

LensStyle: Learning the Aesthetics of Lenses for Controllable Optical Style Rendering

Yachuan Huang , Liwen Xiao , Liao Shen , Qiwen Wang , Huiqiang Sun ^{*}, Zhiyu Pan , and Zhiguo Cao 

School of AIA, Huazhong University of Science and Technology, Wuhan, China
{yachuanhuang, liwenxiao, leoshen, qiwenwang, shq1031, zhiyupan, zgcao}@hust.edu.cn

Abstract. The visual aesthetics of photographs are deeply influenced by lens characteristics such as aperture shape, optical vignetting and optical diffraction, which together define a camera’s unique optical style. Existing lens effect rendering methods primarily focus on accurately simulating the blur transition from small to large apertures but overlook the stylistic aspects of lens effects. As a result, they fail to produce diverse bokeh effects under large apertures or capture distinctive photographic phenomena such as starbursts that emerge under small apertures. In this work, we introduce LensStyle, a unified framework for controllable stylized lens effect rendering that explicitly models lens aesthetics through joint continuous–discrete control. Our model incorporates a Dual-Path Controller that disentangles continuous optical parameter modulation (*e.g.*, focus distance and blur strength) from discrete lens-style conditioning (*e.g.*, circular, polygonal, donut, cat-eye, and starburst effects), enabling fine-grained, interpretable, and physically grounded lens manipulation within a single unified framework. To support model training, we curate a comprehensive MultiLens dataset containing multi-lens image pairs synthesized under real optical constraints. Extensive experiments demonstrate that LensStyle achieves superior realism, controllability, and aesthetic quality compared with existing lens effect rendering approaches and diffusion-based image editing models, advancing computational photography toward multiple-lens-style simulation.

Keywords: Lens Effect Rendering · Controllable Synthesis · Aesthetics of Lens

1 Introduction

The visual aesthetics of a photograph are profoundly influenced by the optical characteristics of the camera lenses. Factors such as aperture geometry, optical vignetting, and diffraction patterns determine how light is transmitted and scattered, collectively defining a camera’s unique lens style. Different optical designs lead to distinctive perceptual outcomes. A circular or polygonal aperture

^{*} Corresponding author.

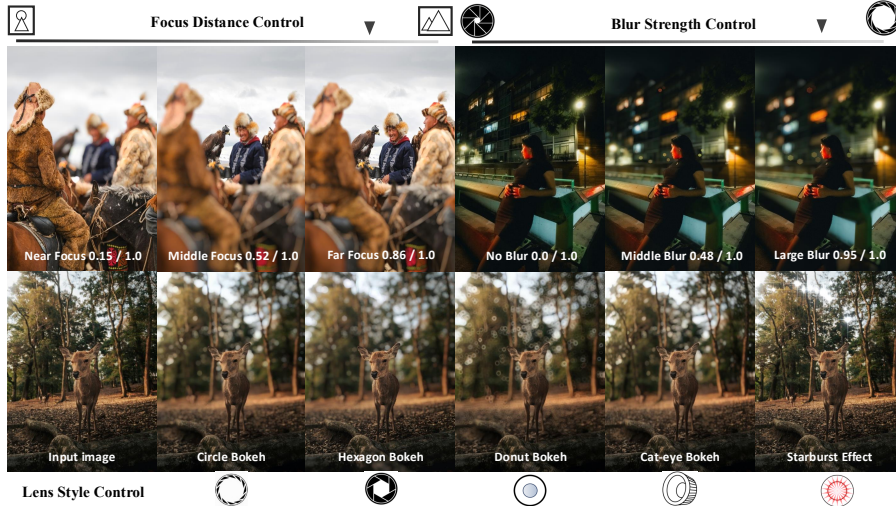


Fig. 1: Controllable multi-style lens effect rendering with LensStyle. Our framework enables unified control over three complementary dimensions of lens aesthetics within a single model. *Top-left*: continuous focus distance control, where the focal plane smoothly shifts from near to far. *Top-right*: continuous blur strength control, adjusting the magnitude of defocus under fixed focus. *Bottom*: discrete lens style control, synthesizing diverse aperture-induced effects including circular, hexagonal, donut, cat-eye bokeh, and diffraction-driven starburst patterns. These results demonstrate that LensStyle jointly models continuous optical parameters and discrete lens geometries in a physically grounded and interpretable manner.

