Large-Scale Light Field Synthesis from Videos Enables Geometrically Consistent Bokeh Editing

Haoming Cai[1†‡] Zhoutong Zhang[2†] Christopher Metzler[1] Shumian Xin[2]

[1] University of Maryland
[2] Adobe

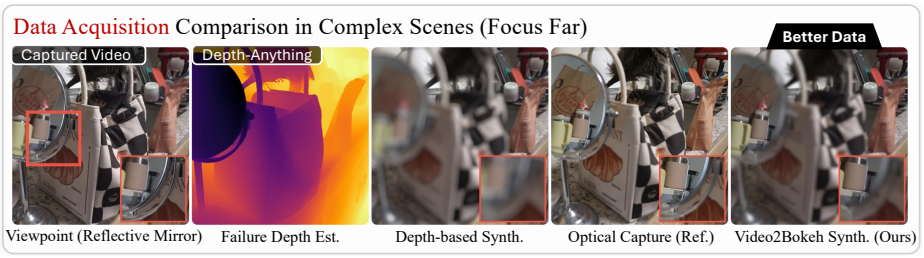
Abstract. Focus and depth of field (DoF) define where attention falls and how much of a scene appears sharp. Adjusting both after capture provides creators two-dimensional control over visual attention. However, existing focus and DoF editing tools fail when applied to scenes with complex geometries, such as reflections and fine features. This failure is a result of fundamental limitations in existing training data generation pipelines: current pipelines either sacrifice geometric fidelity for scalability, or sacrifice scalability for geometric fidelity. To address this, we present Vid2Bokeh, a scalable bokeh data acquisition pipeline that reconstructs dense 25×25 light fields from casual videos via feed-forward 3D reconstruction, bypassing depth estimation entirely and avoiding the geometric errors it introduces. The resulting Vid2Bokeh Dataset comprises over 100K light fields on diverse real-world scenes, simultaneously achieving scene diversity, optical density, and geometric fidelity that no existing bokeh dataset or bokeh data acquisition pipeline provides. By training on this geometrically faithful bokeh data, we introduce a diffusion model for full bidirectional focus–DoF editing that outperforms depth-based baselines on complex geometries. This result demonstrates that geometric fidelity in large-scale training data is the key to geometrically consistent bokeh editing.

Keywords: Bokeh Editing · Light Field Synthesis · Geometric Fidelity

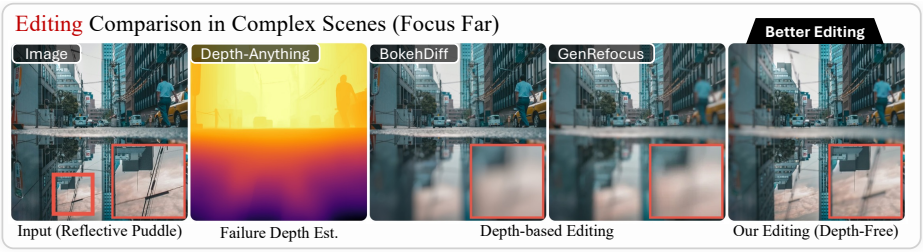
1 Introduction

Focus and depth of field (DoF) are the primary tools of visual storytelling, determining where attention falls and how much of a scene appears sharp. In post-capture bokeh editing [4, 6, 15, 22, 24, 25, 40, 42, 48, 51, 60, 61, 69, 75, 78, 88, 89, 94], users increasingly demand a "full-parameter" mechanism: the ability to simultaneously shift focus and adjust DoF, regardless of whether the input is an all-in-focus image or one containing natural bokeh. However, achieving such versatility on complex geometries remains an open challenge. As illustrated in Fig. 1b, current state-of-the-art tools frequently introduce apparent artifacts around intricate boundaries and reflections, where depth estimation struggles.

[†] Equal contribution.
[‡] Work done during an internship at Adobe.
The project page is available at `https://www.hm-cai.com/vid2bokeh/`.



(a) Vid2Bokeh: Geometrically Faithful Bokeh Data Acquisition. Depth estimation incorrectly maps the mirror reflection as a near plane, causing depth-based bokeh synthesis to produce geometrically incorrect bokeh. This illustrates a fundamental issue: any depth-dependent bokeh data generation philosophy risks introducing inaccurate geometric information as supervision. In contrast, Vid2Bokeh correctly renders bokeh on such complex geometries. The synthesized training data matches the physical optical reference—captured at the same viewpoint with a wide aperture—confirming that far-field reflections are correctly rendered as sharp.



(b) Geometrically Consistent Bokeh Editing. Training on the Vid2Bokeh Dataset enables geometrically consistent bokeh editing on complex geometries, where depth-based baselines introduce severe artifacts due to inaccurate depth estimation.

Fig. 1: Vid2Bokeh acquires geometrically faithful bokeh training data at scale, unlocking geometrically consistent editing on complex geometries.

These consistent failures stem from deficiencies in existing training data [25, 32, 48, 55, 60, 94] . As summarized in Table 1, existing data pipelines suffer from a fundamental limitation: they either sacrifice geometric fidelity for scalability, or sacrifice scalability for geometric fidelity—no existing approach achieves both simultaneously. Physically captured data is geometrically accurate but inherently non-scalable. Synthesis-based pipelines are scalable but invariably rely on depth estimation [85], either directly or indirectly—propagating geometric errors into training data and causing models to consistently fail on complex geometries.

To address this, we propose **Vid2Bokeh**, a scalable bokeh data acquisition pipeline that is simultaneously depth-estimation-free and scalable. At its core, Vid2Bokeh reconstructs dense light fields [7, 36, 44] from casual videos via feed-forward 3D reconstruction [13, 20, 37, 87, 90], bypassing depth estimation entirely. Our key observation is that feed-forward 3D reconstruction has reached a critical performance threshold—the balance between speed and fidelity now enables rapid conversion of in-the-wild videos into dense light fields. For bokeh synthesis, light field representation is inherently more robust than 2.5D depth maps, naturally resolving reflections and intricate occlusions to ensure geometrically consistent blur (Fig. 1a). Moreover, since any casually captured video of a

static scene serves as input, Vid2Bokeh is agnostic to capture setup or data format—making large-scale data acquisition both practical and accessible, whether from internet videos or custom captures.

The result is the **Vid2Bokeh Dataset**, which simultaneously achieves the three dimensions critical for focus–DoF editing: scene diversity, optical density, and geometric fidelity. As shown in Table 1, no existing dataset achieves all three—Vid2Bokeh is the first to do so, providing the geometrically faithful supervision necessary for training bokeh editing models that handle complex geometries.

Leveraging this dataset, we train a diffusion model for full bidirectional control of focus and DoF. Our experiments demonstrate that geometric fidelity in training data is the key enabler of geometrically consistent editing—our model remains robust on intricate geometries (e.g., complex building structures, tennis nets) where current bokeh editing methods consistently fail, achieving bokeh quality that significantly surpasses current state-of-the-art results.

In summary, our contributions are:

- **Vid2Bokeh Pipeline:** A video-to-bokeh data acquisition pipeline that is simultaneously depth-estimation-free and scalable, enabling large-scale generation of geometrically faithful bokeh data for training.
- **The Vid2Bokeh Dataset:** A large-scale light field dataset of 100k diverse scenes with dense 25×25 angular resolution, providing high geometric fidelity for bokeh editing.
- **Robust Bidirectional Editing:** We show that training on Vid2Bokeh enables geometrically consistent, bidirectional focus–DoF control on complex geometries.

