HoloVIC: Large-scale Dataset and Benchmark for Multi-Sensor Holographic Intersection and Vehicle-Infrastructure Cooperative

Supplementary Material

A. Coordinates Transformation

In Sec.3.2, we introduced all the coordinate systems involved in HoloVIC, as shown in Fig.**??**. The coordinate systems include the Lidar Coordinate (l_x, l_y, l_z), Camera Coordinate (c_x, c_y, c_z), and Pixel Coordinate (u, v), which represent the positions in their respective sensors. The Intersection Coordinate $(\sigma_x, \sigma_y, \sigma_z)$ and Ego-Vehicle Coordinate $(\delta_x, \delta_y, \delta_z)$ are used for unifying the coordinate systems of their respective sensors, and the Global Coordinate $(\omega_x, \omega_y, \omega_z)$ mainly aligns the Intersection Coordinate and Vehicle Coordinate. The transformation relationships between all coordinate systems are shown in Fig.A1, where which include both forward and inverse transformations for five processes.

A.1. Lidar⇔Intersection/Vehicle

Intersection to Lidar: Given a point in the Intersection Coordinate: $(\sigma_x, \sigma_y, \sigma_z)$, the transformation of this point from the Intersection Coordinate System to the Lidar Coordinate (l_x, l_y, l_z) is defined as follows:

$$\begin{pmatrix} l_x \\ l_y \\ l_z \\ 1 \end{pmatrix} = RT_{I2L} \begin{pmatrix} \sigma_x \\ \sigma_y \\ \sigma_z \\ 1 \end{pmatrix}, \tag{1}$$

where $RT_{I2L} \in \mathbb{R}^{4\times4}$ is a Rotational Translation of a homogeneous matrix for Intersection Coordinate to Lidar Coordinate, which is formulated as:

$$RT_{I2L}^{4\times4} = \begin{bmatrix} R_L^{3\times3} & T_L^{3\times1} \\ 0 & 1 \end{bmatrix} \tag{2}$$

Lidar to Intersection: Given a point in Lidar Coordinate (l_x, l_y, l_z), the transformation of the point to Intersection Coordinate $(\sigma_x, \sigma_y, \sigma_z)$ is defined as:

$$\begin{pmatrix} \sigma_x \\ \sigma_y \\ \sigma_z \\ 1 \end{pmatrix} = RT_{L2I} \begin{pmatrix} l_x \\ l_y \\ l_z \\ 1 \end{pmatrix}, \quad RT_{L2I} = RT_{I2L}^{-1} \tag{3}$$

where $RT_{L2I} \in \mathbb{R}^{4\times4}$ is the inverse of $RT_{I2L} \in \mathbb{R}^{4\times4}$.

Vehicle to Lidar: Similar to Eq.(1) and Eq.(2), a point $(\delta_x, \delta_y, \delta_z)$ in the Vehicle Coordinate is transformed by using Eq.(1) through the $RT_{V2L} \in \mathbb{R}^{4\times4}$ to obtain the point in the Lidar Coordinate of Vehicle (l_x, l_y, l_z).

Lidar to Vehicle: Similar to Eq.(3), a point (l_x, l_y, l_z) in the Lidar Coordinate of Vehicle is transformed by using Eq.(3) through the $RT_{L2V} \in \mathbb{R}^{4\times4}$ to obtain the point in the Vehicle Coordinate $(\delta_x, \delta_y, \delta_z)$, where RT_{L2V} is the inverse of RT_{V2L}.

A.2. Camera⇔Intersection/Vehicle

Intersection to Camera: Given a point in Intersection Coordinate $(\sigma_x, \sigma_y, \sigma_z)$, the transformation of the point to Camera Coordinate (c_x, c_y, c_z) is defined as:

$$\begin{pmatrix} c_x \\ c_y \\ c_z \\ 1 \end{pmatrix} = S^{4\times4} RT_{I2C} \begin{pmatrix} \sigma_x \\ \sigma_y \\ \sigma_z \\ 1 \end{pmatrix} \tag{4}$$

where $RT_{I2C} \in \mathbb{R}^{4\times4}$ is a Rotational Translation of a homogeneous matrix for Intersection Coordinate to Camera Coordinate, and S is utilized for mapping coordinate axes $(X \to Z, Y \to -X, Z \to -Y)$, which are formulated as:

$$RT_{I2C}^{4\times4} = \begin{bmatrix} R_C^{3\times3} & T_C^{3\times1} \\ 0 & 1 \end{bmatrix}, \quad S^{4\times4} = \begin{bmatrix} 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{5}$$

