A PHYSICAL MEASUREMENT BENCHMARK FOR PASSIVE NON-LINE-OF-SIGHT IMAGING UNDER INDOOR ILLUMINATION AND ENVIRONMENTAL REFLECTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recovering hidden images from indirect light is a challenging problem in physical imaging. In illuminated indoor environments, target-dependent signals are mixed with ambient illumination and reflections from walls, furniture, and other surfaces, making it difficult to determine whether generative models preserve target-specific information or merely produce visually plausible outputs. We introduce a physical benchmark for passive non-line-of-sight image reconstruction under indoor illumination and environmental reflections. The benchmark contains 13,000 physically acquired measurement–target pairs spanning ten STL-10 categories, together with background references, fixed split manifests, and provenance information. The target content consists of 2D STL-10 images projected onto a hidden screen, while the physical contribution lies in the paired measurement acquisition and its evaluation protocol. We support deterministic measurement-to-image reconstruction and coarse-conditioned latent-diffusion refinement, and separately evaluate measurement-only inputs and additional class-conditioned settings. The benchmark provides unified pixel, structural, perceptual, and semantic metrics, together with matched evaluations of deterministic reconstruction, a pretrained latent diffusion model, and a full-UNet fine-tuned variant. Across the matched diagnostic comparisons, domain fine-tuning produces a pronounced perceptual shift at high denoising strength, substantially reducing LPIPS without improving pixel or structural metrics. These results reveal a consistent gap between perceptual plausibility and target-specific reconstruction fidelity rather than a uniform improvement across metrics.

1 INTRODUCTION

Non-line-of-sight (NLOS) imaging is a inverse problem of reconstructing targets outside the camera's view by utilizing indirect optical detection of hidden objects or scenes. This technology can extend the perception range of machines to occluded areas, thus having applications in rescue operations((Velten et al., 2012; Wu et al., 2021)), autonomous((Scheiner et al., 2020; Zhao et al., 2021)), and obstacle avoidance for robots. Unlike active NLOS imaging, which uses a controllable laser light source((Klein et al., 2018; Galindo et al., 2019; Grau Chopite et al., 2020; Chen et al., 2020)), passive NLOS imaging only uses traditional cameras to capture light intensity data to reconstruct images or scenes((Batarseh et al., 2018; Bouman et al., 2017)). The inherent characteristics of passive NLOS—specifically high noise levels and weak light intensity signals—pose significant challenges for target reconstruction.

To accurately recover hidden targets from passive NLOS scenes, a series of physically based methods have been developed ((Bouman et al., 2017; Tanaka et al., 2020; Czajkowski & Murray-Bruce, 2024)). These methods first establish light transport and occlusion models to provide additional spatial constraints, and then estimate the hidden scene from the walls in a regularized inverse problem manner. In recent years, researchers have trained networks by matching wall information and hidden target images to directly learn complex inverse mappings (Geng et al., 2021; Ronneberger et al., 2015; Zhang et al., 2024).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

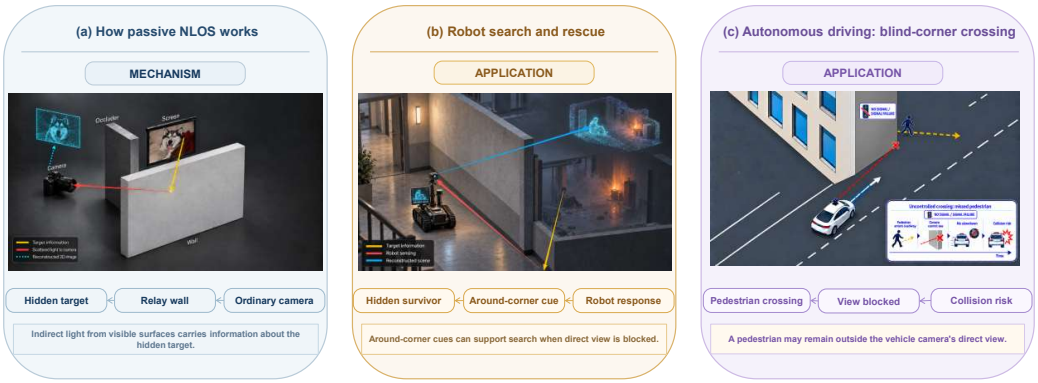


Figure 1: Principle and representative applications of passive non-line-of-sight (NLOS) imaging. (a) Light from a hidden target is reflected by a visible relay wall and captured by an ordinary camera. (b) The resulting around-corner cue supports robot search and rescue. (c) The same capability can reveal pedestrians in blind corners for autonomous driving when the direct view is blocked.

Despite the significant attention given to network-related algorithms in recent years, problems persist. The most prominent issue is that many existing passive NLOS research results are based on relatively idealized assumptions such as specific relay surfaces, lighting conditions, and geometric layouts (Faccio et al., 2020; Bouman et al., 2017; Saunders et al., 2019; Seidel et al., 2021). However, real-world scenarios involve ambient light, multiple reflective objects, occlusions, and uncontrollable background interference, all of which further weaken the already weak target-related signals(Wang et al., 2021; Boger-Lombard & Katz, 2019). Therefore, we can not directly apply existing passive NLOS schemes to common real-world scenarios. Furthermore, the lack of relevant datasets fundamentally hinders the development and evaluation of technologies in fields such as autonomous driving, robot obstacle avoidance, and fire rescue.

To address this gap, we introduce a benchmark with two complementary reconstruction tracks. In the measurement-only track, a deterministic model maps passive measurements to a hidden-target image. In the known-class-conditioned refinement track, a latent diffusion model refines a measurement-derived coarse reconstruction using an externally supplied class label. The label provides only coarse semantic information and does not reveal the target image, instance identity, pose, or location. Both tracks are evaluated on the same measurement–target pairs using pixel-level, structural, perceptual, and distributional metrics. This protocol allows us to test whether generative refinement improves perceptual naturalness while preserving correspondence to the paired target, and how this trade-off changes with denoising strength, guidance, and diffusion-step count.

