MonoDTR: Monocular 3D Object Detection with Depth-Aware Transformer Supplementary Material

Kuan-Chih Huang[1] Tsung-Han Wu[1] Hung-Ting Su[1] Winston H. Hsu[1,2]
[1] National Taiwan University [2] Mobile Drive Technology

A. Depth-Aware Transformer

Transformer architecture. The detailed architecture of depth-aware transformer (DTR) is shown in Figure 1. The encoder aims to generate the encoded context-aware features, while the decoder produces the fused feature from context- and depth-aware features through the multiple self-attention layers. Besides, we supplement two features with the proposed depth positional encoding (DPE) before passing them to the transformer, enabling better 3D reasoning.

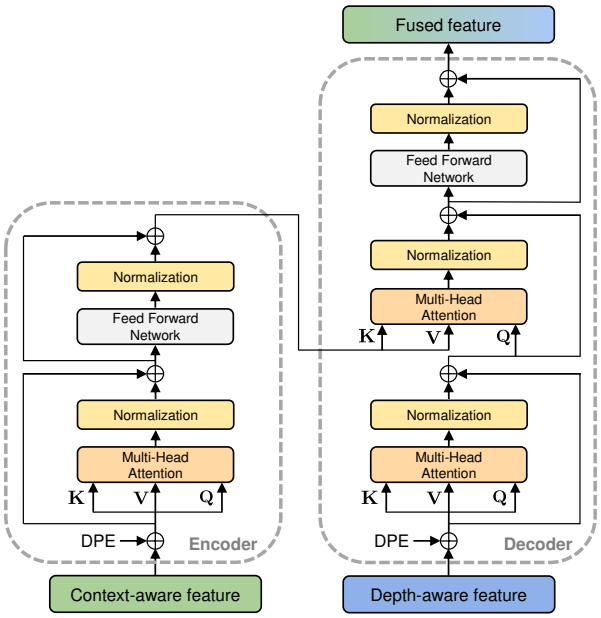

Figure 1. **The architecture of depth-aware transformer (DTR).** DPE is the depth positional encoding proposed in the main paper.

Effectiveness of linear attention. Table 1 shows the results of different self-attention layers on the KITTI dataset, where we can observe that applying linear attention [4] can achieve almost 4 $\times$ faster than vanilla self-attention [6] with comparable performance. Thus, we adopt the linear attention [4] in our transformers for real-time applications.

Attention	Time	AP_{3D}@IoU=0.7			AP_{BEV}@IoU=0.7		
		Easy	Mod.	Hard	Easy	Mod.	Hard
vanilla SA [6]	136 ms	24.38	18.39	16.35	31.57	24.51	21.40
linear SA [4]	37 ms	24.52	18.57	15.51	33.33	25.35	21.68

Table 1. **Comparison of different self-attention mechanisms** on the KITTI validation set for Car category. We follow the same setting and device as in the main paper for running time measurement. Note that 'SA' is the multi-head self-attention. The metric is AP_{40}.

Disc. Method	AP_{3D}@IoU=0.7			AP_{BEV}@IoU=0.7		
	Easy	Mod.	Hard	Easy	Mod.	Hard
UD	23.22	17.67	14.80	31.75	24.32	20.08
SID	23.89	18.10	15.22	32.19	24.76	21.36
LID	**24.52**	**18.57**	**15.51**	**33.33**	**25.35**	**21.68**

Table 2. **Comparison of different discretization methods for auxiliary depth supervision** on the KITTI validation set for Car category. The metric is AP_{40}.

B. Auxiliary Depth Supervision

Depth ground truth generation. We project the LiDAR signals into the image plane to generate the sparse ground truth depth map. Then we apply linear-increasing discretization (LID) [5] method to convert continuous depth d to discretized depth bins. The LID is defined as follows:

$$d = d_{\min} + \frac{d_{\max} - d_{\min}}{D(D+1)} \cdot i(i+1), \; i = \{1, ..., D\}, \quad (1)$$

where i is the depth bin index. The number of depth bins D is set as 96, and the range of depth $[d_{\min}, d_{\max}]$ is set as [1, 80]. Note that the pixels with the depth value outside the range will be marked as invalid and not used for optimization during training.

