A. Architectural Modifications and Details

To reduce the complexity of the CPSeg, all 5×5 and 7×7 convolutions were replaced by 3×3 convolutions with a dilation rate of 2 and 3, respectively. Also, every layer is limited to output a multiple of 24 feature maps instead of 32 in the original CPSeg model. The Task-Aware-Attention (TAM) module was removed due to its negligible effect on the segmentation performance.

Similarly, in our best-performing model, which was based on the single-stage CenterPoint (with VoxelNet backbone), we also reduced the complexity for its 3D object detection backbone. More specifically, we removed one sparse residual block and reduced the number of output feature maps for the second and third sparse sequential blocks by 8 and 32, respectively.

When combining multi-view feature maps from CPSeg and the 3D object detection backbone, the proposed cascade feature fusion module is designed to ensure that the number of RV-based feature maps from CPSeg matches the number of BEV-based feature maps from the 3D backbone. As such, the number of output features of the proposed cascade feature fusion module was set to 256 for SECOND and CenterPoint, and 64 for PointPillars. The boundary refine block in the cascade feature fusion module is composed of a 3×3 convolution layer followed by batch normalization, a ReLU activation function, and another 3×3 convolution layer. The resulting feature maps maintain the same number of channels as input feature maps, and the output feature maps are obtained by adding the resulting feature maps to the input feature maps.

In the RV-BEV feature weighting module, the MLP block is composed of one input layer with 1024 neurons, one hidden layer with 256 neurons, and one output layer with 512 neurons. The ReLU activation function is only used in the hidden layer.

In the class-wise foreground attention module and center density heatmap module, we down-sampled the foreground semantic map and the center density heatmap using three consecutive max-pooling layers.

B. Implementation details

For all experiments on the nuScenes dataset, we set the detection range to $[-51.2m, 51.2m]$ for x- and y-axis, and $[-5m, 3m]$ for z-axis. The voxel size was set to $(0.1m, 0.1m, 0.2m)$ for models based on the VoxelNet backbone such as SECOND or CenterPoint (VoxelNet-based version), and $(0.2m, 0.2m)$ for models based on the PointPillars backbone such as PointPillars and CenterPoint PointPillars-based version).

Data augmentation was exploited for all the compared methods on the nuScenes validation dataset. Random flipping along both x- and y-axis, global scaling with a random factor from $[0.9, 1.1]$, and random global rotation along z-axis between $[-\frac{\pi}{4}, \frac{\pi}{4}]$ were performed. Ground-truth sampling [26] was used to address the long-tail class distribution, which copies and pastes points inside an annotated box from one sample frame to another.

Method	# Params (M)	mAP	NDS
PointPillars [9]	6.1	43.0	56.8
Complex-PointPillars	14.5	43.9	57.3
Multi-Task+PointPillars	14.8	**50.5**	**60.5**

Table 6. Comparing the performance of 3D object detection methods considering their complexity on nuScenes validation set.

C. Quantitative Comparison

In order to demonstrate that the increased complexity is not the key for performance improvement of different BEV-based 3D object detection method under the proposed multi-task framework, we performed another set of experiments. As our PointPillars-based multi-task method has more complexity compared to the original PointPillars [9], we trained a complex version of the PointPillars with almost the same complexity as ours. In Table 6, we compared these three different models. As can be seen, by solely increasing the detector complexity the performance improvement is marginal. However, the proposed method with almost the same complexity as the Complex-PointPillars remarkably performs better.

In Table 7, we demonstrate the performance improvement that the proposed framework provides for different BEV-based 3D object detection methods. For each method, the first row indicates its performance as a standalone model without the panoptic segmentation guidance (PSG). In contrast, the second row provides results when integrated as a part of the multi-task framework with PSG. For each CenterPoint method, the two-stage version is used when PSG is No and the single-stage version is exploited when PSG is Yes. In this table, it can be seen that adding the panoptic segmentation information as guidance improves the overall detection accuracy of all methods tested considerably, regardless of the type of the detection backbone and detection head used.

Waymo Open Dataset [21] is another publicly available large scale 3D object detection dataset. It does not include the panoptic segmentation labels, which makes it less ideal for our framework. However, we prepared panoptic labels using the annotated object boxes to train our method on this dataset. All the points inside the annotated 3D boxes are assigned with their corresponding box semantic labels and instance IDs, while all the points outside of these boxes are labeled as a single background class. We report the mAP and the mean Average Precision weighted by Head-

