Localization Is All You Evaluate: Data Leakage in Online Mapping Datasets and How to Fix It

Supplementary Material

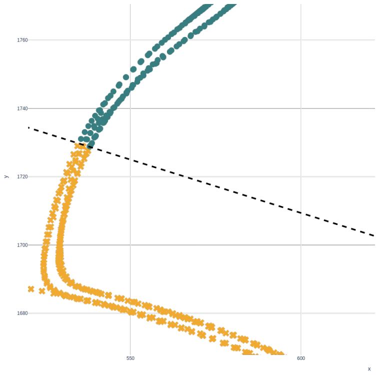


Figure 5. In this example from Singapore Queenstown, the individual samples from a few sequences are divided into training (green) and validation (orange) according to the cut-off border. Some samples from a sequence are put in the training set, whereas the remaining are put in the validation set. The samples close to the dotted black cut-off line are the remaining possible data-leakage samples when using our proposed Near Extrapolation split.

6. Partially Overlapping Maps

Splitting nuScenes for Near Extrapolation on a sequence level requires grouping large areas with similar zone classes together and putting them in a single set, as seen in [24]. This is due to the entangled nature of the sequences where many partially overlap. Instead, we assign each sample individually to a set when a sequence straddles the boundary between two sets (*e.g*. train and val in Fig. 5). We divide the sequence at the boundary, creating two separate partial sequences each with preserved temporal consistency. This maintains the usefulness for object detection and keeps the possibility of using the data for temporal fusion, where having consecutive samples is important. We have kept the number of sequences being cut into multiple parts as low as possible, making the cuts, when necessary, across the road's driving direction.

The sequences in the Argoverse 2 dataset are more spread out compared to nuScenes, and a balanced sequence-wise split is possible to obtain. There is thus no impact on usability for object detection, object tracking, and other temporal fusion applications for the Argoverse 2 split.

Splitting the data geographically ensures that there is no overlap in poses between the different sets. However, as online mapping methods typically predict 30m in front and to the rear there will still be some overlap in the ground truth maps among the samples close to the cut-off border.

Split	HDMapNet		VectorMapNet		MapTR		MapTRv2	
	Val	Test	Val	Test	Val	Test	Val	Test
Near	17.1	21.2	14.0	18.2	19.0	19.7	26.7	26.2
Near∉	17.0	21.4	14.5	18.3	19.0	19.6	26.5	25.8

Table 10. Evaluating the predictions from validation and test sets in the nuScenes' Near Extrapolation split, where samples closer than 60m to a training sample have been removed (indicated by ∉). It can be seen that the impact on performance is negligible. Metrics are IoU for HDMapNet, and mAP for MapTRv2 and VectorMapNet.

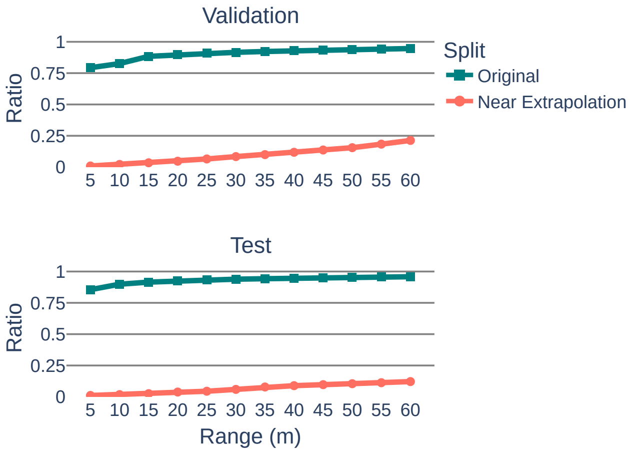


Figure 6. Ratios of validation and test samples within a certain range of training samples for nuScenes. The Geographically disjoint Near Extrapolation split has negligible overlap compared to the, greatly overlapping, Original split.

To see the effects of the remaining overlap in the geographical split of nuScenes we run experiments where the validation and test samples closer than 60 m to a training sample have been filtered out. Tab. 10 demonstrates that these samples have a negligible impact on performance. Furthermore, Fig. 6 shows how the ratio of validation and test samples that are close to a training sample changes with range. For completeness Fig. 7 displays the same information on Argoverse 2.

7. Additional Data Attributes

In this section, we further display the splitting, the number of samples in discretized maps, and different zone classes (*e.g*. residential, commercial, and industrial).

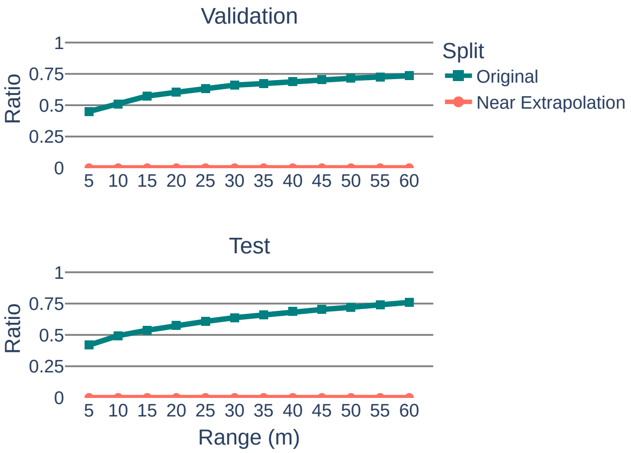


Figure 7. Ratios of validation and test samples within a certain range of training samples for Argoverse 2. The Geographically disjoint Near Extrapolation split has no overlap compared to the, greatly overlapping, Original split.

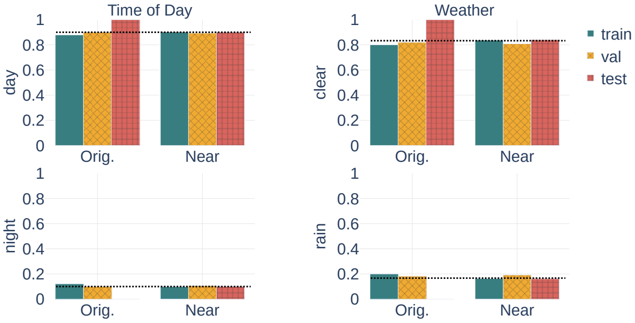


Figure 8. Ratios of weather conditions (clear and rain) as well as time of day (day and night) on the nuScenes dataset. The black dashed lines are the respective ratios over the full dataset.

nuScenes Fig. 8 details that the Near Extrapolation split is balanced across all attributes. This allows for conducting experiments and drawing conclusions on a well-defined dataset. Further, Fig. 10 depicts example images and their position on the map for Boston Seaport. The industrial zones in the south and south-eastern areas have different attributes, *e.g.* type of buildings, lane widths, number of lanes, and frequency of pedestrian crossings, than the commercial and residential zones in the north-western part. It is thus important that these zones are represented in all sets for a fair evaluation of trained methods. Fig. 11 showcases the regions where samples are allocated in each set for all cities. Each set incorporates regions from different parts of the cities to promote diversity. The heatmaps in Fig. 12 depict the distribution of samples within each 60m cell. One can, for instance, observe a concentration of samples in crossings.

