Appendix

1. Autonomous Driving Dataset (ADD)

We describe in more detail the ADD dataset that was used for experiments in the paper.

1.1. 2D Detection

For the task of 2D detection we have annotated a total of 2,261,677 images for the purposes of training and 143,459 images for the test set. This is three orders of magnitude larger than the KITTI 2D detection dataset [2]. This dataset is composed of many dynamic and static object classes that define the scene.

1.2. Panoptic Segmentation

For semantic segmentation we densely annotate groundtruth of the semantic class for each pixel, as part of its instance. In total we annotate 104,587 images for training which is 2 orders of magnitude greated than Cityscapes [1] and 2,363 images for testing.

1.3. Monocular Depth Estimation

For monocular depth estimation, we use the lidar ground points as the ground truth. We use a total of 111,720 images for training 10,968 images for testing, each being at least an order of magnitude larger than KITTI [2]. Each image had approximately 45000 ground truth depths from lidar points. Because the lidars were not co-located with the cameras, some ground truth depths represented the depths of objects behind the actually visible foreground object.

1.4. Qualitative Examples

We also present randomly selected images from our dataset in Figure 1. Our dataset consists of a diversity of cameras, environmental conditions and scenes that result in a high degree of difficulty for the tasks of QuadroNet. ADD is representative of the overall set of problems and difficulty that a perception system for autonomous driving must handle.

2. High Resolution Qualitative Results

Figures 2 and 3 show numerous qualitative examples of the outputs of our network on our Autonomous Driving Dataset (ADD; see subsequent sections for more details): 2D bounding box detections, instance segmentation, semantic segmentation and monocular depth. We include a diverse set of images encompassing different areas and scenarios to demonstrate the generalization ability of our network to complex and varied scenes. Despite the varied input distribution, our network architecture is still able to generalize across multiple cameras, object classes and geographies.

3. Training Details

We evaluate QuadroNet on both Cityscapes and ADD. Each model is trained using Adam optimizer with momentum 0.9 and weight decay 4e-5. Learning rate is linearly increased from 0 to 0.16 in the first training epoch and then annealed down using cosine decay rule. Synchronized batch normalization is added after every convolution with batch norm decay 0.99 and epsilon 1e-3. We use swish activation and exponential moving average with decay 0.9998. We also employ commonly-used focal loss with $\alpha = 0.25$ and $\gamma = 1.5$, and aspect ratio 1/2, 1, 2. We use RetinaNet [3] preprocessing with training-time flipping and scaling.

References

[1] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

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

[3] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2999–3007, 2017.

Figure 1. Randomly sampled images from ADD show the diversity of cameras, environmental conditions and scenes. These factors result in a high degree of difficulty for the tasks that QuadroNet performs.

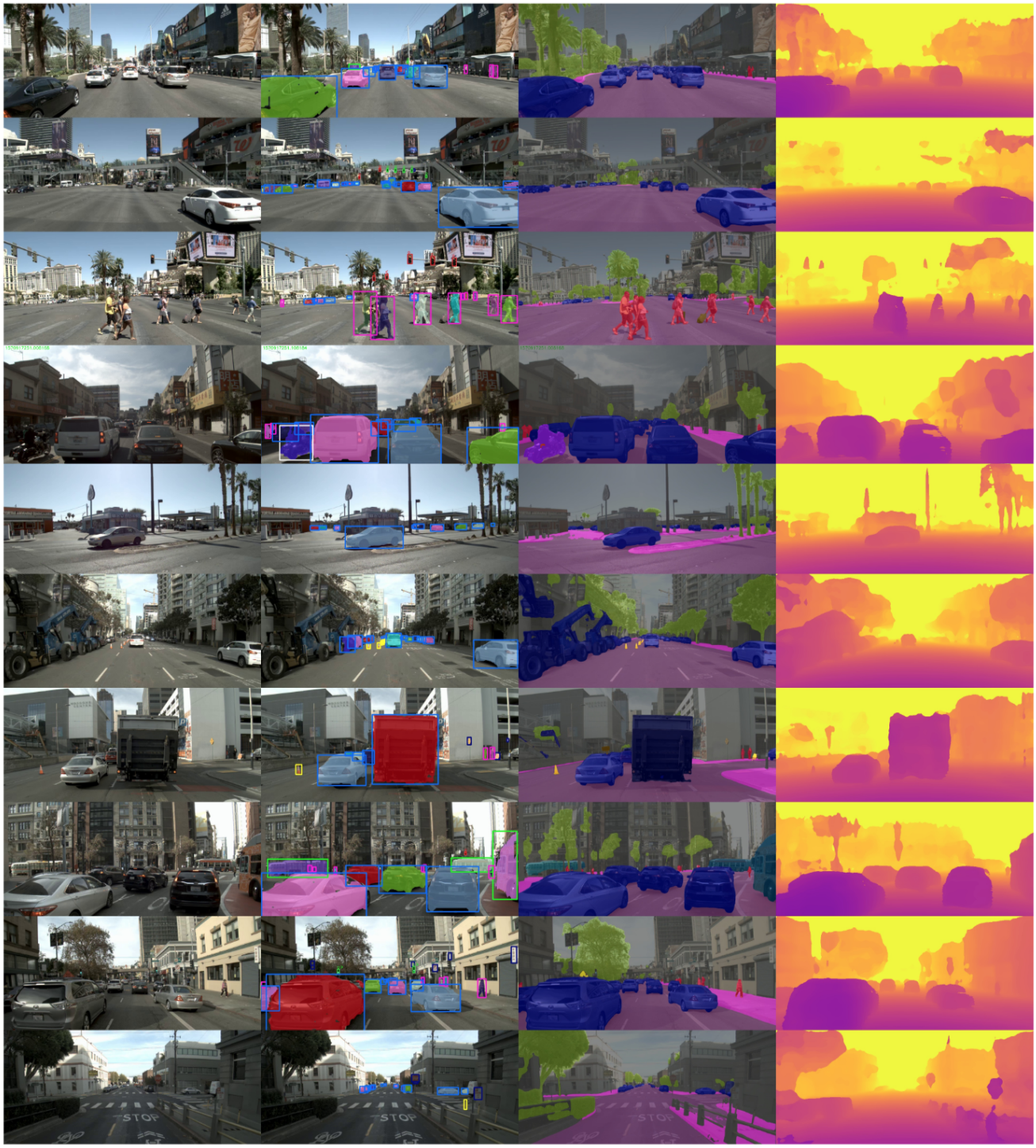

Figure 2. This figure shows additional qualitative examples of the outputs of QuadroNet. From left to right, the columns are original image, 2D detections and instance segmentations, semantic segmentations and egocentric depth map. We show images with a variety of locations and scenarios.

Figure 3. Additional qualitative examples; same display as Figure 2.