2 Related Work

Focus and DoF Editing. Early bokeh-rendering methods simulate lens blur via optical models or layer-based compositing [8,29,34,35,53,72], suffering from step-like transitions, boundary halos, and occlusion failures. Learning-based approaches [14,25,39,48,49,61,75,91] improve realism via end-to-end or hybrid neural–optical rendering, yet still depend on accurate depth and struggle with fine structures. Defocus-blur removal [1–3,32,33,47,55,56,65,84] targets all-in-focus reconstruction, while refocusing remains underexplored [4, 51, 69, 78]. Training data for these tasks is equally constrained: real-captured datasets [2, 25, 33, 60] provide physically accurate blur but scale poorly, whereas synthetic alternatives introduce geometric or photometric errors—depth-based rendering [48,61,75,80] fails near depth boundaries, layer-based compositing produces halos, and physically based renderers suffer domain gaps from lacking real-world textures. In contrast, our pipeline synthesizes light fields from real-world video, inherently modeling occlusions and see-through structures that single-image pipelines cannot reproduce.

Diffusion Models for Bokeh Editing. Diffusion models [19, 23, 52, 54, 57, 64, 66,88] have advanced spatially adaptive digital media manipulation [5,6,9–12,15,

Method	Editing Space		Proposed Training Dataset		Training Data Generation Features		
	Focus (↕)	DoF (↕)	Name	Scenes	Core Philosophy	Depth-Est-Free	Scalable Data
DC^2 [4]	✓	✓	-	100	Physical Capture (Multi-Cam)		
Tedla et al. [69]	✓		Focal Stack Dataset	1,637	Physical Capture (Multi-Cam)	✓	
DPDNet [2]		↑	DPDD Dataset	500	Physical Capture (Multi-Cam)	✓	
AiFNet [55]		↑	LFDOF	100	Physical Capture (Light Field)	✓	
Bokehlicious [60]		↓	RealBokeh	4.4k	Physical Capture	✓	
RVR-LAAF [91]	✓	↓	Aperture Dataset	2942	Physical Capture	✓	
PyNET [24]		↓	EBB!	5k	Physical Capture	✓	
BokehDiff [94]		↓	-	-	Compositional Synthesis	✓	
BokehMe [48]		↓	-	150	Compositional Synthesis	✓	
DiffCamera [80]	✓	✓	-	20k	Depth-based Hybrid Synthesis		✓
GenRefocus [43]	✓	✓	-	-	Depth-based Hybrid Synthesis		✓
Ours	✓	✓	**Vid2Bokeh**	**100k**	**Synthetic Light Field from Video**	✓	✓

Table 1: Comparison of data generation philosophy and editing capabilities. Depth-Est-Free indicates that the training data is prepared without any involvement of depth estimation, neither directly nor indirectly. This ensures geometric fidelity in complex scenes, as depth errors propagate into training data and cause systematic artifacts on complex geometries. As shown, existing methods face a fundamental trilemma: physically captured data is geometrically accurate but non-scalable; compositional synthesis avoids depth estimation but is limited to restricted scene diversity; depth-based synthesis is scalable but compromises geometric fidelity. Vid2Bokeh is the only approach that simultaneously achieves depth-estimation-free data generation and large-scale scalability, providing a geometrically faithful foundation for robust bokeh editing. ↑/↓ in Focus denote shifting the focal plane; ↑/↓ in DoF denote increasing/decreasing blur magnitude; ✓ in DoF denotes full bidirectional control supporting blurry image input.

16,22,31,40,42,45,58,71,73,76,77,82,83,86,92], making them a promising framework for focus–DoF editing. BokehDiff [94] synthesizes bokeh via one-step diffusion with layer-wise supervision; concurrently, DiffCamera [80] performs joint focus–DoF editing but relies on BokehMe [48]'s image-based rendering, which fails under complex occlusions.

Novel View Synthesis (NVS). Early light-field methods [7,17,21,26,28,36,44, 68, 79, 81] require dense sampling or specialized hardware. Classical multi-view and image-based rendering [18,30,50,59] enables sparse-input view interpolation, while pre-NeRF neural representations [38, 46, 62, 63, 67, 93] improve quality but remain resolution-limited. NeRF [41] and 3DGS [27] achieve high fidelity at the cost of per-scene optimization; recent feed-forward methods [13,20,74,87,90] offer a practical quality–efficiency trade-off. We adopt this paradigm to reconstruct dense light fields from real-world videos without per-scene optimization, enabling large-scale focus–DoF supervision.

3 Vid2Bokeh Pipeline

In this section, we present Vid2Bokeh, a scalable bokeh data acquisition pipeline with high geometric fidelity. Using feed-forward novel view synthesis, Vid2Bokeh generates bokeh training data from dense light fields reconstructed directly from real-world videos. These light fields enable continuous variation of focal-plane

position and aperture size, providing geometrically faithful supervision for joint focus–DoF editing on complex geometries.

3.1 Synthesize Light Field from Video

We construct light fields using the DL3DV-10K dataset [37], which contains real-world videos at 960p resolution with calibrated camera poses. As shown in Fig. 2a, we treat each clip as a short multi-view sequence. We randomly select several frames as reference views, and each selected pose defines the center of a light field. Around each center, we create a 25×25 grid of virtual cameras by perturbing the pose along horizontal and vertical axes while keeping intrinsics fixed. Using the LaCT feed-forward 3D model [90], we synthesize RGB images for all virtual viewpoints in a single pass, producing a dense light field $L(u, v, s, t, c)$, where (u, v) are angular coordinates on the aperture plane, (s, t) are spatial pixel coordinates, and c indexes color channels. The resulting light field contains 25×25 aligned views capturing small-baseline variations for the later bokeh rendering and supervision-pair generation.

However, a fixed baseline spacing cannot accommodate the wide variation in depth ranges across scenes. To ensure consistent parallax across scenes, we perform a scene-aware calibration that adjusts the spacing between neighboring sub-aperture cameras in a light field. Because depth ranges vary widely, a fixed baseline may create excessive disparities in near scenes, causing aliasing, or minimal disparities in distant scenes, weakening refocusing. We estimate optical flow between adjacent views using RAFT [70] and scale the baseline such that the mean flow magnitude falls within an empirically chosen range of one to three pixels. This adaptive calibration yields consistent, physically plausible parallax and ensures stable refocusing behavior across diverse scenes.

3.2 Synthesize Bokeh Image from Light Field

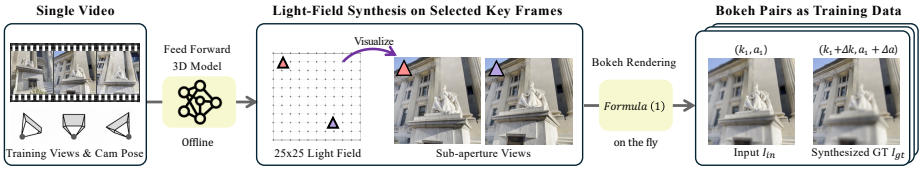
Given a reconstructed light field $L(u, v, s, t, c)$, we render bokeh images *on the fly* using the standard light-field refocusing formulation [36]. We parameterize focal-plane (FP) position by κ and aperture radius (AR) by a, which determines how much each sub-aperture view is shifted before integration. Refocusing with parameters(κ, a) is defined as

$$I(s, t, c;\, \kappa, a) = \iint_{\mathcal{A}(a)} L(u, v,\ s + \kappa u,\ t + \kappa v,\ c)\ \mathrm{d}u\, \mathrm{d}v, \tag{1}$$

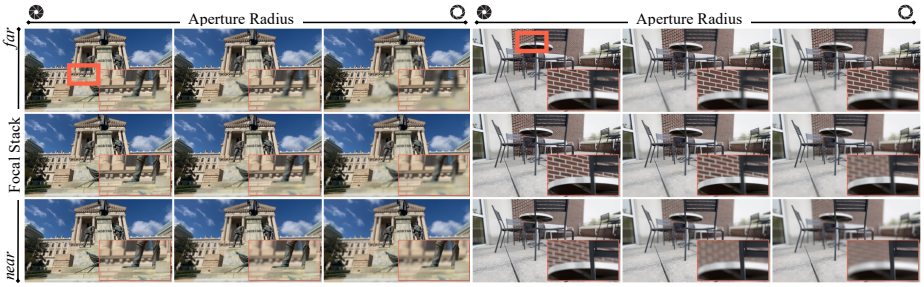
where $\mathcal{A}(a)$ is the circular integration region corresponding to a virtual aperture of radius a. Smaller apertures yield near–AiF images, while larger apertures produce stronger defocus blur because more viewpoints are averaged.

