Supplementary Material for: Planning Paths through Occlusions in Urban Environments

Yutao Han[*]
OPPO US Research Center
{yutao.han}@innopeaktech.com

Youya Xia[*]
Cornell University
{yx454}@cornell.edu

Guo-Jun Qi
OPPO US Research Center
{guojun.qi}@innopeaktech.com

Mark Campbell
Cornell University
{mc288}@cornell.edu

[*] Equal contribution

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

Appendix

We provide details omitted in the main text.

- Section 1 provides more visualization examples and comparisons of planning results using the inpainted semantic map. Specifically, we show visual comparisons to baseline methods (*i.e.*, Pix2Pix [1], Pix2PixHD [2] and DeepFillV2 [3]).
- Section 2 provides ablation studies for our proposed inpainting model. Specifically, following the vision community convention, we show mIOU results for the three ablation models.
- Section 3 contains studies of our planning framework for simulated congested road conditions.
- Section 4 is a brief discussion on failure conditions for our planning framework.
- Our code demo for our inpainting model and path planner: `https://github.com/genplanning/generative_planning.git`.

1 More Visualization Results

Visual comparisons between our inpainting model and other baselines. All images shown are **randomly-selected** from our test set.

Figure 1 shows visual comparisons between the input lidar, ground truth (GT), our model, and Pix2Pix [1]. Our model (third column) consistently plans longer paths than Pix2Pix (rightmost column). No blue dots means no path is found. While Pix2Pix demonstrates a clear improvement over the initial lidar scan (leftmost column), it is clear that our modifications and new loss functions are an improvement on inpainting performance.

Figure 2 shows visual comparisons with Pix2PixHD [2]. All models show a clear performance improvement compared with the initial lidar scans (leftmost column). Our model (third column) shows a clear improvement on Pix2PixHD (rightmost column). For example, in the third row from the bottom, our model is able to inpaint a path much closer to the green dot (goal), and also more similar to the GT, compared to Pix2PixHD.

Figure 3 shows visual comparisons with DeepFillV2 [3]. DeepFillV2 (rightmost column) performs poorly compared with our other baselines qualitatively. It is unable to fill in the sparse semantic lidar points (rightmost column; rows one, four, and seven), and the output of the network is still a sparse BEV point cloud.

2 Ablation Study

We perform a detailed ablation study to assess the performance improvements from each of our proposed components and loss terms. We use mIOU, which is a standard metric for semantic prediction, to evaluate the semantic prediction results in our ablation studies. We evaluate the mIOU results for the **road** class in our evaluation **using the entire test set**, as that is the class used to define the navigable road. In addition to the mIOU results, we evaluate Frechet distance, path length, and average angle distance for the test set of Route 0 (See Table 1 in the paper for more details about training and test sets). Table A1 summarizes the result. We begin with a baseline *paired* translation algorithm (Pix2Pix [1]) (Row 1). By adding our proposed components on top of the baseline algorithm one by one (Row 2-4), we can obtain our proposed model and show the evaluation results for our proposed model in Row 5. We have the following observations from Table A1:

- The evaluation results increase row by row (since the smaller the distance, the better the planned trajectory. So, the decreasing of Frechet distance [4] and angle difference imply improvement of evaluation results). We conclude each of the proposed components for modifying Pix2Pix [1] is indispensable for our proposed model.
- From the degree of improvement row by row, we find that adding the patchNCE loss and replacing the UNet generator [1] with the two stage generator [2] demonstrates the most