In addition, we define the ground plane vector $\phi_C^{1\times4}$ and $\phi_I^{1\times4}$ relative to the Camera Coordinate and Intersection Coordinate, respectively. And we define the the ground plane in Intersection is $Z = 0$, and then the ϕ_I is $[0, 0, 1, 0]$. The points set in ground plane of Intersection Coordinate $\{p : (p_x, p_y, p_z), p_z = 0\}$ and the points set in ground plane of Camera Coordinate $\{q : (q_x, q_y, q_z), q_z = 0\}$ satisfy:

$$\phi_I^{1\times4} \begin{pmatrix} p_x \\ p_y \\ p_z \\ 1 \end{pmatrix} = 0, \quad \phi_C^{1\times4} \begin{pmatrix} q_x \\ q_y \\ q_z \\ 1 \end{pmatrix} = 0 \tag{6}$$

The points in Intersection Coordinate $p : (p_x, p_y, p_z)$ are transformed to Camera Coordinate $q : (q_x, q_y, q_z)$ by Eq.(4):

$$\begin{pmatrix} q_x \\ q_y \\ q_z \\ 1 \end{pmatrix} = S^{4\times4} RT_{I2C} \begin{pmatrix} p_x \\ p_y \\ p_z \\ 1 \end{pmatrix} \tag{7}$$

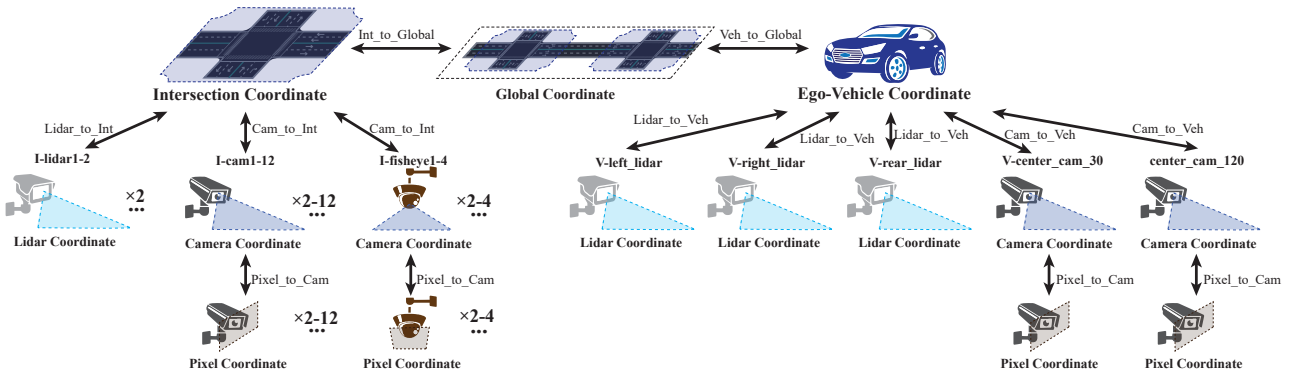

Figure A1. The transformation relationships between all coordinate systems in HoloVIC.

Combining Eq.(6) and Eq.(7) we obtain:

$$\phi_I^{1\times4}[S^{4\times4}RT_{I2C}]^{-1}\begin{pmatrix}q_x\\q_y\\q_z\\1\end{pmatrix}=0 \tag{8}$$

thus we derive the ground plane vector ϕ_C is equal to:

$$\phi_C^{1\times4}=\phi_I^{1\times4}[S^{4\times4}RT_{I2C}]^{-1} \tag{9}$$

Camera to Intersection: Given a point p_c in the Camera Coordinate: (c_x, c_y, c_z), the transformation of this point from the Camera Coordinate System to the Intersection Coordinate $(\sigma_x, \sigma_y, \sigma_z)$ is defined as follows:

$$\begin{pmatrix}\sigma_x\\\sigma_y\\\sigma_z\\1\end{pmatrix}=RT_{C2I}S^{-1}\begin{pmatrix}c_x\\c_y\\c_z\\1\end{pmatrix},\quad RT_{C2I}=RT_{I2C}^{-1} \tag{10}$$

where $RT_{C2I}\in\mathbb{R}^{4\times4}$ is the inverse of $RT_{I2C}\in\mathbb{R}^{4\times4}$.

Vehicle to Camera: Similar to Eq.(4) and Eq.(5), a point $(\delta_x, \delta_y, \delta_z)$ in the Vehicle Coordinate is transformed by using Eq.(4) through the $RT_{V2C}\in\mathbb{R}^{4\times4}$ to obtain the point in the Camera Coordinate of Vehicle (c_x, c_y, c_z).

