Depth-Only Inpainting of Depth Maps for Autonomous Driving

Jagrit Rai
University of Toronto
jagrit.rai@mail.utoronto.ca

***Abstract*—Depth sensors in outdoor environments can produce large missing regions due to reflective surfaces, robot self-occlusion, and hardware limitations, degrading downstream tasks such as SLAM or autonomous driving which depend on reliable sensor measurements. Prior depth inpainting approaches typically rely on RGB guidance or indoor RGB-D statistics, limiting their applicability to outdoor LiDAR. This work investigates *depth-only* inpainting of large structured occlusions in the KITTI dataset. We simulate missing regions by applying random rectangular masks covering 5–20% of each depth map and train a U-Net to infer occluded geometry from surrounding depth context. The network is supervised with global L1 loss, hole-focused L1 loss, and total variation regularization, enabling reconstruction without RGB, monocular priors, or synthetic data. Experiments demonstrate that the model reconstructs plausible outdoor geometry, achieving a global RMSE of 2.88 m RMSE and PSNR of 30.3 dB on a held-out test set. These results show that depth-only inpainting is feasible for outdoor scenes.**

I. Introduction

Depth sensors can fail in real-world environments, producing incomplete depth maps that degrade downstream tasks such as autonomous driving, simultaneous localization and mapping (SLAM), and 3D reconstructions. Such failures can be caused by several factors: reflective surfaces (e.g., water, glass, chrome) can redirect emitted light away so that it never returns to the LiDAR sensor, sensor self-occlusion, where parts of a robot, drone, or vehicle block the sensor, preventing large portions of the scene from being observed, and hardware blind spots or other mechanical limitations may also produce missing regions.

Depth completion methods [1] address sparsity in LiDAR scans by interpolating between widely distributed valid depth samples, achieving strong performance on the KITTI benchmark. However, these models assume that the underlying scene is at least partially observed everywhere, and thus cannot reconstruct large structured occlusions where no valid measurements exist. Other depth inpainting approaches have been developed primarily for indoor RGB-D data and rely heavily on color guidance or smoothness priors, which limits their applicability to outdoor LiDAR scenes where RGB–depth alignment is unreliable and geometric complexity is significantly higher.

In this work, we study depth-only inpainting of large missing regions in outdoor LiDAR depth maps. We use the KITTI Depth Completion dataset and simulate structured occlusions by applying random rectangular masks covering 5-20% of each depth map. Our goal is to learn geometric priors directly from real outdoor depth and to reconstruct occluded regions based solely on surrounding depth context, without relying on RGB imagery, object masks, or synthetic supervision.

We propose a depth-only U-Net model trained with a combination of global L1 loss, hole-focused L1 loss, and total variation (TV) regularization. The network learns to infer missing geometry from spatial context and can reconstruct large occlusions despite the semi-dense and noisy nature of KITTI depth. Our approach requires no RGB guidance nor synthetic data.

This study is limited by the semi-dense nature of KITTI ground truth; we inpaint depth to the level of available supervision rather than producing fully dense maps. Additionally, we are limited by computational constraints that restrict training to a subset of KITTI. Nonetheless, our results demonstrate that depth-only encoder-decoder architectures can learn meaningful geometric priors from outdoor LiDAR data and can serve as a foundation for future work on large-occlusion depth reconstruction.

II. Related Work

Existing work relevant to depth inpainting spans RGB image inpainting, optimization-based depth restoration, and deep learning methods that operate directly on depth. These approaches typically rely on assumptions about dense texture, RGB-depth alignment, or indoor depth statistics that do not hold for outdoor LiDAR data. In this section, we outline key methods in these areas and explain how they differ from the depth-only, large-occlusion setting studied in this work.

A. RGB Image Inpainting

Classical image inpainting techniques focus on propagating local image structures to fill missing regions. Early approaches such as Telea's fast marching method and the Navier-Stokes method [2], [3] propagate local gradients or isophotes inward from the hole boundary to fill missing regions with plausible content. These techniques are effective for small gaps and texture continuation but operate purely in a 2D image-intensity space and cannot reason about underlying 3D structure. Edge-aware filters such as bilateral smoothing [4] enforce local smoothness and preserve sharp transitions, but also remain limited to interpolating photometric information and cannot infer missing geometry.

Deep learning approaches based on encoder–decoder architectures extend inpainting to large missing regions by learning

semantic priors. Yu et al. [5] introduce a contextual attention mechanism that allows the network to reference distant image regions when reconstructing complex textures and structures, generating visually plausible appearance by learning texture and semantic cues.

