Supplementary Material of CramNet: Camera-Radar Fusion with Ray-Constrained Cross-Attention for Robust 3D Object Detection

Jyh-Jing Hwang, Henrik Kretzschmar, Joshua Manela, Sean Rafferty,
Nicholas Armstrong-Crews, Tiffany Chen, Dragomir Anguelov

Waymo

A Appendix

We propose an efficient camera-radar sensor fusion approach for robust 3D object detection for autonomous driving. The method uses a ray-constrained cross-attention mechanism to leverage the range measurements from radar to improve camera depth estimates, leading to improved detection performance. More importantly, the architecture is designed in a way that training with dropout allows the method to fall back to a single modality when one of the sensors malfunctions.

Here, we include more details on the following aspects:

1. We describe the experimental details and present more complete results on the Waymo Open Dataset in A.1.
2. We study the trade-off between latency and foreground thresholds in A.2.
3. We present ablation study on hyperparameters related to the fusion design, e.g., ray-constrained cross-attention and sensor dropout in A.3.
4. We document the detailed architecture of CramNet in A.4.

A.1 Experiment on Waymo Open Dataset

We use the same hyperparameters as on the RADIATE dataset [3] for training a camera-only CramNet on the Waymo Open Dataset [4]. We adopt a longer training procedure, i.e., 60k warm-up steps and 120k total steps, due to the larger size of the dataset. We align our setting with CaDDN [2] to train and evaluate our performance using the front camera. However, we train our model on a lower resolution $(640, 960)$, than in CaDDN [2], $(832, 1248)$.

We report the 3D AP/APH with 0.5 and 0.7 IoU threshold on the LEVEL_1 and LEVEL_2 difficulties in Table 1. We conclude that our camera-only model, CramNet-C, achieves competitive performance among the state-of-the-art models.

We notice that our model performs significantly better $(+50\% \sim 300\%)$ in the longe range region (50 m - ∞). This suggests the sparse operation in 3D after the 2D segmentation filtering can better handle the long range objects, even without implicitly or explicitly modeling depth uncertainty.

Difficulty	Method	3D AP	0 - 30m	30 - 50m	50m - ∞	3D APH	0 - 30m	30 - 50m	50m - ∞
Level 1 (IoU = 0.5)	M3D-RPN [1]	3.79	11.14	2.16	0.26	3.63	10.70	2.09	0.21
	CaDDN [2]	17.54	45.00	9.24	0.64	17.31	44.46	9.11	0.62
	CramNet-C	11.81	32.20	7.24	**2.00**	11.59	31.75	7.08	**1.93**
Level 2 (IoU = 0.5)	M3D-RPN [1]	3.61	11.12	2.12	0.24	3.46	10.67	2.04	0.20
	CaDDN [2]	16.51	44.87	8.99	0.58	16.28	44.33	8.86	0.55
	CramNet-C	10.64	30.29	6.56	**1.76**	10.44	29.86	6.42	**1.69**
Level 1 (IoU = 0.7)	M3D-RPN [1]	0.35	1.12	0.18	0.02	0.34	1.10	0.18	0.02
	CaDDN [2]	5.03	14.54	1.47	0.10	4.99	14.43	1.45	0.10
	CramNet-C	4.14	**15.46**	1.20	**0.15**	4.10	**15.31**	1.19	**0.13**
Level 2 (IoU = 0.7)	M3D-RPN [1]	0.33	1.12	0.18	0.02	0.33	1.10	0.17	0.02
	CaDDN [2]	4.49	14.50	1.42	0.09	4.45	14.38	1.41	0.09
	CramNet-C	3.72	**14.53**	1.09	**0.13**	3.68	**14.38**	1.07	**0.13**

Table 1: Camera-only 3D detection results on the Waymo Open Dataset [4] validation set on the vehicle class, evaluated in terms of 3D AP/APH at 0.5 or 0.7 IoU on the LEVEL_1 or LEVEL_2 difficulties. Baseline numbers are from [2]. Our camera-only model, CramNet-C, achieves competitive performance among state-of-the-art with the best long range detection.

A.2 Ablation Study on Latency and Foreground Threshold

One of the important trade-off in our model hyperparameters is the foreground segmentation threshold. This threshold controls the density of foreground points passed from the 2D to 3D stage. Therefore, we expect the model to perform better with a lower threshold, with the trade-off of a higher latency.

We summarize this ablation study in Figure 1. Our reported performance in the main paper is at the 0.15 threshold with a latency of 46.4 ms. We observe a general trend of lower accuracy and lower latency when setting a lower threshold.

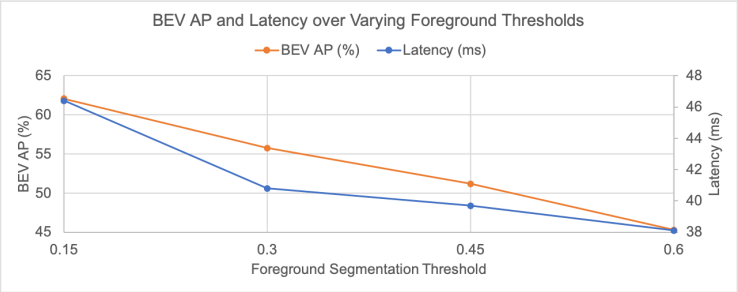

Fig. 1: Ablation study on foreground segmentation thresholds. The model accuracy in BEV AP and latency both decreases as the foreground segmentation threshold decreases.

