RCDN: Towards Robust Camera-Insensitivity Collaborative Perception via Dynamic Feature-based 3D Neural Modeling Supplementary Material

Tianhang Wang
Tongji University
tianya_wang@tongji.edu.cn

Fan Lu
Tongji University
lufan@tongji.edu.cn

Zehan Zheng
Tongji University
zhengzehan@tongji.edu.cn

Guang Chen[*]
Tongji University
guangchen@tongji.edu.cn

Changjun Jiang
Tongji University
cjjiang@tongji.edu.cn

1 OPV2V-N

To facilitate the research on camera-insensitivity for collaborative perception: i) firstly, as we discussed in related works, multi-view based collaborative perception heals the ill-posed of recovering noisy camera images just from single-view. Owing there are no labels of multi-view based overlap regions in existing collaborative perception, we manually collect the multi-view based overlap regions for RCDN experiments, shown in Figure 1. In detail, we will record the corresponding vehicle IDs, camera IDs and duration time t_{start}, t_{end} of the multi-view based overlap regions; ii) secondly, as we need to distinguish the foreground and background for static and dynamic collaborative neural fields, respectively. We extend the OPV2V[1] with more data format, such as the optical flow (supervise the $\mathbf{s}_{fw,bw}$), mask labels, to bridge the gap between neural field and collaborative perception, as shown in Figure 2.

Data analysis. We manually annotate about 65 scenes, which consists of a total of 6138 collaborative samples. Figure 4 presents some statistical analysis results regarding the OPV2V-N dataset. The OPV2V-N covers situations about 61.86%, 33.47%, and 4.66% for two, three, and four V2X collaborative agents, respectively. Meanwhile, before we conduct the corresponding RCDN experiments, we validate whether the random noisy camera data will affect the collaborative perception system. Table 1 shows that i) the noise actually degrades the system performance; ii) compared to static scenes, dynamic vehicles are more susceptible to the influence of noisy data. With this prior knowledge, we decided to explore the RCDN algorithms and need to pay more attention to optimizing the design for dynamic vehicle perception. We also visualize the specific degradation caused by the noisy camera data, shown in Figure 3. Note that Figure 3 (a) degradation with missed vehicle inspections; Figure 3 (b) degradation with missed Dr. area and lane inspections; Figure 3 (c) degradation with both missed vehicle and Dr. area, lane inspections; Figure 3 (d) no degradation. Also, to make sure that the random noisy camera data is always inputted the same way and that performance does not change because of the different types of noise, like blurred or occluded, we replace the manually annotated camera IDs under the camera failure situation[2].

[*] Corresponding Author. Our code is available at: https://github.com/ispc-lab/RCDN.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

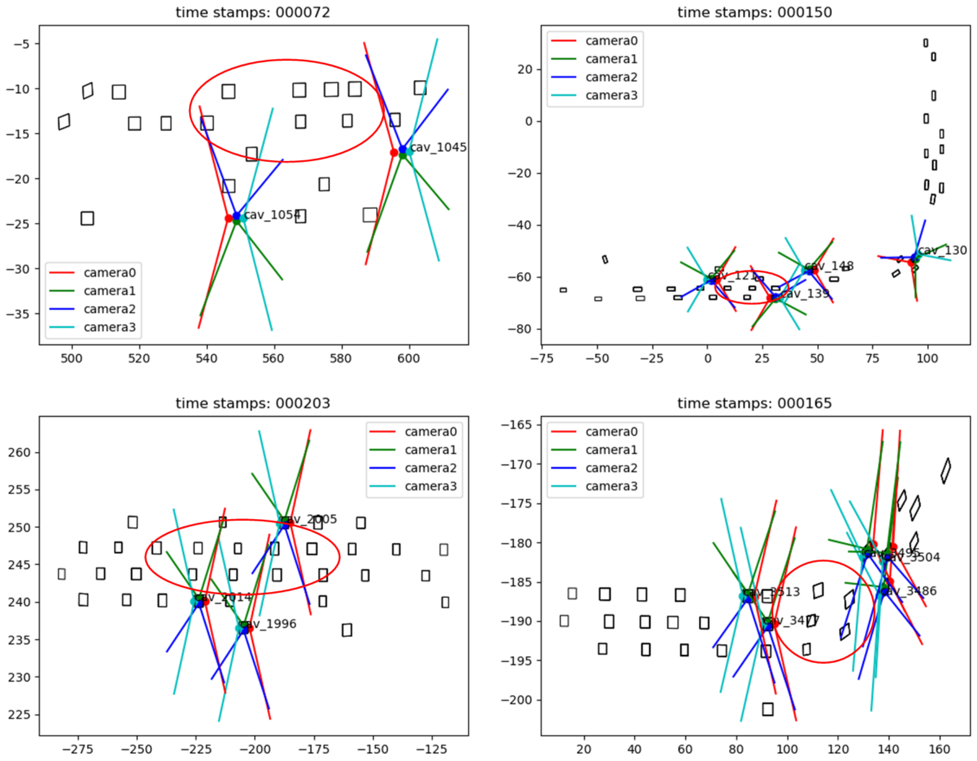
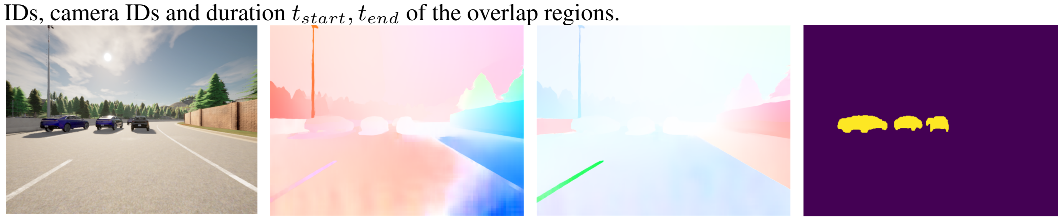


