

RICCARDO: Radar Hit Prediction and Convolution for Camera-Radar 3D Object Detection

– Supplementary Material –

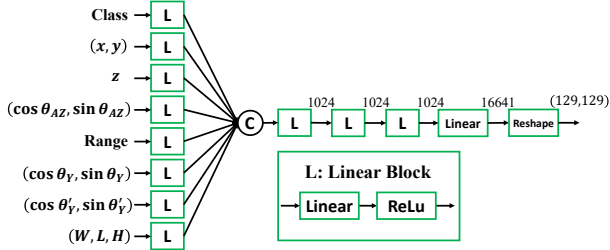


Figure 1. Stage-1 Network Structure. The class input is in one-hot encoding; z represents heights of bounding box bottom faces; θ_{AZ} stands for azimuths of objects in ego coordinates; θ_Y and θ'_Y are object yaws in ego coordinates and relative yaws (*i.e.*, $\theta_Y - \theta_{AZ}$), respectively. “C” represents concatenation, and “Linear” denotes a linear transformation layer. Feature sizes are marked besides network layers.

Table 1. Inference time for SparseBEV with different backbones, radar processing as well as RICCARDO Stages 1 to 3.

Components	SparseBEV (V2.99)	SparseBEV (ResNet101)	Radar Processing	Stage 1	Stage 2	Stage 3
Time (ms)	575.7	211.5	105.4	1.0	12.7	1.2

1. Additional Implementation Details

1.1. Detailed Network Structure

Figs. 1 and 2 show network structures of Stages 1 and 3, respectively.

1.2. Inference Time

Using a NVIDIA V100s GPU and Intel Xeon Platinum 8260 CPUs, we record in Tab. 1 inference time for different components of RICCARDO. Radar processing refers to accumulation and BEV binning of 7 radar sweeps. We can see the monocular component takes most of the inference time and in comparison Stages 1 to 3 are very fast. Radar processing has not been optimized and could be sped up through code optimization.

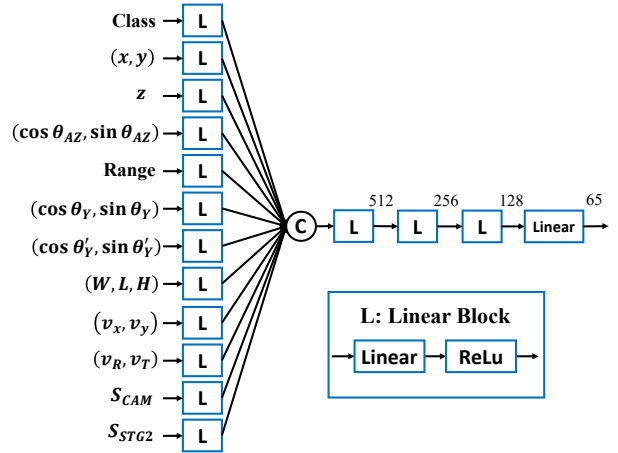


Figure 2. Stage-3 Network Structure. The inputs v_x and v_y are monocular estimated object velocities in ego coordinates; v_R and v_T are monocular velocities in radial and tangential directions, respectively; S_{CAM} and S_{STG2} represent monocular detection scores and Stage-2 matching scores, respectively.

2. Additional Ablation Studies

In the following ablations, we use SparseBEV [3] with backbone ResNet101 [2] for the monocular components in RICCARDO. For efficiency, the data used for evaluation are a subset of nuScenes validation set with 600 random samples.

2.1. Ablation on Velocity Used for Point Motion Compensation

When accumulating 7 radar sweeps in inference, we used estimated radar point velocity to compensate motions of moving points. We implement different velocity estimations and compare resultant detection performance in Tab. 2. The velocity estimations include 0, *i.e.*, no motion compensation, Doppler velocity, Doppler velocity back-projected to estimated object heading direction, Doppler velocity plus tangential component of monocular estimated velocity, and monocular velocity. The geometric relation between full velocity and its tangential and radial compo-

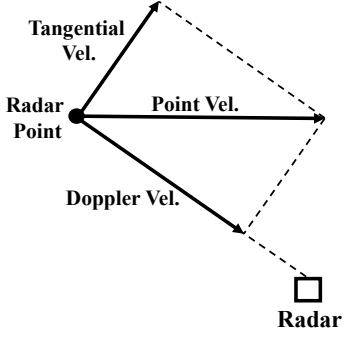


Figure 3. Geometric relation between point velocity and its radial and tangential components.

