BlindSight: Seeing Around Corners with Diffusion Models

Amirmojtaba Sabour
University of Toronto
amsabour@cs.toronto.edu

Abstract

Non-Line-of-Sight (NLOS) imaging, which is the ability to see around corners by analyzing light reflected off walls, has been limited by poor reconstruction quality or the need for specialized hardware and complex active imaging setups. In this paper, we present BlindSight, a new approach that incorporates the generative capabilities of diffusion models to reconstruct hidden scenes from passive indirect reflections. Our method employs a two-stage architecture: first, a reconstruction network maps the wall projections to a learned latent space aligned with Stable Diffusion's VAE, which is decoded into a rough estimation of the target image. Afterwards, an enhancer network based on ControlNet reconstructs a detailed scene through controlled hallucination. Unlike previous approaches that struggle with the inherently ill-posed nature of passive NLOS reconstruction, BlindSight uses the rich prior knowledge of pretrained diffusion models to generate plausible and detailed reconstructions. We evaluate BlindSight on the NLOS-passive dataset and demonstrate significant improvements in both quantitative metrics and qualitative results compared to existing methods. The results suggest that large pretrained generative models can serve as powerful priors for solving challenging inverse problems in computational imaging.

1 Introduction

Non-line-of-sight (NLOS) imaging aims to reconstruct scenes obstructed from direct view by analyzing scattered light on a relay wall, as shown in Figure 1. This capability has broad applications in autonomous driving, robotics, and surveillance. Depending on the presence of a controllable light source, NLOS imaging can be classified into active imaging [2, 12, 15] and passive imaging [28, 20, 21, 32, 1, 22].

Active NLOS imaging relies on illuminating the scene with controlled light sources, such as ultrafast lasers, and measuring the time-of-flight or intensity of light as it reflects off intermediate surfaces. While these methods achieve impressive reconstructions, they often require specialized equipment and complex setups, limiting their practicality. In contrast, passive NLOS imaging eliminates the need for active light sources and instead leverages indirect light captured by an ordinary camera, framing the problem as a challenging image restoration task.

Passive NLOS imaging is inherently challenging due to the extreme blur and information loss in the captured projection image, which makes reconstructing the original scene more difficult. Most existing methods produce highly blurry reconstructions. To address this, we explore the use of additional image and content priors to aid reconstruction and generate high-quality images.

Diffusion models [9] have recently achieved remarkable success in image generation [18], image restoration [14], and conditional generation [29, 27] tasks. Building on this success, we propose BlindSight, a new two-stage passive NLOS imaging method that leverages pretrained image diffusion models as priors to enhance reconstruction quality. In the first stage, we follow prior work and

Preprint. Under review.

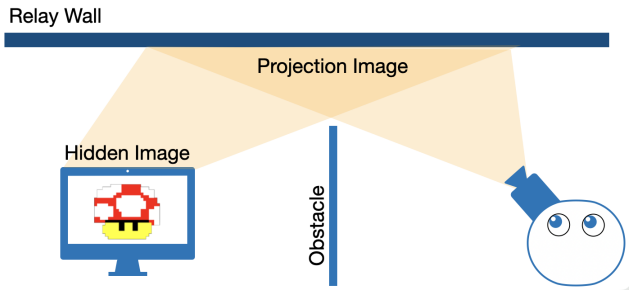

Figure 1: Passive NLOS imaging.

generate a rough and blurry estimate of the hidden image using a reconstruction network. In the second stage, a ControlNet [29] reconstructs the final image conditioned on the rough estimate and an optional content prior in the form of a text prompt describing the hidden image.

We train and evaluate BlindSight on the STL10 [5] subset of the NLOS-Passive [7] dataset and show significant quantitative and qualitative improvements over prior works. We also perform several ablations on the various design decisions made along the way, and qualitatively showcase the impact of the different additional priors used.

2 Related Work

Passive NLOS Imaging Our work focuses on the 2D reconstruction problem in passive NLOS imaging. Existing methods mainly include placing partial occluders [28, 20], using polarizers [21], and applying deep learning methods [32, 1, 22]. Among them, deep learning-based passive NLOS imaging is attractive because the superior representation ability of deep neural networks can greatly improve the reconstruction quality. Most notably, Tancik et al. [22] used a variational autoencoder (VAE) for NLOS imaging, however the model is limited to reconstructing a single specific object.