Our contributions are threefold. First, we introduce a controlled physical benchmark for passive NLOS reconstruction under indoor reflected-light conditions, comprising 13,000 paired measurement–target records with hidden-screen target images spanning ten classes. Second, we define an auditable evaluation protocol with a fixed split of 12,000 training pairs and 1,000 diagnostic pairs, standardized preprocessing, checkpoint logging, reproducibility checks, and complementary target-correspondence, perceptual, and distributional metrics. Third, we conduct a controlled comparison between deterministic reconstruction and latent-diffusion refinement. The results show that pixel fidelity, structural consistency, perceptual naturalness, and distributional similarity need not improve together: deterministic models generally retain stronger paired fidelity, whereas diffusion refinement can improve perceptual and distributional scores without guaranteeing closer correspondence to a particular target. We additionally provide NLOSNet-T as a U-Net-derived coarse initializer for the refinement track and treat DPO(Rafailov et al., 2023; Wallace et al., 2024) and GRPO(Shao et al., 2024; Xue et al., 2025) as auxiliary exploratory analyses.

2 Related Work

Existing NLOS benchmarks cover diverse hidden-scene perception tasks. Klein et al. (2018) introduced a quantitative platform for comparing NLOS reconstruction methods, while Galindo et al.

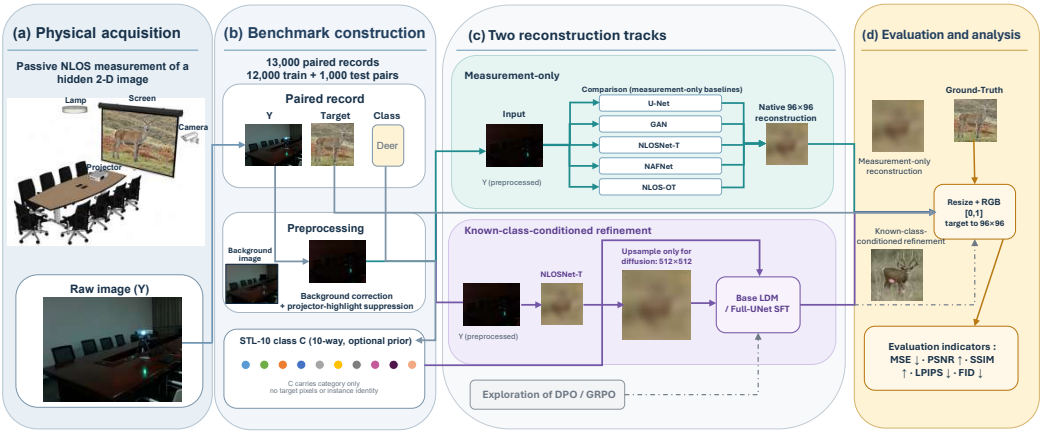


Figure 2: Overview of the passive NLOS reconstruction benchmark and processing pipeline. (a) The physical setup acquires raw observations of hidden 2-D targets. (b) Observations, target images, and category information are organized into paired records and preprocessed. (c) Reconstruction is evaluated through measurement-only and known-class-conditioned tracks. (d) The outputs are compared using paired image metrics and the set-level FID score.

(2019) released a time-resolved dataset for hidden-scene imaging under multi-bounce light transport. Other benchmarks target around-corner road-user detection and hidden-person tracking (Scheiner et al., 2020; Wang et al., 2023), whereas NLOS-Passive(Geng et al., 2021) focus on passive paired reconstruction and event-camera imaging of moving targets, respectively. Because these resources differ in sensing modality, target representation, scene scale, and acquisition protocol, they characterize complementary aspects of NLOS perception rather than a common cross-dataset ranking problem. Our benchmark instead focuses on paired two-dimensional RGB reconstruction from physical passive measurements in a controlled indoor reflected-light setting, with a unified protocol for evaluating target correspondence and perceptual quality.

Existing reconstruction methods can be broadly grouped into physics-based inversion, deterministic learning-based reconstruction, and generative refinement. Physics-based approaches model light transport, occluder geometry, optical coherence, or polarization, but rely on calibrated parameters and simplifying assumptions (Katz et al., 2014; Batarseh et al., 2018; Tanaka et al., 2020). Deterministic learning-based methods learn measurement-to-image mappings and reduce the need for explicit forward modeling, including deep reconstruction networks, learned feature representations, and optimal-transport formulations Grau Chopite et al. (2020); Chen et al. (2020); Geng et al. (2021). More recent latent- and diffusion-based methods introduce stronger generative priors and can improve perceptual realism (Zhang et al., 2024); however, additional semantic or generative information does not necessarily preserve correspondence with a particular target instance. We therefore evaluate measurement-only reconstruction and known-class-conditioned refinement as separate tracks on the same paired measurements, enabling a controlled analysis of the fidelity–perception trade-off.

3 Benchmark and Task Definition

In this section, we introduce four key component of our benchmark. First, we describe the indoor sensing scenario and formulate the passive NLOS reconstruction task, clarifying the connection between our experimental scenario and the practical hidden-target perception. Next, we introduce the specific experimental setup and scalability of Our bench, including scenario configuration, data collection, ambient light noise, and the impact of object reflections. Then, we introduce the baseline we used. Finally, we describe the evaluation metrics used in the experiments.

3.1 APPLICATION SCENARIO AND TASK FORMULATION

We construct our benchmark around an indoor around-corner perception scenario. When a robot operates near a corner or behind an occlusion, the target region may lie outside the camera's direct field of view. The camera can then observe only indirect RGB measurements formed by reflections from visible relay surfaces, including walls, furniture, and other objects. These measurements contain partial information about the hidden target, but they are mixed with ambient illumination, surface reflectance, background structure, and sensor noise. The practical objective is to infer the appearance of the hidden target from these weak and entangled measurements.

