GANESH: Generalizable NeRF for Lensless Imaging

Rakesh Raj Madavan[1][*], Akshat Kaimal[1][*], Badhrinarayanan K V[1][*], Vinayak Gupta[2]

Rohit Choudhary[2], Chandrakala Shanmuganathan[1], Kaushik Mitra[2]

[1]Shiv Nadar University, Chennai   [2]Indian Institute of Technology, Madras

Abstract

Lensless imaging offers a significant opportunity to develop ultra-compact cameras by removing the conventional bulky lens system. However, without a focusing element, the sensor's output is no longer a direct image but a complex multiplexed scene representation. Traditional methods have attempted to address this challenge by employing learnable inversions and refinement models, but these methods are primarily designed for 2D reconstruction and do not generalize well to 3D reconstruction. We introduce GANESH, a novel framework designed to enable simultaneous refinement and novel view synthesis from multi-view lensless images. Unlike existing methods that require scene-specific training, our approach supports on-the-fly inference without retraining on each scene. Moreover, our framework allows us to tune our model to specific scenes, enhancing the rendering and refinement quality. To facilitate research in this area, we also present the first multi-view lensless dataset, LenslessScenes. Extensive experiments demonstrate that our method outperforms current approaches in reconstruction accuracy and refinement quality. Code and video results are available here.

1. Introduction

In recent years, mask-based lensless imaging systems have received significant interest due to their potential to offer compact, lightweight, and cost-efficient alternatives to traditional cameras [18]. Rather than utilizing standard optical lenses, these systems rely on amplitude [2] or phase masks [1, 6] placed in close proximity to the sensor. This design not only minimizes the physical dimensions and mass of the imaging device, i.e., a smaller form factor, but also enables the use of non-traditional sensor geometries, such as spherical, cylindrical, or even flexible configurations [28]. In the absence of a conventional focusing element, the measurements captured by the sensor are not straightforward images of the scene. Instead, they consist
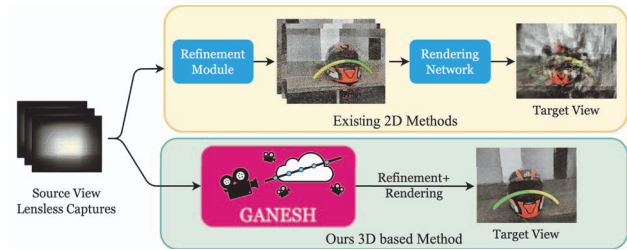
Figure 1. Reconstructing 3D scenes from multi-view lensless captures presents significant challenges. To tackle this, we propose GANESH, a novel framework that refines lensless captures while simultaneously rendering novel views. Existing 2D approaches address this task in a sequential, two-step process, resulting in suboptimal 3D reconstruction quality. In contrast, GANESH integrates these two stages into a unified framework, enabling joint optimization for superior novel view synthesis.

of intricate, multiplexed data that encode the light information from the scene in a highly compressed and non-intuitive form. This image formation model necessitates advanced computational techniques to decode and reconstruct the original scene, as the sensor no longer produces a one-to-one representation of the visual information but rather a superimposition of light intensities across the entire field of view.

Numerous studies have investigated the problem of scene reconstruction from single-image lensless captures [3, 12, 18, 25]. For instance, FlatNet [18] employs a two-step process to recover the scene. Initially, a trainable inversion module is utilized to reconstruct most of the scene's details; however, the resulting output still contains significant noise, which is subsequently addressed by a refinement network. While there are several works in this field of 2D scene reconstruction, there has been little to no work in reconstructing a 3D scene from multi-view lensless captures. This advancement is particularly significant for applications such as endoscopic surgery, where the compact size of lensless cameras offers a substantial advantage. Achieving 3D reconstruction from multi-view lensless images could greatly benefit medical fields and AR/VR applications [13, 17]

Recently, NeRFs have amassed a lot of attention for their

ability to reconstruct 3D scenes from real-world multi-view captures. Most NeRF-based methods rely on RGB images to model the underlying 3D scene. There have been several works in this space which prompt NeRFs with different image modalities like Thermal Images [33], Event Data [14], Multi-spectral captures [35], Single-Photon Data [15]. For instance, in Thermal NeRF [33], they present an approach for reconstructing novel views exclusively from thermal imagery, particularly beneficial in visually degraded robotics scenarios. Ev-NeRF [14] learns to reconstruct multi-view images from a raw stream of event data captured by a neuromorphic camera, which helps in better reconstruction, especially in high dynamic range scenes. Despite these promising developments, a significant limitation of current NeRF models is their reliance on scene-specific training.

One might consider using established refinement techniques such as FlatNet [18] and then feeding the outputs to rendering networks like NeRF or Gaussian Splatting [16]. Though this is a viable option, it comes with its downside. NeRF and Gaussian Splatting operate on the principle of using images solely for supervisory purposes rather than as direct inputs. Such models cannot be trained on paired multi-view lensless and RGB images, resulting in poor novel view synthesis quality. Secondly, while Gaussian Splatting offers improved computational efficiency compared to NeRF, it may be susceptible to overfitting the noisy outputs produced by FlatNet, potentially resulting in suboptimal reconstruction quality. Finally, each model must be re-trained from scratch whenever presented with a new set of images, limiting their scalability and practical application across diverse scenarios.

