Supplementary Material
4D-Former: Multimodal 4D Panoptic Segmentation

Anonymous Author(s)
Affiliation
Address
email

In this supplementary material, we first provide an illustration of the proposed Tracklet Association Module (in Sec. 1) and then present detailed class-wise results on the benchmarks (in Sec. 2). Finally, we show additional qualitative results (in Sec. 3).

1 Tracklet Association Module

We provide an illustration of the proposed Tracklet Association Module (TAM) in Fig.1. The input to our TAM is constructed by concatenating the following attributes of the input tracklet pair along the feature dimension: (1) their (x, y, z) mask centroid coordinates, (2) their respective tracklet queries, (3) the frame gap between them, and (4) their mask IoU. The frame gap and mask centroid coordinates are expanded to 64-D each by applying sine/cosine activations with various frequencies. The concatenated set of features is input to a 4-layer MLP which produces a scalar association score for the input tracklet pair.

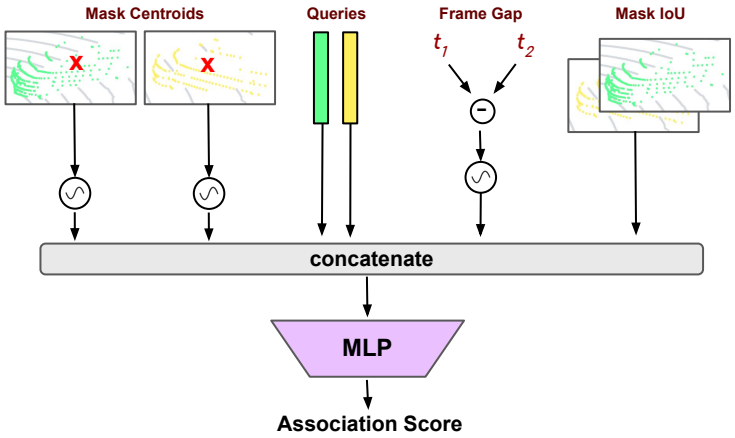

Figure 1: Illustration of the Tracklet Association Module (TAM).

2 Detailed Quantitative Results

We present the detailed per-class results for: nuScenes val set (Tab. 1), nuScenes test set (Tab. 2), and SemanticKITTI val set (Tab.3).

Submitted to the 7th Conference on Robot Learning (CoRL 2023). Do not distribute.

Metric	mean	Barrier	Bicycle	Bus	Car	Construction	Motorcycle	Pedestrain	Traffic Cone	Trailer	Truck	Drivable	Other Flat	Sidewalk	Terrain	Manmade	Vegetation
PTQ	75.17	64.11	74.33	79.05	90.89	64.64	81.87	88.03	83.04	58.67	76.92	95.61	51.92	68.92	54.61	82.46	87.59
sPTQ	75.50	65.25	75.33	79.39	91.16	64.94	82.43	88.47	83.65	59.14	77.26	95.61	51.92	68.92	54.61	82.46	87.59
IoU	78.86	82.74	52.69	90.41	94.31	54.95	88.96	82.66	68.98	65.41	82.57	96.32	71.33	73.36	75.54	91.80	89.75
PQ	77.34	68.59	79.51	80.98	93.51	67.63	86.77	91.71	87.74	61.00	78.91	95.61	51.92	68.92	54.61	82.46	87.59
SQ	89.02	82.53	87.83	93.95	95.73	88.56	91.36	93.59	90.53	86.60	93.66	96.13	84.50	79.75	78.79	91.11	89.64
RQ	86.46	83.11	90.52	86.19	97.68	76.37	94.98	98.00	96.92	70.44	84.26	99.45	61.45	86.42	69.30	90.51	97.72

Table 1: Class-wise results on nuScenes val set. Metrics are provided in [%]

Metric	mean	Barrier	Bicycle	Bus	Car	Construction	Motorcycle	Pedestrain	Traffic Cone	Trailer	Truck	Drivable	Other Flat	Sidewalk	Terrain	Manmade	Vegetation
PTQ	75.47	63.20	73.20	75.21	90.14	62.44	81.01	89.11	84.95	65.46	75.13	97.10	46.13	71.44	58.00	85.16	89.85
sPTQ	75.90	64.63	73.98	75.42	90.45	63.73	81.92	89.57	85.48	66.14	75.43	97.10	46.13	71.44	58.00	85.16	89.85
IoU	80.42	86.66	48.99	92.24	91.72	68.22	79.79	79.84	77.24	85.54	73.81	97.41	66.51	78.50	76.62	93.04	90.62
PQ	77.99	68.63	78.30	77.48	93.01	69.07	86.69	92.64	89.13	68.17	77.05	97.10	46.13	71.44	58.00	85.16	89.85
SQ	89.66	81.69	89.13	94.74	95.80	87.12	92.62	93.94	91.63	88.30	94.29	97.36	85.46	81.85	78.04	91.08	91.51
RQ	86.59	84.01	87.85	81.78	97.09	79.28	93.60	98.62	97.28	77.19	81.71	99.73	53.98	87.28	74.31	93.50	98.19