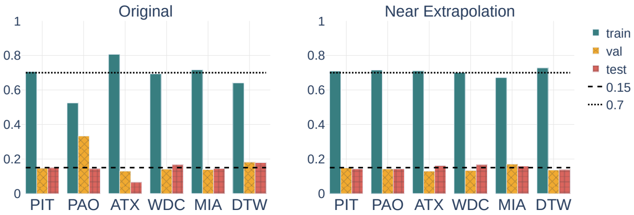


Figure 9. Inter-set city distribution for the Original and Geographically split data of Argoverse 2. The dotted and dashed lines represent the 70% and 15% target ratios respectively.

Argoverse 2 As discussed in Sec. 3.1 it is possible to split Argoverse 2 on a sequence level while preserving zone class diversity. Fig. 9 illustrates the distribution of the number of samples in each city. Fig. 13 further highlights the significance of diverse city areas in all sets by presenting a collection of images from various locations in Washington DC. The downtown area in the southwest exhibits different road characteristics from the sub-urban areas in the north and east. Fig. 14 illustrates how the complete set of city maps are split, ensuring a diverse selection of areas in each set. Heatmaps in Fig. 15 represent the distribution of samples within each 60m cell.

8. Extended Experiments

To further investigate the effects of data splits and the amount of data, we perform additional experiments.

Far Extrapolation We also train and evaluate the segmentation-based methods with city-wise folds. Tab. 11 reports the performance for the folds in the Far Extrapolation split and their cross-validation mean. The performance on these folds are, similarly to the Near Extrapolation splits, lower than on the Original splits. For nuScenes the performance on Near Extrapolation splits is already low, and the performance on the Far Extrapolation folds are on par. For Argoverse 2 the city-wise folds are performing worse than the Near Extrapolation split's validation set, but similar to the test set. This results in the cross-validation mean being consistently lower than the mean of the validation and test performance of the Near Extrapolation splits.

Training set extension To explore how the amount of data affects the performance we extend the training set with the validation samples, effectively increasing the training set with 20%. Tab. 12 shows a boost in the test performance for both segmentation- and vector-based methods. The impact is greater on nuScenes, but Argoverse 2 also benefits from the added data, indicating that more extensive datasets are necessary for learning online mapping. For instance, the extra data has a higher impact for MapTR on Argoverse 2,

	Model	Split	Divider	Boundary	Crossing	Mean	CV
nuScenes	GKT	A	10.6	14.2	0.8	8.5	9.9
		B	14.7	17.2	1.6	11.2	
	CVT	A	13.1	14.1	2.2	9.8	10.7
		B	14.8	17.5	2.6	11.6	
	IPM	A	28.1	34.0	12.1	24.7	26.6
		B	33.6	38.8	13.0	28.5	
	HDMapNet	A	20.1	20.7	7.2	16.0	27.3
		B	24.2	24.4	6.9	18.5	
Argoverse 2	GKT	A	28.2	21.8	7.1	19.0	20.1
		B	30.8	25.8	6.8	21.1	
		C	29.5	23.7	6.9	20.0	
	CVT	A	29.0	21.9	9.5	20.1	20.8
		B	31.9	23.0	7.6	20.8	
		C	30.6	24.0	9.3	21.3	
	IPM	A	43.0	38.2	24.6	35.3	37.4
		B	48.6	43.3	25.8	39.2	
		C	45.1	41.6	26.8	37.8	

Table 11. Segmentation-based methods' IoU on the city-wise folds of the Far Extrapolation split and their corresponding cross-validation mean (CV).

	Model	Split	Divider	Boundary	Crossing	Mean
nuScenes	HDMapNet	Near	15.3	17.3	9.0	13.9
		Near∪	24.4	26.3	14.7	21.8
	VectorMapNet	Near	17.3	21.6	15.7	18.2
		Near∪	18.8	25.3	17.6	20.6
	MapTR	Near	19.9	33.3	5.9	19.7
		Near∪	21.9	36.1	6.8	21.6
	MapTRv2	Near	23.4	40.5	14.8	26.2
		Near∪	25.3	42.1	18.6	28.7
Argoverse 2	VectorMapNet	Near	35.0	32.4	31.3	32.9
		Near∪	37.5	33.1	32.7	34.4
	MapTR	Near	45.2	48.3	50.9	48.2
		Near∪	47.3	49.4	52.0	49.6
	MapTRv2 2D	Near	56.6	53.5	55.6	55.2
		Near∪	59.5	54.7	53.4	55.9

Table 12. Increasing training data by 15% using the union of training and validation samples, marked by ∪, improves test performance. IoU for HDMapNet and mAP for the other methods.

$+1.4$ mAP, than the choice of lifting method, $+0.7$ mAP. On nuScenes, the impact is greater, but also similarly large as using LSS in comparison to GKT for lifting, $+1.9$ and $+2.0$ respectively. The lifting methods are further discussed in Sec. 4.4 and shown in Tab. 8.

Hyperparameter-search For MapTRv2 on the Near Extrapolation split on nuScenes, we investigate various hyperparameters related to overfitting on the training set, *i.e.*, weight decay, learning rate, and training epochs. Interestingly, we can in Tab. 13 observe only minor differences and the parameters initially employed for training on the Original split seem equally effective for the geographically disjoint split.

9. Qualitative Results

nuScenes Fig. 16 portrays three examples with input images, the evaluation prediction, its ground truth, and the closest training sample. Despite not being captured from the exact same pose, these instances demonstrate striking similarities between the evaluation and closest training pose. This underscores that the method, having encountered the closest training sample during training, can achieve accurate predictions through memorization and retrieval of these examples at test time. Additional examples can be seen in the videos to be part of the project webpage.

		LR	
		$6e^{-4}$	$1e^{-4}$
WD	0.05	27.0	26.5
	0.10	26.7	27.2
	0.15	27.1	26.5

Table 13. MapTRv2 show robustness to different hyperparameters, learning rate (LR) and weight decay (WD) on nuScenes Near Extrapolation split.

Fig. 17 compares the predictions of a sample included in the test set of both the Original and Near Extrapolation splits. It demonstrates that a model trained on the Original split can predict dividers, boundaries, and pedestrian crossings occluded by vehicles in the opposing lane accurately. Thus making it tempting to speculate that the method has memorized this information. Furthermore, it shows that the model trained on geographically disjoint data only identifies the dividers near the ego vehicle. These dividers are visible in the images, but absent in the ground truth, indicating that the model has learned to generalize better.

