LENSMASTER: ALL-IN-ONE PHOTOREALISTIC DEPTH-OF-FIELD RENDERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Depth-of-field (DoF) editing remains challenging due to three limitations of existing methods: reliance on error-prone depth or segmentation priors, one-way rendering from all-in-focus inputs, and limited modeling of lens-dependent bokeh characteristics. We present LensMaster, an end-to-end framework that formulates DoF manipulation as a *bidirectional optical-state transformation*. Given a source image and its optical state, LensMaster continuously controls both focus distance and aperture to synthesize an arbitrary target state without explicit geometric priors. To support coherent editing across optical configurations, we introduce optical-state consistency with identity, cycle, and composition constraints. We further disentangle shared DoF transformation from lens-specific appearance through spatially varying lens-character modulation, enabling a single model to reproduce distinct real-lens bokeh characteristics. To support cross-lens learning, we introduce Fuji-Bokeh, a real-world multi-focus, multi-aperture dataset captured with a Fujifilm system. Experiments demonstrate that LensMaster achieves accurate continuous, bidirectional, and lens-aware DoF editing while preserving fine structures and realistic Bokeh appearance.

1 INTRODUCTION

Depth-of-field (DoF) is a fundamental photographic control for directing visual attention and shaping image aesthetics. By jointly adjusting focus distance and aperture, photographers determine both the focal plane and the extent of defocus, while the appearance of Bokeh is further influenced by lens-specific effects such as aperture geometry, optical vignetting, aberrations, and diffraction (Huang et al., 2026b). Reproducing these effects from a single image remains a challenging problem in computational photography.

Existing DoF editing methods are limited in three aspects. First, geometry-assisted pipelines estimate depth maps (Yang et al., 2024; Gui et al., 2025; Li et al., 2025), foreground masks (Kirillov et al., 2023), or other intermediate representations before rendering (Zhou et al., 2018; Carvalho et al., 2018; Xiao et al., 2018; Wang et al., 2018; Peng et al., 2022b; Sheng et al., 2024; Zhu et al., 2025). Errors in these estimates propagate directly into the rendering stage and often cause artifacts around depth discontinuities and fine structures. Second, recent end-to-end methods (Seizinger et al., 2025; Huang et al., 2026a; Wang et al., 2025; Qin et al., 2025; Tuan Mu et al., 2025; Huang et al., 2026b) largely assume a small-aperture or nearly all-in-focus input and synthesize stronger defocus. For example, Bokehlicious (Seizinger et al., 2025) enables continuous aperture control from real paired photographs, but preserves the original focal subject. As a result, existing formulations do not naturally support arbitrary transformations between different focus and aperture configurations.

We instead formulate DoF editing as an *arbitrary-to-arbitrary optical-state transformation*, where an optical state is jointly defined by focus distance and aperture. This formulation covers conventional Bokeh synthesis as a special case, while also supporting DoF expansion, focal-plane relocation, and joint focus–aperture editing. Because increasing defocus discards visual detail whereas reducing defocus requires recovering it, the two directions are inherently asymmetric. We therefore introduce *optical-state consistency*, using identity, cycle, and composition constraints to organize different photographic configurations into a coherent continuous state space.

A further challenge lies in modeling the appearance of real lenses. Synthetic rendering commonly relies on idealized point spread functions, while real lenses exhibit spatially varying characteristics such

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

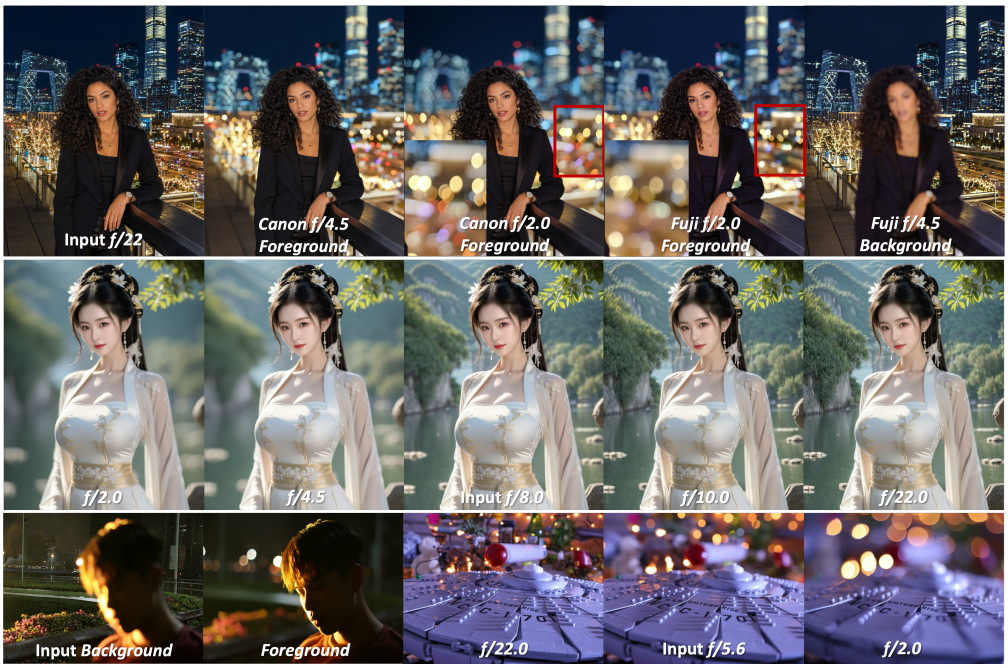

Figure 1: **LensMaster enables unified optical-state and lens-character editing.** Starting from a single photograph, our model supports continuous aperture control, focal-plane relocation, defocus removal, and lens-character rendering within a unified framework.

