

Inverse Image-Based Rendering for Light Field Generation from Single Images

Hyunjun Jung
GIST AI Graduated School
hyunjun.jung@gm.gist.ac.kr

Hae-Gon Jeon
Yonsei University
earboll@yonsei.ac.kr

Abstract

A concept of light-fields computed from multiple view images on regular grids has proven its benefit for scene representations, and supported realistic renderings of novel views and photographic effects such as refocusing and shallow depth of field. In spite of its effectiveness of light flow computations, obtaining light fields requires either computational costs or specialized devices like a bulky camera setup and a specialized microlens array. In an effort to broaden its benefit and applicability, in this paper, we propose a novel view synthesis method for light field generation from only single images, named inverse image-based rendering. Unlike previous attempts to implicitly rebuild 3D geometry or to explicitly represent objective scenes, our method reconstructs light flows in a space from image pixels, which behaves in the opposite way to image-based rendering. To accomplish this, we design a neural rendering pipeline to render a target ray in an arbitrary viewpoint. Our neural renderer first stores the light flow of source rays from the input image, then computes the relationships among them through cross-attention, and finally predicts the color of the target ray based on these relationships. After the rendering pipeline generates the first novel view from a single input image, the generated out-of-view contents are updated to the set of source rays. This procedure is iteratively performed while ensuring the consistent generation of occluded contents. We demonstrate that our inverse image-based rendering works well with various challenging datasets without any retraining or finetuning after once trained on synthetic dataset, and outperforms relevant state-of-the-art novel view synthesis methods.

1. Introduction

The selective control of focus and shallow depth of field (DoF) have been critical tools of photography. Unfortunately, modern devices such as cell phones have struggled to reproduce these effects because of their small sensors and lenses. As a solution to this issue, a concept of 4D light

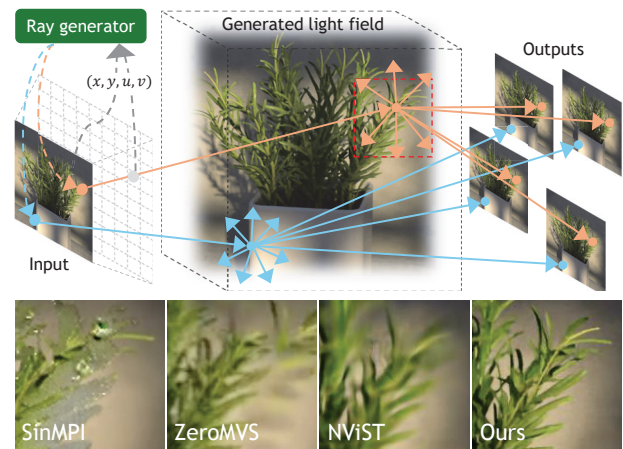


Figure 1. iIBR enables to generate high fidelity and consistent 4D light field from single image

fields [1, 33, 34], taking colors and directions of the light flow in a space, enables to render novel views and photographic effects such as refocusing and shallow depth of field. However, capturing real light fields requires specialized cameras, and suffers from an inherent trade-offs between spatial and angular resolutions of captured images because one sensor should take both of them. The inherent trade-off potentially causes aliasing when we implement the photographic effects from fewer angular resolutions. Light field angular super-resolutions [9, 18, 19, 23, 58] have been proposed to mitigate this trade-off, but still need geometrically well-aligned multiple images as input.

Recent advancements in learning-based methods for novel view synthesis allow us to synthesize angular contents of light field. Techniques like NeRF [31] and 3D Gaussian Splatting [20] facilitate the transformation of photographs of real-world scenes into 3D models by optimizing the underlying geometry and visual properties. However, producing highly detailed scenes is still a demanding task that requires capturing a large number of images. Its inadequate observations can result in models with incorrect geometry and appearance, leading to unrealistic renderings from novel viewpoints.

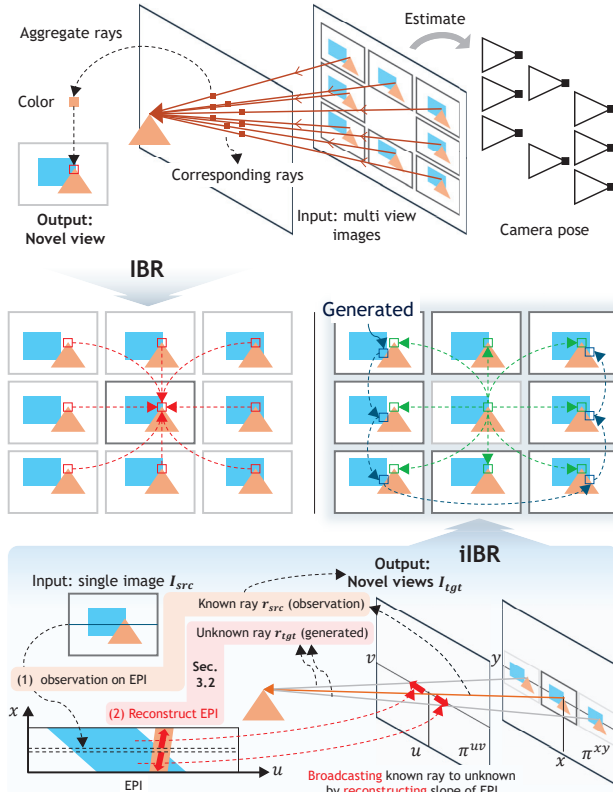


Figure 2. **Difference between IBR and iIBR.** IBR renders target image from multiple source image by aggregating correspondence colors. On the contrary, iIBR renders multiple target images from single image by reconstructing unknown light flows, which is shown as lines in the EPI. EPI is 2D slice of 4D light field. A straight line in EPI represents set of correspondence’s light flows sampled from different angular resolution.

Fortunately, with the recent development of an end-to-end depth-aware view synthesis [27], neural rendering [62], scene approximation into multiple depth planes [11, 25, 53] and image generation [60], we can synthesize images with novel viewpoints from single images to reduce the dependency on dense multi-view captures. In spite of this, they have their own limitations related to quality, efficiency and generality. One of common issues on rendering quality often stems from misaligned geometry and correspondences, particularly in novel view generation approaches. These misalignments in unseen contents of the target viewpoint frequently produce blurry artifacts on 3D objects.