A.3 Ablation Studies on Fusion Hyperparameters

We conduct ablation studies on the effect of hyperparameters w.r.t. the 3D detection performance in BEV AP, summarized in Figure 2. All in all, the ablation studies suggest the model is not too sensitive to hyperparameters.

For the ray constrained cross-attention, we notice the best error rate ($\epsilon = 0.1$) corresponds to the general depth errors. Also, we do not need many samples along the camera ray (from 3 to 5 sampled points) as each sample already covers a large region through feature extraction.

For the sensor dropout, the performance peaks at 0.2 dropout probability and decreases as the probability increases, indicating that too frequent dropout actually hurts the model.

For modality encoding, when removing the modality code, the BEV AP of CramNet degrades by 8.7 percentage points from 62.1% to 53.4%. The indicates the modality encoding is critical for the model to distinguish and utilize features from different sensors.

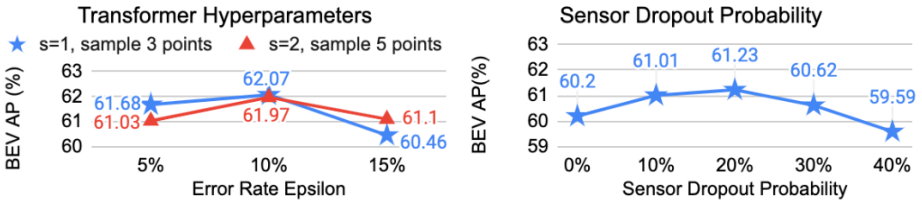

Fig. 2: Left: Ablation study on hyperparameters in ray-constrained cross-attention. **Right**: Abltaion study on hyperparameters in sensor dropout. All in all, the ablation studies suggest the model is not too sensitive to hyperparameters.

A.4 Architecture Details

2D U-Net. A downsampling block $D(B_i, C_i)$ at level i contains B resnet blocks with C-dimensional outputs, with stride 2 in the first convolutional layer. Each upsampling block $U(B_i, C_i)$ at level i also contains B resnet blocks with C-dimensional outputs. The upsampling is performed by a 1×1 convolution layer followed by a bilinear interpolation layer. We connect the same level of the corresponding downsampling block and upsampling block to construct the U-Net.

After applying an initial 1×1 convolution layer on the input with 16-dimensional outputs, we construct a 2-D U-Net with hyperparameters specified in Table 2. We use the exact same 2D U-Net for both camera and radar inputs for simplicity. One can replace them with stronger feature extractor backbones.

3D Sparse U-Net. We reuse the notations as 2D U-Net for 3D U-Net. A residual block in 3D U-Net is replaced by a $3 \times 3(\times 3)$ sparse convolution layer before the

$D(B_1, C_1)$	$D(B_2, C_2)$	$D(B_3, C_3)$	$D(B_4, C_4)$	$U(B_1, C_1)$	$U(B_2, C_2)$	$U(B_3, C_3)$	$U(B_4, C_4)$
$(3, 16)$	$(3, 16)$	$(1, 64)$	$(0, 128)$	$(1, 16)$	$(1, 16)$	$(1, 64)$	$(1, 128)$

Table 2: Detailed hyperparameters to construct a 2D U-Net.

residual connection and by two $3 \times 3(\times 3)$ submanifold sparse convolution layers within the residual block. Unlike the symmetric downsampling and upsampling blocks in the 2D Unet, we employ 2 more downsampling blocks to output a lower resolution objectness heatmap. We summarize the hyperparameters used to construct a 3D sparse U-Net in Table 3.

$D(B_1, C_1)$	$D(B_2, C_2)$	$D(B_3, C_3)$	$D(B_4, C_4)$	$D(B_5, C_5)$	$U(B_1, C_1)$	$U(B_2, C_2)$	$U(B_3, C_3)$
$(1, 96)$	$(2, 96)$	$(2, 96)$	$(1, 96)$	$(1, 96)$	$(0, 96)$	$(2, 96)$	$(2, 96)$

Table 3: Detailed hyperparameters to construct a 3D Sparse U-Net.

References

1. Brazil, G., Liu, X.: M3d-rpn: Monocular 3d region proposal network for object detection. In: ICCV (2019)
2. Reading, C., Harakeh, A., Chae, J., Waslander, S.L.: Categorical depth distribution network for monocular 3d object detection. In: CVPR (2021)
3. Sheeny, M., De Pellegrin, E., Mukherjee, S., Ahrabian, A., Wang, S., Wallace, A.: Radiate: A radar dataset for automotive perception in bad weather. In: ICRA (2021)
4. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al.: Scalability in perception for autonomous driving: Waymo Open Dataset. In: CVPR (2020)