Supplementary Material

A. Overview

This document provides additional technical details, experimental results, theoretical analysis, and qualitative results to the main paper. Specifically, in Section B, we provide more details on the implementation of the depth estimation sub-task, and Section C shows the details and ablations about the proposed IoU oriented loss. Section D provides more discussion which is omitted in the main paper. Finally, Section E presents more visual results.

B. Depth Estimation

Uncertainty modeling. Following [18, 19, 10], we model the heteroscedastic aleatoric uncertainty in the depth estimation sub-task. Specifically, we simultaneously predict the depth $\mathbf{d}$ and the standard deviation σ (or variance σ^2):

$$[\mathbf{d}, \sigma] = f^{\mathbf{w}}(\mathbf{x}), \tag{7}$$

where $\mathbf{x}$ is the input data and f is a convolutional neural network parametrised by the parameters $\mathbf{w}$. Then, we fix a Laplace likelihood to model the uncertainty, and the loss for the depth estimation sub-task can be formulated by:

$$\mathcal{L} = \frac{\sqrt{2}}{\sigma}||\mathbf{d} - \mathbf{d}^*||_1 + \log \sigma, \tag{8}$$

where $||\cdot||_1$ denotes the L1 norm and $\mathbf{d}^*$ is the ground truth value for depth $\mathbf{d}$. Similarly for the Gaussian likelihood:

$$\mathcal{L} = \frac{1}{2\sigma^2}||\mathbf{d} - \mathbf{d}^*||_2 + \frac{1}{2}\log \sigma^2, \tag{9}$$

where $||\cdot||_2$ denotes the L2 norm (please refer to [19] for the derivation of Equation 8 and Equation 9). Note that the uncertainty modeling is not claimed as our contribution.

Experimental results. First, from Figure 6 and Table 9, we can find that uncertainty-based estimation improves the accuracy of depth map, thereby improving the overall performance of monocular 3D detection. Second, the experimental result also show that modeling uncertainty based on the Laplace distribution (all models in the main paper adopted this setting) is more suitable for our task than Gaussian distribution.

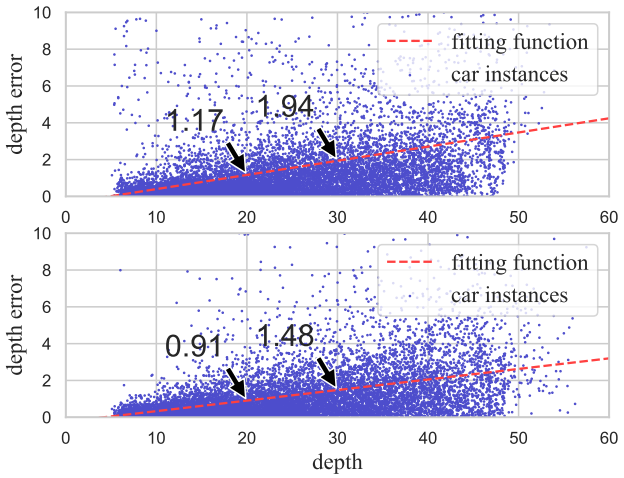

Figure 6: **Errors of depth estimation.** We show the errors of depth estimation as a function of the depth (x-axis) for the plain scheme (*top*) and the uncertainty aware scheme based on the Laplace likelihood (*bottom*).

uncert.	Easy	Mod.	Hard
-	18.56 / 13.18	14.24 / 10.15	12.13 / 8.45
Gaussian	18.68 / 13.20	14.22 / 10.41	12.08 / 8.69
Laplace	20.29 / 14.51	16.15 / 11.12	14.07 / 9.97

Table 9: **Analysis for the designs of depth estimation.** Metrics are AP_{40} of the Car category for BEV/3D detection tasks.

C. IoU Oriented Loss

C.1. Proof of Proposition

This section provides the proof of the following proposition, which is used in Equations 5 and 6 for IoU oriented optimization in Section 3.6.

