Diffusion-Based Depth Reconstruction from Sparse LiDAR Measurements

Aviad Golan Peretz

Abstract—Monocular imaging systems inherently suffer from depth ambiguity, limiting their reliability in robotic and autonomous systems where safety requires accurate three-dimensional environment perception. Traditional approaches attempt to address this limitation using stereo vision systems or learning-based monocular depth estimation. However, stereo systems typically rely on computationally expensive geometric algorithms, while monocular learning-based methods often struggle to preserve fine geometric structures and object boundaries. LiDAR sensors can provide accurate depth measurements, but dense LiDAR systems require significant power and computational resources. In this work, we leverage recent advances in generative artificial intelligence, specifically diffusion models, as a prior for solving the sparse-to-dense depth reconstruction problem. Our approach combines sparse LiDAR measurements with RGB imagery and uses a conditional diffusion process to iteratively refine dense depth estimates. By exploiting the strong generative prior learned by diffusion models, the proposed method produces more accurate and geometrically consistent depth maps compared to conventional learning-based monocular depth estimation methods.

Index Terms—Computational Photography, Deep Learning, Diffusion Models, Perception

1 Introduction

2 Introduction

Navigation is a fundamental and challenging task in robotics, particularly for safety-critical systems that require precise knowledge of the surrounding three-dimensional environment. Reliable navigation depends on accurate 3D maps, which in turn require depth information that cannot be directly obtained from monocular RGB sensors. Bridging this gap for vision-based navigation systems has therefore been an active area of research.

Prior to the recent advances in artificial intelligence, depth estimation was commonly addressed using classical stereo and multi-view geometry methods such as Structure-from-Motion (SfM). These approaches rely on solving large-scale optimization problems and typically require computationally intensive pipelines, for example systems such as COLMAP [1]. While effective in offline settings, these methods are often too expensive for real-time deployment on embedded robotic platforms.

With the rapid development of deep learning for vision tasks, learning-based monocular depth estimation has gained significant attention. Convolutional architectures, particularly U-Net–based models [2], have shown promising results in predicting dense depth maps directly from RGB images. However, purely monocular methods inherently suffer from scale ambiguity and often struggle to preserve fine geometric structures and object boundaries [3], [4].

Recent advances in generative modeling have demonstrated that diffusion models provide powerful priors for solving inverse problems in imaging. Diffusion-based methods have achieved strong performance in tasks such as image reconstruction, inpainting, and restoration [5], [6]. Motivated by these developments, this work explores the use of diffusion models as a generative prior for depth reconstruction.

Specifically, we propose a framework that reconstructs dense depth maps by combining monocular RGB imagery with sparse LiDAR measurements. The sparse LiDAR observations provide non-ambiguous geometric constraints that guide the diffusion process, while the learned generative prior allows the model to infer dense depth structure across the scene. This hybrid approach reduces the reliance on dense LiDAR sensors, which are often power-intensive and impractical for many embedded robotic systems, while also avoiding the heavy optimization pipelines required by traditional stereo methods.

The main contributions of this work are summarized as follows:

- A diffusion-based framework is developed for sparse-to-dense depth reconstruction that combines monocular RGB imagery with sparse LiDAR measurements.
- Sparse measurement is applied as a consistency constraint during reverse diffusion that enforces agreement with LiDAR observations while allowing the generative prior to reconstruct missing depth regions.
- Demonstrated through experiments that the proposed method significantly outperforms a conventional U-Net baseline across multiple reconstruction metrics.
- Modality ablation study is conducted to highlight the complementary roles of RGB imagery and sparse LiDAR measurements in resolving geometric ambiguities.

- *M. Shell is with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 30332. E-mail: see http://www.michaelshell.org/contact.html*
- *J. Doe is with Anonymous University.*

3 Related Work

3.1 Monocular Depth Estimation

Monocular depth estimation aims to recover dense depth maps from a single RGB image. Early deep learning approaches demonstrated that convolutional neural networks could learn to predict depth directly from monocular imagery [3]. While these methods achieve impressive results in many scenarios, monocular depth estimation inherently suffers from scale and depth ambiguity because a single image does not uniquely determine the underlying three-dimensional geometry of a scene, and is only correct up to a scale. As a result, monocular models often struggle in environments where precise geometric reconstruction is required, particularly for safety-critical robotic navigation tasks.

