Supplementary materials of Cross-dataset Sensor Alignment: Making Visual 3D Object Detector Generalize

Anonymous Author(s)
Affiliation
Address
email

In this Supplementary Material, we provide details omitted in the main paper.

Section 1: brief introduction of datasets.

Section 2: data format conversion (maybe include dataset[])

Section 3: Details of training settings

Section 4: 3D object detection using BEVDet

Section 5: additional object detection results using DETR3D

1 Datasets

Argoverse2. The Argoverse2 dataset [1] is collected across six cities in the US, including Pittsburgh, Detroit, Austin, Palo Alto, Miami, and Washington D.C. It encompasses data captured in various weather conditions and at different times of the day. The dataset includes images from two grayscale stereo cameras and seven cameras that provide 360-degree coverage. It offers 3D annotations at a frame rate of 10Hz. To align with the frame rate in the nuScenes dataset, we sub-sample the Argoverse2 dataset, resulting in 21,982 frames for training and 4,705 frames for validation, with a frame rate of 2Hz. **Yicheng: maybe we need to give a reason why?**

KITTI. The KITTI [2] object detection benchmark consists of 7,481 frames for training. These scenes were captured in clear weather and during daytime around Karlsruhe, Germany. The dataset provides images from two RGB cameras and two grayscale cameras, forming 2 stereo pairs. For our study, we solely utilize the left RGB camera. Following [3], we separate the data into 3,712 training frames and 3,769 validation frames.

KITTI-360. The KITTI-360 dataset [4] is significantly larger than KITTI, comprising 61,569 valid frames with 3D annotations at a frame rate of 10Hz. The labeled data is obtained from nine video clips. To create a training and validation split, we utilize the first 80% of each video clip for training and the remaining 20% for validation. This results in a training set containing 49,253 frames and a validation set containing 12,316 frames. Unlike KITTI, the KITTI-360 dataset provides RGB images from two frontal perspective cameras and two side fish-eye cameras. Similar to the KITTI settings, we exclusively use the images from the left frontal camera in our study.

nuScenes. The nuScenes datasets [5] contains 28130 training and 6019 validation key frames. The scenes are collected around Boston, USA and Singapore in multiple weathers and during different time frame. For each frame, the dataset provides six images that collectively cover a 360-degree view.

Lyft. The Lyft Level 5 dataset [6] consists of 22,680 annotated frames captured around Palo Alto, USA, during clear weather conditions and daytime. Each frame within the dataset includes images from six surrounding view cameras as well as a long-focal-length frontal camera. It is important to mention that this dataset is collected using 20 independent vehicles, and the surrounding view images

Submitted to the 7th Conference on Robot Learning (CoRL 2023). Do not distribute.

have two different resolutions. Following the approach outlined in MMDetection3D[], we partition the dataset into 18,900 frames for training and 3,780 frames for validation.

Waymo. The Waymo [7] dataset is collected across Phoenix, Mountain View, and San Francisco, encompassing various weather conditions and different times of the day. It includes images from five cameras and offers 3D annotations at a frame rate of 10Hz. The training set consists of 158,081 frames, while the validation set contains 39,990 frames. To align with a 2Hz frame rate, we sub-sample the dataset, resulting in 31,616 frames for training and 7,998 frames for validation.

2 Converting Datasets into a Unified Format

In this section, we provide a detailed explanation of how we convert Argoverse2, KITTI, KITTI-360, Lyft, nuScenes, and Waymo datasets into a unified format. We specifically focus on the issues related to coordinate systems and 3D annotations that arise when merging these datasets. We converting data under MMDetection3D v1.1.0.

2.1 Coordinate systems

Regarding sensor configuration, the datasets differ in terms of three types of coordinate systems: ego frame, LiDAR frame, and camera frame. The definition of camera is clear so we primarily focus on the former two. Each dataset typically includes at least one LiDAR mounted on the roof of the vehicle. The origin of the LiDAR frame is commonly located at the center of the top LiDAR if no specification.