Proposition. Suppose all predicted items except the 3D sizes (h, w, l) are completely correct, the contribution ratio of each predicted side to the 3D IoU $\frac{\partial IoU}{\partial h} : \frac{\partial IoU}{\partial w} : \frac{\partial IoU}{\partial l}$ can be approximated to $\frac{1}{h} : \frac{1}{w} : \frac{1}{l}$.

Proof. Given the above conditions, the 3D IoU metric can be formulated as:

$$IoU = \frac{\prod_{i\in\{h,w,l\}} \min(i, i^*)}{h \times w \times l + h^* \times w^* \times l^* - \prod_{i\in\{h,w,l\}} \min(i, i^*)}, \tag{10}$$

where (h^*, w^*, l^*) denotes the ground truth of 3D size (h, w, l). With the different relationship between the prediction and the ground truth of the 3D size, we can obtain the following cases:

Case 1: If $h \leq h^*$, $w \leq w^*$, and $l \leq l^*$, the Equation 10 can be simplified as:

$$IoU = \frac{h \times w \times l}{h^* \times w^* \times l^*}, \tag{11}$$

and we further compute the partial derivative of 3D IoU with respect to the variable h as

$$\frac{\partial IoU}{\partial h} = \frac{w \times l}{h^* \times w^* \times l^*}, \tag{12}$$

where $\frac{\partial IoU}{\partial h}$ represents the partial derivative of 3D IoU with respect to the variable h, analogically for $\frac{\partial IoU}{\partial w}$ and $\frac{\partial IoU}{\partial l}$. Then, combining the derivative of 3D IoU with respect to h, w, and l, the contribution ratio of each predicted side can be given as:

$$\frac{\partial IoU}{\partial h} : \frac{\partial IoU}{\partial w} : \frac{\partial IoU}{\partial l} = \frac{1}{h} : \frac{1}{w} : \frac{1}{l}. \tag{13}$$

Case 2: If $h > h^*$, $w > w^*$, and $l > l^*$, the Equation 10 can be simplified as:

$$IoU = \frac{h^* \times w^* \times l^*}{h \times w \times l}, \tag{14}$$

and similar to Equation 12 and 13, we can derive the same conclusion as *Case 1*.

Case 3: If $h > h^*$, $w \leq w^*$, and $l \leq l^*$, then we represent the 3D IoU as:

$$IoU = \frac{h^* \times w \times l}{h \times w \times l + h^* \times w^* \times l^* - h^* \times w \times l}. \tag{15}$$

By calculating the derivative of 3D IoU with respect to h, w, and l respectively, we can get the contribution ratio of each predicted side:

$$\frac{\partial IoU}{\partial h} : \frac{\partial IoU}{\partial w} : \frac{\partial IoU}{\partial l} = \frac{w \times l}{h^* \times w^* \times l^*} : \frac{1}{w} : \frac{1}{l}. \tag{16}$$

Case 4: If $h > h^*$, $w > w^*$, and $l \leq l^*$, similarly, we can get the IoU formulation as:

$$IoU = \frac{h^* \times w^* \times l}{h \times w \times l + h^* \times w^* \times l^* - h^* \times w^* \times l}. \tag{17}$$

Similar to previous steps, the formulation of each side's contribution rate to the 3D IoU is given as:

$$\frac{\partial IoU}{\partial h} : \frac{\partial IoU}{\partial w} : \frac{\partial IoU}{\partial l} = \frac{1}{h} : \frac{1}{w} : \frac{h^* \times w^* \times l^*}{h \times w \times l \times l}. \tag{18}$$

The other cases are similar to *Case 3* and *Case 4*. When $h \approx h^*$, $w \approx w^*$, and $l \approx l^*$, we can get the Equation 5 used in the main paper.

C.2. Experiments

We report the improvement introduced by the proposed loss function in the main paper. To further validate the effectiveness of it, we also implement the 3D GIoU loss [41] for reference. Specifically, we add the 3D GIoU loss as a regularization item as in [41], investigating different weights considered in our baseline model, and the AP_{40} of cars on the moderate setting on KITTI *validation* set (Table 10) show that our IoU oriented optimization improves accuracy but 3D-GIoU with different weights does not.