Generalizable Radiance Fields methods have recently gained traction due to their ability to perform on-the-fly inference on new scenes without specific training. Many of these approaches [29, 30, 36] utilize a set of source views and enforce epipolar constraints across them to generate novel target views. The current state-of-the-art method, GNT [29], employs a transformer-based architecture to aggregate epipolar information from multiple views effectively. It then accumulates these point features along each ray to compute the final pixel color, enabling accurate and efficient rendering of new views. However, as previously noted, most Radiance Fields methods have predominantly focused on using RGB images as input, with limited exploration of alternative modalities such as lensless captures. Given the diverse applications of lensless imaging, incorporating this modality into radiance fields presents significant opportunities for expanding their utility across various fields.

In this paper, we introduce a novel methodology that enables us to reconstruct scenes from multi-view lensless captures in a generalizable setting. Unlike traditional methods that require scene-specific training for each new dataset, our

technique can generalize across various multi-view lensless inputs to render novel views. Our proposed method, **GANESH**, can effectively reconstruct 3D scenes from lensless data. Our model is trained on extensive data of synthetically generated multi-view lensless images. Despite being trained exclusively on synthetic data, the model can refine and render novel views when applied to real multi-view lensless captures. Experimental results illustrate the model's effective generalization to both synthetic and real-world scenes. Additionally, our method allows for scene-specific tuning with minimal finetuning steps, enhancing reconstruction quality. We also present *LenslessScenes*, a dataset of real-world multi-view lensless captures comprising six distinct scenes. These scenes, acquired in a controlled laboratory setting, are accompanied by ground truth data for precise quantitative evaluation. The key contributions of our work are as follows:

- We present a novel framework that simultaneously achieves refinement and rendering of lensless captures.

- Our approach is generalizable, i.e., it can render views on-the-fly without any need for scene-specific training.

- We present *LenslessScenes*, the first dataset of multi-view lensless captures.

- Our experimental results demonstrate that the proposed method outperforms existing techniques that separately handle refinement and novel view synthesis.

2. Related Works

2.1. Lensless Imaging

Lensless imaging refers to capturing images of a scene without using a traditional lens to focus incoming light. Historically, this technique has been widely utilized in X-ray and gamma-ray imaging for astronomical purposes [7, 11], but its application in the visible spectrum has only recently been explored. In lensless systems, the scene is captured either directly by the sensor [19] or after being modulated by a mask element [1, 2, 27]. Our research focuses specifically on mask-based lensless imaging, where replacing the lens with a mask leads to a highly multiplexed sensor capture that does not directly resemble the original scene. Consequently, advanced computational techniques are required to reconstruct the image. FlatNet [18] addresses this by employing a trainable inversion module coupled with a U-Net refinement architecture to recover the scene from a lensless capture. FlatNet3D [3] further extends this by predicting the scene's intensity and depth from a single lensless capture using a neural network. However, no prior works have explored the use of multi-view images in lensless imaging. Our proposed method, GANESH, seeks to refine and render novel views from multi-view lensless
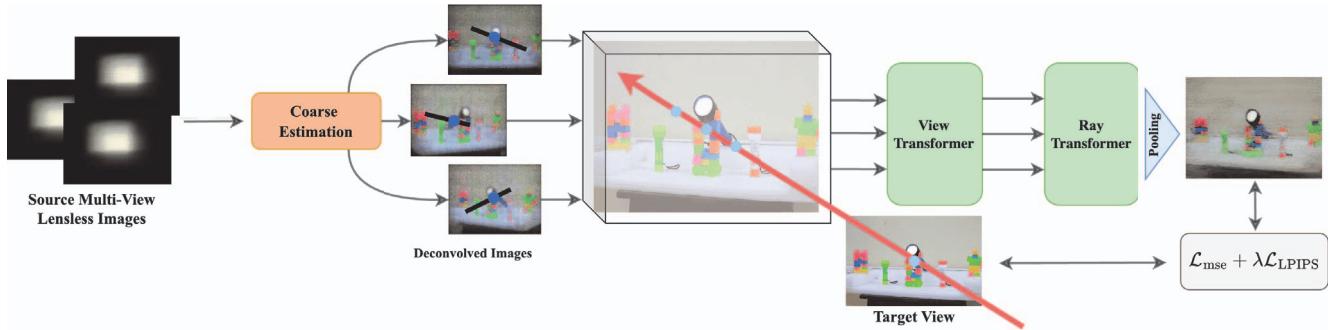
Figure 2. Overview of GANESH: 1) Given multi-view lensless images of a scene, we first Wiener deconvolve the lensless captures to obtain coarse images. 2) These are then passed onto a deep convolutional network to extract features for every input view. 3) Using the source view features, we estimate the target refined rendered view via an epipolar-based rendering pipeline. 4) By supervising this pipeline end-to-end on paired synthetic data, our model learns to inherently refine the coarse estimated images and simultaneously render novel views eliminating the need for a separate refiner. Our method can directly generalize to any new scene during inference.

captures, enabling simultaneous refinement and novel view synthesis.

2.2. Neural Radiance Fields

NeRF [22] utilizes a Multi-Layer Perceptron (MLP) to represent a scene as a continuous 5D function, incorporating both spatial location and viewing direction. This framework encodes the scene's geometry and appearance by mapping a 3D spatial function and a 2D directional function to outputs of color and density. Since its introduction, numerous works have sought to enhance NeRF's rendering capabilities [8, 23, 24, 31]. Mip-NeRF [4, 5], for instance, improves upon the original method by employing an approximate cone tracing approach rather than the ray tracing method used in standard NeRF. PointNeRF [32] further advances this framework by introducing feature point clouds as an intermediate step in the volumetric rendering process, enhancing the overall quality of the rendered output.