In this paper, we focus on correspondence alignment among generated novel views for better photographic reproduction of scenes. Misaligned pixels along an angular axis of the light field can lead to unwanted artifacts when rendering photographic effects. Our key idea comes from a concept of epipolar plane images (EPI) [4] from two-plane parameterized light field [24]. EPI is 2D slices of constant angular and spatial directions in a 4D space. It can be viewed as a 2D image, with spatial resolution along a

horizontal axis and angular resolution along a vertical axis. In an EPI, line structures are visible, and their slopes vary based on the disparity among sub-aperture images, whose example is described in Fig. 2. Pixels along a slope are correspondences between sub-aperture images placed in either one column or row on a regular grid. Therefore, each line in the EPI represents a set of light flows of the rays cast from corresponding pixels. The bottom-sided illustration in Fig. 2 shows that this EPI’s property can be leveraged to generate unknown rays of correspondences from known ray of input image.

To do this, we formulate light field generation from single images as an inverse problem of image-based rendering which typically synthesize a single image by blending colors from multiple correspondences across different views. We propose **iIBRnet**, *inverse Image-Based Rendering* network, a neural renderer that takes single images as input to reconstruct continuous signals of light flows in space, which is the ultimate goal of classical 4D light field imaging, to generate novel views. The concepts of IBR and iIBR are depicted in Fig. 2. To render the light flows to novel views, we utilize the Transformer [54] to compute self-attention scores of ray embeddings from the input image and the target novel viewpoint. This procedure reconstructs angular-consistent, high-fidelity light field images. Additionally, we improve the generality of iIBRnet through pixel-level processing for novel view synthesis. Our model is trained on only synthetic images and tested on real-world images without any re-training or fine-tuning. We demonstrate that our method produces state-of-the-art results in light field generation compared to relevant works, showcasing notable generalization performance.

2. Related Works

2.1. Image-based rendering

Image-based rendering (IBR) [47] has emerged with desire on making free-viewpoint images, given multiple images. It enables the synthesis of novel views from collection of input images. The light field [24] allows us to parameterize incoming light flows from world coordinates to describe scene structures, which are formulated as 4D plenoptic function [16]. However, light field rendering requires a dense sampling of input views to yield high quality images. To mitigate the dense sampling constraint, Lumigraph [10] uses an approximate geometry. Co-operating with explicit geometry [8, 41] shows plausible rendering quality with few image samples. However, learning 3D proxy geometry is challenging, and errors in this process can result in misaligned correspondences during rendering.

Recently, the concept of IBR has contributed to learning radiance fields by aggregating corresponding visual features. Methods for finding correspondences can be categorized as: aligning along an epipolar line and aggregating their colors while leveraging the capabilities of transformers [5, 51]; collecting visual features from input images dur-

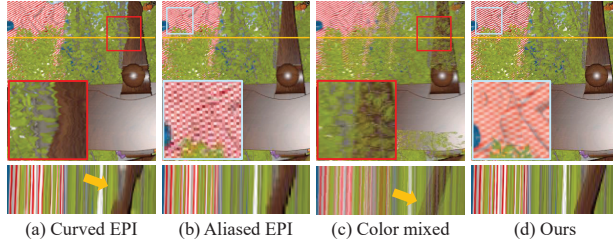


Figure 3. **Importance of accurate EPI generation.** Restoring accurate 2D rays along an EPI in sub-pixel accuracy is importance for novel view synthesis.

ing volumetric sampling [55, 64]; and using plane sweep volumes [6]. They achieve generality, enabling the representation of scenes with only a forward-pass. However, they requires multiple input images and precise camera poses, while our method requires only single images.

2.2. Novel view synthesis from single images

Novel view synthesis from single images is challenging due to a ill-posed nature on representing scene geometry. Previous methods obtain geometric information by using either single-image depth estimation [27, 49] or mesh estimation [15]. With the estimated depth information, works in [11, 39] compute multi-plane images (MPIs) to approximately account for scene geometry, which can be projected onto novel viewpoints. 3D photo conversions from single images [46] separate foreground and background of scenes, using soft occlusion masks to fill missing regions with plausible contents through inpainting. SinNeRF [62] and NViST [17] infer radiance fields for scenes from only single images. However, these methods heavily rely on large-scale datasets, as their networks focus on leveraging useful intuitions about scene information through the visual feature extraction. Our method overcomes this issue by treating pixels as individual light flows.

Given that text-to-image generation models are highly effective at producing visually promising images, methods for obtaining multi-view observations from single images have been proposed [44, 52, 60]. These models offer stronger priors for unseen contexts of scenes using the input images with pose conditioning. However, because they generate views independently, remaining uncertainties among correspondences can lead to performance drops. Multidiff [32] generates novel views by warping true colors from the input image using its corresponding depth map and fills in empty spaces through generation. While the warped true-color pixels are geometrically consistent, the generated regions are not. Our method is free from this problem because we iteratively update inpainted contents at other novel views, ensuring consistent view generation.

3. Methodology

Given a single image, our goal is to reconstruct a 4D light field using our iIBR which is implemented as a neural ren-

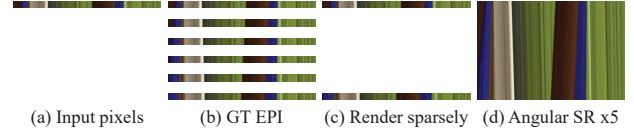


Figure 4. **2D ray generation of iIBR.** iIBR restores continuous 2D rays from single pixels, as demonstrated by the angular super-resolution of light field images.

dering process. Fig. 5 provides an overview of our neural renderer, iIBRnet. We first define a concept of iIBR and provide technical insights how to incorporate it into a neural rendering network. We then describe an architecture and a rendering pipeline, including occlusion detection and handling. To better explain our method, we start with a 2D case of iIBR, involving 1D spatial and 1D angular dimensions, and then extend it into the 4D light field representation.

3.1. Inverse image-based rendering

An inverse rendering [29, 45] typically refers to a process of reversing physically-based rendering, aiming to estimate physical attributes of a scene—such as geometry, material properties and lighting—from images. The concept of iIBR starts from an imagination of an ideal IBR. As illustrated in the upper-sided Fig. 2, the ideal IBR would be possible if exact correspondences are available. Since the ideal IBR is theoretically achievable, its inverse problem allows us to propagate colors from the input image to other sub-aperture images at precise locations of each correspondence.