as off-axis Bokeh deformation, optical vignetting, aperture-dependent highlight shapes, and diffraction. Moreover, models trained on a single camera-lens system tend to entangle these lens-specific effects with the common DoF transformation. To address this, we explicitly disentangle optical state from lens character: a shared backbone models focus- and aperture-dependent transformations, while a spatially varying lens-character module captures lens-specific residual appearance. This enables a single model to reproduce different real-lens Bokeh styles while independently controlling focus and aperture.

Learning such a factorization requires real observations across both optical states and lens domains. RealBokeh (Seizinger et al., 2025) provides high-quality Canon captures with multiple apertures but a fixed focal position within each scene. We therefore introduce FujiBokeh, a complementary Fujifilm dataset with multiple aperture settings and aperture-adaptive focal stacks that sweep the focal plane from foreground to background. Together, the two datasets provide heterogeneous real-lens supervision for learning shared optical transformations and lens-specific appearance.

Our contributions are summarized as follows:

- We formulate single-image DoF manipulation as an end-to-end *arbitrary optical-state transformation*, enabling continuous control of both focus distance and aperture without explicit depth maps or subject masks.
- We introduce bidirectional optical-state consistency with identity, cycle, and composition constraints, allowing coherent editing between arbitrary source and target focus–aperture states.
- We propose a lens-disentangled framework with spatially varying lens modulation, enabling one model to capture shared DoF behavior while reproducing distinct real-lens Bokeh characteristics.
- We introduce FujiBokeh, a real-world multi-focus, multi-aperture Fujifilm dataset that complements Canon-based RealBokeh and enables cross-lens controllable DoF editing.

2 Related Work

2.1 Geometry-Assisted Depth-of-Field Rendering

Classical and learning-based DoF rendering methods commonly decompose the task into scene-geometry estimation and image-space rendering. Depth-aware approaches estimate or acquire a depth map and apply spatially varying blur kernels, layered compositing, or neural rendering according to the distance from the focal plane (Zhou et al., 2018; Wang et al., 2018; Xiao et al., 2018; Peng et al., 2022a; Luo et al., 2024). Other approaches additionally rely on foreground segmentation, alpha mattes or background inpainting to improve foreground–background separation (Peng et al., 2022b; Sheng et al., 2024; Peng et al., 2024; Zhu et al., 2025). Such explicit geometry provides intuitive control over focus and blur strength, but also introduces a strong dependency on intermediate predictions. Depth discontinuities are particularly challenging: even small geometric errors can produce color bleeding or incorrect blur at object boundaries. Thin structures, hair, reflective regions, and partially transparent objects further amplify this problem. Our method instead learns the complete optical transformation end-to-end from real image observations. It does not explicitly predict depth or segmentation, allowing the latent representation to retain image evidence useful for both scene understanding and photographic rendering.

2.2 End-to-End and Controllable Bokeh Rendering

The availability of paired small- and large-aperture photographs has motivated a second family of methods that directly learns Bokeh transformation as an image-to-image mapping (Ignatov et al., 2020; Luo et al., 2023; Conde et al., 2023; Nagasubramaniam & Younes, 2022; Qian et al., 2020). Compared with geometry-assisted pipelines, these approaches reduce explicit preprocessing and can implicitly learn complex image formation effects from data. However, early real-world datasets typically provide only limited aperture settings, and most methods therefore learn fixed transformations. Bokehlicious (Seizinger et al., 2025) advances this direction by introducing RealBokeh and Aperture-Aware Attention, enabling continuous aperture control without explicit depth maps. With the development of diffusion and flow-based generative models (Rombach et al., 2022; Lipman et al., 2022; Esser et al., 2024), recent works further explore richer photographic controls. BokehFlow (Huang et al., 2026a) studies depth-free controllable Bokeh rendering, LensStyle (Huang et al., 2026b) models controllable lens appearance, while CamEdit (Qin et al., 2025), DiffCamera (Wang et al., 2025), and GenRefocus (Tuan Mu et al., 2025) extend generative editing to continuous camera control and refocusing. Our work goes further by jointly modeling source and target focus–aperture states, enabling arbitrary bidirectional DoF editing, including defocus synthesis, DoF expansion, refocusing, and joint focus–aperture manipulation, while also disentangling shared DoF transformation from lens-specific Bokeh appearance.

3 Method

Given a source photograph, our goal is to edit its depth-of-field (DoF) by continuously controlling focus distance and aperture, while optionally changing the characteristic Bokeh appearance of the lens. As illustrated in Fig. 2, LensMaster performs the entire transformation end-to-end without explicitly estimating depth maps, subject masks, or circle-of-confusion maps. The framework consists of a shared optical transformation backbone, continuous source–target optical conditioning, and a lens-specific character branch.