Image restoration Image restoration (IR) addresses the challenge of improving degraded images through tasks like super-resolution, deblurring, and denoising. Traditional methods relied on handcrafted spatial or frequency-based algorithms [11, 17, 3, 6], later overtaken by deep learning approaches using CNNs [13] and Transformers [23]. While these methods have shown strong performance on standard benchmarks, their reconstructions often lack realistic textures. Generative adversarial networks (GANs) introduced adversarial loss to enhance texture realism but are prone to optimization instability and can generate counterfactual artifacts. Diffusion models have recently outperformed GANs in IR by leveraging iterative denoising processes, producing high-fidelity results with fewer artifacts. Supervised approaches, such as SR3 [19] and DeblurDPM [25], train diffusion models from scratch using paired datasets, achieving state-of-the-art performance but requiring extensive labeled data. In contrast, zero-shot methods like ILVR [4] and DDNM [24] utilize pre-trained diffusion models to extract priors, enabling training-free restoration. These approaches are effective in data-scarce scenarios but face challenges in ensuring consistency between degraded and restored images. Our work builds on diffusion-based IR, and using them as a prior to incorporate details into the blurry reconstructions from prior work on non-line-of-sight (NLOS) imaging methods.

3 Method

In this project, we explore how incorporating image and content priors can enhance NLOS imaging reconstruction. Our proposed method reconstructs hidden scenes using a two-stage pipeline, as shown in Figure 2. The first stage, inspired by [7], employs a reconstruction network that processes the projection image and converts it into a rough estimate of the hidden image using a variational autoencoder (VAE). In the second stage, an enhancer network leverages a ControlNet [29] built on a pretrained text-to-image diffusion model [18] to recover lost details using both learned image priors and optional semantic guidance through text prompts.

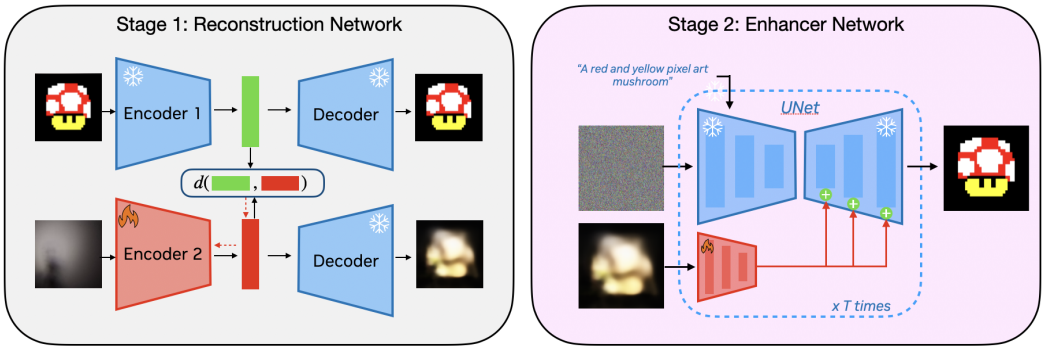

Figure 2: An overview of our two-stage reconstruction pipeline.

More formally, given a projection image I_{proj}, our goal is to reconstruct the hidden image I_{hidden}. In stage 1, the reconstruction network produces a coarse and blurry reconstruction I_{coarse}. In stage 2, a ControlNet built on Stable Diffusion 1.5 generates the final result conditioned on an optional content prior in the form of a text prompt p, expressed as $I_{final} = \Phi_{pretrained}(I_{coarse}, p)$.