Different discretization methods. In Table 2, we investigate the effectiveness of different discretization methods for depth auxiliary supervision. In addition to the LID method, the continuous depth can be discretized using uniform discretization (UD) with fixed bin size: $\frac{d_{\max}-d_{\min}}{D}$, or spacing-increasing discretization (SID) [3] with the increasing bin size in the log space. It can be observed that using the LID strategy can achieve better performance, so we apply it as our discretization method.

Figure 2. **Qualitative results on the KITTI validation set for multi-class 3D object detection.** We utilize orange, blue, and green colors to indicate car, pedestrian, and cyclist categories, respectively.

C. Results on nuScenes Dataset

Table 3 shows the experimental results of deploying our proposed approach on nuScenes [1] val set. Under the same configurations (*e.g.*, backbone and training schedule), our model achieves better performance than two 3D object detection baselines (FCOS3D [7], and PGD [8]), which demonstrates the effectiveness of our approach.

Method	NDS ↑	mAP ↑
FCOS3D [7]	37.7	29.8
PGD [8]	39.3	31.7
Ours	**40.1**	**33.8**

Table 3. **Detection performance on nuScenes val set.** We build our approach based on FCOS3D [7]. The experiments are conducted under the same training settings (trained for 12 epochs). The results of baselines are taken from MMDetection3D [2].

D. Qualitative Visualization

More visualization results. In Figure 2, we provide some qualitative results on the KITTI dataset for multiple-category predictions. In Figure 3, we show the qualitative comparisons of the baseline (without proposed depth-aware modules) and our MonoDTR (full model). It can be observed that our MonoDTR can generate higher quality bounding boxes benefit from the aid of depth cues.

Failure case. We show a representative failure case in Figure 4. The lower-quality 3D bounding box is caused by the inaccurately predicted object depth, which is typical in most monocular 3D object detection tasks.

E. Broader Impacts

Our work aims to develop the monocular 3D object detection approach for autonomous driving. The proposed model may generate inaccurate object depth prediction,

Figure 3. **Qualitative comparison on the KITTI validation set** for the car category. The purple boxes in the image and BEV plane represent the predictions from MonoDTR. The green and pink boxes on BEV are the ground truth and the predictions from baseline (our full model without proposed depth-aware modules), respectively. Best viewed in color and zoomed in.

Figure 4. **Failure case.** The purple and green boxes represent the predictions from MonoDTR and ground truth, respectively. The failure case is caused by the inaccurate object center depth estimation.

leading to incorrect downstream decision-making and potential traffic accidents. Furthermore, we provide a new perspective of leveraging learned depth-aware features to assist monocular 3D object detection. Although considerable progress has been made with our proposed lightweight depth-aware feature extraction module, we believe it is worth further exploring how to learn depth-aware features to effectively improve detection performance.

References

[1] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 2

[2] MMDetection3D Contributors. MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. https://github.com/open-mmlab/mmdetection3d, 2020. 2

[3] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Bat-

manghelich, and Dacheng Tao. Deep Ordinal Regression Network for Monocular Depth Estimation. In *CVPR*, 2018. 1

[4] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *ICML*, 2020. 1

[5] Yunlei Tang, Sebastian Dorn, and Chiragkumar Savani. Center3d: Center-based monocular 3d object detection with joint depth understanding. *arXiv preprint arXiv:2005.13423*, 2020. 1

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 1

[7] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. FCOS3D: Fully convolutional one-stage monocular 3d object detection. In *ICCV Workshops*, 2021. 2

[8] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. Probabilistic and Geometric Depth: Detecting objects in perspective. In *CoRL*, 2021. 2