These works inspire the architecture of our approach, however their underlying assumptions do not transfer to outdoor LiDAR data. RGB image inpainting methods rely on image specific features such as texture, colour and semantic regularities, all of which are absent or unreliable in sparse outdoor depth maps. In contrast, depth-only inpainting requires learning geometric features like metric relationships, surface layouts, and structural continuity without relying on texture or colour.

Our method adapts the strengths of encoder-decoder architectures for reconstruction tasks to this fundamentally different domain, focusing on reconstructing large rectangular occlusions by operating directly on depth rather than appearance information.

B. Depth Inpainting

Classical depth inpainting approaches typically rely on explicit smoothness priors and optimization-based formulations. Herrera et al. [6] propose an inpainting method based on a second-order smoothness prior, which encourages piecewise planar surfaces while avoiding staircase artifacts produced by first-order regularizers. Their formulation incorporates RGB image gradients to prevent oversmoothing across object boundaries and produces high-quality results on Kinect-style RGB-D indoor depth maps. However, the method relies heavily on color guidance and assumes relatively simple planar geometry, making it less effective for outdoor LiDAR data, where depth and RGB are often misaligned and where geometric structure is more complex.

An optimization-based method using non-local Gauss–Markov random fields combined with superpixel-guided search windows is described in [7]. Their approach uses RGB segmentation to restrict non-local patch comparisons within object-consistent regions, and achieved SOTA results for small random holes and structured occlusions on indoor RGB-D benchmarks. Despite strong performance in indoor settings, the method depends on reliable RGB–depth alignment and on the presence of meaningful color boundaries, conditions that do not hold in outdoor LiDAR scenes such as KITTI.

These works demonstrate the value of smoothness priors and non-local constraints for depth inpainting, but both rely on RGB guidance and indoor depth statistics. Neither method directly addresses the challenge studied in this work: depth-only inpainting of large structured occlusions in outdoor LiDAR depth maps, where RGB cues are unavailable and geometric complexity is significantly higher. Our approach is therefore complementary, aiming to learn geometric priors directly from real outdoor depth without relying on color or handcrafted smoothness models.

[8] treats depth inpainting as an inverse reconstruction task solved through optimization; instead of predicting missing pixels directly, their approach utilizes a set of pretrained CNN denoisers as learned regularizers that impose natural image priors on the depth map. While their method operates without RGB guidance, the denoisers are trained on natural images rather than depth data, and their efficacy depends on dense pixel neighborhoods that do not exist in the semi-dense outdoor LiDAR maps of KITTI. As a result, the approach does not naturally extend to large structured occlusions or the highly sparse, irregular depth patterns characteristic of autonomous driving scenes.

Mao et al. [9] propose a depth-only inpainting method based on a U-Net architecture that learns multi-scale features from a single input depth map. Their approach does not rely on RGB guidance and instead trains an encoder–decoder network to predict missing depth values using a combination of hole-region losses and SSIM-based structural constraints. Evaluated on NYUv2, the method reconstructs small and moderate indoor holes effectively, demonstrating that convolutional architectures can learn meaningful geometric structure directly from depth. However, the approach is limited to indoor RGB-D datasets and irregular mask patterns, and is not designed to handle the large structured outdoor occlusions or the sparse LiDAR characteristics of KITTI. In contrast, our work focuses specifically on outdoor, large-occlusion, depth-only inpainting, learning geometric priors tailored to autonomous driving environments.

C. Depth Completion

Depth completion aims to predict a dense depth map from sparse LiDAR measurements and is a standard task on the KITTI benchmark introduced by Uhrig et al. [1]. Unlike depth inpainting which reconstructs large contiguous missing regions, depth completion assumes that valid depth samples are distributed throughout the image and focuses on interpolating between these sparse measurements. Few recent SOTA methods use solely depth data; [10] achieves highly accurate dense depth on KITTI using guidance from monocular foundation models and currently holds the number one spot on the leaderboard. The depth completion task differs fundamentally from our setting, where large structured occlusions must be reconstructed without guidance. We use the KITTI depth completion dataset for training and evaluation, but our problem formulation targets a distinct challenge.

III. Methods

In this section, we describe the depth-only inpainting approach for reconstructing large occluded regions in outdoor depth maps. We first formalize the inpainting problem and define notation. We then present the network architecture used to infer missing geometry from the semi-sparse input depth, followed by the loss functions that supervise the reconstruction, and finally we outline the training strategy used.