Method	PSG	mAP	NDS	Car	Truck	Bus	Trailer	CV	Ped	Motor	Bic	TC	Barrier
PointPillars [9]	No	43.0	56.8	**80.9**	**50.5**	62.1	30.9	11.0	71.8	29.4	5.5	43.8	44.4
	Yes	**50.5**	**60.5**	76.1	50.2	**62.6**	**32.9**	**15.0**	**77.0**	**52.5**	**20.4**	**60.1**	**58.1**
SECOND [26]	No	51.7	62.6	82.6	53.2	65.6	36.3	16.3	79.0	44.4	19.8	60.3	58.9
	Yes	**56.2**	**64.8**	**83.0**	**55.9**	**69.3**	**42.3**	**20.1**	**80.1**	**57.0**	**27.9**	**66.2**	**60.8**
CenterPoint$_{PP}$ [30]	No	50.3	60.2	**84.0**	**53.5**	64.3	31.9	12.5	78.9	44.0	18.2	54.9	60.3
	Yes	**54.3**	**63.1**	79.4	52.9	**69.3**	**34.7**	**13.0**	**82.9**	**53.0**	**29.5**	**66.0**	**62.6**
CenterPoint$_{VN}$ [30]	No	56.4	64.8	84.7	54.8	67.2	35.3	17.1	82.9	57.4	35.9	63.3	65.1
	Yes	**60.3**	**67.1**	**85.1**	**57.1**	**68.3**	**43.6**	**20.5**	**84.7**	**62.5**	**43.6**	**71.5**	**66.0**

Table 7. Performance comparison of different BEV-based 3D object detection methods with/without panoptic segmentation guidance (PSG) based on the nuScenes validation set. In the columns, CV, Ped, Motor, Bic, and TC are abbreviations for Construction Vehicle, Pedestrian, Motorcycle, Bicycle, and Traffic Cone, respectively. CenterPoint$_{PP}$ and CenterPoint$_{VN}$ represent the CenterPoint method with PointPillars and VoxelNet backbones, respectively. For the CenterPoint method, based on both backbones, when PSG is No, the two-stage version is used and when the PSG is Yes, the single-stage version is exploited.

Method	Car L1	Car L2	Ped L1	Ped L2	Cyc L1	Cyc L2
CenterPoint [30]	71.33/70.76	63.16/62.65	72.09/65.49	64.27/58.23	68.68/67.39	66.11/64.87
Ours	**72.72/72.11**	**64.65/64.10**	**73.76/67.51**	**65.86/60.12**	**69.05/67.87**	**66.56/65.42**

Table 8. 3D object detection comparison of the proposed method and the CenterPoint [30] on the Waymo validation set, trained with 20% of training data. The result shown in each column is the mAP and mAPH for each object class. In the columns, Ped, and Cyc are abbreviations for Pedestrian, and Cyclist, respectively.

ing (mAPH) for the 3D object detection task. We trained the proposed model on the 20% of the training data and evaluated on the whole validation data. The comparison results with the CenterPoint model are shown in Table 8. Although this model is not suitable for our framework as it misses the panoptic labels, it can be seen that the proposed model outperforms the CenterPoint in terms of both mAP and mAPH in all class.

The performance of the CPSeg trained under the multi-task framework is shown in Table 9. PQ, RQ, and SQ represent the panoptic quality, recognition quality, and segmentation quality, respectively. Also, Th and St superscript define the things and stuff categories, respectively. Compared to the CPSeg that is trained as a standalone model, the performance of panoptic segmentation is reduced in the proposed framework; however, the difference is insignificant.

D. Qualitative Comparison

In this section, we provide additional qualitative samples that reveal the capabilities of the proposed framework. We first demonstrate how panoptic segmentation complements the object detection task in Figure 10, where sample LiDAR scenes are displayed with the combined panoptic segmentation and object detection outputs from the proposed framework. Then, in Figure 11, we indicate the performance gains of the proposed framework by comparing the object detection predictions from the proposed framework with the predictions from the single-stage CenterPoint.

D.1. Evaluation of the Overall Framework Output

In Figure 10, we observe that the panoptic segmentation result from the framework is highly accurate, as pixels representing each foreground object instance share a single, unique color. This implies that the RV encoder feature maps, foreground semantic labels, and center density heatmap injected into the multi-task framework are strongly dependable. Evidently, the corresponding object detection predictions align with the panoptic predictions, which shows that CenterPoint benefits from the multi-view, multi-task feature fusion.

Specifically, we attribute the framework's robust ability to detect small objects, such as the ones in Figure 10, to the incorporation of dense features representation from RV feature maps. In addition, we note that the multi-task framework excels in detecting distant objects with only few points representation, such as the vehicles on the top right corner of Figure 10 (a) and left of Figure 10 (c), as well as the small object on the top right of Figure 10 (b). These objects are represented by very few points and are thus easily lost during the feature extraction process. The proposed framework is able to detect these objects because the class-wise foreground probability map and center density heatmap generated from its accurate panoptic segmentation prediction enable feature values corresponding to foreground objects to

	PQ	RQ	SQ	PQ^{Th}	RQ^{Th}	SQ^{Th}	PQ^{St}	RQ^{St}	SQ^{St}
CPSeg in Multi-Task Framework	70.0	81.1	85.4	73.8	83.7	87.7	62.6	77.0	81.5
CPSeg as Standalone Model	**70.7**	**81.9**	**86.0**	**74.6**	**84.0**	**88.4**	**64.1**	**78.4**	**82.0**

Table 9. Performance evaluation of the CPSeg model in the multi-task framework and as a standalone model on the nuScenes validation set.