Table 2. Ablation on different velocity estimations for compensating motion during radar sweep accumulation. Keys: Vel.= Velocity; Mono.= Monocular

Point Vel. Estimation	NDS (\uparrow)	mAP (\uparrow)
0	0.614	0.534
Doppler Vel.	0.620	0.544
Back-Projected Doppler Vel.	0.621	0.545
Doppler + Tangential Mono. Vel.	0.621	0.545
Mono. Vel.	0.621	0.546

nents are shown in Fig. 3. From Tab. 2 we can see performing motion compensation improves detection performance and using full velocity estimates achieves better accuracy compared with only applying Doppler velocity. The three full velocity estimates shown on the 4th to 6th rows result in almost the same detection performance.

2.2. Ablation on Range and Score Updating

In Stage-3 inference we update both range and detection score. Detection scores indicate confidence in prediction and have an impact on mAP computation, where predictions with higher scores have priority as true positives to be associated with GT. We update detection scores by adding Stage-3 scores weighted by α to monocular scores. We test different range and score updating options with different α and list resultant detection performance in Tab. 3. We can see both range and score updating improve detection performance while range updating has significantly bigger impacts on performance.

2.3. Ablation on Number of Radar Sweeps

Within 0.5s time window, there are about 7 sweeps of radar points (*i.e.*, 1 current plus 6 past ones) from radars running at 13Hz in nuScenes Dataset [1]. We accumulate multiple radar sweeps during inference and the number of radar sweeps may impact detection performance, as more sweeps

Table 3. Ablation on updating range and detection score with fusion weight α

Update Range	Update Score	α	NDS (\uparrow)	mAP (\uparrow)
		-	0.590	0.501
	\checkmark	0.5	0.593	0.503
\checkmark		-	0.617	0.543
\checkmark	\checkmark	0.2	0.620	0.545
\checkmark	\checkmark	0.5	0.621	0.545
\checkmark	\checkmark	0.8	0.621	0.543
\checkmark	\checkmark	1.0	0.620	0.541

Table 4. Ablation on Number of Radar Sweeps. More radar sweeps result in better detection performance. Key: Num.= Number

Num. of Sweeps	NDS (\uparrow)	mAP (\uparrow)
0	0.590	0.501
1	0.597	0.512
3	0.612	0.531
5	0.618	0.541
7	0.621	0.545

Table 5. Detection performance NDS(\uparrow) / mAP(\uparrow) of RICCARDO and its underlying monocular detector in night, daytime and all validation scenes, respectively.

Scene	Night	Daytime	Overall
SparseBEV [3]	0.526 / 0.400	0.673 / 0.601	0.669 / 0.595
SparseBEV + RICCARDO	0.561 / 0.450	0.704 / 0.642	0.699 / 0.636
Number of Samples	602	5417	6019

provide denser radar measurement used for Stage 2. To verify this, we run RICCARDO multiple times with radar input from 0, 1, 3, 5, and 7 sweeps, respectively and record their detection performance. Note using 0 radar sweep refers to applying only monocular detector without fusion. As shown in Tab. 4, more radar sweeps lead to better detection performance as expected.

3. Performance at Night

Although using radar to handle adverse conditions is a different research focus, we show in Tab. 5 that RICCARDO significantly improves detection performance over the underlying monocular detector under challenging lighting conditions at night. We evaluate on 7 object categories, which appear in night scenes in nuScenes validation set. We also list corresponding daytime and overall performance for reference.