produces distinctive bokeh shapes under large apertures; spherical aberration or mirror-based lens designs give rise to donut-shaped bokeh with hollow highlights; optical vignetting along the image periphery introduces the characteristic cat-eye effect; and small-aperture diffraction generates starburst patterns radiating from bright point sources. Such phenomena not only enhance image realism but also constitute the artistic signature of individual lenses and camera systems, playing a vital role in photographic aesthetics and visual perception.

Recent advances in lens effect rendering have primarily focused on bokeh synthesis. Methods such as classical and neural bokeh rendering [22, 27, 37, 38, 43, 44, 49] or diffusion-driven bokeh generation [6, 12, 29, 50, 55] can simulate the blur transition from small to large apertures, enabling synthetic shallow depth-of-field from all-in-focus inputs, typically with uniform, circular bokeh patterns. In other words, these approaches primarily model the *amount* of blur while neglecting its *style*. In practice, the appearance of bokeh is not only uniform, it also depends on the number of aperture blades, the shape of the aperture and optical aberrations, which jointly determine the texture, shape, and highlight distribution within the defocus region. Moreover, images captured with a small aperture often exhibit aesthetically pleasing starburst effect [20], where radiating

diffraction spikes emerge from light diffraction at the aperture edges—an effect that existing lens rendering approaches have yet to address.

To bridge the gap, we introduce **LensStyle**, a unified flow-based framework for controllable lens effect rendering that enables joint control over continuous optical parameters and diverse lens styles, explicitly modeling the aesthetic characteristics of real-world lenses. Unlike prior works that produce only a single bokeh style, LensStyle models lens behavior as a learnable style space, capturing diverse optical phenomena ranging from circular, polygonal, donut and cat-eye bokeh to diffraction-induced starbursts. At the core of our framework lies a **Dual Path Controller**, which disentangles continuous optical control from discrete lens-style modulation. For continuous attributes such as focus distance and blur strength, we introduce a *Continuous Optical Parameter Modulation* branch. User-specified optical parameters are first encoded through an optics adapter into optical parameter embeddings, which are then transformed into adaptive scaling and shifting factors via lightweight MLPs. These factors modulate intermediate image features in the flow matching network, enabling precise and smooth control over focal plane and defocus magnitude adjustment. In parallel, discrete lens-style categories are handled by a *Discrete Lens Style Cross-Attention* branch. The selected lens style is encoded into text embeddings and injected into the generative backbone via cross-attention, guiding the model to synthesize style-specific optical characteristics such as aperture geometry, vignetting-induced cat-eye distortion, or diffraction-driven starbursts. By decoupling continuous optical parameter modulation from discrete style conditioning, LensStyle achieves fine-grained, interpretable, and unified control over a wide spectrum of lens effects within a single flow-based generative model.

To enable comprehensive lens modeling, we present the MultiLens dataset, a large-scale benchmark containing physically grounded multi-lens synthetic pairs with optical parameters (focus distance and blur strength). As the first dataset designed for unified lens-style rendering, MultiLens provides structured supervision for diverse optical effects, including circular, polygonal, donut, and cat-eye bokeh, along with diffraction-induced starbursts.

Our contributions can be summarized as follows:

- We propose **LensStyle**, a unified flow-based framework for controllable lens effect rendering that jointly models continuous optical parameters and discrete lens-style categories within a single generative architecture.
- We introduce a **Dual Path Controller**, which disentangles continuous optical parameter modulation from discrete lens-style conditioning, enabling fine-grained, interpretable, and photorealistic control over focus distance, blur strength, and diverse optical styles.
- We construct the **MultiLens** dataset, a large-scale, physically grounded lens effect benchmark that offers continuous control over focus distance and blur strength. More importantly, it spans multiple lens styles, covering circular, polygonal, donut, cat-eye bokeh, and diffraction-induced starburst effects.

