Supplementary Material:
A Generic Vehicle-to-Sensor Calibration Framework

Sumin Hu Youngmin Yoo Jeeseong Kim Changsoo Lim Doohyun Cho* Bongnam Kang
StradVision

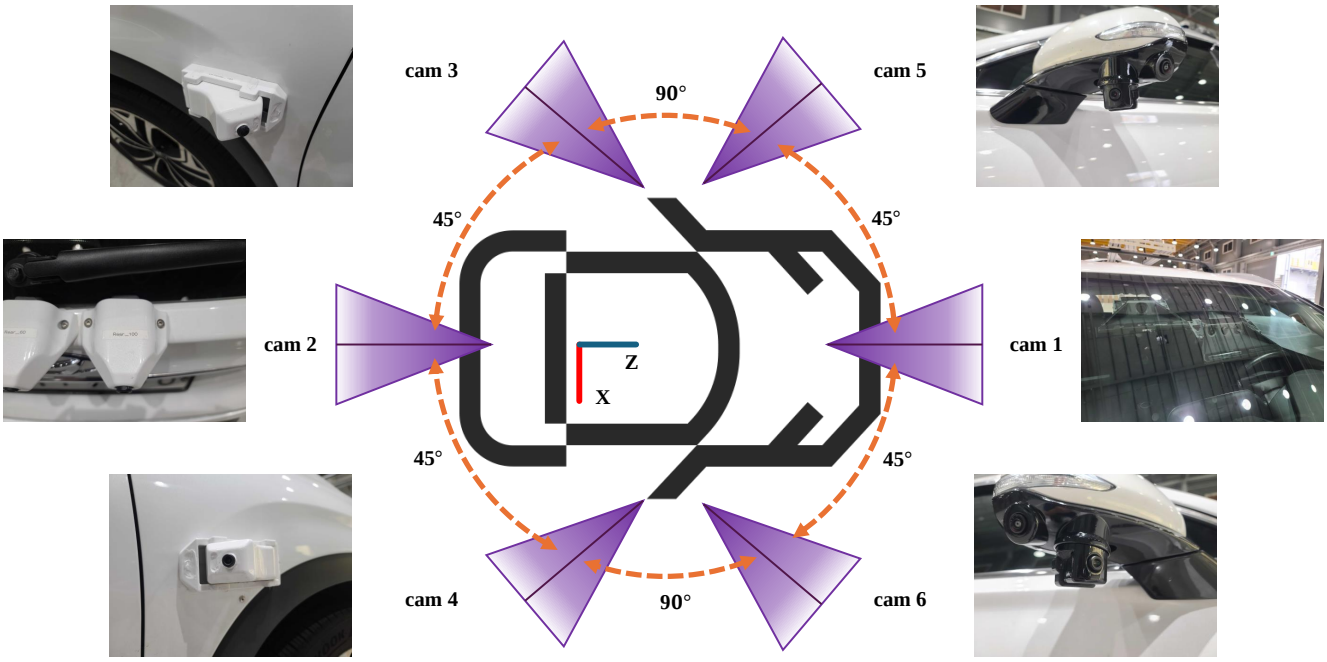

Figure S1. **E3DC dataset's Vehicle Setup.** 6 cameras are installed at angles of approximately 45 or 90 degrees interval, covering 360 degrees. The VGCS is centered at the ground plane just below the vehicle's rear axle's center, easily visible as the intersection of the x- and z-axis.

1. E3DC Dataset

Although we briefly covered some important features of the dataset in the main paper, we provide a complete description here.

Since currently available datasets do not provide accurate vehicle-to-sensor (v2s) calibration results, we have created a custom dataset that has v2s calibration data. We equipped a vehicle with three cameras towards the front at 45-degree intervals and three towards the back at 45-degree intervals, ensuring complete surround-view capture as shown in Fig. S1. The cameras are set to record high-definition video of 1920x1080 resolution at 30 frames per second with a 100-degree field of view. The projection of a 3D point $\mathbf{X}^v = [X, Y, Z, 1]^T$ in Vehicle Ground Coordinate System (VGCS) to a point $\mathbf{x}^{c_i} = [x, y, 1]^T$ in the ith camera image is given as

$$\mathbf{x}^{c_i} = \mathbf{K}_{c_i}\mathbf{D}_{c_i}(\mathbf{T}_{c_i v}\mathbf{X}^v)\,, \tag{S1}$$

where the intrinsic camera matrix $\mathbf{K}_{c_i}$ is

$$\mathbf{K}_{c_i} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}\,, \tag{S2}$$

where f_x and f_y are the focal lengths and c_x and c_y are the principal points. $\mathbf{D}_{c_i}(\cdot)$ is the distortion function with coefficients of $\mathbf{d} \in \mathbb{R}^4$ of the ith camera. We employ the Kannala-Brandt [2] model for the distortion model, but

*Work completed while employed at StradVision.

readers could utilize a distortion model conversion tool [3] to convert the parameters to their own distortion model in use. $\mathbf{T}_{c_i v} \in \mathbb{R}^{4\times4}$ is the extrinsic transformation which is expressed as

$$\mathbf{T}_{c_i v} = \begin{bmatrix} \mathbf{R}_{c_i v} & \mathbf{t}_{c_i v} \\ \mathbf{0}^T & 1 \end{bmatrix}, \tag{S3}$$

where $\mathbf{R}_{c_i v} \in \mathbb{R}^{3\times3}$ and $\mathbf{t}_{c_i v} \in \mathbb{R}^3$ is the extrinsic rotation and translation that transforms points from the VGCS to the Camera Coordinate System (CCS).

