ROI-10D: Monocular Lifting of 2D Detection to 6D Pose and Metric Shape

Anonymous CVPR submission

Paper ID 545

1. Synthetic Data Generation

We show more qualitative examples of our extracted textured shapes in Figure 1. Note that we recover metrically-accurate models, but depict them here at different relative sizes to fit onto the page. We want to stress the high visual fidelity of both the geometry and the projective texturing. This level of quality requires a very precise overlap between 2D pixels and projected 3D shape. In Figure 2 we show additional images from our synthetic augmentation scheme during training.

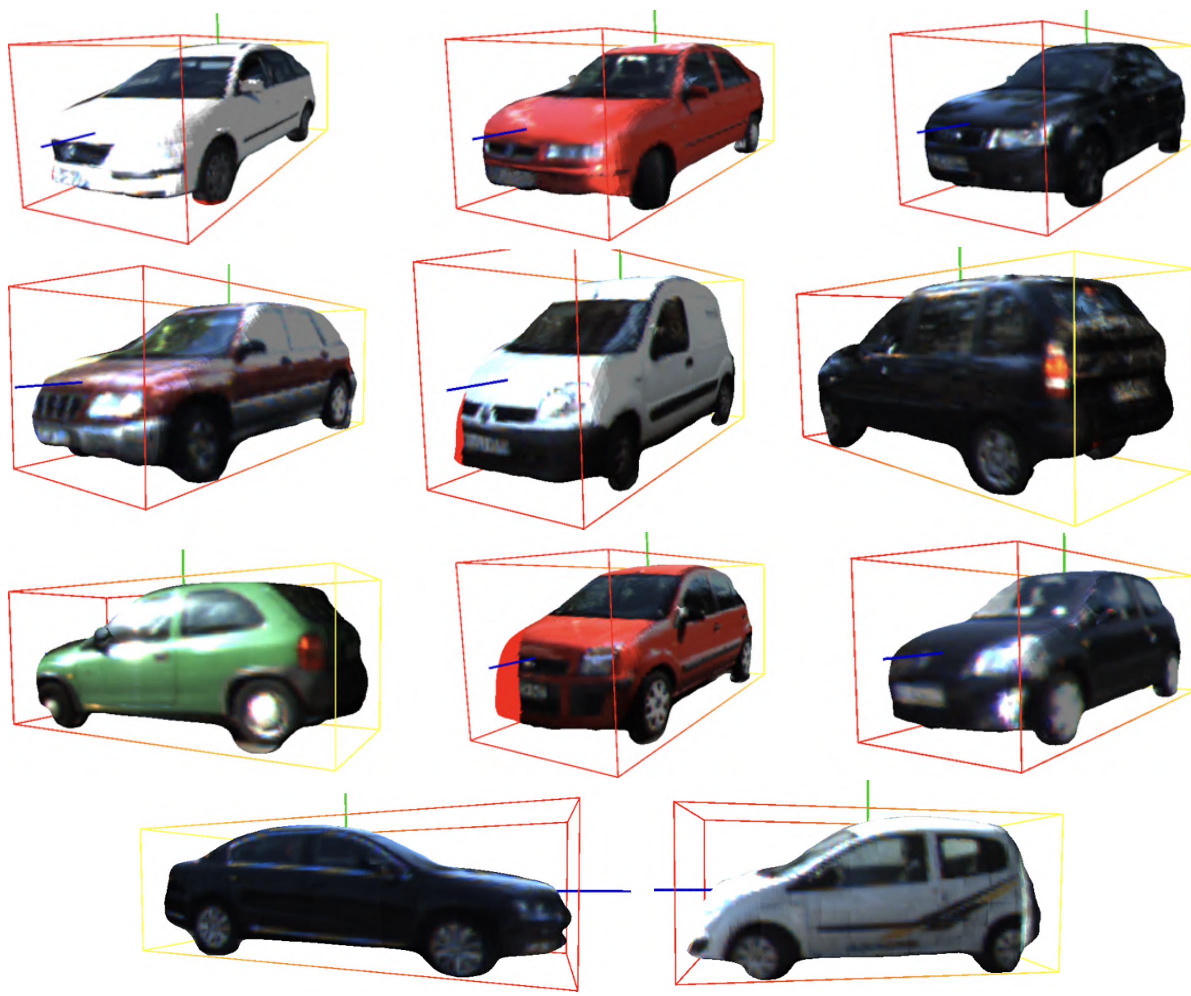


Figure 1: Extracted textured meshes from the train set. Two cars in the center column show red parts that depict missing image information. We inpaint these via texture mirroring along the symmetry axis.