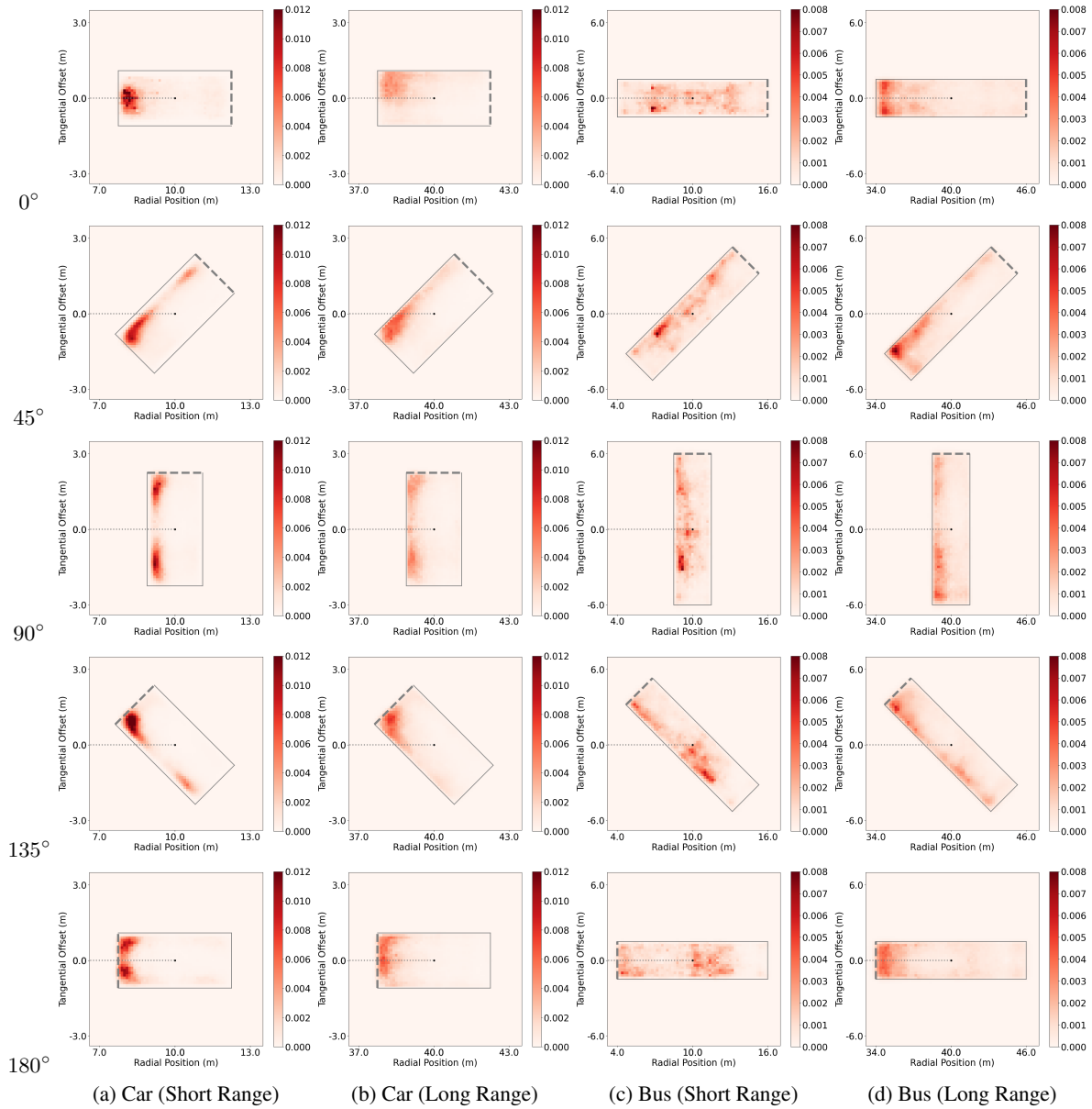


Figure 4. Visualization of predicted radar distributions of (a)(b) Car and (c)(d) Bus viewed from different angles and distances of 10 and 40 meters. X-axis represents radial positions, and Y-axis denotes tangential offsets to object centers. Radial rays are plotted as horizontal dotted lines. Target bounding boxes are shown on top of distributions and dashed lines represent object head.

4. Additional Visualizations

4.1. Visualization of Radar Distributions

To visualize how predicted distribution varies with viewing angles, we simulate object parameters with different orientations and apply Stage-1 model to generate corresponding radar hit distributions. Figs. 4 and 5 shows predicted distributions for car, bus, bicycle, and barrier with different orientations and distances. We can see the distributions vary

with category, orientation and distance. For example, radar distributions are less concentrated spatially at longer range because of larger beam width. We can also notice that distributions of radar points reflected by the tail and head of cars (as shown in the 1st and last row of Fig. 4) are different because of their different surface shapes. More visualizations of predicted radar distributions for objects rotating 360 degrees are shown in the attached video demo.