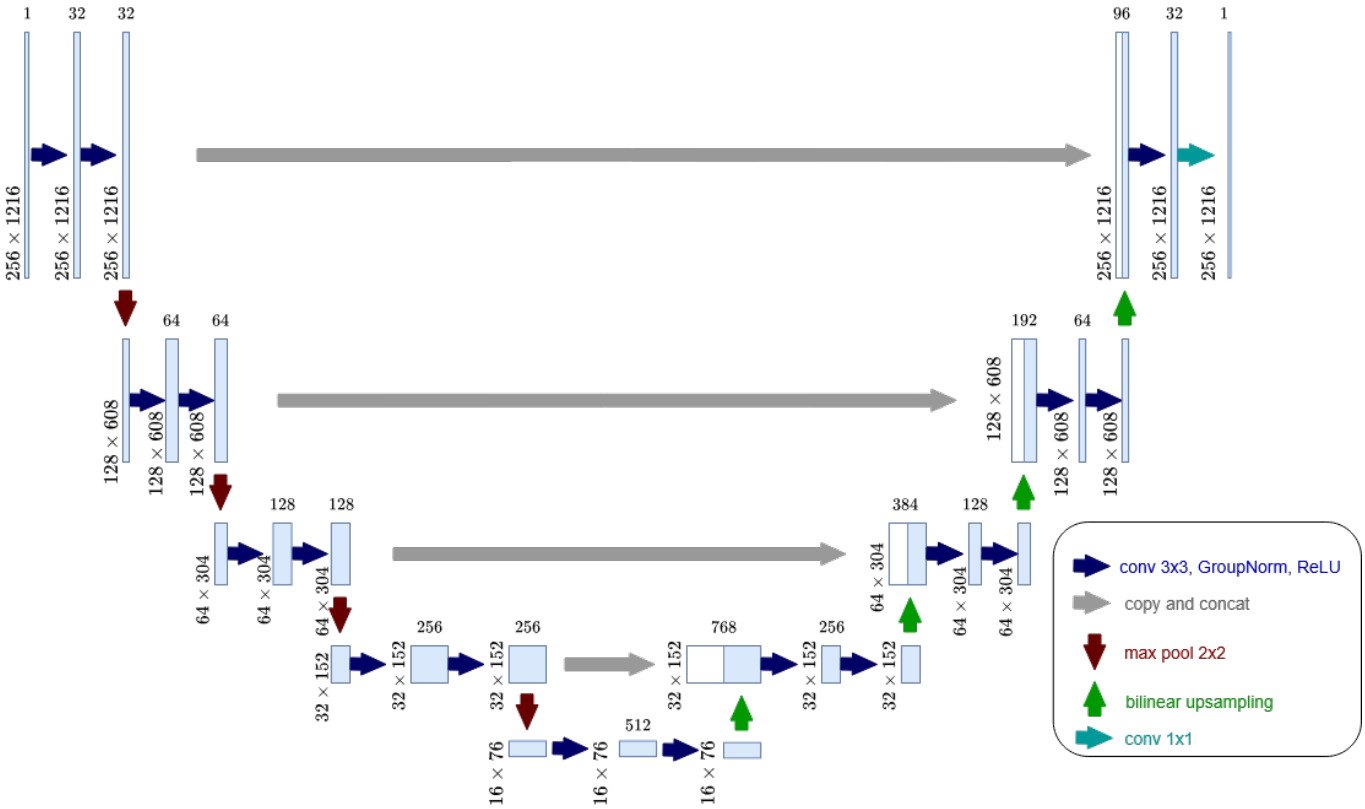

Fig. 1. U-Net architecture used for depth-only inpainting of outdoor scenes. The network takes a semi-sparse KITTI depth map with a simulated rectangular occlusion (5-20% of the image area removed) and predicts a reconstructed depth inside the missing region. The encoder-decoder structure with skip connections enables the network to combine global scene context with fine geometric detail. This figure illustrates the overall architecture.

A. Problem Formulation

We simulate large missing regions by applying random rectangular masks covering 5-20% of the image area to the KITTI ground-truth depth maps. The goal is to reconstruct plausible scene geometry within these occlusions using only surrounding depth information. For each image, let $D_{\text{gt}} \in \mathbb{R}^{H \times W}$ denote the ground-truth depth map, where pixels with no measurement are zero. We define

$$S = \{(i, j) \mid D_{\text{gt}}(i, j) > 0\},$$

be the set of all pixels where KITTI provides ground-truth depth. All loss terms described later are computed only over this set.