3.2 Sparse-to-Dense Depth Completion

To address the limitations of monocular depth estimation, sparse-to-dense depth completion methods combine RGB imagery with sparse LiDAR measurements. These approaches use convolutional neural networks to fuse sparse depth samples with visual features to predict dense depth maps [4]. While effective in many settings, these methods can produce overly smooth reconstructions and may fail to accurately preserve sharp geometric structures or model uncertainty in regions with limited measurements, as will be shown in the experiments section of this paper.

3.3 Diffusion Models for Inverse Problems

Recent advances in generative modeling have introduced diffusion models as powerful tools for solving inverse problems in imaging. Score-based generative models learn the gradient of the data distribution and enable the reconstruction of high-dimensional signals through a stochastic reverse diffusion process [5]. These methods have demonstrated strong performance across a range of image restoration tasks, including denoising, inpainting, and super-resolution. More recently, diffusion models have been applied as generative priors for solving inverse problems in imaging systems, producing high-quality reconstructions even in severely under-constrained settings [6]. These developments motivate the use of diffusion models as priors for sparse-to-dense depth reconstruction in robotic perception systems. Additionally, recent advances in Graphics Processing Units (GPUs) enable deployment of such models onto embedded systems such as Waymo cars. While inference of these models can take many FLOPS, an active research area in the field of diffusion models explore ways to reduce the number of steps required in the reverse diffusion process, to allow for faster, more efficient inference. All of these advances together provide a path forward for deployment of such model on real physical systems.

4 Proposed Method

This work proposes a diffusion-based framework for reconstructing dense depth maps from monocular RGB imagery and sparse LiDAR measurements. The proposed method uses a conditional diffusion model as a generative prior for depth reconstruction. During training, the model learns to predict clean depth maps from noisy observations. During inference, the reverse diffusion process iteratively refines a depth estimate while enforcing consistency with sparse LiDAR measurements. The high-level approach is shown in 1 to provide a simplistic view for the ease of the reader.

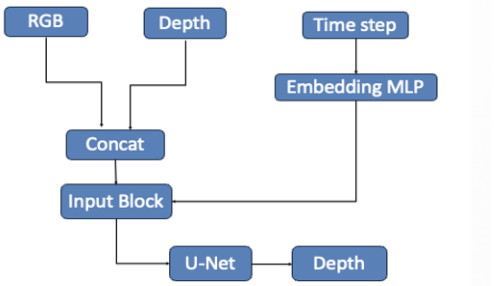

Fig. 1. High-level solution approach. The RGB images and depth maps are being concatenated and the signal is going through an input block together with time embeddings, which is then provided to the UNet for depth map estimation.

4.1 Problem Formulation

Let $I \in \mathbb{R}^{3\times H\times W}$ denote an RGB image and $D_0 \in \mathbb{R}^{1\times H\times W}$ the corresponding ground truth depth map. In addition, we assume sparse depth measurements obtained from a LiDAR sensor, denoted by

$$D_{\mathrm{s}}.$$

A binary measurement mask M is defined as

$$M = \mathbf{1}(D_{\mathrm{s}} > 0),$$

which identifies pixels where LiDAR measurements are available.

The objective is to reconstruct a dense depth map D that is both consistent with the sparse measurements and consistent with the learned depth distribution conditioned on the RGB image.

4.2 Conditional Diffusion Model

The proposed method models the distribution of depth maps using a diffusion process. During training, noise is gradually added to the ground truth depth map using a predefined variance schedule $\{\beta_t\}_{t=1}^{T}$. The forward diffusion process generates noisy depth maps

$$D_t = \sqrt{\bar{\alpha}_t} D_0 + \sqrt{1-\bar{\alpha}_t}\epsilon,$$

where

$$\alpha_t = 1 - \beta_t, \quad \bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i,$$

and ϵ is sampled from a standard Gaussian distribution $\epsilon \sim \mathcal{N}(0, I)$.

The network is trained to predict the clean depth map D_0 from the noisy observation D_t conditioned on the RGB image:

$$\hat{D}_0 = f_\theta(I, D_t, t),$$

where f_θ denotes the diffusion model.