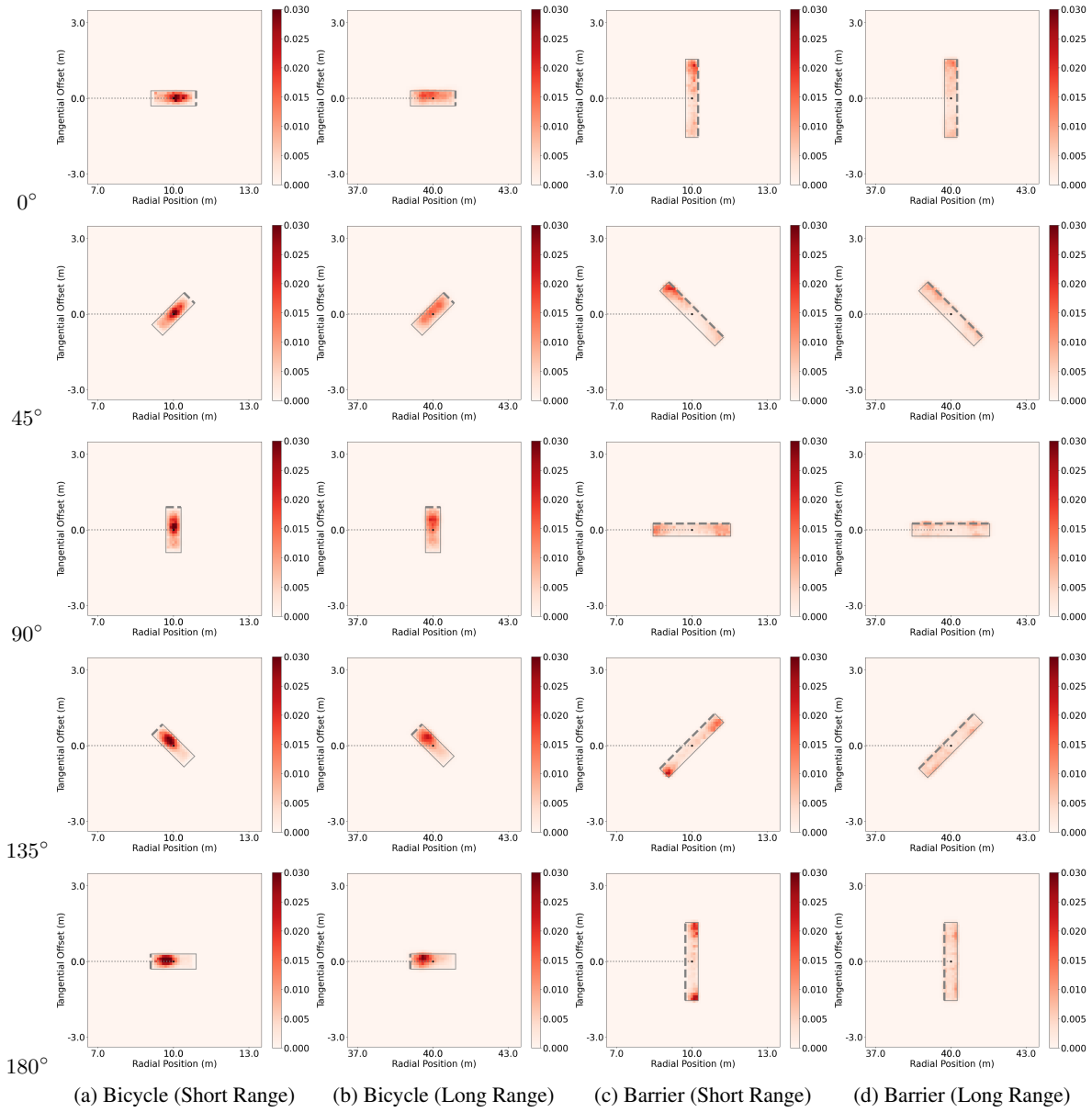


Figure 5. Visualization of predicted radar distributions of (a)(b) bicycle and (c)(d) barrier viewed from five different angles and from distances of 10 and 40 meters.

References

- [1] Holger Caesar, Varun Bankiti, Alex Lang, Sourabh Vora, Venice Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 2
- [2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1
- [3] Haisong Liu, Yao Teng, Tao Lu, Haiguang Wang, and Limin Wang. SparseBEV: High-performance sparse 3D object de-

tection from multi-camera videos. In *ICCV*, 2023. 1, 2