Depth-Only Inpainting of Depth Maps for Autonomous Driving

Jagrit Rai
University of Toronto

Motivation

Depth sensors frequently fail in real-world environments, producing incomplete depth maps that degrade downstream tasks such as **autonomous driving, SLAM, navigation, and 3D reconstruction**.

Real-world depth loss causes:

- Sensor self-occlusion (robot arms, drone body, vehicle hood)
- Reflective surfaces (water, glass, chrome)
- Hardware blind spots

These failures often produce **large, contiguous missing regions**, not just random sparsity.

Most existing depth inpainting literature focuses mainly on **dense indoor RGB-D datasets**, not real autonomous-driving depth with realistic occlusions.

KITTI [1] is a large, real-world dataset collected from a vehicle equipped with LiDAR and stereo cameras. Its fused "ground truth" depth maps are **semi-dense**, created by projecting LiDAR into the camera frame and densifying with stereo cues and geometric heuristics.

We simulate structured occlusions by applying **random rectangular masks** over **real world** KITTI depth maps and train a **U-Net model to recover depth information** inside these masked regions.

The goal: Learn geometric structure from real outdoor depth and reconstruct occluded regions without relying on RGB or synthetic depth.

Related Work

Deep CNN-Based Image Inpainting: Contextual encoder–decoder networks [2] establish that CNNs can inpaint arbitrarily missing regions using synthetic masks. These methods learn spatial priors but are trained for RGB images, not geometric depth structure, and do not generalize well to depth map inpainting.

Depth Inpainting With Smoothness Priors: Classical depth inpainting approaches use explicit geometric regularizers such as second-order smoothness [3] or Gauss–Markov random fields [4]. These methods rely on hand-crafted priors or RGB guidance and do not generalize well to large, structured occlusions.

Learning-Based Depth Map Inpainting: Recent works train CNNs specifically for depth images; for example, depth-denoising priors [5] or U-Net variants for structured holes [6], however these methods are usually evaluated on indoor RGB-D data, depend on semantic masks, or require RGB context.

References

[1] Uhrig, J., Schneider, N., Schneider, L., Franke, U., Brox, T., & Geiger, A. (2017). *Sparsity invariant CNNs*. In International Conference on 3D Vision (3DV).
[2] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 5505–5514.
[3] Herrera C, D., Kannala, J., Ladický, L. U., & Heikkilä, J. (2013, June). Depth map inpainting under a second-order smoothness prior. In *Scandinavian conference on image analysis* (pp. 555-566). Berlin, Heidelberg: Springer Berlin Heidelberg.
[4] Satapathy, S., & Sahay, R. R. (2021). Robust depth map inpainting using superpixels and non-local Gauss–Markov random field prior. *Signal Processing: Image Communication*, *98*, 116378.
[5] Li, Z., & Wu, J. (2019). Learning deep CNN denoiser priors for depth image inpainting. *Applied Sciences*, *9*(6), 1103.
[6] Mao, J., Li, J., Li, F., & Wan, C. (2020, October). Depth image inpainting via single depth features learning. In *2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)* (pp. 116-120). IEEE.

New Technique

Most depth-inpainting work targets **dense indoor RGB-D** or **synthetic** data.
We propose a depth-only U-Net trained exclusively on real KITTI fused ground-truth depth.

- Unlike depth completion (sparse → dense), our task reconstructs **large structured holes** randomly covering 5-20% of the image area, which completion methods cannot resolve.

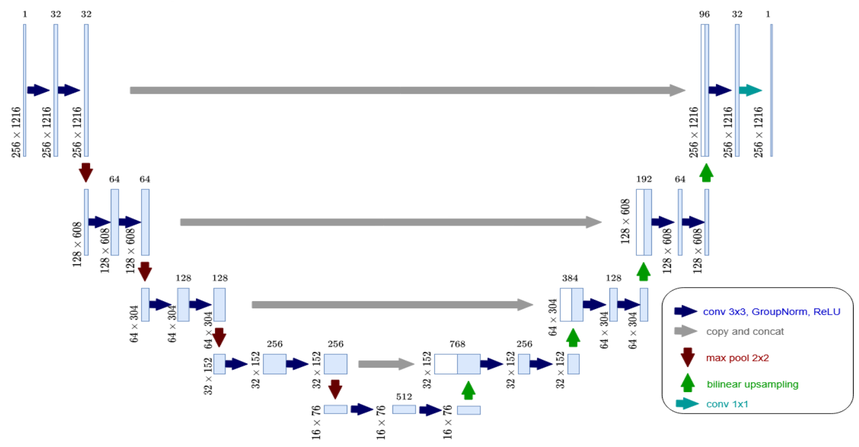


Figure 1: Depth Map U-Net Model Architecture

L1 Hole Loss: Focuses learning within masked region.
L1 All Valid-Pixel Loss: Enforces consistency with all valid KITTI depth.
TV Smoothness: Encourages geometric continuity across boundaries.
Total Loss: L_all + λ L_holes +β TV

Training Methodology:

- Pretrained on **25%** of KITTI for **6 epochs**.
- Fine-tuned on **50%** of KITTI for **5 epochs** (full dataset: 85k+ , computationally prohibitive)

Best model achieved **0.071 RMSE** on masked-region validation.

