A. Implementation Details

In nuScenes, annotations are created every 0.5s, we follow the common practice [3, 35, 36] to transform the Lidar points of non-annotated frames into their following annotated frames to generate denser point clouds, which improves our model by 2.3% PQ on nuScenes validation. For data augmentation, we use global scaling with a random factor within [0.95, 1.05], global rotation with a random factor within $[-\pi/2, \pi/2]$ and random flipping along both X and Y axes of the LiDAR coordinate on both datasets. As mentioned in our paper, we also use copy-paste data augmentation scheme from [35] on SemanticKITTI to alleviate the distribution imbalance among categories.

Our model is trained with a total batch size of 16 on 8 RTX3090 GPUs. To save computation, we train the model for 40 epochs following [41] for semantic segmentation and then train the instance branch for another 20 epochs. All the submissions and ablation experiments are conducted with the same setting.

For our test-time-augmentation version, we follow [36] to apply flip testing, which improves PQ by 0.3% on nuScenes validation. We ensemble five models with inputs of 3D cylindrical size from [240, 180, 32] to [576, 448, 32], which further improves PQ by 1.2%. Note that we only use our single model without any TTA techniques for all the comparisons in our paper, the TTA version is just for reference considering some of methods incorporate TTA. Our single-model version achieves the 1st place on the public leaderboard of SemanticKITTI.

Regarding the hyper-parameters of the center grouping module, we calculate the average size for different categories with corresponding 3D bounding box annotations on KITTI [12] and nuScenes respectively. For a certain *thing* category with the average size [width, length, height], we assign it a radius $r = \min(\text{width}, \text{length})$.

For supervision signals, voxel-wise losses are adopted for both semantic and instance branches. We obtain voxel-wise semantic labels by majority-voting and use the mean offsets of points as the voxel-wise offset labels. We use centers of axis-aligned bounding boxes as instance centers to train the offsets, which is explained later.

B. Discussion

Pseudo Heatmap vs. Learned Heatmap. We compare our clustering pseudo heatmap (PHM) with the learned heatmap (LHM) adopted in Panoptic-PolarNet [40]. We follow [40] to train a heatmap head (with their post-processing) in our framework. Tab. 6 shows our PHM outperforms LHM on both datasets, especially on nuScenes (+6.8% PQ^{Th}). For LHM, there may be inconsistencies in terms of quantity and location between the predicted centers and the clusters of shifted *thing* points. In crowded scenes

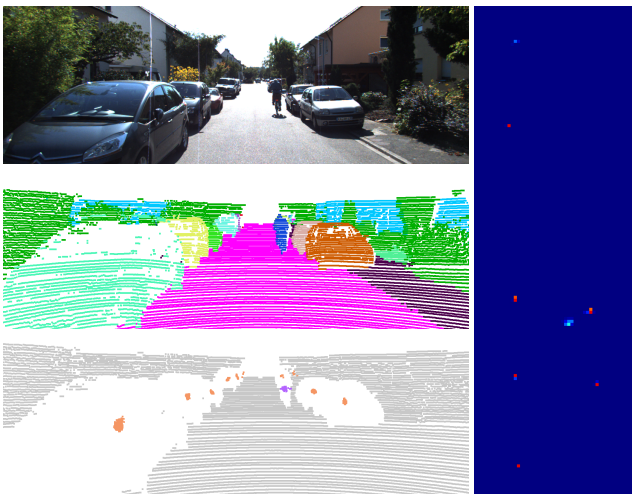

Figure 8. Qualitative example on SemanticKITTI. The left three images are the front view image, panoptic segmentation result and shifted *thing* points. The right image is the corresponding BEV clustering pseudo heatmap.

(commonly seen in nuScenes, while few in semanticKITTI validation), such inconsistency issue has more impacts. Note that the mIoU drops from 77.5 to 67.4 after being refined by LHM centers on nuScenes. It's a strong evidence that LHM is quite inaccurate. On the contrary, our PHM is created from the projection of the shifted *thing* points, where highlights show up certainly as long as there are clustered ones, as illustrated in Fig. 8, so that object-level high recall is achieved. It should be noted that the more accurate the offset regression is, the sparser the clustering pseudo heatmap becomes. As a result, our PHM performs much better on nuScenes.

Dataset	Method	PQ	PQ^{Th}	mIoU	mIoU*
sem.KITTI	LHM	61.1	67.9	65.2	65.1
	PHM (ours)	**61.7**	**69.3 (+1.4)**	65.2	65.1
nuScenes	LHM	69.1	65.7	67.4	77.5
	PHM (ours)	**73.4**	**72.5 (+6.8)**	77.5	77.5

Table 6. Pseodo heatmap vs. learned heatmap. (mIoU*: original semantic results. mIoU: the semantic results refined by instance IDs)

One more thing, a possible weakness of PHM is that there may be multiple center predictions for one object as the shifted *thing* points are not concentrated enough. Fortunately, our proposed center grouping module provides a effective solution.