2 Related Work

2.1 Lens Effect Rendering

Existing lens effect rendering methods mainly focus on generating shallow depth-of-field bokeh images from all-in-focus inputs. Early traditional approaches [43, 49, 53] rely on depth maps and spatially-varying blur kernels to approximate large-aperture effects. Subsequent work, such as Dr.Bokeh [38], introduces an improved compositing formulation that alleviates artifacts under complex occlusion conditions. With the advent of deep learning, depth-dependant neural rendering approaches [22, 24, 27, 28, 44, 47] leverage depth maps and neural networks to produce more physics-aware bokeh effects, while depth-free method bokehlicious [37] synthesizes bokeh patterns based on a large-scale camera-captured paired dataset. More recently generative models [12, 29, 39, 42, 45, 55] further advance the field by leveraging the strong priors from diffusion or flow-based generators and design controllable architectures to achieve bokeh rendering. Beyond general bokeh rendering, several works have explored depth-of-field extension [23] or starburst simulation [20]. However, existing bokeh rendering methods can only reproduce a single type of bokeh pattern, typically the standard circular bokeh or the dataset-specific style, while neglecting the stylistic and aesthetic characteristics of real lenses. In this paper, we propose a unified framework that generates diverse lens-style images and controls continuous focus distance and blur strength.

2.2 Generative Models

Recent diffusion models have demonstrated remarkable success in both unconditional image synthesis [26, 31, 34, 41] and controllable image generation [7, 33, 51]. Representative works such as DDPM [10] and Stable Diffusion [32] enable high-quality generation from textual prompts through iterative denoising. Building on these, numerous diffusion-based editing approaches have emerged, including InstructPix2Pix [3], SDEdit [25], and SuperEdit [18], which support fine-grained semantic modification through conditional diffusion steps or prompt-based control. In parallel, Flow matching [4, 17, 19] and rectified flow [2, 21] models provide an alternative paradigm to diffusion by learning deterministic mappings between data and noise distributions. Unlike diffusion models that rely on iterative denoising, flow-based models directly predict continuous-time velocity fields, enabling highly efficient one-step or few-step sampling with significantly higher efficiency [8, 9, 35, 36]. Recent works [11, 16, 54] extend this formulation to image editing tasks, achieving competitive quality with much lower computational cost. In this work, we adopt a flow-based generative formulation as the foundation of our LensStyle framework. This design allows us to model lens-style rendering as a continuous transport from an all-in-focus image to a stylistically one, efficiently capturing both spatial coherence and optical consistency.

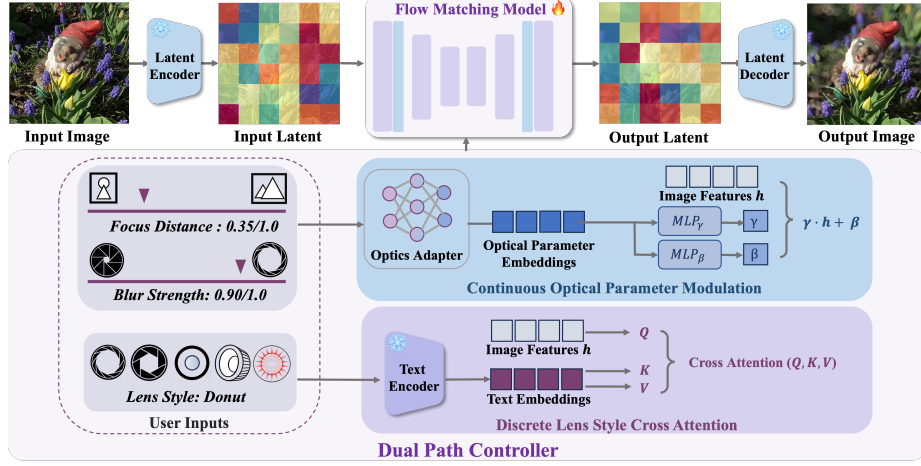


Fig. 2: Overview of LensStyle. The input image is first encoded into latent by a VAE encoder and then processed by a flow matching network for lens effect rendering. A Dual Path Controller conditions the generative backbone through two complementary branches. The *Continuous Optical Parameter Modulation* branch encodes user-specified optical parameters (e.g., focus distance and blur strength) into embeddings that generate adaptive scaling and shifting factors, which modulate intermediate image features for precise continuous control. In parallel, the *Discrete Lens Style Cross-Attention* branch encodes the selected lens style into text embeddings and injects style-specific information into the network via cross-attention. The denoised latent is finally decoded by the VAE decoder to produce the stylized output image.