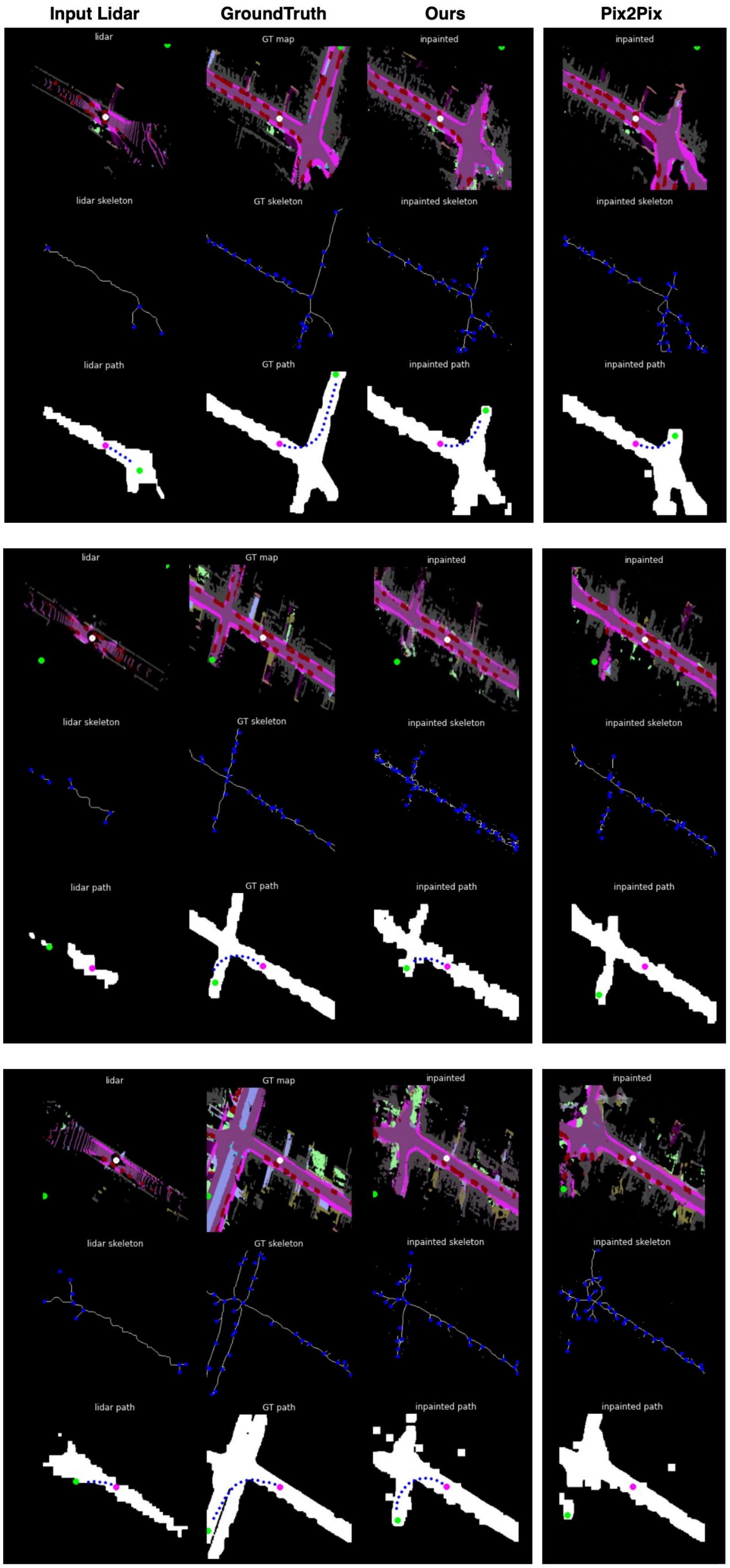


Figure 1: Comparison of **randomly-selected** inpainting and planner performance with our model, Pix2Pix [1], and ground truth (GT). Columns: (leftmost) Lidar scan (OL), (second) GT, (third) Ours and (rightmost) Pix2Pix.