Figure 1: Visualization of manually labeling mechanisms. Note that the red circles represent the multi-view based overlap regions that are suitable for the random noisy situation. We will record the corresponding vehicle IDs, camera IDs and duration t_{start}, t_{end} of the overlap regions.



(a) camera input

(b) forward flow

(c) backward flow

(d) masks

Figure 2: Visualization of extra data format.

2 Detailed Information about Experiments

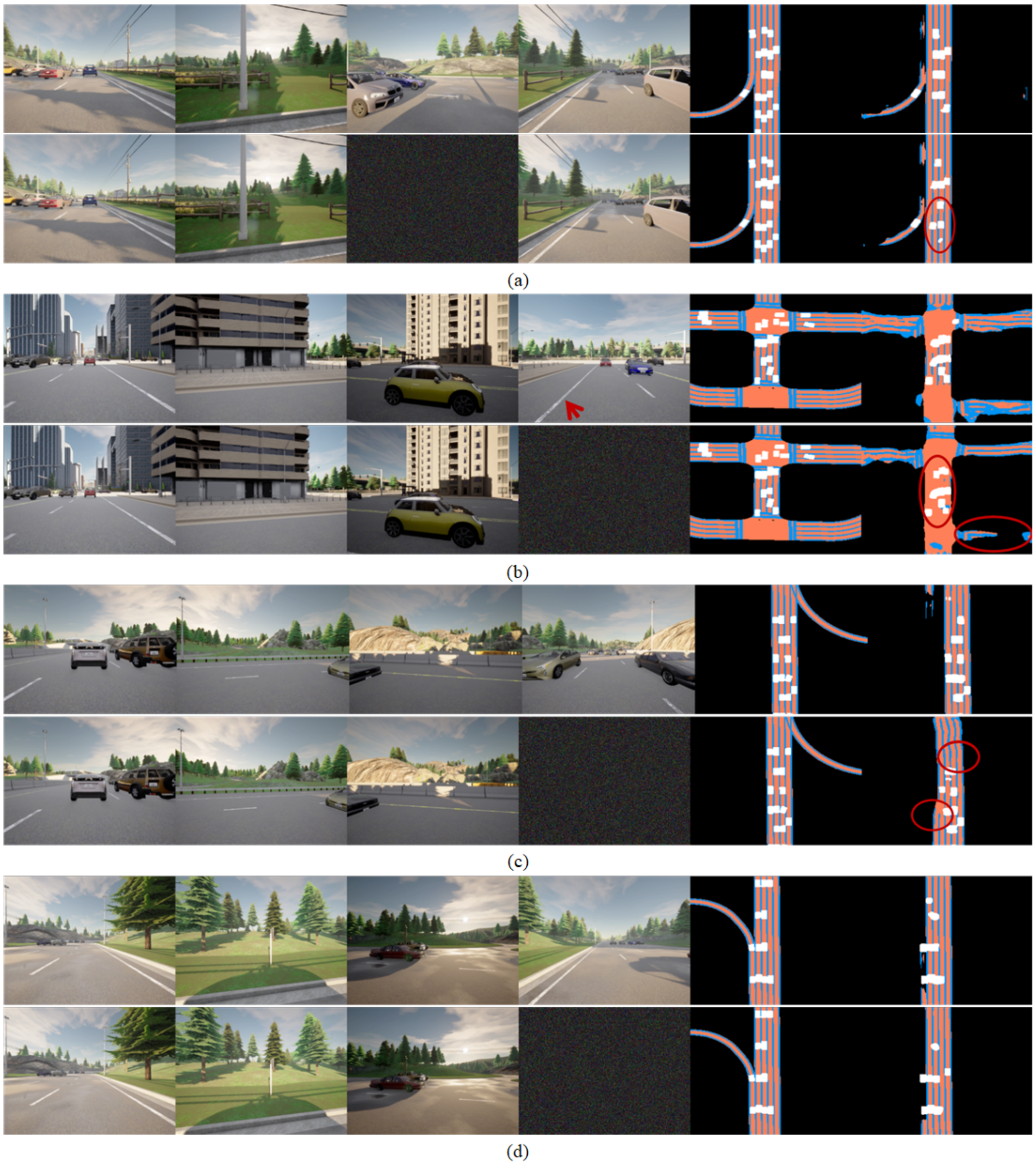
2.1 Implementation Details.

For collaborative perception part, we assume all the AVs have a 70m communication range following[3], and all the vehicles out of this broadcasting radius of ego vehicle will not have any collaboration. We compare with the state-of-the-art multi-agent perception algorithms: F-Cooper, AttFuse, V2VNet, DiscoNet and CoBEVT *w.o/w.* the proposed RCDN.