Choice of Instance Center. There are two types of instance center used in previous researches as the supervision signals of offset regression, i.e., the mass center [40] and the axis-aligned center [14]. In addition, since there are 3D bounding box annotations in nuScenes as external data, the centers of bounding boxes can also be taken as instance centers. We conduct contrast experiments on

both SemanticKITTI and nuScnes datasets for these three choices. As shown in Tab. 7, the difference between the mass center and the axis-aligned center is not obvious on SemanticKITTI. On nuScenes, however, the axis-align center outperforms the mass center by 2.1% PQ, and the annotated center only further improves PQ by 0.1%. It is clear that the annotated center is most beneficial to offset regression due to the highest consistency. Since we do not use external data on nuScenes for comparisons, we adopt axis-aligned center as the final choice. The different results on the two datasets lie in the fact that there are plenty of crowded scenes with more dynamic object instances in nuScenes, where the choice of higher consistent centers performs better.

Dadaset	Mass	Axis-aligned	Annotated
SemanticKITTI	61.6	61.7	-
nuScenes	72.6	74.7	74.8

Table 7. PQ results with different choices of instance center labels on SemanticKITTI and nuScenes validation.

C. Qualitative Results

We show the visualization examples of our Panoptic-PHNet on SemanticKITTI in Fig. 9, as well as on nuScenes in Fig. 10 and Fig. 11. We use the official color map for *stuff* regions and random colors for instance IDs. For nuScenes, we also project panoptic segmentation results onto the front view images. It can be observed that our approach performs well not only for crowded scenes, but also for big objects, which are the focuses of our paper while often ignored in previous studies. Specifically, as shown in the bottom image of Fig. 10, a group of close persons are correctly segmented thanks to our high-quality offset regression and efficient clustering pseudo heatmap.

D. Performance across Classes

We show the detailed class-wise results of our Panoptic-PHNet on SemanticKITTI and nuScenes in Tab. 8, Tab. 9 and Tab. 10.