Argoverse 2 In Fig. 18, we present three examples featuring input images, the evaluation prediction, its ground truth, and the closest training sample. While not being from the exact same pose, *e.g.* in the top example the closest training pose is slightly rotated, and in the bottom from an adjacent lane, it is still plausible for a method to achieve a high score on the test sample by memorizing the map and images from the training sample, and then recall and slightly shift and rotate that map at test time. Additional examples will be available on the project webpage.

Fig. 19 illustrate comparisons between predictions derived from a sample included in both the Original and Geographically disjoint splits' test set, along with the ground truth. Despite the inherent difficulty in predicting objects situated behind a truck on the left side, the model trained on the original split demonstrates commendable accuracy in its estimations. The model also effectively predicts the lane divider to the right of the ego vehicle, when not visibly present in the image but existing in the ground truth. It is worth noting that this may not be due to memorization, as the model could learn, e.g., consistent data annotations and hints from road dividers and road width to accurately predict this non-visible lane divider.

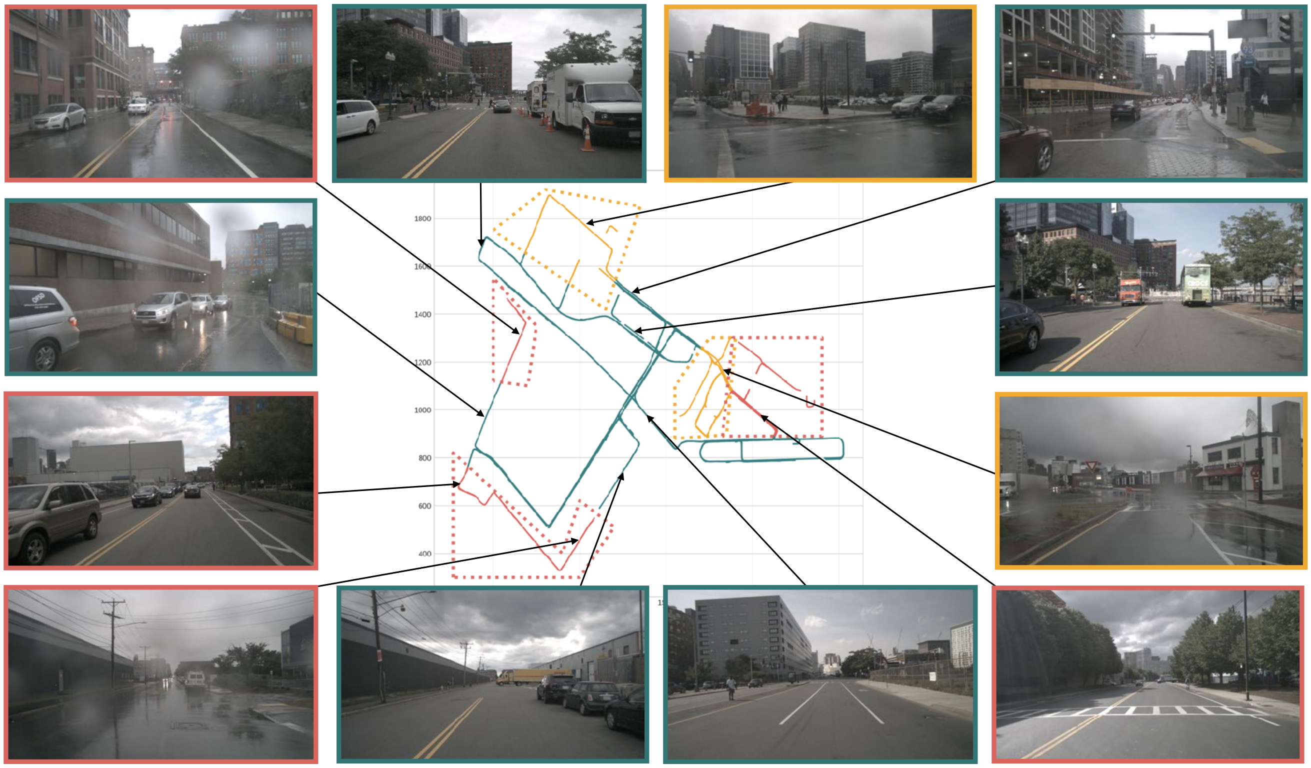
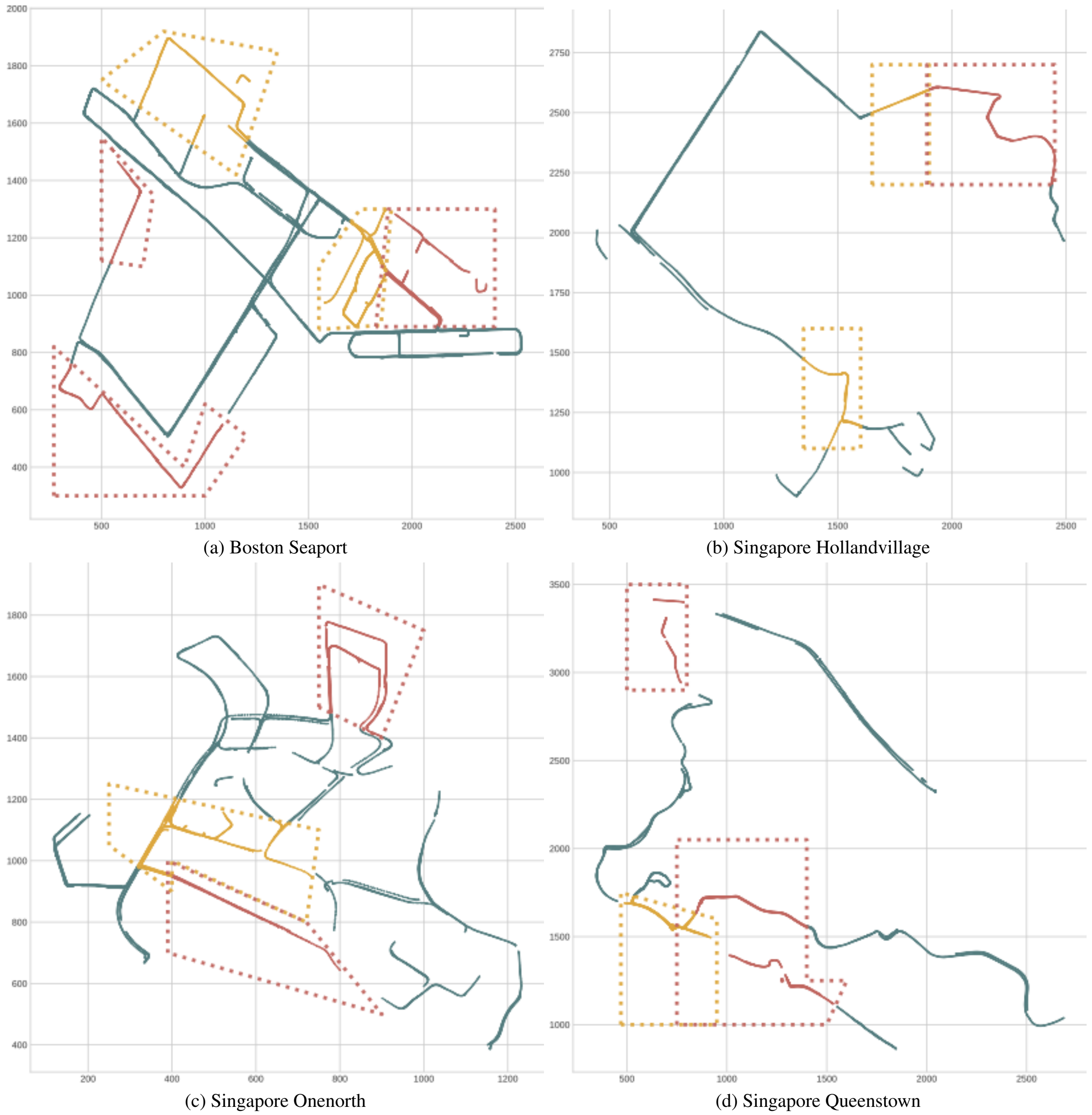