3 Method

3.1 LensStyle Framework

We aim to model the *aesthetic qualities of lenses*, which arise from underlying optical properties such as aperture geometry, focal configuration, vignetting, and diffraction.

As illustrated in Figure 2, our proposed **LensStyle** adopts a flow-based generative architecture composed of three key components: (1) a *latent autoencoder* to encode all-in-focus images into a perceptually aligned latent space and a *latent decoder* to reconstruct images from the latent space; (2) a *Dual Path Controller* that enables joint continuous-discrete optical control; and (3) a *flow-based rendering model* that transforms latent representations of all-in-focus images into corresponding styled ones.

Unlike prior lens rendering approaches that explicitly require disparity or depth maps as additional inputs [27, 28, 38, 55], LensStyle operates in a depth-free manner [29, 37] during training and inference. This avoids dependence on potentially unreliable depth estimation in challenging real-world scenarios. Depth is used only during dataset synthesis (Section 3.3) to ensure physically consistent

supervision. Through learning depth-consistent blur transitions during training, the model implicitly internalizes scene geometry within its latent representation, enabling continuous focus and blur control without explicit depth conditioning.

Let I_A and I_L denote the all-in-focus and lens-style images, respectively. We first obtain latent codes $z_A = \mathcal{E}(I_A)$ and $z_L = \mathcal{E}(I_L)$ using an encoder $\mathcal{E}$.

Inspired by the success of the flow matching paradigm in image generation [4, 5, 19, 21], we directly learn a deterministic vector field N_{fm} that maps z_A to z_L in our method:

$$\mathcal{L}_{FM} = \mathbb{E}_t \|N_{fm}(\phi_t(z_L), z_A, z_C, z_D, t) - v_t(\phi_1(z_L))\|_2^2, \quad (1)$$

where $\phi_t(z_L) = (1-t)z_A + tz_L$ defines the linear interpolation between input z_A and target z_L , z_C denotes continuous optical parameter embedding, and z_D is the discrete lens style conditioning embedding from the Dual Path Controller. The objective $v_t(\phi_1(z_L)) = \phi_1(z_L) - \phi_t(z_L)$ enables the flow matching network to directly predict the residual vector for linear transport from the current distribution to the target distribution, at varying scales.

3.2 Dual Path Controller

The Dual Path Controller enables unified yet disentangled control over (i) continuous optical parameters and (ii) discrete lens-style categories. Instead of treating lens rendering as purely semantic editing, we explicitly separate physically grounded parameter modulation from style-specific feature conditioning.

Continuous Optical Parameter Modulation. Continuous optical attributes, such as focus distance and blur strength, are encoded into optical parameter embeddings $\mathbf{p}$ via an optics adapter. Specifically, the adapter first normalizes the user-specified focus distance and blur magnitude, and then applies Fourier feature encoding to transform them into a high-dimensional embedding vector $\mathbf{p}$.

Given an embedding $\mathbf{p}$, we generate adaptive scaling and shifting coefficients $\gamma(\mathbf{p})$ and $\beta(\mathbf{p})$ through lightweight MLPs, and modulate intermediate features h of the flow backbone as:

$$h' = \gamma(\mathbf{p}) \odot h + \beta(\mathbf{p}). \quad (2)$$

This mechanism enables smooth and interpretable control over defocus magnitude and focal plane adjustment, while preserving spatial coherence of the input image.

Discrete Lens Style Cross-Attention. Discrete lens-style categories (e.g., circular, polygonal, donut, cat-eye, starburst) are encoded into style embeddings z_T using a text encoder.

