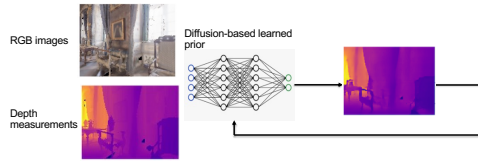


# Sparse Depth Reconstruction Using Diffusion Priors Guided by RGB Observations

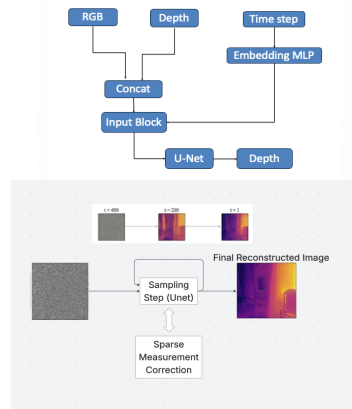
Aviad Golan Peretz  
Stanford University

## Motivation

- Robots and autonomous systems require dense 3D maps to understand and interact with their environments
- LiDAR sensors provide accurate depth, but sparse and expensive. RGB measurements alone are ambiguous and ill-posed
- Diffusion models can act as learned geometric priors



## Diffusion + U-Net Learned Prior Architecture



```

Algorithm 1 Training Diffusion Depth Prior
1 Initialize diffusion model  $f_{\theta}$ 
2 Define noise schedule  $\{\beta_t\}_{t=1}^T$ 
3 Compute  $\alpha_t = 1 - \beta_t$ 
4 Compute  $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$ 
5 while training do
6 Sample RGB image  $I$  and ground truth depth  $D_0$ 
7 Sample timestep  $t \sim \text{Uniform}(1, T)$ 
8 Sample noise  $\epsilon \sim \mathcal{N}(0, I)$ 
9 Generate noisy depth
10 Predict clean depth  $\hat{D}_t = \sqrt{\alpha_t} D_0 + \sqrt{1 - \alpha_t} \epsilon$ 
11 Compute reconstruction loss  $\mathcal{L} = \|D_0 - \hat{D}_t\|_1 + \frac{\lambda}{2} \|D_0 - \hat{D}_t\|_2^2$ 
12 Update  $\theta$  using gradient descent
13 end while

Algorithm 2 Diffusion-Based Sparse-to-Dense Depth Reconstruction
1 Have RGB image  $I$ 
2 Input sparse depth measurements  $D_{\text{obs}}$ 
3 Compute measurement mask  $M = \{D_{\text{obs}} \neq 0\}$ 
4 Initialize depth sample  $D_t = \mathcal{N}(0, I)$ 
5 for  $i = T$  down to 1 do
6 Predict clean depth  $\hat{D}_i = f_{\theta}(I, D_t)$ 
7 Compute color estimate  $\hat{c}_i = \frac{\hat{D}_i - D_t}{\sqrt{1 - \alpha_i}}$ 
8 if  $i > 1$  then
9 Compute DDIM update  $D_{i-1} = \sqrt{\alpha_{i-1}} \left( \frac{\hat{D}_i - \sqrt{1 - \alpha_i} \epsilon}{1 - \alpha_i} + \sqrt{1 - \alpha_{i-1}} \epsilon \right)$ 
10 else
11 Sample noise  $\epsilon \sim \mathcal{N}(0, I)$ 
12 Update  $D_{i-1} = \sqrt{\alpha_{i-1}} D_i + \sqrt{1 - \alpha_{i-1}} \epsilon$ 
13 end if
14 end for
15 Output reconstructed dense-depth  $D_0$ 
    
```

## Related Work

- Monocular depth estimation networks, struggle with depth ambiguity [1]
- Sparse-to-dense depth completion methods fuse LiDAR with RGB images using CNNs, may over-smooth geometry and fail to model uncertainty [2]
- Diffusion models have recently demonstrated strong performance used as generative priors for inverse problems, such as image reconstruction, image inpainting [3, 4]

## References

- [1] D. Eigen, C. Puhrsch, and R. Fergus. Depth Map Prediction from a Single Image using a Multi-Scale Deep Network. *Advances in Neural Information Processing Systems (NIPS)*, 2014.
- [2] F. Ma and S. Karaman. Sparse-to-Dense: Depth Prediction from Sparse Depth Samples and a Single Image. *IEEE International Conference on Robotics and Automation (ICRA)*, 2018.
- [3] Y. Song, J. Sothi-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-Based Generative Modeling through Stochastic Differential Equations. *International Conference on Learning Representations (ICLR)*, 2021.
- [4] B. Kawar, M. Elad, S. Ermon, and J. Song. Denoising Diffusion Restoration Models. *Advances in Neural Information Processing System*, 2022.

## Experimental Results

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

- The diffusion-based architecture outperforms the base U-Net with Sharper edge preservation as can be seen visually, and numerically by the SSIM metric
- Overall, quantitatively higher signal quality is shown by the diffusion-based approach (PSNR), as well as lower reconstruction error (MAE / RMSE)

