8. Supplemental Material

8.1. Ablation Study for Keypoints Confidence Regression

The 2D/3D keypoints regression is a critical component in the proposed framework, however, inaccurate regression of these keypoints is inevitable in the real AD scenario due to many reasons e.g., viewpoint change, occlusion, and labeling noise, etc. Especially, these prediction outliers will greatly affect the results of the linear system described in Eq. 6. In order to handle this problem, we propose to predict a confidence score for each keypoint and employ it as a weight for determining its contribution to the linear system. To verify the effectiveness of the prediction confidence, we set a series of ablations studies on the "Car" category.

We give the results in Tab. 3. From this table, we can see that the 3D object detection performance can be significantly improved by integrating the regressed key points confidences. More importantly, this improvement is independent of the number of keypoints. In addition, for further understanding the actual meaning of this predicted confidence, we have visualized them in Fig. 9. Interestingly, we find that these keypoints with high confidences usually come from the ground point (the intersection point between the tire and the ground) and these distinguished shape border points. These points will give more contribution to the object pose estimation.

Figure 9: Visualization of keypoints confidences. Here, the blue represents score "1" and yellow represents score "0" and the color changing from blue to yellow represents the confidence score decreasing from "1" to "0". This figure is better to view in color print.

Num. Kps.	Kps. Confi.	Car 3D Det.		
		Easy	Mod.	Hard
16		16.49	12.31	10.54
	✓	**19.59**	**14.50**	**11.88**
48		16.85	12.39	10.04
	✓	**20.09**	**14.65**	**12.07**

Table 3: Keypoints confidence ablation experiments on KITTI *val* set using $AP|_{R_{40}}$ metric.

8.2. Multi-classes Detection

Currently, the designed Autoshape model can't generate the keypoints annotation for "Pedestrian" and "Cyclist" due to the lack of CAD models. Here, we simply transform the 3D keypoints from the mean "Car" template to the "Pedestrian" and "Cyclist" by normalize them first and re-scale them to the bounding box's size of other categories. By generating these keypoints, then the object's pose can easily solve as the "Car" category. We evaluate multi-class 3d detection on the KITTI *test* sever and the performances are shown in Tab. 4. From this table, we can find that the proposed framework performs relatively well even though the keypoints annotation is not very accurate of "Pedestrian" and "Cyclist". Interestingly, we find that the cyclist gives much better results than the "Pedestrian" and this is because the "Cyclist" can be considered as a rigid object to some extent. On the contrary, the "Pedestrian" is a non-rigid object and the location of these keypoints varies a lot with different object pose.

Methods	3D Det.					
	Pedestrian			Cyclist		
	Easy	Mod.	Hard	Easy	Mod.	Hard
M3D-RPN[1]	4.92	3.48	2.94	0.94	0.65	0.47
MonoPair[4]	**10.02**	**6.68**	**5.53**	3.79	2.12	1.83
MonoFlex[43]	9.43	6.31	5.26	4.17	2.35	2.04
Ours	5.46	3.74	3.03	**5.99**	**3.06**	**2.70**

Table 4: Quantitative results for "Pedestrian" and "Cyclist" on KITTI *test* set with $AP|_{R_{40}}$ metric.