Ego frame is more confusing as the origin is defined differently across the datasets. In Argoverse2, nuScenes and Waymo, the ego origin is located at the center of the rear axle of the car. In Argoverse2, it is approximately 33cm above the ground, while in the latter two datasets, it is projected onto the ground plane. Lyft does not explicitly specify the location; however, based on the statistical analysis of 3D annotations, it is also considered to be on the ground. These four datasets have corrected their axes, ensuring that the z-axis consistently points upwards from the road surface. On the other hand, for KITTI and KITTI-360, the ego frame is defined by the Inertial Measurement Unit (IMU). Across all the datasets, the x-axis aligns with the car's longitudinal direction, while the y-axis points to the left.

Regarding LiDAR point clouds and 3D annotations, Argoverse2 and Waymo define them in the ego frame, while KITTI, KITTI-360, Lyft, and nuScenes define them in the LiDAR frame. Consequently, during training, we consider the LiDAR centers of the latter datasets as the 'ego centers'.

In terms of ego frame alignment, for Argoverse2, KITTI, and KITTI-360, we simply lower their ego centers by 0.33m, 1.73m, and 1.73m, respectively, to align them with the road surface. For Lyft and nuScenes, we transform the entire coordinate system to their original ego frames, which are also pressed against the road.

2.2 Object filtering

To ensure consistency and data quality, we discard object annotations that fall outside the camera view. This is accomplished by projecting the eight corners of each object's 3D bounding box onto the image plane. If all eight corners are outside the image boundary, the object annotation is removed. Additionally, we filter annotations based on a specific range in the x, y, and z coordinates, namely $[-51.2, 51.2] \times [-51.2, 51.2] \times [-5.0, 4.0]$. As every dataset includes LiDAR data, we also discard annotations that have no LiDAR points within the 3D bounding box since they may be occluded.

2.3 Merging categories

To unify the category labels across datasets, we merge the categories within each dataset into three classes: vehicle, pedestrian, and bicycle. This taxonomy closely resembles Waymo's classification,

Table 1: Cross-dataset testing results of BEVDet[9] trained on single dataset.

Setting	src/dst	N	A	L	K	K360	W	avg.
Direct	N	25.4	0.0	0.1	0.0	0.0	0.0	4.2
	A	0.0	29.3	0.0	0.0	0.0	3.6	5.5
	L	0.0	0.0	28.4	0.1	0.0	0.0	4.8
	K	0.0	0.0	0.1	11.7	0.4	0.0	2.0
	K360	0.0	0.0	0.1	0.6	18.4	0.0	3.2
	W	0.0	2.2	0.0	0.0	0.0	39.0	6.9

Table 2: Multi-dataset training results by BEVDet

Setting	src/dst	N	A	L	K	K360	W	avg.
Direct	N	25.4	0.0	0.1	0.0	0.0	0.0	4.2
	+A	26.5	31.1	0.1	0.0	0.0	1.5	9.9
	+L	27.1	32.4	31.5	0.2	0.1	2.8	15.7
	+K	28.1	32.9	32.9	22.2	0.3	2.2	19.8
	+K360	26.8	33.5	32.8	27.0	21.4	3.1	24.1
	+W	29.2	36.7	33.9	28.8	22.0	44.5	32.5

but with a little difference in the bicycle category. Waymo excludes bicycles without a rider, whereas we include such objects when relabeling the other datasets. Table **??** shows the mapping of all categories to the three classes. Any types not listed in the table are discarded during the merging process.

3 Training Details

For DETR3D, we use the original training policy, while for BEVDet, we use adamW[8] with weight decay 1×10^{-7} as optimizer, and train it for 24 epochs with batch size 64 and inital learning rate 2×10^{-4}, which will be decreased 10 times on 20th and 24th epoch.