3.1 Optical-State Formulation

We describe a photographic observation as

$$\mathbf{I} = \mathbf{I}(\mathbf{S}, o, \ell), \qquad o = (f, a), \tag{1}$$

where $\mathbf{S}$ denotes scene content, f the focus coordinate, a the aperture, and ℓ the lens identity. Given a source image $\mathbf{I}_s$ captured at optical state $o_s = (f_s, a_s)$, our objective is to synthesize the appearance under a target state $o_t = (f_t, a_t)$ and target lens ℓ_t:

$$\hat{\mathbf{I}}_t = \mathcal{G}_\theta(\mathbf{I}_s, o_s, o_t, \ell_t). \tag{2}$$

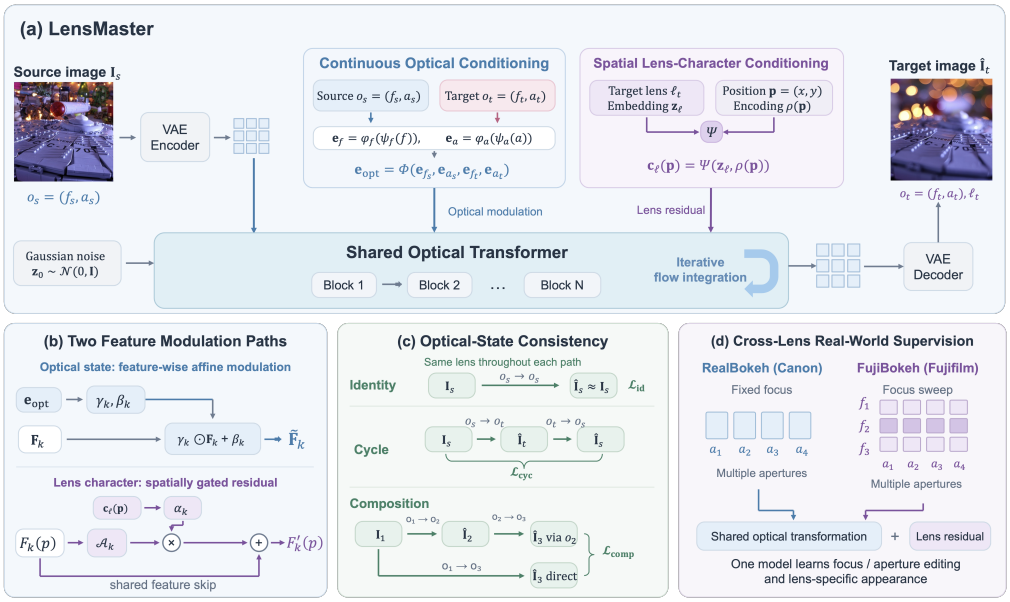

Figure 2: **Overview of LensMaster.** Source and target optical states continuously control focus and aperture, while spatial lens modulation captures lens-specific Bokeh appearance.

This formulation generalizes conventional Bokeh rendering, which typically assumes an approximately all-in-focus source and only varies the target aperture. Instead, Eq. equation 2 supports arbitrary transformations

$$(f_s, a_s) \to (f_t, a_t), \tag{3}$$

including aperture opening and closing, focal-plane relocation, and their joint manipulation.

We operate in the latent space of a pretrained image autoencoder. The source image is encoded as conditioning tokens, while the target latent is optimized through conditional flow matching. The resulting transformer jointly models scene content and the desired optical transformation in a shared latent representation.

3.2 Continuous Optical-State Conditioning

Focus distance and aperture play complementary roles in image formation: focus determines the location of the focal plane, whereas aperture primarily controls the extent of defocus around that plane. We therefore encode them separately:

$$\mathbf{e}_f = \phi_f(\psi_f(f)), \qquad \mathbf{e}_a = \phi_a(\psi_a(a)), \tag{4}$$

where ψ_f and ψ_a normalize the corresponding optical parameters and ϕ_f, ϕ_a are learnable continuous embeddings.

Importantly, we condition on both the source and target states rather than the target alone:

$$\mathbf{e}_{\text{opt}} = \Phi(\mathbf{e}_{f_s}, \mathbf{e}_{a_s}, \mathbf{e}_{f_t}, \mathbf{e}_{a_t}). \tag{5}$$

This removes an inherent ambiguity in arbitrary DoF editing: the same target state may require synthesizing additional defocus, recovering previously blurred detail, or relocating the focal plane depending on the source state.

The resulting optical embedding modulates intermediate transformer features:

$$\tilde{\mathbf{F}}_k = \gamma_k(\mathbf{e}_{\text{opt}}) \odot \mathbf{F}_k + \beta_k(\mathbf{e}_{\text{opt}}), \tag{6}$$

where γ_k and β_k are learned modulation functions. This continuous conditioning allows a single model to represent different source-to-target trajectories throughout the focus–aperture space.

3.3 Lens-Disentangled Character Modeling

Optical-state transformation and lens appearance are related but distinct. While the relationship between scene depth, focus, and aperture is largely shared across lenses, real lenses exhibit different

Bokeh characteristics due to aperture geometry, optical vignetting, aberrations, and diffraction. We therefore factorize the representation into a shared optical component and a lens-specific character component.