Figure 10. Selected poses from the Boston Seaport map in nuScenes dataset, with marked training (green), validation (blue), and test (red) poses according to the Near Extrapolation split. Dotted polygons mark the boundaries of the validation and test zones. To ensure diversity in zone types within each set, regions from various parts of the city are included. The industrial zones in the south and south-eastern areas have different attributes than the commercial and residential zones in the north-western part.



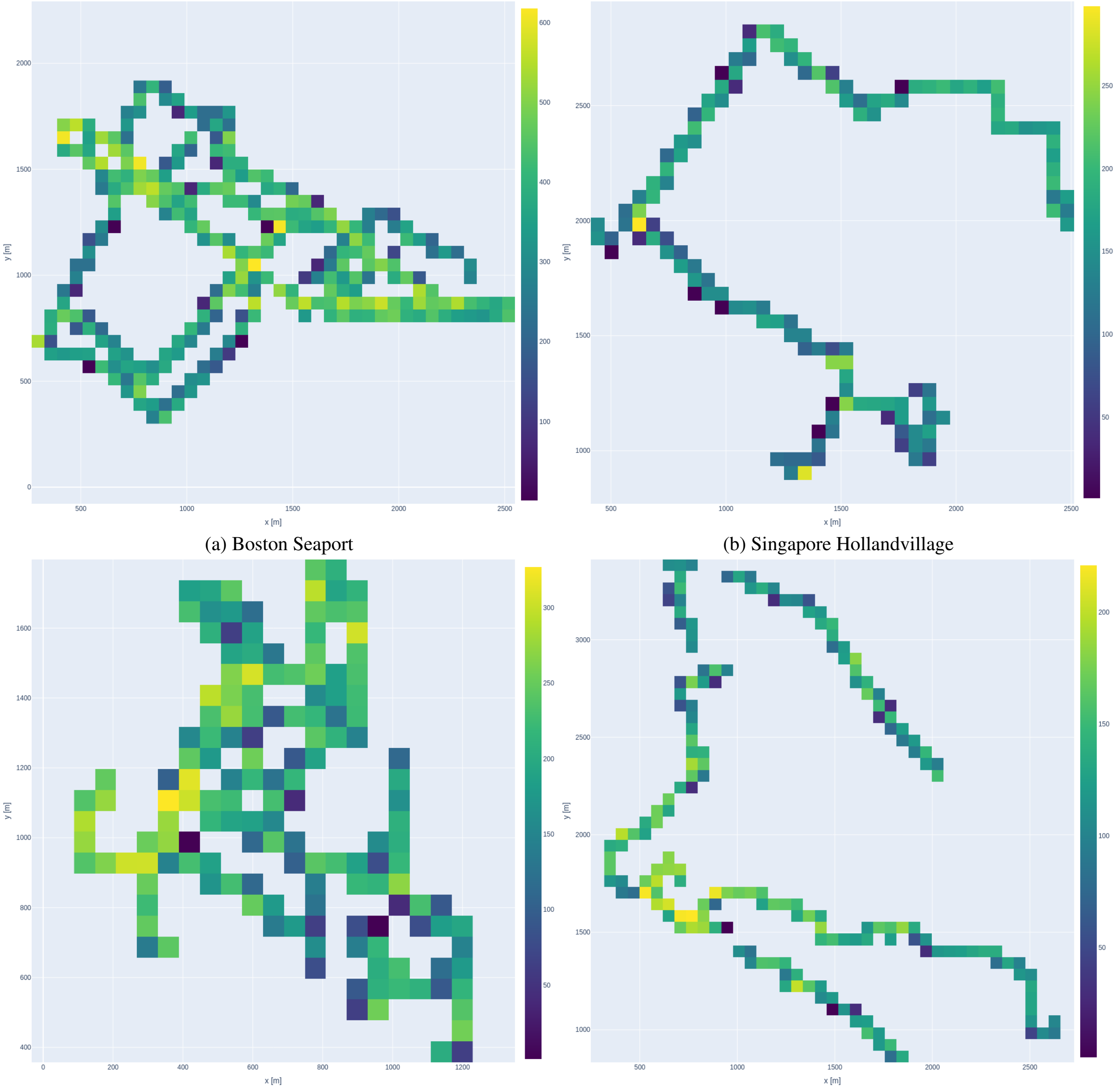
(a) Boston Seaport

(b) Singapore Hollandvillage

(c) Singapore Onenorth

(d) Singapore Queenstown

Figure 11. Positions of samples in the nuScenes dataset, with the geographical areas of the Near Extrapolation split outlined by dotted polygons. Training, validation, and test sets are distinguished by green, orange, and red colors, respectively. Areas from various parts of the cities are present in each set.



(a) Boston Seaport

(b) Singapore Hollandvillage

(c) Singapore Onenorth

(d) Singapore Queenstown

Figure 12. Heatmaps depicting the distribution of samples within 60m cells in the nuScenes dataset, revealing a high amount of samples in many cells, especially concentrated within crossings.