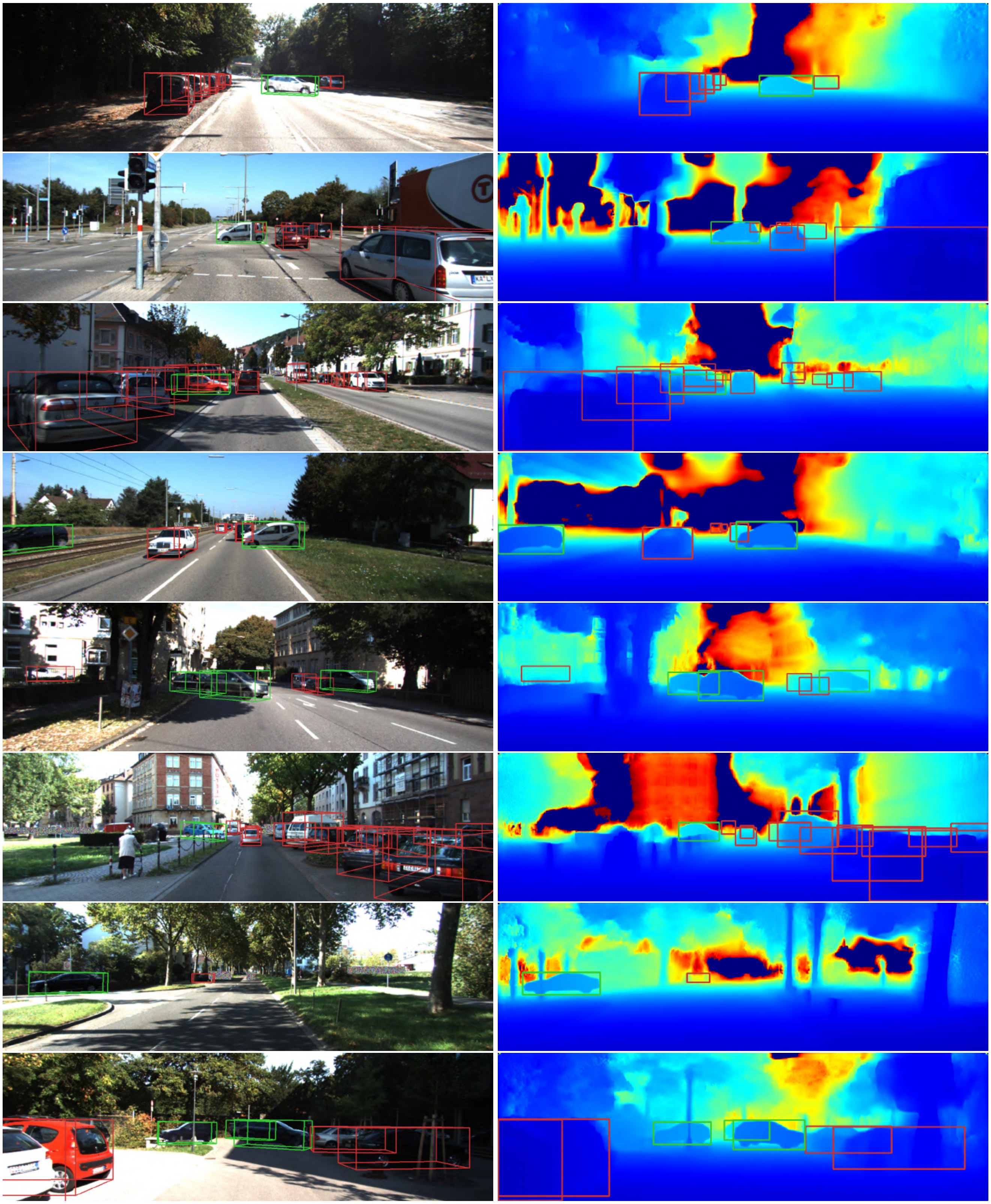


Figure 2: Synthetic training images. The red boxes illustrate the original ground truth instances. The green boxes show the synthetically-added data via rendering random instances from our generated car collection in new poses. The noisy patterns in some images enforce 'ignore'-annotated parts of the image to not be used by negative mining during training.