Physically, each pixel in the input images represents a ray of light flows carrying the pixel color. A pixel can be cast into a structured 4D ray space defined by two planes π^{xy} and π^{uv} , where the local plane coordinates at the intersections are $(x, y) \in \pi^{xy}$ and $(u, v) \in \pi^{uv}$. Consequently, the set of correspondences for the ray (x, y, u, v) from horizontally aligned sub-aperture images can be defined as $S = \{(x_i, y, u_i, v) \mid x_i \neq x, u_i \neq u, i \in I\}$, where I is an index set. This point of view simplifies the problem of finding correspondences by reducing it to a task of calculating pairs (x_i, u_i) , which functionally serves to obtain the set S . Here, one of our significant contribution is that EPIs are used as a powerful tool for solving this problem because they are constructed along two axes, x and u ; i.e., constructing accurate EPIs directly addresses the challenge of calculating the set S . Generating EPIs from a single angular content via 2D image processing is highly ill-posed, but representing EPIs by the slopes of pixels could provide a viable solution. The approach involves casting the ray of each input pixel into either $x - u$ or $y - v$ 2D space with the orientation defined by the pixel’s slope.

3.2. 2D inverse Image-Based Rendering

For easy-to-understand our 2D iIBR, we first consider 2D ray space. The sub-problem of synthesizing a 4D light field via iIBR can be represented in the reduced dimensionality of the ray space. This is achieved by selecting one spatial and one angular domain, thereby considering the 2D ray

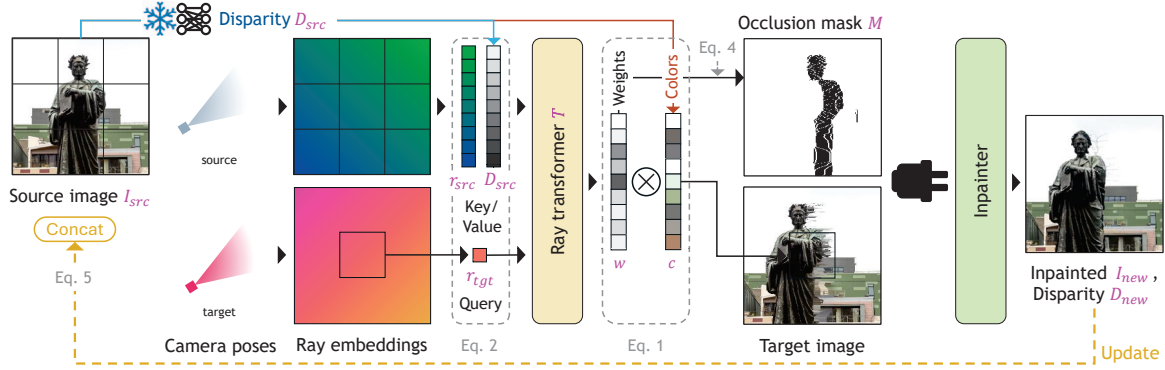


Figure 5. **An overview of iIBRnet.** Given a single image, iIBRnet generates novel views. iIBRnet operates in two key stages: (1) the ray transformer, which calculates the relationships between rays from different viewpoints, acting as the mechanism for rendering the light flows of cast rays from input image, and (2) an inpainting process that handles occlusions and updates the input for iIBRnet.

space, without a loss of generality. Our goal is to calculate a set S by casting the ray of an input pixel into 2D ray space, represented as EPIs. Imagine a baby drawing straight lines from top to bottom on a blank page with crayons. These straight lines will form an EPI, with the colors of crayons matching the colors of the input pixels.

Drawing a line in the EPI precisely, i.e., casting a 2D ray in the 2D ray space, is critical for synthesizing accurate correspondences from the input image to sub-aperture images. Challenges arise when the drawn line is not straight (i.e. curved), or when the line needs to be drawn at sub-pixel coordinates. These lead to geometrically inconsistent results and aliasing, as illustrated in Fig. 3 (a) and (b), respectively.

To address this, our neural renderer, iIBRnet, is designed to render each angular content step-by-step, progressing from the input to the next angular content and so on. It renders a pixel by a weighted summation of the colors obtained through tracking the surrounding 2D rays. With this consideration, iIBRnet is capable of generating continuous 2D rays with sub-pixel accuracy, enabling anti-aliased rendering and unlimited angular super-resolution. It could also render views sparsely, where the step of angular content generation can be skipped, as demonstrated in Fig. 4.

Rendering a color of 2D ray c_i^j in the i -th spatial and j -th angular dimension is achieved by aggregating all spatial contents in the $(j-1)$ -th angular dimension, each associated with their respective weights w , as shown below:

$$c_i^j = \sum_{a \in A_{j-1}} w_a^{j-1} c_a^{j-1}, \quad (1)$$

where A_{j-1} is a set of spatial indices of pixels in the $(j-1)$ -th angular dimension. iIBRnet is designed to predict the weight w rather than directly predicting the color c_i^j .

In the context of aggregation, we use Transformer architecture [54] to predict w . Transformers are widely used in neural rendering [51, 55] due to their effectiveness in aggregating visual information. However, to be more precise, our iIBRnet focuses on investigating the relationships between virtually projected rays rather than feature aggregation. In

physical terms, predicting w involves establishing connections between each 2D ray in the set A_{j-1} and the 2D ray associated with c_i^j , and determining how closely they are related. Therefore, our ray transformer in iIBRnet uses only ray coordinates as inputs, without any visual feature. The ray coordinates are sampled from a camera of the input image (typically with the view matrix defined as an identity matrix) and from the camera positions of the novel views to be rendered.

The representation of rays is also essential to predict w . Since we assume a 4D ray space with a two-plane parameterized light flow, the ray coordinates are defined as a light slab [24], denoted by (x, y, u, v) . In order to generalize new scenes without relying on specific camera configurations, we choose to parameterize rays using Plücker coordinates which has been used to model a neural field [48]. Plücker coordinates represent a ray that originates from a point $o \in \mathbb{R}^3$ and casts in the direction $d \in \mathbb{R}^3$ as $r = (d, o \times d)$. This representation spans four degrees of freedom and two scale factors within six dimensions, allowing us to uniquely process and define rays.