4.3 Network Architecture

The diffusion model is implemented as a conditional U-Net architecture. The network receives the RGB image and the noisy depth map as input and predicts the clean depth estimate.

The RGB image I and noisy depth D_t are concatenated along the channel dimension to form a four-channel input tensor. This tensor is processed by an encoder–decoder architecture composed of residual convolutional blocks.

The encoder consists of a sequence of residual blocks followed by downsampling operations, progressively reducing the spatial resolution while increasing feature dimensionality. The network includes three downsampling stages. A bottleneck module composed of two residual blocks operates at the lowest spatial resolution.

The decoder mirrors the encoder using transposed convolution layers to progressively restore spatial resolution. Skip connections link encoder and decoder features at corresponding scales, allowing high-frequency spatial information to propagate through the network. A final fusion block combines decoder features with early encoder representations before producing the final output through a 1×1 convolution layer.

To condition the network on the diffusion timestep, a sinusoidal timestep embedding is computed and injected into each residual block using a learned linear projection. This conditioning allows the network to adapt its predictions to the current diffusion stage.

4.4 Training Objective

The model is trained using an x_0-prediction formulation, where the network directly predicts the clean depth map from the noisy input. The training objective combines an L_1 reconstruction loss and a mean squared error loss:

$$\mathcal{L} = \|D_0 - \hat{D}_0\|_1 + \frac{1}{2}\|D_0 - \hat{D}_0\|_2^2.$$

The model parameters are optimized using the Adam optimizer.

4.5 Sparse-Guided Reverse Diffusion

At inference time, the reverse diffusion process is used to iteratively reconstruct a dense depth map. The process begins from a random depth sample

$$D_T \sim \mathcal{N}(0, I).$$

At each diffusion step, the model predicts the clean depth estimate

$$\hat{D}_0 = f_\theta(I, D_t, t).$$

A DDIM-style update is then used to compute the next depth estimate

$$D_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\hat{D}_0 + \sqrt{1 - \bar{\alpha}_{t-1}}\hat{\epsilon},$$

where $\hat{\epsilon}$ is recovered from the predicted clean depth ($\hat{\epsilon}$ is the noise component).

To enforce consistency with the LiDAR measurements, the depth estimate is projected onto the sparse observations:

$$D_{t-1} = (1 - M) \odot D_{t-1} + M \odot (\lambda D_{\text{sparse}} + (1 - \lambda)D_{t-1}).$$

This projection step ensures that the reconstructed depth map remains consistent with the available measurements while allowing the diffusion model to infer missing regions. And M, again, is the measurement mask.

The final output D_0 represents the reconstructed dense depth map. Figure 2 provides a visual for the inference process.

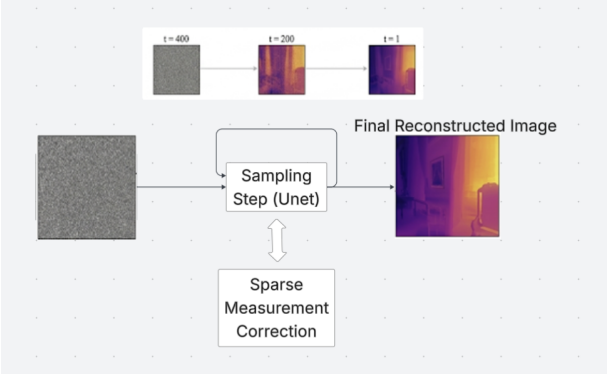

Fig. 2. Reverse Diffusion process for the inverse problem solution.

5 Experimental Results

This section evaluates the proposed diffusion-based depth reconstruction framework and compares it against a supervised U-Net baseline. The experiments are designed to measure reconstruction accuracy, robustness to sparse measurements, and the contribution of each sensing modality.

5.1 Experimental Setup

All the data used for training and evaluating during this project was generated through the Habitat simulation environment [7]. All experiments are conducted using paired RGB images and ground-truth depth measurements from the simulation environment built-in depth sensor simulators. To simulate realistic sparse LiDAR measurements, a sub-sampling process was applied keeping one measurement every 10 pixels in both spatial directions. This was arbitrarily chosen as a sparse measurements. There were 4000 samples used for the training due to the limited compute resources in this project.