To simulate sensing failures such as self-occlusion, reflective surfaces, and blind spots, we introduce a second binary mask $H \in \{0, 1\}^{H \times W}$, where $H_{ij} = 1$ indicates a synthetic occlusion region and $H_{ij} = 0$ elsewhere. We generate H by sampling between one and three axis-aligned rectangles per image, each covering between 5% and 20% of the image area. The corrupted input depth map is then formed as

$$D_{\text{in}} = (1 - H) \odot D_{\text{gt}}, \tag{1}$$

where $\odot$ denotes element-wise multiplication. Inside the synthetic holes, D_{in} is set to zero, even if valid ground-truth depth exists in D_{gt}; the validity mask V is never modified and therefore still marks these pixels as supervised.

The goal of depth inpainting is to learn a function

$$\hat{D} = f_\theta(D_{\text{in}}) \tag{2}$$

that predicts a completed depth map $\hat{D} \in \mathbb{R}^{H \times W}$ which reconstructs plausible geometry in the occluded regions H while remaining consistent with the original measurements in V. In contrast to depth completion, which densifies sparse LiDAR measurements, this setting requires inferring unseen background geometry behind large contiguous holes where no depth samples are observed by the sensor.

B. Model Architecture

Our depth inpainting network is a U-Net encoder-decoder architecture [11] adapted for single-channel depth image reconstruction. The model receives a corrupted depth map as input and predicts a completed depth image $\hat{D}$ of the same spatial resolution. Figure 1 illustrates the full architecture.

The encoder consists of a sequence of convolutional blocks with progressively increasing channel width and spatial downsampling. Each block contains two 3×3 convolutions followed by Group Normalization and ReLU activations, allowing the

model to extract increasingly abstract geometric features each layer.

The decoder mirrors the encoder by a sequence of bilinear upsampling layers followed by convolutional blocks. At each decoder layer, the features are concatenated with the corresponding encoder activations via skip connections. These skip connections are used to preserve geometric details and fine-grained depth boundaries that might otherwise be lost through repeated downsampling. The concatenated feature maps are then refined using the same convolutional blocks used in the encoder.

To isolate the importance of skip connections, we also train an ablated Depth Autoencoder, which retains the same fundamental architecture as described but removes all encoder-decoder skip connections. This results in a strict bottlenecked network where all information must pass through the final deepest latent representation. Comparing this model with the full U-Net architecture enables us to quantify the contribution of skip connections to the geometric consistency and hole-region reconstruction quality.

C. Loss Functions and Training Objective

Our model is trained to reconstruct the missing depth within synthetic occlusion regions while remaining consistent with available ground-truth measurements. The total training loss combines three components described below.

1) Valid Pixel L1 Loss: The primary reconstruction term penalizes errors on valid ground-truth pixels. Since KITTI depth annotations are semi-dense, we compute the L1 error only at pixels where ground-truth depth is available. The valid-pixel L1 loss is then

$$\mathcal{L}_{\text{all}} = \frac{1}{|S|} \sum_{(i,j)\in S} \left| \hat{D}_{ij} - D_{\text{gt},ij} \right|. \tag{3}$$

This term ensures that the model learns accurate per-pixel reconstructions on all supervised regions of the scene.

2) Hole L1 Loss: The L1 hole loss is added on top of the global L1 loss in order to emphasize reconstruction inside the synthetically occluded regions. Without this term, it may be possible for the model to minimize error primarily outside of the hole and effectively "ignore" the missing region. The hole L1 loss is

$$\mathcal{L}_{\text{hole}} = \frac{1}{|S_{\text{hole}}|} \sum_{(i,j)\in S_{\text{hole}}} \left| \hat{D}_{ij} - D_{\text{gt},ij} \right|. \tag{4}$$

This encourages the network to focus its learning capacity on accurately reconstructing missing depth values inside the occluded regions.

3) Total Variation Smoothness: To promote locally smooth geometry while preserving edges, we apply a Total Variation (TV) prior. Given horizontal and vertical forward differences

$$\nabla_x \hat{D}_{ij} = \hat{D}_{i,j+1} - \hat{D}_{ij}, \qquad \nabla_y \hat{D}_{ij} = \hat{D}_{i+1,j} - \hat{D}_{ij},$$

the anisotropic TV penalty is defined as

$$\mathcal{L}_{\text{TV}} = \frac{1}{2} \left(\left\| \nabla_x \hat{D} \right\|_1 + \left\| \nabla_y \hat{D} \right\|_1 \right). \tag{5}$$

Total Variation smoothness encourages locally coherent depth predictions by penalizing abrupt changes between neighboring pixels. This is well-suited to real-world depth maps, where most surfaces (roads, walls, building facades, vehicles) are piecewise smooth with relatively small curvature.