Another benefit of leveraging angular information in an EPI is to provide geometric information. Its angle of each slope directly represent disparity, while the slopes are formed due to the uniform sampling of corresponding light flows. From a perspective of an inverse problem, geometric priors help guide the correct formation of EPI slopes and resolve the challenge of distinguishing foreground and background pixels in regions where pixel slopes intersect. To incorporate such geometric priors, we focus on the positional encoding of the ray transformer. In GPNR [51], a novel positional encoding is proposed for the Transformer architecture to retain the spatial position of visual information, epipolar geometry and relative camera positions. Similarly, we introduce a positional encoding to embed them on the matching direction of 2D rays. This gives us an insight how the ray transformer in iIBRnet functions: it learns to establish a set of potentially matched correspondences based on the 2D ray direction. We choose disparity information

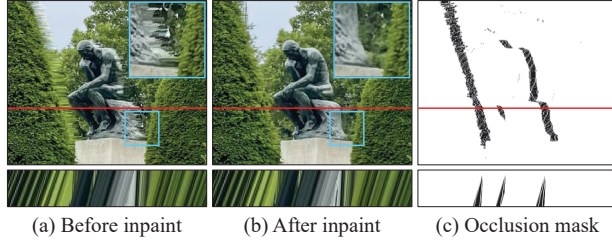


Figure 6. **Occlusion handling on 2D iIBR.** We can effectively capture occluded regions. The inpainted background is then regenerated to form 2D rays through the 2D iIBR.

for the positional encoding because disparity values have broader purposes than just representing scene depths. In the EPI, disparity values represent the displacement along the u -axis for each unit of movement along the x -axis, effectively capturing the 2D ray direction. The disparity is directly inferred from a depth foundation model $\mathcal{F}$ (we use DepthAnythingv2 [63] in this work), and is further refined using scale factors α and shift factors β predicted by a simple convolutional neural network. The final disparity value D is calculated as: $D_{src} = \alpha\mathcal{F}(I_{src}) + \beta$, where I_{src} is the input image. Note that any depth estimation model can easily be available in our framework, and recent metric depth estimation [38] may further minimize the need for scale/shift parameters. Finally, our w prediction using the ray transformer can be formulated as follows:

$$w^j = T(\{[r_{tgt,a}^j \parallel r_{src,a}^j \parallel D_{src,a}^j] \mid a \in A_j\}), \quad (2)$$

where T refers to the ray transformer, and r_x^j and d_x^j denote the 2D slices of r and d at the j -th angular dimension, respectively.

To optimize T , the objective function of iIBRnet consists of three loss terms: The first is L_2 loss on the rendered color, $\mathcal{L}_c = \|c - \hat{c}\|_2^2$ where $\hat{c}$ is its ground truth color. The second term is an entropy loss on w , $\mathcal{L}_w = -\sum w \log(w)$, ensuring that a dominant 2D ray contributes—though this may not exist in occluded regions—to rendering the target 2D ray. This helps prevent the issue on blending background and foreground contents, as illustrated in Fig. 3 (c). The last term is L_1 loss on the local structure tensor [57] of the rendered EPI ζ , $\mathcal{L}_{epi} = \|J(\zeta) - J(\hat{\zeta})\|$, where J and $\hat{\zeta}$ denote the structure tensor operator and ground truth EPI, respectively. This loss encourages the rendering of 2D rays to have straight linear structures on EPIs, and assists in predicting the scale/shift value of disparity by refining the local slopes of the rendered EPIs. In total, our objective function is formulated as follows:

$$\mathcal{L} = \lambda_c \mathcal{L}_c + \lambda_w \mathcal{L}_w + \lambda_{epi} \mathcal{L}_{epi}. \quad (3)$$

3.3. Occlusion handling

While our ray transformer is capable of accurately rendering 2D rays, challenges remain in restoring contents in occluded regions. To address this issue, we need to generate

unseen contexts in the occluded regions. Thanks to recent advances in recent generative models [42, 43, 65], inpainting occlusions, seamlessly matching the surrounding context, is feasible.

As the first step, we detect occlusions in the synthesized novel view. In Fig. 6, we present an example of an EPI. One thing to note is that in the initially rendered EPI from the ray transformer, occlusions appear as a blend of adjacent foreground and background contents. This implies that the weight w used to render pixels in occluded regions may not accurately target the dominant 2D rays corresponding to the occlusion contents. Therefore, we interpret w as an uncertainty in rendering and detect an occlusion mask M using the following formulation:

$$M_i^j = \begin{cases} 0 & \text{if } -\sum_{a \in A_{j-1}} w_a^{j-1} \log(w_a^{j-1}) < k \\ 1 & \text{else} \end{cases}, \quad (4)$$

where k is a threshold, and empirically set to 2.3. For occlusions, M is 0. We then restore the occluded content by inpainting the image from a novel viewpoint for the masked regions. However, since 2D iIBR renders only a single slice of a 2D image at a time, we render the complete 2D novel view image at first, and then inpaint the occlusions. We use Latent diffusion model [42] for inpainting.

Here, we aim to achieve consistent inpainting across all sub-aperture images. After the complete novel view is generated through inpainting, we treat each newly generated pixel as a new ray, allowing us to cast it into the 2D ray space using 2D iIBR. We first infer disparity as $D_{new} = \alpha\mathcal{F}(I_{new}) + \beta$, where I_{new} is the novel view. We then assign ray coordinates and the corresponding disparity to the generated pixels from the masked region, and incorporate them into the input set $\{r_{src,x}^j, D_{src,x}^j \mid x \in A_j\}$ for the ray transformer. This process is iteratively repeated whenever we encounter an occlusion that needs to be generated.

3.4. 4D light field generation

With the solution to the sub-problem of 2D iIBR, it is straightforward to extend this into (2+1)D iIBR, where one spatial dimension is expanded. The process involves simply repeating the 2D iIBR steps in the extended domain. For instance, if we apply iIBR into a 2D image along a single angular dimension, a viable solution would be to perform 2D iIBR on each horizontal slice of the input image, and then merge all the slices.

For a 4D ray space where includes an additional extension in the angular dimension, we opt to expand the positional encoding accordingly. Since the input ray coordinate is already defined in the 4D ray space, the ray transformer for 4D iIBR is formulated as follows:

$$w^{j_1, j_2} = T(\{[r_{tgt,a,b}^{j_1, j_2} \parallel r_{src,a,b}^{j_1, j_2} \parallel D_{x,src,a,b}^{j_1, j_2} \parallel D_{y,src,a,b}^{j_1, j_2}] \mid a \in A^{j_1}, b \in A^{j_2}\}), \quad (5)$$

where D_x and D_y denote the disparity between sub-aperture images along the horizontal and vertical angular

