Towards Generalization Across Depth for Monocular 3D Object Detection Supplementary Material

Andrea Simonelli[2,3], Samuel Rota Buló[1], Lorenzo Porzi[1], Elisa Ricci[2,3], and Peter Kontschieder[1]

[1] Facebook
{rotabulo, porzi, pkontschieder}@fb.com
[2] University of Trento, Trento, Italy
{andrea.simonelli, e.ricci}@unitn.it
[3] Fondazione Bruno Kessler, Trento, Italy

Abstract. We provide the following, additional contributions for our ECCV 2020 main paper:
– Implementation details of our method
– Qualitative 3D detection results obtained on KITTI3D

1 Implementation details

In this section we provide additional details about the implementation of the virtual views as well as additional information about the hyper-parameters.

3D Detection Head. We adopt 18 anchors with six aspect ratios $\{ \frac{1}{3}, \frac{1}{2}, \frac{3}{4}, 1, 2, 3\}$ and three different scales $\{2s_i 2^{\frac{j}{3}} : j \in 0, 1, 2\}$, where s_i is the downsampling factor of the FPN level f_i. Each anchor is considered positive if its IoU with a ground truth object is > 0.5. To account for the presence of objects of different categories and therefore of potentially different 3D extent, we create class-specific *reference anchors*. Each *reference anchor* has been obtained by analyzing dataset statistics about the training set labels. We define the reference *Car* size as $W_0 = 1.63m$, $H_0 = 1.53m$, $D_0 = 3.84m$, the *Pedestrian* reference as $W_0 = 0.63m$, $H_0 = 1.77m$, $D_0 = 0.83m$ and the *Cyclist* reference as $W_0 = 0.57m$, $H_0 = 1.73m$, $D_0 = 1.78m$.

Losses. We used a weight of 1 for the 2D confidence loss and for the 3D regression loss, while we set at 0.5 the weight of the 2D regression and 3D confidence loss. The Huber parameter is set to $\delta_H = 3.0$ and the 3D confidence temperature to $T = 1$.

Optimization. We used SGD with a learning rate of 0.2 and a weight decay of 0.0001 to all parameters but scale and biases of iABN. We did not optimize the parameters in conv1 and conv2 of the ResNet34. Due to the reduced resolution of the virtual views, we are able to train with a batch size of 2048 on 4 NVIDIA V-100 GPUs (32GB) for 20k iterations, decreasing the learning rate by a factor of 0.1 at 16k and 18k iterations, respectively.

2 Qualitative results

We provide additional qualitative results by visualizing the predictions obtained with MoVi-3D on KITTI3D validation images (using the split described in the main submission), in Fig. 1-2. Following the Official KITTI3D evaluation protocol, we would like to stress that methods are not supposed to detect any vehicle other than *Car* and are also not penalized in case of a false positive detection on *Van* vehicles. Moreover, objects which appear behind or fairly near to the camera as well as objects which are consistently truncated are not supposed to be detected and also not penalized in case of false positives according to the official evaluation protocol. The same holds for objects which are too distant from the camera, as well as objects which are in cluttered areas such as bicycle racks or parking lots.

Fig. 1: Example results of our MoVi-3D model on KITTI3D validation images.

Fig. 2: Further example results of our MoVi-3D model on KITTI3D validation images.