Camera to Vehicle: Similar to Eq.(10), a point (c_x, c_y, c_z) in the Camera Coordinate of Vehicle is transformed by using Eq.(10) through the $RT_{C2V}\in\mathbb{R}^{4\times4}$ to obtain the point in the Vehicle Coordinate $(\delta_x, \delta_y, \delta_z)$, where RT_{C2V} is the inverse of RT_{V2C}.

A.3. Pixel⇔Camera

Camera to Pixel: Given a position point p_c in the camera coordinate system: (c_x, c_y, c_z), the projection of this point from the camera coordinate system to undistorted image in the pixel coordinate system is defined as follows:

$$Z_c\begin{pmatrix}u\\v\\1\end{pmatrix}=K^{3\times3}\begin{pmatrix}c_x\\c_y\\c_z\end{pmatrix} \tag{11}$$

where $K\in\mathbb{R}^{3\times3}$ indicates the intrinsic matrix of camera which is calibrated by Chessboard Calibration. f_x, f_y denote the focal of the camera in x-axis, y-axis. u_0, v_0 represent the center of image. Z_c indicates the distance from point p_c to the projection plane of camera.

Pixel to Camera: Since the transformation from the pixel coordinate system to the camera coordinate system is a 2D to 3D process, let's assume that we select the points (u, v) in the image that corresponds to a point on the ground plane in real scene, and it is projected onto the camera coordinate system as (c_x, c_y, c_z). Before the projection, we have to calculate distance between the points on Camera Coordinate to plane of camera Z_c, which is calculated as:

$$Z_c=\frac{-d}{(u[a,b,c]K_{|0}^{-1}+v[a,b,c]K_{|1}^{-1}+[a,b,c]K_{|2}^{-1})} \tag{12}$$

where $\phi_C\in\mathbb{R}^{1\times4}:[a,b,c,d]$ is the ground plane vector for Camera Coordinate, which is introduced in Sec.A.2. $K^{-1}\in\mathbb{R}^{3\times3}$ is the inverse of camera intrinsic K, and in Eq.(12), $K_{|i}^{-1}\in\mathbb{R}^{3\times1}$ indicates the i-th column of the K^{-1}. The transformation from Pixel Coordinate to Camera Coordinate is defined as:

$$\begin{pmatrix}c_x\\c_y\\c_z\\1\end{pmatrix}=\begin{bmatrix}K^{-1}&0\\0&1\end{bmatrix}^{4\times4}\begin{pmatrix}Z_cu\\Z_cv\\Z_c\\1\end{pmatrix} \tag{13}$$

A.4. Global⇔Intersection:

Global to Intersection: Both of Global Coordinate and Intersection belong to East-North-Up (ENU) Coordinate. Given a point in Global Coordinate $(\omega_x, \omega_y, \omega_z)$, the transformation from Global Coordinate to Intersection Coordinate $(\sigma_x, \sigma_y, \sigma_z)$ is defined as:

$$\begin{pmatrix}\sigma_x\\\sigma_y\\\sigma_z\\1\end{pmatrix}=\begin{bmatrix}E^{3\times3}&T_{G2I}^{3\times1}\\0&1\end{bmatrix}\begin{pmatrix}\omega_x\\\omega_y\\\omega_z\\1\end{pmatrix} \tag{14}$$

$$T_{G2I}^{3\times1} = \begin{pmatrix}\omega_{x0}\\ \omega_{y0}\\ \omega_{z0}\end{pmatrix} - \begin{pmatrix}\sigma_{x0}\\ \sigma_{y0}\\ \sigma_{z0}\end{pmatrix} \tag{15}$$

where $E \in \mathbb{R}^{3\times3}$ is a identity matrix, $T_{G2I}^{3\times1}$ indicates the translation matrix between Global Coordinate and Intersection, $(\omega_{x0}, \omega_{y0}, \omega_{z0})$ and $(\sigma_{x0}, \sigma_{y0}, \sigma_{z0})$ denote the original of Global and Intersection, respectively.