NeRF has also been extended to non-RGB input modalities, such as in Thermal-NeRF [33] and Hyperspectral NeRF [10] etc. Thermal-NeRF [33] reconstructs 3D scenes using infrared (IR) images as input, focusing on preserving thermal characteristics more accurately. Similarly, Hyperspectral NeRF [10] adapts NeRF for hyperspectral imaging, which captures data across a broad range of the electromagnetic spectrum. In this work, we explore a related approach by reconstructing 3D scenes from lensless captures, leveraging this alternative imaging modality to extend NeRF's capabilities.

2.3. Generalizable Radiance Fields

A significant drawback of NeRF is its lack of generalization, as each model is specifically trained for a single scene and cannot be easily transferred to new, unseen scenes. Various approaches, such as MVSNeRF [9], IBR-Net [30], and Generalizable NeRF Transformer (GNT) [29],

have addressed this limitation by developing models capable of generalizing across different scenes. MVSNeRF [9] enhances novel view synthesis speed by incorporating multi-view stereo methods. IBRNet [30] offers a generalizable image-based rendering framework that generates novel views from arbitrary inputs without requiring per-scene optimization. GNT [29] leverages a transformer-based architecture to synthesize novel views by aggregating the information across source views based on epipolar constraints. Building on these advancements, we propose a generalizable model designed explicitly for novel view synthesis from lensless captures.

3. Preliminary: Generalizable NeRF Transformer

Our method is built on the Generalizable NeRF Transformer (GNT) framework, aimed at generating accurate images from noisy input views. Given N calibrated source views with known pose information $\{I_i, P_i\}_{i=1}^N$, the objective is to synthesize a novel target view I_T, even for scenes not encountered during training, ensuring generalizability. To achieve this, deep convolutional features F_i are extracted from each source view. During the rendering process, rays are cast into the scene, with K points sampled along each ray, defined by the camera center o and ray direction d. Each point is projected onto the source views using a projection operator Π_i, and the nearest features in the image plane are retrieved.

These features are combined into a point feature $f(t)$ using a permutation-invariant aggregation function $\mathcal{F}_{view}$ that is trained to handle occlusions.

$$f(t) = \mathcal{F}_{view}(\{F_{\Pi_i(r(t))}\}_{i=1}^N) \qquad (1)$$

Finally, the accumulated point features are used to compute the target color $c(r)$ via an aggregation function

$\mathcal{F}_{point}$, resulting in the rendered image.

$$c(r) = \mathcal{F}_{point}(\{\boldsymbol{f}(t_i)\}_{i=1}^K) \qquad (2)$$

We leverage GNT's capabilities to generate target novel views and visually enhance the reconstructions of lensless captures, avoiding the need for a refinement network before novel view synthesis.

4. GANESH

Overview. We introduce GANESH, a novel framework to perform generalizable novel view synthesis from lensless captures, as illustrated in Fig. 2. The task is to generate refined novel views from N calibrated multi-view lensless images of a scene, with known camera poses, while ensuring the model generalizes to unseen scenes. Our method builds upon existing GNT architecture [29] but conditions the scene representation and rendering processes based on the captured multi-view lensless images. First, these lensless captures are passed through a simple Wiener deconvolution filter to obtain a coarse estimate of the scene (Sec 4.1). The deconvolved outputs of this filter are then passed on to a generalizable view synthesis model, which performs both refinement and rendering simultaneously. Such a pipeline can be trained end-to-end on synthetically generated scenes (Sec 4.2) and directly transferred to any real scene without additional optimization (Sec. 4.3).

4.1. Coarse Scene Estimation

Given the global multiplexing of lensless captures, we cannot directly feed them into the radiance fields model to render novel views. Hence, to reconstruct the RGB image from the lensless captures, these need to be deconvolved with the lensless camera's point spread function (PSF) to obtain coarse reconstructed images. For this, we utilize wiener deconvolution, which accepts the lensless capture and the point spread function as the input and returns the reconstructed image. For an RGB image I and a PSF kernel H, the observed lensless image is given by:

$$G(x,y) = (I * H)(x,y) + n(x,y) \qquad (3)$$

The wiener deconvolution will produce the following estimate for the RGB image:

$$\hat{I}(\omega_x, \omega_y) = \frac{H^*(\omega_x, \omega_y)}{|H(\omega_x, \omega_y)|^2 + K} G(\omega_x, \omega_y), \qquad (4)$$

where $\hat{I}(\omega_x, \omega_y)$ is the Fourier transform of the estimated original image, $H(\omega_x, \omega_y)$ is the Fourier transform of the PSF, $H^*(\omega_x, \omega_y)$ is the complex conjugate of $H(\omega_x, \omega_y)$, $G(\omega_x, \omega_y)$ is the Fourier transform of the lensless capture, and K is the noise-to-signal ratio. Note that the deconvolved outputs are extremely noisy and require refinement to reconstruct the scene.

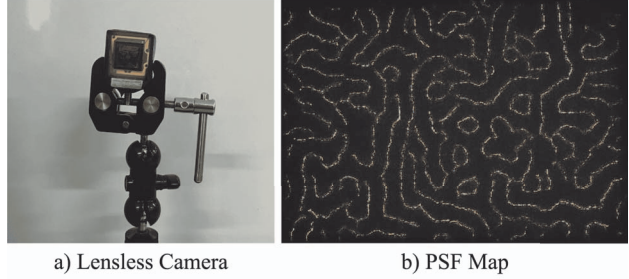a) Lensless Camera   b) PSF Map

Figure 3. a) Lensless camera setup used to capture the real-world dataset *LenslessScenes*. b) The calibrated Point Spread Function (PSF) is used for simulating lensless captures in our synthetically generated dataset.