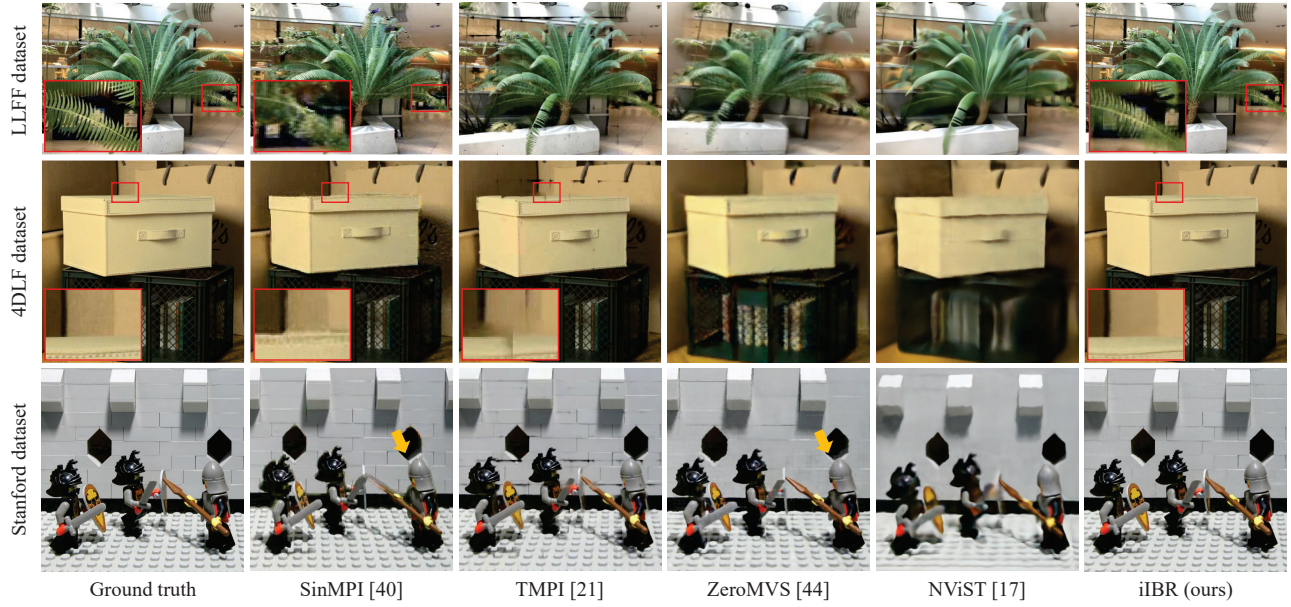


Figure 7. **Qualitative comparison.** Our iBR consistently synthesizes novel views, an archives highest quality.

axes, respectively. After rendering through the ray transformer, inpainting occlusions and iterative update for the set of the input 4D ray are also performed. To reduce the number of iterations and enhance the efficiency, we first generate the farthest views from the center view, and then update the generated occlusion rays.

4. Experiments

Further evaluations and analyses are provided in our **supplementary material**, which includes: (1) *Video results* demonstrating the qualitative results of our novel view synthesis across various scenes; (2) *BRDF rendering* details, explaining how our learned iBRnet is used for rendering specularities; and (3) additional experimental results and analysis on our rendering pipeline.

4.1. Implementation details

We implement our network using a public Pytorch [36] framework. For training, we use Adam [22] optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The learning rate and batch size are set to 0.0001 and 1, respectively. Our network is trained on a single NVIDIA Tesla V100 GPU for a day, 630K iterations. To avoid an overfitting problem, we adopt a data augmentation in [13]. The balance terms for the loss functions are set to $\lambda_c = 100$, $\lambda_w = 1$, and $\lambda_{cpi} = 0.1$. For memory-efficient training on GPU, we select five source ray coordinates closest to the target ray coordinate to learn T .

Dataset. We evaluate ours and state-of-the-art methods on four datasets: three light field datasets used for training and one real-world dataset to assess real-world performance.

(1) *Pov-ray dataset* [12]: Pov-ray dataset is a synthetic light field dataset with 11×11 sub-aperture images on a regular

grid. The dataset contains 900 scenes, and we divide them into 800 training scenes and 100 test scenes, following the authors’ split.

(2) *Stanford dataset* [50]. We use the (new) Stanford Light Field Archive dataset, which was captured in a camera array. This dataset has 17×17 sub-aperture images for each scene. The dataset is primarily used for evaluation, but we also report fine-tuned results on this dataset.

(3) *4D light field dataset (4DLF)* [14]: The synthetic dataset consists of 9×9 sub-aperture images. Similar with the Stanford dataset, this dataset is mainly utilized for evaluation and fine-tuned results.

(4) *NeRF LLFF dataset* [30]: This dataset provides unstructured multiview images captured in real-world scenarios. To evaluate our 4D iBR on this dataset, we first densely construct a structured 4D light field (128×128) from the view closest to a center of all cameras. We then render free views for evaluation by utilizing an approach from [16]. This dataset is mainly utilized for evaluation and fine-tuned results as well. To fine-tune the dataset to our method, we use a coarse 4D light field rendered through ZipNeRF [2].

Evaluation protocol. We compare our 4D iBR with reproducible state-of-the-art methods that are capable of synthesizing full scenes, not just object-centric scenes. For a fair comparison, all the compared methods, including ours, use the same resolution of input images and depth maps from DepthAnythingV2 [63]. Since SinMPI [39] utilizes an off-the-shelf latent diffusion model checkpoint [42], therefore we apply our inpainter fine-tuned with each dataset.

To evaluate performances, we use common quantitative measures of the image quality: peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM) [56] and learned perceptual image patch similarity (LPIPS) [66].

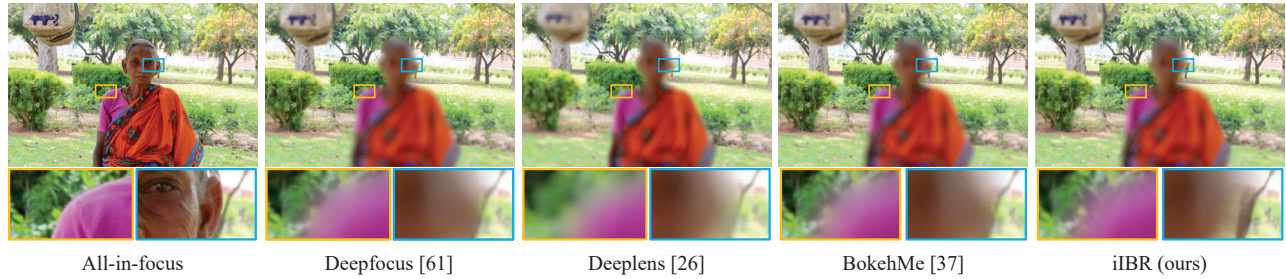


Figure 8. **Qualitative comparison for digital refocusing.** Since iBR generates 4D light flows in space and rendering light field image, physical and realistic digital refocusing is available.

