Supplementary Materials for Waymo Open Dataset: Panoramic Video Panoptic Segmentation

Jieru Mei[1⋆], Alex Zihao Zhu[2], Xinchen Yan[2], Hang Yan[2], Siyuan Qiao[3]
Yukun Zhu[3], Liang-Chieh Chen[3], and Henrik Kretzschmar[2]

[1]Johns Hopkins University [2]Waymo LLC [3]Google Research

In the supplementary materials, we provide additional dataset statistics (Sec. 1), the semantic categories definition (Sec. 2), and video visualizations for our dataset annotations and baseline predictions (see other attachments). We also provide additional information about the panorama generation (Sec. 3), and discuss the proposed weighted STQ (wSTQ) in Sec. 4. Finally, we discuss the current limitations (failure modes) of our baselines and the challenges presented in our dataset in Sec. 5.

1 Additional Dataset Statistics

In this section, we provide more detailed statistics of our WOD-PVPS dataset, including instance association quality (Sec. 1.1) and the distribution of object instances (Sec. 1.2).

1.1 Instance Association Quality

As mentioned in the main paper, we exploit the existing 3D and 2D bounding box annotations in WOD [6] to associate instance IDs across cameras and frames. In this subsection, we analyze the effect of 3D and 2D bounding box associations (*i.e.*, step 2 and step 3 of Fig. 4 in main paper).

On the training set, we have 580k object instances in total with 207 instances per sequence. Our 3D and 2D association assigns 376k and 161k object instances with unique IDs within each sequence, respectively. The remaining 43k object instances (7.4% of the population) are treated as crowd and we exclude them in our experiments (*e.g.*, loss and metric computation).

On the validation set, we have 120k object instances in total with 6k instances per sequence on average. Our 3D and 2D association assigns 79k and 31k object instances with unique IDs, respectively. Similarly, the remaining 10k instances (8.3% of the population) are treated as crowd and we exclude them in our experiments.

To provide additional references for the instance Association Quality (AQ), we treat the final assignment of instance labels as ground-truth (which we show in the video demo) and compute the AQ metric on the 3D-only and 2D-only

⋆ Work done as an intern at Waymo.

Table A: Distribution of object instances. We report the statistics of the number of pixels covered by object instances in our training and validation set.

	Area < 100	Area $\in [100, 32^2)$	Area $\in [32^2, 96^2)$	Area $>= 96^2$
Train	4.1%	37.9%	36.7%	21.3%
Val	4.2%	38.7%	37.2%	19.8%

associations. The AQ is 71.06% for 3D boxes association only and 72.17% for 2D boxes association only, while a perfect match should have AQ of 100.00%. We can see the instance association with 3D or 2D ground-truth boxes alone is not sufficient, which suggests the hybrid association step is required for our challenging dataset.

1.2 Distribution of Object Instances

We provide additional statistics regarding the object instances of our dataset. The object instances in our training set have an average span of 3.6 temporal frames (out of 5 frames), while the object instances in our validation set have an average span of 21.6 temporal frames (out of 100 frames). Our validation set is challenging due to its emphasis on the longer-term consistency. We also compute the number pixels covered by each object instance per camera. Our training instances cover 20,349 pixels on average (and 1,560 pixels as the median number), while our validation instances cover 18,174 pixels on average (and 1,471 pixels as the median number). This suggests that object instances in our training and validation set share similar distributions in terms of object size. Finally, we show the detailed distribution in Tab. A, where we adopt the thresholds used in the literature [4].

2 Semantic Categories Definition

In Tab. B, we provide detailed definitions of the 28 semantic categories on WOD-PVPS dataset. As seen in the table, we follow the definition of the existing public dataets (*e.g.*, Cityscapes [1]) for the common classes (*e.g.*, `person`). We summarize the key differences in our dataset as follows. First, we have a fine-grained definition on the `vehicle` super-class, which contains 8 semantic categories, as the vehicle object instances play an important role in the urban driving applications. Compared to the existing public datasets, we replace the 'train' category with `other_large_vehicle` as a more generic class. Second, we have `lane_marker` and `road_marker` separated from `road`. This separation provides additional information for context modeling in the holistic scene understanding. For example, the `lane_marker` and `road_marker` are important signals for reasoning the vehicle object's motion (*e.g.*, lane change) in the scene. Third, `bird` and `ground_animal` are introduced as they have different motion patterns. In summary, our dataset

Table B: Semantic categories definition of our WOD-PVPS dataset. †: Classes that contain instance annotations consistent across cameras and time.