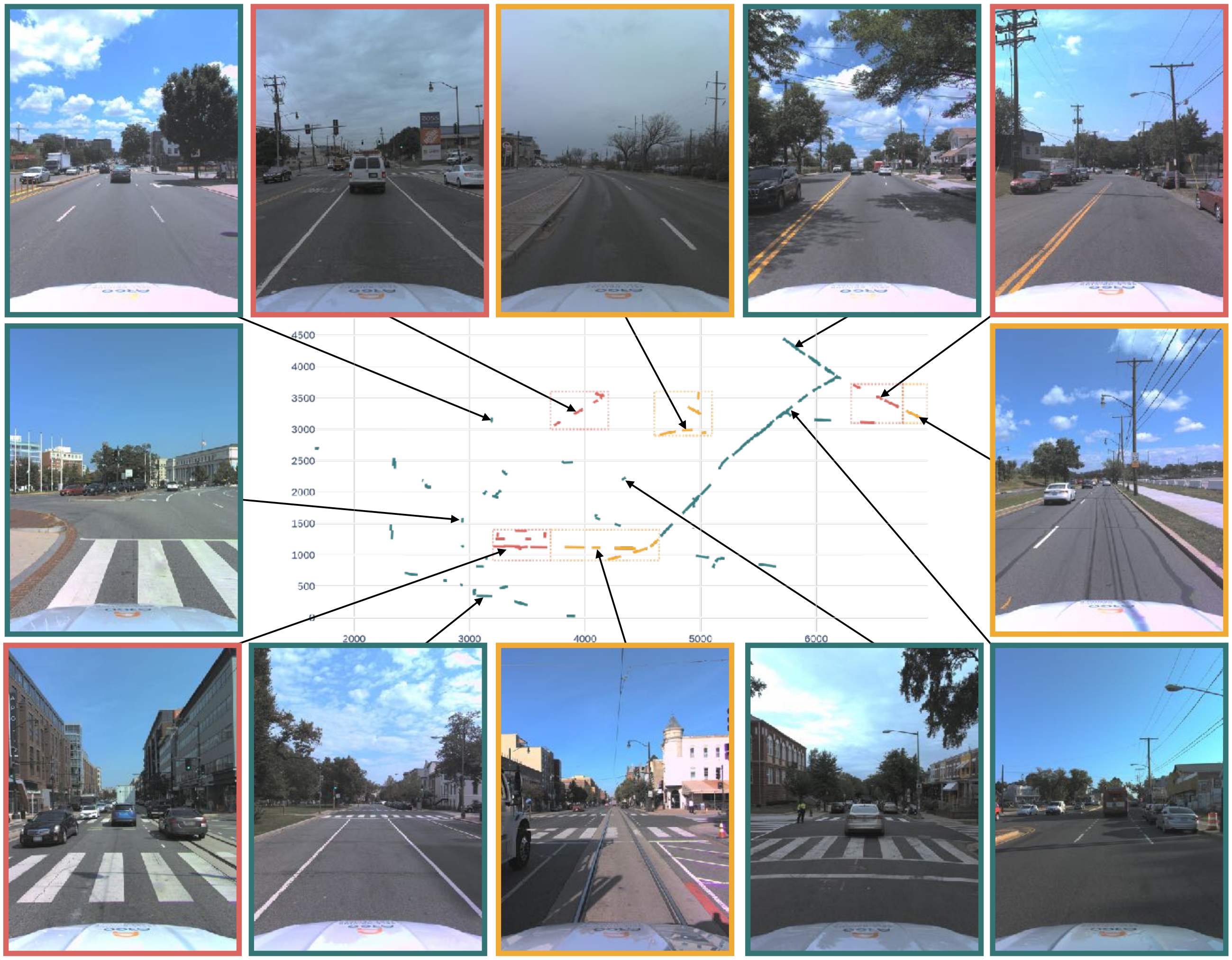
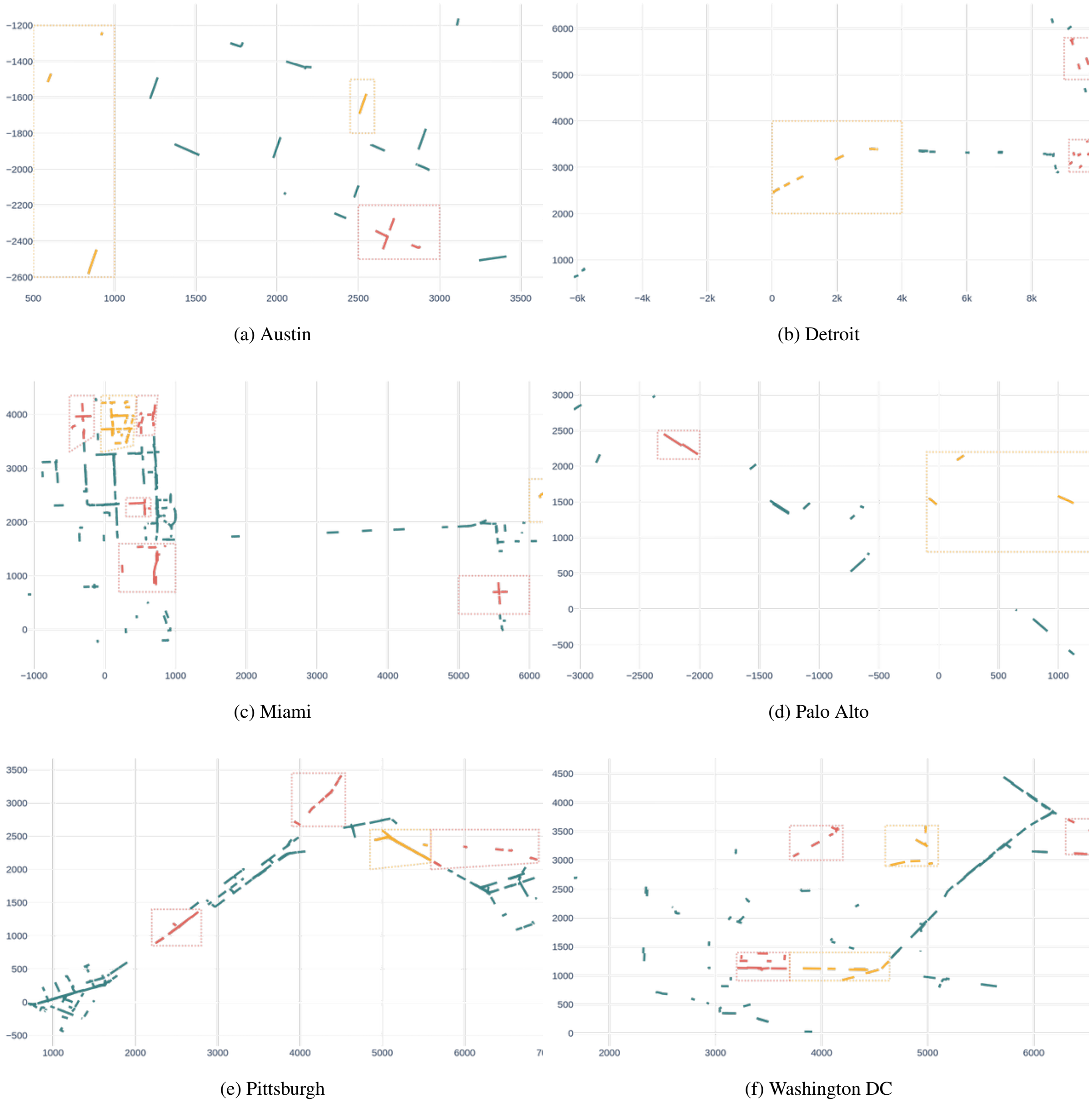


Figure 13. Samples from the Washington DC map in Argoverse 2 dataset, with marked training (green), validation (blue), and test (red) pose according to the Near Extrapolation split. Dotted polygons mark the boundaries of the validation and test zones. To enhance diversity in zone types within each set, regions from different parts of the city are incorporated. The downtown area in the southwest has different road characteristics from the sub-urban areas in the north and eastern parts.



