Appendix of "Semi-Supervised Monocular 3D Object Detection by Multi-View Consistency"

In the appendix, we first present the details of generating the relative depth for structure from motion and then provide the comparison between 2D and 3D object detection in the semi-supervised training framework.

A Details of the relative depth module

As illustrated in the main paper, we utilize the bounding boxes parameters $b \in \mathcal{R}^{3\times 8}$ to generate the relative depth for each point $p = [u, v]^T$ in the image. We first define the camera intrinsic as:

$$K = \begin{pmatrix} f_u & 0 & c_u \\ 0 & f_v & c_v \\ 0 & 0 & 1 \end{pmatrix}. \tag{1}$$

Then we can represent the ray emit from the camera origin to point p as a vector with: $\vec{op} = [\frac{u-c_u}{f_u}, \frac{v-c_v}{f_v}, 1]$. For the surface vector in the bounding box i,

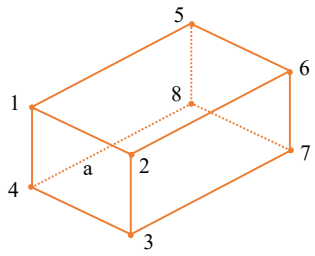

Fig. 1. Visualization of the notation in the 3D bounding box

we present an example to calculate the vector of the plane a with $b\vec{b^{ia}}$. The points in the bounding box is denotes as $\{p^{b_i}\}_{i=1}^{8}$. For plane a, it can be represented as the cross product between the vector of $p^{b_1}\vec{p^{b_2}}$ and $p^{b_1}\vec{p^{b_3}}$:

$$b\vec{b^{ia}} = p^{b_1}\vec{p^{b_2}} \times p^{b_1}\vec{p^{b_3}} = (p^{b_2} - p^{b_1}) \times (p^{b_3} - p^{b_1}). \tag{2}$$

With the vector of each surface, the intersections can be computed based on the method described in the main paper. Based on the cube-shaped assumption [1], the basic relative depth can be generated as described above. And then we add a parallel head to estimate the surface-to-plane offset ΔZ to fill the gap between the relative depth and the true surface depth. This offset is added to the depth calculated in the relative depth module and regenerates the final 3D location for each pixel.

Table 1. Experimental results of 2D and 3D detection performance on the KITTI validation set. The metrics of $AP|_{R40}$ with IoU threshold=0.7 on three difficulties (easy, moderate, and hard) are reported. We randomly sample 10%, 50%, and 100% data from the KITTI training set as the labeled split.

Setting	10%			50%			100%		
	Easy	Mod	Hard	Easy	Mod	Hard	Easy	Mod	Hard
2D detection	91.14	81.15	72.75	96.13	89.41	83.34	96.46	90.75	85.10
3D detection	10.13	7.25	6.24	18.52	14.56	12.53	21.99	16.32	14.48

B Comparison between 2D and 3D object detection

In this section, we present the 2D and 3D detection performance of the baseline detector (MonoDLE) on the KITTI validation set in Table 1.

Compared to 3D detection task, the detector shows strong ability in the 2D detection tasks, even when the labeled data is scarce. However, the detector performs much worse in the 3D detection task. Since MonoDLE utilizes the same classification head for the 2D and 3D tasks, the performance gap can indicate that the main weakness in 3D detection is the 3D bounding boxes regression. Furthermore, the strong performance in 2D detection also helps us for matching 3D bounding boxes in different images.

B.1 Effectiveness of the photometric loss

To evaluate the effectiveness of the designed photometric loss, we visualize the corresponding loss landscape to check if it can represent the bounding box localization error. Specifically, we first uniformly sample the object depth with an offset of [-5m, 5m] around the ground truth. Then we generate the corresponding bounding boxes with the ground truth dimension, orientation, projection location and the sampled depth. We display the photometric loss computed by the relative depth module with and without the predicted offset. Figure 2 shows the landscape with averaging from different instances on the KITTI dataset. Ideally, the photometric loss should achieve the minimum value when the depth offset is 0. The shape of the loss landscape also should be smooth. As visualized, the loss without offset can achieve minimum when the depth error is near to 0. And with adding the offset, its loss landscape can better match with what we expected.

C Ablation study in the post-processing setting

In this section, we provide the experimental results of applying our relative depth module to the post-processing stage in Stereo R-CNN [1]. In Stereo R-CNN, the post-processing stage utilizes the photometric loss with the cube-shape assumption to search optimal object depth for 3D bounding box detection. In our work, we first start from the cube-shaped assumption in Stereo R-CNN, and

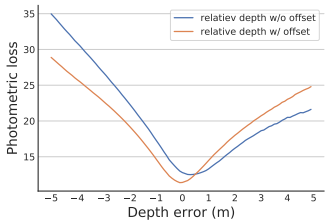

Fig. 2. Visualization of the loss landscape with different ways in computing object-level photometric loss.

progressively learn the finer object shape (surface-to-plane offset) by our semi-supervised training framework. Compared to Stereo R-CNN, the photometric loss in our framework is further optimized with object shape, which can better reflects the depth error. In Table 2, we display the experimental results of applying our relative depth component in Stereo R-CNN for searching optimal object depth.

Table 2. Experimental results of stereo-based detector on the KITTI dataset. The car class with 3D $AP|_{40}$ metric is reported.

Setting	Easy	Mod	Hard
Stereo R-CNN (cube-shape)	47.31	29.23	25.31
+ our relative depth (finer shape)	**51.50**	**32.96**	**27.93**

References

1. Li, P., Chen, X., Shen, S.: Stereo r-cnn based 3d object detection for autonomous driving. In: CVPR (2019)