	Pov-ray dataset			4DLF dataset			Stanford dataset			NeRF LLFF dataset		
(ft: fine-tuning)	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
SinMPI [39]	22.419	0.697	0.213	22.432	0.701	0.201	22.381	0.674	0.218	17.101	0.530	0.581
TMPI [21]	23.619	0.726	0.170	24.832	0.780	0.132	24.013	0.749	0.162	17.561	0.569	0.422
ZeroMVS [44]	20.835	0.710	0.229	20.759	0.689	0.231	20.481	0.664	0.228	11.211	0.403	0.623
NViST [17]	19.889	0.553	0.290	19.842	0.540	0.325	18.548	0.471	0.493	15.341	0.437	0.701
iBR (ours) (zero-shot)	28.407	0.931	0.037	28.095	0.926	0.046	27.889	0.910	0.068	24.910	0.810	0.183
iBR (ours) (ft)	28.407	0.931	0.037	28.382	0.929	0.041	27.682	0.902	0.052	25.534	0.883	0.095

Table 1. **Qualitative comparisons.** iBR outperforms all compared methods across datasets and metrics. We tested iBR in a zero-shot setting using the model pretrained on the Pov-ray dataset, as well as fine-tuned versions for each specific dataset.

Test set $\rightarrow$	Pov-ray			4DLF			Stanford		
Train set $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
Pov-rayz	28.407	0.931	0.037	28.095	0.926	0.046	27.889	0.910	0.068
4DLF	28.183	0.926	0.048	28.382	0.929	0.041	27.691	0.903	0.068
Stanford	27.301	0.901	0.069	27.163	0.892	0.069	27.682	0.902	0.052

Table 2. **Zero-shot cross validation.** iBR demonstrates consistent performance across various datasets, showing high generalizability for novel view synthesis.

	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
w/o disparity positional encoding	12.008	0.312	0.808
w/o inpainting	24.899	0.735	0.163
w/o entropy loss	25.210	0.831	0.139
w/o EPI structure tensor loss	26.910	0.882	0.116
replace depth model to MiDaS [40]	27.990	0.857	0.063
replace depth model to ZoeDepth [3]	28.310	0.901	0.043
replace depth model to UniDepth [38]	28.372	0.930	0.040
Ours	28.407	0.931	0.037

Table 3. **Ablation study.** An ablation study was conducted on the Pov-ray dataset, demonstrating that each key component of our method contributes to the performance of iBR.

4.2. Qualitative results

Novel view synthesis. In Fig. 7, we present a qualitative comparison for novel view synthesis. We render top-left sub-aperture image from bottom-left sub-aperture image, while a center view is used for NeRF LLFF dataset. Since our method is designed to accurately reconstruct correspondences for novel view synthesis, it consistently produces precise results, even in areas with fine structures. Although 4D iBR computes relationships among rays, its individual ray processing enables our model to handle any input image, regardless of scene structures or contexts. Additionally, The better performance of our method on the NeRF LLFF dataset comes from the capability of the dense light field prediction and outpainting. In contrast, the compar-

ison methods struggle to reconstruct fine details because they rely heavily on visual features. While SinMPI does not heavily depend on visual information, it exhibits an issue on depth scale misalignment when generating out-painted images, even when the depth estimator is fine-tuned for the scene. SinMPI often fails to produce accurate inpainted texture, given a lack of details in the occlusion mask.

Zero-shot result. Our iBR can synthesize novel views with any image and demonstrates consistent performance regardless of the scene context. In Fig. 9, we present zero-shot novel view synthesis results on mobile phone images captured directly by ourselves. We also show results using material-edited images generated with [7]. These results demonstrate iBR can synthesize novel views not only with various scene contents but also with different materials.

Digital refocusing. We introduce an interesting application of our 4D iBR to photographic effects. In Fig. 8, we show the digital refocusing results. We render the refocusing image after making dense light fields. Compared to the relevant works, Deepfocus [61], Deeplens [26] and BokehMe [37], our 4D iBR produces the realistic refocusing effect because the defocus blur is made from an integration of light rays over the lens aperture. The background content is also visible through the defocus blur in our result which enhances the realism of refocusing, as it is generated and cast as the 4D ray.

Simulating specularity. All novel views generated by iBRnet assumes a Lambertian surface. In Fig. 10, we report specularity simulation. Specular surfaces can be rendered through our iBRnet if the BRDF and light source is provided (or defined by users). The theoretical basis is as follows: First, specular surfaces appear as curved structures on the EPI as discussed in [30], indicating that BRDFs

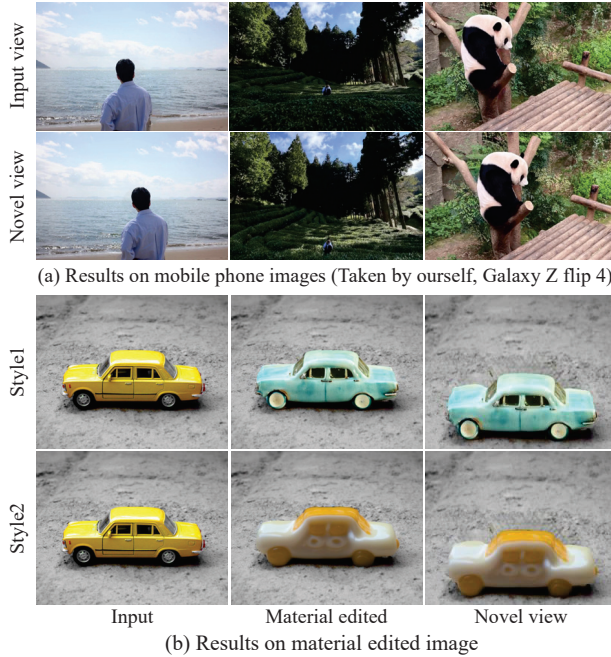


Figure 9. **Zero-shot novel view synthesis results.** Our iIBR can synthesize novel views with any image and demonstrates consistent performance regardless of the scene context.

can be synthesized on the EPI. Second, we refer to the bottom-sided Fig. 2 again. By distorting the position of corresponding spatial rays on the EPI (the two gray-colored arrows), view-dependent effects can be rendered based on the rendering equation Eq. (1), using the distorted weight from Eq. (5), which is obtained by inputting ray embeddings from different positions than their original locations.

4.3. Quantitative results

Novel view synthesis. We evaluate the performance of our model on novel view synthesis. For this evaluation, we use the sub-aperture image from the center view as input and assess the quality of synthesized views over all other sub-aperture images. The result in Table 1 demonstrates that our method outperforms the comparison methods. We note that the performance drop for all methods on the NeRF LLFF dataset is due to the large baselines between target images. This highlights that unstructured viewpoints and large baselines pose significant challenges to the comparison works.