Global to Intersection: The transformation from Intersection to Global is formulated as:

$$\begin{pmatrix}\omega_x\\ \omega_y\\ \omega_z\\ 1\end{pmatrix} = \begin{bmatrix}E^{3\times3} & T_{I2G}^{3\times1}\\ 0 & 1\end{bmatrix} \begin{pmatrix}\sigma_x\\ \sigma_y\\ \sigma_z\\ 1\end{pmatrix}, \; T_{I2G}^{3\times1} = -T_{G2I}^{3\times1} \tag{16}$$

A.5. Vehicle⇔Global

Global to Vehicle: Given a point in Global Coordinate $(\omega_x, \omega_y, \omega_z)$, the rotation and translation matrixes are computed according to GPS, orientation and accelerate of the vehicle by RTK and E-Compass. We directly provide the RT_{G2V} matrix from Global Coordinate to Vehicle Coordinate $(\delta_x, \delta_y, \delta_z)$, which is defined as:

$$\begin{pmatrix}\delta_x\\ \delta_y\\ \delta_z\\ 1\end{pmatrix} = \begin{bmatrix}R_{G2V}^{3\times3} & T_{G2V}^{3\times1}\\ 0 & 1\end{bmatrix} \begin{pmatrix}\omega_x\\ \omega_y\\ \omega_z\\ 1\end{pmatrix} \tag{17}$$

Vehicle to Global: The transformation from Vehicle Coordinate $(\delta_x, \delta_y, \delta_z)$ to Global Coordinate $(\omega_x, \omega_y, \omega_z)$ is formulated as:

$$\begin{pmatrix}\omega_x\\ \omega_y\\ \omega_z\\ 1\end{pmatrix} = RT_{V2G} \begin{pmatrix}\delta_x\\ \delta_y\\ \delta_z\\ 1\end{pmatrix}, \; RT_{V2G} = RT_{G2V}^{-1} \tag{18}$$

B. Annotation Process

B.1. Device Time Synchronization

All devices at each intersection are connected to a switch, and their time synchronization is achieved through NTP (Network Time Protocol), which ensure the time error is less than 5ms. When collecting data, the cameras and fisheyes capture at a frequency of 25Hz, while the LiDAR operates at 10Hz.

To construct the dataset, we establish timestamp anchors at a frequency of 10Hz along the timeline. We then select the nearest frame from each device around each anchor and package them into frame batches. Each batch contains synchronized data from all devices at that specific time, and is assigned a corresponding frame index.

B.2. Calibration

Once the sensors are deployed at each intersection and time synchronization is complete, we need to calibrate all sensors. For cameras and fisheyes, calibration involves determining the distortion, intrinsic, and extrinsic parameters. For LiDARs, only the extrinsic parameters need to be calibrated. These extrinsic parameters establish the transformation relationship between the intersection/vehicle coordinate system and the device coordinate system, as explained in Sec. A.

To make it easier for researchers to use our dataset, all images captured by the cameras and fisheyes are undistorted. The coordinate transformation is based on undistorted pixel coordinates and the intersection/vehicle coordinate system. Furthermore, the coordinate transformation between devices can be achieved by linking their extrinsic parameters to the intersection/vehicle coordinate system.

B.3. Global Annotation

We merge all the point clouds within each frame batch. Using Eq.(3) from the supplementary material, we project each individual point cloud onto the intersection coordinate system. The two sets of point clouds are concatenated together. And then, we annotate the 3D boxes for the targets that appear in the concatenated point cloud scene. This annotation process involves determining the positions, orientations, and categories of the 3D boxes.

Afterwards, the annotated 3D boxes are projected onto the images using the corresponding extrinsic parameters for each camera and fisheye. Annotators are then tasked with performing supplementary annotations for each camera, especially in cases where the target's point cloud is occluded or extends beyond the range captured by the LiDAR, resulting in missed annotations. The annotated boxes are subsequently projected onto the intersection/vehicle coordinate system. However, it is important to note that due to calibration errors, there may be a minuscule number of boxes in the dataset that have inaccurate projections onto the intersection position.

Annotators associate a global ID with the annotated 3D boxes in the timeline. Subsequently, all the 3D boxes are reprojected onto all the devices. Annotators then determine the visibility of each box, indicating which devices can see the box. For example, if a box is only visible in Lidar-2, Camera-1, Camera-3, and Fisheye-1, then the visibility information for that box will have "True" ("Visible") for those devices and "False" ("Invisible") for the rest. There are several situations where a box may be marked as "Invisible": If the global 3D box is outside the field of view of a device; If the object is occluded by other objects that exceeds more than 80%; If the object appears too small in the image (far from the device).

