Supplementary Material for DID-M3D: Decoupling Instance Depth for Monocular 3D Object Detection

Liang Peng[1,2], Xiaopei Wu[1], Zheng Yang[2], Haifeng Liu[1], and Deng Cai[1,2] ✉

[1] State Key Lab of CAD&CG, Zhejiang University, China
{pengliang, wuxiaopei, haifengliu}@zju.edu.cn dengcai@cad.zju.edu.cn
[2] Fabu Inc., Hangzhou, China
yangzheng@fabu.ai

Considering the space limitation of the main text, we provided more results and discussion in this supplementary material, which is organized as follows:

- Section A: results on Waymo dataset.
- Section B: detailed ablation analysis and discussion.
  - Section B.1: decoupled instance depth.
  - Section B.2: grid design.
  - Section B.3: affine-based data augmentation.
  - Section B.4: instance depth uncertainty.
  - Section B.5: instance depth aggregation.
- Section C: qualitative results.
  - Section C.1: analysis on attribute depth and visual depth uncertainty.
  - Section C.2: more qualitative results of 3D box predictions.
  - Section C.3: failure cases and discussion.

A Results on Waymo Dataset

We perform experiments on Waymo dataset [4], which is a large-scale modern dataset for self-driving. It contains 798 sequences for training and 202 sequences for validation. We use the same data split strategy proposed in CaDDN [3]. The processed training dataset includes approximately 50K training samples. We show the results in Table 1. We can see that our method performs the best on most metrics. It further validates the effectiveness of the proposed method.

B Detailed Ablation Analysis and Discussion

To better understand the effect of each component in our method, we perform more detailed ablation studies. The results are shown in Table 2. We perform 5 groups of experiments, *i.e.*, experiment ([a,b,c], [d,e,f,g], [h,i,j,k], [l,m,n,o], [p,q,r,s,t,u]), to compare the decoupled depth with the original prediction. We also extend our grid design, affine transformation based data augmentation, depth uncertainty, and depth aggregation to the baseline for comprehensive comparisons. For comparison convenience, we copy the results from experiment (e, g, i, k) to experiment (m, o, q, s), respectively.

Table 1. Results on Waymo *val* set. *"Runtime*"* in the table refers to the runtime reported on KITTI. DID-M3D performs the best.

Methods	Venue	Runtime*	3D mAP/mAPH			
			Overall	0−30m	30−50m	50m−∞
Under Level 1 (IoU=0.5)						
PatchNet [2]	*ECCV20*	488ms	2.92/2.74	10.03/9.75	1.09/0.96	0.23/0.18
CaDDN [3]	*CVPR21*	630ms	**17.54/17.31**	**45.00/44.46**	**9.24/9.11**	0.64/0.62
PCT [5]	*NeurIPS21*	445ms	4.20/4.15	14.70/14.54	1.78/1.75	0.39/0.39
MonoJSG [1]	*CVPR22*	42ms	5.65/5.47	20.86/20.26	3.91/3.79	**0.97/0.92**
DID-M3D	*ECCV22*	40ms	**20.66/20.47**	**40.92/40.60**	**15.63/15.48**	**5.35/5.24**
Under Level 2 (IoU=0.5)						
PatchNet [2]	*ECCV20*	488ms	2.42/2.28	10.01/9.73	1.07/0.94	0.22/0.16
CaDDN [3]	*CVPR21*	630ms	**16.51/16.28**	**44.87/44.33**	**8.99/8.86**	0.58/0.55
PCT [5]	*NeurIPS21*	445ms	4.03/3.99	14.67/14.51	1.74/1.71	0.36/0.35
MonoJSG [1]	*CVPR22*	42ms	5.34/5.17	20.79/20.19	3.79/3.67	**0.85/0.82**
DID-M3D	*ECCV22*	40ms	**19.37/19.19**	**40.77/40.46**	**15.18/15.04**	**4.69/4.59**

B.1 Decoupled Instance Depth

As shown in Table 2, for every group of experiments, we investigate the effect of the decoupled instance depth. We can easily see that the decoupled design consistently improves the overall performance with a significant margin. For example, on the naive baseline, the decoupled design boosts the performance from 13.43/8.70 to 16.49/10.94 (experiment b→c) under the moderate setting. When employing other strategies, it also improves the AP from 19.82/14.47 to 22.76/16.12 (experiment t→u) under the moderate setting. These improvements validate its effectiveness.