Baseline	GIoU (w=0.5)	GIoU (w=1)	GIoU (w=5)	Ours
11.12	10.17	10.19	8.48	11.74

Table 10: **Ablation study** for the proposed loss function and 3D GIoU loss on the KITTI *validation* set. Metric is AP_{40} of the Car category under moderate setting.

Range	Easy	Moderate	Hard	UnKnown	Total
$[5m, 15m]$	2,131	1,428	963	1,457	5,979
$[10m, 20m]$	2,639	1,840	1,670	558	6,707

Table 11: **Data distribution** for the car samples located in $[5m, 15m]$ and $[10m, 20m]$. The data is collected from the KITTI *trainval* set.

D. Performance for the Close Objects

The Figure 1 in the main paper provides lots of insights to us. Except for the observations analyzed in the main paper, we also found that the performance degrades for the very close object. Here we provides our analysis for this. In particular, there are three main reasons in total. a) The close-range objects tend to have larger center misalignment (see Figure 3 for the statistics). b) The objects at closer ranges are usually more truncated, *e.g.* the red car (depth=3.7, truncation=0.88) and the black car (depth=6.2, truncation=0.34) in Figure 7. c) The training samples in the close range are fewer. For example, there are 5,979 cars in $[5m, 15m]$ and 6,707 cars in $[10m, 20m]$ on the KITTI *trainval* set, and the distribution for those samples are summarized in Table 11. Note that the KITTI annotate the difficulty of each samples according to its size of 2D bounding box, occlusion, and truncation. The instance with 'unKnown' tag usually means that it is extremely difficult to detect and is ignored in evaluation. With that in mind, the effective samples of those two ranges are 4,522 and 6,149. In summary, the low performance of the very close objects is caused by the limited training samples (c) and the large proportion of hard cases (a, b).

E. More Visualizations

E.1. Learned features

From Figure 4 in the main paper, we can see there is a misalignment between the center of the 2D bounding box and the projected center of the 3D object, especially for close objects (see Figure 3 and Figure 4). Accordingly, we propose our solution for this problem. Here we visualize the learned features of coarse center detection branch in Figure 7 to show the effectiveness of the proposed method. The qualitative results clearly show that using projected 3D center as ground truth can make the coarse center more accurate, thereby improving the localization accuracy.

Figure 7: **Qualitative comparison for the learned features of coarse center detection task on the KITTI *validation* set.** *Top:* the input image. *Middle:* the features of the coarse center detection branch supervised by 2D center. *Bottom:* the features of the coarse center detection branch supervised by projected 3D center. We use the write circle to highlight the ground truth projected 3D center for better comparison. Best viewed in color with zooming in.

E.2. Comparison of qualitative results

Visualizations in the image plane. We show more qualitative results of M3D-RPN (the best of all open-source standard monocular 3D detector) and the proposed method in Figure 8. We use red circle to highlight the main differences of each pair of images, and we can find that our method performs better than M3D-RPN for dense objects.

Visualizations in the 3D world space. We also visualize the 3D bounding boxes in the 3D world space for better presentation. As shown in Figure 9, the proposed model outputs better results than M3D-RPN, especially for the orientation estimation.

Representative failure case. We show a typical error pattern in monocular 3D object detection in Figure 10. We can observe that the projected 3D bounding boxes fit the object's appearance tightly in the image plane. However, from the visualization results in the 3D world space, this is a clear false positive because the depth is inaccurate (the outline of the object can be perceived through the point clouds, best viewed with zooming in). Note that this problem is common in the monocular 3D detection task, which suggests that depth estimation is a key factor restricting this task.