Each camera's 3D position is measured with laser line guides w.r.t. the ground right below the vehicle's rear axle center, which represents the origin of the VGCS. Then, each camera's intrinsic parameters, *i.e.*, $\mathbf{D}_{c_i}$ and $\mathbf{K}_{c_i}$, are calibrated with a checkerboard.

A larger chessboard is positioned so it visible from the camera of interest and the distance is measured w.r.t. the VGCS. Then, the checkerboard's image corner points are manually annotated. With the known information, we use Perspective-n-Point [1] to acquire the extrinsic rotation $\mathbf{R}_{c_i v}$. These processes are repeated 5 times to get the best estimate.

The dataset encompasses two distinct environments: an urban neighborhood and a parking lot, with each location yielding about 9000 frames from a 5-minute drive at speeds adjusted for the respective driving location. In total, 6 videos, 1 video from each camera, were recorded in each location. The dataset directory tree is structured as shown below:

```
.
├── Parking_Lot_1
│   └── 20230516_103000
│       ├── cam1.h264
│       ├── cam2.h264
│       ├── cam3.h264
│       ├── cam4.h264
│       ├── cam5.h264
│       └── cam6.h264
├── Parking_Lot_2
│   └── 20230516_104500
│       └── ...
├── Neighborhood_1
│   └── 20230516_094000
│       └── ...
├── Neighborhood_2
│   └── 20230516_095500
│       └── ...
└── Calib
    ├── cam1_front.json
    ├── cam2_back.json
    ├── cam3_left_back.json
    ├── cam4_right_back.json
    ├── cam5_left_front.json
    └── cam6_right_front.json
```

2. Epipoles in the VGCS

Here, we derive how we arrived at Eqs. (18) and (19). From Fig. 1, transformation render the following relationship:

$$\mathbf{T}_{sv}\mathbf{T}^v_{ji} = \mathbf{T}^s_{ji}\mathbf{T}_{sv} \; . \tag{S4}$$

Then,

$$\mathbf{T}^s_{ji} = \mathbf{T}_{sv}\mathbf{T}^v_{ji}\mathbf{T}^T_{sv} \; . \tag{S5}$$

From the equation above, we can derive

$$\mathbf{R}^s_{ji} = \mathbf{R}_{sv}\mathbf{R}^v_{ji}\mathbf{R}^T_{sv} \; . \tag{S6}$$

and

$$\mathbf{t}^s_{ji} = \mathbf{R}_{sv}\left(\mathbf{t}^v_{ji} + \left(\mathbf{I} - \mathbf{R}^v_{ji}\right)\mathbf{R}^T_{sv}\mathbf{t}_{sv}\right) \; . \tag{S7}$$

According to Eq. (16), each epipole can be represented with the camera intrinsic matrix, rotation, and translation from frame i to frame j. By substituting the corresponding terms with Eqs. (S6) and (S7), we get

$$\begin{aligned} \mathbf{e}^s_1 &= \mathbf{K}^{-1}\mathbf{e}^{s,I}_1 && \text{(S8)} \\ &= \mathbf{R}^{s\,T}_{ji}\mathbf{t}^s_{ji} && \text{(S9)} \\ &= \mathbf{R}_{sv}\mathbf{R}^v_{ij}\mathbf{R}^T_{sv}\mathbf{R}_{sv}\left(\mathbf{t}^v_{ji} + \left(\mathbf{I} - \mathbf{R}^v_{ji}\right)\mathbf{R}^T_{sv}\mathbf{t}_{sv}\right) \; . && \text{(S10)} \end{aligned}$$

Since, $\mathbf{t}_{vs} = -\mathbf{R}^T_{sv}\mathbf{t}_{sv}$ and using simple rotation matrix characteristics that $\mathbf{R}^T\mathbf{R} = \mathbf{I}$,

$$\mathbf{e}^s_1 = \mathbf{R}_{sv}\mathbf{R}^v_{ij}\left(\left(\mathbf{R}^v_{ji} - \mathbf{I}\right)\mathbf{t}_{vs} + \mathbf{t}^v_{ji}\right) \; . \tag{S11}$$

By multiplying $\mathbf{R}^T_{sv}$ to both sides, which transforms the CCS defined epipole to VGCS and utilizing the fact that $\mathbf{t}^v_{ij} = -\mathbf{R}^{v\,T}_{ji}\mathbf{t}^v_{ji} = -\mathbf{R}^v_{ij}\mathbf{t}^v_{ji}$, we get

$$\mathbf{e}^v_1 = \left(\mathbf{I} - \mathbf{R}^v_{ij}\right)\mathbf{t}_{vs} - \mathbf{t}^v_{ij} \; , \tag{S12}$$

which finally becomes

$$\mathbf{e}^v_1 = \mathbf{t}^v_{ij} + \left(\mathbf{R}^v_{ij} - \mathbf{I}\right)\mathbf{t}_{vs} \; , \tag{S13}$$

since epipoles are correct up to scale.

