HUM3DIL: Semi-supervised Multi-modal 3D Human Pose Estimation for Autonomous Driving
Supplementary material

Andrei Zanfir[†] **Mihai Zanfir**[†] **Alexander Gorban**[‡] **Jingwei Ji**[‡] **Yin Zhou**[‡]

Dragomir Anguelov[‡] **Cristian Sminchisescu**[†]

[†]**Google Research**
{andreiz, mihaiz, sminchisescu}@google.com

[‡] **Waymo Research**
{gorban, jingweij, yinzhou, dragomir}@waymo.com

1 Additional Ablations

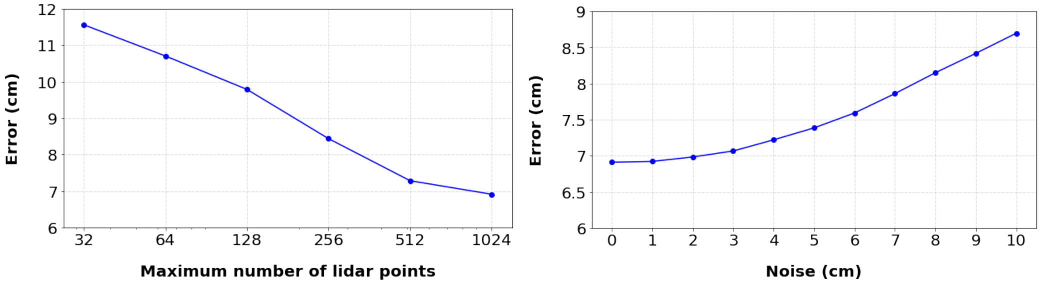

Figure 1: **Left** Network error with respect to the number of lidar points used in inference. **Right** Network error with respect to random gaussian noise added to lidar points (only on the depth component) during testing. The noise is in centimeters and represents the standard deviation.

Method	MPJPE (cm) ↓
GT Bounding Boxes	6.72
GT Bounding Boxes + 1xNoise	7.14
Predicted Bounding Boxes [1]	7.37
GT Bounding Boxes + 3xNoise	7.80

Table 1: Test time MPJPE metric when evaluating our method with different bounding boxes for human detection.

Lidar Points. In Fig. 1, we plot the error with respect to the maximum number of LiDAR points (see left) and random noise added on the depth dimension of the LiDAR points (see right). In the former, we observe that we can get competitive performance even with a small number of available LiDAR points (this relates to a person being too far from the camera or generally, having a small visible surface). In the latter, we show that the network gracefully handles noise and remains robust for even high levels of noise.

Human Detection. Our method takes as input an image crop around a 2D detected bounding box and a subset of lidar points within a 3D detected bounding box. In the main paper, we tried to separate the problem of object detection, as to not be a confounding factor, and evaluated against ground

6th Conference on Robot Learning (CoRL 2022), Auckland, New Zealand.

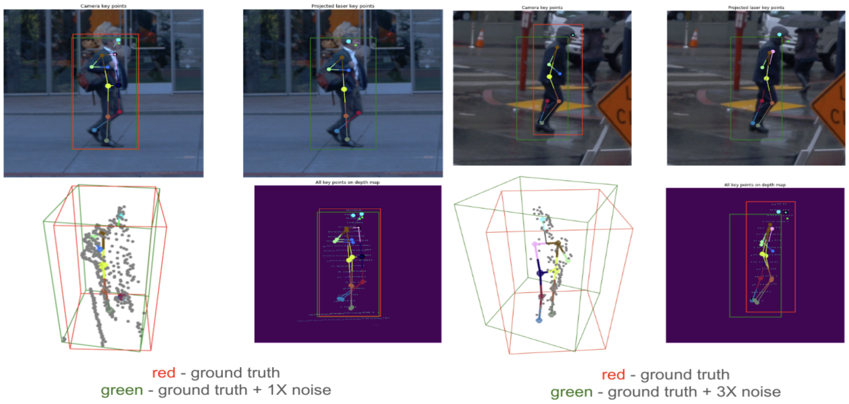
Camera key points
Projected laser key points
Camera key points
Projected laser key points
All key points on depth map
All key points on depth map
red - ground truth
green - ground truth + 1X noise
red - ground truth
green - ground truth + 3X noise

Figure 2: Examples of ground truth bounding boxes used at test time, with added noise.

truth bounding boxes. Here, we try to ablate the performance of our method when using predicted or noisy detections. We consider three additional scenarios: (1) Bounding boxes are predicted by the detection method of [1], which is state-of-the-art for offline object detection using LiDAR sensors. (2) Added a small amount of Gaussian noise to ground truth boxes (std = 0.03 for 3D boxes, std = 7 for 2D boxes). We call this ablation 1xNoise. (3) Added three times more Gaussian noise to the ground truth boxes (std = 0.09 for 3D boxes and std = 21 for 2D boxes). We call this ablation 3xNoise. Examples for noisy bounding boxes are shown in Figure 2. In Table 1 we show comparisons of our method's performance when using different detections at testing time. We can conclude that the reported MPJPE using GT boxes is a bit lower, but somewhat comparable to both the evaluation settings with 1xNoise and to that using predicted boxes, while quality notably degrades if we add too much noise to the GT boxes (3xNoise). We conclude that our model is robust to different levels of noise in the human detections.