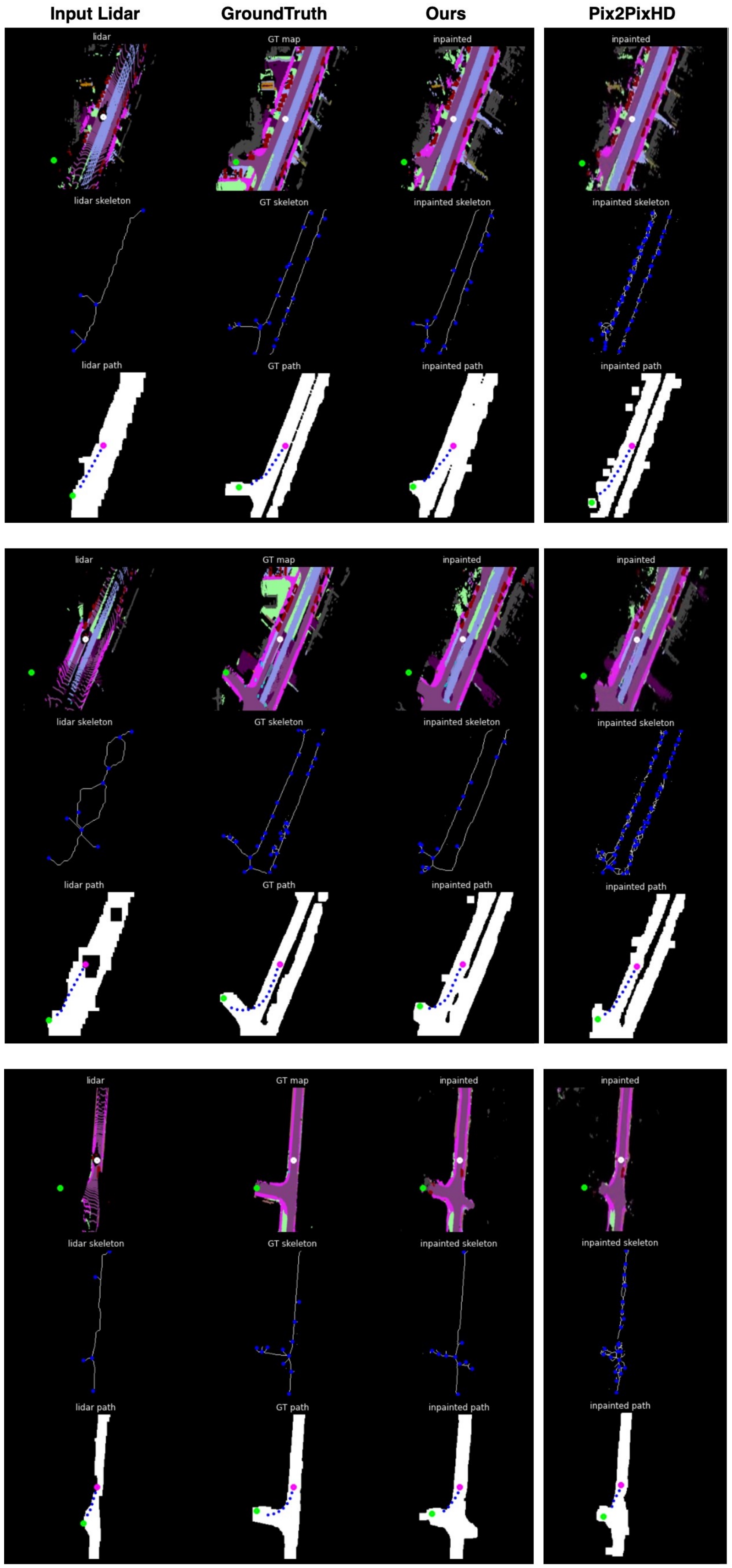


Figure 2: Comparison of **randomly-selected** inpainting and planner performance with our model, Pix2PixHD [2], and ground truth (GT). Columns: (leftmost) OL, (second) GT, (third) Ours and (rightmost) Pix2PixHD.

