Supplementary Material: Kinematic 3D Object Detection in Monocular Video

Garrick Brazil[1], Gerard Pons-Moll[2], Xiaoming Liu[1], and Bernt Schiele[2]

[1] Michigan State University, Computer Science & Engineering
[2] Max Planck Institute for Informatics, Saarland Informatics Campus
[1] {brazilga, liuxm}@msu.edu, [2] {gpons, schiele}@mpi-inf.mpg.de

	AP_{3D} (IoU $\geq$ [0.7/0.5])			AP_{BEV} (IoU $\geq$ [0.7/0.5])		
	Easy	Mod	Hard	Easy	Mod	Hard
2 bins	12.83/46.46	9.47/33.78	7.93/26.85	19.17/52.06	14.72/37.54	11.38/31.16
4 bins	12.65/44.01	10.02/33.27	7.87/26.27	19.09/49.86	14.55/37.90	11.14/30.44
10 bins	14.27/49.71	10.74/36.12	8.29/28.62	21.12/54.70	15.37/39.72	11.60/31.75
Our decomp.	16.66/51.47	12.10/38.58	9.40/30.98	23.15/56.48	17.43/42.53	13.48/34.37

Table 1. Orientation. We compare our orientation decomposition to bin-based orientation following the high-level concepts within [3–5, 7], using AP_{3D} and AP_{BEV}. We evaluate our performances on the KITTI validation set [2] using IoU $\geq$ 0.7/0.5.

1 Orientation Ablations

We provide detailed experiments on 3D object detection and Bird's Eye View tasks to compare our orientation decomposition performance with bin-based approaches such as [3–5, 7] within Tab. 1. Recall that bin-based orientation first classifies the best bin for orientation then predicts an offset with respect to the bin. In contrast, our method disentangles the bin classification into a distinct explainable objectives such as an axis classification and a heading classification. For such experiments we change our formulation to use bins of $[2,\ 4,\ 10]$, where 4 bins has a similar representational power as two binary classifications $[\theta_a,\ \theta_h]$. The bins are spread uniformly from $[0,\ 2\pi]$ and an offset is predicted afterwards. We use the settings in Sec. 3.4 in main paper. We emphasize that our method outperforms the bin-based approaches between $\approx 1.36 - 2.63\%$ on AP_{3D} and $\approx 2.06 - 2.71\%$ on AP_{BEV} using the standard moderate setting and IoU $\geq$ 0.7.

2 Kalman Forecasting

Since our method uses ego-motion and a 3D Kalman filter to aggregate temporal information, the approach can be modified to act as a box forecaster. Although our method was not strictly designed for the tracking and forecasting task, we evaluate the 3D object detection and Bird's Eye View performance after forecasting $n_f = [1,\ 2,\ 3,\ 4]$ frames into the future. We assume a static ego-motion for unknown frames and otherwise use the Kalman equations described in the main paper Sec. 3.3 to forecast the tracked boxes.

For all forecasting experiments we process 4 temporally adjacent frames before forecasting. Since KITTI only provides a current frame and 3 proceeding frames, we carefully map images back to the raw dataset in order to forecast. For instance, when $n_f = 2$ we infer using frames $[-5, -4, -3, -2]$ then forecast ego-motion and Kalman n_f times. We then evaluate with respect to frame 0 which is the standard timestamp KITTI provides images and 3D labels for. We provide detailed performances on AP_{3D} in Tab. 2 and AP_{BEV} in Tab. 3. We find that the forecasting performance degrades through time but performs reasonably $1-2$ frames ahead, being competitive in magnitude to state-of-the-art methods on the KITTI test dataset as reported in Tab. 1 of the main paper. For instance, forecasting 1 and 2 frames results in 10.64% and 5.10% AP_{3D} respectively, which are competitive to methods [1, 3–8] on the test dataset.