4.2. Simulating Lensless Imaging

Given the absence of a large-scale paired dataset comprising multi-view lensless captures and corresponding ground-truth RGB images, which is essential for training our generalizable model, we propose to simulate lensless data using existing multi-view RGB datasets. Specifically, we approximate lensless captures by convolving each ground-truth image with the lensless camera's point spread function (PSF). However, lensless measurements are frequently corrupted by noise in real-world applications, necessitating refinement steps to recover the original scene. We artificially introduce 40dB of Gaussian noise to the convolved lensless captures to simulate these practical conditions. This addition of noise ensures that the model is exposed to noisy data during training, helping it learn robust features that transfer well from synthetic to real-world scenarios.

An important design choice in our simulation process involves using a grayscale PSF map instead of an RGB PSF map for the convolution operation. Through empirical studies, we found that the grayscale PSF map more accurately mimics real-world lensless captures, which inherently lack color-channel-specific information due to the absence of a lens. As a result, the grayscale PSF map provides a more realistic approximation of the sensor measurements in lensless imaging systems, leading to improved reconstruction quality during inference, as demonstrated in the ablation study presented in Sec. 5.6.

4.3. Training and Inference

To optimize the network end-to-end, we use 2 losses to ensure view consistent rendering and accurate refinement.

Mean squared error: We use MSE to measure the distortion between the ground truth and the rendered output given by.

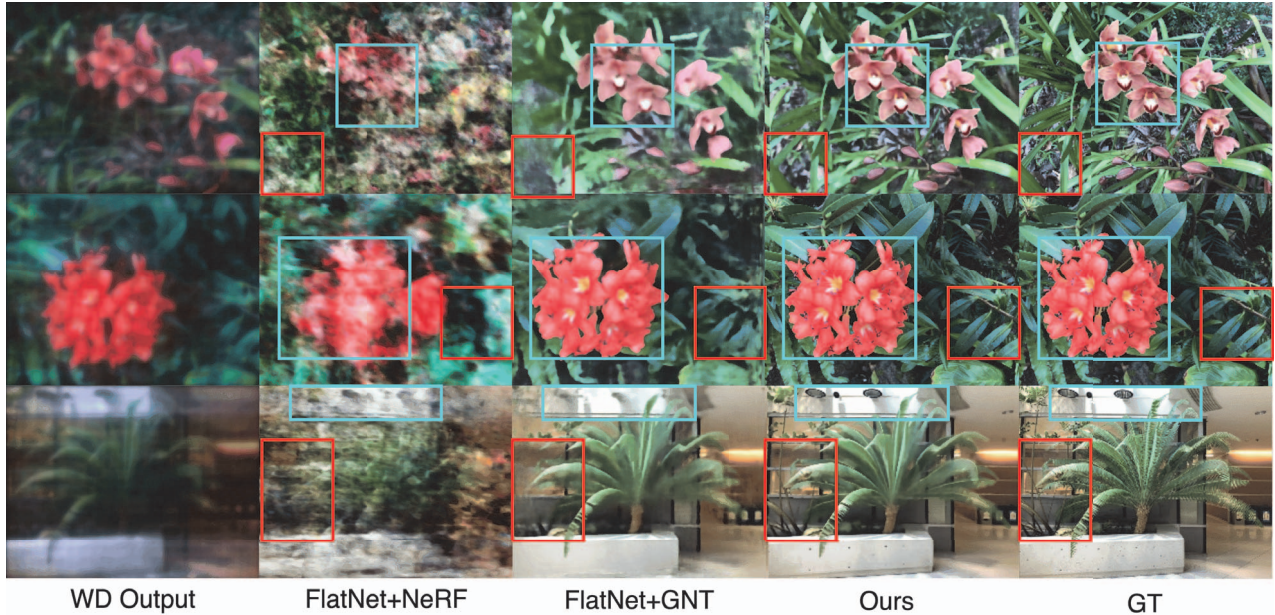$$\mathcal{L}_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{x}_i)^2 \qquad (5)$$

Figure 4. Qualitative results for scene-specific experiment on the synthetic NeRF-LLFF dataset. FlatNet+NeRF baseline exhibits significant artifacts and fails to preserve critical scene geometry. While FlatNet+GNT improves scene geometry reconstruction, it introduces excessive smoothing, resulting in the loss of high-frequency details. In contrast, our proposed method accurately reconstructs scene geometry and renders novel views, preserving high-frequency details and delivering superior visual fidelity. Note that the input to all the baselines and our model is the direct lensless capture. In the first column of this figure and all the subsequent figures, we show the Wiener Deconvolution (WD) output just for visualisation.

where, x and $\hat{x}$ represents the ground truth and the predicted images, respectively.

Perceptual Loss [34]: In addition to MSE loss, we employ a perceptual loss to capture higher-level feature similarities between the ground truth and the rendered output. We use a pre-trained VGG-19 network to achieve this, extracting features from both the ground truth and predicted views at various layers. The perceptual loss is formulated as follows:

$$\mathcal{L}_{\text{Perceptual}} = \sum_l \lambda_l \frac{1}{N_l} \sum_{i,j} (F_l(x)_{i,j} - F_l(\hat{x})_{i,j})^2 \quad (6)$$

where F represents the VGG network and λ_l represents the weight given for each layer in the network.