The proposed diffusion model is trained for 80 epochs using the Adam optimizer with a learning rate of 10^{-4} and a batch size 16. The diffusion process uses $T = 400$ time steps with a cosine noise schedule. During training, the network predicts the clean depth map x_0 from a noisy observation x_t conditioned on the RGB image. The loss function is a weighted combination of L1 and mean squared error (MSE):

$$\mathcal{L} = \|\hat{x}_0 - x_0\|_1 + 0.5\|\hat{x}_0 - x_0\|_2^2 \quad (1)$$

which was found to stabilize optimization for small datasets (L2) and also try to introduce sparsity in the solution (L1).

During inference, depth maps are reconstructed using a deterministic DDIM-style reverse diffusion process with up to 200 sampling steps.

5.2 Diffusion Reconstruction

The proposed method reconstructs dense depth maps by combining the diffusion prior with sparse LiDAR measurements. During reverse diffusion, the predicted depth map is projected onto the known LiDAR measurements using a simple consistency constraint:

$$x = (1 - \mathrm{M}) \cdot x + \mathrm{M} \cdot (\lambda_{\text{data}} D_s + (1 - \lambda_{\text{data}}) x) \qquad (2)$$

where M is a binary mask indicating valid LiDAR measurements and λ_{data} controls the strength of measurement enforcement. In these experiments we use $\lambda_{\text{data}} = 1.0$, corresponding to hard projection onto the sparse measurements, so that the provided LiDAR measurements do not get "diluted" during the diffusion process.

The encoder begins with an initial residual block that maps the four-channel input to a base feature dimensionality of 64 channels. The network then progressively increases the number of feature channels while reducing spatial resolution through strided convolution layers. Specifically, the encoder stages expand the representation from $64 \rightarrow 128 \rightarrow 256$ channels while downsampling the spatial resolution by factors of two, ultimately producing a feature representation at $1/8$ of the input resolution. Each stage consists of residual convolutional blocks using 3×3 convolutions, Group Normalization, and SiLU activation functions, with the timestep embedding added through a learned MLP.

At the bottleneck, two residual blocks operate on the lowest-resolution feature representation (256 channels), enabling the network to capture global context across the scene. The decoder then mirrors the encoder structure.

Finally, a fusion block combines the highest-resolution encoder and decoder features (128 channels) and maps them back to the base feature dimensionality of 64 channels. The reconstructed depth map is produced using a 1×1 convolution layer applied after Group Normalization and SiLU activation, yielding a single-channel prediction of the clean depth map x_0.

5.3 Evaluation Metrics

Reconstruction performance is evaluated using several standard depth estimation metrics:

- Mean Absolute Error (MAE)
- Root Mean Square Error (RMSE)
- Peak Signal-to-Noise Ratio (PSNR)
- Structural Similarity Index (SSIM)

These metrics capture both pixel-level accuracy and perceptual reconstruction quality, as well as signal quality (PSNR).

5.4 Baseline: U-Net Depth Estimation

As a baseline, U-Net architecture for monocular depth estimation was used alone (without any diffusion process attached to it). The network follows an encoder–decoder design with symmetric skip connections that preserve spatial information across multiple feature scales, as the original UNet paper [2].

The encoder progressively reduces the spatial resolution using max-pooling, while increasing the number of feature channels through Conv2D blocks. Each Conv2D block consists of two 3×3 convolution layers followed by batch normalization and ReLU activation. Starting from the RGB input, the feature dimensionality is increased from 32 to 256 channels across successive downsampling stages.

The decoder mirrors the encoder structure and reconstructs the spatial resolution using transposed convolution layers. At each stage, feature maps from the corresponding encoder level are concatenated via skip connections, allowing the network to retain fine spatial details that may otherwise be lost during down-sampling.

The network outputs a dense depth map through a final 1×1 convolution layer that maps the reconstructed feature representation to a single-channel depth prediction.

To evaluate the effectiveness of the proposed diffusion-based reconstruction framework, we compare it against a supervised U-Net baseline trained to directly predict dense depth maps from RGB images.

Table 1 reports the reconstruction performance of both models using standard depth estimation metrics.