Considering that many researchers often use trajectory

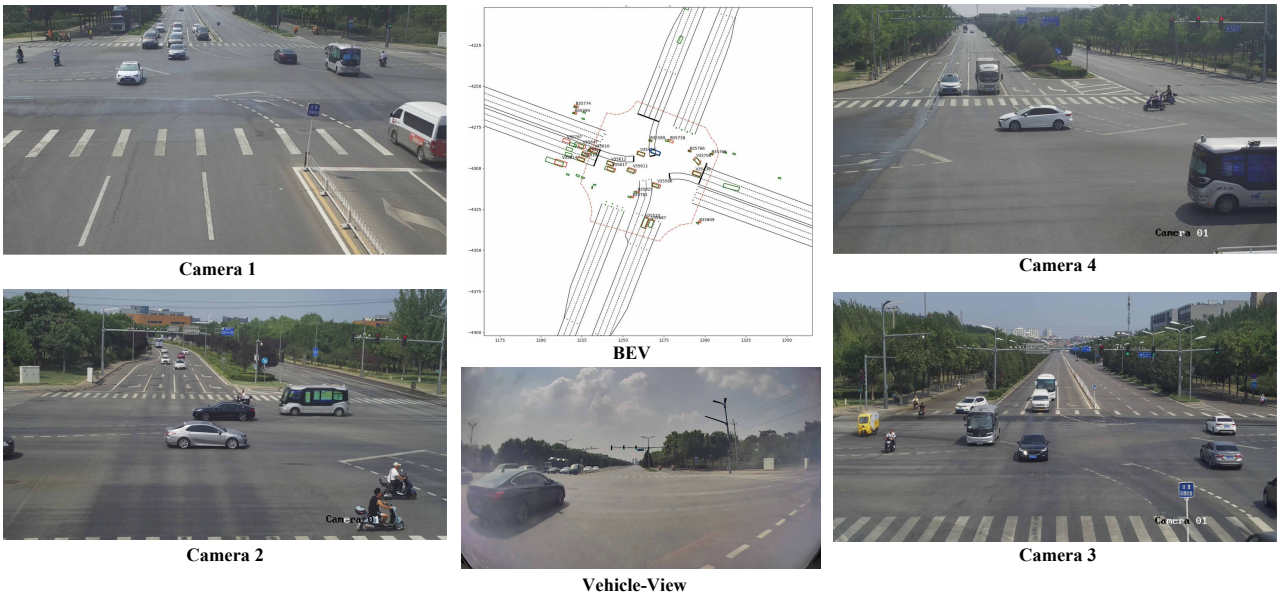

Figure A2. The illustration of Vehicle-Infrastructure Cooperative in VIC-1 at the 4590-th frame. In the BEV view, the blue rectangular box indicates the ego-vehicle position, the red and green boxes indicate the targets from the Vehicle Coordinate and Intersection Coordinate, respectively.

inpainting based on temporal to handle occlusions in tracking tasks, it is worth noting that even though an object may be occluded and not visible in a specific frame, the inpainted boxes can still accurately outline its correct position. Therefore, during evaluation, all of the boxes are labels as "Invisible" are not counted as false positives or false negatives for calculating mAP, MOTA, IDF1, etc., regardless of whether the prediction boxes provided or not.

Both the vehicle-side and the road-side undergo the calibration process described above. Afterwards, all the 3D boxes are transformed onto the global coordinate system. The global ID association is then determined based on the Intersection over Union (IOU) between the 3D boxes of the vehicles and the road.

C. Visualization of HoloVIC

We show all of the visualization results involving all of intersections Int-1/VIC-1 to Int-5/VIC-5. The distribution of all intersections in the HoloVIC Dataset in HD-Map is shown in Fig.A3. The red dashed box identifies the corresponding intersection number for each intersection (Int-1/VIC-1)-(Int-5/VIC-5), which share the same Global Coordinate System. The illustration of Vehicle-Infrastructure Cooperative is illustrated in Fig.A2. The illustration of intersections with different sensor layouts (Type A-D) are shown in Fig.A4-A7.