Final Loss. A weighted sum of the MSE and Perceptual Loss is taken to compute the final loss.

$$\mathcal{L} = \mathcal{L}_{\text{mse}} + \lambda \mathcal{L}_{\text{perceptual}} \quad (7)$$

Through training on synthetically generated lensless scenes, we observe that GANESH successfully generalizes to real-world data without additional finetuning. This generalization capability aligns with results observed in prior 2D-based methods [3, 18], extending the hypothesis into the 3D domain and confirming its applicability to lensless image reconstruction.

5. Experiments and Results

5.1. Datasets

We leverage the IBRNet [30] and LLFF [21] datasets to create a synthetic dataset comprising lensless images and their corresponding ground truth RGB images. These datasets, which are well-established benchmarks for novel view synthesis (NVS), include a total of 110 scenes. For validation purposes on synthetic scenes, we utilize the NeRF-LLFF dataset [22].

5.2. LenslessScenes Real-World Dataset

To complement the synthetic data and test our model's robustness, we collected the first real-world multi-view lensless dataset. A set of 7 scenes was collected in the lab environment, consisting of an average of around 20 frames, each collected in a forward-facing setting. We replicate the setup of FlatNet [18] to collect real-data captures along with ground truth labels using a monitor capture setup. We used the *BASLER Ace acA4024-29uc* with 12.2MP *Sony IMX226* sensor with a pixel size of $1.85\mu m$ to capture the scenes, see Fig. 3. This data collection involved multiple scenes, each captured with a detailed environmental setup. A point-sized light source was employed to calibrate the camera's point spread function (PSF), which is essential for accurate recon-

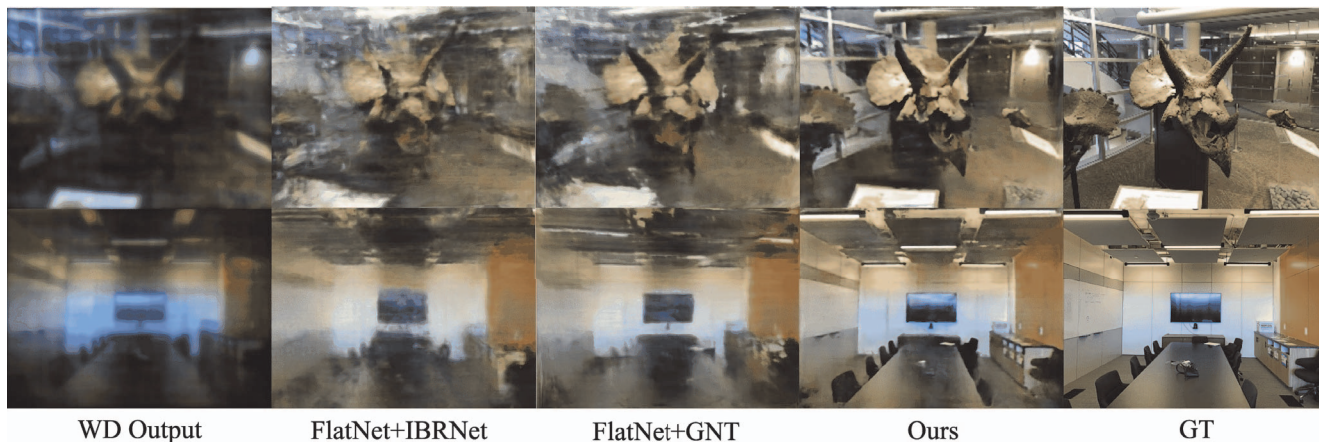| WD Output | FlatNet+IBRNet | FlatNet+GNT | Ours | GT |

Figure 5. Qualitative results for generalizable setting conducted on the synthetic NeRF-LLFF dataset. We observe that the FlatNet+IBRNet and FlatNet+GNT baselines fall short in rendering high-fidelity novel views compared to our method. Our approach demonstrates superior recovery of fine geometry and textures.

struction. Additionally, a white display screen was used to capture environmental noise, which was subsequently subtracted from the images to enhance the quality of the captured data.

5.3. Implementation Details

Our entire pipeline is trained end-to-end using datasets of multi-view posed images. For consistency, we adopt the same input view sampling strategy employed by IBR-Net [30], selecting 8 to 12 source views during training while fixing the number of source views to 10 during inference for novel scenes. Instead of training the model from scratch, we initialize our network with a pretrained checkpoint from GNT [29], allowing us to leverage its generalization capabilities.

The optimization of our model for rendering clean images is carried out using the Adam optimizer [20], with an initial learning rate set to 5×10^{-4}, which decays progressively over 300k training iterations. In each iteration, 576 rays are cast, with each ray sampling 192 points. The weight λ in our loss function is assigned a value of 0.4, while the parameter K from the Wiener Deconvolution process is 0.00045. All experiments are conducted on a single NVIDIA RTX 3090 GPU, with the entire training process taking approximately 24 hours to complete. As we cannot run COLMAP [26] directly on the lensless captures, we use the images recovered from FlatNet to run COLMAP and extract camera poses and bounds.

5.4. Comparisons

In the absence of prior research specifically addressing novel view synthesis for lensless imagery, we propose several baseline approaches to evaluate our method.

FlatNet+NeRF. This approach involves first applying FlatNet to refine lensless captures, followed by utilizing NeRF for rendering. This is a major downside of using scene-specific methods like NeRFs since they rely on the supervision of images and hence cannot be trained explicitly to refine the lensless captures. Additionally, this baseline does not offer generalization across different scenes.

FlatNet+IBRNet. Here, we replace NeRF with the generalizable IBRNet for rendering, while maintaining FlatNet as the refinement module.

FlatNet+GNT. This baseline adopts a similar strategy to the previous ones but uses GNT instead of IBRNet for rendering. Both FlatNet+IBRNet and FlatNet+GNT are designed to generalize across different scenes.