Zero-shot cross validation. To demonstrate that our 4D iIBR generalizes well across different datasets, we report zero-shot cross-validation results in Table 2. Our 4D iIBR consistently performs well regardless of the training data, thanks to our pixel-wise processing nature. However, there is a slight performance drop on the Stanford dataset due to specularities. The specularities make the ray transformer hard to synthesize the corresponding ray coordinates, even though the disparity positional encoding provides ray directional information for light flows.

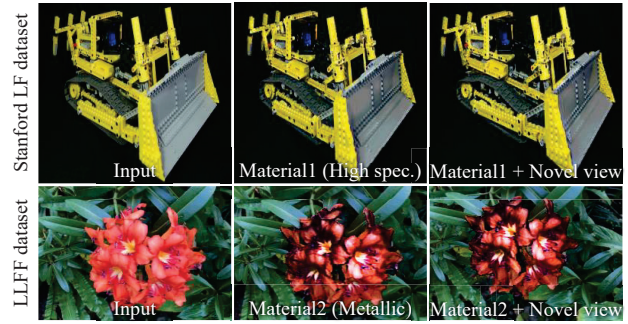


Figure 10. **Qualitative results for specularly simulation.** Our iIBR can synthesize novel views with specularly based on user-defined parameters.

Ablation study. To evaluate the impact of our contributions, we conduct a series of ablation experiments on the Pov-ray dataset in Table 3. Key components of our iIBR, such as disparity positional encoding and inpainting, are tested, along with the loss terms, including the entropy loss and the EPI structure tensor loss. The results confirm that each component contributes to achieving better performance. We also test several off-the-shelf depth foundation models in our pipeline. The result implies that our pipeline is effective with any model because all the models provide certain quality of output depth maps nowadays. Note that UniDepth [38] which offers approximate metric depth maps, is beneficial with respect to minimizing the need of learning scale/shift parameters.

5. Conclusion

We introduce a novel method to generate 4D light fields from single images, called inverse image-based rendering (iIBR). Through iIBR, we demonstrate that the inverse rendering of any light flow can be inferred from EPI pixel slope orientations. This insight allows us to generate and render continuous light flows of novel view images from single images. Additionally, we propose effective occlusion handling, enabling us to generate realistic rays in unseen areas. Through extensive evaluations, we show that our method outperforms recent novel view synthesis methods from single images and provides better generalization performance.

Limitation and future direction. Several directions exist for improving iIBR. One key challenge arises when moving objects are visible in scenes because of its temporal inconsistency of correspondences. Recent methods [28, 35, 59], incorporating deformation fields, can be a good solution to reconstruct light flows of moving or deforming subjects over time. Another limitation occurs when we attempt to render 360 degree images. This stems from the two-plane parameterization of light field photography, which causes the number of pixel colors that iIBRnet can reference to decrease as the significant viewpoint shifts away from the input image. Nevertheless, we believe that it is feasible by extending our approach to 360 geometry using a two-sphere parameterization, which is one of our future works.

Acknowledgment This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(RS-2024-00338439) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (RS-RS-2021-II212068, Artificial Intelligence Innovation Hub and RS-2025-25441838, Development of a human foundation model for human-centric universal artificial intelligence and training of personnel)