4 BEVDet Results

We present object detection results using the BEVDet model, which leverages depth estimation to project 2D image features back to 3D space. In Table 1, we observe that BEVDet achieves favorable performance when trained and tested within the same dataset. We also see a clear performance drop when training a model on one dataset and evaluating it on others. Interestingly, this performance drop is more severe compared to the DETR3D detector in many cases. We surmise that the dense feature projection operation makes the model more prone to domain gaps across datasets. Consequently, BEVDet becomes highly reliant on data augmentation techniques. Table 2 demonstrates a similar trend when considering the results from a multi-dataset perspective.

5 Additional Results from DETR3D

In this section, we first provide additional results on monocular 3D detecton using DETR3D.

5.1 Ablation studies by cropping the input images

To investigate whether the model relies on other visual cues for object detection, we conduct an experiment where we crop the input images at different positions during testing. In Table 6, we find no performance drop in DETR3D, whereas BEVDet shows a substantial decrease. This indicates that DETR3D does not rely on objects' position in the images. On the other hand, BEVDet, which

Table 3: Ablation study of synchronizing focal length to different values.

Train	focal length	N	A	L	K	K360	W
N	not synced	36.3	0.8	1.8	0.0	0.0	1.1
	1260	35.7	21.9	13.8	27.1	16.9	19.2
	2070	40.8	25.5	18.6	29.7	18.0	23.4
	3100	41.7	26.2	18.7	28.5	17.7	25.8
	4140	43.3	26.4	19.4	31.8	17.5	26.3
A	not synced	0.2	48.0	0.1	0.0	0.0	17.4
	1780	11.8	46.8	7.2	6.4	5.4	38.4
	2070	13.2	51.4	7.5	6.6	4.6	38.8
	3100	12.1	51.1	8.9	7.5	5.1	40.7
	4140	11.5	53.9	9.2	6.1	4.0	40.9
L	not synced	0.5	0.1	37.3	0.4	0.0	0.1
	2070	1.0	1.3	44.0	8.1	5.7	1.5
	3100	1.1	1.3	41.0	8.2	4.8	1.4
	4140	1.0	1.6	42.9	10.5	3.2	1.6
K360	not synced	0.1	0.2	0.0	3.2	26.1	0.1
	550	8.2	4.5	3.3	18.4	25.9	5.5
	2070	14.6	14.7	7.3	34.6	34.7	8.2
	3100	14.0	15.4	9.4	33.6	35.9	7.5
W	not synced	0.1	8.9	0.0	0.0	0.0	58.8
	2070	14.5	37.8	14.3	9.4	5.6	57.7
	3100	14.6	38.1	13.6	7.7	3.2	62.6
	5170	10.1	38.7	10.7	8.8	2.0	62.1

incorporates a depth network in its architecture, is more reliant on this kind of pictorial cues, as suggested in prior work (Dijk et al., 2019) [reference].

5.2 Synchronize focal length to different values

To further investigate the intrinsic synchronization module, we perform an ablation study by increasing the synchronized focal length value during both training and testing. In Table 3, we observe that enlarging the images does improve mAP; however, the extent of improvement diminishes as the input image size increases. We argue that this phenomenon can be attributed to the fact that smaller objects become easier to detect in larger images.

5.3 Ablation studies on sensor alignment approaches

We evaluate the combinations of the modules on all datasets and report the results in Table 4. We observe that each component contributes to the overall performance; however, it is only after aligning the intrinsic parameters that the extrinsic and ego coordinate system start to impact the detection performance. While the Extrinsic Aware Module (EAM) may cause a drop in performance, we argue that this module provides robustness in real-world scenarios where the extrinsic parameters are subject to change.

We evaluate various combinations of modules across all datasets and present the results in Table 4. We observe that each component contributes to the overall performance; however, it is only after aligning the intrinsic parameters that the extrinsic and ego coordinate system start to impact the detection performance. While the Extrinsic Aware Module (EAM) may cause a drop in performance, we argue that this module provides extrinsic robustness in real-world scenarios.