Table 1. Quantitative results for scene-specific experiment on synthetic dataset averaged across the 8 scenes from the NeRF-LLFF dataset. The best scores and second best scores are highlighted with their respective colors

Models	PSNR↑	SSIM↑	LPIPS↓
FlatNet+NeRF	13.7	0.057	0.705
FlatNet+GNT	20.2	0.27	0.59
Ours	22.8	0.71	0.27

5.5. Results

Scene-Specific Finetuning. We aim to synthesize refined novel views from multi-view lensless images by training the model with supervision from corresponding ground truth RGB images. To evaluate the effectiveness of our approach, we compare the results with two baseline methods: FlatNet+NeRF and FlatNet+GNT. We supervise the NeRF model using FlatNet outputs as it is not possible to supervise

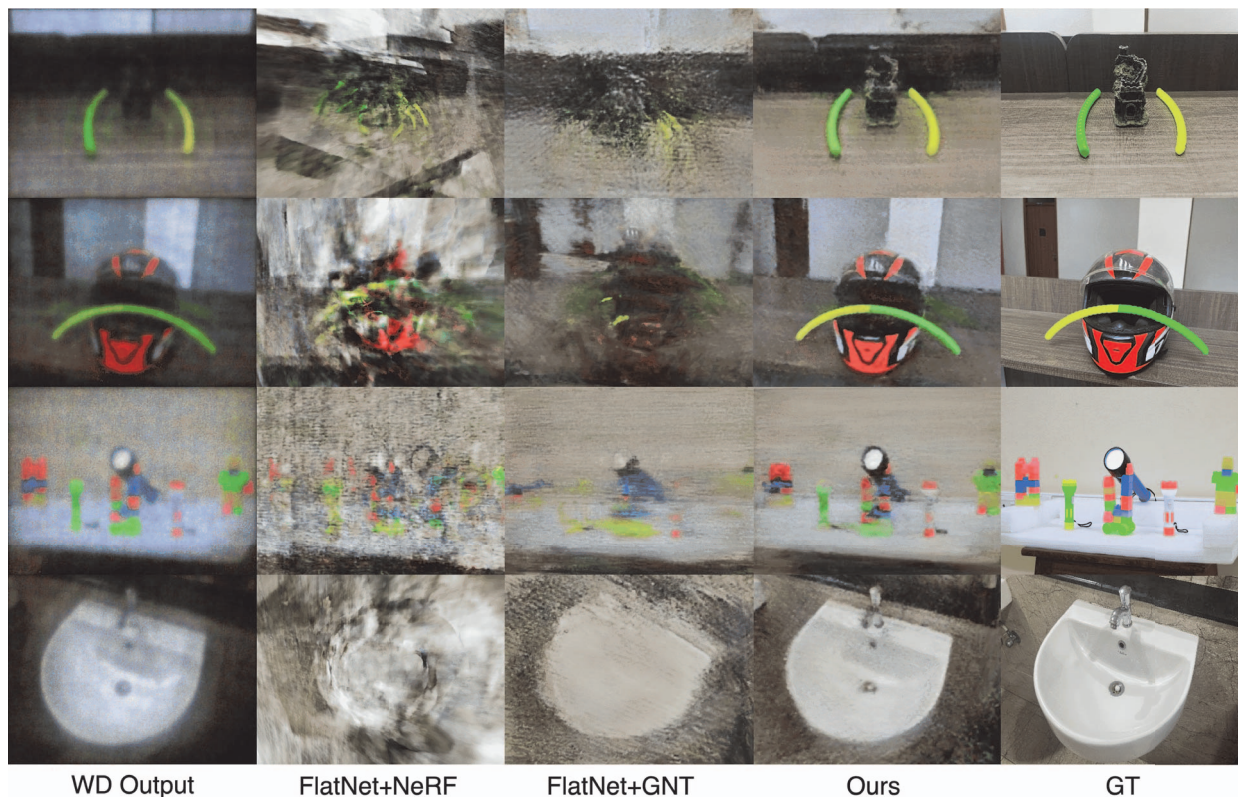| WD Output | FlatNet+NeRF | FlatNet+GNT | Ours | GT |

Figure 6. Qualitative results on the real-world LenslessScenes dataset. We show results for 4 scenes from the LenslessScenes dataset. Even though our model was trained on synthetic data, it learns to generalize to the real world captures and outperforms both the baselines in terms of render quality.

it with ground truth labels since it takes in coordinates and viewing directions as input rather than an image. In contrast, for the FlatNet+GNT baseline, the outputs of FlatNet are provided as input to GNT, and supervision is conducted using ground truth RGB images. The evaluation is carried out on the synthetic NeRF-LLFF dataset, with the results detailed in Table 1. Our proposed method, GANESH, demonstrates superior performance across all three metrics compared to both baseline models. This improvement can be attributed to the joint refinement and rendering strategy, which enhances the overall reconstruction quality. In a multi-view setup like ours, information lost in one view can be recovered from another, a feature that our method leverages. In contrast, the baseline methods perform refinement and rendering sequentially, missing out on the potential benefits of joint optimization.

Figure 4 illustrates the qualitative differences in performance. The FlatNet+NeRF model suffers from prominent ghosting artifacts, likely due to inconsistencies in the outputs generated by FlatNet, which are used for supervising NeRF. While FlatNet+GNT improves upon these issues through its more complex architecture, it still exhibits excessive smoothening effects. In contrast, our

method achieves superior refinement and rendering, producing high-quality novel views. This demonstrates that our joint approach to refining and rendering in a 3D context significantly enhances the accuracy of novel view synthesis in lensless imaging.