To build training supervision, we randomly sample two optical configurations (κ_1, a_1) and $(\kappa_2, a_2) = (\kappa_1 + \Delta\kappa; a_1 + \Delta a)$, and render both using Eq. (1):

$$I_{\text{in}} = I(s, t, c;\, \kappa_1, a_1), \qquad I_{\text{gt}} = I(s, t, c;\, \kappa_2, a_2). \tag{2}$$



(a) Vid2Bokeh data generation pipeline. Real-world videos are converted into dense light fields via feed-forward 3D reconstruction, enabling geometrically faithful bokeh rendering under arbitrary focus and aperture settings. Starting from a single video with estimated camera poses, we use a feed-forward novel view synthesis network [90] to synthesize a dense 25×25 light field around selected key frames, where each sub-aperture view corresponds to a slightly shifted viewpoint. During diffusion model training, we render bokeh images on the fly using the light-field refocusing formulation (Eq. 1), sampling arbitrary focal-plane (FP) and aperture radius (AR) configurations. This yields supervision pairs such as (κ_1, a_1) for the input and $(\kappa_1{+}\Delta\kappa,\ a_1{+}\Delta a)$ for the synthesized ground truth.



(b) Example data rendered from reconstructed light fields. Rows vary the focal plane to bring different scene depths into focus. Columns vary the aperture radius to control defocus strength.

Fig. 2: Overview of the Vid2Bokeh scalable bokeh data acquisition pipeline for geometrically faithful bokeh synthesis.

Each pair depicts the same scene under two distinct focus-aperture settings, matching the notation $(\mathrm{FP}_1, \mathrm{AR}_1)$ and $(\mathrm{FP}_1 + \Delta\kappa; \mathrm{AR}_1{+}\Delta a)$ used in Fig. 2a. Continuous sampling of (κ_1, a_1) and offsets $(\Delta\kappa, \Delta a)$ densely covers the joint focus–DoF space with physically consistent supervision.

Figure 2b shows representative examples: varying κ changes which depth layer comes into focus, while varying a controls the defocus blur magnitude. The rendered samples exhibit smooth, physically plausible transitions across both dimensions of focus and DoF.

3.3 The Vid2Bokeh Dataset

As established in Table 1, Vid2Bokeh is the only data acquisition pipeline that simultaneously achieves depth-estimation-free data generation and large-scale scalability. The Vid2Bokeh Dataset is the direct result of addressing both constraints, providing comprehensive coverage across three dimensions critical for focus–DoF editing: scene diversity, optical density, and geometric fidelity.

In terms of scene diversity, rooted in DL3DV, Vid2Bokeh dataset contains over 100K real-world light fields (scenes) spanning a wide range of environments, lighting conditions, and geometric complexity, enabled by the scalability of our pipeline. In terms of optical density, each scene supports continuous variation

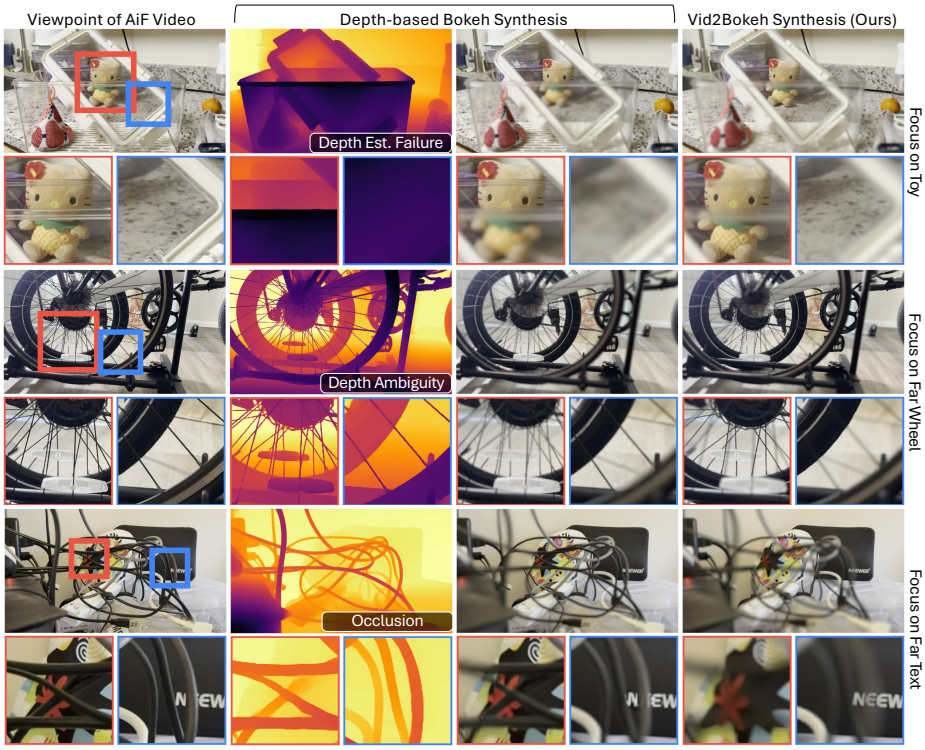


Fig. 3: Vid2Bokeh synthesizes geometrically faithful bokeh on complex geometries where depth estimation fails. (Top) Transparent container: depth estimation produces an incorrect depth map; Vid2Bokeh correctly captures the 3D structure behind transparent surfaces. (Middle) Bicycle wheels: depth estimation assigns similar depths to spokes at different depths; Vid2Bokeh's light field naturally resolves this ambiguity. (Bottom) See-through effect. Vid2Bokeh synthesizes physically correct see-through bokeh from video, capturing occlusion complexity that depth-based methods cannot reproduce at scale.

of aperture radius and focal-plane position, densely covering the entire joint focus–DoF space. Our (25×25) angular resolution further exceeds that of existing light field datasets, enabling finer aperture control and smoother defocus transitions. In terms of geometric fidelity, our light field representation is reconstructed directly from real-world videos, preserving physically accurate parallax and occlusion without any involvement of depth estimation—directly translating the Depth-Est-Free property of our pipeline into geometrically faithful training data. Furthermore, our (25×25) angular resolution significantly exceeds that of existing light field datasets, enabling finer aperture control and smoother defocus transitions.

Achieving this scale and diversity is made practical by the efficiency of the Vid2Bokeh pipeline. Using the feed-forward LaCT model [90], synthesizing a

Fig. 4: Validating data acquisition fidelity against Blender ground truth. For a Blender scene, we render a near all-in-focus video and process this video through the Vid2Bokeh pipeline. Column 1: reference viewpoint rendered in Blender (near AIF). Column 2: physically accurate bokeh rendered via Blender path tracing (ground truth). Column 3: bokeh synthesized by Vid2Bokeh from the rendered video. Across all scenes, Vid2Bokeh's synthesized bokeh closely matches the path-tracing ground truth, confirming the high fidelity of our data acquisition pipeline.