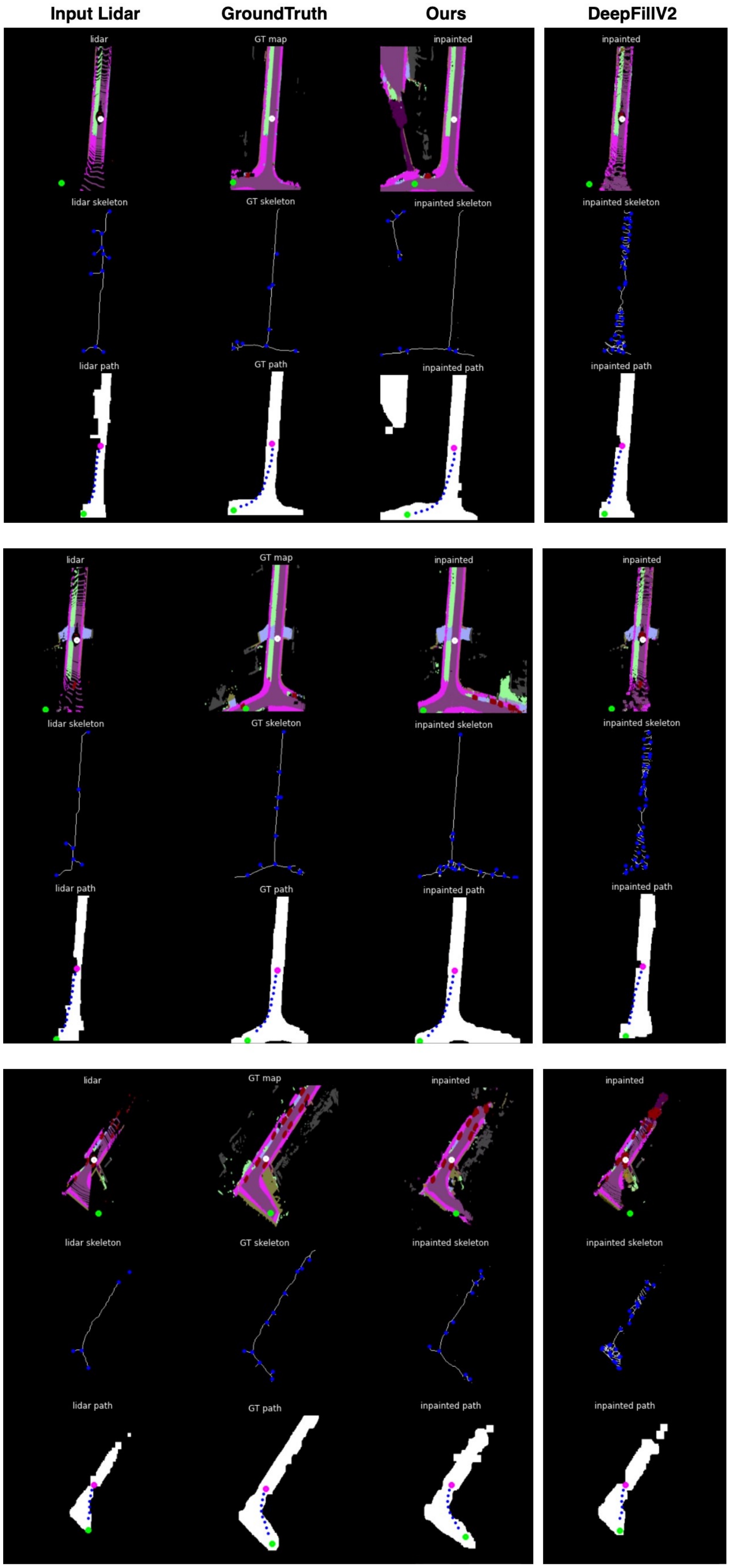


Figure 3: Comparison of **randomly-selected** inpainting and planner performance with our model, DeepFillV2 [3], and ground truth (GT). Columns: (leftmost) OL, (second) GT, (third) Ours and (rightmost) DeepFillV2.

Table A1: Ablation study of our proposed models.

Arch+loss	mIOU ↑	Frechet distance (pixel) ↓	Path length (%) ↑	Angle Difference (°) ↓
U-Net encoder-decoder generator [1]+PatchGAN discriminator+GAN loss [1]+L1	55.11	12.8015	75.92	11.87
Two stage generator [2]+PatchGAN discriminator [1]+GAN loss+ L1	58.25	9.219	77.38	9.76
Two stage generator [2]+multi-scale discriminator [2]+GAN lossr+ L1	61.93	9.08	78.48	8.67
Two stage generator [2]+multi-scale discriminator [2]+GAN loss+ inpainting-targeted L1	62.50	8.99	78.85	7.18
Two stage generator [2]+multi-scale discriminator [2]+GAN loss+ inpainting-targeted L1+PachNCE (**Ours**)	65.66	8.17	83.32	6.53

obvious improvement. We conclude they are the two most important components for our proposed inpainting model.

3 Experiments for Congested Road Conditions

To simulate a congested road, we use additional occlusion masks on the original lidar (OL) scans (rectangles of size 200×50 pixels) to simulate additional occlusions along the road. We run our inpainting framework on the masked OL scans. We run experiments for one, two, and three occlusion masks to simulate increasing scene difficulty for 180 frames from the test set. We also calculate the percentage of the simulated occlusion to road ratio in terms of image pixels. The percentages are 6.71%, 10.96%, and 16.05% for one, two, and three masks respectively.