5.4 Multi-dataset training beginning with Waymo

Table 5, where we begin with Waymo as the first dataset. It is worth noting that the order of dataset addition does not affect the observations made in the main paper. Here, KITTI-360 drags down

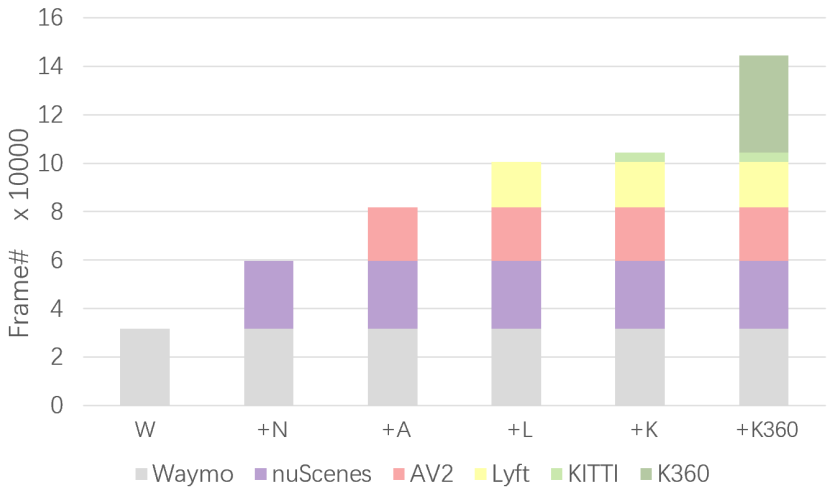

Figure 1: Data volume of each dataset under monocular detection setting

Table 4: Ablation study on the effectiveness of each module in sensor alignment approaches. All models are trained on the Waymo dataset.

Focal	EAM	Ego	N	A	L	K	K360	W	avg
			0.1	8.9	0.0	0.0	0.0	58.8	11.3
		✓	0.0	8.4	0.0	0.0	0.0	58.3	11.1
	✓		0.0	5.0	0.0	0.0	0.0	59.4	10.7
	✓	✓	0.0	3.5	0.0	0.0	0.0	59.4	10.5
✓			14.5	37.8	14.3	9.4	5.6	57.7	23.2
✓		✓	24.7	39.4	33.0	21.2	13.6	57.7	31.6
✓	✓		14.1	37.7	17.0	9.3	5.8	57.6	23.6
✓	✓	✓	25.4	38.2	33.6	21.2	11.7	57.6	31.3

the performance again. This decline can be attributed to a significant amount of discordant data, as illustrated in Figure 1, which provides statistics on the data volume across different datasets.

5.5 Per-class and per-location evaluation results

Table 8 and Table 7 are the extended version of Table 2 and Table 3 in the main paper, showing per-class mAP on each datasets. Furthermore, Table 9 shows the evaluation result of best model in Table 3 on each city in each dataset.

5.6 Results from surrounding view detection

We conduct extensive experiment under surrounding view settings to better understand the behavior of the detectors. All input images are of 1/4 resolution, and the perception range is bigger, including the area behind our ego car.

Ablation studies on sensor alignment approaches. We examine our sensor alignment under the surrounding view settings. In Table 10, DETR3D is trained on Argoverse2, nuScenes and Waymo, and tested in six datasets.

Data diversity vs. data volume We test if the model benefits more from data diversity more than data volume. Given that Waymo and nuScene are of similar data volume, we used different percentage of training data from the 2 datasets and test the model on 3 datasets. As shown in Table 11, mixing the data achieves better performance.

Table 5: Multi-dataset training results by DETR3D, beginning with Waymo.

