BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning

Supplementary Material

1. Dataset Details

We present more details about our dataset and annotations in this section.

1.1. Image Tagging

Figure 1 shows the distribution of weather, scene, and time of day attributes in BDD100K. The distribution demonstrates visual diversity of the images and thus provides an opportunity to study visual transfer between different domains.

1.2. Object Detection

	Caltech [1]	KITTI [3]	City [6]	Ours
# persons	1,273	6,336	19,654	86,047
# per image	1.4	0.8	7.0	1.2

Table 1: Comparisons on number of pedestrians with other datasets. The statistics are based on the training set in each dataset. Our dataset has more examples of pedestrians, but because our dataset contains non-city scenes such as highways, the number of person per image is lower than Cityscapes.

1.3. Lane Marking and Drivable Area

Our choice of annotated lane attributes is based on their influence on driving decisions. The continuity of a lane marking is essential for making a "driving-across" decision, so we labeled it independently as an important attribute. Similarly, the direction of a lane marking is also significant for autonomous driving. For example, if a lane marking is parallel to the passing car, it may serve to guide cars and separate lanes; if it is perpendicular, it can be treated as a sign of deceleration or stop. The distribution of the number of annotations in varied driving scenes are shown in Figure 3a, Figure 3b, and Figure 3c. The detailed evaluation results for the lane marking benchmark are in Table 5.

Drivable area detection is a new task, so we show results of a baseline method on the task here. First, the drivable area detection is converted to 3-way segmentation task (background, directly, and alternatively drivable) by ignoring the region ID. Then, we train DRN-D-22 model [5] on the 70,000 training images. We find that after learning from the large-scale image dataset, the model learns to split the road according to the lanes and extrapolate the drivable area to unmarked space. The mIoU for directly and alternatively drivable areas is 79.4% and 63.3%. However, the same model can achieve much higher accuracy on road segmentation, which indicates that techniques beyond segmentation may be required to solve the drivable area problem.

	Total	person	rider	car	truck	bus	train	motorcycle	bicycle
Tracks	131K	24K	1.0K	97K	4.6K	1.5K	34	685	1.8K
Boxes	3.3M	440K	23K	2.6M	177K	67K	1.9K	13K	33K
Truncated	346K	18K	1.4K	284K	26K	11K	557	1.4K	3.1K
Occluded	2.2M	253K	18K	1.7M	132K	51K	1.9K	9.1K	26K

Table 2: Annotations of the BDD100K MOT dataset by category.

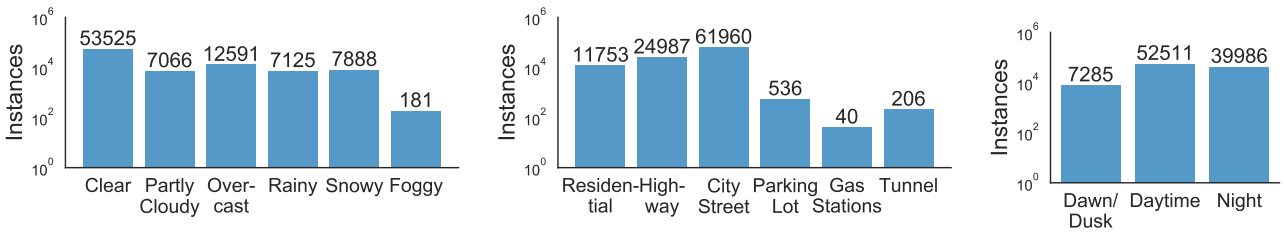
1.4. GPS Trajectory

Figure 4 shows GPS trajectories of example sequences. Our data presents diverse driving behaviors, like starting, stopping, turning and passing. The data is suitable to train and test imitation learning algorithms on real driving data.