Directly collecting paired data for every real search-and-rescue or robotic deployment scenario would require a large number of scene configurations and carefully controlled physical targets. We therefore instantiate the application scenario as a controlled hidden-screen reconstruction task. Specifically, a two-dimensional RGB image is projected onto a screen outside the camera's direct field of view, while the camera records the corresponding passive measurement in an illuminated and furnished indoor environment. This design preserves the key difficulties of passive NLOS perception, including indirect light transport, occlusion, ambient illumination, and reflections from surrounding objects, while providing an exactly paired target image and a repeatable acquisition protocol.

Let I_h denote the hidden two-dimensional RGB target, and let Y_{raw} denote the raw camera measurement. Let Y_{bg} denote the background reference acquired under the same camera and room configuration. The processed measurement is defined as

$$\widetilde{Y} = P(Y_{\mathrm{raw}}, Y_{\mathrm{bg}}), \tag{1}$$

where P is the fixed background-correction and normalization pipeline. The benchmark defines two reconstruction tracks according to their inference-time information.

For the **measurement-only track**, the model receives only the processed measurement $\widetilde{Y}$:

$$\widehat{I}_{h,\mathrm{measurement}} = R_\theta(\widetilde{Y}), \tag{2}$$

where R_θ denotes a registered measurement-to-image reconstruction baseline.

For the **known-class-conditioned refinement track**, the same measurement is first mapped to a coarse bridge and then refined using an externally supplied class cue:

$$\widehat{I}_{h,\mathrm{conditioned}} = D_\phi\left(\mathcal{U}\left(B_{\theta_0}(\widetilde{Y})\right), C, z\right), \tag{3}$$

where B_{θ_0} is the fixed NLOSNet-T initializer, D_ϕ is the diffusion refinement operator, $\mathcal{U}$ is the bridge-resizing operation, C is the externally supplied category cue, and z is the diffusion noise variable.

The hidden target is an image rather than a free-space three-dimensional reconstruction. In a search or robotic-inspection setting, C may come from mission context, operator input, or an upstream perception module while the target appearance, instance identity, pose, and location remain unknown. The class cue contains no target pixels or instance-specific information. The two tracks therefore answer complementary questions: how much target information is recoverable from passive measurements alone, and whether a coarse semantic cue changes perceptual quality while preserving correspondence to the paired target.

3.2 PHYSICAL SETUP AND DATA ACQUISITION

To instantiate the task, we construct a controlled physical acquisition setup that preserves the key components of the application scenario: a hidden target, indirect light transport through visible room surfaces, ambient illumination, and passive camera measurements. The setup does not attempt to reproduce every real deployment environment. Instead, it provides a repeatable physical proxy in which the hidden target, acquisition parameters, and measurement–target correspondence are explicitly controlled for quantitative evaluation.

3.2.1 INDOOR PHYSICAL CONFIGURATION

We conduct the experiments in an illuminated indoor conference room, as shown in Fig. 2. The room has an ambient illuminance of approximately 300 lx. A 100-inch projection screen is placed at

a projection distance of approximately 3 m and serves as the hidden target surface. The screen is outside the direct field of view of the camera. The camera observes visible relay surfaces, including the wall, furniture, and other reflecting objects, and records the weak RGB signals produced by indirect light transport from the hidden screen. The camera resolution is 640×480, and the camera position, focal length, exposure, and other acquisition parameters are fixed throughout data collection.

The projector is not physically blocked during acquisition. Its light is projected onto the screen and contributes to the illumination scattered into the surrounding scene. Because the camera also captures a high-intensity projector highlight at the projector's image location, we suppress this component by applying a fixed image-space black mask at that location during preprocessing. The mask is applied consistently to all camera measurements, while the target images remain unchanged.

3.2.2 Hidden Target and Paired Data Collection

To obtain paired measurement–target samples, we project two-dimensional RGB images onto the hidden screen and capture the corresponding passive camera observations. We use STL-10(Coates et al., 2011) as the source of hidden target images, covering ten object categories. For each target image $I_{h,i}$, the projected image is treated as the ground-truth hidden target, while the corresponding passive camera observation is recorded as Y_i^{raw}. The category label C_i is retained as optional semantic information and is used only in the known-class-conditioned refinement track defined in Section 3.1.

3.2.3 Measurement Preprocessing and Data Split

The raw camera measurement is first processed using the fixed background and direct-light suppression procedure described above. The processed observation is then resized according to the input contract of each reconstruction model. All deterministic models produce native 96×96 RGB outputs. For diffusion refinement, the coarse reconstruction produced by the fixed initializer is resized to the diffusion working resolution, and the generated image is resized to 96×96 for the common paired evaluation. The fixed model contracts in Appendix A.1.1.

The benchmark contains 13,000 paired records from ten target classes. We use 12,000 pairs for training and reserve 1,000 pairs as a diagnostic-test split. The split is fixed before training, and each measurement–target pair remains indivisible. The benchmark protocol does not define an additional validation split. Checkpoints, training schedules, and diffusion operating points are fixed before diagnostic evaluation, and the split is used only for final comparison.

A separate train-only auxiliary screen contains 500 records, with 50 records per class, and evaluates eight predeclared diffusion configurations. This auxiliary screen supports diffusion sensitivity analysis and operating-point documentation. It does not access the diagnostic-test targets and is not treated as a validation or formal test set. The record schema is provided in Appendix A.1.