B.2 Grid Design

Most previous monocular works produce a single instance depth prediction. Our method divides the RoI into grids to decouple the instance depth. The grid design produces multiple predictions, which may be unfair to the single prediction. Therefore, we extend the grid design to the baseline for fair comparisons. We can see that the baseline benefits from this grid design with slight improvements. The naive baseline obtains 1.0/0.24 AP improvements (experiment a→b) under the moderate setting. However, when the network is equipped with other useful components, gains from the grid design are weakened (*e.g.*, 19.68/14.13→19.82/14.47 (experiment j→t) under the moderate setting).

B.3 Affine-based Data Augmentation

We extend the affine-based data augmentation to the baseline detector, Please note, in this process we directly scale the instance depth, as the baseline uses the direct instance depth prediction. When using affine-based data augmentation, the baseline benefits from it a lot (*e.g.*, 12.43/8.54→16.25/11.20 (experiment a→f) under the moderate setting). Even if without a correct instance depth

Table 2. Detailed ablation studies. "E.": experiments; "Dec. ID.": decoupled instance depth; "G.": grid; "Aff. Aug.": affine-based data augmentation; "Tr. ID.": transformed instance depth; "ID. U.": instance depth uncertainty; "ID. C.": instance depth confidence; "ID. AA.": instance depth adaptive aggregation. The transformed instance depth ("Tr. ID.") refers to the depth transformation in the affine-based data augmentation.

E.	Dec. ID.	G.	Aff. Aug.	Tr. ID.	ID. U.	ID. C.	ID. AA.	AP_{BEV}/AP_{3D} (IoU=0.7)$\|_{R_{40}}$ Easy	Moderate	Hard
(a)								15.52/10.55	12.43/8.54	11.42/7.14
(b)		✓						16.23/10.94	13.43/8.70	11.88/7.98
(c)	✓	✓						**19.86/13.13**	**16.49/10.94**	**14.40/9.89**
(d)			✓					16.91/11.21	13.20/9.08	12.07/8.16
(e)	✓	✓	✓					17.98/12.11	15.34/10.72	14.11/9.11
(f)			✓	✓				19.84/14.07	16.25/11.20	14.13/9.97
(g)	✓	✓	✓	✓				**22.98/16.95**	**18.72/13.24**	**16.57/11.23**
(h)			✓	✓	✓			21.75/15.94	17.87/12.64	15.45/11.16
(i)	✓	✓	✓	✓	✓			25.23/18.14	20.06/13.91	17.63/12.52
(j)			✓	✓	✓	✓		25.56/18.91	19.68/14.13	16.94/12.32
(k)	✓	✓	✓	✓	✓	✓		**29.34/21.51**	**21.53/15.57**	**18.55/12.84**
(l)		✓	✓					17.60/12.14	14.09/9.37	12.36/8.42
(m)	✓	✓	✓					17.98/12.11	15.34/10.72	14.11/9.11
(n)		✓	✓	✓				21.01/14.67	16.79/11.24	15.41/10.18
(o)	✓	✓	✓	✓				**22.98/16.95**	**18.72/13.24**	**16.57/11.23**
(p)		✓	✓	✓	✓			22.76/16.56	18.40/12.99	16.15/11.06
(q)	✓	✓	✓	✓	✓			25.23/18.14	20.06/13.91	17.63/12.52
(r)		✓	✓	✓	✓	✓		26.61/19.44	19.72/14.44	16.85/12.13
(s)	✓	✓	✓	✓	✓	✓		29.34/21.51	21.53/15.57	18.55/12.84
(t)		✓	✓	✓	✓	✓	✓	26.72/19.48	19.82/14.47	16.93/12.15
(u)	✓	✓	✓	✓	✓	✓	✓	**31.10/22.98**	**22.76/16.12**	**19.50/14.03**

transformation, the performance is still boosted (12.43/8.54→13.20/9.08 (experiment a→d) under the moderate setting), which can be attributed to the improvements on the robustness for a simple baseline.

On the other hand, for our decoupled manner, the instance depth is decoupled in a more intuitive and reasonable way, thus incorrect depth transformation damages the accuracy (downgrading from 16.49/10.94 to 15.34/10.72 (experiment c→e) under the moderate setting). With the correct depth transformation, our method obtains significant improvements via the affine-based data augmentation (2.23/2.30 AP gains (experiment c→g) under the moderate setting).