1.5. Semantic Instance Segmentation

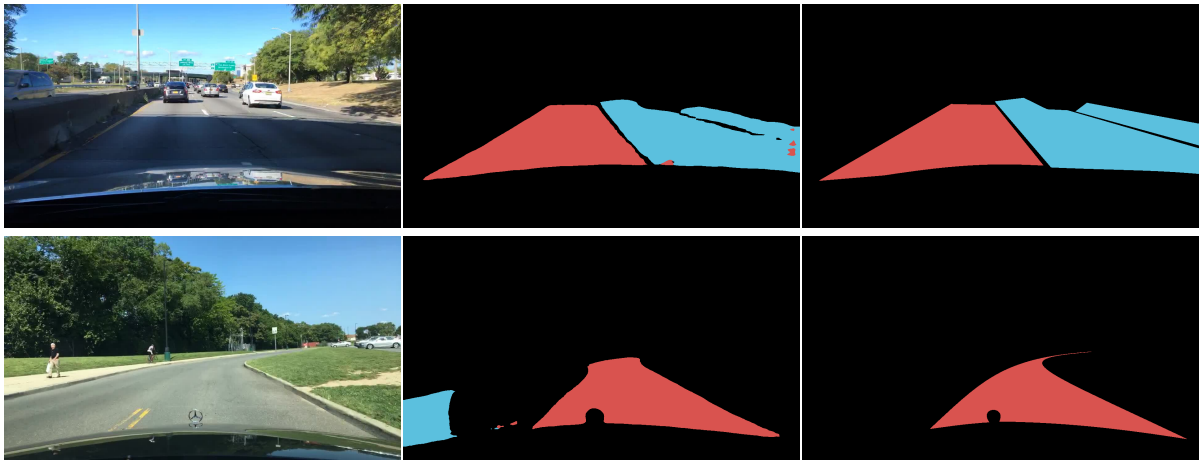
Figure 5 shows the distribution of number of instances observed in the segmentation annotations. BDD100K has a good coverage on rare categories (e.g. trailer, train) and large number of instances of common traffic objects such as persons and cars. We also observe long-tail effects on our dataset. There are almost 60 thousand car instances, a few hundred rider and motorcycle instances, and mere dozens of trailer and train instances.

Figure 9 in the main paper shows some segmentation examples produced by DRN-D-38. They also reveal some interesting properties of various domains. Probably because of the infrastructure differences between Germany and the US, the models trained on Cityscapes confuse some big structures in an unreasonable way, such as segmenting the sky as building as shown in the third row of the figure. The model is also confused by the US highway traffic sign. However, the same model trained on our dataset does not



(a) Weather (b) Scene (c) Hours

Figure 1: Distribution of images in weather, scene, and day hours categories.



(a) Image (b) DRN (c) Ground truth

Figure 2: Drivable area prediction by segmentation. The segmentation predicts the drivable area with lanes well, as shown in the top row. Also, we find that the segmentation model learns to interpolate in areas that has no lane markings.

suffer these problems. Also, the model of Cityscapes may over-fit the hood of the data collecting vehicle and produces erroneous segmentation for the lower part of the images.

1.6. Multiple Object Tracking and Segmentation

Table 2 and Table 3 shows the label distributions by categories. Our bounding box tracking annotations cover more than one hundred thousand instances with more than two million bounding boxes, and the segmentation tracking set contains more than six thousand instances with one hundred thousand polygons. We showed in the paper submission, our tracking annotation set is one of the largest out there, in addition to our advantage in multitask and diversity.

	Total	person	rider	car	truck	bus	train	motorcycle	bicycle
Tracks	6.3K	1.8K	31	4.0K	215	93	4	21	76
Masks	129K	22K	894	93K	7.6K	4.0K	117	369	1.4K
Truncated	15K	833	45	12K	1.3K	743	8	49	70
Occluded	85K	13K	793	61K	5.7K	3.1K	116	292	970

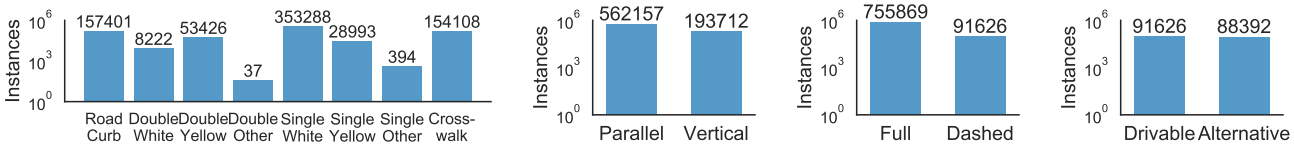
Table 3: Annotations of BDD100K MOTS by category.

2. Model Details

In this section, we present more implementation details for benchmark models.