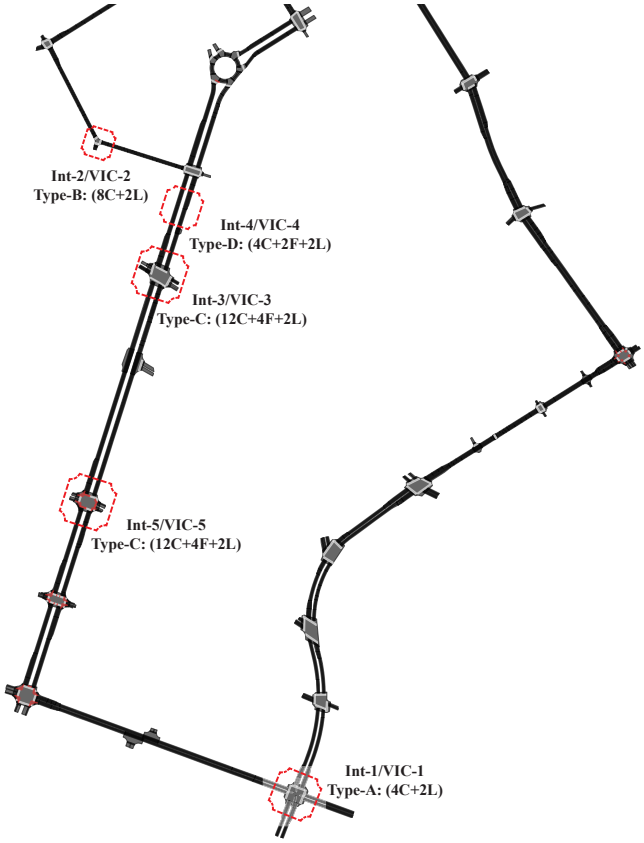

Figure A3. The distribution of all intersections in the HoloVIC Dataset in HD-Map

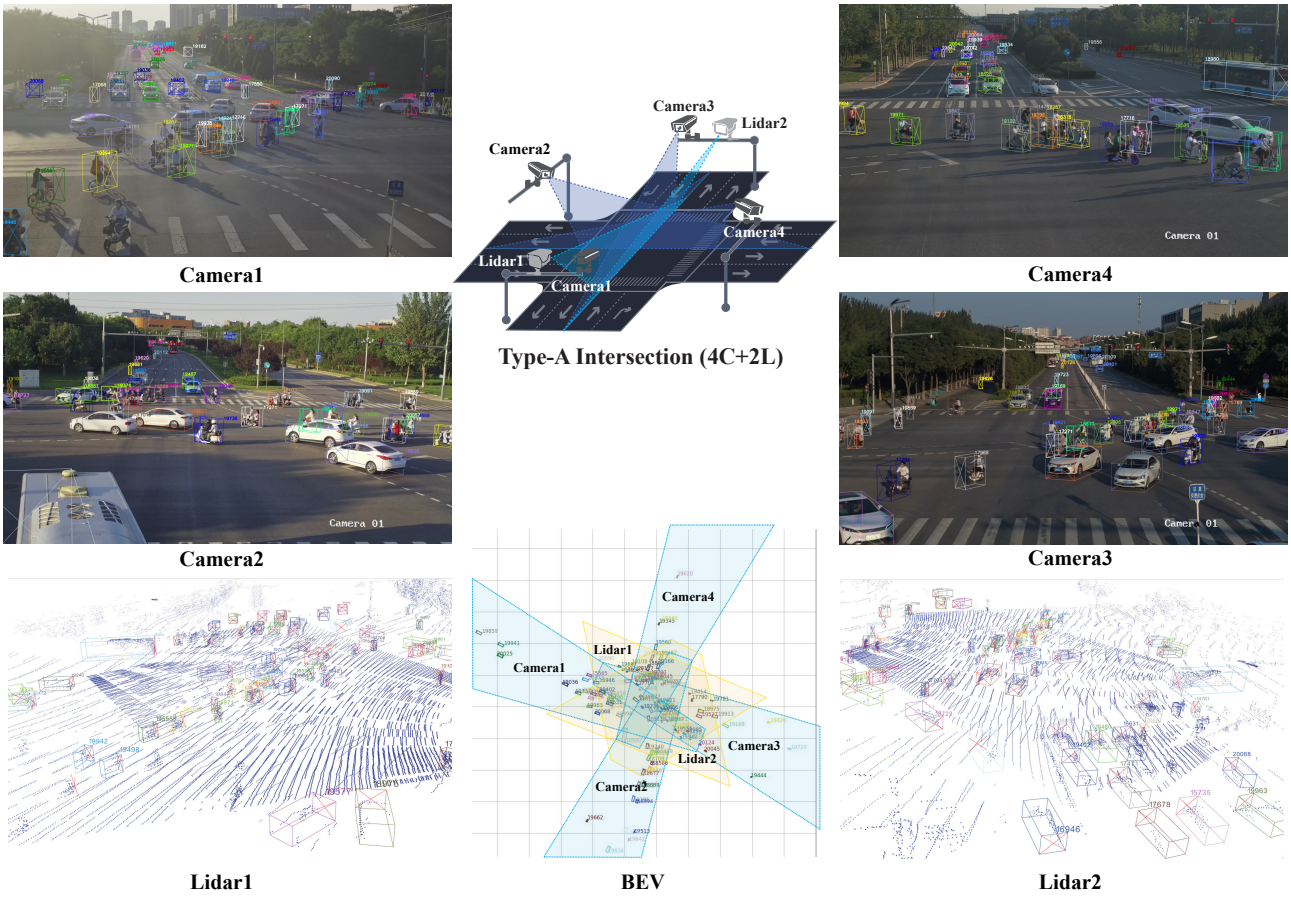

Figure A4. The illustration of Type-A intersection (4C+2L) in Int-1 at the 376-th frame.