Super-Class	Class	Definition
vehicle	car†	A small vehicle such as sedan, SUV, pickup truck, minivan, and golf cart.
	bus†	A large vehicle that carries more than 8 passengers.
	truck†	A large vehicle that carries cargo.
	other_large_vehicle†	A large vehicle that is not a truck nor a bus.
	trailer†	A smaller or larger trailer attached to another vehicle or horse.
	ego_vehicle	The ego vehicle.
	motorcycle	Motorcycles with no rider.
	bicycle	Bicycles with no rider.
human	person†	A pedestrian. Does not include objects that are sticking out of the contour of the pedestrian, such as suitcases, strollers or cars.
	cyclist†	A bicycle with rider.
	motorcyclist†	A motorcycle with rider.
animal	ground_animal	Animals that run on the ground such as dog, cat, cow, etc.
	bird	Birds.
object	pole	Permanent horizontal and vertical lamp poles, traffic-sign poles, etc.
	sign	Signs related to traffic, including front and back facing signs.
	traffic_light	The box that contains traffic lights regardless of front or back facing.
	construction_cone	Cones and short poles related to construction.
	pedestrian_object	Large objects carried by a pedestrian that are sticking out of their contour.
building	building	Permanent buildings and walls, including solid fences.
flat	road	Drivable road with proper markings, including parking lots and gas stations.
	sidewalk	Paved walkable surface for pedestrians, including curbs.
	road_marker	All markings on the road other than lane markers.
	lane_marker	Markings on the road that are parallel to the ego vehicle and defines lanes.
vegetation	vegetation	Vegetation including tree trunks/branches, bushes, tall grasses, flowers etc.
sky	sky	The sky, including clouds.
void	ground	Other horizontal surfaces that are drivable or walkable.
	static	Permanent object that does not belong to any of above classes.
	dynamic	Objects that are not permanent in their current position and do not belong to any of above classes.

annotates 28 semantic classes, among which 8 classes contain instance annotations that are consistent across cameras and time, presenting a more challenging scenario than existing datasets [1,3,7].

3 Panorama Generation

We provide additional information on our method that generates equirectangular panorama and explain our considerations in the streaming prediction setting [5]. In order to unproject each pixel into the 3D space, one has to estimate the corresponding pixel depth. However, estimating accurate depth map could potentially cause serious model prediction latency in the streaming prediction setting. Instead, we consider a simplified version where we assume each pixel is captured at 100 metres from the camera center and generate the panorama this way for our baseline method. This can be a potentially unfavorable factor to the panorama baseline regarding the performance, as mapping from pixels of the close-range objects may not be very accurate. Indeed, this is another challenging fact of our benchmark that we leaves as an open research topic for future study.

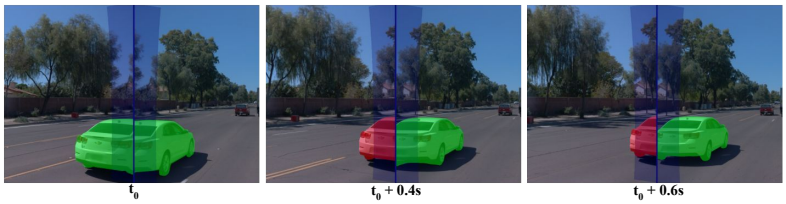

Fig. A: A toy example for comparing wSTQ and STQ. The overlapping regions between two cameras' field-of-view are colored in blue. The model makes an accurate prediction at t_0 (the accurate prediction is colored in green), but predicts a wrong semantic class in the field-of-view of the left camera (colored in red) in the following time steps. The existing metric STQ over-penalizes this case (since the pixels are covered by two cameras), while the proposed wSTQ properly balances the results.

4 wSTQ *vs*. STQ

In this section, we discuss our proposed weighted STQ (wSTQ), particularly how wSTQ alleviates the pixel double counting issue when naïvely applying STQ [7] to the task of panoramic video panoptic segmentation.

We use a toy example, illustrated in Fig. A, to better understand the pixel double counting issue. For simplicity, we only consider two semantic classes: car and background. As shown in the figure, the large white car (in the center) spans the field-of-views of two cameras (the overlapping regions between cameras are marked in blue). Let's suppose the model makes a perfect prediction (colored in green) across two cameras at the time step t_0. However, in the following time steps (*i.e.*, $t_0 + 0.4$s and $t_0 + 0.6$s), the model predicts a wrong semantic class in the field-of-view of the left camera (colored in red). Consequently, as shown in Tab. C, the semantic errors (colored in red) are penalized twice by STQ (once for left camera, and once for right camera), leading to a much lower performance compared to wSTQ, which properly weights the pixels according to their coverage by cameras. This avoids the metric computation biased in the overlapping regions.

Table C: wSTQ *vs*. STQ for the toy example in Fig. A. The errors in the overlapping region are doubled counted in STQ (2nd row: non-weighted), while the proposed weighted STQ (1st row: weighted) considers the panoramic property of our dataset, and thus properly reflects the segmentation and tracking quality.