Let $\mathbf{z}_\ell$ denote a learnable embedding for lens ℓ. A global lens code alone is insufficient because many real-lens effects vary across the image plane. For example, optical vignetting causes peripheral Bokeh highlights to deform differently from those near the optical axis. We therefore construct a spatially varying lens-character field

$$\mathbf{c}_\ell(\mathbf{p}) = \Psi\big(\mathbf{z}_\ell, \rho(\mathbf{p})\big), \qquad \mathbf{p} = (x, y), \tag{7}$$

where $\rho(\mathbf{p})$ is a positional encoding.

The lens-character field modulates the shared optical features through a lightweight residual branch:

$$\mathbf{F}'_k(\mathbf{p}) = \mathbf{F}_k(\mathbf{p}) + \alpha_k(\mathbf{c}_\ell(\mathbf{p})) \odot \mathcal{A}_k(\mathbf{F}_k(\mathbf{p})). \tag{8}$$

The shared backbone therefore models focus- and aperture-dependent scene transformations, while the lens branch captures residual appearance specific to each real lens.

This factorization also enables heterogeneous supervision from different datasets. RealBokeh provides dense aperture variation from a Canon system with a fixed focal position, whereas FujiBokeh provides both aperture variation and focal stacks from a Fujifilm system. Both datasets supervise the shared aperture-dependent transformation, while FujiBokeh additionally provides direct supervision for focal-plane relocation; lens-dependent appearance is absorbed by the corresponding character branch. Thus, a single model can learn from both datasets without forcing their Bokeh characteristics into one shared appearance representation.

3.4 Flow Matching and Optical-State Consistency

Let

$$\mathbf{z}_1 = \mathcal{E}(\mathbf{I}_t) \tag{9}$$

be the latent representation of the target image and $\mathbf{z}_0 \sim \mathcal{N}(0, \mathbf{I})$ a Gaussian sample. We construct the linear probability path

$$\mathbf{z}_t = (1 - t)\mathbf{z}_0 + t\mathbf{z}_1, \qquad t \sim \mathcal{U}(0, 1), \tag{10}$$

with target velocity

$$\mathbf{v}^* = \mathbf{z}_1 - \mathbf{z}_0. \tag{11}$$

Conditioned on the source image, optical states, and target lens, the network predicts

$$\mathbf{v}_\theta = \mathbf{v}_\theta(\mathbf{z}_t, t, \mathbf{I}_s, o_s, o_t, \ell_t), \tag{12}$$

and is optimized by

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{t,\mathbf{z}_0}\left[\|\mathbf{v}_\theta - \mathbf{v}^*\|_2^2\right]. \tag{13}$$

Since arbitrary DoF editing includes both information-removing and information-recovering transformations, direct paired supervision alone does not guarantee a coherent optical-state space. We therefore introduce three complementary consistency constraints.

For identical source and target states, the identity loss preserves the input:

$$\mathcal{L}_{\mathrm{id}} = \|\mathcal{G}_\theta(\mathbf{I}_s, o_s, o_s, \ell_s) - \mathbf{I}_s\|_1. \tag{14}$$

For bidirectional editing, cycle consistency encourages an edited image to retain information useful for the reverse transformation:

$$\mathcal{L}_{\mathrm{cyc}} = \|\mathcal{G}_\theta(\mathcal{G}_\theta(\mathbf{I}_s, o_s, o_t, \ell_s), o_t, o_s, \ell_s) - \mathbf{I}_s\|_1. \tag{15}$$

We do not interpret this constraint as physical invertibility; rather, it regularizes scene-content preservation during bidirectional editing.

Finally, for three optical states o_1, o_2, o_3, composition consistency encourages direct and intermediate transformations to agree:

$$\begin{aligned}\mathcal{L}_{\mathrm{comp}} = \big\|&\mathcal{G}_\theta(\mathbf{I}_1, o_1, o_3, \ell) \\ &- \mathcal{G}_\theta(\mathcal{G}_\theta(\mathbf{I}_1, o_1, o_2, \ell), o_2, o_3, \ell)\big\|_1.\end{aligned} \tag{16}$$

The final objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{FM}} + \lambda_{\mathrm{id}}\mathcal{L}_{\mathrm{id}} + \lambda_{\mathrm{cyc}}\mathcal{L}_{\mathrm{cyc}} + \lambda_{\mathrm{comp}}\mathcal{L}_{\mathrm{comp}}. \tag{17}$$



Figure 3: **Samples from FujiBokeh.** Each row shows a real focal stack with the focus plane swept continuously from foreground to background.

Table 1: Comparison of Bokeh rendering datasets. FujiBokeh is distinguished by real multi-aperture captures with dense focus-distance variation for each scene.

Dataset	# Images	Apertures	Multi-Focus	Resolution	Real
EBB! (Ignatov et al., 2020)	4,694	$f/1.8$	No	1536×1024	Yes
Aperture (Zhang et al., 2019)	2,942	$f/8.0,\ f/2.0$	No	1200×750	Yes
BETD (Conde et al., 2023)	20,000	$f/2.0,\ f/1.8$	No	1920×1080	No
RealBokeh (Seizinger et al., 2025)	23,000	$f/22.0$–$f/2.0$	No	2000×1500	Yes
FujiBokeh (Ours)	**40,000**	$f/22.0$–$f/2.8$	**Yes**	3000×2000	**Yes**

4 FujiBokeh

Learning arbitrary DoF transformations requires observations that vary in both aperture and focus distance. Existing real-world Bokeh datasets usually fix the focal subject; FujiBokeh provides complementary supervision for continuous optical-state editing and cross-lens Bokeh modeling.