3.1 Coarse Reconstruction Network

The reconstruction network's primary role is generating an initial rough estimate of the hidden image. Following prior work [7], this is achieved by aligning projection images I_{proj} and hidden images I_{hidden} in the latent space of a pretrained autoencoder. We specifically use the VAE from Stable Diffusion 1.5 as our pretrained model. This VAE employs an encoder that compresses images of size $512 \times 512 \times 3$ into a compact latent representation of size $64 \times 64 \times 4$, which can then be decoded back into the original image space. For our reconstruction network, we initialize a new encoder E_{proj} using the pretrained weights of the original encoder $E_{pretrained}$ and train it to align the latent representations of projection images with those of the hidden images. This alignment can be formally expressed as:

$$E^*_{proj} = \arg\min |E_{proj}\left[I_{proj}\right] - E_{pretrained}\left[I_{hidden}\right]|^2_2 \tag{1}$$

Once training is complete, we can obtain a rough estimation by passing the aligned latents through the pretrained decoder:

$$I_{coarse} = D_{pretrained}(E_{proj}(I_{proj})) \tag{2}$$

However, due to the substantial information loss inherent in the projection images, the resulting output is typically very blurry and lacks fine details. To fix this, we use an enhancer network that recovers these details by leveraging the rich knowledge and priors in pretrained diffusion models.

3.2 Enhancer Network

The enhancer network's role is to leverage a pretrained diffusion model to hallucinate missing details that were lost during the NLOS projection process. We investigate two complementary approaches: using the pretrained diffusion model's inherent image prior alone, and augmenting it with an additional content prior in the form of a text prompt describing the hidden image. When this content prior is provided, it significantly constrains the solution space of possible images that could have produced the observed projection, leading to more accurate and effective reconstructions. For a detailed analysis of how different priors affect reconstruction quality, please see Section 4.

To condition the pretrained diffusion model on the coarse reconstruction I_{coarse}, we train a ControlNet [29] on top of Stable Diffusion 1.5. During training, we randomly drop both the text prompts and the conditioning signal I_{coarse} with probabilities of 10% and 50% respectively, enabling classifier-free guidance [10] during inference. This technique is vital for ensuring the final outputs maintain consistency with both the provided content prior and the structural information present in the coarse reconstruction.

Table 1: Quantitative evaluation and ablation study results. Higher PSNR (↑) and lower LPIPS/FID (↓) are better. Our full method achieves the best perceptual metrics while ablations demonstrate the importance of each component.

Method	PSNR ↑	LPIPS ↓	FID ↓
Reconstruction network	**17.1**	0.649	194.799
Ours	13.6	**0.581**	**20.147**
Ours - No text	13.0	0.661	114.420
Ours - No projection	9.4	0.708	24.183

4 Experiments

4.1 Experiment setting

Data The NLOS-Passive dataset [7] is used for training and evaluation. This dataset contains four groups with different image sources: STL10, MNIST digits, Anime faces, and Supermodel faces, each paired with their corresponding projections. We focus on the STL10 group as it presents the most diverse and challenging scenarios. We use 100K image-projection pairs for training and 5K pairs for quantitative evaluation. Text descriptions for all images are generated using the Florence-2 multimodal vision-language model [26]. All hidden and projection images are resized to 512×512 resolution for both training and inference.

Metrics We evaluate reconstruction quality using both standard metrics and distribution-based measures. For direct image comparison, we use PSNR and LPIPS [30] against ground truth images. However, given the ill-posed nature of NLOS imaging and the substantial information loss in the projection process, these reconstruction metrics alone are insufficient. We therefore also employ Frechet Inception Distance (FID)[8] on the 5K evaluation set. For inference, we use the UniPC sampler[31] with 20 steps and apply classifier-free guidance to both text and coarse reconstruction signals (I_{coarse}) with a guidance scale of 5.

Implementation Details The reconstruction network is trained for 50K iterations using an L2 loss between latent representations. We use the AdamW optimizer [16] with a learning rate of 1e-4 and batch size of 32. For the enhancer network, we train a ControlNet on top of Stable Diffusion 1.5, applying dropout to the conditioning signals with probabilities of 10% for text prompts and 50% for coarse reconstructions to enable classifier-free guidance during inference. The training is performed for 20K steps with a batch size of 32. All training is performed on 8 NVIDIA A100 GPUs.

4.2 Results and Ablations

We evaluate BlindSight through comprehensive experiments and ablation studies to understand the contribution of each component. Table 1 presents the quantitative results of our method compared to baselines and ablated variants.

Reconstruction Quality The baseline reconstruction network achieves the highest PSNR (17.1), primarily because it learns to predict the mean of the possible image distribution for a given projection. While this minimizes MSE loss, it results in blurry reconstructions lacking fine details, reflected in poor LPIPS (0.649) and FID (194.799) scores. In contrast, our full method produces more detailed and realistic reconstructions, demonstrated by significant improvements in perceptual metrics (LPIPS: 0.581, FID: 20.147), despite a lower PSNR (13.6) due to the inherent variation in generated details.