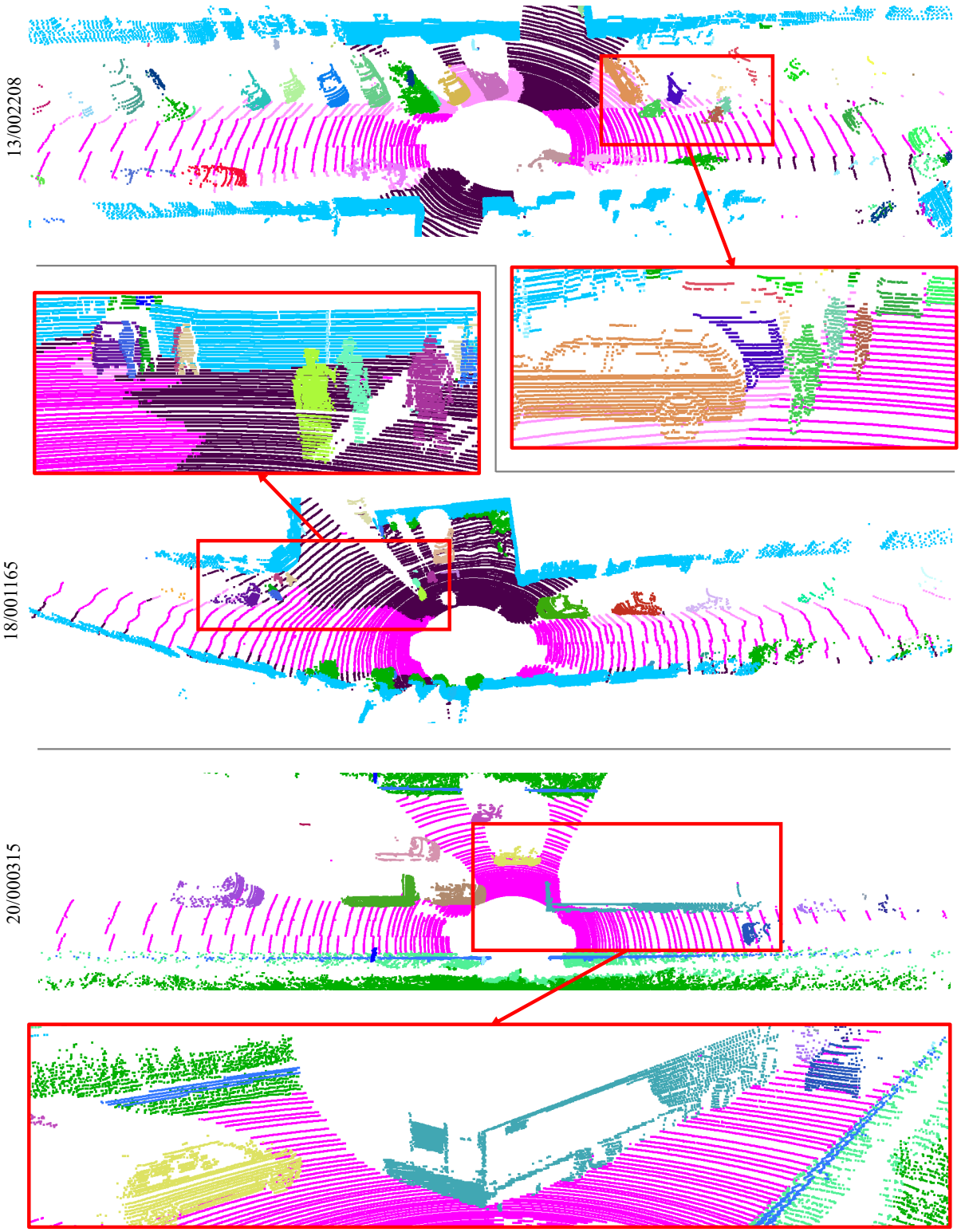

Figure 9. Qualitative examples on SemanticKITTI. The top two examples show the performance of our method in crowded scenes. The bottom example focuses on the big object segmentation.