We capture each scene with a Fujifilm XT-2 and XF16-55mm *f/2.8 R* zoom lens on a fixed tripod. Automated capture records all configurations within a short interval, minimizing scene motion and photometric variation. Rather than an aperture sequence at a fixed focal plane, each scene samples a two-dimensional grid over focus samples $\mathcal{F} = \{f_1, \ldots, f_M\}$ and apertures $\mathcal{A} = \{a_1, \ldots, a_N\}$:

$$\mathcal{I} = \{\mathbf{I}(f_i, a_j) \mid f_i \in \mathcal{F}, a_j \in \mathcal{A}\}. \tag{18}$$

Any pair in this stack defines a valid transformation, supporting aperture editing, refocusing, DoF expansion and reduction, and joint manipulation.

FujiBokeh contains approximately 40,000 3000×2000 images from 200 scenes spanning day and night, indoor and outdoor settings, people, and general objects. Each scene is captured at four apertures with a foreground-to-background focal sweep. The number of focal planes is aperture-adaptive: larger apertures, whose shallower DoF changes more rapidly with focus, are sampled more densely. Each stack contains 1–100 images, yielding real observations across both optical dimensions.

5 Experiments

We evaluate LensMaster on RealBokeh (Seizinger et al., 2025) and the proposed FujiBokeh dataset. We report PSNR and SSIM for reconstruction fidelity and LPIPS for perceptual similarity. All quantitative results are computed on held-out scenes without training-test content overlap.

5.1 Experimental Setup

We evaluate three complementary capabilities: (i) controllable DoF rendering from an approximately all-in-focus (AIF) input, (ii) defocus removal from already blurred inputs, and (iii) arbitrary transformations between source and target focus–aperture states. On RealBokeh, the focal position is

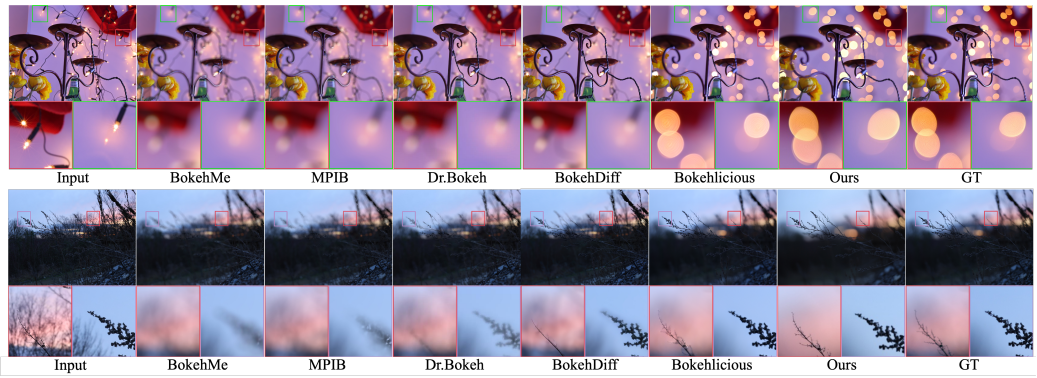

Figure 4: **Qualitative comparison on RealBokeh for $f/22 \rightarrow f/2.0$ rendering.** Our method better reproduces the characteristic Canon Bokeh, including peripheral cat-eye effects, while preserving foreground structures and fine details more faithfully.

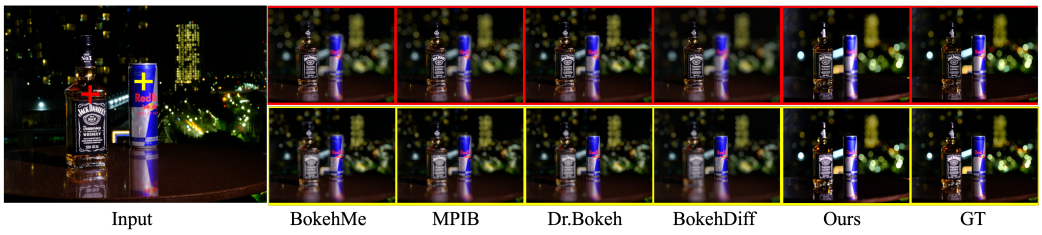

Figure 5: **Qualitative comparison on FujiBokeh for $f/22 \rightarrow f/2.8$ with focus plane varying.** Red and yellow crosses indicate the target focus locations; our method achieves more accurate focal-plane relocation and more faithful Fujifilm Bokeh.

Table 2: Controllable DoF rendering from AIF inputs. On RealBokeh, only aperture varies; on FujiBokeh, both focus distance and aperture vary.