Similarly, $\mathbf{e}^v_2$ can be derived as follows:

$$\begin{aligned} \mathbf{e}^s_2 &= \mathbf{K}^{-1}\mathbf{e}^{s,I}_2 && \text{(S14)} \\ &= \mathbf{t}^s_{ji} && \text{(S15)} \\ &= \mathbf{R}_{sv}\left(\mathbf{t}^v_{ji} + \left(\mathbf{I} - \mathbf{R}^v_{ji}\right)\mathbf{R}^T_{sv}\mathbf{t}_{sv}\right) \; . && \text{(S16)} \end{aligned}$$

Again, by multiplying both sides by $\mathbf{R}^T_{sv}$, we transform epipole $\mathbf{e}^s_2$ in the CCS to the VGCS:

$$\begin{aligned} \mathbf{e}^v_2 &= \mathbf{t}^v_{ji} + \left(\mathbf{I} - \mathbf{R}^v_{ji}\right)\mathbf{R}^T_{sv}\mathbf{t}_{sv} && \text{(S17)} \\ &= \mathbf{t}^v_{ji} + \left(\mathbf{R}^v_{ji} - \mathbf{I}\right)\mathbf{t}_{vs} \; . && \text{(S18)} \end{aligned}$$

3. LiDAR calibration Results on KITTI

Since KITTI does not provide GT v2s calibration results, we evaluate LiDAR calibration results by indirectly calculating the camera-to-LiDAR (c2L) extrinsic rotation errors for both `camera 2` and `camera 3`. Note that this result is indirectly estimated from vehicle-to-camera (v2c) results and vehicle-to-LiDAR (v2L) results, thus errors in both v2c and v2L calibration results propagate to the c2L result. Despite such limitation, we notice Euler angle errors are approximately 0.15, 0.3, and 0.6 degrees for pitch, yaw, and roll, respectively. Given that most sequences are too short for convergence in roll, we argue that c2L calibration can be used for initial calibrations and the fact that time synchronization is not required is a large benefit. We also present the LiDAR point projection onto the images from `camera 2` in Fig. S2.

Sequence	Frames	abs(Δ Pitch) (deg)		abs(Δ Yaw) (deg)		abs(Δ Roll) (deg)	
		cam2	cam3	cam2	cam3	cam2	cam3
00	4539	0.031	0.011	0.402	0.432	0.687	0.370
01	1099	0.092	0.072	0.315	0.397	0.145	0.183
02	4659	0.055	0.031	0.318	0.305	0.709	0.478
03	799	0.091	0.048	0.098	0.068	0.862	0.354
04	269	0.866	0.901	0.830	0.853	69.900	72.552
05	2759	0.137	0.031	0.279	0.263	0.478	0.274
06	1099	0.082	0.116	0.291	0.271	0.409	0.088
07	1099	0.014	0.203	0.317	0.317	0.722	0.659
08	4069	0.083	0.125	0.268	0.275	0.141	0.060
09	1589	0.014	0.111	0.271	0.300	1.028	0.454
10	1199	0.308	0.210	0.209	0.250	0.374	0.335
11	919	0.085	0.099	0.239	0.281	0.577	0.856
12	1059	0.153	0.127	0.182	0.192	1.677	1.823
Avg. Abs. Error	-	0.155	0.160	0.309	0.323	0.651*	0.494*

Table S1. **Indirect evaluation of vehicle-to-LiDAR (v2L) via camera-to-LiDAR (c2L) on the KITTI dataset for E3DC.** The asterisk notation (*) denotes values that are calculated by removing an outlier case of sequence `04`, in which there were no driving scenarios where roll could be estimated.

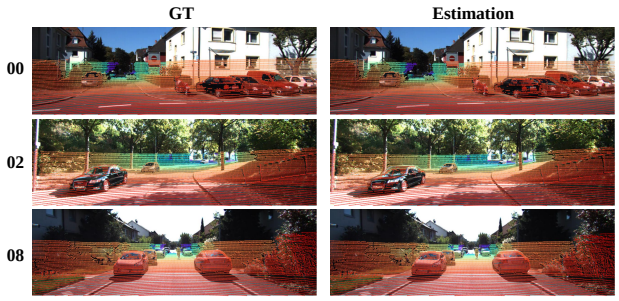

Figure S2. **LiDAR Reprojections on images from `camera 2`.**

References

[1] Xiao-Shan Gao, Xiao-Rong Hou, Jianliang Tang, and Hang-Fei Cheng. Complete solution classification for the perspective-three-point problem. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(8):930–943, 2003. 2

[2] Juho Kannala and Sami S. Brandt. A generic camera model and calibration method for conventional, wide-angle, and fish-eye lenses. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(8):1335–1340, 2006. 1

[3] Sangjun Lee. Fisheye-calib-adapter: An easy tool for fisheye camera model conversion, 2024. 2