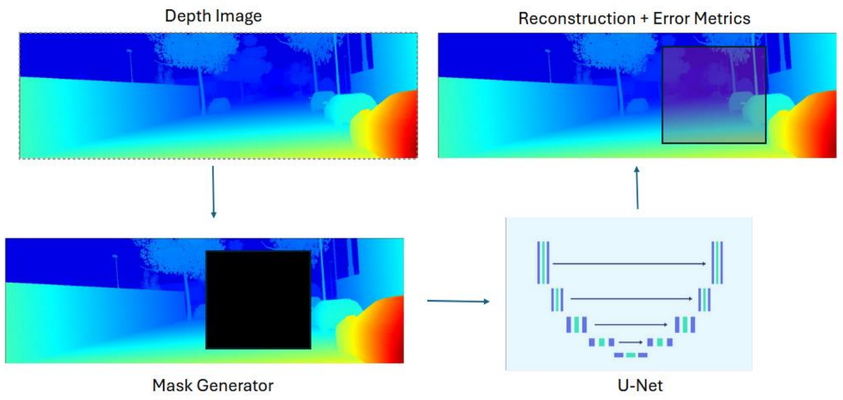


Figure 2: Inpainting Process Flow Chart

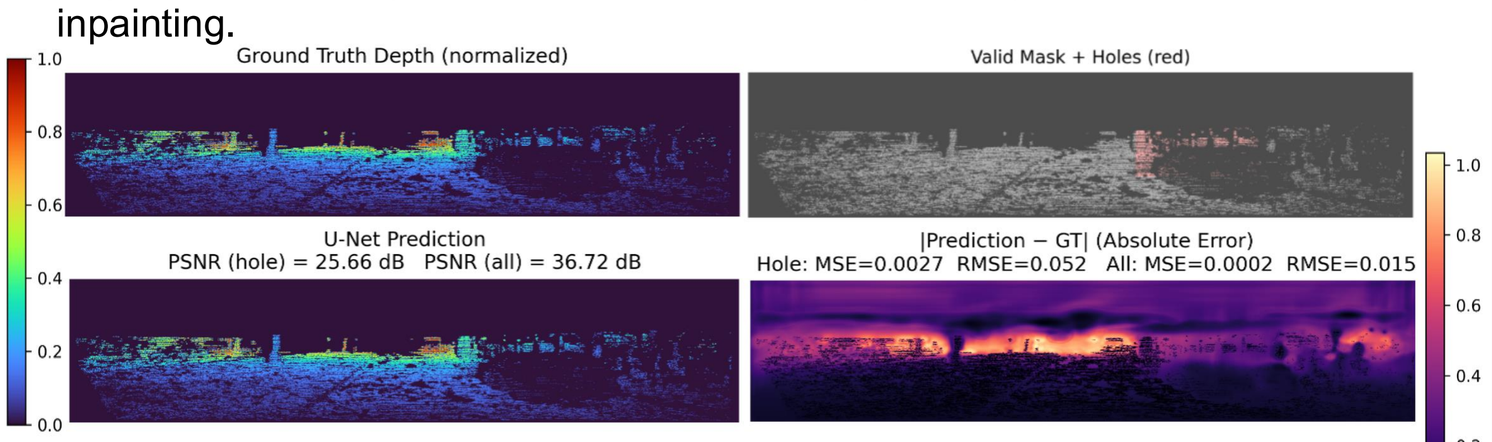
Experimental Results

We evaluate three variants of our architecture:
(1) U-Net w/ TV, **(2)** Autoencoder (no skip connections), **(3)** U-Net w/out TV.
λ=2.0, β=0.01 for all training.
Metrics are averaged across a held-out test set (80% of the validation set)

Model	RMSE (holes)	PSNR (holes)	RMSE (all)	PSNR (all)
U-Net (with TV)	0.068	24.7 dB	**0.036**	**30.3 dB**
Autoencoder	**0.067**	**24.9 dB**	0.038	29.3 dB
U-Net (No TV)	0.072	24.2 dB	0.039	29.2 dB

Key Findings:

- **Skip connections preserve global geometric structure:** Removing them gives similar hole-region RMSE but worse global RMSE and PSNR because the network must reconstruct large-scale depth gradients from heavily compressed features, losing fine spatial cues.
- **TV regularization improves stability**: Without TV, both hole and global RMSE degrade, confirming that smoothness priors significantly help outdoor depth inpainting.



In this example, **only the car region is masked**. Inside the masked region, the model achieves **RMSE of 0.052 (~4m error)** and **PSNR of 25.66 dB**. Globally, the image achieves **PSNR of 36.72 dB and RMSE of 0.015 (~1.2m error)**, indicating the reconstruction is visually plausible and comparable to KITTI Depth Completion leaderboard stats, which range from 0.67m - 2.31m [1].

TV encourages low-frequency continuity, which leads the network to interpolate smoothly across the sky/ground boundary. This produces visually plausible geometric transitions, even though the ground truth itself can be noisy at the horizon due to sensor sparsity.

The model **recovers the entire car structure** from only contextual cues, illustrating strong geometric priors learned from real KITTI depth.