Impact of Priors To understand the impact of different priors, we conduct two ablation studies. First, we remove the content prior by using empty text prompts during generation. This results in significant degradation across all metrics (PSNR: 13.0, LPIPS: 0.661, FID: 114.420), highlighting the importance of semantic guidance in producing coherent reconstructions.

To verify that our method truly adheres to the projection information rather than merely using the pretrained diffusion model, we test generation without the coarse conditioning signal. This ablation

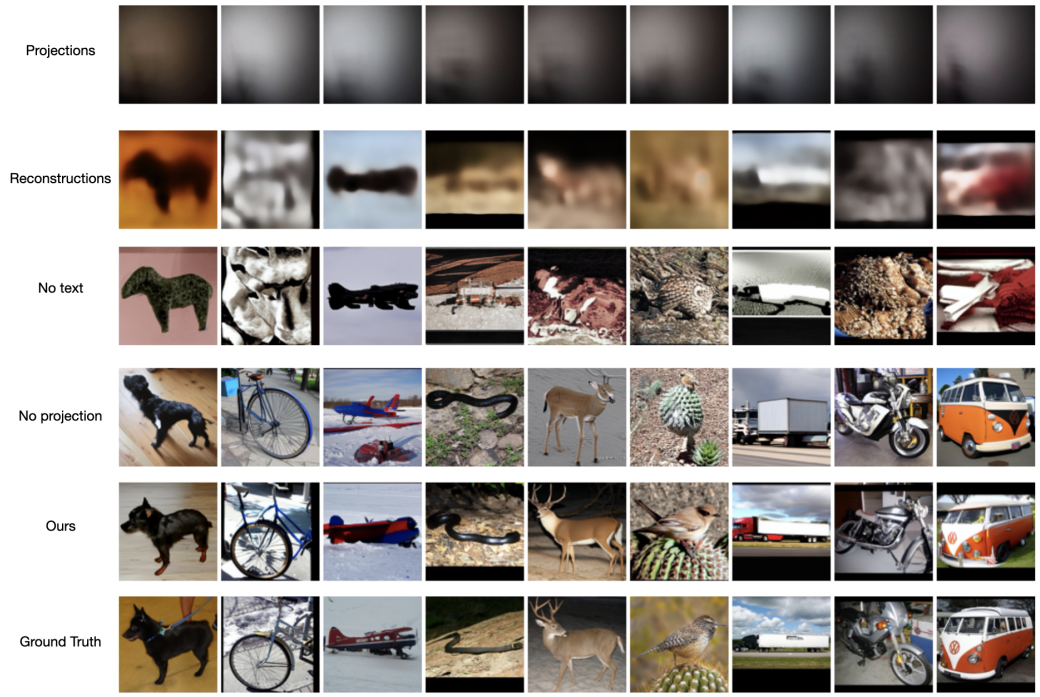

Figure 3: Qualitative comparison of NLOS reconstruction methods. Top to bottom: input projections, baseline reconstructions, ablations (no text, no projection), our full method, and ground truth. Our approach generates more detailed and realistic reconstructions while preserving scene content.

shows the most severe performance drop in reconstruction metrics (PSNR: 9.4, LPIPS: 0.708) and a slight deterioration in distribution matching (FID: 24.183). These results confirm that while the diffusion model's prior contributes significantly to the visual quality, the conditioning signal from the reconstruction network is crucial for maintaining a semblance of consistency to the original image. The quantitative results are supported by qualitative comparisons shown in Figure 3, where our method consistently produces more detailed and realistic reconstructions while preserving the essential structure of the hidden scene.

5 Conclusion

In this paper, we presented BlindSight, a new approach to passive non-line-of-sight imaging that leverages the generative capabilities of diffusion models. Our two-stage architecture, combining a reconstruction network with a diffusion-based enhancer network, demonstrates that pretrained generative models can serve as powerful priors for solving complex inverse problems in computational imaging. Rather than fighting the inherently ill-posed nature of passive NLOS reconstruction, our method embraces controlled hallucination, using the rich prior knowledge encoded in pretrained diffusion models to generate plausible and detailed reconstructions. The experimental results show that BlindSight significantly outperforms existing methods in both quantitative metrics and qualitative results while requiring only standard RGB cameras, suggesting that similar approaches could be valuable for other ill-posed inverse problems.