2. ROI-10D Results on KITTI RAW

Additionally to the 10D results on some KITTI RAW sequences in the supplementary video, we show some recovered meshes in more detail in Figure 3. Note that although these images have not been seen during training, we can retrieve accurate poses and shapes, and in consequence, textured meshes. Even for highly occluded or far-away instances our predictions for pose and shape are quite accurate. For these cases, though, projective texturing can lead to visual artifacts such as overlaps or pixelation.

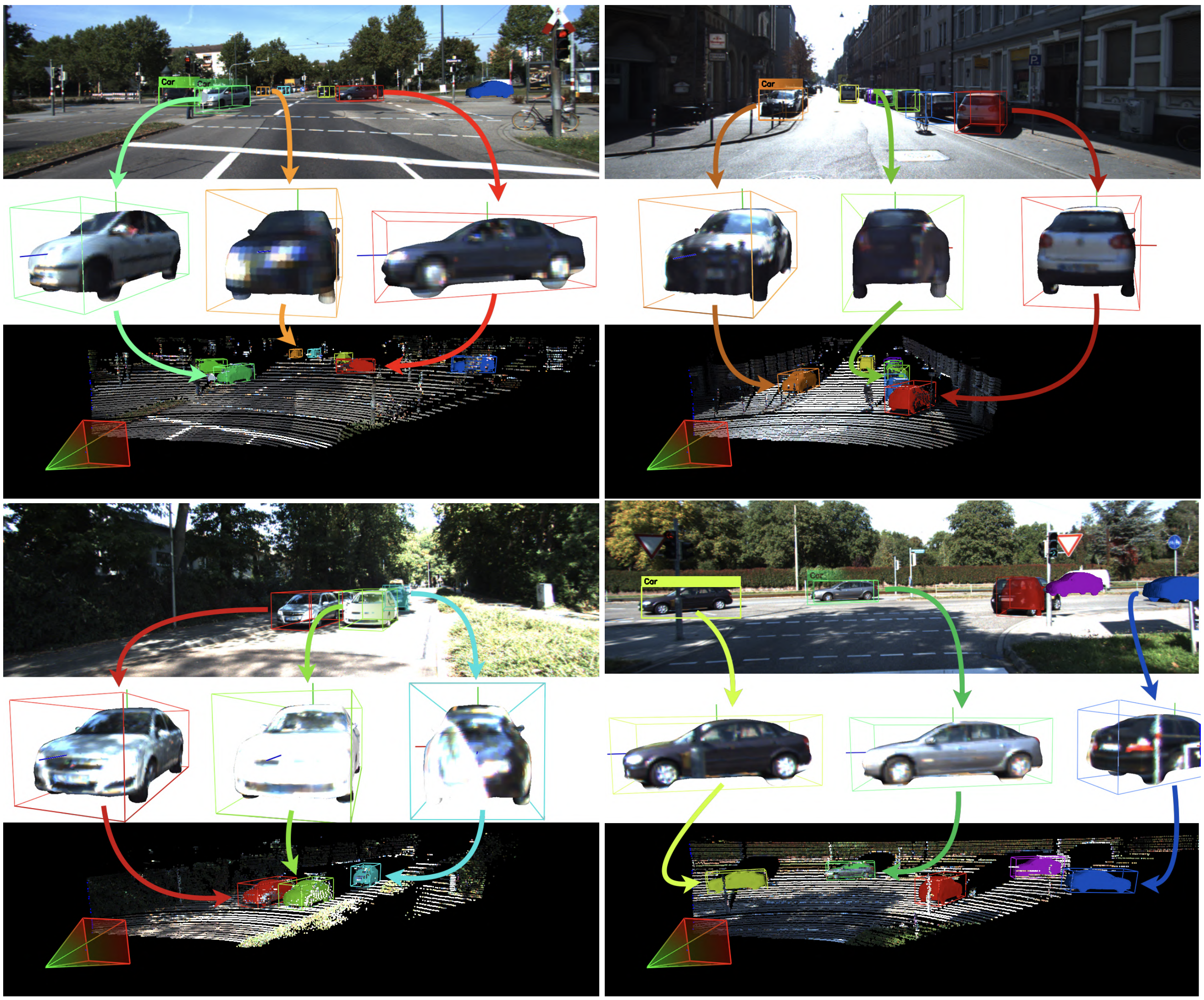


Figure 3: 10D detections and recovered meshes on KITTI RAW.

3. Shape space dimensionality