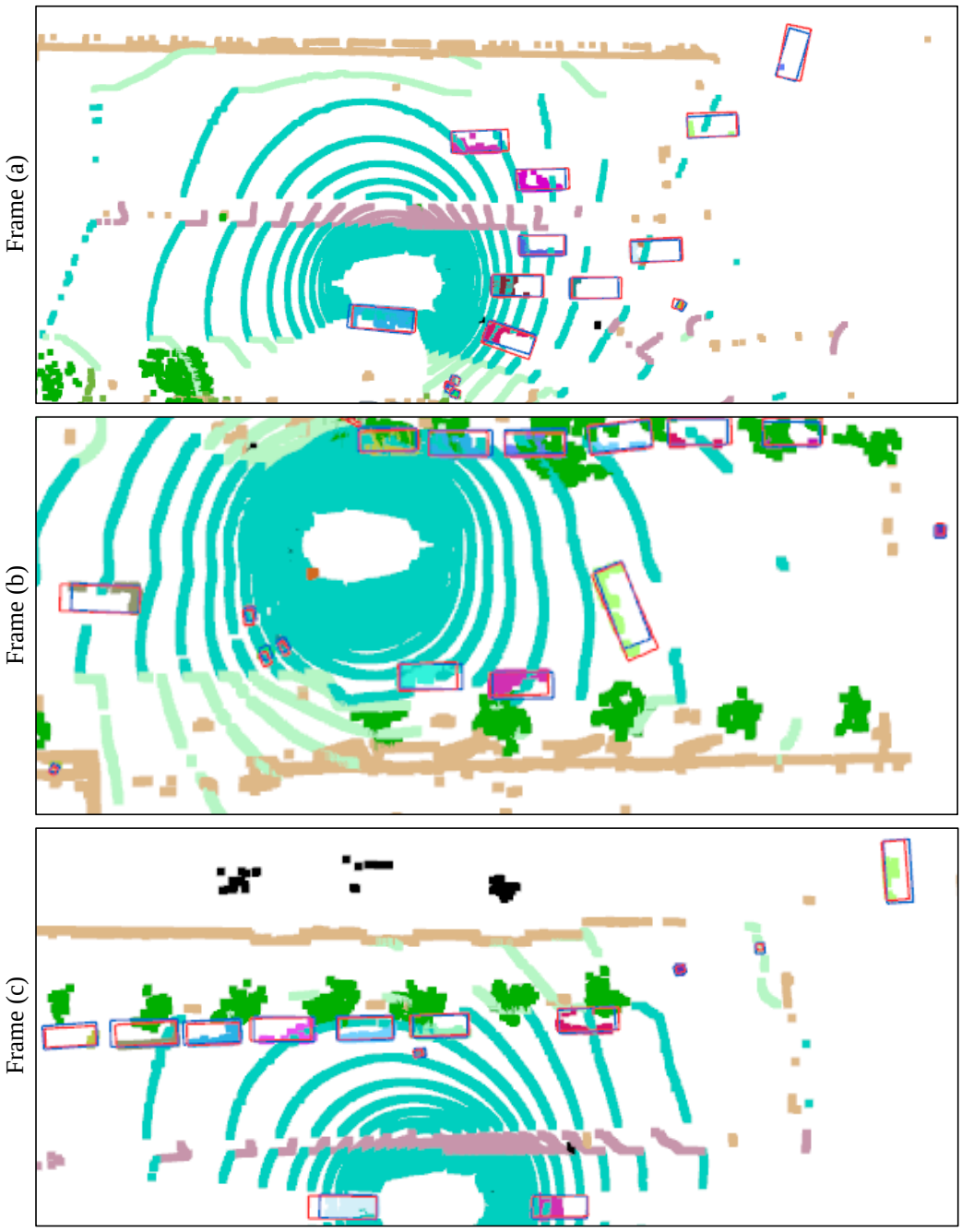

Figure 10. Examples of qualitative results, containing the predicted bounding boxes (in blue), ground truth bounding boxes (in red), and panoptic segmentation results. Best viewed in color.

