Are we Missing Confidence in Pseudo-LiDAR Methods for Monocular 3D Object Detection? Supplementary Material

Andrea Simonelli[1] Samuel Rota Bulò[2] Lorenzo Porzi[2] Peter Kontschieder[2] Elisa Ricci[1]
[1]University of Trento, Fondazione Bruno Kessler [2]Facebook Reality Labs

Abstract

We provide the following, additional contributions for our ICCV 2021 main paper:

- *Additional details of the creation of our geographically separated (GeoSep) depth splits*
- *Additional 3D confidence ablation results*
- *Additional implementation details of our method*
- *Additional qualitative 3D detection results*

1. Additional details about GeoSep

As explained in Sec. 4.2 of the main paper, we create novel training and validation depth splits (*GeoSep)* by introducing geographical separation between the commonly used depth training set (*Eigen et al.* [6]) and the object detection dataset (KITTI3D).

To create the *GeoSep* splits we exploit the GPS information included in the available KITTI Benchmark data, and define two separation criteria. We withhold all images i) captured closer than 200m from any KITTI3D training or validation detection image, and ii) belonging to any of the KITTI3D detection sequences. From a total available amount of 47962 images the aforementioned filtering process yields 22954 images, which we divide in 22287 for the training set and 667 images for the validation set. The distance threshold has been chosen to ensure that our novel *GeoSep* splits would have approximately the same number of images of the *Eigen et al.* [6] splits, which are 23488 for training and 697 for validation, respectively.

2. Additional 3D confidence ablation results

In Tab. 1 we perform a sensitivity study and analyze the behaviour of the absolute 3D confidence with respect to changes in the temperature value β. We chose a temperature of 0.1, 1 and 10 and computed results on the KITTI3D validation set. The significant change in performance demonstrates that this type of absolute confidence is indeed sensitive to hyperparameter tuning.

	Wang et al. Car 3D AP			*PatchNet Car* 3D AP		
β	Easy	Mod.	Hard	Easy	Mod.	Hard
0.1	28.51	18.78	15.85	33.69	21.00	16.42
1	32.44	20.84	17.26	37.04	23.26	18.78
10	30.72	19.82	16.43	32.05	19.06	15.47

Table 1: *Absolute 3D Confidence* ablation results.

3. Additional implementation details

In this section we provide additional details about the implementation and additional information about the hyperparameters. Since our method is subdivided into multiple branches, we provide details of each one namely *2D Detection*, *Pseudo-LiDAR* and *3D Detection*. In all our experiments, we trained our models on a single NVIDIA GTX 1080 Ti with 11GB of memory.

2D Detection. As described in Sec. 6 of the main paper we do not train a 2D detector but instead rely on precomputed 2D detections. In our experiments we used, for both validation and test set, the 2D detections used in PatchNet [18].

Pseudo-LiDAR. We took the open-source code of BTS [12] and selected the DenseNet161-based estimator. For our results on the Eigen et al. [6] we used the model trained by the authors[1]. For the trainings on our our novel *GeoSep* splits, we used the ImageNet [33] pre-trained model and followed the official schedule and hyperparameters.

3D Detection. The architecture of our proposed models, *i.e.* the ones based on Wang et al. [29] and PatchNet [18],

[1] https://cogaplex-bts.s3.ap-northeast-2.amazonaws.com/bts_eigen_v2_pytorch_densenet161.zip

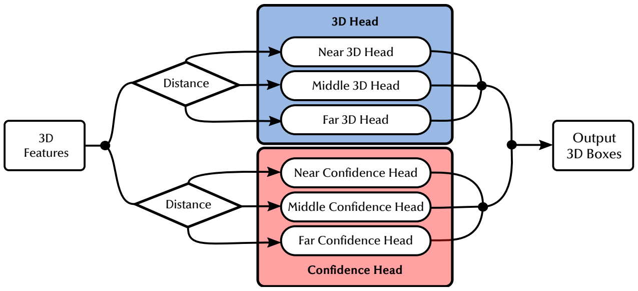

Figure 1: Example of final part of the 3d detection architecture, where we introduced our proposed *3D Confidence Head* in PatchNet [18]. The implementation of the *3D Confidence Head* (red) follows the one of the *3D Head* (blue).

