Table 5: Comparison of utilizing different ground truth depth maps on the KITTI validation set.

Depth Map Type	AP 3D Car@IoU=0.7		
	easy	mod	hard
Dense Depth Map [10, 8]	21.37	16.08	13.14
Foreground Depth Map [9, 24]	**23.34**	**16.78**	**13.70**

A Details of Depth-aware Transformer

Previous works have proposed various depth-aware transformers [35, 27, 9, 8] and achieved great results. In our work, we follow the settings from [9], which contains a visual encoder, a depth encoder, and an object decoder that merges cross-view attention and cross-depth attention to generate depth-aware features. Inspired by [27], we first feed the learnable object queries $F_Q \in \mathbb{R}^{N \times C}$ to the transformer, where N is the number of underlying 3D bounding boxes. In each transformer layer, the queries interact with image features and depth embeddings through the cross-view attention followed by the cross-depth attention. After L layers, we obtain the depth-aware object queries.

B Implementation Details

We utilize the AdamW optimizer with a weight decay of 10^{-4} and an initial learning rate of 10^{-5}. To warm up the training process, we employ cosine warm-up scheduling for the first 5 epochs. Afterward, the learning rate is set to 2×10^{-4} and decays at epoch 125 with a rate of 0.1. We set $[d_{min}, d_{max}] = [10^{-3}, 60]$ and the number of depth bins $N_D = 80$ for the LID method. The input image size is 1280×384 for both training and testing. Following [3], we apply the photometric distortion, random horizontal flipping, and random cropping techniques. The random scaling factor is 0.4 while the random cropping factor is 0.25. Our model predicts up to $N = 50$ bounding boxes for each image sample without requiring Non-Maximum Suppression (NMS).

C Additional Results

Comparison of different types of ground truth depth maps. Table 5 shows the performance of the KITTI validation set of utilizing different types of ground truth depth maps. Previous works, such as CaDDN [10], generate dense depth maps by projecting point clouds onto the image plane followed by depth completion [36]. On the other hand, Foreground Depth Map [9] is built by the ground truth 2D bounding boxes and object center depths. Our model performs better with the Foreground Depth Map. We argue that the traditional dense depth map only contains the depth of the object surface, which cannot be directly employed in 3D object detection tasks. In contrast, Foreground Depth Map stores the depth of the object center, making depth features generate more relevant information for the detection head.

Qualitative Results. Figure 4 illustrates the qualitative results of the KITTI validation set. Our model demonstrates accurate object localization for objects at near and moderate distances. However, it faces challenges when dealing with heavily occluded objects or those located far from the image center. These challenges arise due to missing information on the image and object distortion, which are common limitations of monocular-based methods.

To further demonstrate the effectiveness of our model, particularly in detecting distant objects, we compare our prediction results with those of MonoDTR [8]. Figure 5 showcases the 3D detection results of our model and MonoDTR in both the image and LiDAR coordinates. Our model exhibits superior performance in localizing distant objects compared to MonoDTR. Moreover, we are able to detect far objects that MonoDTR fails to detect.