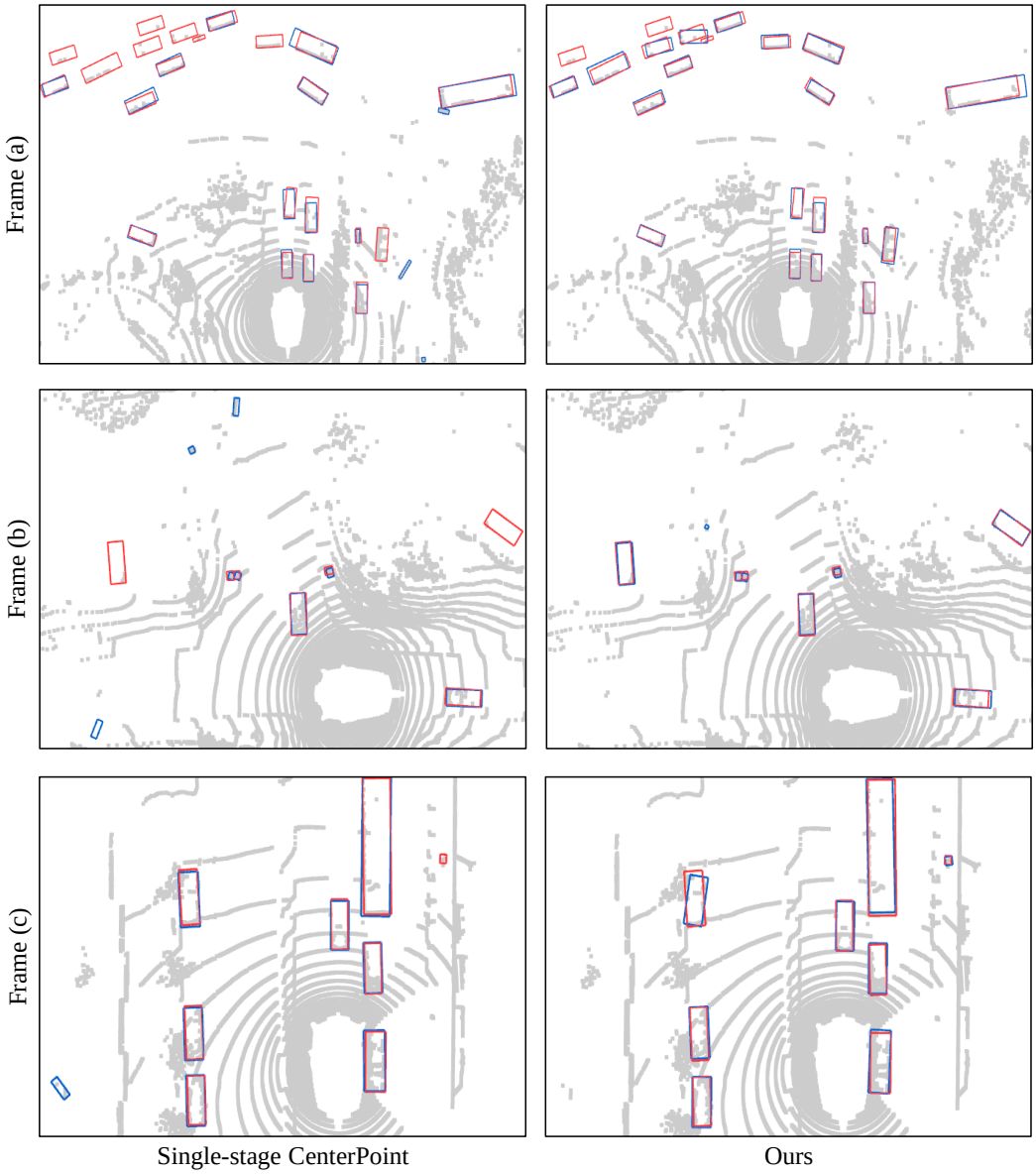

Figure 11. Qualitative comparison of the single-stage CenterPoint (left) and the proposed framework (right) on nuScenes validation set, containing predicted bounding boxes (in blue) and ground truth bounding boxes (in red). Best viewed in color.

be highlighted, so they are easily picked up by the detection head.

D.2. Detection Result Comparison with Baseline

As shown in Figure 11, the comparison of detection performance between the proposed framework and the single-stage CenterPoint further exemplifies the usefulness of the injection of panoptic segmentation information.

In all examples, due to the lack of context-rich features representation, CenterPoint frequently introduces false positive predictions, such as the errors on the bottom left corner and top of Figure 11 (b). In contrast, in the proposed framework, most of the potential false detections are suppressed during the class-wise foreground attention and center density modules. More importantly, it is evident that the Single-stage CenterPoint struggles to detect distant and small objects, such as vehicles on the top left corner of Figure 11 (a), the vehicles on the left and right of Figure 11 (b), and the pedestrian on the top right of Figure 11 (c). In comparison, with the attention-based feature weighting module that combines RV and BEV features, although not perfect, the proposed framework detects more objects that are difficult to be detected if only BEV features representations are considered.