SRCN3D: Sparse R-CNN 3D for Compact Convolutional Multi-View 3D Object Detection and Tracking

Yining Shi[1], Jingyan Shen[1], Yifan Sun[1], Yunlong Wang[1],
Jiaxin Li[1], Shiqi Sun[1], Kun Jiang[1*], Diange Yang[1*]
[1] Tsinghua University * Correspondence author

1. Model Appendix

1.1. Preliminaries

3D object detection for autonomous driving perception systems aims to classify objects of interest into according categories and predict 3D bounding boxes, given camera images and corresponding camera parameters. Basically, each 3D bounding box is represented by parameters including its translation $[x, y, z]$, dimension $[w, l, h]$, rotation θ and velocity $[v_x, v_y]$. On the other hand, multi-object tracking associates detected objects in the temporal dimension, as well as records unique labels and past trajectories of each object of interest.

1.2. Set Prediction Loss

SRCN3D adopts set prediction loss for learning. In a set-to-set prediction, the loss function is optimized based on an optimal bipartite matching between predictions and ground truth sets. Hungarian algorithm is implemented to produce the optimal assignment. Assuming there are N predictions denoted by $\{(\hat{c}_i, \hat{b}_i)\}_{i=1}^{N}$, and the ground truth set $\{(c_i, b_i)\}_{i=1}^{N}$ is padded to the same size with $\emptyset$. Then the matching cost can be written as $\sum_{i=1}^{N}[-log\hat{p}_{\sigma^*(i)}(c_i) + \mathbb{I}_{\{c_i \neq \emptyset\}}\mathcal{L}_1(b_i, \hat{b}_{\sigma^*(i)})]$, where $\sigma^* = \arg\min_{\sigma} \sum_{i=1}^{N}[\hat{p}_{\sigma(i)}(c_i) + \mathbb{I}_{\{c_i \neq \emptyset\}}\mathcal{L}_1(b_i, \hat{b}_{\sigma(i)})]$ denotes the optimal index ranking.

2. Data Appendix

2.1. Data Pre-processing

Experiments are conducted on nuScenes dataset [1]. We follow common pre-processing practices as [3]. For both training and inference phrases, we first load multi-view image files with camera parameters, then apply normalization, padding and multi-scale flipping to each input image for data augmentation. During training, GridMask [2] is employed to randomly remove some pixels. During inference phrase, test time augmentation is not applied.

3. Experimental Details

3.1. Pipelines

Pipeline. The bounding box of each detected object is predicted in ego vehicle coordinate and transformed into world coordinate for multi-object detection and tracking. Score thresholds for both tasks are both 0.2.

3.2. Training Details

Hyper-parameters of our model is set as Table.1, which are selected based on experiments.

Hyper-parameters	Settings
FPN input channels	$[256, 512, 768, 1024]$
FPN output channels	256
Number of cascade stages	6
Number of proposal boxes	900
Number of heads	8
FFN layers	2
Classification branch layers	2
Regression branch layers	3
Optimizer	AdamW
Initial learning rate	$2e^{-4}$
Learning rate policy	Cosine annealing
Warm up steps	500
Warm up ratio	$1/3$
Batch size	2
Training epochs	24

Table 1. Hyper-parameters for SRCN3D.

4. More Visualization

In this section, we present more visualization results on nuScenes [1] dataset to illustrate the effectiveness of our approach. Fig.1 and Fig.2 display the visualization results on nuScenes detection test set where true annotations are held out for the leaderboard. In each figure, predicted boxes are drawn in different colors corresponding to predicted categories. As shown in Fig.1, our model is able to classify