2 Qualitative results WOD

In Fig. 3 we show qualitative reconstructions for test images in the Waymo Open Dataset. The predicted joints are: 'NOSE', 'LEFT EYE', 'LEFT EAR', 'LEFT SHOULDER', 'LEFT ELBOW', 'LEFT WRIST', 'LEFT HIP', 'LEFT KNEE', 'LEFT ANKLE', 'RIGHT EYE', 'RIGHT EAR', 'RIGHT SHOULDER', 'RIGHT ELBOW', 'RIGHT WRIST', 'RIGHT HIP', 'RIGHT KNEE', 'RIGHT ANKLE', 'MOUTH', 'FOREHEAD', 'HEAD CENTER'. Note that **FOREHEAD** is only available in 2D annotations, while **HEAD CENTER** only in 3D.

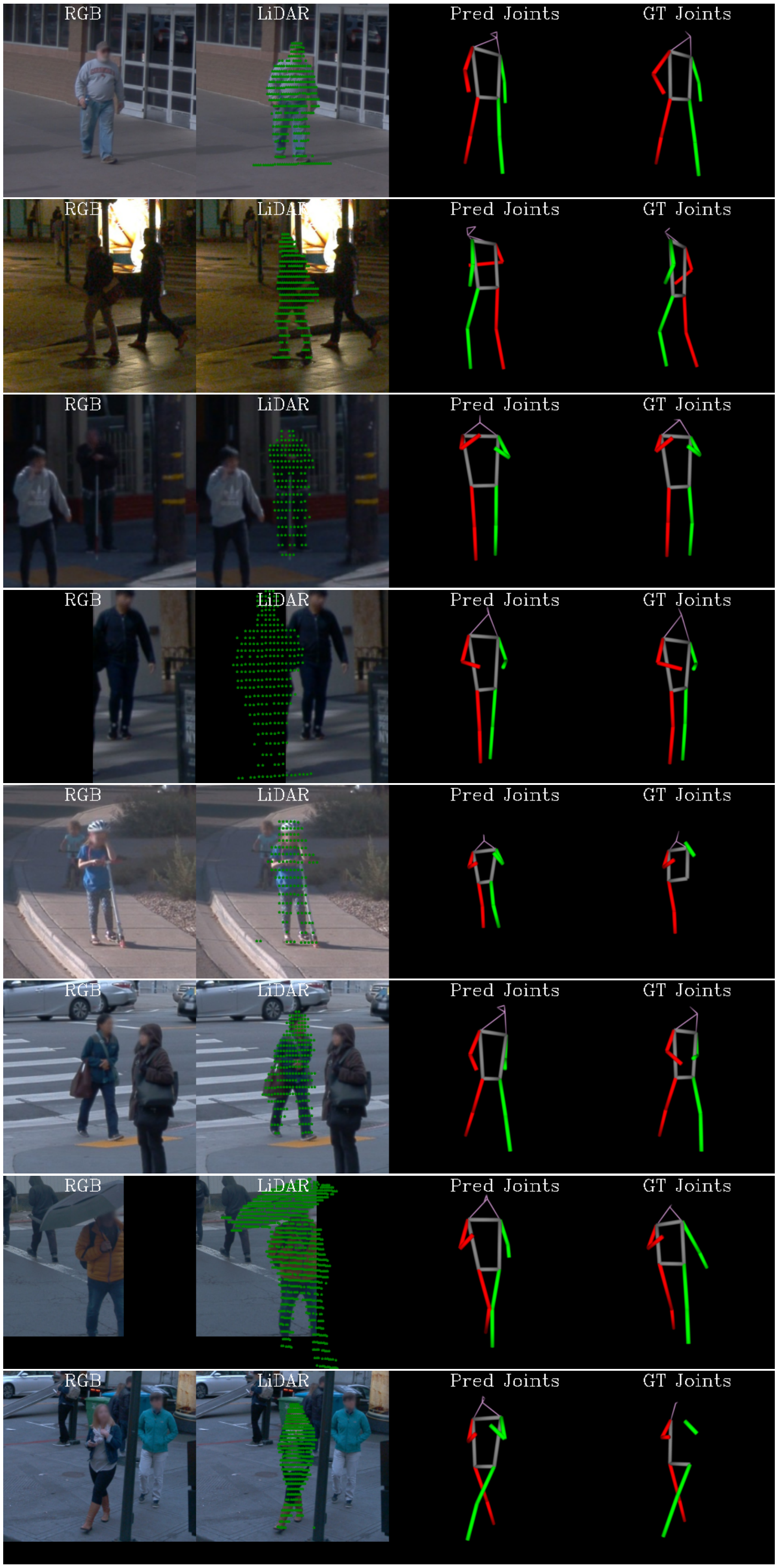

Figure 3: Qualitate results for HUM3DIL on WOD. Left to right: Input RGB, Input LiDAR, Predicted 3D joints, GT 3D joints.

References

[1] C. R. Qi, Y. Zhou, M. Najibi, P. Sun, K. Vo, B. Deng, and D. Anguelov. Offboard 3d object detection from point cloud sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6134–6144, 2021.