Method	RealBokeh			FujiBokeh		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
MPIB (Peng et al., 2022b)	27.84	0.9012	0.1645	24.68	0.8234	0.2457
BokehMe (Peng et al., 2022a)	28.91	0.9128	0.1489	25.14	0.8351	0.2283
Dr.Bokeh (Sheng et al., 2024)	28.36	0.9075	0.1537	25.42	0.8416	0.2196
BokehDiff (Zhu et al., 2025)	<u>31.72</u>	<u>0.9349</u>	<u>0.1038</u>	<u>27.86</u>	<u>0.8827</u>	<u>0.1541</u>
Bokehlicious (Seizinger et al., 2025)	31.45	0.9317	0.1096	–	–	–
Ours	**32.58**	**0.9436**	**0.0867**	**29.31**	**0.9018**	**0.1264**

fixed within each scene, enabling evaluation of aperture control and defocus removal. FujiBokeh additionally varies focus distance, allowing evaluation of focal-plane relocation and joint focus–aperture editing. We compare against MPIB (Peng et al., 2022b), BokehMe (Peng et al., 2022a), Dr.Bokeh (Sheng et al., 2024), BokehDiff (Zhu et al., 2025), Bokehlicious (Seizinger et al., 2025), DiffCamera (Wang et al., 2025), and GenRefocus (Tuan Mu et al., 2025), according to the capabilities supported by each method.

5.2 SOTA Comparison

Controllable Rendering from AIF Inputs. We first evaluate controllable DoF rendering from approximately all-in-focus inputs. On RealBokeh, we render from $f/22$ to different target apertures while keeping the focal plane fixed, i.e., $(f_s, a_{\text{AIF}}) \rightarrow (f_s, a_t)$. On FujiBokeh, both focus and aperture are changed, $(f_s, a_{\text{AIF}}) \rightarrow (f_t, a_t)$. As shown in Tab. 2, our method achieves the best performance on both datasets, showing that the generalized optical-state formulation preserves strong aperture-control performance while enabling focal-plane manipulation. Qualitatively, Fig. 4 shows that our method better reproduces Canon-specific Bokeh, including peripheral cat-eye effects, while preserving foreground details. On FujiBokeh (Fig. 5), our method more accurately moves the focal plane to the indicated target positions and produces Bokeh consistent with the Fujifilm reference.

Table 3: Defocus removal on RealBokeh. Larger source apertures correspond to stronger defocus.

Method	$f/4.0 \rightarrow f/22$			$f/2.0 \rightarrow f/22$		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Bokehlicious (Seizinger et al., 2025)	29.42	0.9018	0.1216	25.91	0.8269	0.2013
DiffCamera (Wang et al., 2025)	29.86	0.9097	0.1132	26.37	0.8395	0.1867
GenRefocus (Tuan Mu et al., 2025)	<u>30.31</u>	<u>0.9165</u>	<u>0.1014</u>	<u>26.94</u>	<u>0.8512</u>	<u>0.1698</u>
Ours	**31.08**	**0.9274**	**0.0847**	**28.13**	**0.8749**	**0.1372**

Table 4: Arbitrary optical-state editing on FujiBokeh.

Method	Aperture Opening			Aperture Closing		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
DiffCamera (Wang et al., 2025)	28.07	0.8894	0.1512	26.21	0.8526	0.1987
GenRefocus (Tuan Mu et al., 2025)	<u>28.62</u>	<u>0.8987</u>	<u>0.1375</u>	<u>27.04</u>	<u>0.8694</u>	<u>0.1746</u>
Ours	**29.73**	**0.9152**	**0.1086**	**28.41**	**0.8927**	**0.1419**
Method	Refocusing			Joint Editing		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
DiffCamera (Wang et al., 2025)	25.46	0.8371	0.2194	24.91	0.8243	0.2368
GenRefocus (Tuan Mu et al., 2025)	<u>26.83</u>	<u>0.8661</u>	<u>0.1812</u>	<u>26.17</u>	<u>0.8517</u>	<u>0.2043</u>
Ours	**28.26**	**0.8914**	**0.1451**	**27.68**	**0.8786**	**0.1648**

Arbitrary Optical-State Editing. We next evaluate arbitrary transformations $(f_s, a_s) \rightarrow (f_t, a_t)$ from already defocused inputs. On RealBokeh, where focus is fixed, we evaluate defocus removal from $f/4.0$ and $f/2.0$ to $f/22$. As shown in Tab. 3, all methods degrade as source defocus increases, while our advantage becomes larger in the more challenging $f/2.0 \rightarrow f/22$ setting. Our method improves PSNR from 26.94 dB to 28.13 dB and reduces LPIPS from 0.1698 to 0.1372 over the second-best method. Fig. 6 further shows that our method recovers finer scene details and reproduces the small-aperture starburst pattern around point light sources rather than simply sharpening the blurred input.

FujiBokeh enables the full arbitrary-state setting, including aperture opening, aperture closing, refocusing, and joint focus–aperture editing. Tab. 4 shows that our method performs best in all four settings, with particularly clear gains for aperture closing, refocusing, and joint editing, where previously suppressed details must be recovered. The refocusing comparisons in Fig. 7 further demonstrate this capability in both directions: from foreground focus to background focus and vice versa. At the large aperture $f/2.8$, competing methods struggle to remove the extreme source defocus and reconstruct missing details, whereas our method successfully restores the newly focused regions and simultaneously produces realistic Fujifilm Bokeh in the newly defocused areas. Together, these results validate our formulation of DoF manipulation as a unified bidirectional optical-state transformation.