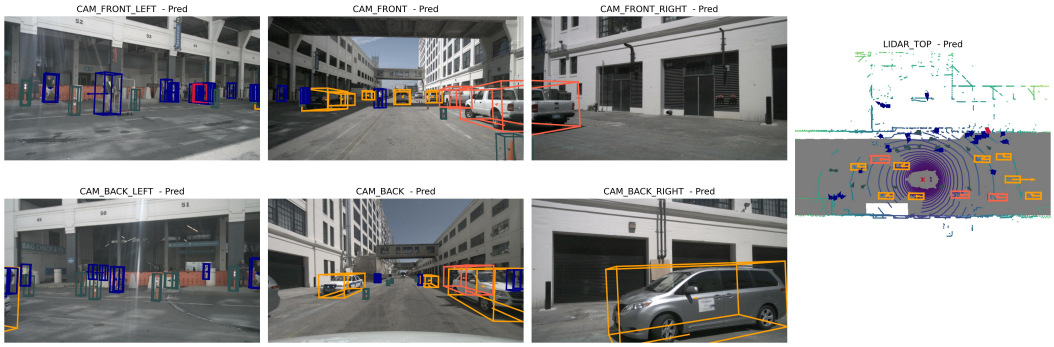

Figure 1. Visualization of predictions on detection test set. Objects like traffic cones, vehicles, and trails are distinctly detected in the scene. Different colors indicate classifications of objects.

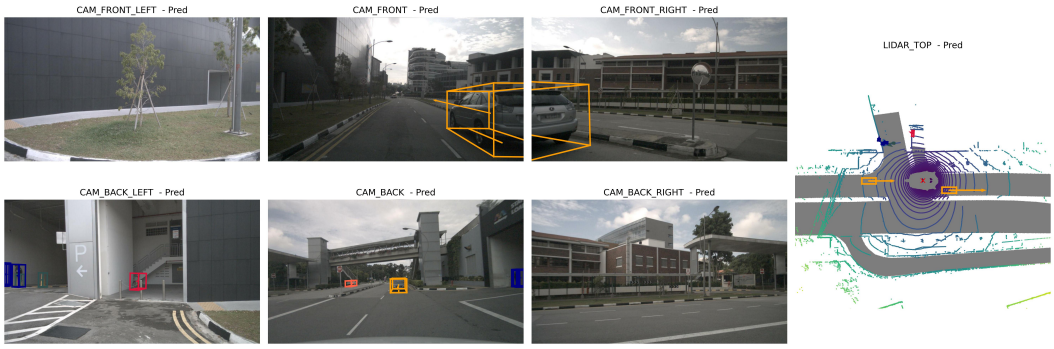

Figure 2. Visualization of a case of cross-view fusion on nuScenes test set. As a successful example of cross-view fusion, the white car appearing in both the front and front-right view is recognized as a whole.

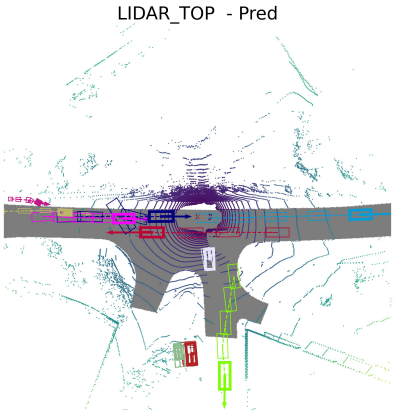

Figure 3. Tracking visualization for an intersection scenario, in which surrounding vehicles are tracked accurately.

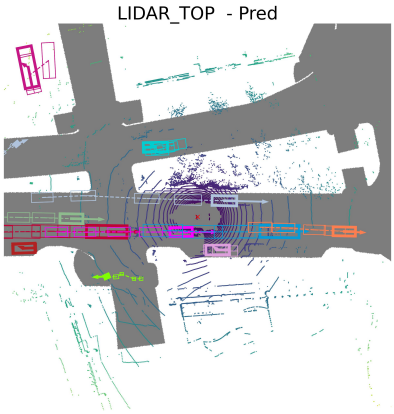

Figure 4. Tracking visualization for a more crowded crossroad, in which near and remote objects have continuous unique IDs, implying good track continuity.

and locate different objects precisely. It successfully detects small objects like traffic cones, as well as objects in the distance.