Table 2: Inference time for each chunk.

Modules	Time Cost
f_{static}	4.47±0.11ms
$f_{dynamic}$	3.94±0.21ms
f_{render}	20.98±0.22ms

Meanwhile, to make a fair comparison, we first employ CVT to extract the BEV feature from camera rigs for all methods. The transmitted BEV intermediate representation has a resolution of 32×32×128; For collaborative neural fields part, we pretrain the BEV decoder with the mcp encoder for better performance, and the geometry collaborative volume feature has a resolution of 128×128×128. Same as [4], we select $(t-1, t, t+1)$ as the mini training unit and train



(a)

(b)

(c)

(d)

Figure 3: Visualization of different performance degradation with random noisy camera data.

the whole model with the Adam[5] optimizer and cosine annealing learning rate scheduler with initial learning rate of 5e-4 on a single RTX 3090 24G GPU with AMD Ryzen Threadripper 3960X. As for the inference time, we record the corresponding time in Table 2. Note that chunk is the smallest unit of parallel processing of the image, *e.g.*, if the image size is $(400, 400)$, the chunk size is 4096 pixels, the number of each image's parallel chunks is about 40.

2.2 Discussion on RCDN.

Theoretically, the RCDN reconstructs the entire collaborative scenario field, according to radial field theory[6], so whichever camera has the noise problem can actually be recovered. In this regard, we experimentally validate the RCDN using CoBEVT and V2VNet, and the corresponding results are in Table 3. We can see that i) if all the RCDN reconstructed cameras are used, the performance is much better compared to using all the noisy camera data, *e.g.* as for CoBEVT, about $62.38\%/123.29\%/262.70\%$ increment for the Dr. area, lanes and dynamic vehicles, respectively. ii) compared to using all normal cameras, using all reconstructed RCDN cameras will degrade the

Table 1: The validation experiments on whether random noise will affect collaborative perception systems. Note that we utilize the current SOTA map-segmentation method, CoBEVT.

OPV2V-N	w. random noisy	w.o. random noisy
Dr. Area	49.37	52.64
Lanes	34.80	37.96
Dynamic Veh.	39.81	47.49

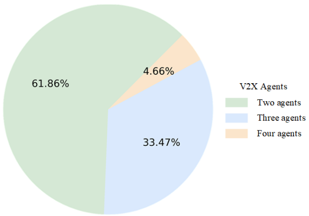


Figure 4: The distributions of V2X collaborative agents.

performance, *e.g.*, as for V2VNet, about $20.94\%/21.70\%/35.73\%$ decrement for the Dr. area, lanes and dynamic vehicles, respectively. To address this phenomenon, we visualize the perspective of the reconstructed camera views, shown in Figure 5, and it is not difficult to find that there is a domain gap[7, 8] between the reconstructed and the normal cameras. Meanwhile, the backbone used to extract the BEV is trained by using the normal camera, so if all reconstructed cameras are used, it does cause a certain degree of degradation. Thanks to the development of abnormal detection algorithms[9, 10], it is easy to find noisy camera data. Hence, we only replace the corresponding noisy data without replacing all data for better performance.

Table 3: Performance comparison (Dr. Area/Lanes/Dynamic Veh.)

methods/setting	all Nor. data	all RCDN data	all noisy data
CoBEVT	51.96/34.19/56.61	36.78/20.90/44.54	22.65/9.36/12.28
V2VNet	41.70/27.14/42.57	32.97/21.25/27.36	18.12/8.76/6.78

2.3 Multi-agents Collaborative Perception

The MCP module stands for the Multi-agent Collaborative Perception process. Existing state-of-the-art (SoTA) MCP modules share a common pipeline: an encoder-fusion-decoder architecture. To ensure fairness in collaborative perception experiments, different MCP modules use the same encoder-decoder architecture but differ in the fusion process. The fusion process is responsible for the bird-eye view (BEV) feature aggregation. Therefore, the MCP module can be replaced by simply switching between different BEV feature aggregation processes.

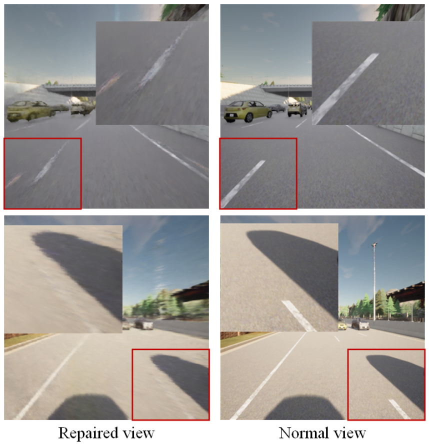


Figure 5: Visualization of domain gap between normal view and RCDN repaired view.

2.4 Benchmarks

We conduct extensive experiments on current collaborative perception methodologies with the proposed RCDN. Table 4 presents the segmentation performance under the expectation of random noisy camera numbers from 0 to 3 on OPV2V-N, which corresponds to the numerical results shown in Figure **??** in the main text. We see that RCDN can be portable to other baseline methods and stabilize the performance even under the extreme camera-insensitivity setting. We also visualize some training scene samples, shown in Figure 6.