These embeddings are injected into the flow backbone through cross-attention:

$$\text{CrossAttn}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V, \quad (3)$$

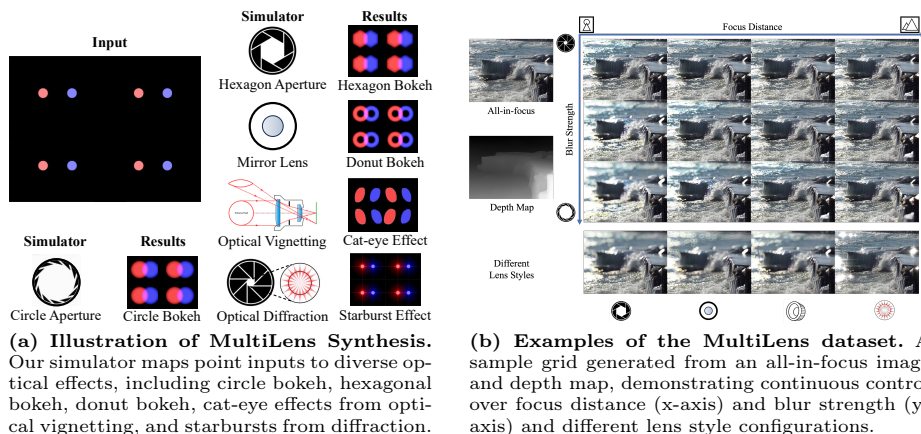


Fig. 3: MultiLens Dataset Overview

with $Q = W_Q h'$, $K = W_K z_T$, and $V = W_V z_T$.

This branch governs high-level stylistic characteristics such as aperture geometry, vignetting-induced distortion, and diffraction spike patterns, ensuring consistent rendering of discrete lens aesthetics.

By disentangling continuous optical modulation from discrete style conditioning, the Dual Path Controller enables fine-grained, interpretable, and physically grounded lens effect rendering within a single generative model.

3.3 MultiLens Dataset Synthesis

To enable learning of diverse and physically grounded optical styles, we construct the **MultiLens** dataset using a unified optical forward model that jointly simulates defocus blur and diffraction effects.

The all-in-focus images I_A are sourced from ReDWeb [46], a real-world monocular depth benchmark that provides ground-truth depth maps $Z(p)$. The availability of accurate depth enables physically consistent synthesis of spatially varying defocus, producing realistic blur transitions during training. Notably, depth is used only for data synthesis. At inference time, our model does not require depth as input, making depth-free lens effect rendering feasible through physically grounded supervision.

Unified Optical Forward Model. Let $A(x, y)$ denote the aperture transmission function defined on the aperture plane, where (x, y) are aperture-plane coordinates and $A(x, y) \in \{0, 1\}$ indicates whether light passes through the aperture at location (x, y) . Under the thin-lens approximation, image formation is modeled as the combination of (i) defocus blur described by geometric optics and (ii) diffraction effects described by wave optics.

Defocus PSF. The defocus point spread function (PSF) is proportional to the projected aperture:

$$B(x, y) = \frac{A(x, y)}{\sum_{x', y'} A(x', y')}, \quad (4)$$

where $B(x, y)$ is a normalized blur kernel satisfying energy conservation.

Diffraction PSF. To synthesize starburst effects under small apertures, we adopt Fraunhofer diffraction:

$$D(u, v) = |\mathcal{F}\{A(x, y)\}|^2, \quad (5)$$

where $\mathcal{F}$ denotes the 2D Fourier transform and (u, v) are spatial-frequency coordinates proportional to angular deviations on the sensor plane. This formulation yields diffraction spikes whose count and orientation are determined by the aperture blade number N_b and rotation θ_0 .

Aperture Functions for Multi-style Bokeh. Different bokeh styles are modeled by different analytical forms of the aperture transmission $A(x, y)$.

(1) *Circular aperture.* For a circular aperture with radius r_0 :

$$A_{\text{circ}}(x, y) = \begin{cases} 1, & x^2 + y^2 \leq r_0^2, \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

(2) *Polygonal aperture.* For a regular N_b -sided polygon aperture rotated by θ_0 :

$$A_{\text{poly}}(x, y) = \begin{cases} 1, & \forall k, x \cos \theta_k + y \sin \theta_k \leq r_0 \cos\left(\frac{\pi}{N_b}\right), \\ 0, & \text{otherwise,} \end{cases} \quad (7)$$

where $\theta_k = \frac{2\pi k}{N_b} + \theta_0$ specifies the orientation of each blade edge.