Table 2: Class-wise results on nuScenes test set. Metrics are provided in [%]

Metric	mean	Car	Bicycle	Motorcycle	Truck	Other Vehicle	Person	Bicyclist	Motorcyclist	Road	Parking	Sidewalk	Other Ground	Building	Fence	Vegetation	Trunk	Terrain	Pole	Traffic Sign
Assoc	80.9	89.0	32.0	63.0	88.0	56.0	49.0	82.0	31.0	-	-	-	-	-	-	-	-	-	-	-
IoU	67.6	97.0	61.0	78.0	84.0	73.0	83.0	95.0	0.0	96.0	44.0	80.0	4.0	88.0	56.0	89.0	71.0	76.0	66.0	45.0

Table 3: Class-wise results on SemanticKITTI val set. Metrics are provided in [%]. Note that association metrics are not available for 'stuff' classes.

3 Qualitative Comparison (LiDAR-only vs. Fusion)

Figures 2 and 3 provide a qualitative comparison of our proposed method with the LiDAR-only baseline (Tab. 4, row 1 in the main text). We provide the segmentation results in the LiDAR domain for both LiDAR-only and fusion models in the first two columns, respectively, and the corresponding camera view in the third column. The region of interest in each case is highlighted in red.

In the first example (Fig. 2), the baseline wrongly segments the building at range as vegetation due to the limited information obtained from the LiDAR input. By contrast, the final model with fusion effectively leverages the rich contextual information from the camera (highlighted by the red box) and segments the correct class.

In the second example (Fig. 3), the baseline fails to track pedestrians when they are close to each other (the two pedestrians on the left are merged together as a single instance). By contrast, the camera view provides distinct appearance cues for each pedestrian, enabling our model to accurately segment and track them.

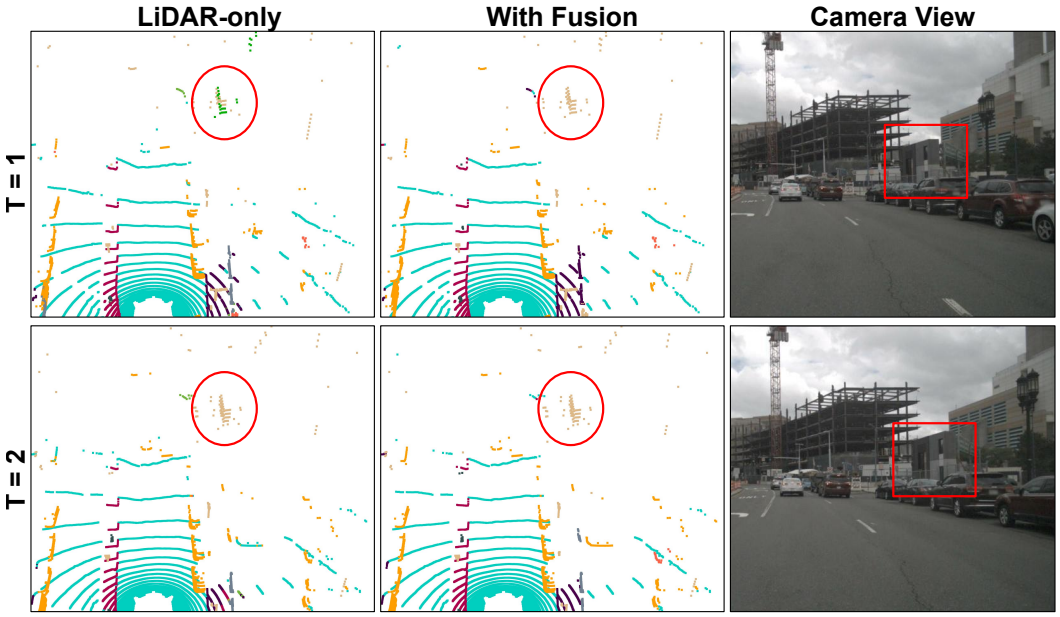

Figure 2: Qualitative comparison of semantic segmentation for LiDAR-only vs. fusion model on sequence 0105 from nuScenes.

Figure 3: Qualitative comparison of instance segmentation and tracking for LiDAR-only vs. fusion model on sequence 0003 from nuScenes.