(a) Austin

(b) Detroit

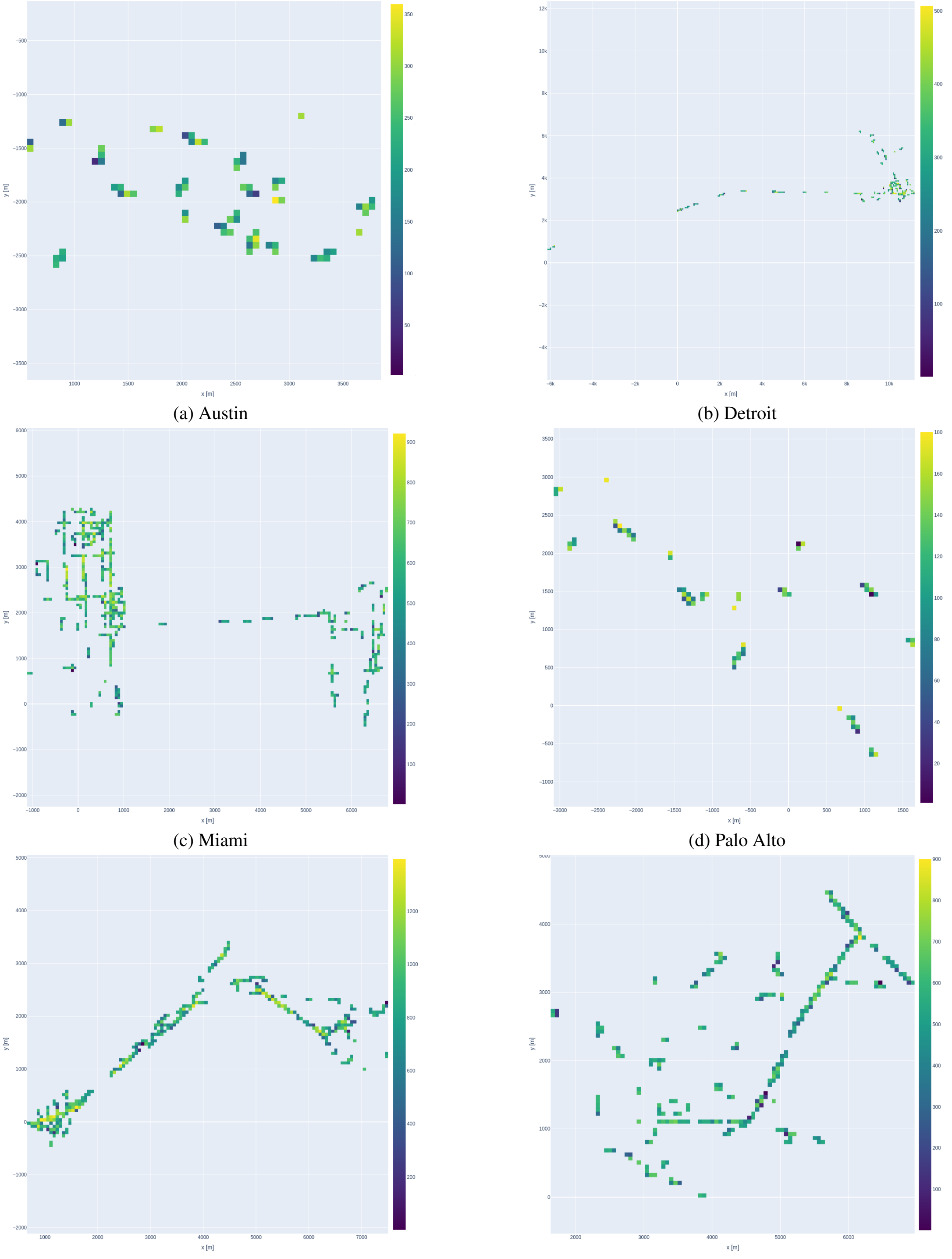
(c) Miami

(d) Palo Alto

(e) Pittsburgh

(f) Washington DC

Figure 14. Near Extrapolation. Positions of samples in the nuScenes dataset, with the geographical areas of the validation and test sets outlined by dotted polygons. Training, validation, and test sets are distinguished by green, orange, and red colors, respectively. Regions from different parts of the cities are present in each set.



(a) Austin

(b) Detroit

(c) Miami

(d) Palo Alto

(e) Pittsburgh

(f) Washington DC

Figure 15. Heatmaps for the number of samples within 60m cells for Argoverse 2 dataset. Many cells contain a lot of samples, with the maximum number of samples in a single cell being 1398.