As mentioned in the paper, we trained a 3D convolutional autoencoder with a latent dimensionality of 6 for our shape space. We tried different dimensionalities and found 6 to be a good compromise between feature compactness as well as expressional power and detail preservation. We depict in Figure 4 the shape interpolation between two median shapes, similar to what has been shown in the paper, but for different latent dimensionalities. As can be seen, even for shape spaces trained with a single latent dimension (top row), we are able to traverse the manifold in a smooth, non-destructive way. In fact, the visual differences are marginal: lower dimensions lead to smoother surfaces and identical side mirrors whereas higher dimensions allow for harder edges and generally more irregularity.

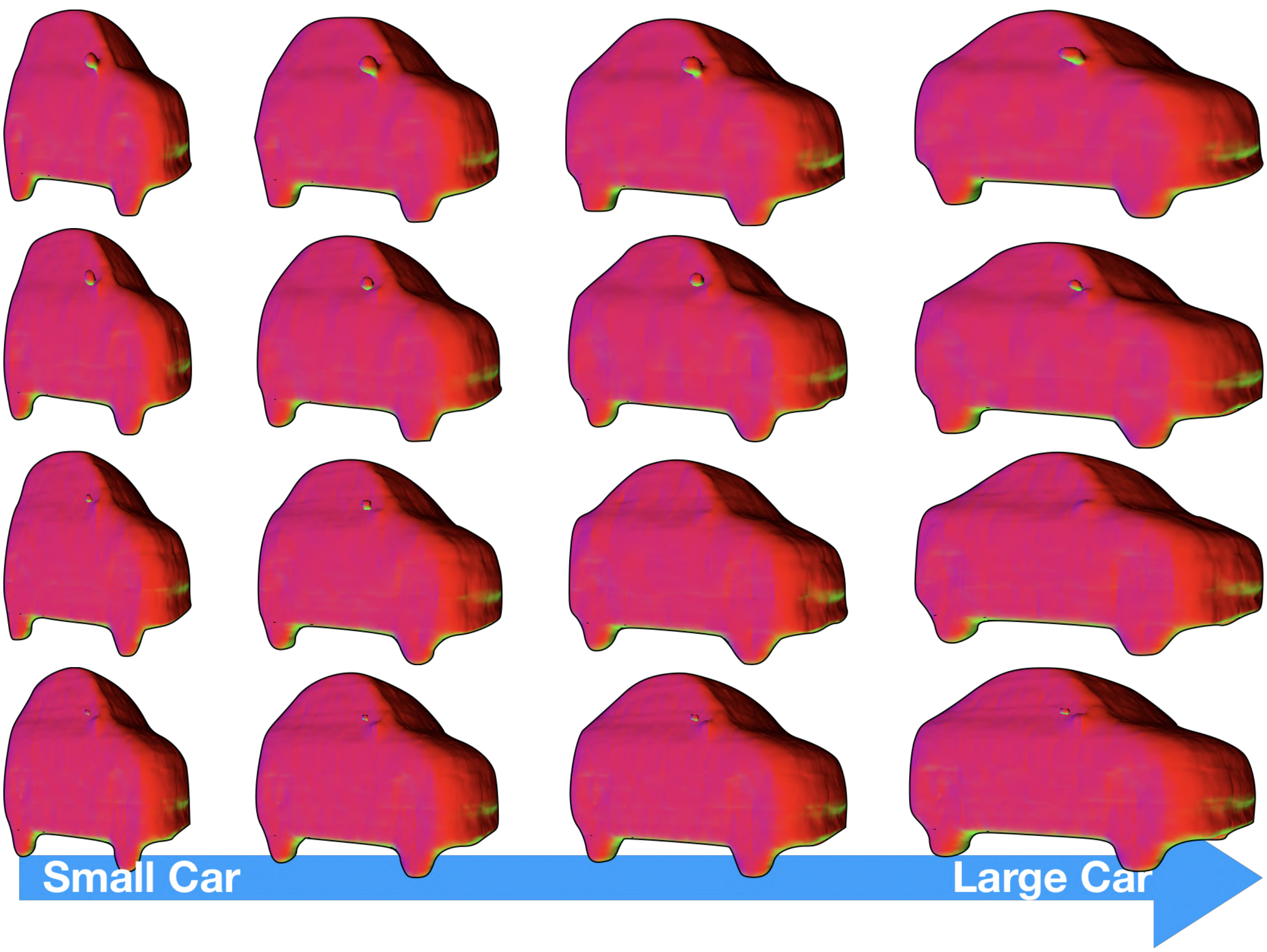
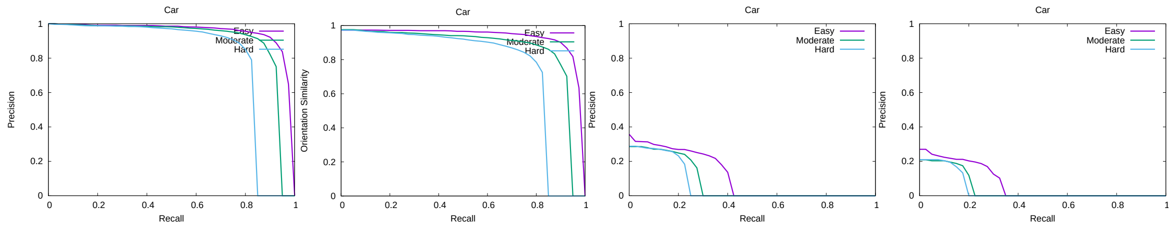