2.5 PSNR Results

We also record the corresponding PSNR results of different baseline methods *w.* RCDN's reconstruction's image view, as shown in Table 5. Note that the term peak signal-to-noise ratio (PSNR) is an expression for the ratio between the maximum possible value (power) of a signal and the power of distorting noise that affects the quality of its representation. Hence, the higher PSNR, the better image quality.

Table 4: Performance of RCDN with other baseline methods. Note that − represents the failed results.

Number / Performance / Method		Dr. Area/Lanes/Dynamic Veh.							
		w.o. RCDN				w. RCDN			
		n=0	n=1	n=2	n=3	n=0	n=1	n=2	n=3
F-Cooper	scene1	34.03/17.55/55.56	19.37/6.67/28.96	20.41/7.77/24.39	22.68/9.27/17.56	34.03/17.55/55.56	35.46/16.82/52.69	36.85/16.23/44.45	35.31/16.88/43.49
	scene2	48.14/24.45/55.32	44.55/16.81/20.89	35.79/17.10/24.98	35.83/16.34/22.39	48.14/24.45/55.32	48.50/24.75/52.66	47.54/22.84/52.86	46.90/19.58/54.19
	scene3	35.00/29.86/66.79	25.43/18.01/38.00	26.11/14.99/26.28	21.94/11.56/24.74	35.00/29.86/66.79	35.93/30.10/67.70	34.40/29.51/61.14	36.77/28.86/67.09
	scene4	65.36/63.79/77.24	30.95/17.74/31.34	27.02/18.48/16.31	-/-/7.03	65.36/63.79/77.24	64.98/62.73/77.77	65.93/62.15/67.66	-/-/70.18
	scene5	44.67/30.20/61.74	24.03/20.50/29.32	25.57/15.51/15.49	24.03/17.55/17.32	44.67/30.20/61.74	39.56/26.73/58.00	38.20/23.83/52.49	38.51/19.31/52.02
	Avg	**45.44/33.17/63.33**	**28.87/15.95/29.70**	**26.98/14.77/21.49**	**26.12/13.68/17.808**	**45.44/33.17/63.33**	**44.89/32.23/61.76**	**44.58/30.91/55.72**	**39.37/21.16/57.39**
AttFuse	scene1	32.29/20.38/43.32	19.07/14.04/26.07	17.58/10.94/14.55	17.28/12.77/11.68	32.29/20.38/43.32	29.32/17.16/42.11	28.25/15.05/37.25	27.05/16.39/35.11
	scene2	49.38/20.84/52.16	35.85/16.16/20.36	30.96/18.91/10.93	27.28/17.40/12.14	49.38/20.84/52.16	51.08/22.75/47.93	53.02/26.69/41.57	53.15/26.38/40.16
	scene3	38.79/32.05/57.73	32.64/26.01/36.76	30.67/24.36/22.91	25.48/17.04/15.75	38.79/32.05/57.73	38.19/29.78/56.18	36.75/28.60/46.56	38.86/31.12/42.80
	scene4	58.33/57.97/63.95	24.26/14.31/20.98	19.26/12.38/14.84	-/-/9.18	58.33/57.97/63.95	58.65/55.44/63.11	54.56/48.79/60.59	-/-/62.64
	scene5	49.18/37.55/53.56	28.15/23.32/19.65	23.22/18.34/18.09	23.15/18.47/19.30	49.18/37.55/53.56	44.65/32.35/51.41	41.89/25.55/47.17	39.74/21.54/46.21
	Avg	**45.59/33.76/54.14**	**27.99/18.77/24.76**	**24.34/16.99/16.26**	**23.30/16.42/13.61**	**45.59/33.76/54.14**	**44.38/31.50/52.15**	**42.89/28.94/46.63**	**39.70/23.86/45.38**
DiscoNet	scene1	45.31/23.68/43.26	15.73/6.68/7.47	11.85/7.38/3.71	13.15/9.24/5.35	45.31/23.68/43.26	34.20/19.25/38.88	30.34/13.73/38.84	23.42/7.85/22.83
	scene2	36.58/8.20/37.10	31.87/8.96/11.59	24.74/10.80/11.43	22.70/11.36/7.62	36.58/8.20/37.10	33.60/8.37/34.90	35.72/11.28/32.53	30.72/11.83/29.87
	scene3	28.61/17.05/35.23	19.51/10.48/2.50	14.87/25.98/3.56	14.42/8.73/-	28.61/17.05/35.23	28.89/18.44/28.63	25.98/16.46/22.40	27.33/18.86/21.04
	scene4	50.28/37.95/62.15	32.42/20.69/11.13	24.30/12.44/7.84	-/-/3.46	50.28/37.95/62.15	49.50/38.23/59.25	51.25/40.25/55.49	-/-/54.21
	scene5	50.74/34.33/55.07	22.03/14.64/13.54	22.59/13.58/16.91	27.55/14.70/14.06	50.74/34.33/55.07	46.50/30.58/53.51	41.86/28.56/48.83	44.22/25.07/48.82
	Avg	**42.30/24.24/46.56**	**24.31/12.29/9.25**	**19.67/14.04/8.69**	**21.25/12.32/6.10**	**42.30/24.24/46.56**	**38.54/22.97/43.03**	**37.03/22.06/39.62**	**37.33/23.10/35.35**
V2VNet	scene1	39.83/19.49/28.06	20.52/7.15/8.21	11.41/7.68/5.68	9.54/7.33/3.94	39.83/19.49/28.06	39.28/19.95/28.72	36.13/15.55/28.25	30.56/15.21/19.67
	scene2	40.21/14.45/39.60	37.38/7.19/14.73	29.05/6.84/6.77	19.64/6.66/8.58	40.21/14.45/39.60	38.45/14.96/39.45	34.61/8.07/39.17	33.40/12.29/33.17
	scene3	33.24/26.83/31.89	20.11/6.11/2.93	12.39/7.28/8.37	9.48/7.01/1.49	33.24/26.83/31.89	29.53/20.95/35.15	24.75/16.82/26.92	27.94/16.59/20.52
	scene4	49.76/36.24/57.83	28.50/17.23/11.52	26.74/18.85/9.37	14.54/12.00/8.00	49.76/36.24/57.83	48.34/36.74/58.46	47.00/36.37/57.31	53.40/40.01/56.45
	scene5	45.47/38.71/55.45	33.46/14.91/18.99	29.66/17.04/16.38	24.23/14.99/14.10	45.47/38.71/55.45	42.98/33.59/52.06	39.04/28.94/52.93	39.21/28.99/47.79
	Avg	**41.70/27.14/42.57**	**27.99/10.52/11.28**	**21.85/11.54/9.31**	**15.49/9.60/7.22**	**41.70/27.14/42.57**	**39.72/25.24/42.77**	**36.31/21.15/40.92**	**36.90/22.62/35.52**
CoBEVT	scene1	30.59/12.43/47.62	24.55/10.07/30.67	23.23/8.19/18.83	22.99/11.63/14.59	30.59/12.43/47.62	27.37/10.63/46.65	27.22/11.31/41.38	26.96/10.59/36.99
	scene2	39.28/10.72/54.25	37.47/14.26/28.43	32.32/12.94/16.57	23.43/8.41/11.93	39.28/10.72/54.25	37.47/9.64/50.48	44.97/15.85/49.88	39.51/9.77/46.46
	scene3	47.80/37.85/60.40	24.91/11.90/33.73	18.34/11.15/18.29	15.78/11.56/63.49	47.80/37.85/60.40	49.45/37.51/60.02	50.96/40.09/57.53	49.85/37.13/63.49
	scene4	73.62/61.44/61.70	46.69/20.88/38.91	34.26/12.43/12.10	-/-/8.67	73.62/61.44/61.70	70.27/59.07/62.01	68.48/56.13/53.58	-/-/56.71
	scene5	68.50/48.53/59.06	26.80/15.13/30.32	28.38/8.42/14.19	23.51/11.26/10.67	68.50/48.53/59.06	51.39/30.88/56.35	52.02/30.53/51.86	46.52/25.31/53.37
	Avg	**51.96/34.19/56.61**	**32.08/14.45/32.41**	**27.31/10.63/15.99**	**22.05/10.53/11.52**	**51.96/34.19/56.61**	**47.19/29.55/55.10**	**48.73/30.78/50.85**	**46.10/27.78/51.40**