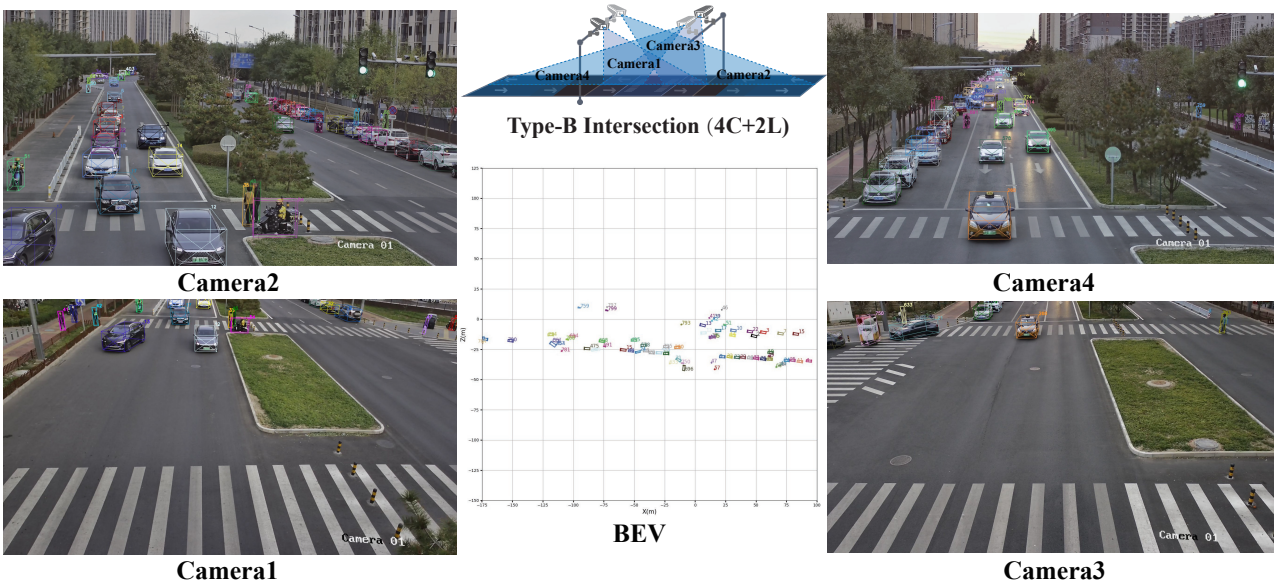

Figure A5. The illustration of Type-B intersection with two opposite viewpoints (4C+2L) in Int-2 at the 754-th frame.