	Blender Direct Bokeh					Blender LF-based Bokeh				
Method	PSNR↑	SSIM↑	MS-SSIM↑	LPIPS↓	DISTS↓	PSNR↑	SSIM↑	MS-SSIM↑	LPIPS↓	DISTS↓
Blender LF-based (Upperbound)	31.03	0.956	0.978	0.069	0.054	∞	1.000	1.000	0.000	0.000
BokehDiff	28.70	0.900	0.948	0.141	0.066	30.72	0.889	0.954	0.216	0.101
Vid2Bokeh	29.71	0.929	0.971	0.108	0.071	33.37	0.931	0.984	0.147	0.091

Table 2: Quantitative validation of data acquisition fidelity using Blender-simulated scenes. We define two references: *Blender Direct Bokeh*, rendered via Blender path tracing as the physical ground truth, and *Blender LF-based Bokeh*, synthesized from an ideal 25×25 light field extracted directly from Blender, serving as the theoretical upper bound of light field representation for bokeh synthesis. Vid2Bokeh closely matches both references across all metrics, confirming the high fidelity of our data acquisition pipeline.

complete (25×25) light field takes under 30 seconds on a single NVIDIA A100 GPU, allowing the entire dataset to be generated on demand and stored within 10 TB. Vid2Bokeh thus provides a practical and scalable data acquisition pipeline for training models that require dense, geometrically faithful bokeh supervision.

Notably, unlike hardware-based pipelines that require DSLR cameras, tripods, pixel-wise alignment, and complex multi-view alignment, Vid2Bokeh requires only a casual handheld video, dramatically lowering the barrier for users to scale up their own bokeh training data.

4 Validation of Vid2Bokeh Data Acquisition

We validate the Vid2Bokeh pipeline across two complementary evaluations: geometric fidelity on real-world complex scenes, and quantitative accuracy against Blender path-tracing ground truth.

4.1 Data Acquisition Fidelity on Complex Geometries

Light field synthesis avoids the geometric approximations inherent in depth-based pipelines, making it naturally robust to scenes that challenge single-view

Fig. 5: Robust Bokeh Rendering Despite Reconstruction Artifacts via LF View Averaging. To stress-test this robustness, we deliberately feed LaCT sparse-view input to induce reconstruction errors (red arrows) in the synthesized LF views. Despite these per-view artifacts, the final rendered bokeh remains artifact-free, as errors across individual views cancel out during LF integration (Eq. 1).

depth estimation. To validate this intrinsic advantage, we compare bokeh data synthesized by the Vid2Bokeh pipeline against that synthesized by depth-based pipelines on identical real-world scenes (Fig. 3). As shown, depth estimation collapses on complex geometries such as transparent surfaces and intricate occlusions, introducing structural artifacts into the synthesized training data for editing model. By contrast, Vid2Bokeh leverages light fields reconstructed directly from casual videos, naturally resolving these challenging geometries and producing geometrically faithful training data where depth-dependent data acquisition systematically fail.

4.2 Evaluating Acquisition Accuracy in Blender

To assess the end-to-end fidelity of the Vid2Bokeh pipeline, we leverage Blender as a controlled environment for generating physically accurate ground truth. We define two references: *Blender Direct Bokeh*, rendered via native path tracing as the physical ground truth, and *Blender LF-based Bokeh*, synthesized from an ideal 25×25 light field extracted directly from the Blender scene, serving as the theoretical upper bound of our light field representation.

Our evaluation protocol simulates a real-world acquisition workflow by rendering a near all-in-focus video of the Blender scene and processing it through the Vid2Bokeh pipeline. We include BokehDiff as a depth-based baseline to contextualize our results. As shown in Fig. 4 and summarized in Table 2, where metrics are averaged over 100 diverse viewpoints, Vid2Bokeh closely matches both references across qualitative and quantitative metrics, consistently outperforming the depth-based baseline. These results confirm the high fidelity of our data acquisition pipeline, demonstrating its reliability for generating geometrically faithful bokeh data as supervision at scale.

4.3 Robustness to 3D Reconstruction Error

Although Vid2Bokeh relies on feed-forward 3D reconstruction, the final bokeh quality is robust to NVS reconstruction imperfections for two reasons. First, light

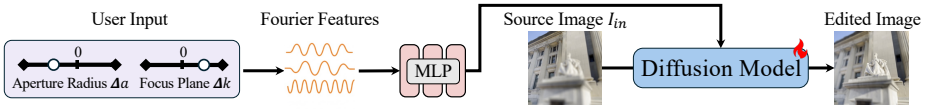


Fig. 6: Brief architecture of our editing framework. The network takes a source image I_{in} and target optical offsets $(\Delta a, \Delta \kappa)$ as input. These numerical offsets specify the desired change in focal plane position and aperture radius relative to the input image. Components marked with a snowflake symbol remain frozen during training

Method	Refocus from Near to Far				Refocus from Far to Near			
	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	PSNR↑	SSIM↑	LPIPS↓	DISTS↓
DRBNet + Dr.Bokeh	28.03	0.841	0.294	0.169	27.90	0.838	0.299	0.171
DRBNet + BokehMe	28.18	0.846	0.286	0.166	28.05	0.843	0.291	0.168
DRBNet + BokehDiff	28.82	0.834	0.232	0.132	28.67	0.861	0.238	0.145
DiffCamera	28.32	0.864	0.202	0.125	28.84	0.872	0.201	0.138
Ours	**29.62**	**0.907**	**0.198**	**0.108**	**29.51**	**0.893**	**0.183**	**0.121**

Table 3: Quantitative comparison of refocusing. We test refocusing in both forward and backward directions on the Vid2Bokeh test set. Metrics include PSNR↑, SSIM↑, LPIPS↓, and DISTS↓ averaged across all test scenes. Our model achieves the highest scores on all measures, outperforming all baseline methods.

field synthesis requires only narrow camera baselines and a small angular range. This is a significantly easier regime than general novel view synthesis, where multi-view inconsistencies rarely arise. Second, bokeh rendering integrates across all 25×25 sub-aperture views; residual per-view artifacts can be averaged out during integration, leaving the final rendered bokeh artifact-free. We stress-test this robustness by deliberately inducing reconstruction errors via sparse-view inputs (Fig. 5): despite visible artifacts in individual LF views, the rendered bokeh remains geometrically consistent.

5 Enable Geometrically Consistent Bokeh Editing

We evaluate our method across four tasks: refocusing, bokeh rendering, defocus-blur removal, and joint focus–DoF editing. Together, these experiments demonstrate that training on geometrically faithful Vid2Bokeh data enables robust, geometrically consistent focus–DoF bidirectional editing, including on images that already contain natural defocus. Implementation details, extended comparisons, ablation studies, and interactive html are provided in the supp. material.