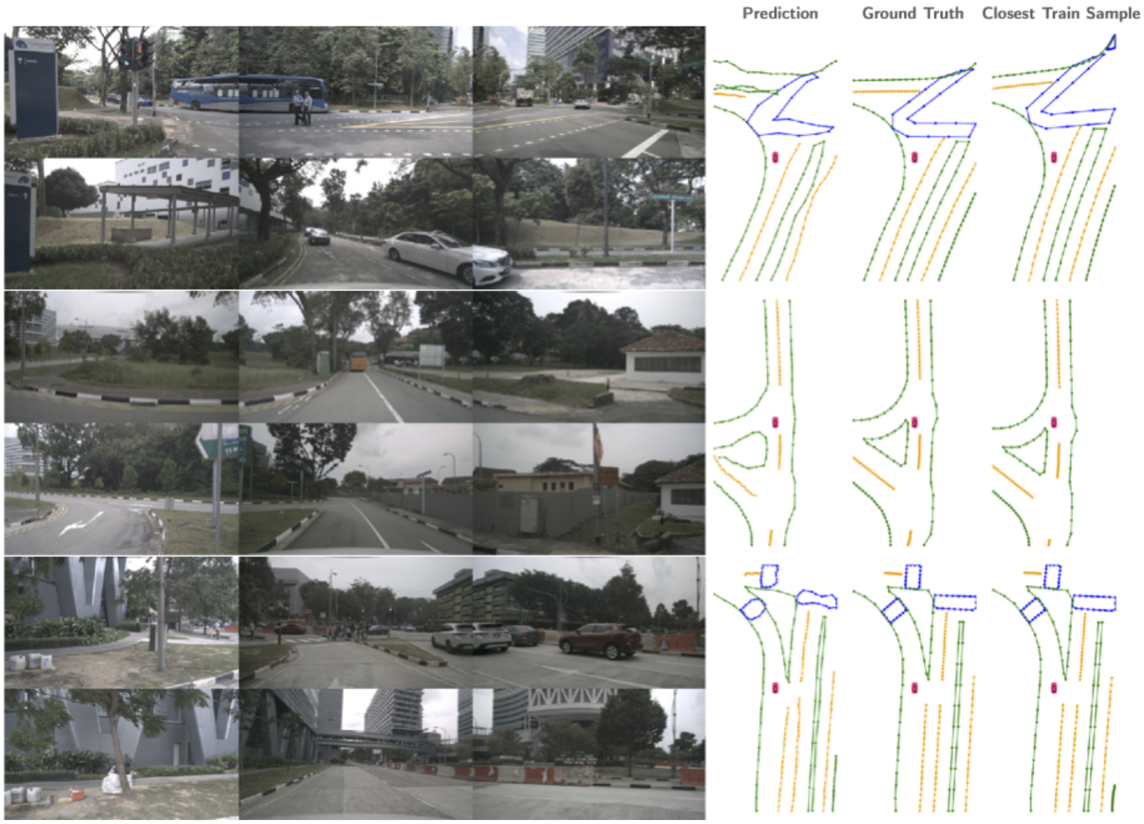


Figure 16. Multiple examples of validation or test prediction from MapTR on nuScenes, corresponding ground truth, and the closest training sample's ground truth. The close similarities between the closest training samples and the evaluation samples are evident in each example.

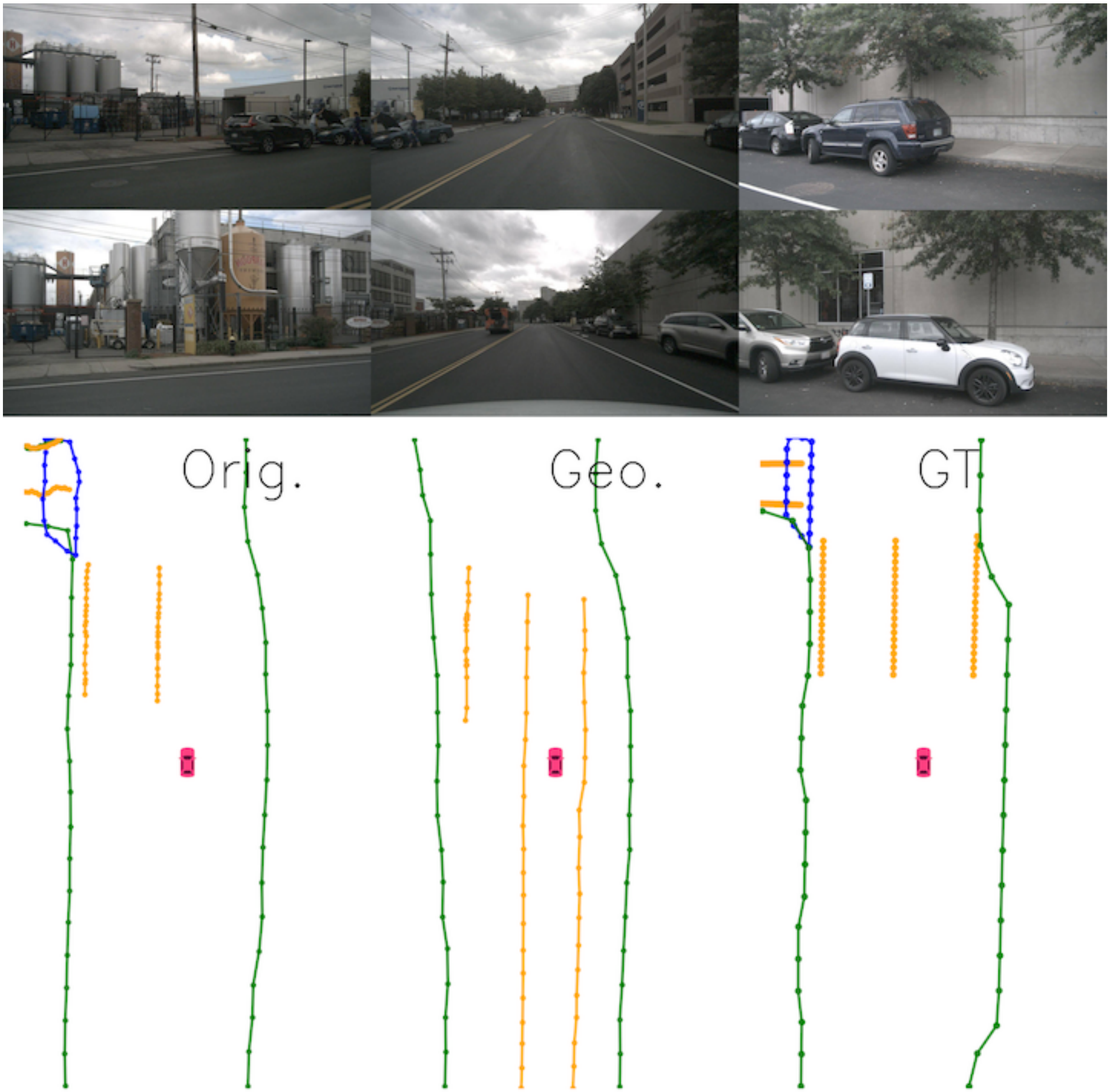


Figure 17. nuScenes test prediction from MapTR trained on Original (Orig.) and Geographically disjoint (Geo.), here Near Extrapolation, along with the ground truth (GT). Dividers, Boundaries, and Pedestrian crossings are visualized in orange, green, and blue respectively. Despite occlusion on the left side by opposing lane vehicles, the method trained on the original split accurately predicts them. In contrast, the model trained on geographically disjoint splits fails to detect them. On the other hand, the model trained on geographically disjoint split data successfully identifies dividers near the ego vehicle, even though they are absent in the ground truth.