The diffusion model consistently outperforms the U-Net baseline across all metrics. In particular, the proposed method reduces the mean absolute error (MAE) by nearly 50%, indicating substantially improved reconstruction accuracy. The improvements in SSIM and PSNR further suggest that the diffusion model better preserves structural details and depth discontinuities, which are often smoothed out by deterministic convolutional architectures.

TABLE 1
Quantitative comparison between the baseline U-Net and the proposed diffusion model for sparse-to-dense depth reconstruction over 100 validation samples. Lower is better for MAE and RMSE, while higher is better for PSNR and SSIM.

Model	MAE	RMSE	PSNR	SSIM
U-Net Baseline	0.0278	0.0537	28.16	0.823
Diffusion (ours)	**0.0141**	**0.0352**	**33.45**	**0.949**

Additionally, the diffusion-based method achieves a higher peak signal-to-noise ratio (PSNR), suggesting that the generated depth maps are closer to the ground truth in terms of overall reconstruction fidelity. These results highlight the advantage of using diffusion models as generative priors for inverse problems, allowing the model to iteratively refine depth estimates while maintaining global consistency.

Overall, the results demonstrate that the proposed approach produces more accurate and structurally coherent depth reconstructions compared to a conventional convolutional encoder–decoder architecture. Visually, the results also show the clear superiority of the diffusion-based UNet in its capability of edge preserving.

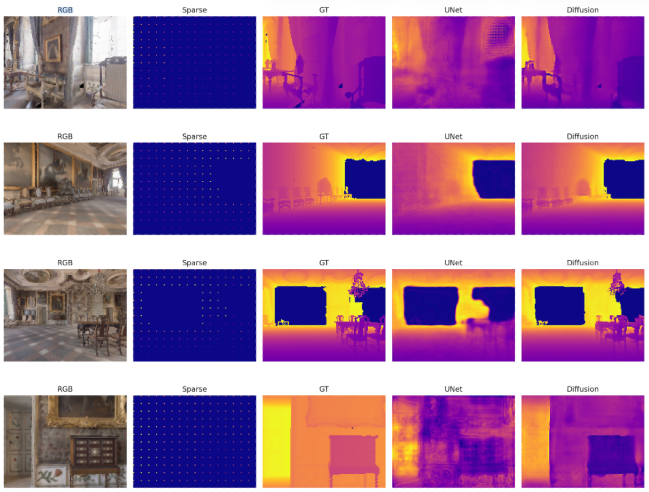

Fig. 3. Comparison between the baseline U-Net and the diffusion-based approach to solve the inverse depth-map problem

5.5 Modality Ablation Study

To better understand the contribution of each sensing modality, a modality ablation study comparing three configurations was conducted:

- RGB + LiDAR (full system)
- LiDAR only
- RGB only

The diffusion reconstruction process is evaluated under several sampling configurations, varying the number of reverse diffusion steps and stochasticity parameter η. Specifically:

- 100, 200, and 400 reverse diffusion steps
- $\eta \in \{0.0, 0.05\}$

For fair comparison, all configurations use the same random initialization and diffusion sampling parameters. Reconstructions are generated using identical diffusion schedules, and reconstruction errors are measured relative to the ground-truth depth maps.

This experiment highlights the complementary roles of visual context and sparse depth measurements in resolving geometric ambiguities.

Figure 4 shows the results of the ablation study, emphasizes the need for an RGB sensor on board. Clearly, the lack of RGB seems impossible to overcome by the neural network.

The table with the corresponding MAE values is shown in 2

TABLE 2
Modality ablation study. Mean Absolute Error (MAE) for different sensing configurations and diffusion sampling parameters.

η	Steps	λ_{data}	RGB + LiDAR	LiDAR Only	RGB Only
0.0	100	1.0	**0.0224**	0.2937	0.0228
0.0	200	1.0	**0.0224**	0.2937	0.0227
0.0	400	1.0	**0.0223**	0.2937	0.0227
0.05	200	1.0	**0.0227**	0.2936	0.0230

6 Conclusion

This work presented a diffusion-based framework for reconstructing dense depth maps from monocular RGB imagery and sparse LiDAR measurements. By leveraging diffusion models as generative priors, the proposed method is able to iteratively refine depth estimates while enforcing consistency with sparse geometric observations. The approach combines the strengths of learning-based vision methods with the reliability of LiDAR measurements, enabling accurate depth reconstruction without requiring dense LiDAR sensors or computationally expensive multi-view reconstruction pipelines.