2.1. Tracking

We use a modified Faster R-CNN [4] architecture for tracking similar with Feichtenhofer *et al*. [2]. Like Feichtenhofer *et al*. [2], we use a correlation module and a bounding box propagation (regression) head to estimate the bounding box offset between two frames for short-term association. We also implement an association head based on appearance to learn embeddings for instance re-identification. During training, we sample a pair of frames within the interval of $t = 3$ frames. During inference, we first perform detection for the first frame. For each subsequent frame, we use the propagation head to associate detected bounding boxes with boxes from the previous frame based on overlap. We then use the association head based on appearance to associate the rest with the unmatched boxes in the previous 15 frames using dot product of the embeddings followed by softmax.



(a) Lane category (b) Lane direction (c) Lane continuity (d) Drivable areas

Figure 3: Distribution of different types of lane markings and drivable areas.

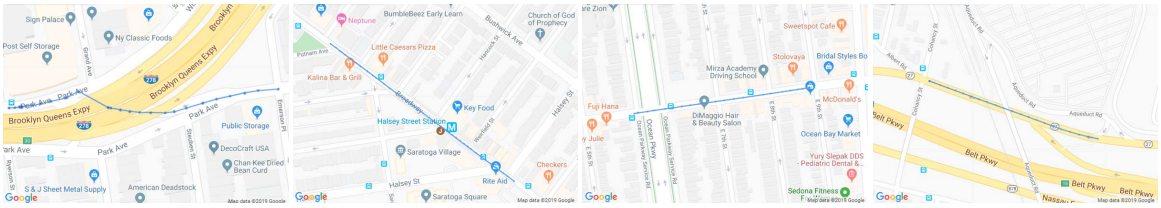


Figure 4: Trajectories of example driving videos.

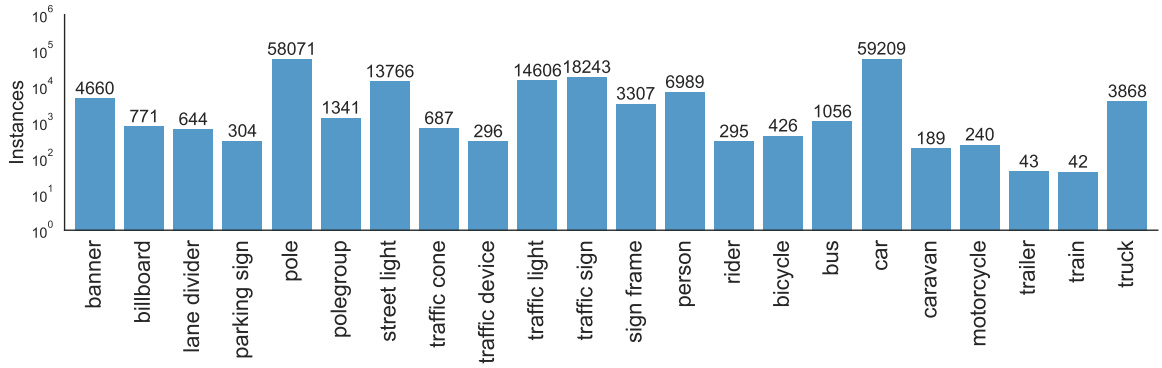


Figure 5: Distribution of classes in semantic instance segmentation. It presents a long-tail effect with more than 10 cars and poles per image, but only tens of trains in the whole dataset.

Figure 6: Example annotations for BDD100K MOTS. Frames are down-sampled for visualization.

Train (30K)	Test	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
City	City	**29.5**	**55.3**	**27.2**	**14.4**	**32.8**	**47.2**
Non-City		24.9	48.6	22.1	11.7	28.1	40.7
Random		28.7	54.5	25.8	13.7	31.9	47.0
City	Non-City	26.5	49.3	**25.5**	13.5	**32.1**	47.0
Non-City		24.3	46.0	22.4	13.3	30.0	42.0
Random		**26.6**	**49.8**	24.4	**14.4**	31.8	**47.4**
City	Val	**28.8**	54.1	**26.8**	**13.8**	**32.7**	**47.0**
Non-City		24.9	48.3	22.2	11.8	28.7	41.2
Random		28.7	**54.5**	25.8	13.7	31.9	**47.0**
Daytime	Daytime	**30.6**	**56.0**	**28.7**	**16.4**	**35.6**	**50.7**
Non-Daytime		25.9	49.6	23.2	12.8	30.7	42.7
Random		29.5	55.0	27.2	15.7	34.5	48.7
Daytime	Non-Daytime	23.6	46.1	21.2	10.5	25.4	41.3
Non-Daytime		25.3	49.9	22.0	11.6	26.6	43.4
Random		**26.0**	**50.9**	**22.5**	**12.8**	**27.4**	**45.1**
Daytime	Val	28.1	52.8	**25.9**	13.3	**31.9**	**47.0**
Non-Daytime		25.6	49.8	22.6	11.5	29.1	42.5
Random		**28.7**	**54.5**	25.8	**13.7**	**31.9**	**47.0**

Table 4: Full evaluation results of the domain discrepancy experiments with object detection.