Setting	src/dst	W	N	A	L	K	K360	avg.
Direct	W	58.8	0.1	8.9	0.0	0.0	0.0	11.3
	+N	60.4	38.4	11.2	0.2	0.0	0.0	18.4
	+A	63.0	42.7	54.8	0.1	0.0	0.0	26.8
	+L	60.2	44.4	53.1	47.3	0.0	0.0	34.2
	+K	63.1	45.5	53.4	49.2	44.3	1.9	42.9
	+K360	61.9	46.2	53.7	49.4	39.5	29.7	46.7

Table 6: Results of DETR3D and BEVDet, which are trained on Waymo, using images cropped at different positions during testing.

Cropped height	[192,992]	[288,1088]	[384,1184]	[480,1280]	Origin
DETR3D	56.9	59.3	59.3	59.7	57.7
BEVDet	20.8	30.5	36.4	31.5	39.0

Table 7: Multi-dataset training results by DETR3D, beginning with nuScenes (full version). The performance is reported in terms of LET-3D-AP for all, vehicles, pedestrians, and bicycles, denoted as a(b/c/d).

Setting	src/dst	N	A	L	K	K360	W
Direct	N	36.3 (56.7/36.7/15.7)	0.8 (0.9/1.4/0.3)	1.8 (1.6/1.1/2.7)	0.0 (0.0/0.0/0.0)	0.0 (0.0/0.0/0.0)	1.1 (0.6/1.2/1.4)
	+A	40.5 (60.4/38.9/22.1)	49.2 (76.0/42.3/29.4)	0.5 (0.7/0.5/0.3)	0.0 (0.0/0.0/0.0)	0.0 (0.0/0.0/0.0)	5.2 (4.5/4.2/6.9)
	+L	41.6 (61.0/40.7/23.1)	50.5 (78.4/42.3/30.9)	43.7 (74.5/26.3/30.4)	0.0 (0.0/0.0/0.0)	0.0 (0.0/0.0/0.0)	3.8 (4.2/4.2/2.9)
	+K	41.5 (62.7/39.6/22.1)	49.7 (78.5/42.4/28.3)	46.0 (75.8/28.1/34.2)	41.4 (62.1/38.7/23.5)	1.1 (1.2/1.2/1.0)	3.6 (4.3/4.2/2.2)
	+K360	42.6 (64.2/40.8/22.8)	54.3 (78.6/43.8/40.6)	46.8 (76.7/26.5/37.1)	36.3 (51.3/35.0/22.7)	29.7 (60.6/8.6/20.0)	3.3 (4.2/3.8/1.8)
	+W	46.2 (66.6/42.7/29.2)	53.7 (79.8/47.9/33.5)	49.4 (76.7/33.0/38.4)	39.5 (54.2/36.1/28.0)	29.7 (60.6/9.4/19.2)	61.9 (82.0/55.4/48.2)