B.4 Instance depth Uncertainty

We also investigate the impact brought by the uncertainty. The uncertainty can stabilize the training process as it allows the network to learn more reasonable objects. We can observe that this strategy brings improvements for both the coupled and the decoupled manner (*e.g.*, 16.25/11.20→17.87/12.64 (experiment f→h) and 18.72/13.24→20.06/13.91 (experiment g→i) under the moderate setting). Intuitively, once the network has learned the depth uncertainty, we can use it to represent the confidence of instance depth estimation. This depth confidence can also express the 3D location confidence, which is used to combine

with 2D detection confidence as the final 3D detection confidence. Based on this, the performance is further boosted (*e.g.*, 1.81/1.49 gains (experiment h→j) and 1.47/1.66 gains (experiment i→k) under the moderate setting).

B.5 Instance Depth Aggregation

In previous experiments, for the grid design, the final instance depth is the average value in the RoI grids. Given that every instance depth estimation in each grid has the corresponding uncertainty, we can use uncertainties in the grids to adaptively obtain the final instance depth. Thus we perform experiments r→t (for the original coupled manner) and s→u (for our decoupled manner). Interestingly, we can observe that this depth aggregation in the grid cannot bring significant improvements to the original coupled manner (only 0.1/0.03 gains under the moderate setting). This is because all direct instance depth estimates are very close. By contrast, for our decoupled manner, we obtain obvious improvements (1.23/0.55 gains under the moderate setting). It indicates that different parts of the objects can produce different features, to conduct different instance depth predictions with associated uncertainties. This experiment further validates the presence of the coupled nature in instance depth and demonstrates the effectiveness of our method.

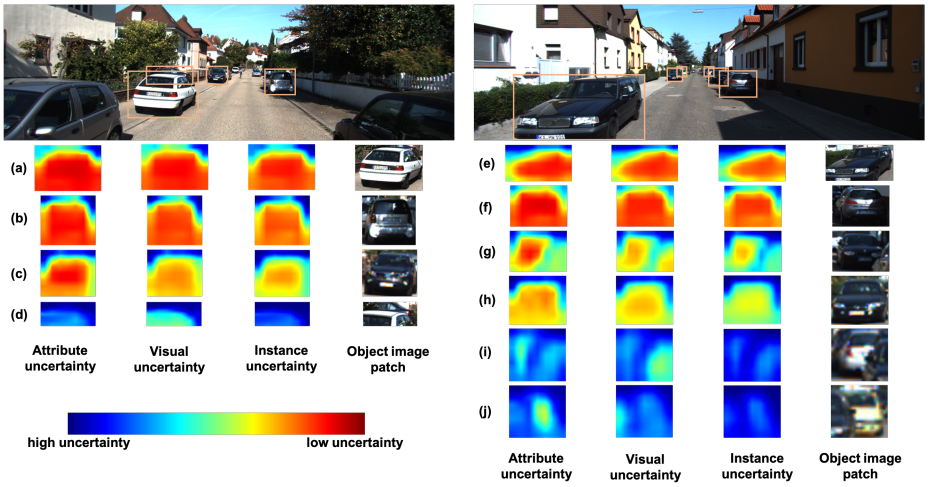

Fig. 1. Attribute depth and visual depth uncertainty. For simple objects, they show similar distributions. While they have different characteristics for difficult objects. Please refer to Section C.1 for more discussion. Best viewed in color with zoom in.

C Qualitative Results

C.1 Analysis on Attribute Depth and Visual Depth Uncertainty

To better understand how the attribute depth and visual depth work, we illustrate their uncertainties on two typical scenes, as shown in Figure 1. • First, we can easily see that all background areas of RoIs have high uncertainties. It is expected because the background area does not have important clues to the estimation of foreground objects, and its visual depths and attribute depths are hard to predict. • Second, regarding simple objects such as objects (a, b, e, f), their attribute depth uncertainties and visual depth uncertainties have similar distributions, since both two types of depths are easy to estimate. • Third, for the far objects (c, h), we know that the visual depth is less confident than the attribute depth. This is reasonable as the object texture is obvious, and the network can be confident in estimating attribute depths. By contrast, predicting absolute visual depths for far objects is difficult. • Finally, concerning the occluded objects (d, g, i, j), we can observe that visual depths and attribute depths have different interest areas. Visual depths mainly focus on closer objects because of the prediction simplicity. In contrast to visual depths, attribute depths are more interested in the target object, which even is heavily occluded (objects (i, j)). When the object is nearly invisible, attribute depths will give all areas high uncertainties (object (d)).