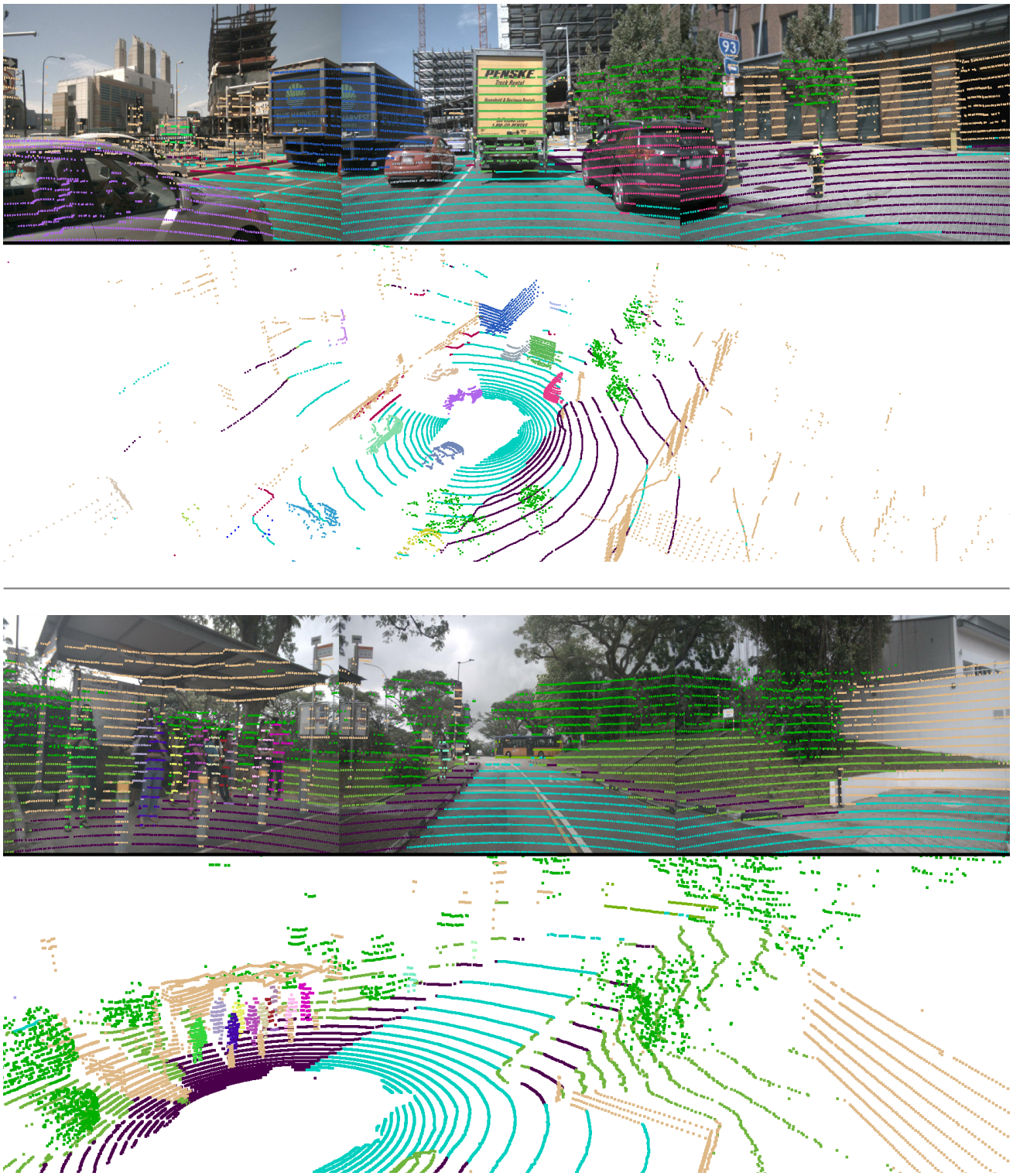

Figure 10. Qualitative examples on nuScenes. The two examples show a driving scene and a group of crowded people respectively.

Figure 11. Qualitative examples on nuScenes, including small objects that are close to each other (row1), big objects (row2), as well as cloudy and rainy day (row3).

Metrics	Car	Truck	Bicycle	Motorcycle	Other Vehicle	Person	Bicyclist	Motorcyclist	Road	Sidewalk	Parking	Other Ground	Building	Vegetation	Trunk	Terrain	Fence	Pole	Traffic Sign	Mean
PQ	94.0	45.1	54.6	62.4	51.2	74.4	76.3	52.0	89.9	70.6	49.4	11.7	87.8	79.4	57.2	45.0	52.6	54.5	61.2	61.5
RQ	98.6	47.5	71.8	69.9	54.9	82.8	83.2	54.6	96.0	85.3	63.2	15.6	93.4	95.0	77.0	59.0	68.6	72.5	80.2	72.1
SQ	95.4	95.0	76.0	89.3	93.3	89.8	91.7	95.2	93.6	82.8	78.1	75.0	94.1	83.6	74.3	76.3	76.6	75.2	76.3	84.8
IoU	96.3	56.4	59.4	55.5	48.0	66.2	70.0	22.9	92.1	77.5	67.9	33.0	92.8	84.9	69.3	69.8	68.5	61.2	62.2	66.0

Table 8. Class-wise LiDAR panoptic segmentation resuts on the **test** set of SemanticKITTI. All scores are in [%].

Metrics	Barrier	Bicycle	Bus	Car	Construction Vehicle	Motorcycle	Pedestrian	Traffic Cone	Trailer	Truck	Driveable Surface	Other Flat	Sidewalk	Terrain	Manmade	Vegetation	Mean
PQ	68.0	77.6	75.4	95.5	75.9	91.1	94.9	94.8	71.7	76.4	97.7	52.0	75.2	59.4	88.8	86.5	80.1
RQ	82.4	86.4	78.9	97.9	82.5	96.1	99.0	99.1	78.8	80.2	100.0	59.2	90.0	75.5	98.6	96.5	87.6
SQ	82.5	89.8	95.7	97.5	92.1	94.7	95.8	95.7	90.9	95.2	97.7	87.8	83.6	78.7	90.1	89.7	91.1
IoU	84.3	35.6	84.9	93.1	70.1	88.0	82.0	81.1	86.6	73.2	97.7	68.4	80.6	76.1	92.2	88.7	80.2

Table 9. Class-wise LiDAR panoptic segmentation resuts on the **test** set of nuScenes. All scores are in [%].

Metrics	Barrier	Bicycle	Bus	Car	Construction Vehicle	Motorcycle	Pedestrian	Traffic Cone	Trailer	Truck	Driveable Surface	Other Flat	Sidewalk	Terrain	Manmade	Vegetation	Mean
PQ	53.5	77.5	75.4	90.8	48.6	87.3	91.0	87.0	56.5	72.6	96.7	58.3	72.4	54.9	88.7	84.8	74.7
RQ	67.7	89.4	80.6	95.5	60.5	95.0	97.4	95.2	65.4	78.6	99.8	67.8	88.0	69.6	99.0	97.0	84.2
SQ	79.1	86.7	93.5	95.0	80.4	91.9	93.5	91.3	86.4	92.3	96.8	85.9	82.3	78.9	89.6	87.4	88.2
IoU	77.9	52.4	93.5	93.0	57.0	88.1	83.9	69.9	69.6	86.3	96.9	75.3	76.3	75.3	90.7	88.7	79.7

Table 10. Class-wise LiDAR panoptic segmentation resuts on nuScenes validation. All scores are in [%].