Threshold	Training Set	Direction			Continuity			Category							total avg.
		parallel	vertical	avg.	continuous	dashed	avg.	crosswalk	double white	double yellow	road curb	single white	single yellow	avg.	
$\tau = 1$	Lane 10K	28.41	28.35	28.38	28.31	26.32	27.31	27.48	6.5	32.99	19.92	28.51	27.09	23.75	26.48
	Lane+Drivable 10K	31.19	32.46	31.83	31.89	28.84	30.36	31.35	14.41	37	24.28	30.4	28.6	27.68	29.95
	Lane 20K	34.45	36.62	35.54	34.58	33.61	34.09	35.73	20.75	39.7	27.59	34.53	33.5	31.97	33.87
	Lane+Drivable 20K	34.45	36.32	35.38	34.51	33.32	33.92	35.34	20.14	39.69	27.59	34.42	33.4	31.76	33.69
	Lane 70K	34.57	36.92	35.74	34.62	33.85	34.23	36.17	21.51	39.88	27.91	34.62	33.77	32.31	34.1
	Lane+Drivable 70K	34.48	36.60	35.54	34.49	33.62	34.05	35.78	20.7	39.69	27.87	34.4	33.47	31.99	33.86
$\tau = 2$	Lane 10K	35.76	36.63	36.19	35.48	33.91	34.70	35.85	7.76	39.31	26.64	35.61	32.73	29.65	33.51
	Lane+Drivable 10K	38.79	41.26	40.03	39.28	37.01	38.14	40.26	16.94	43.34	31.78	37.78	34.78	34.15	37.44
	Lane 20K	42.44	46.03	44.23	42.32	42.41	42.37	45.31	24.89	46.35	35.76	42.41	40.34	39.18	41.93
	Lane+Drivable 20K	42.42	45.65	44.03	42.22	42.06	42.14	44.78	24.07	46.38	35.77	42.23	39.99	38.87	41.68
	Lane 70K	42.56	46.40	44.48	42.32	42.71	42.51	45.8	25.44	46.54	36.09	42.48	40.47	39.47	42.15
	Lane+Drivable 70K	42.48	46.00	44.24	42.18	42.46	42.32	45.32	24.6	46.39	36.08	42.25	40.09	39.12	41.89
$\tau = 10$	Lane 10K	49.35	49.22	49.29	48.32	47.39	47.85	49.25	9.37	46.62	44.14	46.41	38.72	39.08	45.41
	Lane+Drivable 10K	54.07	53.87	53.97	52.61	52.57	52.59	53.37	20.64	51.05	50.27	50.58	41.98	44.65	50.4
	Lane 20K	56.34	58.38	57.36	54.71	56.99	55.85	58.68	30.71	54.48	52.73	54.49	48.19	49.88	54.36
	Lane+Drivable 20K	56.31	58.07	57.19	54.59	56.69	55.64	58.24	29.39	54.68	52.86	54.22	47.62	49.5	54.11
	Lane 70K	56.3	58.70	57.50	54.59	57.16	55.87	59.1	30.87	54.86	53.05	54.34	48.28	50.08	54.48
	Lane+Drivable 70K	56.41	58.29	57.35	54.53	56.98	55.76	58.53	29.63	54.6	53.06	54.22	47.72	49.63	54.24

Table 5: Full evaluation results of the individual lane marking task and the joint training of lane marking and the drivable area detection. We report the ODS-F scores with different thresholds $\tau = 1, 2, 10$ pixels of direction, continuity as well as each category.

References

[1] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 304–311. IEEE, 2009. 1

[2] C. Feichtenhofer, A. Pinz, and A. Zisserman. Detect to track and track to detect. 2017. 2

[3] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 1

[4] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 2

[5] F. Yu, V. Koltun, and T. Funkhouser. Dilated residual networks. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[6] S. Zhang, R. Benenson, and B. Schiele. Citypersons: A diverse dataset for pedestrian detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1