

# Deep Line Encoding for Monocular 3D Object Detection and Depth Prediction: Supplementary Material

Ce Liu<sup>1</sup>

ce.liu@vision.ee.ethz.ch

Shuhang Gu<sup>2</sup>

shuhangu@gmail.com

Luc Van Gool<sup>1,3</sup>

vangool@vision.ee.ethz.ch

Radu Timofte<sup>1,4</sup>

radu.timofte@vision.ee.ethz.ch

<sup>1</sup> Computer Vision Lab

ETH Zurich

Switzerland

<sup>2</sup> University of Sydney

Australia

<sup>3</sup> ESAT

KU Leuven

Belgium

<sup>4</sup> University of Würzburg

Germany

## 1 Overview

In this supplementary document, we provide  $AP_{BEV}$  scores for the configurations ablation study in the 3D object detection task. As for the depth prediction task, we present results on the KITTI validation set [1] and the NYU Depth V2 test set [2]. Moreover, we present detailed qualitative comparison to demonstrate the effectiveness of deep line encoding.

## 2 $AP_{BEV}$ in Ablation Study

$AP_{BEV}$  is a slightly easier metric than  $AP_{3D}$ , because the vertical component is ignored when computing the bound box overlap. As shown in Table 1, we observe a significant improvement by adding the coordinate map or the line vector to the VisualDet3D [3], which is consistent with the results in  $AP_{3D}$ . When combining the above two techniques, the performance for easy case is further improved slightly.

## 3 Depth Prediction on Different Datasets

We present experiment results on the KITTI validation set [1] in Table 2. After adding the deep line encoding to the backbone, the performance of GAC [4] is improved under all the metrics, demonstrating the effectiveness of our design.

We further evaluate the deep line encoding on the NYU Depth V2 data set [2]. It consists of 120,000 images captured in indoor scenes. Following the official split, we use 249 scenes

Configuration	Coordinate	Line Vector	$AP_{BEV}$		
			Easy	Moderate	Hard
a			29.70	20.98	16.20
b	✓		33.12	22.54	16.81
c		✓	33.25	<b>22.64</b>	<b>17.01</b>
d	✓	✓	<b>33.65</b>	22.34	16.94

Table 1:  $AP_{BEV}$  scores for different configurations ablation study.

Method	SILog	sqErrorRel	absErrorRel	iRMSE
GAC [9]	9.64	1.22	6.43	7.33
<b>+ Line Encoding (ours)</b>	<b>9.36</b>	<b>1.13</b>	<b>6.14</b>	<b>7.21</b>

Table 2: Performance on KITTI validation set for depth prediction.

for training and 215 scenes (654 images) for testing. In the training set, 24,231 images and depth maps are associated and sampled using timestamps by even-spacing in time. We train and test on the center cropping proposed by Eigen *et al.* [10].

The network structure and the training hyper-parameters are the same as in KITTI data set, except that we removed the ground-aware convolution because it’s designed for autonomous driving scenarios.

NYU Depth V2 data set [11] is captured in indoor scenes by Microsoft Kinect. Thereby the camera poses with respect to the ground plane are more variable than in KITTI, and the simplified projection model might be inapplicable in a lot of images. However, we still observe an improvement in Table 3 with deep line encoding. The improvement shows that the line information is beneficial even in indoor scenes.

Method	SILog	sqErrorRel	absErrorRel	iRMSE
GAC [9]	13.72	3.79	13.15	8.67
<b>+ Line Encoding (ours)</b>	<b>12.71</b>	<b>3.20</b>	<b>12.44</b>	<b>8.22</b>

Table 3: Performance comparison on NYU Depth V2 test set.

## 4 Qualitative Comparison for 3D Object Detection

We present qualitative results in Figure 1 and 2 on KITTI validation set [12] for 3D object detection. A notable observation from Figure 1 is the predicted depth of distant cars is more accurate with the help of deep line encoding. In Figure 1 (a) and (b), the predicted bounding boxes from deep line encoding have larger overlap with the ground truth. In Figure 1 (c) and (d), VisualDet3D [13] even filters out the prediction for distant cars due to low confidence.

Figure 2 shows the cases where there are slopes or steps. In Figure 2 (a) and (b), deep line encoding helps to predict more accurate bounding boxes for cars on slopes. When there are steps, as shown in Figure 2 (c) and (d), more cars are located correctly with the help of deep line encoding.