Sync Intrinsic	N	40.8 (58.5/42.1/21.8)	25.5 (45.6/25.1/5.7)	18.6 (39.8/14.4/1.6)	29.7 (41.1/24.2/23.7)	18.0 (37.6/11.3/4.9)	23.4 (37.4/24.8/7.9)
	+A	45.5 (64.5/45.0/27.0)	50.0 (77.9/44.0/28.1)	25.1 (49.2/18.0/8.1)	35.8 (48.1/26.2/33.1)	21.3 (42.5/15.0/6.5)	44.2 (67.9/35.8/28.7)
	+L	46.8(64.3/47.1/28.9)	53.2(79.5/46.6/33.6)	55.1(82.6/36.4/46.2)	37.8(50.7/30.6/31.9)	23.1(44.6/16.9/7.8)	45.3(69.2/37.0/29.6)
	+K	47.4 (64.5/48.0/29.8)	53.5 (79.3/45.6/35.6)	53.6 (82.5/34.3/43.9)	57.8 (77.7/48.7/46.9)	21.8 (36.4/17.2/11.6)	44.4 (69.4/37.2/26.7)
	+K360	50.2 (68.0/47.5/34.9)	54.4 (80.7/48.3/34.3)	54.0 (83.7/36.4/42.0)	60.2 (81.9/47.2/51.4)	39.6 (72.7/16.9/29.3)	44.7 (71.6/38.2/24.2)
	+W	51.8 (68.2/49.8/37.3)	55.3 (82.1/48.8/34.9)	56.6 (84.4/38.4/47.1)	61.9 (83.0/51.5/51.1)	40.7 (73.6/19.9/28.7)	63.7 (83.2/55.0/52.8)
Sync Extrinsic and Ego center	N	43.1(63.0/45.0/21.4)	33.6(62.4/30.4/7.9)	32.8(60.8/19.3/18.4)	33.0(49.0/28.2/21.7)	18.4(37.0/12.8/5.5)	33.0(47.8/28.4/22.8)
	+A	52.1(68.1/50.8/37.4)	52.7(77.9/47.3/32.9)	38.4(70.4/23.5/21.2)	42.2(54.9/33.6/37.9)	23.2(43.0/16.5/10.3)	40.7(64.1/35.6/22.5)
	+L	52.6(68.9/50.2/38.6)	53.2(79.1/47.4/33.2)	59.5(85.6/45.5/47.4)	46.1(61.6/37.8/38.9)	26.1(47.6/19.7/10.9)	43.6(67.1/35.6/28.0)
	+K	51.0(67.9/51.8/33.3)	54.7(79.8/47.7/36.5)	60.2(85.6/44.5/50.5)	63.9(83.2/58.0/50.6)	28.4(48.8/22.9/13.5)	44.6(67.1/35.4/31.4)
	+K360	50.0(70.5/50.4/29.3)	55.0(81.4/47.4/36.2)	59.8(86.9/43.9/48.4)	65.0(85.4/54.5/55.1)	42.7(75.5/20.3/32.3)	45.2(68.6/36.2/30.9)
	+W	54.8(72.7/52.5/39.1)	56.4(82.3/49.0/38.0)	60.5(87.4/45.7/48.4)	66.8(85.2/58.1/57.2)	43.4(76.3/22.3/31.5)	62.7(83.4/56.9/47.9)