4.1. Visualization of Sparse Local Attention

Fig.5 visualizes all 3D proposal boxes of each stage in camera front views. The iterations of proposal boxes show a tendency from arbitrary distribution in the whole images to target areas and appropriate sizes. This process shows that

(a) Stage 1 (b) Stage 3

(c) Stage 5 (d) Stage 6

Figure 5. Visualization of learnable proposal boxes of certain stages on nuScenes val set. As the stage increases, boxes in images locate more densely on areas of objects. Distribution of proposal boxes shows a thorough coverage of possible locations of detected objects.

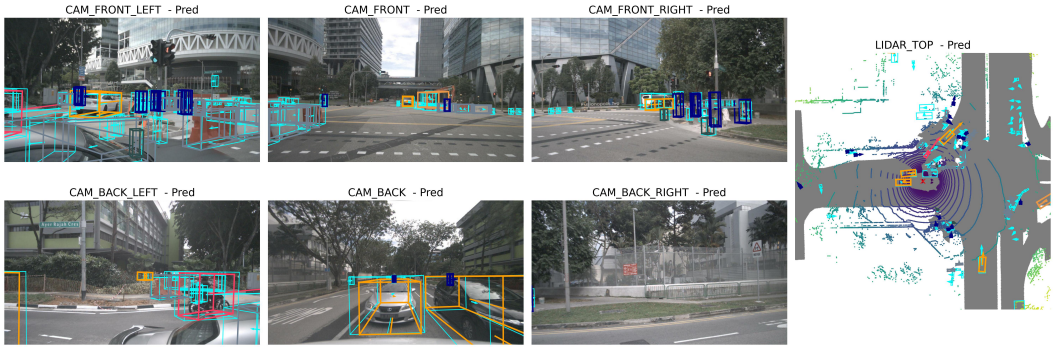

Figure 6. Visualization of a failure scene of SRCN3D on nuScenes val set. Objects in this crowded scene are densely distributed and concentrated. In the front-left camera view, it shows quantities of overlapped and redundant predicted boxes. There also exists many occlusions. For instance, some blocked objects in the back-left view are not captured in our predictions.

SRCN3D head captures semantic characteristics well with small sets of regional proposals and local attention mechanism, so that convergence is guaranteed during the cascade refinement.

4.2. Visualization of Cross-view Fusion

The effect of cross-view fusion is that objects appearing in multiple perspectives can be detected as a whole through boxes partially inside the views. Fig.2 shows a case of cross-view fusion prediction on the test set. It can be observed that the same object that appears in two perspectives is framed as one. This demonstrates that the RoI aggregation strategy in the sparse feature sampling module successfully realizes cross-view fusion detection.

4.3. Visualization of Tracking

Fig.3 and Fig.4 present two more visualization results on nuScenes tracking validation set in the bird's eye view. The tracked objects are shown in the past five keyframes of the same scene. Boxes in different colors refer to different tracked instances. SRCN3D succeeds in tracking distant objects continuously given inaccurate detection results.

4.4. Visualization of Failure Cases

Finally we report a failure case of SRCN3D detection on the validation set in comparison to the ground truth boxes. As observed in Fig.6, in several perspectives, SRCN3D fails to distinguish objects that are seriously blocked (e.g. vehicles and pedestrians in the back-left view). Besides, our approach still remains limited in detecting objects in some densely distributed regions, where predicted boxes are commonly overlapped. It tends to make redundant predictions and misclassify the attributes.

References

[1] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar 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. 1

[2] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Gridmask data augmentation. *arXiv preprint arXiv:2001.04086*, 2020. 1

[3] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. DETR3d: 3d object detection from multi-view images via 3d-to-2d queries. In *5th Annual Conference on Robot Learning*, 2021. 1