Table 2. Quantitative results for generalizable setting on synthetic NeRF-LLFF dataset.

Models	PSNR↑	SSIM↑	LPIPS↓
FlatNet+IBRNet	15.31	0.42	0.645
FlatNet+GNT	16.47	0.43	0.63
Ours	21.75	0.68	0.366

Generalizable Setting. We evaluate our approach in a generalization scenario where the model is tested on unseen scenes, using eight scenes from the NeRF-LLFF dataset. The average results across these scenes are summarized in Table 2. Our method demonstrates clear superiority over both the FlatNet+IBRNet and FlatNet+GNT baselines, achieving higher performance across all three evaluation metrics. Figure 5 provides a visual comparison of the novel views rendered by our method and the baselines. Our

model successfully recovers intricate scene details, such as the subtle grooves on the plastic fortress and the veins on the leaves. This highlights the effectiveness of our method in rendering high-quality novel views, even in challenging generalization settings.

Table 3. Quantitative results on the real-world *LenslessScenes* dataset. Our method shows strong generalisable capability across real scenes compared to the other baselines.

Models	PSNR↑	SSIM↑	LPIPS↓
FlatNet+NeRF	11.2	0.34	0.67
FlatNet+GNT	13.56	0.44	0.60
Ours	18.45	0.62	0.47

Quantitative results on Real-world dataset To assess the robustness of our model on real-world data, we evaluate its performance on the *LenslessScenes* dataset. We compare the results against the FlatNet+NeRF and FlatNet+GNT baselines. As in previous experiments, the FlatNet+NeRF baseline supervises NeRF using the output from FlatNet, while FlatNet+GNT is applied directly to the real scenes without any finetuning for the real-world data. Quantitative and qualitative evaluations of these comparisons are presented in Table 3 and Fig. 6. Qualitatively, our model demonstrates a superior ability to recover fine scene details compared to the baselines. For example, the shape and geometry of the toys surrounding the torch in the figure are visible, contrasting with the results of the baseline methods, where they are hardly discernable. Quantitative results further support this observation, showcasing improved performance when transferring from synthetic to real-world data. These findings highlight the efficacy of our joint refinement and rendering approach, which significantly enhances 3D scene reconstruction compared to methods that treat refinement and rendering as independent tasks.

Models	PSNR↑/SSIM↑/LPIPS↓
RGB PSF	22.43/0.62/0.47
Gray PSF	22.75/0.63/0.39

Table 4. Ablation Study: Grayscale vs RGB PSF.

Models	PSNR↑/SSIM↑/LPIPS↓
Ours w/o noise	11.2/0.34/0.67
Ours with noise	18.45/0.62/0.47

Table 5. Ablation Study: Noise Augmented Training.

5.6. Ablation Studies

We conduct the following ablations to validate our lensless simulation pipeline and provide quantitative results evaluated on real world data to test the effectiveness of our simulation.

Grayscale vs RGB PSF map. Our experimental results indicate that utilizing a grayscale PSF for reconstructing lensless images consistently outperforms using a 3-channel RGB PSF, as demonstrated in Table 4. We hypothesize that this advantage stems from the fact that the grayscale PSF more closely mirrors how actual lensless cameras capture images. Consequently, using the grayscale PSF map during the simulation process yields superior results when tested on to real-world datasets.

Synthetic Noise. While our model is resilient to low levels of noise, training with synthetic noise added to the lensless images proved crucial for adapting to real-world scenes. Gaussian noise was artificially introduced before reconstruction during the training on synthetic captures as read noise. We tested this model on real-world data, and the results are presented in Table 5. The significant noise observed in real-world captures necessitated this approach, and incorporating noise in the training pipeline enhanced the model's robustness and ability to generalize to real-world scenarios.

6. Discussion

Limitations and Future Work. While GANESH demonstrates the ability to refine and render novel views from lensless captures in both scene-specific and generalizable settings, it is not without limitations. A primary challenge it faces is that the model requires substantial training time when generalizing across diverse scenes, and its inference speed is not optimized for real-time applications, posing a limitation for on-the-fly rendering tasks. Finally, GANESH is an entirely data-driven model trained on an extensive dataset to mimic the reconstruction task. Integrating physical light transport models into radiance fields could be a promising direction for future improvements, combining data-driven approaches with physical principles for more accurate and efficient lensless rendering. We depend on Flatnet to get camera poses and we aim to develop pose free rendering in the future.

Conclusions. In this work, we present GANESH, a novel framework that integrates refinement and novel view rendering from multi-view lensless captures within a generalizable framework, demonstrating robustness in real-world scenarios. While existing approaches such as FlatNet for refinement and NeRF or Gaussian Splatting (GS) for rendering could be employed sequentially, they are fundamentally constrained by their reliance on image supervision, making training on extensive synthetic datasets impractical. In contrast, GANESH enables joint refinement and rendering, addressing this limitation and achieving superior performance in novel view synthesis. This approach is crucial for various applications, including medical imaging (e.g., endoscopy), augmented and virtual reality (AR/VR), and wearable technologies.