(3) *Donut aperture.* To model mirror-lens or strong spherical-aberration cases, we use a ring-shaped aperture:

$$A_{\text{donut}}(x, y) = \begin{cases} 1, & r_1^2 \leq x^2 + y^2 \leq r_2^2, \\ 0, & \text{otherwise,} \end{cases} \quad (8)$$

where r_2 is the outer radius and $r_1 = \mu r_2$ controls the inner void with $\mu \in [0, 1)$.

(4) *Cat-eye aperture.* The cat-eye effect caused by vignetting under off-axis viewing is approximated by an elliptically clipped aperture:

$$A_{\text{cat}}(x, y) = \begin{cases} 1, & \frac{(x - \delta_x)^2}{a^2} + \frac{y^2}{b^2} \leq 1, \\ 0, & \text{otherwise,} \end{cases} \quad (9)$$

where δ_x models the off-axis shift, and a, b are ellipse axes with $b/a < 1$ controlling lateral compression.

Rendering with Depth-consistent Defocus and Diffraction. Given I_A and depth map $Z(p)$, we define a spatially varying defocus magnitude:

$$K(p) = \alpha |Z(p) - Z_f|, \quad (10)$$

where Z_f is the focal depth and α controls blur strength. The final stylized image is synthesized by spatially varying convolution with a kernel scaled by $K(p)$:

$$I_L(p) = \sum_q I_A(q) \cdot B_{K(p)}(p - q) + \lambda_d \cdot H(I_A)(p) \cdot D(p), \quad (11)$$

where $B_{K(p)}$ denotes the defocus PSF whose effective radius is proportional to $K(p)$, $H(I_A)$ extracts high-intensity regions (e.g., saturated highlights) to trigger visible diffraction spikes, $D(p)$ is the diffraction response projected onto the image plane, and λ_d controls diffraction strength.

Dataset Composition. MultiLens comprises 303,000 paired samples derived from 3,000 all-in-focus scenes. The dataset covers (1) 300,000 defocus pairs, generated by varying 4 aperture geometries $\{A_{\text{circ}}, A_{\text{poly}}, A_{\text{donut}}, A_{\text{cat}}\}$ across 5 focus distances Z_f and 5 blur strengths α ; and (2) 3,000 diffraction pairs capturing starburst effects by modulating the diffraction parameter λ_d . This structured composition ensures a comprehensive coverage of physically grounded lens effects with precisely controlled optical parameters.

3.4 Optimization.

We supervise the model with a combination of flow consistency and perceptual reconstruction losses. The overall objective is:

$$\mathcal{L} = \mathcal{L}_{FM} + \lambda_1 \mathcal{L}_{rec} + \lambda_2 \mathcal{L}_{perc}, \quad (12)$$

where $\mathcal{L}_{FM}$ is the flow matching loss described in Eq. (1), $\mathcal{L}_{rec} = \|D(\hat{z}_L) - I_L\|_1$ enforces pixel-level fidelity via the decoder D , $\mathcal{L}_{perc}$ is a perceptual loss computed using a pretrained VGG network [40] to align high-level aesthetics.

4 Experiments

4.1 Experimental Settings

Baselines and Metrics. For the standard circular bokeh, we adopt several open-source state-of-the-art models as baselines, including MPIB [28], BokehMe [27], Dr.Bokeh [38], and BokehDiff [55]. Since our work is the first to explicitly consider bokeh style, no existing bokeh rendering methods can directly generate polygonal, donut, or cat-eye bokeh effects. Therefore, we further compare our approach with state-of-the-art instruction-based image editing methods, including UltraEdit [54] and SuperEdit [18]. As they do not take lens parameters into

consideration, we further retrain SuperEdit and UltraEdit on our MultiLens dataset to enable lens style editing, denoted as SuperEdit* and UltraEdit*, for fair comparison. Training details can be found in our appendix. For the starburst effect, although prior work [20] has explored the physical simulation of starbursts, it has not released its implementation. Hence, we also employ image editing models for comparison in this setting. Following prior work [27, 55], we report the PSNR and SSIM scores. Since PSNR is insensitive to blurring [38], we additionally include LPIPS [52], a perceptual metric that better aligns with human visual perception.