3.3 Evaluation Metrics

Because each passive measurement is paired with a hidden target image, the benchmark supports direct instance-level evaluation. We report four paired fidelity metrics—MSE, PSNR, SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018)—on the $N = 1{,}000$ diagnostic-test pairs. Predictions and targets are converted to RGB, resized to 96×96 with bicubic interpolation, normalized to $[0, 1]$, and evaluated using the same pair identifiers. Metrics are computed per pair and averaged over the test set. MSE and PSNR quantify pixel-level error, SSIM measures structural consistency, and LPIPS measures perceptual similarity.

We additionally report Fréchet Inception Distance (FID) (Heusel et al., 2017) as a supplementary distribution-level diagnostic. FID compares reconstructed and target feature distributions over the complete diagnostic-test set without using measurement–target correspondence. It therefore characterizes distributional similarity rather than instance-level fidelity. Exact definitions and implementation details are given in Appendix B.

3.4 Reconstruction Models and Evaluation Tracks

We organize the reconstruction methods according to their role in the benchmark and their inference-time input permissions. The measurement-only track evaluates reconstruction from passive measurements alone, whereas the known-class-conditioned refinement track additionally uses a coarse reconstruction and an externally provided class condition. The detailed quantitative comparisons are presented in Section 4.2.

3.4.1 Measurement-only Track

The measurement-only track evaluates how much information can be recovered from passive NLOS measurements alone. At inference time, each model receives only the preprocessed measurement $\widetilde{Y}$. The paired target image is used for training supervision and final evaluation, but is not provided during inference or test-time optimization.

NLOSNet-T. NLOSNet-T is our proposed deterministic measurement-to-image reconstruction model. It produces the coarse reconstruction used in the measurement-only track and provides the fixed upstream initializer for the known-class-conditioned refinement track. Its detailed architecture and training configuration are provided in Appendix A.1.1. **U-Net** (Ronneberger et al., 2015) is a general encoder–decoder architecture with skip connections. It serves as a standard image-to-image reconstruction baseline and evaluates the ability of a generic convolutional architecture to recover the paired hidden target from passive measurements. **Pix2Pix-derived conditional GAN** We include a Pix2Pix-derived conditional GAN with a PatchGAN discriminator(Isola et al., 2017). The generator is trained with a binary adversarial objective and an L_1 reconstruction term with weight $\lambda = 100$. It receives the same measurement-only input and is evaluated using the same paired metrics and diagnostic-test split as the other deterministic methods. The evaluation metrics are therefore comparable even though the GAN training objective differs from the MSE objectives used by some reference models. **NAFNet** (Chen et al., 2022) is an image-restoration architecture designed for efficient feed-forward reconstruction. We include it to examine whether a restoration-oriented backbone can improve reconstruction fidelity under the weak and spatially mixed NLOS measurements. **NLOS-OT** (Geng et al., 2021) is a task-specific optimal-transport reconstruction method designed for NLOS image recovery. It provides a comparison with a reconstruction formulation that incorporates task-specific transport structure in addition to standard image-to-image learning.

3.4.2 Known-class-conditioned Refinement Track

Base LDM and Full-UNet SFT receive the 512×512 coarse bridge and a category prompt. Base LDM denotes the frozen Stable Diffusion v1.5 image-to-image prior(Rombach et al., 2022). It receives the bridge produced by NLOSNet-T and an externally supplied category prompt; Full-UNet SFT starts from the same SD 1.5 initialization and fine-tunes the U-Net under the training contract in the appendix's 'Models and fixed training contracts' section. DPO and GRPO are reported only as exploratory adaptations of the SFT branch.

The class condition provides only coarse category information. It does not contain target pixels, instance identity, pose, spatial location, or target-specific appearance. The refinement track includes a frozen Base latent diffusion model and a Full-UNet supervised fine-tuning variant. Both models use the same fixed coarse initializer, category encoding, preprocessing contract, data split, and evaluation protocol.

Full-UNet supervised fine-tuning. The Full-UNet SFT model adapts the diffusion denoiser using clean target images and fixed STL-10 class prompts. Passive measurements and coarse reconstructions are not used as training inputs for SFT; the measurement-derived coarse image is introduced only during inference. Its training and inference configuration is provided in Appendix A.1.1.

3.4.3 Exploratory Preference Optimization

DPO and GRPO are treated as exploratory preference-optimization extensions rather than primary reconstruction baselines. Both methods inherit the fixed NLOSNet-T initializer and the Full-UNet SFT lineage. Their adaptation uses a train-only auxiliary subset of 500 records, while final metrics are reported on the diagnostic-test pairs. These methods are reported separately in Section 4.4 and are excluded from the primary reconstruction ranking.

Because the refinement track receives additional semantic information and uses a different reconstruction objective, comparisons between the measurement-only and refinement tracks are not interpreted as direct rankings of model architectures.

4 Experiments and Analysis

4.1 Experiment setup

In this section, we evaluate the reconstruction methods in the two benchmark tracks defined in Section 3.4: measurement-only reconstruction with deterministic networks, and known-class-conditioned diffusion refinement with the Base LDM and Full-UNet SFT models.

All methods are evaluated on the same diagnostic-test pairs using the paired fidelity metrics and the distribution-level diagnostic introduced in Sec. 3.3. The main comparison reports MSE, PSNR, SSIM, LPIPS, and FID, while Detailed preprocessing, training, and sampling configurations are provided in Appendix A.1.1.

We compare the representative deterministic and refinement methods under their respective input conditions. Then we isolate the effect of the class condition, examine the sensitivity of diffusion refinement to its operating parameters, and finally report DPO and GRPO as exploratory preference-optimization extensions.

4.2 Quantitative Results under Different Information Conditions

We first introduce benchmark results using configurations fixed before the final test evaluation. Model architectures, training settings, and diffusion sampling parameters are pre-defined or determined using training data only, and are then kept fixed for the test evaluation. Target images from the test split are not used for model selection or test-time optimization.