5.1 Bokeh Editing Framework

Training. We fine-tune a diffusion model based on SDXL [52] for bidirectional focus–DoF editing using the Vid2Bokeh dataset. For each training iteration, we sample a light field from Vid2Bokeh and render a pair $(I_{\text{in}}, I_{\text{gt}})$ under two randomly sampled optical configurations (κ_1, a_1) and (κ_2, a_2) via Eq. (1), where the

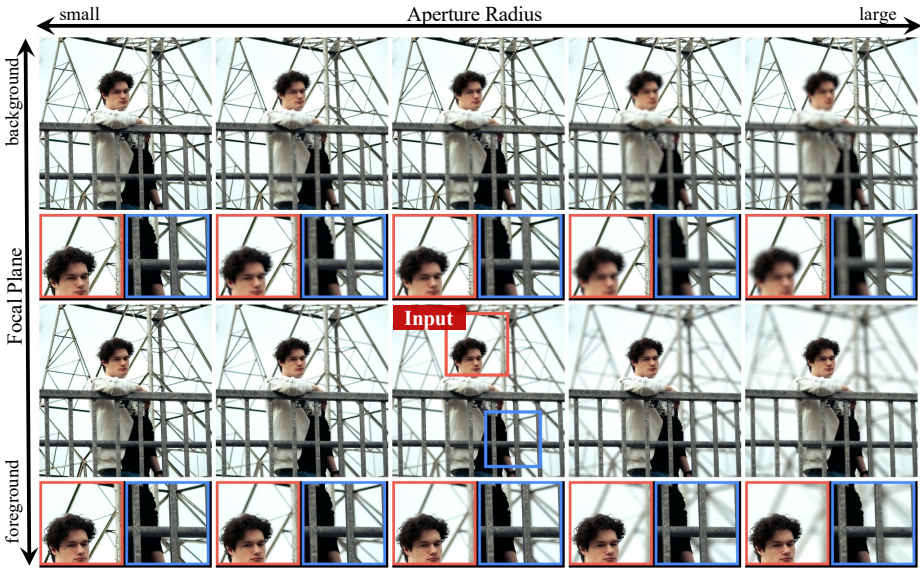


Fig. 7: Geometrically consistent joint focus–DoF editing. Starting from a single input image, our method supports bidirectional refocusing and continuous aperture control. Left: focused on the foreground subject with increasing aperture radius. Right: refocused on the background structure with increasing aperture radius. Our model consistently preserves fine geometric details.

offsets $(\Delta\kappa, \Delta a) = (\kappa_2 - \kappa_1,\, a_2 - a_1)$ define the desired focus–DoF transformation. The offsets $(\Delta\kappa, \Delta a)$ are encoded via Fourier features and a lightweight MLP into a conditioning vector involved in cross-attention. Both images are encoded into latents z^s and z^e via the frozen VAE encoder. Following a spatial-concatenation design inspired by InstructMove [12], standard diffusion training is performed on the concatenated latent. Architectural details are provided in the supp. material.

Inference. At test time, the user provides a source image I_{in} and desired bokeh settings offsets $(\Delta\kappa, \Delta a)$. The source image is encoded into z^s via the frozen VAE encoder, while a pure-noise latent is initialized for the target branch. The U-Net iteratively denoises the target branch conditioned on z^s and $(\Delta\kappa, \Delta a)$, producing the edited image I_{out} at the requested focus–DoF setting offsets.

Training Resources. We train for 100K iterations on eight NVIDIA A100 GPUs, which takes approximately four days. Full architectural specifications, loss functions, and hyperparameters are provided in the supp. file.

5.2 Refocusing

We evaluate refocusing on the Vid2Bokeh test set. To assess controllability, we fix the aperture radius and sweep the focus offset $\Delta\kappa$ in both directions, refocusing from near to far and vice versa. Baselines include DiffCamera [80], which

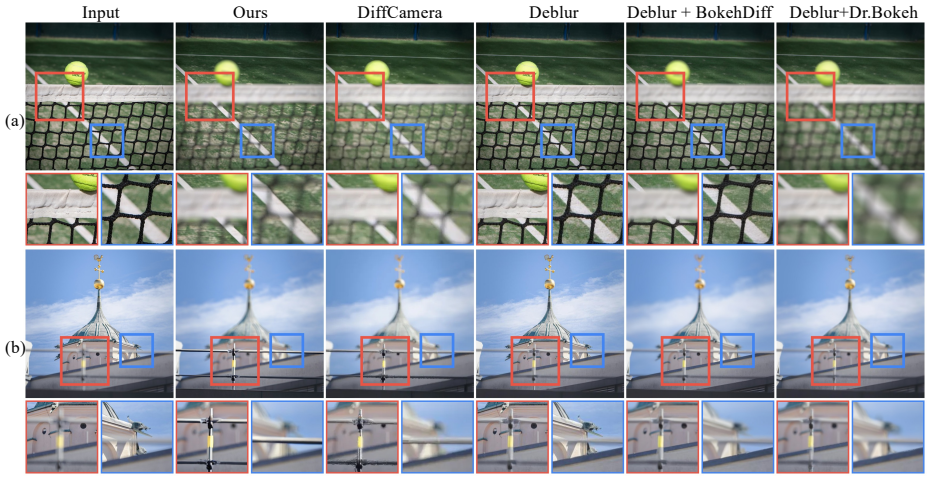


Fig. 8: Refocusing comparisons on in-the-wild images. Deblurring is performed using DRBNet [56]. Since most of bokeh rendering methods cannot create shallower DoF from an image that is already blurred, bokeh rendering baselines follow a deblur-then-refocus pipeline, which is why we include the intermediate deblurred result for reference. DiffCamera [80] supports bidirectional editing directly. (a) Tennis-net: the task is to shift focus from the foreground net to the background court. Our model achieves clean background focus and realistic see-through behavior, while all baselines fail even with a blur-free input: Dr.Bokeh [61] and DiffCamera [80] treat the net as a continuous surface and blur background lines that should remain visible, and BokehDiff [94] primarily defocuses the tennis ball. (b) Rod: the task is to shift focus from the background building to a thin foreground rod. Baselines exhibit depth errors around the slender structure, whereas ours restores sharp focus on the rod.

directly performs refocusing, and deblurring-then-refocusing pipelines combining DRBNet [55] with Dr.Bokeh [61], BokehMe [48], Bokehlicious [60], and BokehDiff [94]. Following the evaluation protocol of BokehDiff, we search for the best focal depth via binary search for baseline methods. Results are reported in Table 3, with qualitative comparisons on in-the-wild images in Fig. 8. Our method outperforms baselines quantitatively and qualitatively, and produce sharper and more realistic refocused images on complex geometries.

5.3 Joint Focus–DoF Editing

We demonstrate joint control of focus and DoF from a single input in Fig. 7. Unlike prior methods that require separate models or sequential inference for refocusing and DoF adjustment, our approach performs both within a unified diffusion framework. The model remains stable across large blur ranges, preserves identity and structure, and produces coherent 2D focus–DoF grids.

Method	EBB! [25]				Vid2Bokeh (Avg, 5 AR)			
	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	PSNR↑	SSIM↑	LPIPS↓	DISTS↓
BokehMe [48]	22.73	0.8146	0.264	0.140	26.91	0.833	0.261	0.194
Dr.Bokeh [61]	22.61	0.8117	0.314	0.189	27.10	0.838	0.269	0.180
Bokehlicious [60]	24.34	0.8071	0.182	0.098	27.42	0.842	0.198	0.152
BokehDiff [94]	**24.38**	0.8155	0.274	0.151	27.36	0.846	0.231	0.148
Ours	24.02	**0.8175**	**0.169**	**0.085**	**28.05**	**0.897**	**0.129**	**0.092**

Table 4: Quantitative comparisons of bokeh rendering. Rightmost columns report the mean metrics computed over five aperture-radius settings. All methods start from the same near–all-in-focus input and render bokeh at five progressively larger apertures, and the metrics are averaged across these five settings.

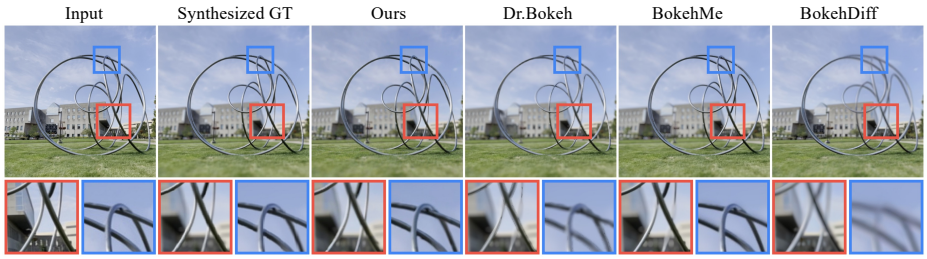


Fig. 9: Bokeh rendering on the Vid2Bokeh test set. We compare against the synthesized ground truth and depth-based baselines on a complex scene. Our method closely matches the GT, while baselines introduce artifacts at complex boundary.