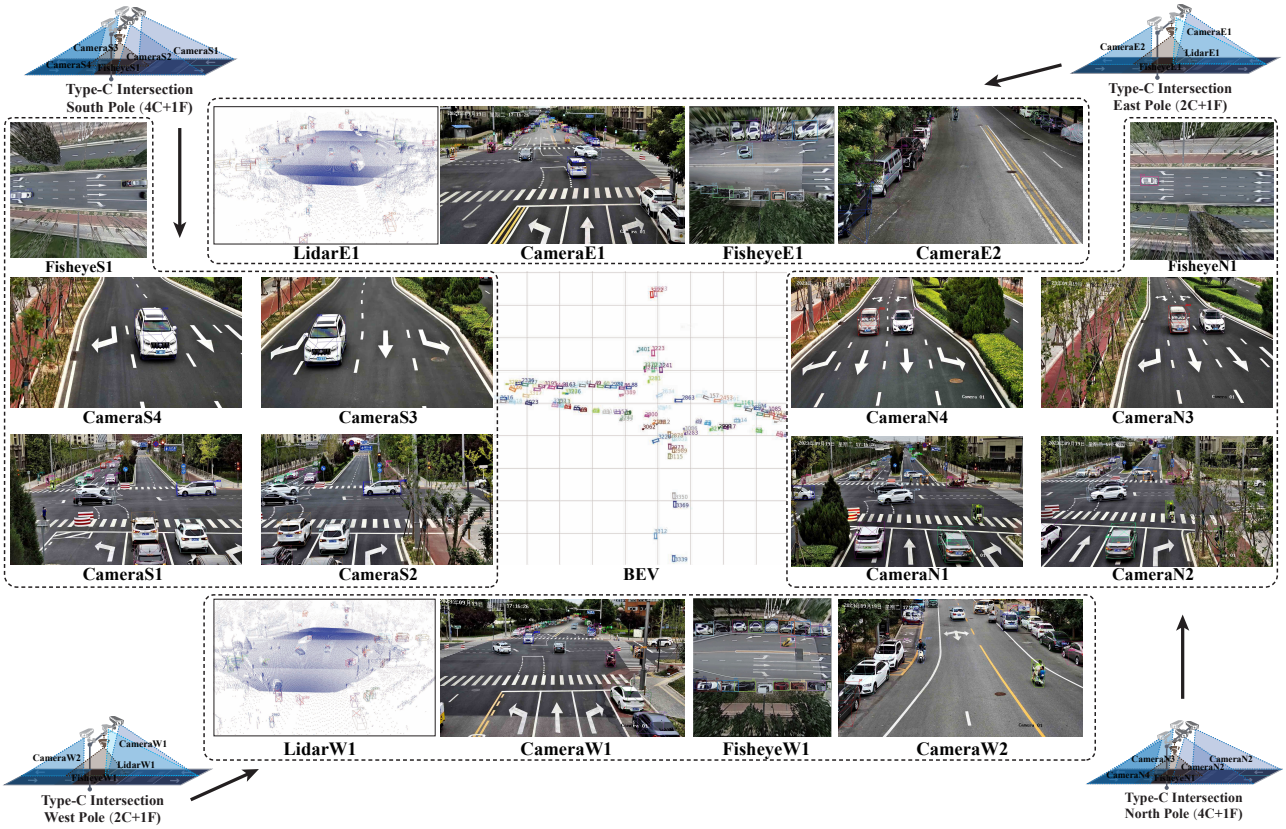

Figure A6. The illustration of Type-C intersection (12C+4F+2L) in Int-3 at the 1105-th frame.

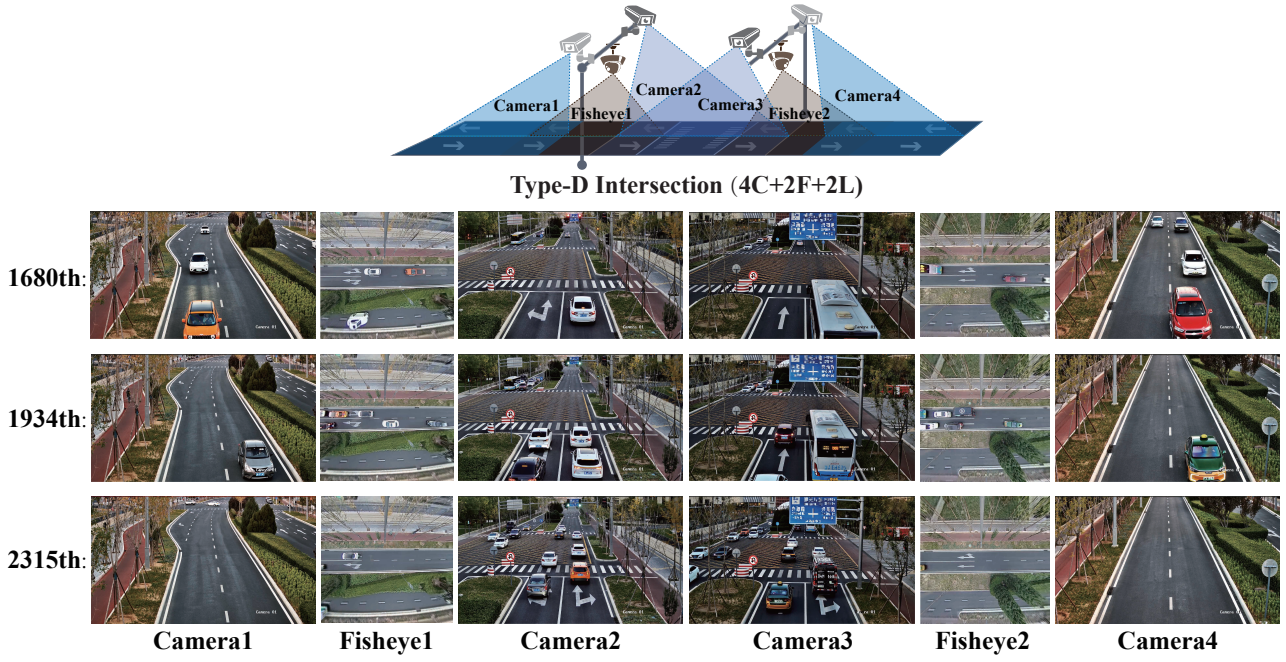

Figure A7. The illustration of Type-D intersection (4C+2F+2L) in Int-4 at the 1680-th, 1934-th and 2315-th frame.