## 5 Qualitative Comparison for Depth Prediction

We present qualitative results in Figure 3. The first column shows the input image. The second and third columns show the error map of deep line encoding and GAC [9] respectively. The error map indicates the `absErrorRel` for the predicted depth map. Following Uhrig *et al.* [9], correct estimates are in blue while wrong estimates are in red color tones.

The first four rows show the cases where there are slopes. With deep line encoding, the prediction for guardrails, road surface, and trees is more accurate. In the next four rows, although the road is approximately horizontal, we notice the deep line encoding can help to refine the prediction for vertical structures, such as the wall. The rest of the examples show our method yields better accuracy on the cars and bushes that are on the steps.

## References

- [1] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.
- [2] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012.
- [3] Yuxuan Liu, Yuan Yixuan, and Ming Liu. Ground-aware monocular 3D object detection for autonomous driving. *IEEE Robotics and Automation Letters*, 6(2):919–926, 2021.
- [4] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pages 746–760. Springer, 2012.
- [5] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *2017 international conference on 3D Vision (3DV)*, pages 11–20. IEEE, 2017.

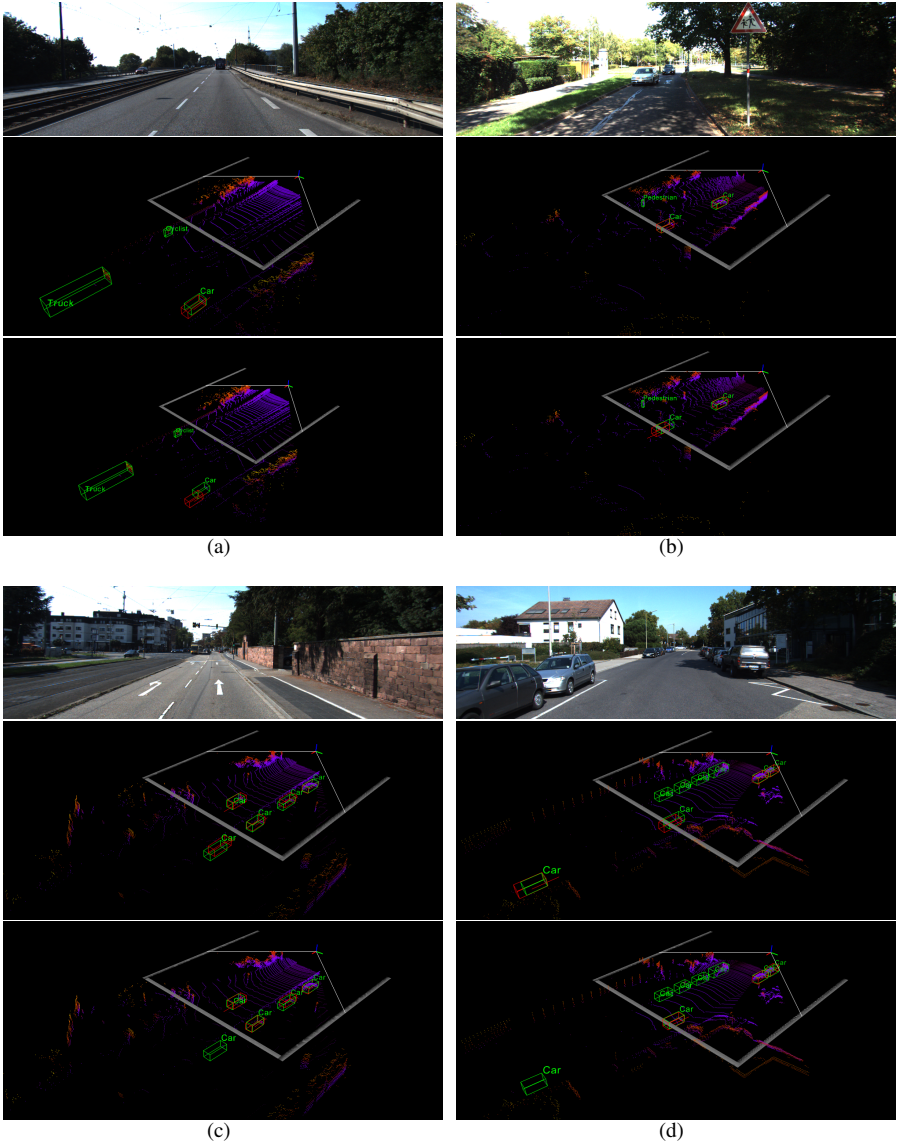


Figure 1: Qualitative comparison in examples of distant cars. From top to bottom: input image, predictions from deep line encoding and VisualDet3D [8]. The green and red boxes represent the ground truth and prediction respectively.