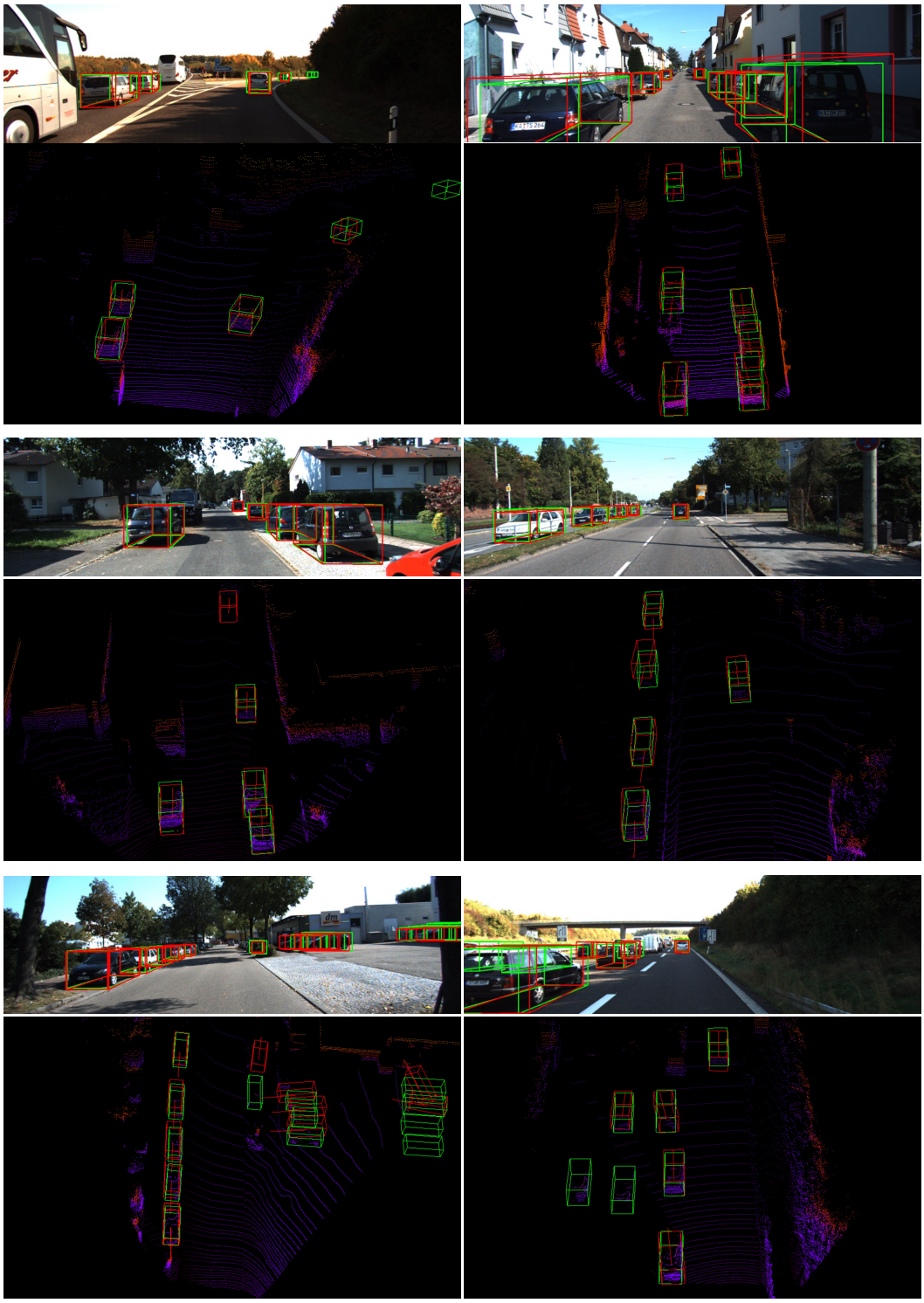

Figure 4: Qualitative results on the KITTI validation set. The red bounding boxes represent the prediction of our model, while the green ones denote the ground truth bounding boxes. Best viewed in color and zoom-in.

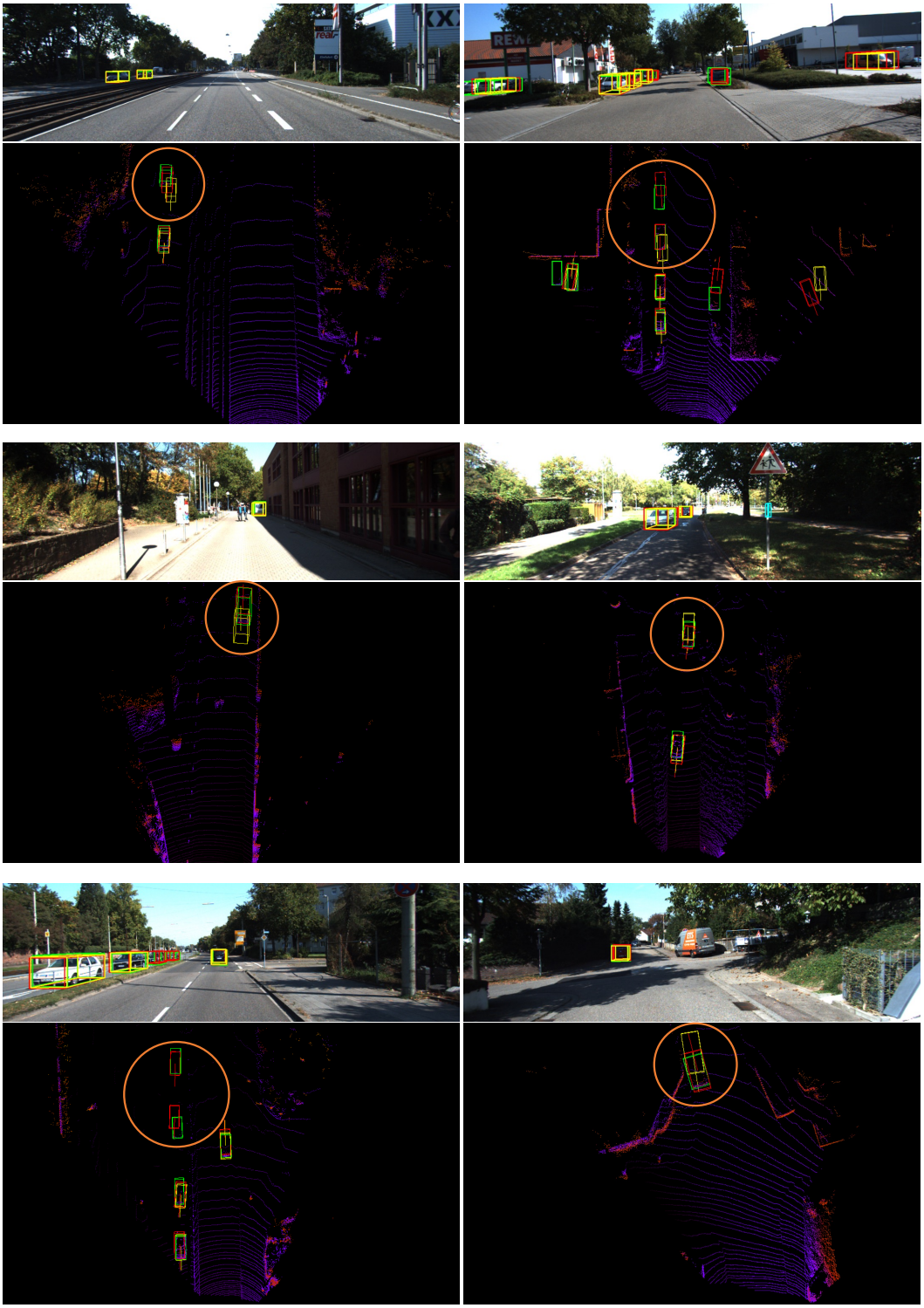

Figure 5: Comparison of our MonoHBD and MonoDTR [8] on the KITTI validation set. The green bounding boxes represent the ground truth boxes. The red, yellow boxes denote the prediction of our model and MonoDTR, respectively. Best viewed in color and zoom-in.