References

- [1] Edward H Adelson and John YA Wang. Single lens stereo with a plenoptic camera. *IEEE Trans. Pattern Anal. Mach. Intell.*, 14(2):99–106, 1992. [1](#)
- [2] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. *Int. Conf. Comput. Vis.*, 2023. [6](#)
- [3] Shariq Farooq Bhat, Reiner Birkel, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. [7](#)
- [4] Robert C Bolles, H Harlyn Baker, and David H Marimont. Epipolar-plane image analysis: An approach to determining structure from motion. *International journal of computer vision*, 1(1):7–55, 1987. [2](#)
- [5] David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 19457–19467, 2024. [2](#)
- [6] Yuedong Chen, Haoqi Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In *Eur. Conf. Comput. Vis.*, 2024. [3](#)
- [7] Ta-Ying Cheng, Prafull Sharma, Andrew Markham, Niki Trigoni, and Varun Jampani. Zest: Zero-shot material transfer from a single image. In *Eur. Conf. Comput. Vis.*, 2024. [7](#)
- [8] Inchang Choi, Orazio Gallo, Alejandro Troccoli, Min H Kim, and Jan Kautz. Extreme view synthesis. In *Int. Conf. Comput. Vis.*, 2019. [2](#)
- [9] Chen Gao, Youfang Lin, Song Chang, and Shuo Zhang. Spatial-angular multi-scale mechanism for light field spatial super-resolution. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. [1](#)
- [10] Steven J Gortler, Radek Grzeszczuk, Richard Szeliski, and Michael F Cohen. The lumigraph. In *SIGGRAPH*, 1996. [2](#)
- [11] Yuxuan Han, Ruicheng Wang, and Jiaolong Yang. Single-view view synthesis in the wild with learned adaptive multi-plane images. In *SIGGRAPH*, 2022. [2, 3](#)
- [12] Stefan Heber and Thomas Pock. Convolutional networks for shape from light field. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. [6](#)
- [13] Stefan Heber, Wei Yu, and Thomas Pock. Neural epi-volume networks for shape from light field. In *Int. Conf. Comput. Vis.*, 2017. [6](#)
- [14] Katrin Honauer, Ole Johannsen, Daniel Kondermann, and Bastian Goldluecke. A dataset and evaluation methodology for depth estimation on 4d light fields. In *Asian Conf. Comput. Vis.* Springer, 2016. [6](#)
- [15] Ronghang Hu, Nikhila Ravi, Alexander C. Berg, and Deepak Pathak. Worldsheet: Wrapping the world in a 3d sheet for view synthesis from a single image. In *Int. Conf. Comput. Vis.*, 2021. [3](#)
- [16] Aaron Isaksen, Leonard McMillan, and Steven J Gortler. Dynamically reparameterized light fields. In *SIGGRAPH*, pages 297–306, 2000. [2, 6](#)
- [17] Wonbong Jang and Lourdes Agapito. Nvst: In the wild new view synthesis from a single image with transformers. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10181–10193, 2024. [3, 7](#)
- [18] Jing Jin, Junhui Hou, Hui Yuan, and Sam Kwong. Learning light field angular super-resolution via a geometry-aware network. In *Assoc. Advancem. Artif. Intell.*, 2020. [1](#)
- [19] Jing Jin, Junhui Hou, Jie Chen, Huanqiang Zeng, Sam Kwong, and Jingyi Yu. Deep coarse-to-fine dense light field reconstruction with flexible sampling and geometry-aware fusion. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(04): 1819–1836, 2022. [1](#)
- [20] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. In *SIGGRAPH*, 2023. [1](#)
- [21] Numair Khan, Lei Xiao, and Douglas Lanman. Tiled multiplane images for practical 3d photography. In *Int. Conf. Comput. Vis.*, 2023. [7](#)
- [22] Diederick P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Int. Conf. Learn. Represent.*, 2015. [6](#)
- [23] Anat Levin and Fredo Durand. Linear view synthesis using a dimensionality gap light field prior. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2010. [1](#)
- [24] Marc Levoy and Pat Hanrahan. Light field rendering. In *SIGGRAPH*, 1996. [2, 4](#)
- [25] Qinbo Li and Nima Khademi Kalantari. Synthesizing light field from a single image with variable mpi and two network fusion. *ACM Trans. Graph.*, 39(6):229–1, 2020. [2](#)
- [26] Wang Lijun, Shen Xiaohui, Zhang Jianming, Wang Oliver, Lin Zhe, Hsieh Chih-Yao, Kong Sarah, and Lu Huchuan. Deeplens: Shallow depth of field from a single image. *ACM Trans. Graph.*, 37(6):6:1–6:11, 2018. [7](#)
- [27] Miaomiao Liu, Xuming He, and Mathieu Salzmann. Geometry-aware deep network for single-image novel view synthesis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4616–4624, 2018. [2, 3](#)
- [28] Zhicheng Lu, Xiang Guo, Le Hui, Tianrui Chen, Min Yang, Xiao Tang, Feng Zhu, and Yuchao Dai. 3d geometry-aware deformable gaussian splatting for dynamic view synthesis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024. [8](#)
- [29] Stephen Robert Marschner. *Inverse rendering for computer graphics*. Cornell University, 1998. [3](#)
- [30] 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. In *SIGGRAPH*, 2019. 6, 7
- [31] 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 *Eur. Conf. Comput. Vis.*, pages 405–421, 2020. 1
- [32] Norman Müller, Katja Schwarz, Barbara Rössle, Lorenzo Porzi, Samuel Rota Bulò, Matthias Nießner, and Peter Kotschieder. Multidiff: Consistent novel view synthesis from a single image. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10258–10268, 2024. 3
- [33] Ren Ng. Fourier slice photography. In *SIGGRAPH*, pages 735–744, 2005. 1
- [34] Ren Ng, Marc Levoy, Mathieu Brédif, Gene Duval, Mark Horowitz, and Pat Hanrahan. *Light field photography with a hand-held plenoptic camera*. PhD thesis, Stanford University, 2005. 1
- [35] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Int. Conf. Comput. Vis.*, 2021. 8
- [36] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 6
- [37] Juewen Peng, Zhiguo Cao, Xianrui Luo, Hao Lu, Ke Xian, and Jianming Zhang. Bokehme: When neural rendering meets classical rendering. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 16283–16292, 2022. 7
- [38] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10106–10116, 2024. 5, 7, 8
- [39] Guo Pu, Peng-Shuai Wang, and Zhouhui Lian. Sinmpi: Novel view synthesis from a single image with expanded multipane images. In *SIGGRAPH*, pages 1–10, 2023. 3, 6, 7
- [40] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(3):1623–1637, 2020. 7
- [41] Gernot Riegler and Vladlen Koltun. Stable view synthesis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12216–12225, 2021. 2
- [42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10684–10695, 2022. 5, 6
- [43] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *SIGGRAPH*, 2022. 5
- [44] Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, and Jiajun Wu. ZeroNVS: Zero-shot 360-degree view synthesis from a single real image. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 3, 7
- [45] Yoichi Sato, Mark D Wheeler, and Katsushi Ikeuchi. Object shape and reflectance modeling from observation. In *SIGGRAPH*, pages 379–387, 1997. 3
- [46] Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 3d photography using context-aware layered depth inpainting. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 3
- [47] Harry Shum and Sing Bing Kang. Review of image-based rendering techniques. In *Visual Communications and Image Processing 2000*, pages 2–13. SPIE, 2000. 2
- [48] Vincent Sitzmann, Semon Rezhchikov, Bill Freeman, Josh Tenenbaum, and Fredo Durand. Light field networks: Neural scene representations with single-evaluation rendering. In *Neural Inform. Process. Syst.*, 2021. 4
- [49] Pratul P Srinivasan, Tongzhou Wang, Ashwin Sreelal, Ravi Ramamoorthi, and Ren Ng. Learning to synthesize a 4d rgbd light field from a single image. In *Int. Conf. Comput. Vis.*, 2017. 3
- [50] Stanford CGLab. The (new) stanford light field archive. <http://lightfield.stanford.edu/lfs.html>, 2008. 6
- [51] Mohammed Suhail, Carlos Esteves, Leonid Sigal, and Ameesh Makadia. Generalizable patch-based neural rendering. In *Eur. Conf. Comput. Vis.*, pages 156–174. Springer, 2022. 2, 4
- [52] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *Eur. Conf. Comput. Vis.*, 2024. 3
- [53] Richard Tucker and Noah Snavely. Single-view view synthesis with multipane images. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 2
- [54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neural Inform. Process. Syst.*, 2017. 2, 4
- [55] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4690–4699, 2021. 3, 4
- [56] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4): 600–612, 2004. 6
- [57] Sven Wanner and Bastian Goldluecke. Globally consistent depth labeling of 4d light fields. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2012. 5

- [58] Gaochang Wu, Yebin Liu, Lu Fang, and Tianyou Chai. Re-visiting light field rendering with deep anti-aliasing neural network. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(9): 5430–5444, 2021. [1](#)
- [59] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024. [8](#)
- [60] Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P Srinivasan, Dor Verbin, Jonathan T Barron, Ben Poole, et al. Reconfusion: 3d reconstruction with diffusion priors. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 21551–21561, 2024. [2](#), [3](#)
- [61] Lei Xiao, Anton Kaplanyan, Alexander Fix, Matt Chapman, and Douglas Lanman. Deepfocus: Learned image synthesis for computational display. In *SIGGRAPH*, 2018. [7](#)
- [62] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Humphrey Shi, and Zhangyang Wang. Sinnerf: Training neural radiance fields on complex scenes from a single image. In *Eur. Conf. Comput. Vis.*, 2022. [2](#), [3](#)
- [63] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv:2406.09414*, 2024. [5](#), [6](#)
- [64] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. [3](#)
- [65] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Int. Conf. Comput. Vis.*, 2019. [5](#)
- [66] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. [6](#)