C.2 More Qualitative Results of 3D Box Predictions

We provide more qualitative results in Figure 2. For better visualization, the 3D box predictions are drawn in the 3D space and on the RGB image simultaneously. We can see that our method works well in most scenes, which proves its effectiveness and robustness.

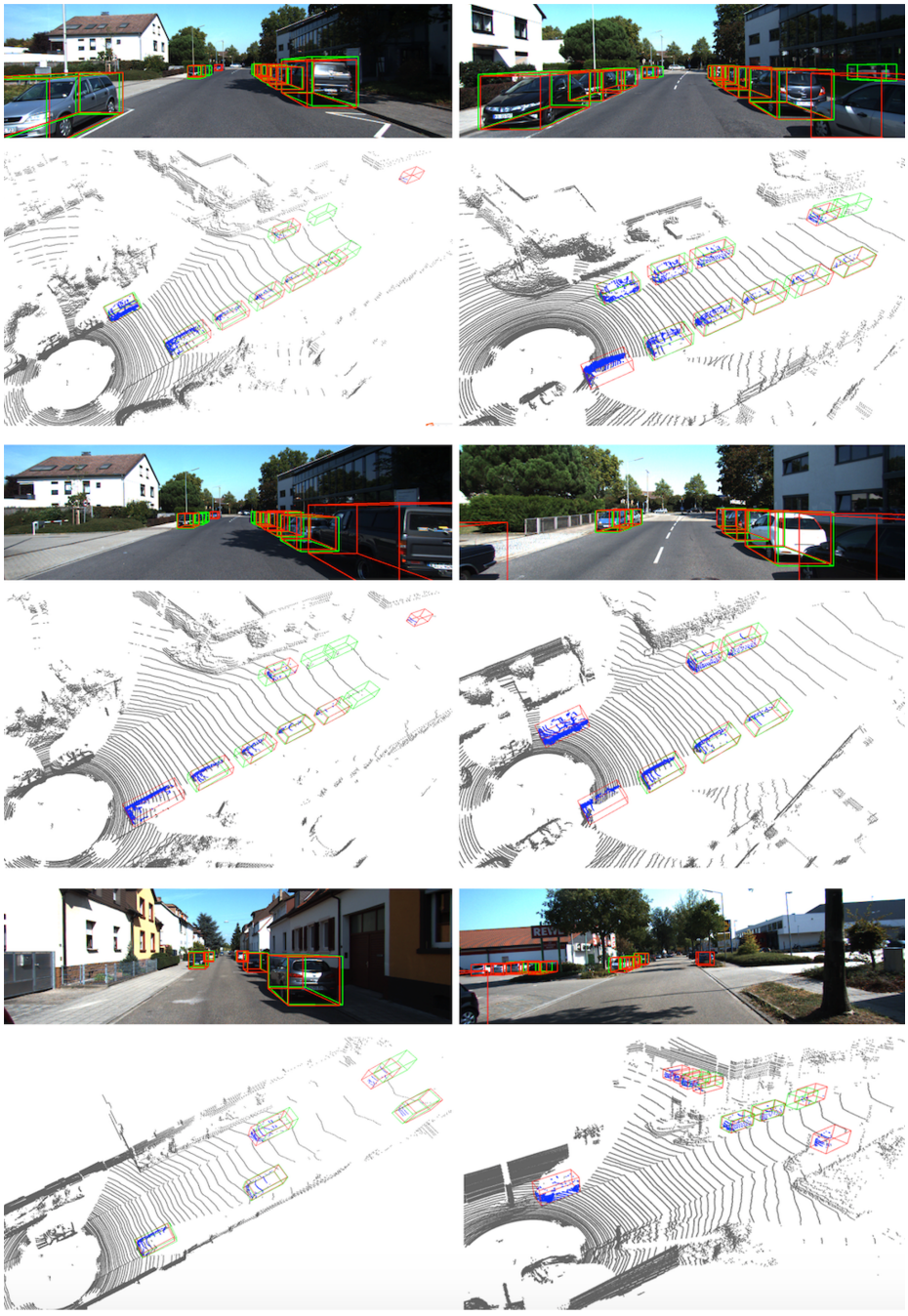

Fig. 2. Qualitative results on KITTI *val* set. Red: ground-truth 3D boxes; Green: our predictions. We can observe that most 3D box predictions are quite accurate. The LiDAR point clouds are only used for visualization. Best viewed in color with zoom in.