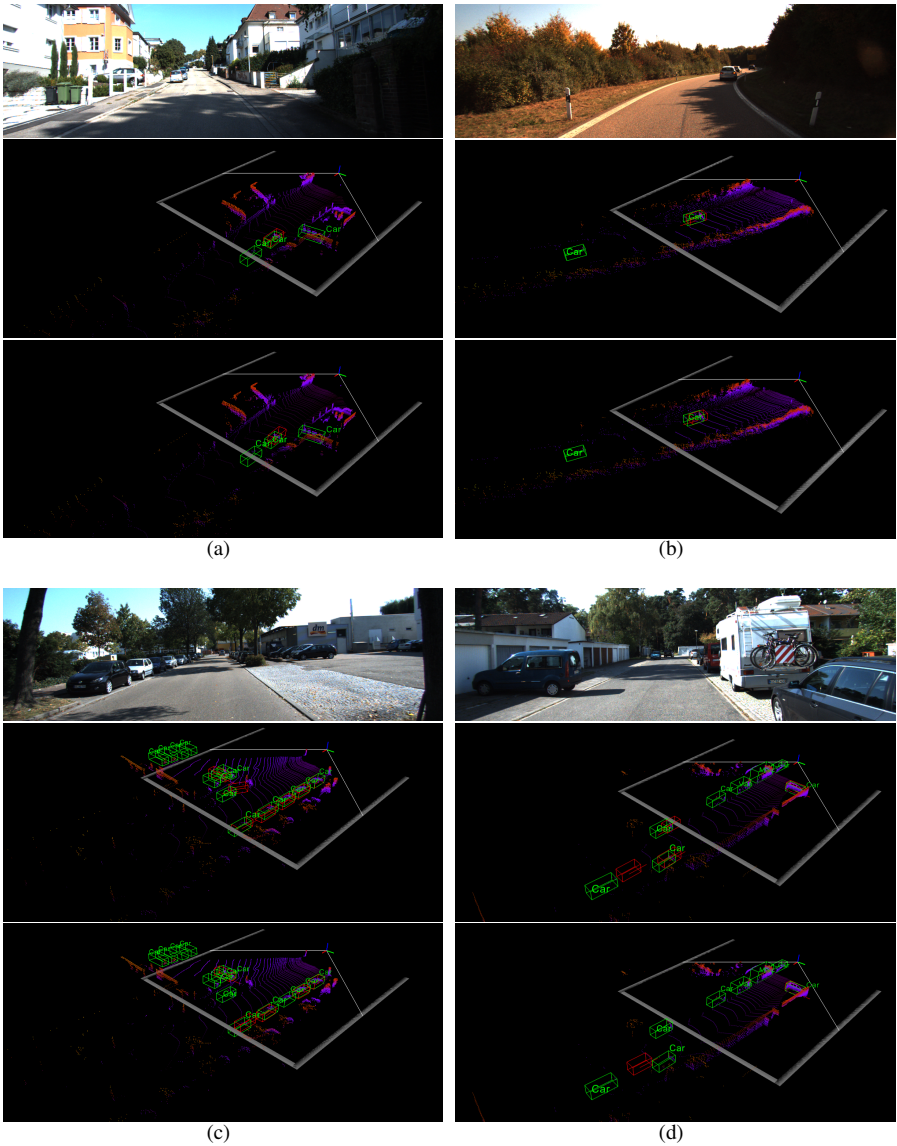


Figure 2: Qualitative comparison in examples of slopes and steps. From top to bottom: input image, predictions from deep line encoding and VisualDet3D [8]. The green and red boxes represent the ground truth and prediction respectively.



Figure 3: Qualitative comparison to GAC [9] on the validation set of KITTI single-image depth prediction benchmark.