always follow the official one with the only exception of the introduction of our proposed *3D Confidence Head*. The implementation of this particular head closely follows the one of the respective 3D Head. In particular, for our implementation based on Wang et al. [29] we introduced a series of three fully-connected layers with *512-D*, *512-D*, and *1-D* dimensions respectively. For the implementation of PatchNet we introduced three distance-specific heads composed by a series of three fully-connected layers with *512-D*, *512-D*, and *1-D* dimensions respectively. We depict the PatchNet *3D Head* (blue), along with our proposed *3D Confidence Head* (red), in Fig. 1. We trained our model with the Adam optimizer with a learning rate of 0.001 and a batch size of 64 for 100 epochs, decreasing the learning rate by a factor of 0.1 at the 20^{th} and 40^{th} epoch.

4. Additional qualitative results

After a initial description of the visualization method, we provide additional qualitative results of our detections on KITTI3D.

Visualization method. We visualize our results by superimposing our *PatchNet + Relative 3D Confidence* 3d bounding box detections on the input RGB image as well as rendered Pseudo-LiDAR pointcloud, as presented in Fig. 2.

In the top part of the images we visualize our detections on the rendered Pseudo-LiDAR pointcloud, where each point has been colored with its corresponding RGB value (if available), and consequently visualized our predicted 3d bounding-boxes in green for *Car*, cyan for *Cyclist* and red for *Pedestrian*. The presence of black pixels (*e.g.* on top of objects) is due to the fact that we rendered the scene from a point-of-view which is different from the one of the KITTI3D RGB camera. This change of pose inevitably introduces these black pixels on regions which were not visible from the RGB camera pose.

Qualitative results on images. In Fig. 3,4 we show our results on KITTI3D test set images. Our proposed confidence is demonstrated to reliably determine the overall quality of the predicted 3D bounding box. The confidence is in fact higher on nearer and not occluded objects, *i.e.* where the estimation is more reliable, and seems to degrade with distance and occlusions. We also included some failure cases in which our confidence is shown to be less reliable. In particular, we have identified some imprecise or empty detections that still have fairly high confidence.

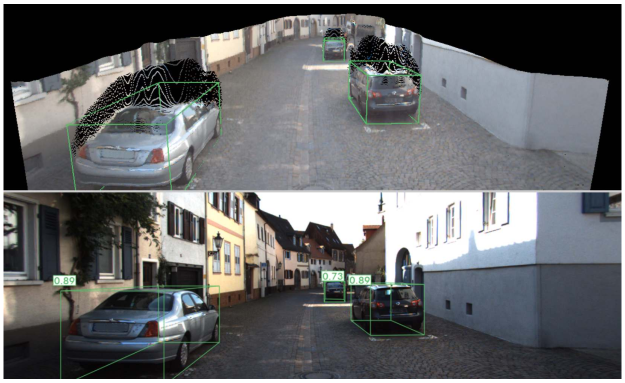

Figure 2: Example of output visualization. Top: Visualization of our predictions on the colored Pseudo-LiDAR pointcloud. Bottom: Visualization of our predictions, with corresponding confidence score, on the input RGB image.

Qualitative results on video sequences. We further provide a qualitative video[2] by showing our predictions on complete KITTI3D sequences taken from the KITTI3D validation set. Unfortunately, it was not possible to provide videos on the KITTI3D test set sequences due to the unavailability of test set sequence information. The predictions are computed for each frame in an independent manner, without exploiting temporal information in any way. Despite the presence of failure cases, *e.g.* when object are too near/far/occluded, our confidence score is shown to generally reflect the quality of the 3d detection.