Table 8: Cross-dataset testing results of DETR3D trained on single dataset (full version). The performance is reported in terms of LET-3D-AP for all, vehicles, pedestrians, and bicycles, denoted as a(b/c/d).

Setting	src/dst	N	A	L	K	K360	W
Direct	N	36.3 (56.7/36.7/15.7)	0.8 (0.9/1.4/0.3)	1.8 (1.6/1.1/2.7)	0.0 (0.0/0.0/0.0)	0.0 (0.0/0.0/0.0)	1.1 (0.6/1.2/1.4)
	A	0.2 (0.1/0.5/0.1)	48.0 (73.8/38.7/31.7)	0.1 (0.0/0.1/0.1)	0.0 (0.0/0.1/0.0)	0.0 (0.0/0.1/0.0)	17.4 (19.7/14.1/18.3)
	L	0.5 (0.9/0.7/0.0)	0.1 (0.1/0.1/0.0)	37.3 (70.5/16.6/24.9)	0.4 (0.5/0.7/0.1)	0.0 (0.0/0.1/0.0)	0.1 (0.0/0.3/0.0)
	K	2.8 (4.8/3.0/0.5)	1.2 (0.9/1.3/1.5)	0.0 (0.1/0.1/0.0)	24.5 (40.2/25.1/8.3)	1.1 (0.9/2.1/0.4)	0.7 (0.2/0.4/1.3)
	K360	0.1(0.1/0.2/0.0)	0.2(0.0/0.2/0.4)	0.0(0.0/0.0/0.0)	3.2(0.9/6.7/2.2)	26.1(60.2/4.5/13.7)	0.1(0.0/0.1/0.2)
	W	0.1 (0.0/0.1/0.0)	8.9 (14.5/9.2/3.1)	0.0 (0.0/0.0/0.0)	0.0 (0.0/0.0/0.0)	0.0 (0.0/0.0/0.0)	58.8 (78.1/50.3/47.9)
Sync Intrinsic	N	35.7(56.1/36.3/14.8)	21.9(40.9/19.3/5.3)	13.8(30.4/9.6/1.4)	27.1(41.7/24.6/14.9)	16.9(36.4/10.0/4.2)	19.2(32.0/18.9/6.7)
	A	11.8(19.1/12.9/3.4)	46.8(73.4/38.2/28.8)	7.2(18.2/2.5/1.0)	6.4(10.1/2.1/6.9)	5.4(13.7/1.1/1.5)	38.4(59.2/29.5/26.4)
	L	1.0 (1.8/1.2/0.1)	1.3 (3.3/0.4/0.1)	44.0 (76.4/20.9/34.6)	8.1 (17.3/4.7/2.3)	5.7 (12.7/2.0/2.4)	1.5 (3.2/1.0/0.2)
	K	1.2 (0.9/1.9/0.7)	0.2 (0.4/0.0/0.1)	0.6 (0.8/0.8/0.1)	24.7 (40.0/26.1/8.0)	6.4 (13.8/3.5/1.9)	0.1 (0.3/0.0/0.1)
	K360	14.6 (34.0/9.3/0.4)	14.7 (36.4/2.7/4.9)	7.3 (20.0/1.6/0.2)	34.6 (59.9/25.1/18.9)	34.7 (69.4/10.0/24.7)	8.2 (22.7/0.7/1.1)
	W	14.5(25.9/13.3/4.4)	37.8(67.3/34.1/11.9)	14.3(25.5/8.5/8.8)	9.4(6.3/4.8/17.0)	5.6(13.1/1.1/2.6)	57.7(78.0/50.2/45.0)
Sync Extrinsic and Ego center	N	43.1(63.0/45.0/21.4)	33.6(62.4/30.4/7.9)	32.8(60.8/19.3/18.4)	33.0(49.0/28.2/21.7)	18.4(37.0/12.8/5.5)	33.0(47.8/28.4/22.8)
	A	24.4(42.0/23.0/8.3)	48.1(75.3/41.7/27.3)	34.1(61.5/22.4/18.3)	18.1(22.0/17.3/15.0)	8.7(16.1/3.1/6.8)	37.4(58.4/28.8/24.9)
	L	15.7(35.1/10.3/1.9)	19.6(42.6/13.1/3.0)	47.1(79.3/21.2/40.8)	20.0(38.7/11.5/9.9)	12.9(31.8/4.0/2.9)	18.9(30.2/11.2/15.2)
	K	7.1(12.6/7.9/0.8)	8.7(19.2/3.8/2.9)	10.2(20.7/4.2/5.6)	29.1(49.5/28.1/9.6)	9.3(20.2/4.9/2.7)	2.4(3.6/2.5/1.1)
	K360	13.9(33.2/8.3/0.3)	17.7(41.7/5.6/5.8)	16.6(38.4/4.2/7.3)	39.1(67.6/26.5/23.2)	36.7(72.4/10.9/26.6)	8.4(18.2/2.8/4.3)
	W	25.4(45.0/26.3/5.0)	38.2(68.3/34.6/11.7)	33.6(63.3/16.8/20.7)	21.2(13.5/25.3/24.7)	11.7(25.6/6.2/3.3)	57.6(77.9/50.9/44.2)

Table 9: Evaluation results per location, using DETR3D with all sensor alignment approaches. The performance is reported in terms of LET-3D-AP for all, vehicles, pedestrians, and bicycles, denoted as a(b/c/d).