2.6 Geometry BEV Volume Feature

We utilize the geometry BEV volume feature to speed up the training process and improve the generality of the collaborative neural fields. We observe that with f_{geo_bev} the RCDN can obtain higher PSNR initial values and a shorter training process, as shown in Figure 7.

Table 5: The PSNR results.

	PSNR
F-Cooper	26.11
AttFuse	25.79
DiscoNet	25.86
V2VNet	26.14
CoBEVT	25.89

2.7 Loss Functions

Similar to [11], we regularize the collaborative scene flow to be spatially smooth by minimizing the difference between neighboring 3D points' scene flow. To regularize the consistency of the collaborative scene flow, we have the scene flow cycle consistency regularization as follows:

$$\begin{aligned}\mathcal{L}_{cyc} &= \sum \|\mathbf{s}_{fw}(\mathbf{r}, t) + \mathbf{s}_{bw}(\mathbf{r} + \mathbf{s}_{fw}(\mathbf{r}, t), t+1)\|_2^2 && \text{(1a)}\\ &+ \sum \|\mathbf{s}_{bw}(\mathbf{r}, t) + \mathbf{s}_{fw}(\mathbf{r} + \mathbf{s}_{bw}(\mathbf{r}, t), t+1)\|_2^2 && \text{(1b)}\end{aligned}$$

As for the weights of different losses in Eq **??**, we set $\lambda_1, \lambda_2, \lambda_4 = 1.0$ and $\lambda_3 = 0.1$ for training. Meanwhile, we train the whole network for about 1000 epochs, which takes about 20 to 30 minutes.

3 Visualization

We visualize some segmentation results of different baseline methods *w.o/w* RCND under the different scenes, shown in Figure 8-12.