Within the measurement-only track, each deterministic model maps the same preprocessed passive measurement to a reconstruction, which is compared with its paired hidden target. Table 1 reports the results of NLOSNet-T, U-Net, NAFNet, GAN, and NLOS-OT using pixel-level, structural, perceptual, and distribution-level metrics.

The refinement track takes the fixed NLOSNet-T output as its starting point and supplies each method with the same externally provided class condition. Since this track uses additional inputs and a different reconstruction procedure, its results are analyzed separately rather than combined with the measurement-only methods in a single ranking.

Table 1: Main quantitative comparison on the test set. The best and second-best results are shown in bold and underlined, respectively.

Track	Method	MSE ↓	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
Measurement only	U-Net	0.0302	15.51	0.259	0.716	326.02
	GAN	0.0407	14.23	0.160	0.558	200.80
	NAFNet	0.0166	18.08	0.363	0.680	238.59
	NLOS-OT	0.0193	17.47	0.347	0.615	293.61
	NLOSNet-T	**0.0164**	**18.15**	**0.383**	0.801	283.78
Known-class conditioned	Base LDM	0.0223	16.80	0.335	0.566	85.49
	Full-UNet SFT	0.0239	16.45	0.294	**0.477**	**53.05**

4.3 Effect of Known-Class Conditioning

Because the refinement track also introduces a generative reconstruction prior, the main experience in section 4.2 does not by itself isolate the contribution of the semantic condition, We therefore conduct a controlled comparison in which the coarse initializer, refinement model, sampling procedure, and target pairs are kept fixed, while only the class condition is changed.

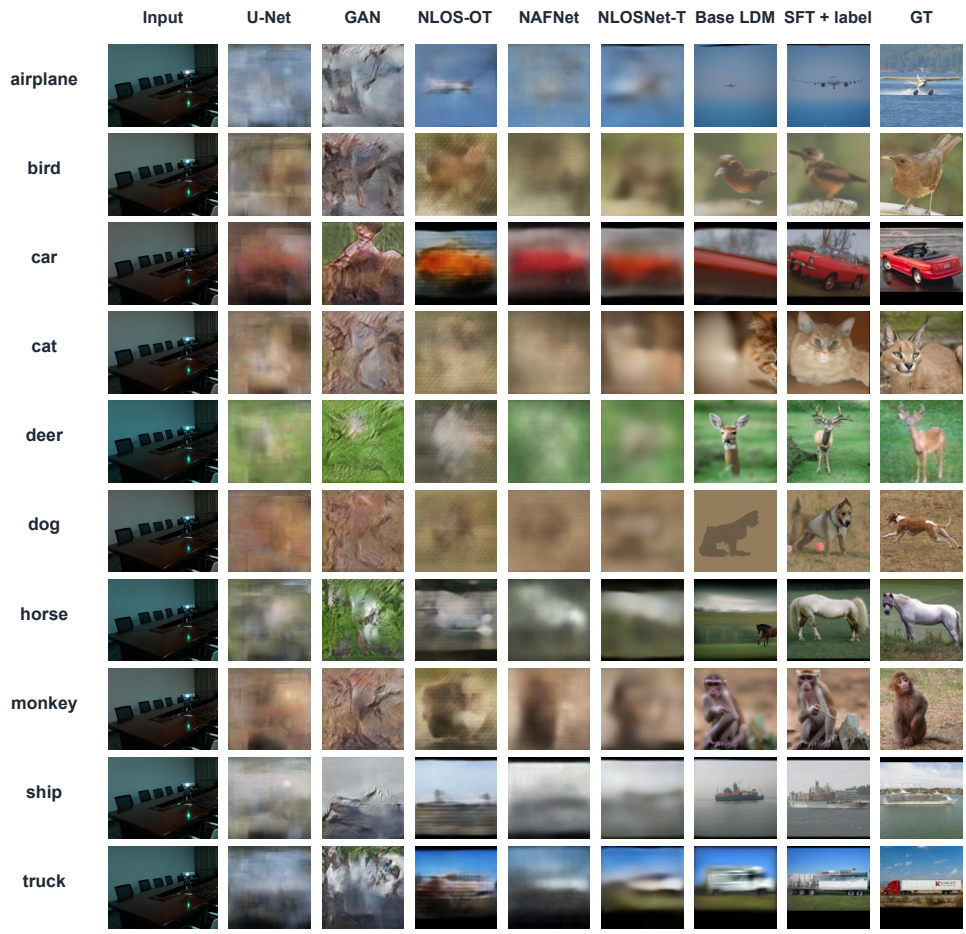


Figure 3: Qualitative comparison of reconstructed images across ten target classes. Each row corresponds to one class, and the columns show the input, U-Net, GAN, NLOS-OT, NAFNet, NLOSNet-T, Base LDM, SFT with the class label, and the ground-truth image (GT). The grid provides a visual comparison of structural, detail, and class-semantic recovery across methods.

This comparison evaluates whether coarse category information improves perceptual naturalness without providing target-specific appearance information. We report the changes in paired fidelity, structural similarity, perceptual distance, and distributional discrepancy.

To isolate the role of semantic side information, we analyze two inference conditions while keeping the measurement, coarse initializer, model checkpoint, sampler, sampling parameters, and per-sample random seed fixed. In the known-category setting, the model receives the category associated with the target image. In the null-condition setting, the category input is replaced by a fixed neutral embedding. The category condition contains no target pixels, instance identity, pose, or spatial location.

As shown in Table 2, , the correct category condition reduces LPIPS and FID for both refinement models, with LPIPS reductions of approximately 13% and FID reductions of more than 50%. These improvements are accompanied by decreases in PSNR and SSIM, confirming a trade-off between perceptual plausibility and paired-target fidelity.

These results indicate that semantic conditioning steers the reconstruction toward category-aligned perceptual features and image-set statistics, but does not necessarily improve fidelity to the specific

Table 2: Known-class conditioning ablation on the current parent-bound diagnostic split. C_{true} is the paired category label and C_{null} removes the category prompt while keeping the measurement and sampling contract fixed.