Implementation Details. Training is performed on the MultiLens dataset at a resolution of 512×384 for 10K iterations using the Adam optimizer [14]. The initial learning rate is set to 3×10^{-5} and linearly decayed to 3×10^{-7} after 3K iterations. We adopt an $8\times$ downsampling VAE [15] that produces a 4-channel latent representation. The flow-based backbone is implemented as a conditional U-Net with 8 input channels and 4 output channels, and is initiated from pretrained SD2.1 [32] using the Diffusers library. In the Continuous Optical Parameter Modulation branch, both MLP_γ and MLP_β are implemented as 3-layer multilayer perceptrons with SiLU activations. For the Discrete Lens Style Cross-Attention branch, we employ a pretrained CLIP text encoder [30] to extract style embeddings. The loss coefficients in Eq. (12) are set to $\lambda_1 = 0.2$ and $\lambda_2 = 0.5$. For diffusion-based UltraEdit and SuperEdit, the number of inference steps is set to 50, with the image guidance scale and text guidance scale configured to 1.5 and 7.5, respectively. All experiments are conducted on the NVIDIA RTX A6000 GPU.

4.2 Results and Comparisons

Quantitative Comparison. As shown in Table 2, for circular bokeh rendering, LensStyle consistently outperforms all neural approaches on both synthetic **MultiLens** and the real-world **EBB400** [13] dataset, achieving the highest PSNR and SSIM scores and the lowest LPIPS scores. Although the SSIM is lower than that of diffusion-based BokehDiff, our method achieves better performance in terms of LPIPS, which is more insensitive to blur. More importantly, quantitative comparisons of other lens styles are reported in Table 1. Existing diffusion-based image editing models (e.g., SuperEdit [18], UltraEdit [54]) fail to produce physically accurate and stylistically consistent effects, as they lack knowledge of lens geometry, vignetting, and diffraction. After being fine-tuned on our MultiLens dataset, their performance improves but still fall short of ours due to the lack of lens-aware design. In contrast, our model achieves a significant improvement of over 1.1–1.8 dB in PSNR and 15–25% relative reduction in LPIPS, indicating better perceptual fidelity and aesthetic consistency. This validates the importance of the proposed Dual Path Controller and its continuous-discrete conditioning mechanism in capturing both the optical parameters and semantic aspects of real lens styles.

Table 1: Quantitative comparison across polygonal, donut, cat-eye, and starburst styles on MultiLens dataset. Higher PSNR/SSIM and lower LPIPS are better. LensStyle constantly outperforms diffusion-based image editing methods.

Method	Hexagon			Donut			Cat-eye			Starburst		
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
UltraEdit [54]	14.57	0.3808	0.5971	14.52	0.3845	0.5504	14.43	0.3790	0.5833	14.32	0.3785	0.5821
SuperEdit [18]	17.27	0.4900	0.4320	16.24	0.4608	0.4180	17.58	0.4981	0.4043	17.42	0.4975	0.4021
UltraEdit* [54]	19.11	0.6109	0.3240	19.21	0.5977	0.2811	18.60	0.5738	0.3436	18.67	0.5725	0.3445
SuperEdit* [18]	<u>22.77</u>	<u>0.7230</u>	<u>0.1835</u>	<u>22.27</u>	<u>0.6723</u>	<u>0.1871</u>	<u>22.73</u>	<u>0.7106</u>	<u>0.1692</u>	<u>22.69</u>	<u>0.7118</u>	<u>0.1652</u>
Ours	23.38	0.7389	0.1718	23.09	0.6989	0.1700	23.41	0.7300	0.1652	23.45	0.7312	0.1641

Table 2: Quantitative comparison of circular bokeh rendering on synthetic MultiLens and real-world EBB400 dataset.