Table A2 shows the metrics described in the experimental evaluation section of our paper comparing performance given number of simulated occlusions. We can see that performance drops as the number of simulated occlusions increases. However, performance is still better than using the baseline lidar map (original OL) for planning across all metrics, except for path length (%) for three occlusions, even when the baseline map has no simulated masks.

Figure 4 shows qualitative inpainting results with occlusion masks. Even with three occlusion masks which completely block out the intersection region (Figure 4(c)), our predictive model is still able to fill in the rough form of the intersection (Figure 4(f)).

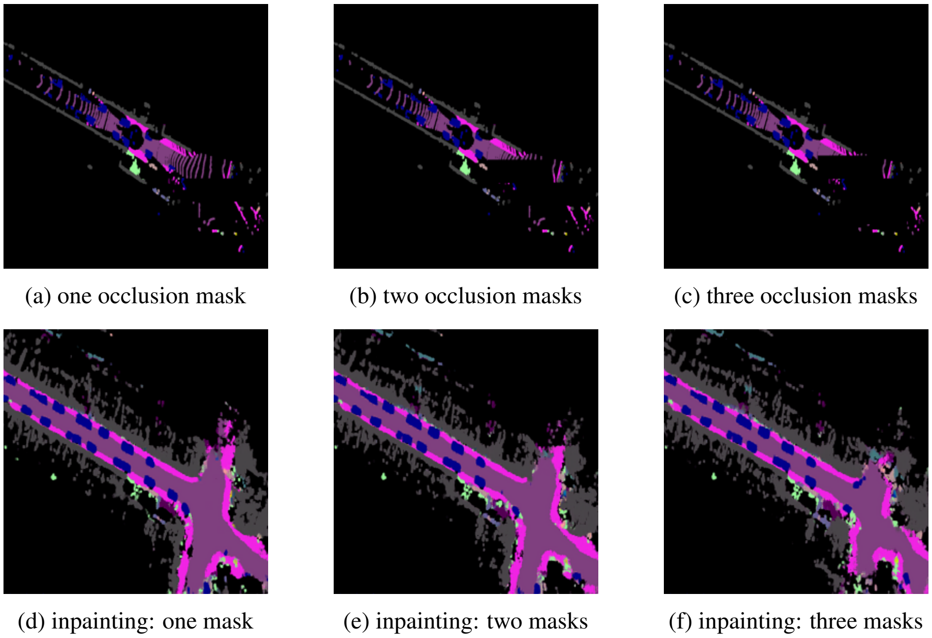
Table A2: Evaluation results for simulated occlusions

	Frechet distance	Average Angle difference	Path length (%)	Major branch prediction (%)
No occlusions	9.03	7.12	81.11	92.09
One occlusion	9.32	8.64	78.21	89.99
Two occlusions	10.72	10.09	70.73	86.05
Three occlusions	11.15	12.62	61.32	83.00
No occlusions (original OL)	21.32	14.57	62.06	60.87

4 Failure Case Evaluation

The main failure case in our experiments is when the predictive model is unable to predict a turn or an intersection. This leads the predictive model to behave like the baseline original lidar (OL) planner, so while the predictive model is not helpful in this case, it does not worsen the performance compared to the baseline. In our experiments on our model, this case occurs only 1.3% of the time. However, for the baseline OL planner, this failure case occurs 72.2% of the time.

The other potential failure case is if the model predicts a turn that does not exist. This is more dangerous because if the robot plans a turn that does not exist, it could lead to collisions. However, in our experiments this case does not occur.



(a) one occlusion mask

(b) two occlusion masks

(c) three occlusion masks

(d) inpainting: one mask

(e) inpainting: two masks

(f) inpainting: three masks

Figure 4: Some examples of simulated occlusion masks. (Top Row): Lidar scans with occlusion masks. (Bottom Row): Inpainting results on occlusion masks.

References

[1] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, 2017.

[2] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.

[3] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4471–4480, 2019.

[4] H. Alt and M. Godau. Computing the fréchet distance between two polygonal curves. *International Journal of Computational Geometry & Applications*, 5(01n02):75–91, 1995.