Model	Condition	MSE↓	PSNR↑	SSIM↑	LPIPS↓	FID↓
Base LDM	C_{true}	0.0223	16.80	0.335	0.566	85.49
Base LDM	C_{null}	**0.0190**	**17.48**	**0.367**	0.653	184.35
Full-UNet SFT	C_{true}	0.0239	16.45	0.294	**0.476**	**53.05**
Full-UNet SFT	C_{null}	0.0204	17.17	0.333	0.548	118.82

target instance. In the underdetermined passive NLOS setting, the category prior can therefore lower LPIPS and FID while reducing pixel-level and local-structure agreement with the paired target.

Accordingly, the known-category-conditioned model is treated as an auxiliary refinement track, while measurement-only reconstruction remains the primary setting for passive NLOS imaging without semantic side information.

4.4 Exploratory Preference Optimization

We finally examine whether preference optimization can further improve the Full-UNet SFT refinement model. DPO and GRPO inherit the same NLOSNet-T coarse initializer and SFT initialization, and are adapted using the train-only auxiliary subset. Their results are evaluated on the diagnostic-test pairs and are summarized in Table 3.

Table 3: Quantitative comparison of SFT, DPO, and GRPO.

Method	MSE ↓	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
SFT	0.0242	16.40	0.291	0.471	51.75
DPO	**0.0241**	**16.43**	**0.292**	**0.470**	**51.65**
GRPO	0.0243	16.39	0.290	0.471	51.94

Relative to the SFT model, DPO produces only marginal changes across the reported metrics, while GRPO remains close to the SFT baseline. Under the current reward definition, auxiliary-subset size, and optimization budget, neither preference-optimization method provides a material improvement in reconstruction quality. These findings motivate further investigation into how preferences should be constructed for passive NLOS reconstruction.

5 Conclusion

In this work, we introduce a physical benchmark for passive non-line-of-sight image reconstruction under indoor illumination and environmental reflections. The benchmark contains passive measurements and hidden-screen target images from ten categories, a fixed training and diagnostic-test split, two complementary reconstruction tracks, and reproducible preprocessing and evaluation protocols. Our experiments show that deterministic reconstruction generally preserves correspondence to the paired target more reliably, whereas known-class-conditioned diffusion refinement can improve perceptual and distribution-level quality without guaranteeing higher pixel-level or structural fidelity. These results highlight the distinction between perceptual plausibility and target-specific reconstruction fidelity in passive NLOS imaging. We hope this benchmark will support future research on more diverse physical scenes, target conditions, and reconstruction objectives.

AI USE STATEMENT

In this work, we used generative AI tools for paper editing and discussion, optimization of diagrams, and disscussion of experimental failures. We have not used generative AI tools for experimental design, physical experimental setup, data collection and generation, or to fabricate experimental metrics and results.

Additionally, we used generative AI tools to research papers in some areas. We checked LLM-generated code, papers and writing content.

REPRODUCIBILITY STATEMENT

It is important that the work published in ICLR is reproducible. Authors are strongly encouraged to include a paragraph-long Reproducibility Statement at the end of the main text (before references) to discuss the efforts that have been made to ensure reproducibility. This paragraph should not itself describe details needed for reproducing the results, but rather reference the parts of the main paper, appendix, and supplemental materials that will help with reproducibility. For example, for novel models or algorithms, a link to an anonymous downloadable source code can be submitted as supplementary materials; for theoretical results, clear explanations of any assumptions and a complete proof of the claims can be included in the appendix; for any datasets used in the experiments, a complete description of the data processing steps can be provided in the supplementary materials. Each of the above are examples of things that can be referenced in the reproducibility statement.

REFERENCES

Mufeed Batarseh, Sergey Sukhov, Zhiqin Shen, Heath Gemar, Reza Rezvani, and Aristide Dogariu. Passive sensing around the corner using spatial coherence. *Nature communications*, 9(1):3629, 2018.

Jeremy Boger-Lombard and Ori Katz. Passive optical time-of-flight for non line-of-sight localization. *Nature communications*, 10(1):3343, 2019.

Katherine L Bouman, Vickie Ye, Adam B Yedidia, Frédo Durand, Gregory W Wornell, Antonio Torralba, and William T Freeman. Turning corners into cameras: Principles and methods. 2017.

Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *European conference on computer vision*, pp. 17–33. Springer, 2022.

Wenzheng Chen, Fangyin Wei, Kiriakos N Kutulakos, Szymon Rusinkiewicz, and Felix Heide. Learned feature embeddings for non-line-of-sight imaging and recognition. *ACM Transactions on Graphics (ToG)*, 39(6):1–18, 2020.

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*, pp. 215–223. JMLR Workshop and Conference Proceedings, 2011.

Robinson Czajkowski and John Murray-Bruce. Two-edge-resolved three-dimensional non-line-of-sight imaging with an ordinary camera. *Nature Communications*, 15(1):1162, 2024.

Daniele Faccio, Andreas Velten, and Gordon Wetzstein. Non-line-of-sight imaging. *Nature Reviews Physics*, 2(6):318–327, 2020.

Miguel Galindo, Julio Marco, Matthew O'Toole, Gordon Wetzstein, Diego Gutierrez, and Adrian Jarabo. A dataset for benchmarking time-resolved non-line-of-sight imaging. In *ACM SIGGRAPH 2019 Posters*, pp. 1–2. 2019.

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.

Javier Grau Chopite, Matthias B Hullin, Michael Wand, and Julian Iseringhausen. Deep non-line-of-sight reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 960–969, 2020.

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.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 5967–5976. Ieee, 2017.