C.3 Failure Cases and Discussion

This paragraph aims to investigate failure cases in our method. We show some examples in Figure 3. These failure cases can be roughly divided into four categories, *i.e.*, faraway, occluded, truncated, and poor illumination. Faraway objects are very difficult to be precisely predicted due to the dramatically decreased information along with the depth in monocular imagery. Some occluded and truncated objects provide few valuable clues on the image to infer their locations and orientations. As for objects in poor illumination, their texture features are weakened, thus bringing difficulty in estimating their 3D boxes. In future works, we will further explore these failure cases, to mitigate their adverse impacts.

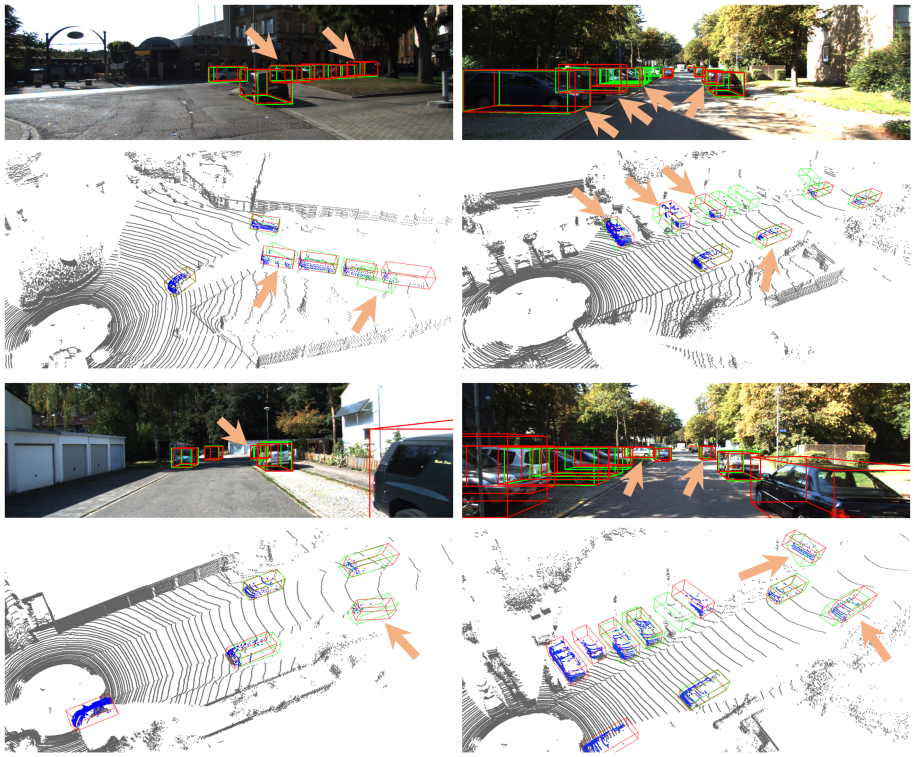

Fig. 3. Failure cases on KITTI *val* set. Red: ground-truth 3D boxes; Green: our predictions. We use arrows to indicate failure cases. We can see that failure objects usually are faraway, occluded, truncated, or suffer from poor illumination. The LiDAR point clouds are only used for visualization. Best viewed in color with zoom in.

References

1. Lian, Q., Li, P., Chen, X.: Monojsg: Joint semantic and geometric cost volume for monocular 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1070–1079 (2022)
2. Ma, X., Liu, S., Xia, Z., Zhang, H., Zeng, X., Ouyang, W.: Rethinking pseudo-lidar representation. arXiv preprint arXiv:2008.04582 (2020)
3. Reading, C., Harakeh, A., Chae, J., Waslander, S.L.: Categorical depth distribution network for monocular 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8555–8564 (2021)
4. 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2446–2454 (2020)
5. Wang, L., Zhang, L., Zhu, Y., Zhang, Z., He, T., Li, M., Xue, X.: Progressive coordinate transforms for monocular 3d object detection. Advances in Neural Information Processing Systems **34** (2021)