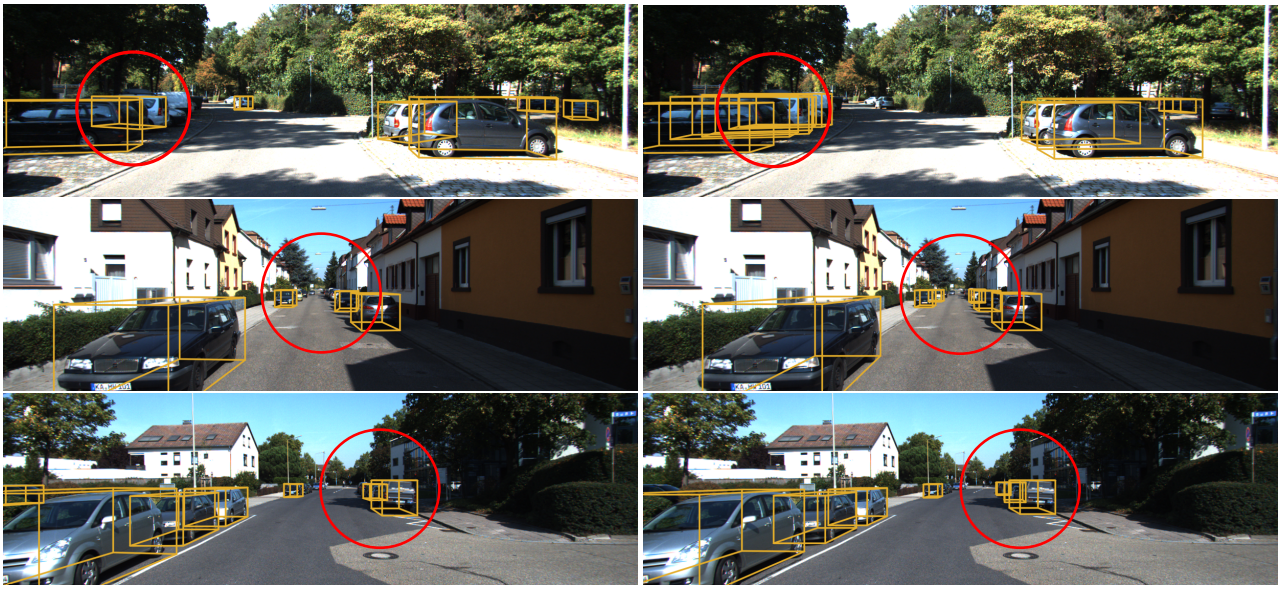

Figure 8: **Qualitative comparison on the KITTI *validation* set**. We visualize the 3D bounding boxes in the image plane. Results are from M3D-RPN (*left*) and our method (*right*).