Metric	STQ	AQ	SQ
weighted	84.98	74.50	96.29
non-weighted	79.42	69.16	91.20

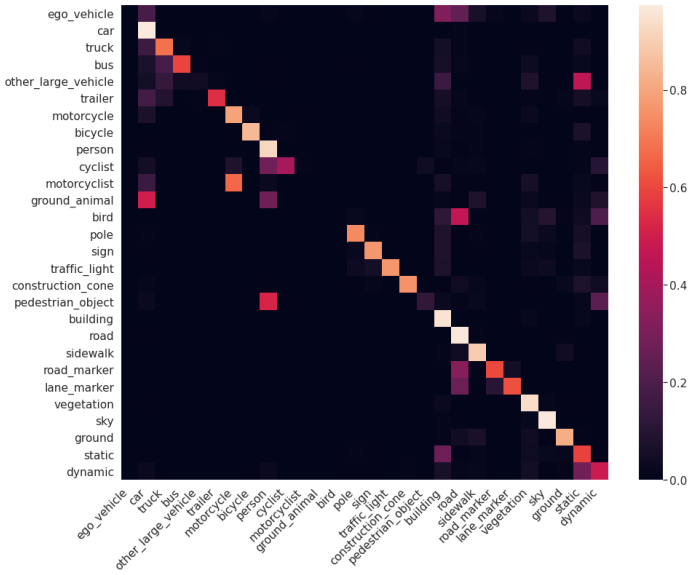

Fig. B: Confusion matrix for View-View (trained on View scheme and evaluated on View scheme) ViP-DeepLab baseline.

5 Failure Cases Analysis

Thanks to the decomposition property of wSTQ (inherited from STQ [7]), we are able to separately analyze the performance in terms of segmentation quality and association quality. We look into the failure cases of each quality in the following subsections.

5.1 Segmentation Quality

Segmentation quality is the typical mean Intersection-over-Union [2]. To better understand the error patterns, we visualize the confusion matrix in Fig. B. As shown in the figure, we summarize the error patterns into two modes, namely, intra-class errors and inter-class errors. First, the intra-class errors happen more frequently for classes such as `motorcyclist`, `bird`, and `ground_animal`. We noticed the model has difficulty in predicting accurate segmentation, mostly due to their relatively small regions or low pixel frequencies (*i.e.*, only a small amount of pixels for them) in the dataset. Second, our detailed semantic category definition presents a new challenge. For classes, such as `road`, `lane_marker` and `road_marker`, the model needs to understand the full scene in a holistic way in order to make accurate predictions for these classes.

In summary, our proposed WOD-PVPS dataset presents new challenges (*e.g.*, the class-imbalance and detailed semantic category definitions) to the research community.

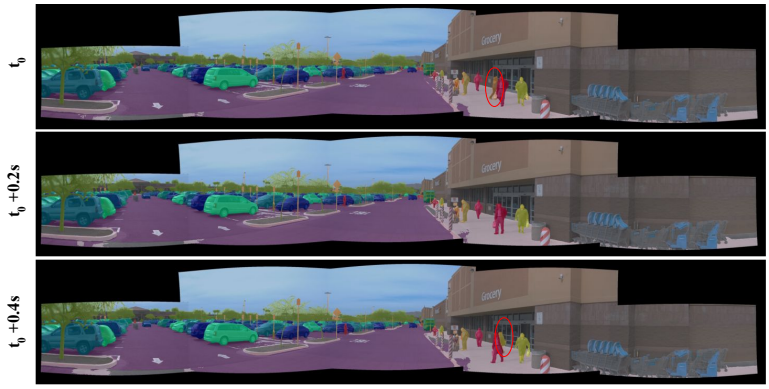

Fig. C: Failure case for association quality: an ID switch occurs due to the occlusion.

Fig. D: Failure case for association quality: Over-segmented instances.

5.2 Association Quality

The association quality aims to measure long-term tracking. To better understand the failure modes for association quality, we visualize them in Fig. C and Fig. D. In Fig. C, the person in orange (highlighted in a cycle) is occluded by the person in red at $t_0 + 0.2$s, and re-appears in the next frame at $t_0 + 0.4$s. As the ViP-DeepLab baseline performs "panoptic stitching over time" frame-by-frame, it could not recover from the lost tracking, resulting in an ID switch for that person. Fig. D shows another failure case, where the model over-segments a large object into multiple small instances, failing to properly associate the large object. As shown in the figures, our dataset presents new challenges for long-term tracking (*e.g.*, instance segmentation and tracking in a crowded region, and tracking and segmentation for large objects).

References

1. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The Cityscapes Dataset for Semantic Urban Scene Understanding. In: CVPR (2016) 2, 3
2. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The Pascal Visual Object Classes (VOC) Challenge. IJCV (2010) 5
3. Kim, D., Woo, S., Lee, J.Y., Kweon, I.S.: Video Panoptic Segmentation. In: CVPR (2020) 3
4. Kisantal, M., Wojna, Z., Murawski, J., Naruniec, J., Cho, K.: Augmentation for small object detection. arXiv preprint arXiv:1902.07296 (2019) 2
5. Li, M., Wang, Y.X., Ramanan, D.: Towards streaming perception. In: ECCV (2020) 3
6. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al.: Scalability in perception for autonomous driving: Waymo Open Dataset. In: CVPR (2020) 1
7. Weber, M., Xie, J., Collins, M., Zhu, Y., Voigtlaender, P., Adam, H., Green, B., Geiger, A., Leibe, B., Cremers, D., Osep, A., Leal-Taixe, L., Chen, L.C.: Step: Segmenting and tracking every pixel. In: NeurIPS Track on Datasets and Benchmarks (2021) 3, 4, 5