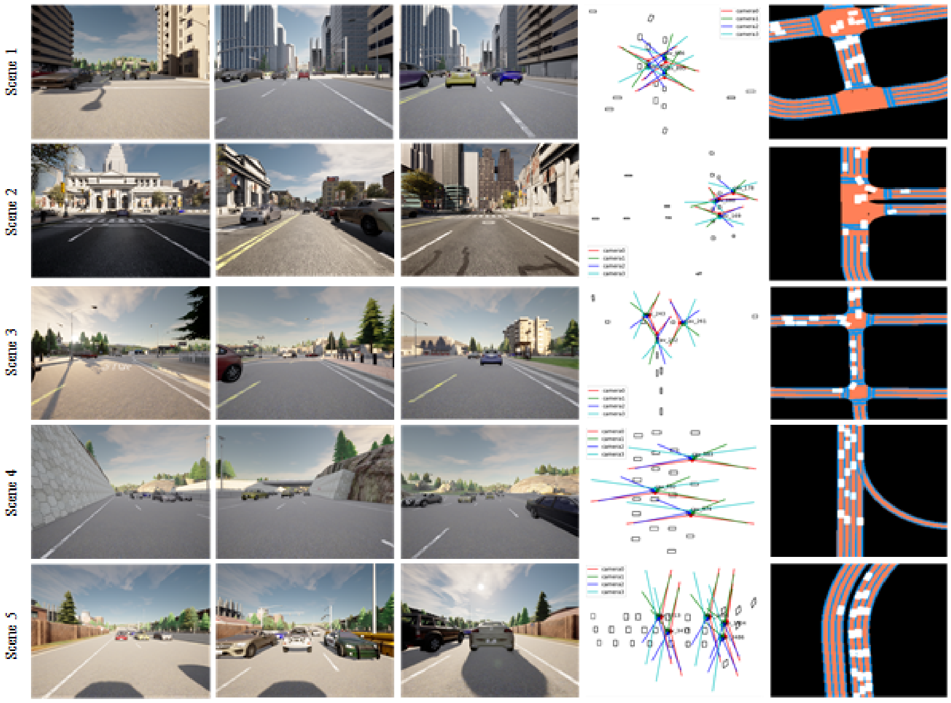
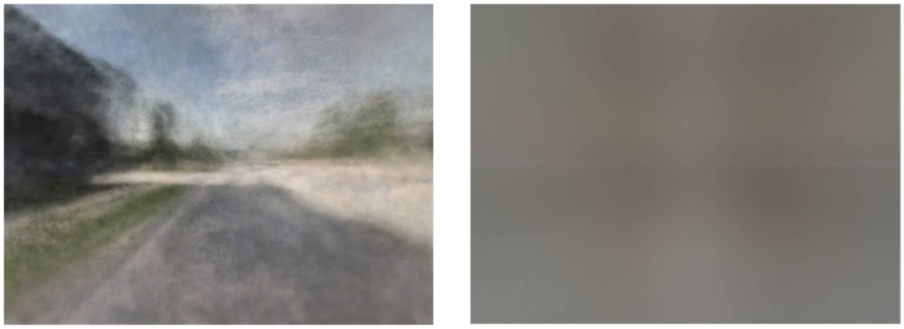


Figure 6: Visualization of some training scenes samples. Note that we cover the classical scenes, including the four-way Intersection, T Intersection, Midblock, Entrance Ramp and Curvy Segment.



w. geometry feature modeling
(PSNR=14.79)

w.o. geometry feature modeling
(PSNR=4.64)

Figure 7: Visualization of *w.o/w.* geometry BEV volume feature modeling.

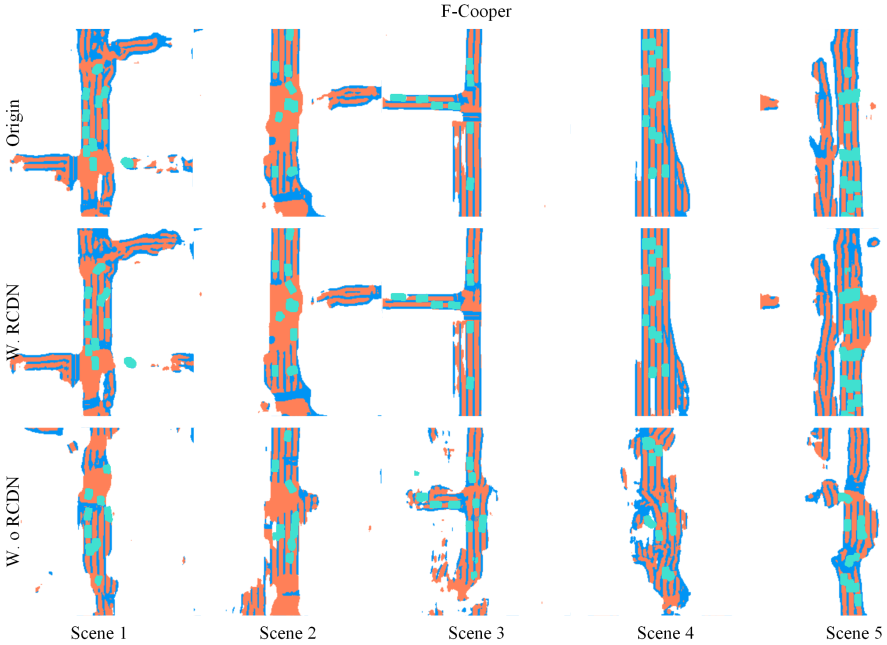


Figure 8: Visualization of baseline method of F-Cooper *w.o/w.* RCDN with one random camera failure.