5.4 Bokeh Rendering

We examine bokeh rendering by increasing aperture radius while keeping the focal plane fixed. Experiments are conducted on Vid2Bokeh test set and the EBB validation set [25]. On Vid2Bokeh, we start from near-AiF images and render a sequence of aperture settings. The EBB dataset has a different aperture profile with much stronger defocus blur than our training distribution. Following Bokehlicious [60], we perform a lightweight fine-tuning step on the EBB training set to align the bokeh profile with one training epoch. Quantitative results are shown in Table 4, and qualitative results on both in-the-wild and our test set images are shown in Fig. 10 and Fig. 9, respectively. Our model achieves the best scores in most metrics and produces more geometrically consistent bokeh than baselines.

5.5 Defocus Blur Removal

We evaluate defocus-blur removal by reconstructing an all-in-focus (AiF) image from a blurred input, with the target aperture set to near zero. Experiments are conducted on the Vid2Bokeh test set, DPDD [2], and real-world images. Our method achieves competitive performance against the deblurring networks DRBNet and Restormer, with results summarized in Table 5. Qualitative com-

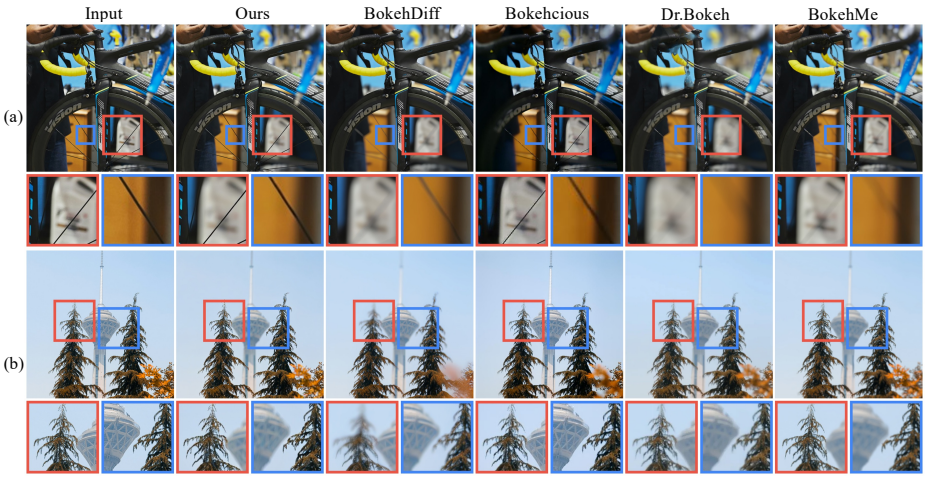


Fig. 10: Bokeh rendering on real images. (a) Bike spokes: the goal is to blur the background while preserving the thin wire structure. Our method produces geometry-aware defocus, whereas baselines fail due to inaccurate depth around the thin spokes. (b) Tree and tower: the goal is to blur the distant tower while preserving the delicate tree branches in the foreground. Our method accurately renders bokeh on fine structures, while baselines over-blur or distort the intricate branch boundaries.

parisons Fig. 11 on the real-world images further show that our model produces sharper and more structurally consistent results than the baselines.

6 Discussion

Vid2Bokeh Pipeline Scope and Extensibility. The current pipeline adopts a circular aperture model, as non-circular aperture shapes are primarily an aesthetic consideration and do not affect geometric fidelity, which is the core focus of this work. Supporting alternative aperture profiles requires only a modification to the integration region in Eq. (1). Separately, faithful reproduction of highlight circle-of-confusion effects requires RAW-space processing, which current feed-forward 3D reconstruction methods do not support. This limitation can be addressed by substituting a RAW-capable feed-forward model as reconstruction techniques mature. Meanwhile, RawNeRF already offers this capability, though at greater computational cost. For 4K extension, LaCT can be replaced with a 4K-capable feed-forward model once available. These substitutions require no changes to the rest of the pipeline, as Vid2Bokeh's modular design decouples the reconstruction backbone from the bokeh rendering and training stages.

Editing Model Design Choice. This work focuses on bokeh data acquisition, and model architecture engineering is outside its scope. We use a standard diffusion model fine-tuning setup to validate the utility of Vid2Bokeh data on complex geometries. One direction worth pursuing is training the model to produce per-

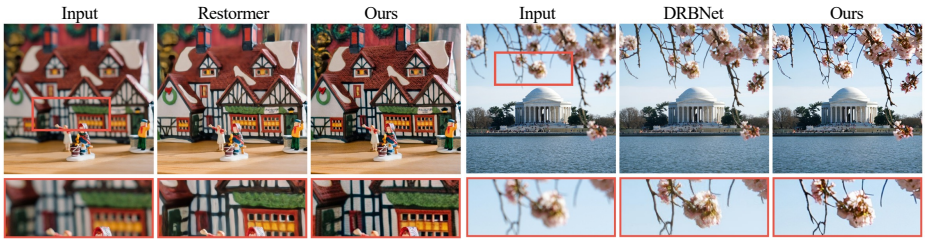


Fig. 11: Defocus blur removal on real images. Our method restores sharp details.

Method	DPDD Dataset [2]				Vid2Bokeh (Avg, 5 AR)			
	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	PSNR↑	SSIM↑	LPIPS↓	DISTS↓
DRBNet	26.67	0.818	0.137	0.091	26.45	0.792	0.248	0.178
Restormer	**26.77**	0.835	0.128	0.093	27.81	0.802	0.230	0.172
Ours	26.70	**0.841**	**0.105**	**0.078**	**28.58**	**0.853**	**0.145**	**0.103**

Table 5: Quantitative comparisons of defocus-blur removal. We evaluate on DPDD dataset and our Vid2Bokeh testing set. For Vid2Bokeh testing set, each method is tested on five defocus-blur levels, where inputs are generated at five different aperture radii and the goal is to recover the same near–all-in-focus target.

ceptually smooth, monotonically varying transitions across the full focus–DoF range without relying on depth or defocus maps. Architectures designed to better exploit Vid2Bokeh's geometrically faithful supervision remain an open direction.

Broader Impact for Downstream Applications. Beyond focus–DoF editing, the Vid2Bokeh approach of converting videos into dense, geometrically faithful light fields at scale could support several other tasks. It could benefit generative models for spatial image or photo re-framing. Additionally, the densely sampled angular information could prove useful for occlusion separation and depth estimation. The resulting light fields capture accurate parallax and occlusion relationships that are difficult to obtain from single images or depth maps.

7 Conclusion

We present Vid2Bokeh, a scalable pipeline that converts real-world videos into geometrically faithful bokeh training data via light field synthesis. It tackles the scalability–fidelity trade-off in bokeh data acquisition. Training on the resulting Vid2Bokeh Dataset, we show that even a standard diffusion model achieves geometrically consistent, bidirectional focus–DoF editing on complex geometries where depth-based methods consistently fail. These results indicate that geometric fidelity in large-scale training data, rather than model architecture alone, is the key factor behind robust bokeh editing.