4) Final Loss: The total training loss is a weighted sum of the three components:

$$\mathcal{L} = \mathcal{L}_{\text{all}} + \lambda\, \mathcal{L}_{\text{hole}} + \beta\, \mathcal{L}_{\text{TV}}. \tag{6}$$

In all experiments, we use $\lambda = 2.0$ and $\beta = 0.01$, which were found to provide a good balance between global accuracy, hole-region reconstruction quality, and smoothness.

Training details: We optimize the network using the Adam optimizer with a learning rate of 10^{-3}, batch size 4 (the maximum that our machine allowed without incurring Out-of-Memory error), and weight decay 10^{-6}. To reduce computational cost while still exposing the model to diverse scenes, we first train on a randomly selected 25% subset of the KITTI depth-annotated training split for 6 epochs, using the valid-pixel and weighted hole L1 losses ($\lambda = 2.0$) with TV regularization ($\beta = 0.01$). Depth values are normalized by a fixed maximum range so that both inputs and targets lie approximately in $[0, 1]$. We then warm-start the inpainting network from a separately trained depth autoencoder and fine-tune it for an additional 5 epochs on a larger 50% subset of the training data, using a smaller learning rate of 5×10^{-4}. During both stages, synthetic rectangular holes are regenerated on the fly for each sample and epoch. The best checkpoint is selected based on the validation RMSE measured only inside the hole regions.

IV. Results

We evaluate the proposed depth inpainting approach using quantitative metrics on a held-out test set consisting of 5,482 KITTI depth maps (80% of the official validation split), along with qualitative visualizations of representative scenes. Our experiments assess two aspects of the method: (1) the impact of architectural and loss-function choices through ablations of the U-Net model, and (2) the effectiveness of learned models relative to classical 2D inpainting baselines. We further analyze the model's behaviour on real KITTI scenes to understand its strengths and failure modes when inferring large missing regions of outdoor depth. The following subsections present quantitative and qualitative results supporting these findings.

A. Quantitative Results

We present quantitative comparisons across the learned models and classical baselines. Table I summarizes the quantitative performance of the three learned depth-inpainting models evaluated on the held-out KITTI test split. Each model was trained for six epochs on a randomly selected 25% subset of the dataset and then fine-tuned for five additional epochs on a separate 50% subset, using the 25% model as a warm start.

This training strategy was adopted to balance computational cost while still exposing the networks to a sufficiently diverse range of scenes. We evaluate three architectural variants: a full U-Net using skip connections and TV regularization, an autoencoder without skip connections, and a U-Net trained without the TV loss.

TABLE I
QUANTITATIVE EVALUATION ON KITTI DEPTH INPAINTING ACROSS U-NET VARIANTS AND AN AUTOENCODER BASELINE.

Model	RMSE (holes)	PSNR (holes)	RMSE (all)	PSNR (all)
U-Net (with TV)	5.44 m	24.7 dB	**2.88 m**	**30.3 dB**
Autoencoder	**5.36 m**	**24.9 dB**	3.04 m	29.3 dB
U-Net (No TV)	5.76 m	24.2 dB	3.12 m	29.2 dB

Across all metrics, the learned models substantially outperform the classical baselines (Table II), demonstrating that learning geometric priors from real data is crucial for this task. Comparing the learned models to one another reveals several important architectural effects. The autoencoder achieves the lowest RMSE inside the hole region (5.36 m), slightly outperforming the full U-Net (5.44 m). This behaviour is explained by the architectural differences; without skip connections, the model is forced to encode all relevant geometric cues into the latent bottleneck, encouraging the encoder to extract stronger global structural features that are relevant to reconstructing the hole-region, since the decoder cannot rely on local high-resolution passed from earlier layers. However, the autoencoder performs worse at global reconstruction, with a higher all-pixel RMSE (3.04 m vs. 2.88 m). This is expected: without the high resolution feature concatenation from skip connections, the decoder must reconstruct the entire image from heavily compressed latent information, leading to loss of fine-grained depth structure. Skip connections in the U-Net mitigate this issue by directly passing high-resolution spatial cues to the decoder, yielding superior global PSNR and RMSE despite slightly weaker performance within the hole region. While the autoencoder beats U-Net by 0.08 m in hole reconstruction, U-Net outperforms globally by 0.16 m, making U-Net preferable when overall scene accuracy and preservation of fine geometric structure are prioritized.