Figure 4: Interpolation between two median shapes with a shape space dimensionality of (from top to bottom) 1, 3, 6, 16.

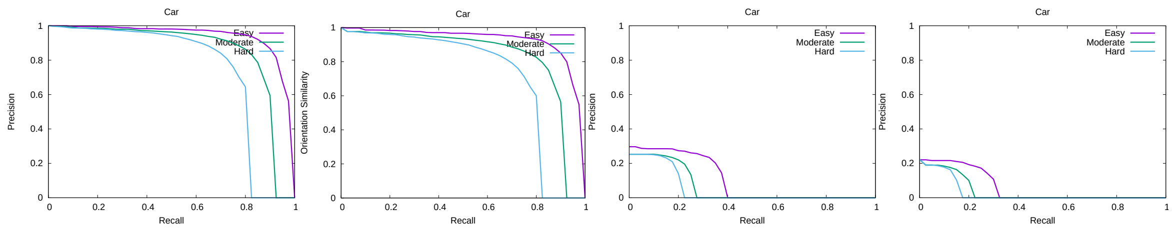
4. 2D Detection and 6D Pose Metrics

We first show the plots produced by the offline evaluation tool for the 'val' set from split of [1] in Figure 5. Additionally, we show the plots provided by the official servers for the test set in Figure 6.



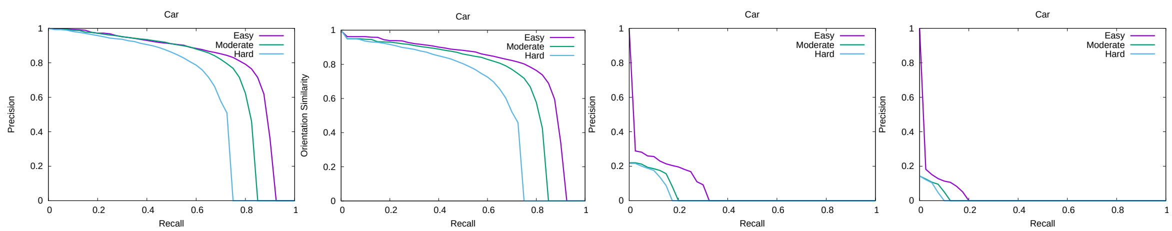
(a) 2D Detection AP (b) Orientation AP (c) Bird's Eye View AP (d) 3D Detection AP