References

1. Abuolaim, A., Afifi, M., Brown, M.S.: Improving single-image defocus deblurring: How dual-pixel images help through multi-task learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1231–1239 (2022)
2. Abuolaim, A., Brown, M.S.: Defocus deblurring using dual-pixel data. In: European conference on computer vision. pp. 111–126. Springer (2020)
3. Abuolaim, A., Delbracio, M., Kelly, D., Brown, M.S., Milanfar, P.: Learning to reduce defocus blur by realistically modeling dual-pixel data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2289–2298 (2021)
4. Alzayer, H., Abuolaim, A., Chan, L.C., Yang, Y., Lou, Y.C., Huang, J.B., Kar, A.: Dc2: Dual-camera defocus control by learning to refocus. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21488–21497 (2023)
5. Alzayer, H., Zhang, Y., Geng, C., Huang, J.B., Wu, J.: Coupled diffusion sampling for training-free multi-view image editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 43686–43696 (2026)
6. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18392–18402 (2023)
7. Buehler, C., Bosse, M., McMillan, L., Gortler, S., Cohen, M.: Unstructured lumigraph rendering. In: Proceedings of the 28th annual conference on Computer graphics and interactive techniques. pp. 425–432 (2001)
8. Busam, B., Hog, M., McDonagh, S., Slabaugh, G.: Sterefo: Efficient image refocusing with stereo vision. In: Proceedings of the IEEE/CVF international conference on computer vision workshops. pp. 0–0 (2019)
9. Cai, H., Huang, T.W., Gehlot, S., Feng, B.Y., Shah, S., Su, G.M., Metzler, C.: Parametric shadow control for portrait generation in text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18207–18217 (2025)
10. Cai, X., You, Z., Zhang, Z., Xue, T.: Da-vae: Plug-in latent compression for diffusion via detail alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18703–18713 (2026)
11. Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., Zheng, Y.: Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 22560–22570 (2023)
12. Cao, M., Zhang, X., Zheng, Y., Xia, Z.: Instruction-based image manipulation by watching how things move. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2704–2713 (2025)
13. Chen, Y., Zheng, C., Xu, H., Zhuang, B., Vedaldi, A., Cham, T.J., Cai, J.: Mvsplat360: Feed-forward 360 scene synthesis from sparse views. Advances in Neural Information Processing Systems **37**, 107064–107086 (2024)
14. Conde, M.V., Kolmet, M., Seizinger, T., Bishop, T.E., Timofte, R., Kong, X., Zhang, D., Wu, J., Wang, F., Peng, J., et al.: Lens-to-lens bokeh effect transformation. ntire 2023 challenge report. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1643–1659 (2023)
15. Couairon, G., Verbeek, J., Schwenk, H., Cord, M.: Diffedit: Diffusion-based semantic image editing with mask guidance. arXiv preprint arXiv:2210.11427 (2022)

16. Dai, Y., Zhang, Z., Zhou, S., Zhao, N.: Linear image generation by synthesizing exposure brackets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16206–16215 (2026)
17. Davis, A., Levoy, M., Durand, F.: Unstructured light fields. In: Computer Graphics Forum. vol. 31, pp. 305–314. Wiley Online Library (2012)
18. Debevec, P.E., Taylor, C.J., Malik, J.: Modeling and rendering architecture from photographs: A hybrid geometry-and image-based approach. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2, pp. 465–474 (2023)
19. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
20. Gao, R., Holynski, A., Henzler, P., Brussee, A., Martin-Brualla, R., Srinivasan, P., Barron, J.T., Poole, B.: Cat3d: Create anything in 3d with multi-view diffusion models. arXiv preprint arXiv:2405.10314 (2024)
21. Gortler, S.J., Grzeszczuk, R., Szeliski, R., Cohen, M.F.: The lumigraph. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2, pp. 453–464 (2023)
22. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022)
23. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
24. Ignatov, A., Patel, J., Timofte, R.: Rendering natural camera bokeh effect with deep learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 418–419 (2020)
25. Ignatov, A., Timofte, R., Qian, M., Qiao, C., Lin, J., Guo, Z., Li, C., Leng, C., Cheng, J., Peng, J., et al.: Aim 2020 challenge on rendering realistic bokeh. In: European Conference on Computer Vision. pp. 213–228. Springer (2020)
26. Kalantari, N.K., Wang, T.C., Ramamoorthi, R.: Learning-based view synthesis for light field cameras. ACM Transactions on Graphics (TOG) **35**(6), 1–10 (2016)
27. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139–1 (2023)
28. Kim, C., Zimmer, H., Pritch, Y., Sorkine-Hornung, A., Gross, M.H.: Scene reconstruction from high spatio-angular resolution light fields. ACM Trans. Graph. **32**(4), 73–1 (2013)
29. Kraus, M., Strengert, M.: Depth-of-field rendering by pyramidal image processing. In: Computer graphics forum. vol. 26, pp. 645–654. Wiley Online Library (2007)
30. Kutulakos, K.N., Seitz, S.M.: A theory of shape by space carving. International journal of computer vision **38**(3), 199–218 (2000)
31. Kwon, G., Ye, J.C.: Diffusion-based image translation using disentangled style and content representation. arXiv preprint arXiv:2209.15264 (2022)
32. Lee, J., Lee, S., Cho, S., Lee, S.: Deep defocus map estimation using domain adaptation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12222–12230 (2019)
33. Lee, J., Son, H., Rim, J., Cho, S., Lee, S.: Iterative filter adaptive network for single image defocus deblurring. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2034–2042 (2021)
34. Lee, S., Eisemann, E., Seidel, H.P.: Real-time lens blur effects and focus control. ACM Transactions on Graphics (TOG) **29**(4), 1–7 (2010)
35. Lee, S., Kim, G.J., Choi, S.: Real-time depth-of-field rendering using point splatting on per-pixel layers. In: Computer Graphics Forum. vol. 27, pp. 1955–1962. Wiley Online Library (2008)

36. Levoy, M., Hanrahan, P.: Light field rendering. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2, pp. 441–452 (2023)
37. Ling, L., Sheng, Y., Tu, Z., Zhao, W., Xin, C., Wan, K., Yu, L., Guo, Q., Yu, Z., Lu, Y., et al.: Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22160–22169 (2024)
38. Lombardi, S., Simon, T., Saragih, J., Schwartz, G., Lehrmann, A., Sheikh, Y.: Neural volumes: Learning dynamic renderable volumes from images. arXiv preprint arXiv:1906.07751 (2019)
39. Mandl, D., Mori, S., Mohr, P., Peng, Y., Langlotz, T., Schmalstieg, D., Kalkofen, D.: Neural bokeh: Learning lens blur for computational videography and out-of-focus mixed reality. In: 2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR). pp. 870–880. IEEE (2024)
40. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. arXiv preprint arXiv:2108.01073 (2021)
41. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
42. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text inversion for editing real images using guided diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6038–6047 (2023)
43. Mu, C.W.T., Huang, J.B., Liu, Y.L.: Generative refocusing: Flexible defocus control from a single image. arXiv preprint arXiv:2512.16923 (2025)
44. Ng, R., Levoy, M., Brédif, M., Duval, G., Horowitz, M., Hanrahan, P.: Light field photography with a hand-held plenoptic camera. Ph.D. thesis, Stanford university (2005)
45. Nie, S., Guo, H.A., Lu, C., Zhou, Y., Zheng, C., Li, C.: The blessing of randomness: Sde beats ode in general diffusion-based image editing. arXiv preprint arXiv:2311.01410 (2023)
46. Niemeyer, M., Mescheder, L., Oechsle, M., Geiger, A.: Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3504–3515 (2020)
47. Pan, L., Chowdhury, S., Hartley, R., Liu, M., Zhang, H., Li, H.: Dual pixel exploration: Simultaneous depth estimation and image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4340–4349 (2021)
48. Peng, J., Cao, Z., Luo, X., Lu, H., Xian, K., Zhang, J.: Bokehme: When neural rendering meets classical rendering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16283–16292 (2022)
49. Peng, J., Zhang, J., Luo, X., Lu, H., Xian, K., Cao, Z.: Mpib: An mpi-based bokeh rendering framework for realistic partial occlusion effects. In: European Conference on Computer Vision. pp. 590–607. Springer (2022)
50. Penner, E., Zhang, L.: Soft 3d reconstruction for view synthesis. ACM Transactions on Graphics (TOG) **36**(6), 1–11 (2017)
51. Pertuz, S., Puig, D., Garcia, M.A.: Analysis of focus measure operators for shape-from-focus. Pattern Recognition **46**(5), 1415–1432 (2013)

52. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
53. Potmesil, M., Chakravarty, I.: A lens and aperture camera model for synthetic image generation. ACM SIGGRAPH Computer Graphics **15**(3), 297–305 (1981)
54. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
55. Ruan, L., Chen, B., Li, J., Lam, M.L.: Aifnet: All-in-focus image restoration network using a light field-based dataset. IEEE Transactions on Computational Imaging **7**, 675–688 (2021)
56. Ruan, L., Chen, B., Li, J., Lam, M.: Learning to deblur using light field generated and real defocus images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16304–16313 (2022)
57. Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., Norouzi, M.: Palette: Image-to-image diffusion models. In: ACM SIGGRAPH 2022 conference proceedings. pp. 1–10 (2022)
58. Sasaki, H., Willcocks, C.G., Breckon, T.P.: Unit-ddpm: Unpaired image translation with denoising diffusion probabilistic models. arXiv preprint arXiv:2104.05358 (2021)
59. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4104–4113 (2016)
60. Seizinger, T., Vasluianu, F.A., Conde, M.V., Wu, Z., Timofte, R.: Bokehlicious: Photorealistic bokeh rendering with controllable apertures. arXiv preprint arXiv:2503.16067 (2025)
61. Sheng, Y., Yu, Z., Ling, L., Cao, Z., Zhang, X., Lu, X., Xian, K., Lin, H., Benes, B.: Dr. bokeh: Differentiable occlusion-aware bokeh rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4515–4525 (2024)
62. Sitzmann, V., Thies, J., Heide, F., Nießner, M., Wetzstein, G., Zollhofer, M.: Deepvoxels: Learning persistent 3d feature embeddings. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2437–2446 (2019)
63. Sitzmann, V., Zollhöfer, M., Wetzstein, G.: Scene representation networks: Continuous 3d-structure-aware neural scene representations. Advances in neural information processing systems **32** (2019)
64. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. pmlr (2015)
65. Son, H., Lee, J., Cho, S., Lee, S.: Single image defocus deblurring using kernel-sharing parallel atrous convolutions. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2642–2650 (2021)
66. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
67. Srinivasan, P.P., Tucker, R., Barron, J.T., Ramamoorthi, R., Ng, R., Snavely, N.: Pushing the boundaries of view extrapolation with multiplane images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 175–184 (2019)

68. Srinivasan, P.P., Wang, T., Sreelal, A., Ramamoorthi, R., Ng, R.: Learning to synthesize a 4d rgbd light field from a single image. In: Proceedings of the IEEE international conference on computer vision. pp. 2243–2251 (2017)
69. Tedla, S., Zhang, Z., Zhang, X., Xin, S.: Learning to refocus with video diffusion models. In: SIGGRAPH Asia 2025 Conference Proceedings. ACM (2025), to appear
70. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: European conference on computer vision. pp. 402–419. Springer (2020)
71. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1921–1930 (2023)
72. Vaidyanathan, K., Munkberg, J., Clarberg, P., Salvi, M.: Layered light field reconstruction for defocus blur. ACM Transactions on Graphics (TOG) **34**(2), 1–12 (2015)
73. Wallace, B., Gokul, A., Naik, N.: Edict: Exact diffusion inversion via coupled transformations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22532–22541 (2023)
74. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5294–5306 (2025)
75. Wang, L., Shen, X., Zhang, J., Wang, O., Lin, Z., Hsieh, C.Y., Kong, S., Lu, H.: Deeplens: Shallow depth of field from a single image. arXiv preprint arXiv:1810.08100 (2018)
76. Wang, T., Zhang, T., Zhang, B., Ouyang, H., Chen, D., Chen, Q., Wen, F.: Pretraining is all you need for image-to-image translation. arXiv preprint arXiv:2205.12952 (2022)
77. Wang, T., Xie, M., Cai, H., Shah, S., Metzler, C.A.: Flash-split: 2d reflection removal with flash cues and latent diffusion separation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5688–5698 (2025)
78. Wang, T., Piao, Y., Li, X., Zhang, L., Lu, H.: Deep learning for light field saliency detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8838–8848 (2019)
79. Wang, T.C., Zhu, J.Y., Kalantari, N.K., Efros, A.A., Ramamoorthi, R.: Light field video capture using a learning-based hybrid imaging system. ACM transactions on graphics (TOG) **36**(4), 1–13 (2017)
80. Wang, Y., Chen, X., Xu, X., Liu, Y., Zhao, H.: Diffcamera: Arbitrary refocusing on images. arXiv preprint arXiv:2509.26599 (2025)
81. Wanner, S., Goldluecke, B.: Globally consistent depth labeling of 4d light fields. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 41–48. IEEE (2012)
82. Wu, J., Cai, H., Fermuller, C., Metzler, C., Aloimonos, Y.: Real2sam2real: Generative 3d caches as complementary context for video diffusion. arXiv preprint arXiv:2606.00299 (2026)
83. Wu, J., Wang, T., Siddique, M.A.B., Islam, M.J., Fermuller, C., Aloimonos, Y., Metzler, C.A.: Single-step latent diffusion for underwater image restoration. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
84. Xin, S., Wadhwa, N., Xue, T., Barron, J.T., Srinivasan, P.P., Chen, J., Gkioulekas, I., Garg, R.: Defocus map estimation and deblurring from a single dual-pixel image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2228–2238 (2021)

85. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10371–10381 (2024)
86. You, Z., Wang, K., Zhang, H., Cai, X., Gu, J., Xue, T., Dong, C., Zhang, Z.: Photoframer: Multi-modal image composition instruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10197–10207 (2026)
87. Yu, W., Xing, J., Yuan, L., Hu, W., Li, X., Huang, Z., Gao, X., Wong, T.T., Shan, Y., Tian, Y.: Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. arXiv preprint arXiv:2409.02048 (2024)
88. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023)
89. Zhang, L., Rao, A., Agrawala, M.: Scaling in-the-wild training for diffusion-based illumination harmonization and editing by imposing consistent light transport. In: The Thirteenth International Conference on Learning Representations (2025)
90. Zhang, T., Bi, S., Hong, Y., Zhang, K., Luan, F., Yang, S., Sunkavalli, K., Freeman, W.T., Tan, H.: Test-time training done right. arXiv preprint arXiv:2505.23884 (2025)
91. Zhang, X., Matzen, K., Nguyen, V., Yao, D., Zhang, Y., Ng, R.: Synthetic defocus and look-ahead autofocus for casual videography. arXiv preprint arXiv:1905.06326 (2019)
92. Zhao, M., Bao, F., Li, C., Zhu, J.: Egsde: Unpaired image-to-image translation via energy-guided stochastic differential equations. Advances in Neural Information Processing Systems **35**, 3609–3623 (2022)
93. Zhou, T., Tucker, R., Flynn, J., Fyffe, G., Snavely, N.: Stereo magnification: Learning view synthesis using multiplane images. arXiv preprint arXiv:1805.09817 (2018)
94. Zhu, C., Fan, Q., Zhang, Q., Chen, J., Zhang, H., Xu, C., Shi, B.: Bokehdiff: Neural lens blur with one-step diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9508–9518 (2025)