5.3 Ablation Studies

We ablate source–target optical conditioning, optical-state consistency, and lens-character modeling in Tab. 5. Explicitly incorporating the source optical state substantially improves over target-only conditioning, while disentangling focus and aperture provides a further gain. Identity, cycle, and composition consistency progressively improve arbitrary DoF editing, supporting the proposed coherent optical-state formulation. Finally, spatial lens modulation clearly outperforms both lens-agnostic and global lens conditioning, confirming that real lens character is inherently field dependent.

Table 5: Ablation of optical conditioning, consistency regularization, and lens-character modeling.

Optical Conditioning			Consistency			Lens Modeling		
Variant	PSNR↑	LPIPS↓	Variant	PSNR↑	LPIPS↓	Variant	PSNR↑	LPIPS↓
Target state	26.74	0.1873	Direct	26.84	0.1876	No lens	27.18	0.1738
Source–Target	27.39	0.1619	$+\mathcal{L}_{\text{id}} + \mathcal{L}_{\text{cyc}}$	27.42	0.1697	Global	27.54	0.1587
Disentangled	**27.68**	**0.1548**	$+\mathcal{L}_{\text{comp}}$	**27.68**	**0.1648**	Spatial	**28.19**	**0.1326**



Figure 6: **Defocus removal from $f/2.0$ to $f/22$ on RealBokeh.** Our method recovers fine scene details more faithfully and, in the presence of point light sources, reproduces the characteristic starburst diffraction pattern of a small-aperture lens.

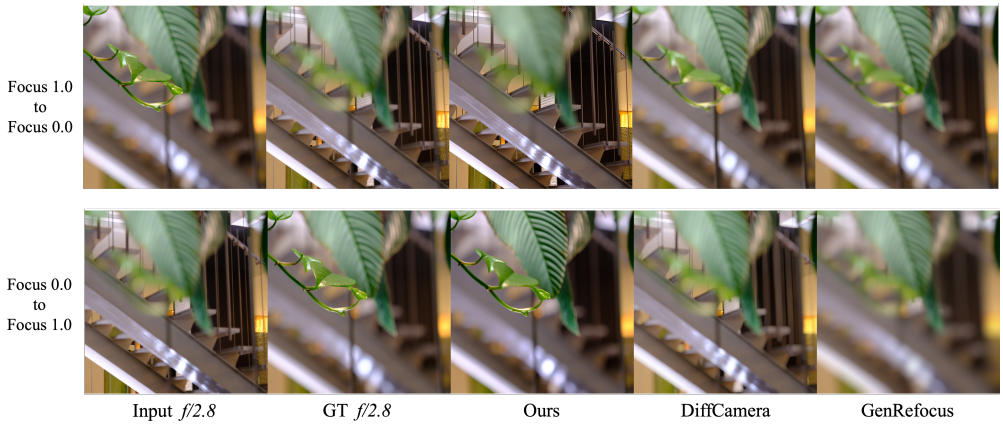

Figure 7: **Refocusing on FujiBokeh at $f/2.8$.** Our method successfully recovers details from heavily defocused regions when shifting focus between foreground and background, while other methods struggle to remove extreme defocus blur.

6 Conclusion

We presented LensMaster, a unified framework for end-to-end, bidirectional, and lens-aware depth-of-field editing. By conditioning on both source and target optical states, our method supports continuous manipulation of focus distance and aperture without explicit depth or segmentation priors. Optical-state consistency promotes coherent transformations across arbitrary focus–aperture configurations, while lens-disentangled modeling separates shared DoF behavior from spatially varying lens character. We further introduced FujiBokeh, a real-world multi-focus, multi-aperture Fujifilm dataset that complements Canon-based RealBokeh for cross-lens learning. Together, these components extend DoF editing from one-way Bokeh synthesis to controllable manipulation of focal plane, aperture, and real-lens appearance within a single framework.

THE USE OF GENERATIVE AI

We used generative AI tools solely as auxiliary assistants for writing and software development. For manuscript preparation, they were used for language polishing, including grammar and spelling correction, improving clarity and flow, shortening overly long sentences, and assisting with LaTeX drafting and revision. They were also used to assist with code editing, debugging, and identifying potential implementation issues.

All AI-generated text and code suggestions were manually reviewed, verified, and, when necessary, substantially revised by the authors before use. All technical content, mathematical formulations, experimental procedures, implementations, quantitative results, and scientific conclusions were independently checked and validated by the authors. The authors take full responsibility for the correctness and final content of this work.

REFERENCES

Marcela Carvalho, Bertrand Le Saux, Pauline Trouvé-Peloux, Andrés Almansa, and Frédéric Champagnat. Deep depth from defocus: how can defocus blur improve 3d estimation using dense neural networks? In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pp. 0–0, 2018.

Marcos V Conde, Manuel Kolmet, Tim Seizinger, Tom E Bishop, Radu Timofte, Xiangyu Kong, Dafeng Zhang, Jinlong Wu, Fan Wang, Juewen Peng, 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.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Ming Gui, Johannes Schusterbauer, Ulrich Prestel, Pingchuan Ma, Dmytro Kotovenko, Olga Grebenkova, Stefan Andreas Baumann, Vincent Tao Hu, and Björn Ommer. Depthfm: Fast generative monocular depth estimation with flow matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3203–3211, 2025.