References

[1] Miika Aittala, Prafull Sharma, Lukas Murmann, Adam Yedidia, Gregory Wornell, Bill Freeman, and Fredo Durand. Computational mirrors: Blind inverse light transport by deep matrix factorization. *Advances in neural information processing systems*, 32, 2019.

[2] Mauro Buttafava, Jessica Zeman, Alberto Tosi, Kevin Eliceiri, and Andreas Velten. Non-line-of-sight imaging using a time-gated single photon avalanche diode. *Optics express*, 23(16):20997–

21011, 2015.

[3] Giannis Chantas, Nikolaos P Galatsanos, Rafael Molina, and Aggelos K Katsaggelos. Variational bayesian image restoration with a product of spatially weighted total variation image priors. *IEEE transactions on image processing*, 19(2):351–362, 2009.

[4] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*, 2021.

[5] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011.

[6] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on image processing*, 16(8):2080–2095, 2007.

[7] Ruixu Geng, Yang Hu, Zhi Lu, Cong Yu, Houqiang Li, Hengyu Zhang, and Yan Chen. Passive non-line-of-sight imaging using optimal transport. *IEEE Transactions on Image Processing*, 31:110–124, 2021.

[8] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[10] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[11] Kwang In Kim and Younghee Kwon. Single-image super-resolution using sparse regression and natural image prior. *IEEE transactions on pattern analysis and machine intelligence*, 32(6):1127–1133, 2010.

[12] Ahmed Kirmani, Tyler Hutchison, James Davis, and Ramesh Raskar. Looking around the corner using ultrafast transient imaging. *International journal of computer vision*, 95:13–28, 2011.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[14] Xin Li, Yulin Ren, Xin Jin, Cuiling Lan, Xingrui Wang, Wenjun Zeng, Xinchao Wang, and Zhibo Chen. Diffusion models for image restoration and enhancement–a comprehensive survey. *arXiv preprint arXiv:2308.09388*, 2023.

[15] Xiaochun Liu, Ibón Guillén, Marco La Manna, Ji Hyun Nam, Syed Azer Reza, Toan Huu Le, Adrian Jarabo, Diego Gutierrez, and Andreas Velten. Non-line-of-sight imaging using phasor-field virtual wave optics. *Nature*, 572(7771):620–623, 2019.

[16] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[17] Julien Mairal, Guillermo Sapiro, and Michael Elad. Learning multiscale sparse representations for image and video restoration. *Multiscale Modeling & Simulation*, 7(1):214–241, 2008.

[18] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[19] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4713–4726, 2022.

[20] Charles Saunders, John Murray-Bruce, and Vivek K Goyal. Computational periscopy with an ordinary digital camera. *Nature*, 565(7740):472–475, 2019.

[21] Kenichiro Tanaka, Yasuhiro Mukaigawa, and Achuta Kadambi. Polarized non-line-of-sight imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2136–2145, 2020.

[22] Matthew Tancik, Guy Satat, and Ramesh Raskar. Flash photography for data-driven hidden scene recovery. *arXiv preprint arXiv:1810.11710*, 2018.

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[24] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. *arXiv preprint arXiv:2212.00490*, 2022.

[25] Jay Whang, Mauricio Delbracio, Hossein Talebi, Chitwan Saharia, Alexandros G Dimakis, and Peyman Milanfar. Deblurring via stochastic refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16293–16303, 2022.

[26] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. arxiv (nov. 2023), 2023.

[27] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

[28] Adam B Yedidia, Manel Baradad, Christos Thrampoulidis, William T Freeman, and Gregory W Wornell. Using unknown occluders to recover hidden scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12231–12239, 2019.

[29] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[30] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[31] Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

[32] Chen Zhou, Cheng-Yu Wang, and Zhiwen Liu. Non-line-of-sight imaging off a phong surface through deep learning. *arXiv preprint arXiv:2005.00007*, 2020.