Ori Katz, Pierre Heidmann, Mathias Fink, and Sylvain Gigan. Non-invasive single-shot imaging through scattering layers and around corners via speckle correlations. *Nature photonics*, 8(10): 784–790, 2014.

Jonathan Klein, Martin Laurenzis, Dominik Ludewig Michels, and Matthias B. Hullin. A quantitative platform for non-line-of-sight imaging problems. In *British Machine Vision Conference*, 2018.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

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

Nicolas Scheiner, Florian Kraus, Fangyin Wei, Buu Phan, Fahim Mannan, Nils Appenrodt, Werner Ritter, Jurgen Dickmann, Klaus Dietmayer, Bernhard Sick, et al. Seeing around street corners: Non-line-of-sight detection and tracking in-the-wild using doppler radar. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2068–2077, 2020.

Sheila W. Seidel, John Murray-Bruce, Yanting Ma, Christopher Yu, William T. Freeman, and Vivek K Goyal. Two-dimensional non-line-of-sight scene estimation from a single edge occluder. *IEEE Transactions on Computational Imaging*, 7:58–72, 2021.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Kenichiro Tanaka, Yasuhiro Mukaigawa, and Achuta Kadambi. Polarized non-line-of-sight imaging. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2133–2142. IEEE, 2020.

Andreas Velten, Thomas Willwacher, Otkrist Gupta, Ashok Veeraraghavan, Moungi G Bawendi, and Ramesh Raskar. Recovering three-dimensional shape around a corner using ultrafast time-of-flight imaging. *Nature communications*, 3(1):745, 2012.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8228–8238. IEEE, 2024.

Yangyang Wang, Yaqin Zhang, Meiyu Huang, Zhao Chen, Yi Jia, Yudong Weng, Lin Xiao, and Xueshuang Xiang. Accurate but fragile passive non-line-of-sight recognition. *Communications Physics*, 4(1):88, 2021.

Yihao Wang, Zhigang Wang, Bin Zhao, Dong Wang, Mulin Chen, and Xuelong Li. Propagate and calibrate: real-time passive non-line-of-sight tracking. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 972–981. IEEE, 2023.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Cheng Wu, Jianjiang Liu, Xin Huang, Zheng-Ping Li, Chao Yu, Jun-Tian Ye, Jun Zhang, Qiang Zhang, Xiankang Dou, Vivek K Goyal, et al. Non–line-of-sight imaging over 1.43 km. *Proceedings of the National Academy of Sciences*, 118(10):e2024468118, 2021.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025.

Jiarui Zhang, Ruixu Geng, Xiaolong Du, Yan Chen, Houqiang Li, and Yang Hu. Passive non-line-of-sight imaging with light transport modulation. *IEEE Transactions on Image Processing*, 34: 410–424, 2024.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

Weicheng Zhao, Ruisi He, Bo Ai, Zhangdui Zhong, and Haoxiang Zhang. Vehicle localization based on hypothesis test in nlos scenarios. *IEEE Transactions on Vehicular Technology*, 71(2): 2198–2203, 2021.

A Appendix

A.1 Benchmark composition and record schema

Each record stores a fixed sample identifier, the raw camera measurement Y_{raw}, the background reference used for preprocessing, the fixed direct-light mask, the processed measurement $\widetilde{Y}$, the hidden target image I_h, and its category label C. The category label contains only the coarse STL-10 category; it contains no target pixels, instance identity, pose, location, or target-specific appearance.

A.1.1 Models and fixed training contracts

All model parameters and sampling settings are fixed before the 1,000 test records are evaluated. NLOSNet-T is a U-Net-derived measurement-to-image initializer whose native output is 96×96. It is evaluated as a measurement-only reconstruction and supplies the fixed bridge for the refinement track. U-Net, NAFNet, GAN, and NLOS-OT are measurement-only baselines. Base LDM is frozen, while Full-UNet SFT is trained using the 12,000 training records and evaluated with the same bridge, category condition, and sampling contract.

B Evaluation metrics and aggregation

All predictions and targets are converted to RGB, resized to 96×96 with bicubic interpolation, and normalized to $[0, 1]$. Paired metrics are computed for each of the 1,000 test pairs and then averaged. For image i, with D scalar RGB values,

$$\mathrm{MSE}_i = \frac{\|I_{h,i} - \widehat{I}_{h,i}\|_F^2}{D}, \qquad \mathrm{PSNR}_i = 10 \log_{10}\left(\frac{1}{\mathrm{MSE}_i}\right). \tag{A.1}$$

SSIM measures local structural consistency and LPIPS measures learned-feature distance. Their directions are SSIM ↑ and LPIPS ↓.

FID is computed once on the complete set of 1,000 predicted images and the complete set of 1,000 paired target images using the declared Inception feature extractor. FID is a distribution-level statistic and does not use the measurement–target pairing; it is therefore not interpreted as instance-level reconstruction accuracy. No class-wise FID is reported.

The resolution contract is fixed across the benchmark: raw captures are 640×480 RGB; NLOSNet-T consumes a 256×192 processed measurement and produces a 96×96 coarse image; U-Net, NAFNet, GAN, and NLOS-OT use 384×384 signed-residual adapters and produce96×96 RGB outputs; the diffusion bridge upsamples the coarse image to 512×512 with antialiased bilinear interpolation; Base LDM and Full-UNet SFT sample at 512×512 under the fixed sampler contract; and all predictions and targets are converted to RGB, bicubically resized to 96×96, and normalized to [0,1] for paired evaluation. .

We report paired MSE, PSNR, SSIM, and LPIPS, together with set-level FID. MSE and PSNR quantify pixel fidelity, SSIM local structural agreement, and LPIPS learned perceptual similarity. FID compares the reconstructed and target distributions over the complete diagnostic set without using measurement–target correspondence, so it is treated only as a supplementary diagnostic.