Yachuan Huang, Xianrui Luo, Qiwen Wang, Liao Shen, Jiaqi Li, Huiqiang Sun, Zihao Huang, Wei Jiang, and Zhiguo Cao. Bokehflow: Depth-free controllable bokeh rendering via flow matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 5167–5175, 2026a.

Yachuan Huang, Liwen Xiao, Liao Shen, Qiwen Wang, Huiqiang Sun, Zhiyu Pan, and Zhiguo Cao. Lensstyle: Learning the aesthetics of lenses for controllable optical style rendering. In *European Conference on Computer Vision*, pp. 135–152. Springer, 2026b.

Andrey Ignatov, Jagruti Patel, and Radu Timofte. 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.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 3992–4003. IEEE, 2023.

Jiaqi Li, Yiran Wang, Jinghong Zheng, Junrui Zhang, Liao Shen, Tianqi Liu, and Zhiguo Cao. Ch3depth: Efficient and flexible depth foundation model with flow matching. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7222–7232, 2025.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Xianrui Luo, Juewen Peng, Ke Xian, Zijin Wu, and Zhiguo Cao. Defocus to focus: Photo-realistic bokeh rendering by fusing defocus and radiance priors. *Information Fusion*, 89:320–335, 2023.

Yawen Luo, Min Shi, Liao Shen, Yachuan Huang, Zixuan Ye, Juewen Peng, and Zhiguo Cao. Video bokeh rendering: Make casual videography cinematic. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 7677–7685, 2024.

Hariharan Nagasubramaniam and Rabih Younes. Bokeh effect rendering with vision transformers. 2022.

Juewen Peng, Zhiguo Cao, Xianrui Luo, Hao Lu, Ke Xian, and Jianming Zhang. Bokehme: When neural rendering meets classical rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16283–16292, 2022a.

Juewen Peng, Jianming Zhang, Xianrui Luo, Hao Lu, Ke Xian, and Zhiguo Cao. Mpib: An mpi-based bokeh rendering framework for realistic partial occlusion effects. In *European Conference on Computer Vision*, pp. 590–607. Springer, 2022b.

Juewen Peng, Zhiguo Cao, Xianrui Luo, Ke Xian, Wenfeng Tang, Jianming Zhang, and Guosheng Lin. Bokehme++: Harmonious fusion of classical and neural rendering for versatile bokeh creation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Ming Qian, Congyu Qiao, Jiamin Lin, Zhenyu Guo, Chenghua Li, Cong Leng, and Jian Cheng. Bggan: Bokeh-glass generative adversarial network for rendering realistic bokeh. In *European Conference on Computer Vision*, pp. 229–244. Springer, 2020.

Xinran Qin, Zhixin Wang, Fan Li, Haoyu Chen, Renjing Pei, Wenbo Li, and Xiaochun Cao. Camedit: Continuous camera parameter control for photorealistic image editing. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

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

Tim Seizinger, Florin-Alexandru Vasluianu, Marcos V Conde, Zongwei Wu, and Radu Timofte. Bokehlicious: Photorealistic bokeh rendering with controllable apertures. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8908–8917. IEEE, 2025.

Yichen Sheng, Zixun Yu, Lu Ling, Zhiwen Cao, Xuaner Zhang, Xin Lu, Ke Xian, Haiting Lin, and Bedrich Benes. Dr. bokeh: differentiable occlusion-aware bokeh rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4515–4525, 2024.

Chun-Wei Tuan Mu, Jia-Bin Huang, and Yu-Lun Liu. Generative refocusing: Flexible defocus control from a single image. *arXiv preprint arXiv:2512.16923*, 2025.

Lijun Wang, Xiaohui Shen, Jianming Zhang, Oliver Wang, Zhe Lin, Chih-Yao Hsieh, Sarah Kong, and Huchuan Lu. Deeplens: Shallow depth of field from a single image. *arXiv preprint arXiv:1810.08100*, 2018.

Yiyang Wang, Xi Chen, Xiaogang Xu, Yu Liu, and Hengshuang Zhao. Diffcamera: Arbitrary refocusing on images. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, SA Conference Papers '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400721373. doi: 10.1145/3757377.3763827. URL https://doi.org/10.1145/3757377.3763827.

Lei Xiao, Anton Kaplanyan, Alexander Fix, Matt Chapman, and Douglas Lanman. Deepfocus: Learned image synthesis for computational display. In *ACM SIGGRAPH 2018 Talks*, pp. 1–2. 2018.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024.

Xuaner Zhang, Kevin Matzen, Vivien Nguyen, Dillon Yao, You Zhang, and Ren Ng. Synthetic defocus and look-ahead autofocus for casual videography. *arXiv preprint arXiv:1905.06326*, 2019.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018.

Chengxuan Zhu, Qingnan Fan, Qi Zhang, Jinwei Chen, Huaqi Zhang, Chao Xu, and Boxin Shi. Bokehdiff: Neural lens blur with one-step diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9508–9518, 2025.