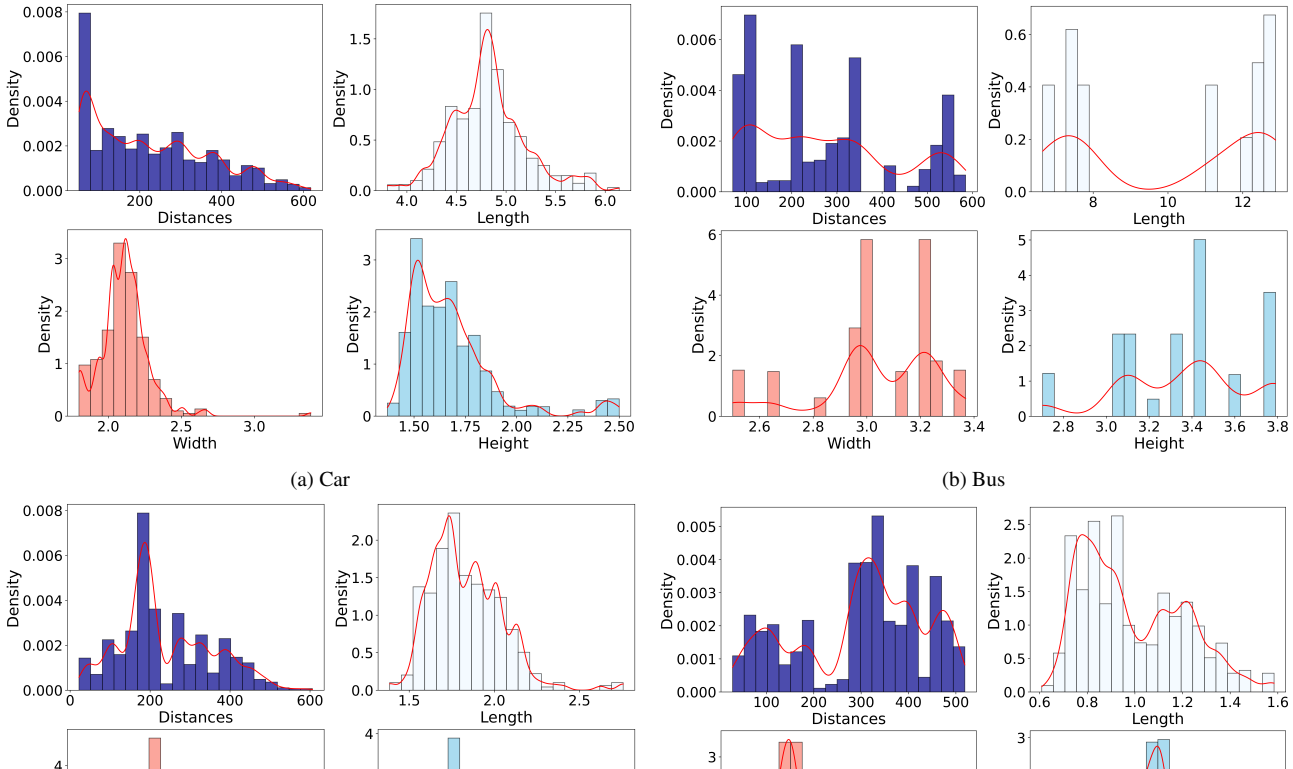
V2XScenes has an average of around 10k and 3k 3D bounding boxes for each scene under roadside and vehicle-side view respectively. Our V2XScenes is characterized with a wide variety and quantity of roadside participants due to the large perception range.



(a) Front left camera (b) Front right camera (c) Left camera

(d) Back left camera (e) Back right camera (f) Right camera

Figure 16. **Visualization of the annotations in vehicle-side camera.**



(a) Car (b) Bus

(c) Bicycle (d) Pedestrian

Figure 17. **Distribution of labeled bounding boxes for Car, Bus, Bicycle and Pedestrian.** The red line represent the density estimation based on the Gaussian kernel. We calculate the distances between the origin position and labeled positions in global coordinates.