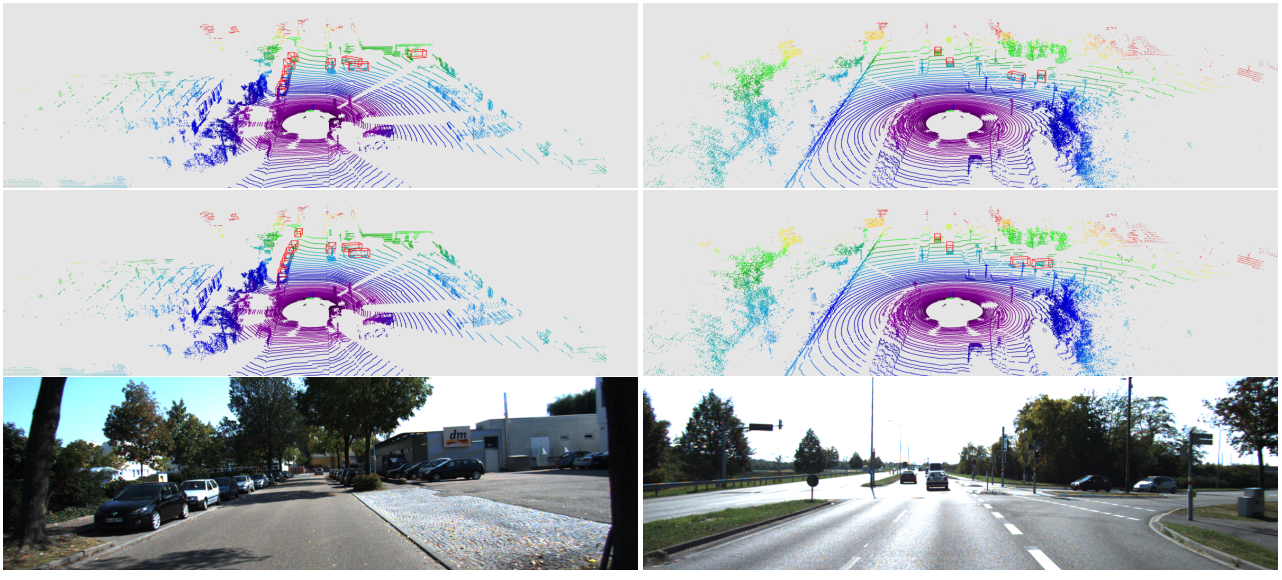

Figure 9: **Qualitative comparison on the KITTI *validation* set**. We visualize the 3D bounding boxes in the 3D world space. Results are from M3D-RPN (*top*) and our method (*middle*). We also show the corresponding 2D image (*bottom*) for reference. Best viewed in color with zooming in.

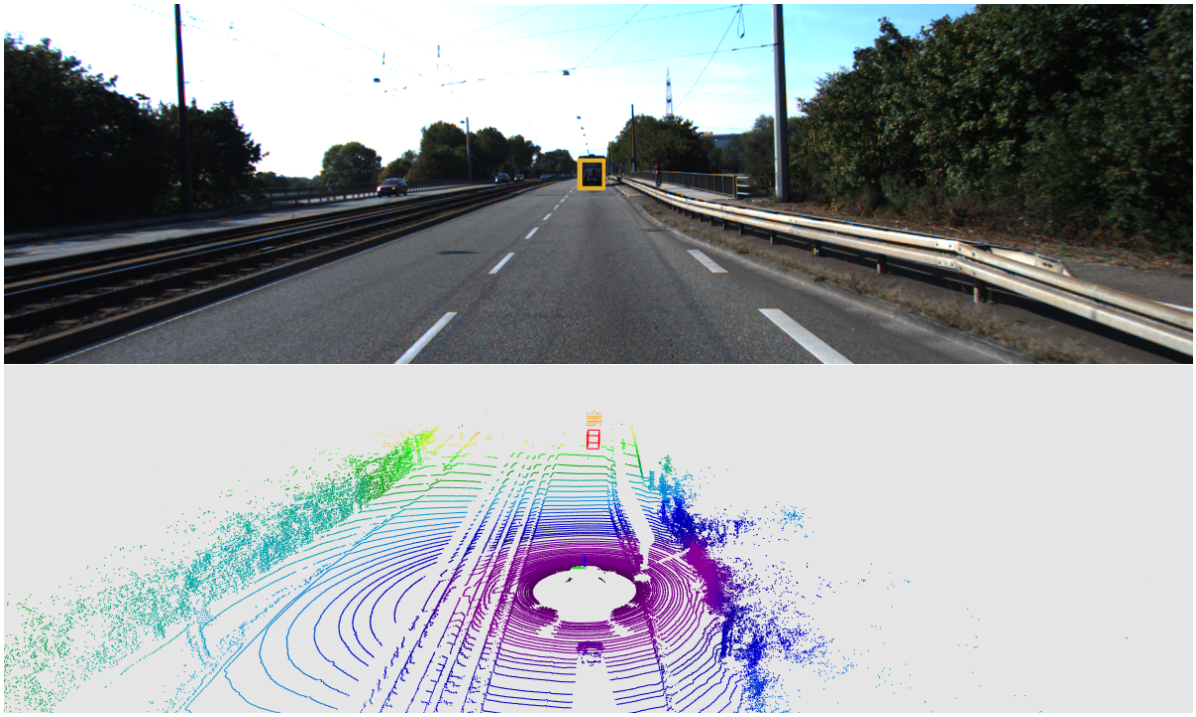

Figure 10: **Failure case**. We show a representative failure case which is caused by the inaccurate depth estimation.