B.1 Complete quantitative comparison

The two blocks have different inference information: measurement-only methods receive the processed measurement, whereas refinement methods additionally receive the NLOSNet-T bridge and a known category condition. The blocks are therefore read as complementary tracks rather than as a single unconditional ranking.

B.2 Known-class conditioning ablation

To isolate the effect of the category input, we keep the measurement, coarse initializer, checkpoint, sampler, sampling settings, and test pairs fixed. The known-category condition C_{true} is compared

with a neutral condition C_{null}. The condition contains only the category and no instance-specific target information.

The true category condition lowers LPIPS and FID for both refinement models, while the neutral condition gives higher paired pixel and structural scores in this matched comparison. This result separates perceptual/distributional quality from correspondence to a particular test target; it does not show that the category condition recovers target identity, pose, or location.

B.3 Qualitative reconstruction and failure modes

Qualitative panels should use fixed test identifiers and the same paired target for every method. A recommended row is *target / processed measurement / NLOSNet-T coarse / NAFNet / Base LDM +* C_{true} */ Full-UNet SFT +* C_{true} */ Full-UNet SFT +* C_{null}. Deterministic outputs are native 96×96 images; diffusion outputs are 512×512 images and should be displayed at a common canvas size without implying additional measured detail.

The fixed test examples should document three recurring cases: a category is correct while instance appearance is wrong; a generated image is perceptually natural but mismatched to its paired target; and stronger generative refinement improves perceptual or distributional scores while reducing paired pixel or structural agreement. These examples are evidence for the metric trade-off and the limits of the controlled benchmark.

C Reproducibility record

The release accompanying this appendix should contain the 12,000/1,000 split manifest, the raw and processed measurement mapping, target paths, background and mask artifacts, train-statistics cache, model configuration files, checkpoint hashes, per-sample prediction files, metric CSV/JSON files, and the evaluation receipt. The receipt should record the split contract hash, sample count, prediction and target resolution, metric implementation, and the fact that the 1,000 test targets were accessed only for final evaluation.

Table A.1: Per-class quantitative comparison on the diagnostic split.

Class	Method	MSE ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Airplane	U-Net	0.0255	16.08	0.364	0.711
	GAN	0.0360	14.61	0.233	0.626
	NLOS-OT	0.0148	18.60	0.555	0.531
	NAFNet	**0.0117**	**19.52**	0.566	0.591
	NLOSNet-T	0.0122	19.37	**0.598**	0.701
	SFT	–	17.95	0.522	**0.432**
Bird	U-Net	0.0261	15.98	0.280	0.726
	GAN	0.0352	14.72	0.163	0.556
	NLOS-OT	0.0183	17.57	0.360	0.658
	NAFNet	0.0156	18.23	0.382	0.688
	NLOSNet-T	**0.0153**	**18.29**	**0.402**	0.823
	SFT	–	16.55	0.308	**0.535**
Car	U-Net	0.0407	14.09	0.195	0.732
	GAN	0.0526	13.01	0.129	0.572
	NLOS-OT	0.0294	15.44	0.254	0.611
	NAFNet	**0.0248**	**16.14**	0.276	0.693
	NLOSNet-T	0.0257	16.00	**0.278**	0.779
	SFT	–	14.81	0.211	**0.466**
Cat	U-Net	0.0246	16.29	0.250	0.716
	GAN	0.0329	14.98	0.162	**0.537**
	NLOS-OT	0.0157	18.20	0.315	0.647
	NAFNet	0.0143	18.57	0.327	0.716
	NLOSNet-T	**0.0137**	**18.81**	**0.348**	0.865
	SFT	–	16.99	0.256	0.541
Deer	U-Net	0.0211	16.86	0.251	0.716
	GAN	0.0307	15.27	0.146	0.507
	NLOS-OT	0.0130	18.99	0.312	0.687
	NAFNet	0.0119	19.34	0.320	0.720
	NLOSNet-T	**0.0115**	**19.54**	**0.339**	0.856
	SFT	–	17.38	0.238	**0.468**
Dog	U-Net	0.0269	15.87	0.256	0.709
	GAN	0.0347	14.74	0.155	0.517
	NLOS-OT	0.0170	17.86	0.332	0.627
	NAFNet	0.0156	18.19	0.336	0.681
	NLOSNet-T	**0.0150**	**18.33**	**0.363**	0.823
	SFT	–	16.34	0.254	**0.476**
Horse	U-Net	0.0319	15.22	0.221	0.705
	GAN	0.0393	14.27	0.139	0.502
	NLOS-OT	0.0217	16.78	0.266	0.611
	NAFNet	0.0195	17.20	0.282	0.699
	NLOSNet-T	**0.0190**	**17.32**	**0.296**	0.829
	SFT	–	15.70	0.213	**0.453**
Monkey	U-Net	0.0271	15.84	0.234	0.688
	GAN	0.0350	14.74	0.147	**0.494**
	NLOS-OT	0.0165	17.97	0.298	0.605
	NAFNet	0.0153	18.27	0.312	0.678
	NLOSNet-T	**0.0148**	**18.43**	**0.330**	0.832
	SFT	–	16.37	0.238	0.499
Ship	U-Net	0.0335	15.08	0.316	0.728
	GAN	0.0491	13.48	0.189	0.659
	NLOS-OT	0.0193	17.47	0.467	0.557
	NAFNet	0.0141	18.76	0.498	0.630
	NLOSNet-T	**0.0137**	**18.88**	**0.529**	0.719
	SFT	–	17.48	0.450	**0.421**
Truck	U-Net	0.0448	13.75	0.226	0.726
	GAN	0.0617	12.46	0.137	0.606
	NLOS-OT	0.0272	15.83	0.315	0.615
	NAFNet	**0.0228**	**16.55**	0.330	0.699
	NLOSNet-T	0.0231	16.53	**0.348**	0.784
	SFT	–	14.95	0.252	**0.479**