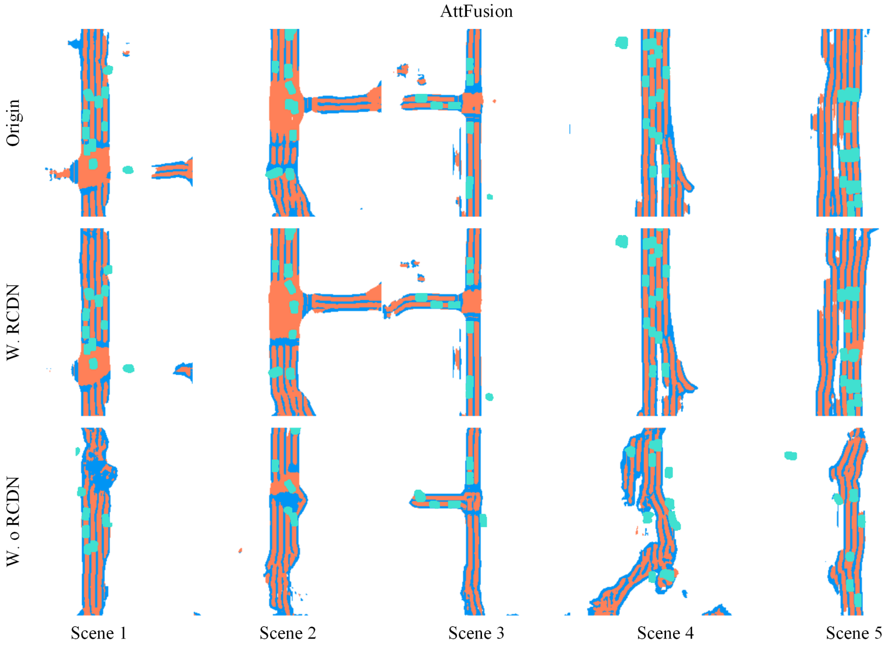


Figure 9: Visualization of baseline method of AttFuse *w.o/w.* RCDN with one random camera failure.

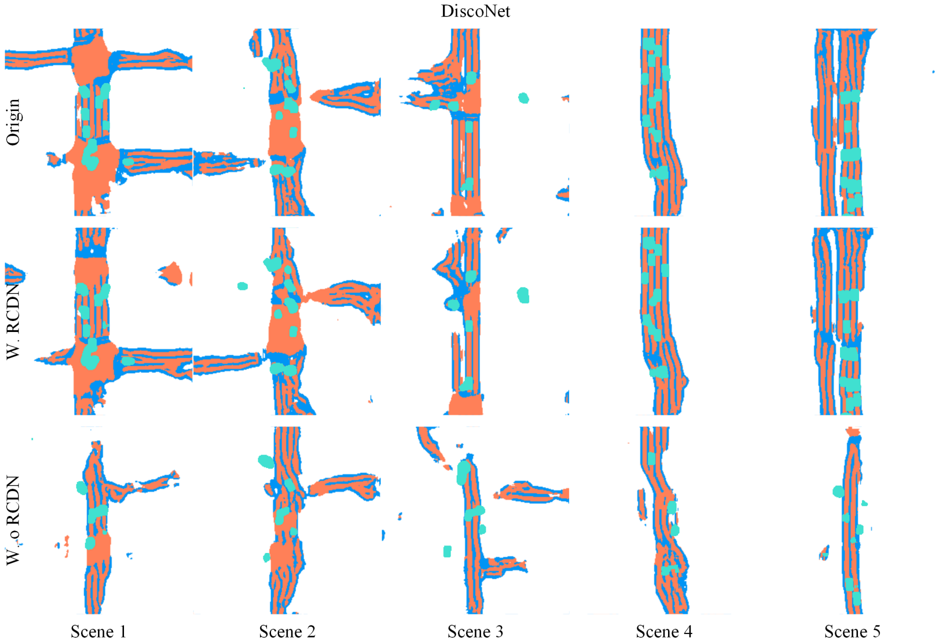


Figure 10: Visualization of baseline method of DiscoNet *w.o/w.* RCDN with one random camera failure.