Training the U-Net without Total Variation regularization degrades both hole-region metrics and global metrics. Without TV, predictions exhibit noisier depth gradients and less geometric coherence, increasing RMSE by roughly 0.3–0.4 m across evaluation regions. This confirms that a smoothness prior is valuable for outdoor depth inpainting, where large surfaces (roads, facades, vehicles) are typically piecewise smooth. Among all variants, the full U-Net with TV provides the best global reconstruction quality, while the autoencoder offers insight into the role of skip connections for recovering occluded depth.

As a sanity-check baseline, we also evaluate classical inpainting methods on our held-out test set, with quantitative results shown in Table II. These methods were expected to perform poorly on this task, as they are designed for RGB image inpainting and do not attempt to model 3D structure or geometric continuity. We use the OpenCV implementations of Telea and Navier-Stokes inpainting, only modifying pixels inside the hole region and keeping the remaining pixel values the same. We additionally include an iterative bilateral inpainting baseline, implemented by repeatedly applying OpenCV's bilateral filter and copying filtered values into the hole while keeping known depth fixed; this follows the edge-preserving smoothing behavior originally introduced by [4] and serves as a simple depth-only interpolation method. Across all three baselines, performance is substantially worse than any learned model, particularly within the hole regions where geometric inference is required: Telea and Navier–Stokes reach hole RMSE values of approximately 16.5 m, and the bilateral method remains above 14 m, compared to 5.4 m for our best U–Net model. Because none of these classical techniques incorporate learned priors or any understanding of scene geometry, they are unable to reconstruct plausible depth structure when large regions of the map are missing.

TABLE II
COMPARISON OF CLASSICAL DEPTH INPAINTING METHODS ON KITTI, EVALUATED AGAINST OUR BEST LEARNED MODEL (U–NET WITH TV).

Method	RMSE (holes)	PSNR (holes)	RMSE (all)	PSNR (all)
Telea (fast marching)	16.49 m	14.59 dB	8.28 m	20.95 dB
Navier–Stokes inpainting	16.52 m	14.56 dB	8.30 m	20.93 dB
Iterative bilateral filter	13.96 m	16.25 dB	7.04 m	22.61 dB
U-Net (with TV, ours)	**5.44 m**	**24.70 dB**	**2.88 m**	**30.30 dB**

B. Qualitative Results

Figure 2 presents a qualitative example of the depth inpainting performance of our U-Net model with TV regularization. In this example, a large synthetic occlusion is applied over the car region in the semi-dense KITTI depth map. The U-Net successfully reconstructs the geometry of the occluded object and the surrounding road surface using only contextual depth cues. The predicted depth map closely matches the ground-truth structure, and the absolute error visualization highlights that most errors remain small and spatially localized. Errors naturally concentrate along depth discontinuities (such as object boundaries and the horizon) where minor misalignments between the predicted and ground-truth depth are amplified by steep depth gradients. Within the masked region, the model achieves an RMSE of 0.052 (approximately 4 m) and PSNR of 25.66 dB, indicating strong inpainting accuracy for large missing regions. Over the full image, the model achieves an RMSE of 0.015 (1.2 m) and PSNR of 36.72 dB, demonstrating that global scene geometry is well preserved.

The strengths of the model are most evident in the accurate reconstruction of the vehicle silhouette and roadway geometry. The model recovers the car's overall shape, relative depth, and boundaries despite having no depth measurements inside the masked region. This illustrates the strong geometric priors learned by the network from real KITTI data. Total Variation regularization enhances this behaviour by encouraging smooth, low-frequency depth transitions across the occlusion, which