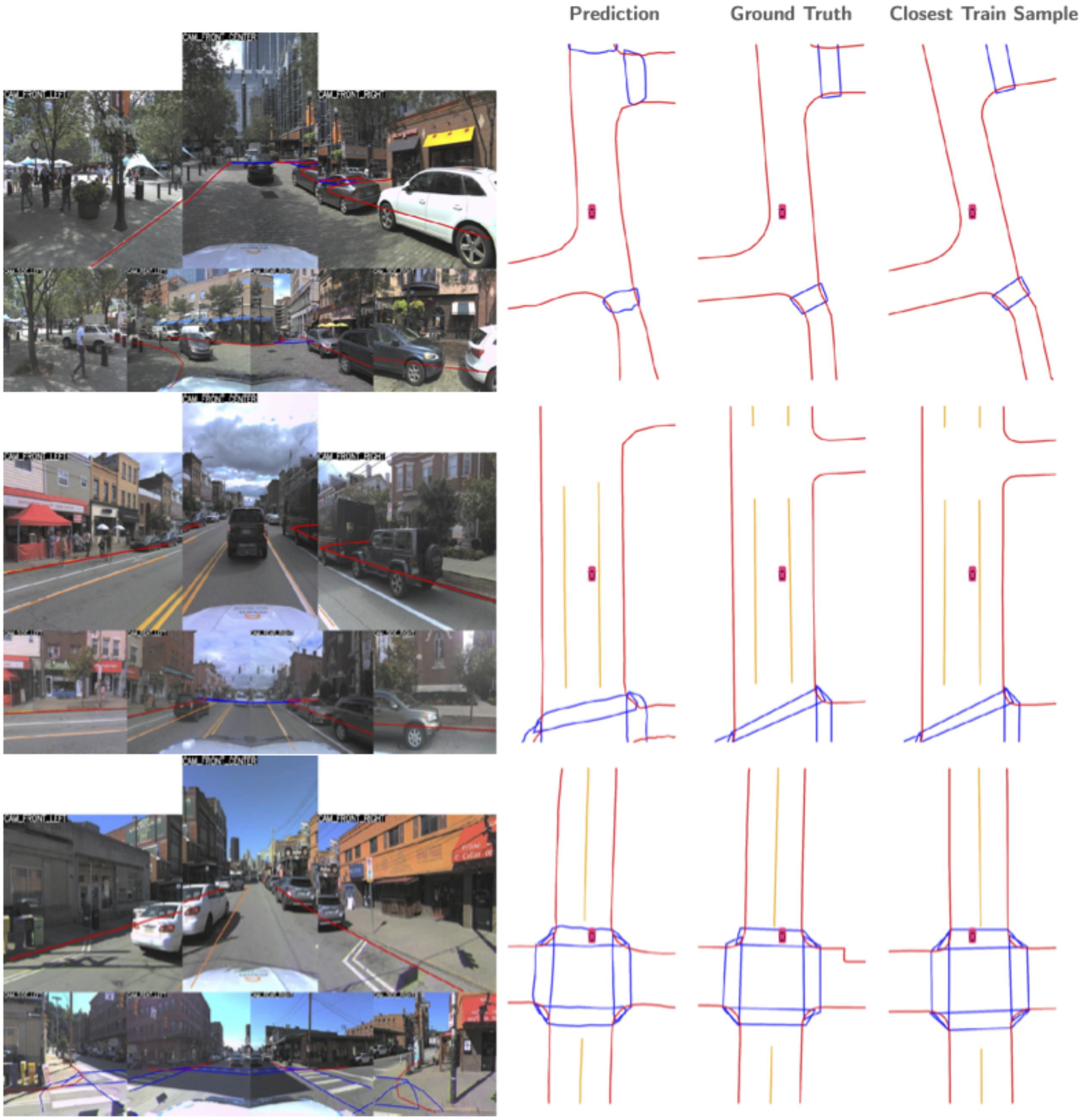


Figure 18. Multiple instances of validation or test predictions from MapTRv2 on Argoverse 2, alongside corresponding ground truth and the ground truth of the nearest training sample. The close similarities between the nearest training samples and the evaluation samples are apparent in each example. In the top illustration, the closest training sample exhibits a slight rotation, but the positions are very similar. In the bottom example, the closest training sample is from the lane adjacent to the evaluation sample

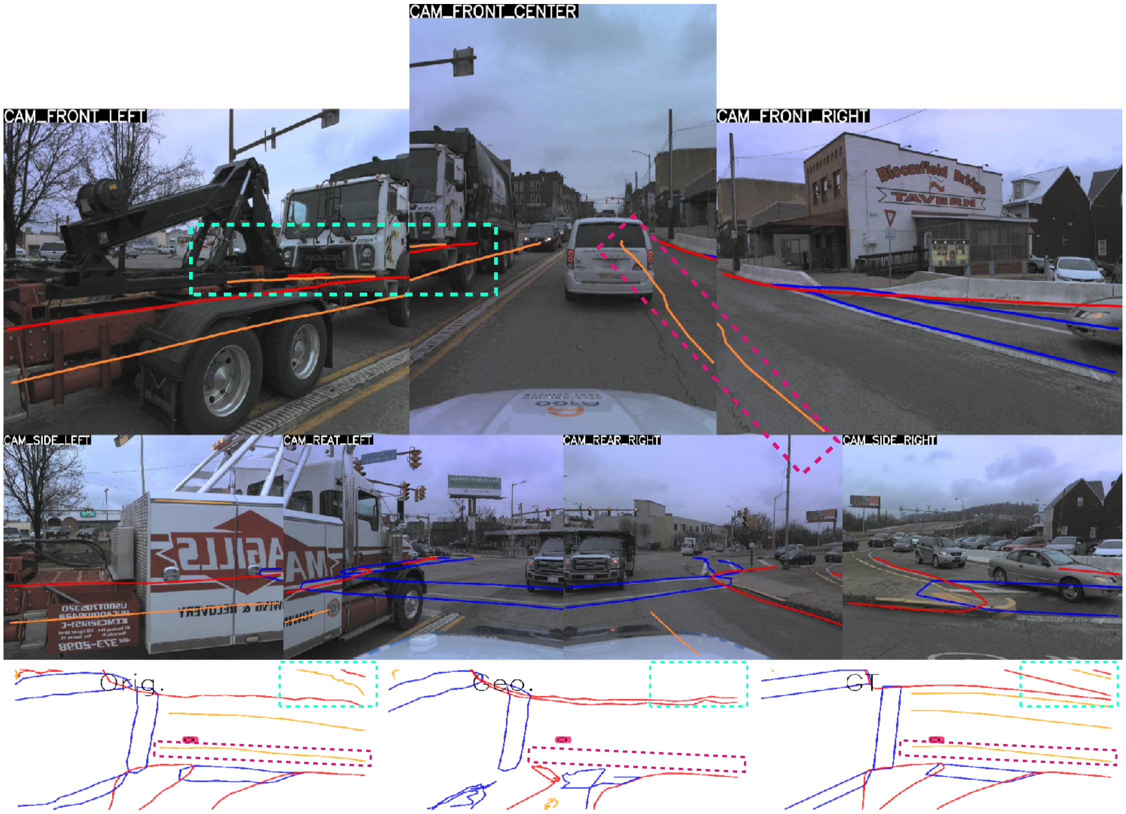


Figure 19. Argoverse 2 test prediction by mapTRv2 trained on Original (Orig.) and Geographical (Geo.), here Near Extrapolation, splits along with the ground truth (GT). Dividers, Boundaries, and Pedestrian crossings are visualized in orange, red, and blue, respectively. The predictions in the image view are from training on the Original split. Here, the predictions behind the truck on the left side, most notably the divider and boundary highlighted with the teal box, ought to be difficult to predict. Additionally, the model effectively predicts the lane divider to the right of the ego vehicle, highlighted by the pink box, even though there is no visible lane divider present in the image. It is worth noting that this may not solely be due to memorization, as the model could learn, *e.g.*, consistent data annotations and hints from road dividers and road width to accurately predict this non-visible lane divider.