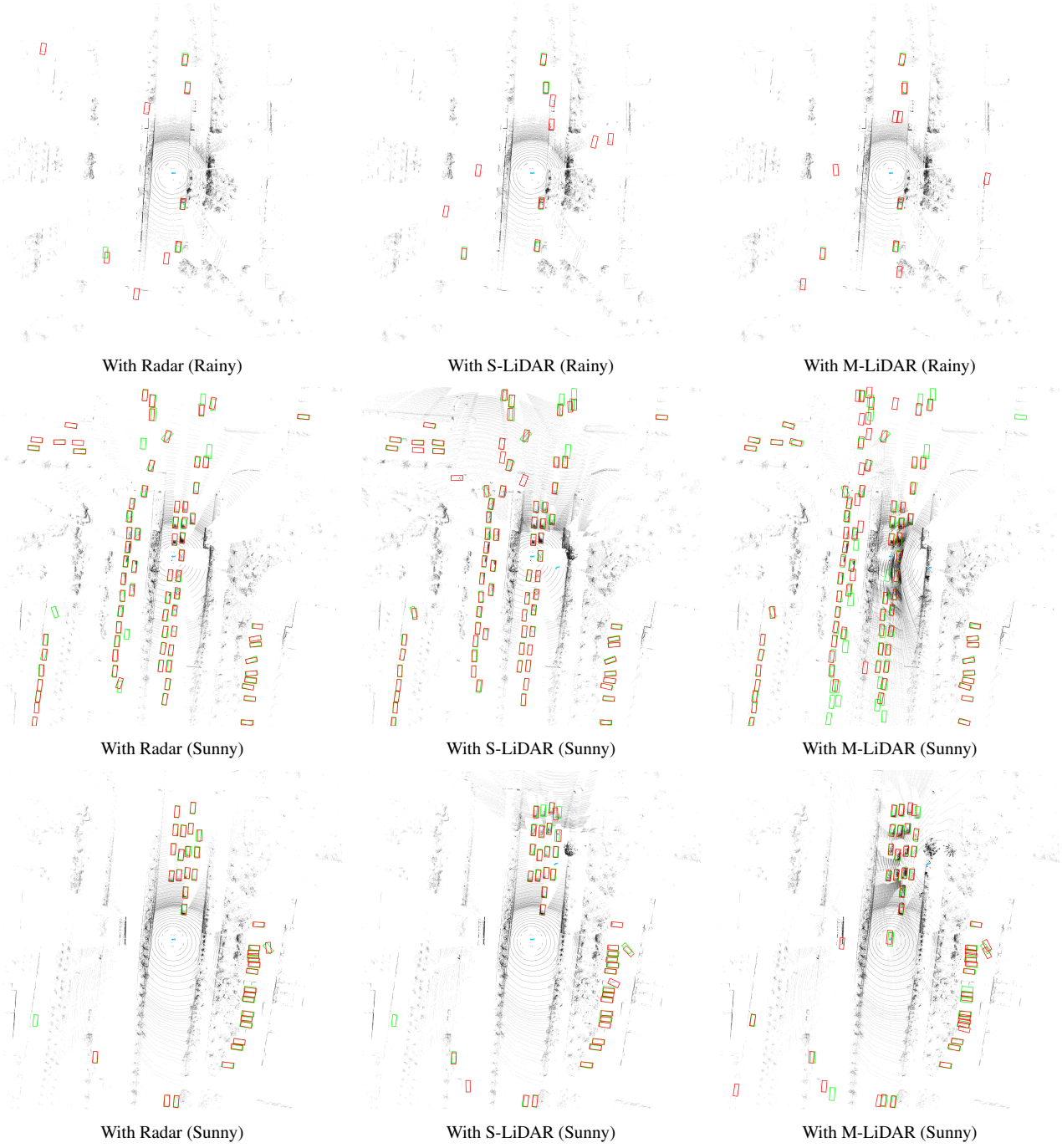


Figure 18. **Visualization of collaborating with heterogeneous data under different weather conditions.** The fused point cloud is colored by black. The ground truth and predictions are represented by **green** and **red** respectively.

E. More Experiments and Visualization

E.1. More Experimental Results

To investigate the impact of heterogeneous data (e.g. different types of point cloud data) on the collaboration perception, we add ablation experiments using three different point cloud (from 128-beams mechanical rotating LiDAR, 125-beams solid-state LiDAR and 4D millimeter-wave Radar) in the intersection B of V2XScenes. We set up the same intermediate fusion method to compare the results of fusion perception by providing different types of roadside point-based data. The results of ablation are presented in Tab. 6