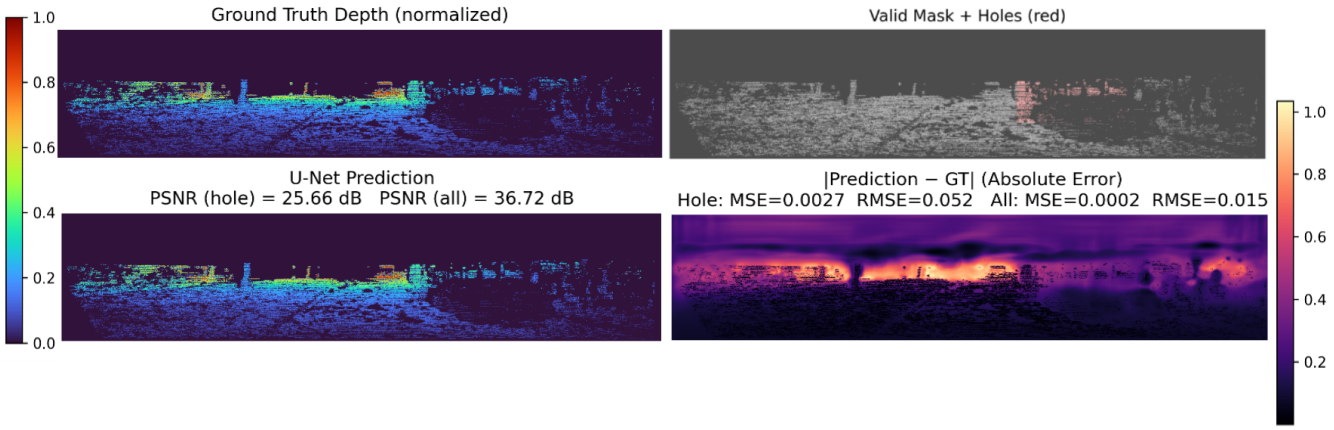

Fig. 2. Qualitative depth inpainting results on a KITTI sample. Top-left: semi-dense ground-truth depth (normalized). Top-right: valid-pixel mask with synthetic hole regions highlighted in red. Bottom-left: U-Net prediction trained using our depth-only inpainting U-Net with TV. Bottom-right: absolute error map showing reconstruction accuracy inside and outside the hole region. Reported PSNR and **normalized** RMSE values are computed separately for hole pixels and for all valid ground-truth pixels.

allows the network to interpolate plausible structure while still preserving major depth discontinuities.

The primary failure mode appears near the sky/ground boundary on the left side of the scene, where the absolute error map shows increased error relative to other regions. This is expected: KITTI's fused ground truth is sparse and noisy near the horizon, where LiDAR returns become unreliable and stereo matching fails. Both bilinear upsampling in the decoder and the TV smoothness prior can also contribute to oversmoothing and slight depth bias in these large low-texture regions. Despite these limitations, the qualitative results demonstrate that the U-Net produces visually plausible and structurally coherent reconstructions using only geometric context, successfully inpainting large missing regions without RGB guidance.

V. Discussion

Our results demonstrate that a depth-only encoder-decoder network can learn meaningful geometric features from outdoor LiDAR data and reconstruct large missing regions using only surrounding depth context. This shows that large-occlusion inpainting is feasible in outdoor scenes, even without RGB guidance. More broadly, the work highlights an open research gap: unlike depth completion, outdoor depth inpainting lacks established benchmarks, standardized mask distributions, and comparable baseline methods, whereas RGB image inpainting is a similar task and a well established field of study. Finally, because self-occlusion and LiDAR dropout are common in robotics, depth-only inpainting has potential practical utility as a preprocessing step for downstream perception tasks, enabling systems to recover plausible geometry in unobserved regions of the scene. Below, we address areas of future work, as well as limitations this study faces.

A. Limitations

Several limitations constrain the current applicability of this work. First, although the model achieves strong qualitative performance, reconstruction errors inside large occlusions are on the order of several metres. While this is reasonable given the scale of the training scenes and the absence of direct observations within the occluded regions, such errors are likely too large for safety-critical tasks such as autonomous navigation. In addition, training was limited to a subset of the KITTI dataset due to computation constraints, and a more exhaustive hyperparameter exploration (particularly of the hole-loss and TV-regularization weights) were not feasible, but may yield further improvements.

The approach also inherits limitations from the dataset and inpainting setup. Our inpainting masks are synthetic rectangular occlusions, which approximate but do not fully capture the structure of real-world sensor failures such as LiDAR dropout patterns, specular reflections, or self-occlusion from ego-vehicle geometry. Because KITTI provides only semi-dense supervision, our method reconstructs missing regions up to the density of the fused ground truth rather than producing a fully dense depth map. Moreover, the model operates solely on depth and lacks semantic or photometric cues; without RGB guidance or learned appearance features, it cannot resolve ambiguous structures or fine-grained object detail, and its predictions may oversmooth depth discontinuities due to TV regularization and decoder upsampling.