	AP_{3D} (IoU $\geq$ [0.7/0.5/0.3])		
	Easy	Mod	Hard
Forecast → 4	1.16 / 18.47 / 47.26	0.84 / 11.21 / 29.22	0.62 / 8.97 / 23.40
Forecast → 3	3.72 / 28.97 / 58.46	2.32 / 18.05 / 37.82	1.75 / 13.88 / 29.80
Forecast → 2	7.84 / 39.40 / 68.87	5.10 / 25.48 / 48.30	4.14 / 20.20 / 37.84
Forecast → 1	16.09 / 49.66 / 75.88	10.64 / 34.18 / 55.26	8.14 / 26.62 / 44.01
No Forecast	19.76 / 55.44 / 79.81	14.10 / 39.47 / 60.57	10.47 / 31.26 / 48.95

Table 2. Forecasting - 3D Object Detection. We evaluate our forecasting performance on AP_{3D} within the KITTI validation [2] set and using IoU $\geq$ 0.7/0.5/0.3.

	AP_{BEV} (IoU $\geq$ [0.7/0.5/0.3])		
	Easy	Mod	Hard
Forecast → 4	5.48 / 29.40 / 54.52	3.54 / 18.13 / 36.13	2.90 / 14.71 / 28.49
Forecast → 3	11.03 / 39.08 / 64.87	6.89 / 24.01 / 43.52	5.67 / 18.85 / 34.91
Forecast → 2	17.02 / 47.07 / 72.33	10.76 / 31.62 / 51.67	8.37 / 25.47 / 40.79
Forecast → 1	23.58 / 55.99 / 77.48	15.79 / 39.33 / 58.05	12.54 / 31.22 / 46.59
No Forecast	27.83 / 61.79 / 81.20	19.72 / 44.68 / 63.44	15.10 / 34.56 / 49.84

Table 3. Forecasting - Bird's Eye View. We evaluate our forecasting performance on AP_{BEV} within the KITTI validation [2] set and using IoU $\geq$ 0.7/0.5/0.3.

3 Qualitative Video

We further provide a qualitative demonstration video at http://cvlab.cse.msu.edu/project-kinematic.html. The video demonstrates our framework's ability to determine a full scene understanding including 3D object cuboids, per-object velocity and ego-motion. We compare to a related monocular work of M3D-RPN [1], plot ground truths, image view, Bird's Eye View, and the track history.

References

1. Brazil, G., Liu, X.: M3D-RPN: Monocular 3D region proposal network for object detection. In: ICCV. IEEE (2019) 2
2. Chen, X., Kundu, K., Zhu, Y., Berneshawi, A.G., Ma, H., Fidler, S., Urtasun, R.: 3D object proposals for accurate object class detection. In: NeurIPS (2015) 1, 2
3. Ku, J., Pon, A.D., Waslander, S.L.: Monocular 3D object detection leveraging accurate proposals and shape reconstruction. In: CVPR. IEEE (2019) 1, 2
4. Li, B., Ouyang, W., Sheng, L., Zeng, X., Wang, X.: GS3D: An efficient 3D object detection framework for autonomous driving. In: CVPR. IEEE (2019) 1, 2
5. Liu, L., Lu, J., Xu, C., Tian, Q., Zhou, J.: Deep fitting degree scoring network for monocular 3D object detection. In: CVPR. IEEE (2019) 1, 2
6. Ma, X., Wang, Z., Li, H., Zhang, P., Ouyang, W., Fan, X.: Accurate monocular 3D object detection via color-embedded 3d reconstruction for autonomous driving. In: ICCV. IEEE (2019) 2
7. Manhardt, F., Kehl, W., Gaidon, A.: ROI-10D: Monocular lifting of 2D detection to 6D pose and metric shape. In: CVPR. IEEE (2019) 1, 2
8. Simonelli, A., Bulo, S.R., Porzi, L., López-Antequera, M., Kontschieder, P.: Disentangling monocular 3D object detection. In: ICCV. IEEE (2019) 2