No Weighting



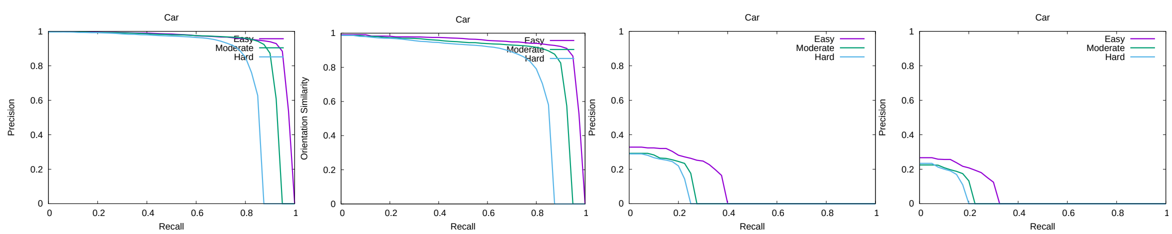
(a) 2D Detection AP (b) Orientation AP (c) Bird's Eye View AP (d) 3D Detection AP

Multi-Task Weighting



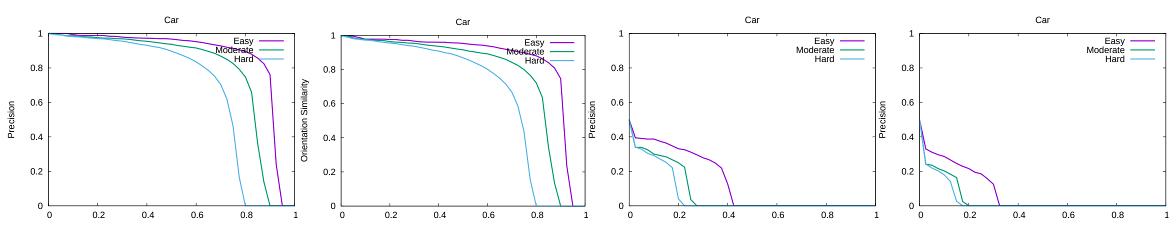
(a) 2D Detection AP (b) Orientation AP (c) Bird's Eye View AP (d) 3D Detection AP

ROI-10D Standard formulation (without SuperDepth module)



(a) 2D Detection AP (b) Orientation AP (c) Bird's Eye View AP (d) 3D Detection AP

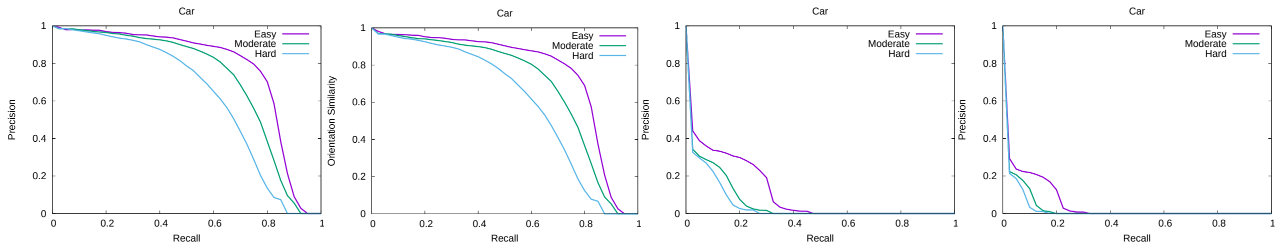
ROI-10D Standard formulation



(a) 2D Detection AP (b) Orientation AP (c) Bird's Eye View AP (d) 3D Detection AP

ROI-10D Standard formulation with additional synthetic training data

Figure 5: Plots of the ablative evaluation on the 'val' split from [1] for different configurations of our method.



(a) 2D Detection AP

(b) Orientation AP

(c) Bird's Eye View AP

(d) 3D Detection AP

Figure 6: Results of our synthetically-augmented model on the official test set. [2]

References

[1] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, and R. Urtasun. Monocular 3d object detection for autonomous driving. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[2] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.