References

[1] Nick Antipa, Grace Kuo, Reinhard Heckel, Ben Mildenhall, Emrah Bostan, Ren Ng, and Laura Waller. Diffusercam: lensless single-exposure 3d imaging. *Optica*, 5(1):1–9, 2017. 1, 2

[2] M Salman Asif, Ali Ayremlou, Aswin Sankaranarayanan, Ashok Veeraraghavan, and Richard G Baraniuk. Flatcam: Thin, lensless cameras using coded aperture and computation. *IEEE Transactions on Computational Imaging*, 3(3):384–397, 2016. 1, 2

[3] Dhruvjyoti Bagadthey, Sanjana Prabhu, Salman S Khan, D Tony Fredrick, Vivek Boominathan, Ashok Veeraraghavan, and Kaushik Mitra. Flatnet3d: intensity and absolute depth from single-shot lensless capture. *JOSA A*, 39(10):1903–1912, 2022. 1, 2, 5

[4] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021. 3

[5] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5470–5479, 2022. 3

[6] Vivek Boominathan, Jesse K Adams, Jacob T Robinson, and Ashok Veeraraghavan. Phlatcam: Designed phase-mask based thin lensless camera. *IEEE transactions on pattern analysis and machine intelligence*, 42(7):1618–1629, 2020. 1

[7] Ezio Caroli, JB Stephen, G Di Cocco, L Natalucci, and A Spizzichino. Coded aperture imaging in x-and gamma-ray astronomy. *Space Science Reviews*, 45(3):349–403, 1987. 2

[8] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European conference on computer vision*, pages 333–350. Springer, 2022. 3

[9] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14124–14133, 2021. 3

[10] Gerry Chen, Sunil Kumar Narayanan, Thomas Gautier Ottou, Benjamin Missaoui, Harsh Muriki, Cédric Pradalier, and Yongsheng Chen. Hyperspectral neural radiance fields. *arXiv preprint arXiv:2403.14839*, 2024. 3

[11] RH Dicke. Scatter-hole cameras for x-rays and gamma rays. *Astrophysical Journal, vol. 153, p. L101*, 153:L101, 1968. 2

[12] Haoyang Ge, Qiao Feng, Hailong Jia, Xiongzheng Li, Xiangjun Yin, You Zhou, Jingyu Yang, and Kun Li. Lpsnet: End-to-end human pose and shape estimation with lensless imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1471–1480, 2024. 1

[13] Stéphanie Guérit, Siddharth Sivankutty, John Aldo Lee, Hervé Rigneault, and Laurent Jacques. Compressive lensless endoscopy with partial speckle scanning. *arXiv preprint arXiv:2104.10959*, 2021. 1

[14] Inwoo Hwang, Junho Kim, and Young Min Kim. Ev-nerf: Event based neural radiance field. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 837–847, 2023. 2

[15] Sacha Jungerman and Mohit Gupta. Radiance fields from photons. *arXiv preprint arXiv:2407.09386*, 2024. 2

[16] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), July 2023. 2

[17] Salman S Khan, Vivek Boominathan, Ashok Veeraraghavan, and Kaushik Mitra. Designing optics and algorithm for ultra-thin, high-speed lensless cameras. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1583–1588. IEEE, 2023. 1

[18] Salman Siddique Khan, Varun Sundar, Vivek Boominathan, Ashok Veeraraghavan, and Kaushik Mitra. Flatnet: Towards photorealistic scene reconstruction from lensless measurements. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):1934–1948, 2020. 1, 2, 5

[19] Ganghun Kim, Kyle Isaacson, Rachael Palmer, and Rajesh Menon. Lensless photography with only an image sensor. *Applied optics*, 56(23):6450–6456, 2017. 2

[20] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[21] Ben Mildenhall, Pratul P. Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 2019. 5

[22] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I*, pages 405–421, 2020. 3, 5

[23] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022. 3

[24] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. In *ACM Conference and Exhibition on Computer Graphics and Interactive Techniques in Asia (SIGGRAPH Asia)*, 2021. 3

[25] Joshua D Rego, Karthik Kulkarni, and Suren Jayasuriya. Robust lensless image reconstruction via psf estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 403–412, 2021. 1

[26] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 6

[27] David G Stork and Patrick R Gill. Lensless ultra-miniature cmos computational imagers and sensors. *Proc. Sensor-comm*, pages 186–190, 2013. 2

[28] Eric J Tremblay, Ronald A Stack, Rick L Morrison, and Joseph E Ford. Ultrathin cameras using annular folded optics. *Applied optics*, 46(4):463–471, 2007. 1

[29] Mukund Varma, Peihao Wang, Xuxi Chen, Tianlong Chen, Subhashini Venugopalan, and Zhangyang Wang. Is attention all that nerf needs? In *The Eleventh International Conference on Learning Representations*, 2023. 2, 3, 4, 6

[30] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2021. 2, 3, 5, 6

[31] Suttisak Wizadwongsa, Pakkapon Phongthawee, Jiraphon Yenphraphai, and Supasorn Suwajanakorn. Nex: Real-time view synthesis with neural basis expansion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8534–8543, 2021. 3

[32] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Pointnerf: Point-based neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5438–5448, 2022. 3

[33] Tianxiang Ye, Qi Wu, Junyuan Deng, Guoqing Liu, Liu Liu, Songpengcheng Xia, Liang Pang, Wenxian Yu, and Ling Pei. Thermal-nerf: Neural radiance fields from an infrared camera. *arXiv preprint arXiv:2403.10340*, 2024. 2, 3

[34] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5

[35] Haoyi Zhu. X-nerf: Explicit neural radiance field for multi-scene 360deg insufficient rgb-d views. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5766–5775, 2023. 2

[36] Haidong Zhu, Tianyu Ding, Tianyi Chen, Ilya Zharkov, Ram Nevatia, and Luming Liang. Caesarnerf: Calibrated semantic representation for few-shot generalizable neural rendering. *arXiv preprint arXiv:2311.15510*, 2023. 2