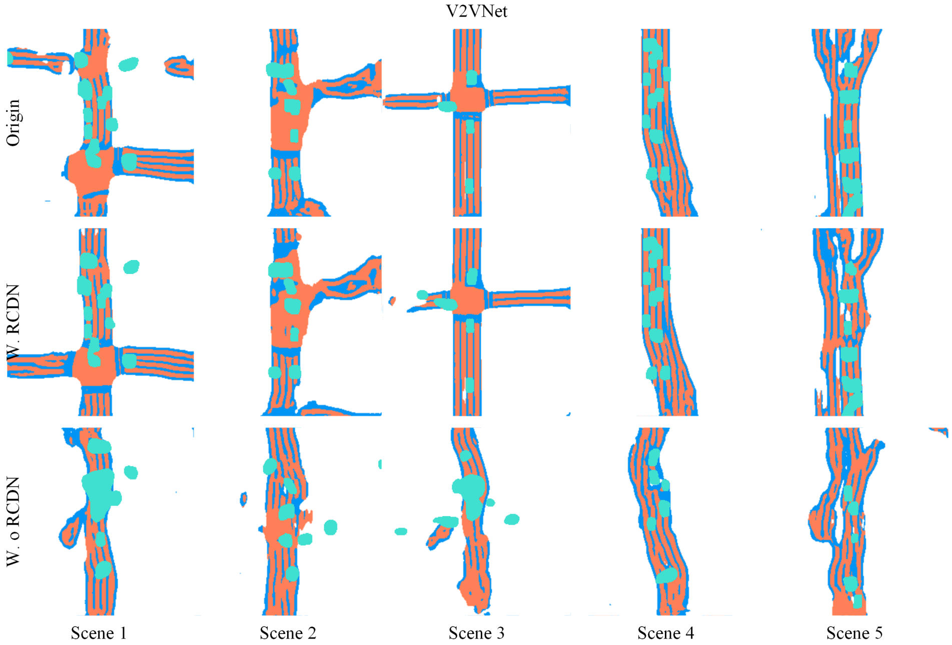


Figure 11: Visualization of baseline method of V2VNet *w.o/w.* RCDN with one random camera failure.

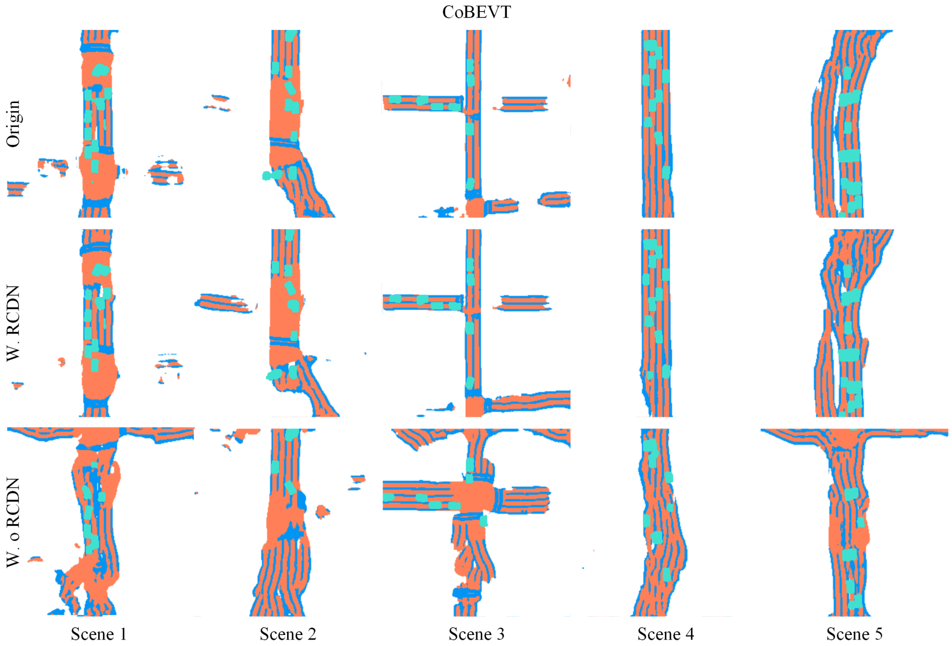


Figure 12: Visualization of baseline method of CoBEVT *w.o/w.* RCDN with one random camera failure.

References

[1] Runsheng Xu, Hao Xiang, Xin Xia, Xu Han, Jinlong Li, and Jiaqi Ma. Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2583–2589. IEEE, 2022.

[2] Kaicheng Yu, Tang Tao, Hongwei Xie, Zhiwei Lin, Tingting Liang, Bing Wang, Peng Chen, Dayang Hao, Yongtao Wang, and Xiaodan Liang. Benchmarking the robustness of lidar-camera fusion for 3d object detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3188–3198, 2023.

[3] Runsheng Xu, Hao Xiang, Zhengzhong Tu, Xin Xia, Ming-Hsuan Yang, and Jiaqi Ma. V2x-vit: Vehicle-to-everything cooperative perception with vision transformer. *ArXiv*, abs/2203.10638, 2022.

[4] A. Schmied, T. Fischer, M. Danelljan, M. Pollefeys, and F. Yu. R3d3: Dense 3d reconstruction of dynamic scenes from multiple cameras. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3193–3203, Los Alamitos, CA, USA, oct 2023. IEEE Computer Society.

[5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[6] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

[7] Runsheng Xu, Jinlong Li, Xiaoyu Dong, Hongkai Yu, and Jiaqi Ma. Bridging the domain gap for multi-agent perception. *arXiv preprint arXiv:2210.08451*, 2022.

[8] Sanqing Qu, Tianpei Zou, Florian Röhrbein, Cewu Lu, Guang Chen, Dacheng Tao, and Changjun Jiang. Upcycling models under domain and category shift. In *CVPR*, 2023.

[9] Anuroop Gaddam, Tim Wilkin, and Maia Angelova. Anomaly detection models for detecting sensor faults and outliers in the iot - a survey. In *2019 13th International Conference on Sensing Technology (ICST)*, pages 1–6, 2019.

[10] Tianhang Wang, Kai Chen, Guang Chen, Bin Li, Zhijun Li, Zhengfa Liu, and Changjun Jiang. Gsc: A graph and spatio-temporal continuity based framework for accident anticipation. *IEEE Transactions on Intelligent Vehicles*, 9(1):2249–2261, 2024.

[11] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021.