Experimental results demonstrate that the proposed diffusion-based model significantly outperforms a conventional U-Net baseline across multiple evaluation metrics, including MAE, RMSE, PSNR, and SSIM. In particular, the method achieves nearly a twofold reduction in mean absolute error while producing depth maps with improved structural fidelity and edge preservation. The modality ablation study further highlights the complementary roles of RGB imagery and sparse depth measurements, showing that visual context is essential for resolving depth ambiguities while sparse LiDAR observations provide reliable geometric constraints.

These results indicate that diffusion models provide a powerful prior for solving sparse-to-dense depth reconstruction problems in robotic perception systems. By combining sparse measurements with learned generative priors, the proposed framework offers a promising direction for improving depth estimation in resource-constrained robotic platforms.

Future work will focus on extending this approach to larger datasets and real-world sensor data, as well as exploring more efficient diffusion sampling techniques to reduce inference time. Such improvements could enable real-time deployment of diffusion-based depth reconstruction methods in autonomous robotic systems.

Acknowledgments

I would like to thank EE367 course staff for the useful reference material provided, and to the course instructor for the suggestions to conduct insightful experiments.

References

[1] J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.

[3] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

[4] F. Ma and S. Karaman, "Sparse-to-dense: Depth prediction from sparse depth samples and a single image," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2018.

[5] Y. Song, J. Sohl-Dickstein, D. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *International Conference on Learning Representations (ICLR)*, 2021.

[6] B. Kawar, M. Elad, S. Ermon, and J. Song, "Denoising diffusion restoration models," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

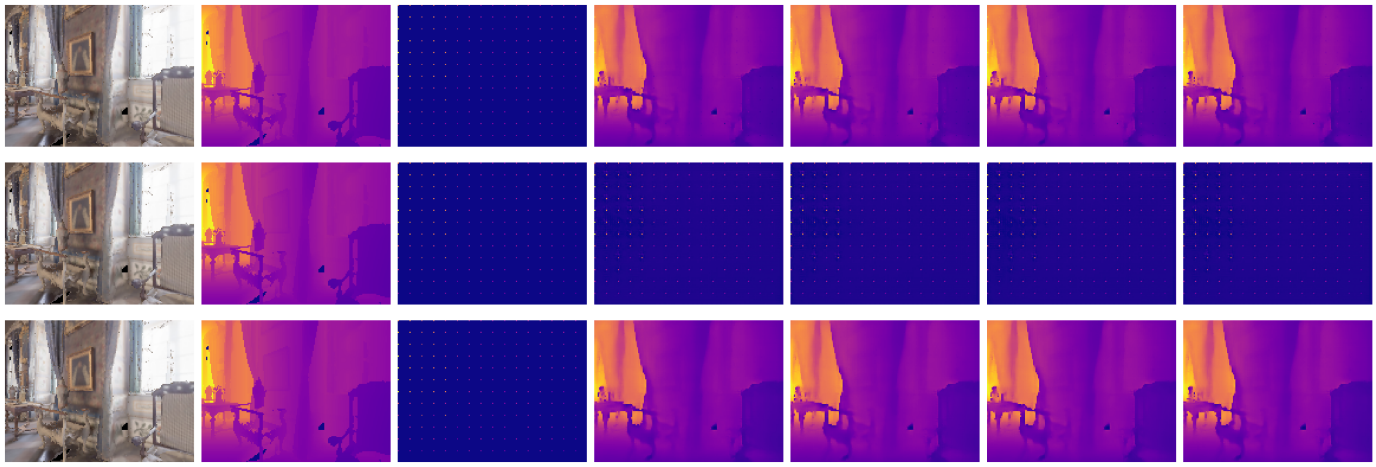

Fig. 4. Ablation study results grid: Left 3-columns are ground truth RGB, depth map, and the sparse LiDAR measurements given to the model. Top row provides the results for both RGB+LiDAR, while the middle row provides LiDAR only, and the bottom row provides RGB images only.

[7] M. Savva, A. Kadian, O. Maksymets *et al.*, "Habitat: A platform for embodied ai research," in *IEEE International Conference on Computer Vision (ICCV)*, 2019.