References

[33] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In CVPR, 2009. 1

[2] https://www.youtube.com/watch?v=yMWNRbJLw6A

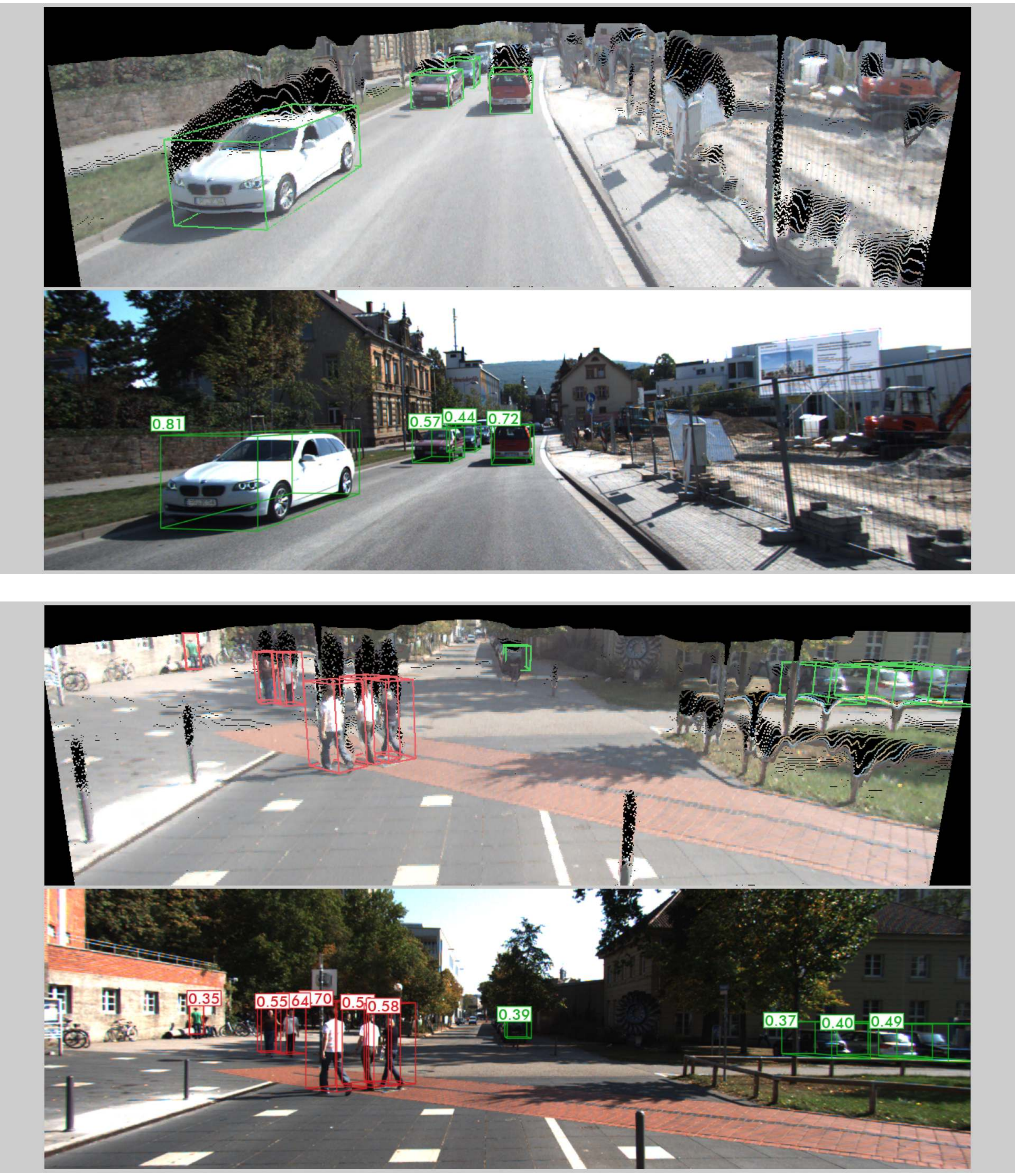

Figure 3: Additional qualitative results of our 3d bounding box detections on the KITTI3D test set.