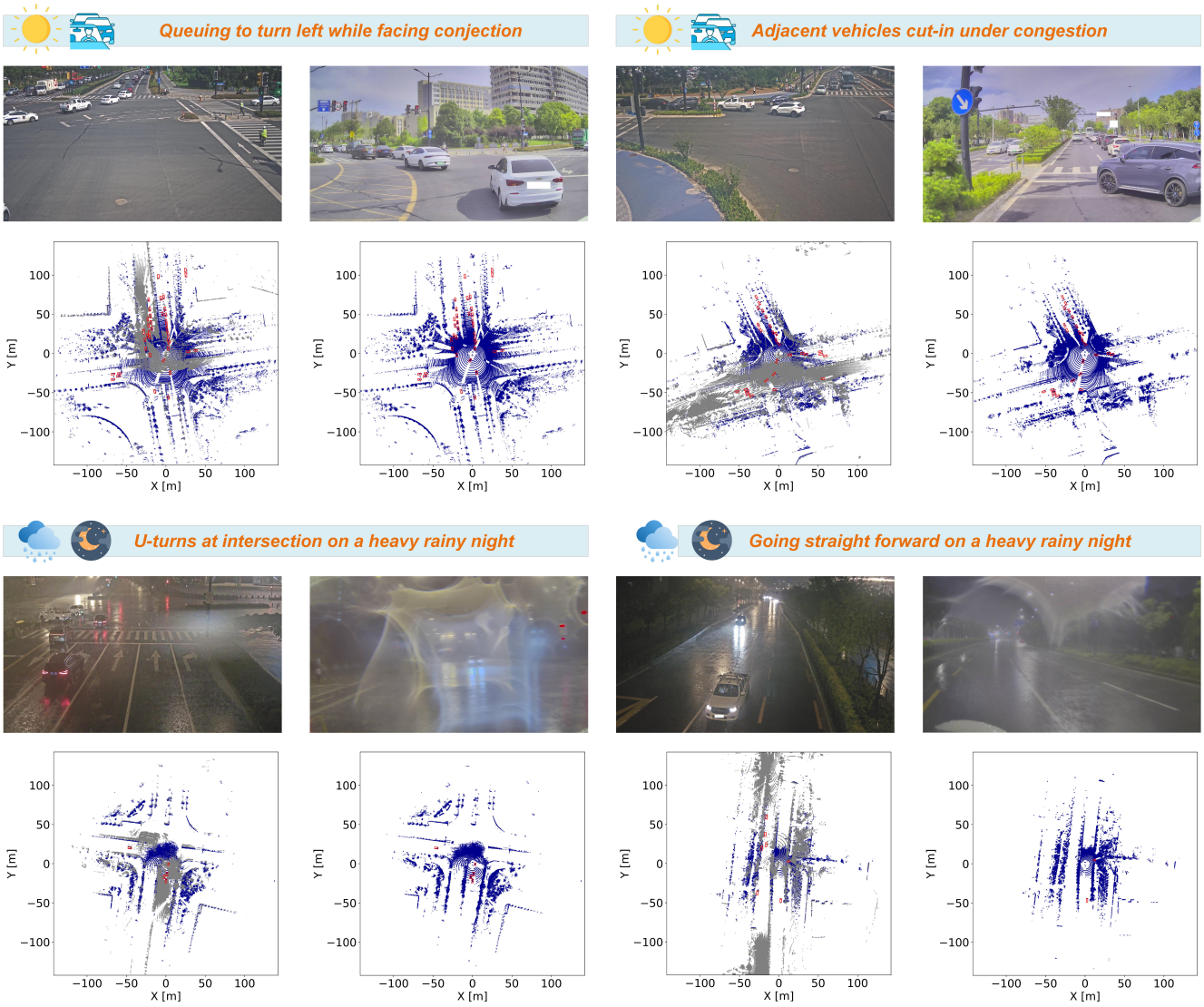


Figure 19. **Illustration of more example scenes under multiple challenging conditions.**

and we also provide the visualization of cooperative 3D object detection under different fused heterogeneous data as shown in Fig. 18.

E.2. More Examples of Multi-Condition Scenes

In this section, we provide more example scenarios of the multi-condition challenging data sequences in V2XScenes. As shown in Fig. 19, each scene is labeled with a specific description, where it can be observed that we collect a high quantity of traffic congestion scenarios under different weather conditions. For instance, the camera of the vehicle is basically covered by water when driving in heavy rainy night, which seriously affects driving safety. Despite this, the roadside sensors can still provide the key perception information for single vehicle to ensure the safety.

Table 6. **Ablation study on collaboration with heterogeneous data.** "S" and "M" denote solid-state LiDAR and mechanical LiDAR, respectively. CoBEVT is used for all settings.

Fused source	Car AP_{3D} of Sunny		
	IoU=0.3	IoU=0.5	IoU=0.7
With Radar	92.08%	89.16%	67.26%
With S-LiDAR	91.04%	88.81%	72.64%
With M-LiDAR	82.54%	76.97%	50.15%
Fused source	Car AP_{3D} of Rainy		
	IoU=0.3	IoU=0.5	IoU=0.7
With Radar	46.12%	27.60%	11.74%
With S-LiDAR	46.26%	34.19%	16.41%
With M-LiDAR	50.66%	34.98%	11.51%