A further limitation of this study is in the lack of established baselines for outdoor depth inpainting. As discussed earlier, most prior depth inpainting work focuses on small indoor scenes or object-centric datasets, and very few methods are applicable to large structured occlusions in outdoor LiDAR depth. In contrast to depth completion, where standardized benchmarks enable consistent comparison across methods, depth inpainting on outdoor LiDAR data has no established

benchmark and few comparable models. Classical RGB inpainting algorithms are only used as sanity-check references, but are not designed to address the underlying geometric reasoning required by this task. The closest point of reference is depth completion literature, where models must also learn geometric priors to interpolate between sparse LiDAR points and are benchmarked using RMSE against semi-dense ground truth. While completion and inpainting are fundamentally different tasks, depth completion performance provides a rough sense of the achievable error range on KITTI and helps contextualize the reconstruction accuracy of our approach. In the qualitative example shown in Figure 2, the reconstructed image achieves a PSNR of 36.72 dB and an RMSE of 0.015 (approximately 1.2 m), indicating that the reconstruction is visually plausible and falls within the range of RMSE values reported by the best performing models on the KITTI depth completion challenge leaderboard, which span 0.67 m to 2.31 m [1]. As shown in Table I, our best model achieves a global RMSE of 2.66 m on our held-out test set. While depth completion cannot serve as a direct baseline for inpainting, these comparisons nonetheless demonstrate that our model achieves reconstruction errors on the same order of magnitude as established KITTI methods, highlighting the effectiveness of the learned geometric priors for outdoor depth inference.

B. Future Work

Several promising directions exist for future work. More realistic occlusion patterns could be simulated using semantic segmentation rather than axis-aligned rectangles. A related line of work is a KITTI-based LiDAR inpainting model designed for static background map generation by removing segmented foreground objects [12]. Although operating on a similar range-image representation, its objective and mask distribution differ significantly from ours, and integrating such object-aware inpainting frameworks represents promising future work that we were unable to explore within the scope of this project. Further, extending the model to operate over video sequences could enable multi-frame reasoning and geometric cues unavailable in single images. Diffusion-based generative models or monocular-depth-guided architectures as in [10] also represent interesting avenues for improving reconstruction fidelity. Finally, jointly performing depth completion and inpainting within a unified framework could allow the model to exploit both sparse LiDAR structure and large occlusion reasoning simultaneously.

VI. Conclusion

This work presents a depth-only approach for inpainting large rectangular occlusions in KITTI ground truth depth maps, addressing a gap left by existing RGB-guided or indoor focused methods. By training a U-Net architecture with hole-aware L1 supervision and TV regularization on the KITTI dataset, we demonstrated that meaningful geometric priors can be learned directly from semi-dense outdoor depth. Our experiments show that the model reconstructs plausible scene structure from depth context alone and achieves error levels near the range of KITTI depth completion models, despite tackling a significantly harder problem. While limitations remain, including reconstruction accuracy, reliance on synthetic mask patterns, and lack of semantic cues, the results indicate that such depth-only inpainting is feasible and potentially valuable for robotics or other sensing applications. Future work should explore more realistic occlusion distributions, joint completion–inpainting frameworks, and the development of standardized benchmarks to advance depth-only inpainting.

References

[1] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, and A. Geiger, "Sparsity invariant cnns," in *2017 international conference on 3D Vision (3DV)*. IEEE, 2017, pp. 11–20.

[2] A. Telea, "An image inpainting technique based on the fast marching method," *Journal of graphics tools*, vol. 9, no. 1, pp. 23–34, 2004.

[3] M. Bertalmio, A. L. Bertozzi, and G. Sapiro, "Navier-stokes, fluid dynamics, and image and video inpainting," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 1. IEEE, 2001, pp. I–I.

[4] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Sixth international conference on computer vision (IEEE Cat. No. 98CH36271)*. IEEE, 1998, pp. 839–846.

[5] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5505–5514.

[6] D. Herrera C, J. Kannala, L. Ladickỳ, and J. Heikkilä, "Depth map inpainting under a second-order smoothness prior," in *Scandinavian conference on image analysis*. Springer, 2013, pp. 555–566.

[7] S. Satapathy and R. R. Sahay, "Robust depth map inpainting using superpixels and non-local gauss–markov random field prior," *Signal Processing: Image Communication*, vol. 98, p. 116378, 2021.

[8] Z. Li and J. Wu, "Learning deep cnn denoiser priors for depth image inpainting," *Applied Sciences*, vol. 9, no. 6, p. 1103, 2019.

[9] J. Mao, J. Li, F. Li, and C. Wan, "Depth image inpainting via single depth features learning," in *2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. IEEE, 2020, pp. 116–120.

[10] Y. Liang, Y. Hu, W. Shao, and Y. Fu, "Distilling monocular foundation model for fine-grained depth completion," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 22 254–22 265.

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[12] J. Lee, S. Hwang, W. J. Kim, and S. Lee, "Sam-net: Lidar depth inpainting for 3d static map generation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 12 213–12 228, 2021.