Dataset	Location	LET-3D-AP
argoverse2	ATX	44.6 (69.7/64.3/0.0)
	DTW	71.5 (79.3/60.3/75.1)
	MIA	46.8 (84.0/41.1/15.2)
	PAO	64.3 (91.6/51.4/49.7)
	PIT	59.9 (82.0/53.2/44.4)
	WDC	39.9 (80.1/39.7/0.0)
kitti	Germany	66.8 (85.2/58.1/57.2)
kitti-360	Germany	43.3 (76.3/22.3/31.4)
lyft	Palo Alto	60.5 (87.4/45.7/48.5)
nuscenes	boston-seaport	61.4 (75.4/51.7/57.1)
	singapore-hollandvillage	30.3 (67.4/23.4/0.1)
	singapore-onenorth	50.5 (66.8/53.0/31.8)
	singapore-queenstown	51.6 (68.5/58.9/27.5)
waymo	other	44.9 (72.4/44.8/17.6)
	phx	63.3 (87.0/55.7/47.1)
	sf	65.1 (83.7/58.8/52.8)

Table 10: Surrounding view 3D detection results: ablation study on the effectiveness of each module in sensor alignment approaches. All models are trained on AV2, nuScenes, Waymo.

Focal	EAM	Ego	A	N	W	L	K	K360	avg.
			48.0	40.4	54.8	0.6	6.2	0.7	25.1
		✓	48.6	41.0	53.8	0.0	3.6	0.0	24.5
	✓		49.5	39.7	54.7	1.8	7.4	1.8	25.8
	✓	✓	47.4	40.8	53.3	0.2	4.4	0.6	24.4
✓			52.2	46.5	55.2	22.2	22.0	11.0	34.9
✓		✓	50.1	47.2	54.2	30.1	36.3	22.1	40.0
✓	✓		52.0	46.7	55.5	31.1	26.2	15.8	37.9
✓	✓	✓	52.1	47.5	54.8	31.4	39.7	24.0	41.6

Table 11: Surrounding view 3D detection results: models are trained on different combinations of Waymo and nuScenes.

test/ train(W+N)	0.00+1.00	0.01+0.99	0.10+0.90	0.33+0.67	0.5+0.5	0.67+0.33	1.00+0.00
A	1.0	6.4	9.1	13.0	14.5	16.0	4.3
N	32.5	32.5	32.3	32.1	31.5	30.4	0.2
W	0.3	23.4	36.7	45.8	45.6	47.0	46.2

Table 12: The original categories in each dataset that we include into the vehicle, pedestrain and bicycle categories.

Dataset	Vehicle	Pedestrian	Bicycle
argoverse2	REGULAR VEHICLE, LARGE VEHICLE, BUS, BOX TRUCK, TRUCK, MOTORCYCLE, VEHICULAR TRAILER, TRUCK CAB, SCHOOL BUS	PEDESTRIAN, WHEELED RIDER, OFFICIAL SIGNALER	BYCYCLE, BYCYCLIST
kitti	Car, Van, Trunk, Tram	Pedestrian, Person Sitting	Cyclist
kitti-360	bus, car, caravan, motorcycle, trailer, train, truck, unknown vehicle	person	bicycle, rider
lyft	car, truck, bus, emergency vehicle, other vehicle, motorcycle	pedestrian	bicycle
nuscenes	car, truck, construction vehicle, bus, trailer, motorcycle	pedestrian	bicycle
waymo	Car	Pedestrian	Cyclist

References

[1] B. Wilson, W. Qi, T. Agarwal, J. Lambert, J. Singh, S. Khandelwal, B. Pan, R. Kumar, A. Hartnett, J. K. Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.00493*, 2023.

[2] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[3] X. Chen, K. Kundu, Y. Zhu, A. G. Berneshawi, H. Ma, S. Fidler, and R. Urtasun. 3d object proposals for accurate object class detection. *Advances in neural information processing systems*, 28, 2015.

[4] Y. Liao, J. Xie, and A. Geiger. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[5] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.

[6] R. Kesten, M. Usman, J. Houston, T. Pandya, K. Nadhamuni, A. Ferreira, M. Yuan, B. Low, A. Jain, P. Ondruska, et al. Lyft level 5 av dataset 2019. *urlhttps://level5. lyft. com/dataset*, 1:3, 2019.

[7] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020.

[8] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[9] J. Huang, G. Huang, Z. Zhu, Y. Ye, and D. Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021.