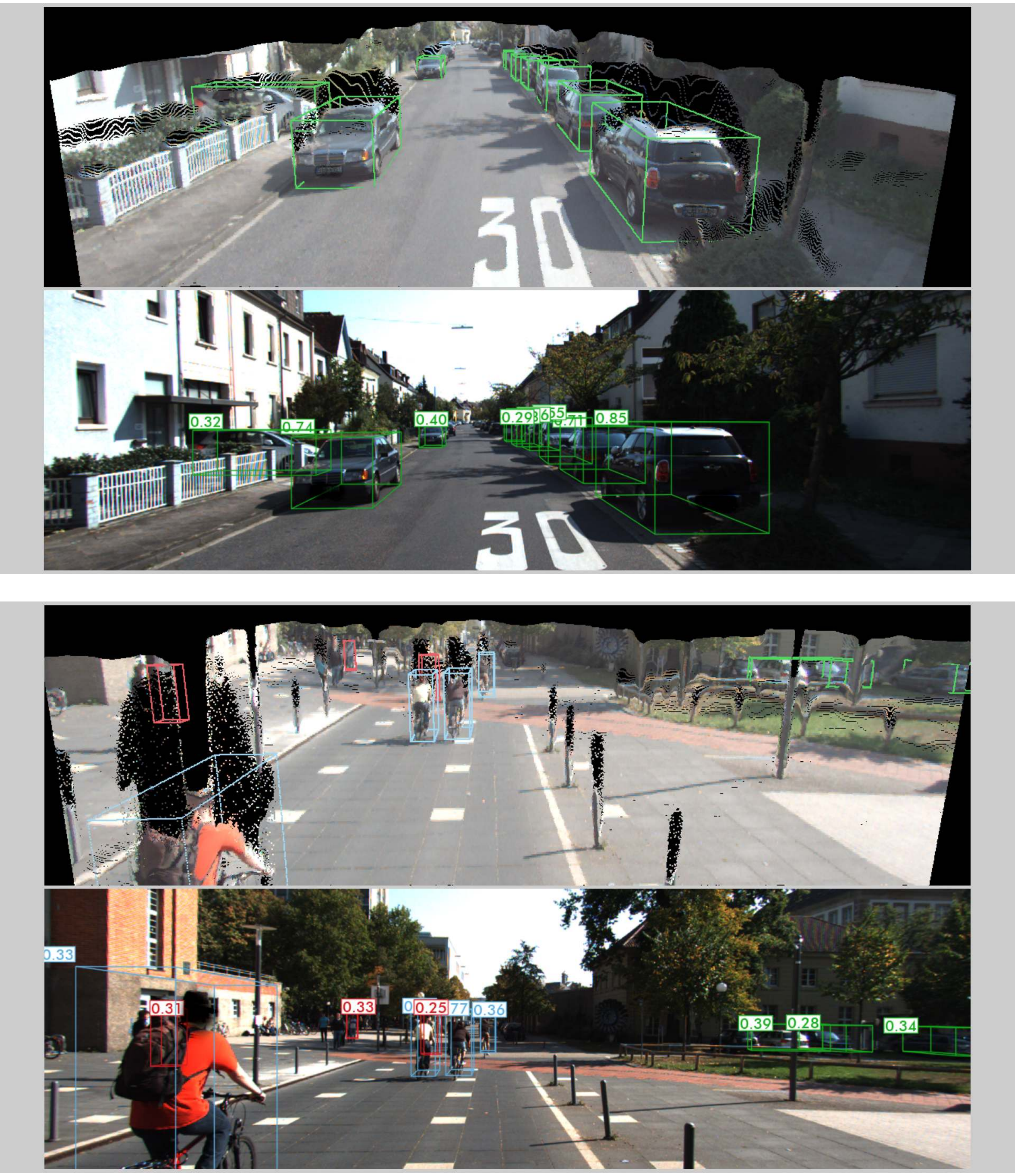

Figure 4: Additional qualitative results of our 3d bounding box detections on the KITTI3D test set.