Method	MultiLens-Circle			EBB400 [13]		
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
MPIB [28]	21.22	0.6797	0.4284	23.08	0.7855	0.3032
BokehMe [27]	21.34	0.6749	0.4407	23.15	0.7880	0.2984
Dr.Bokeh [38]	19.83	0.6397	0.4629	22.27	0.7501	0.3510
BokehDiff [55]	<u>22.70</u>	<u>0.7642</u>	<u>0.2368</u>	<u>23.31</u>	<u>0.8029</u>	<u>0.2558</u>
Ours	23.33	0.7387	0.1743	23.58	0.7892	0.2398

Table 3: User study results indicate that users prefer ours regarding visual quality.

Comparison	Human Preference
Ours vs. SuperEdit [18]	76.8% / 23.2%
Ours vs. UltraEdit [54]	88.4% / 11.6%
Ours vs. SuperEdit* [18]	60.8% / 39.2%
Ours vs. UltraEdit* [54]	76.5% / 23.5%

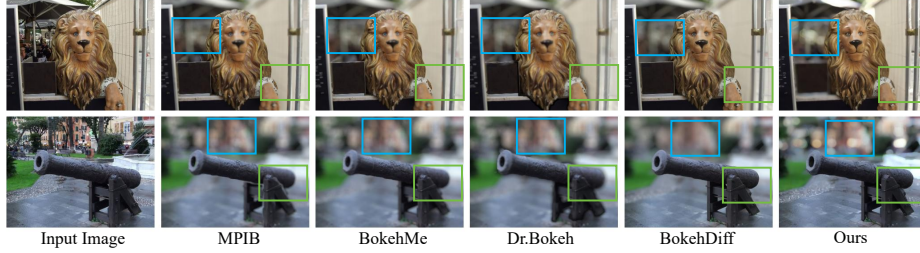


Fig. 4: Qualitative Comparison against bokeh rendering methods for standard circle bokeh. All baselines take a depth map predicted by Depth Anything V2 [48] as input. As a result, when depth map is inaccurate, the focused region of output can be blurred by mistake (green box area). Moreover, ours produces more aesthetically pleasing circle bokeh (blue box area).

Qualitative Comparison. For standard circle bokeh rendering, Figure 4 presents visual comparisons among different bokeh rendering methods. All baselines share the same depth map predicted by Depth Anything V2 [48], while ours is in a depth-free manner. As a result, when the depth input is inaccurate in some scenarios, the focused region can be mistakenly blurred, such as the green box areas in Figure 4. Furthermore, benefiting from the designed Lens Style Controller, ours produces more physically consistent and visually pleasing circular bokeh, such as blue box areas.

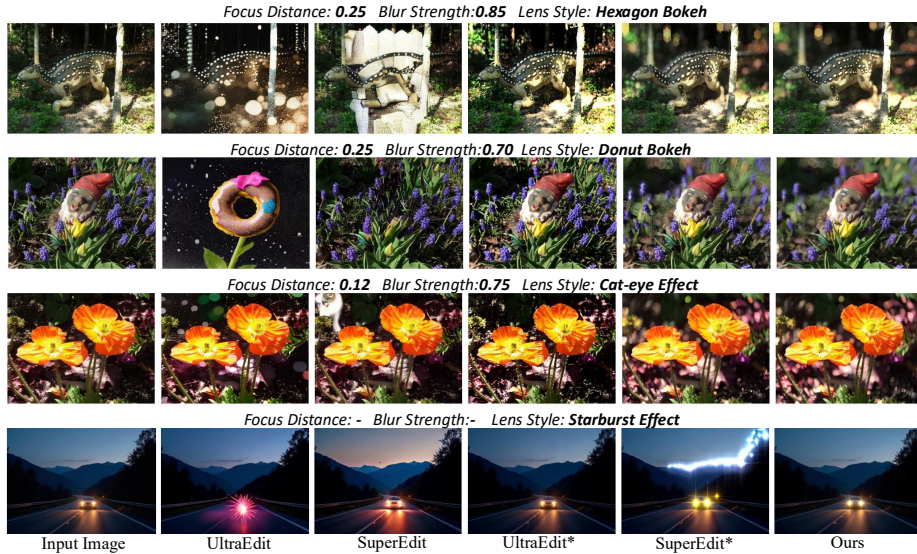


Fig. 5: Qualitative Comparison against diffusion-based image editing methods. UltraEdit and SuperEdit lack lens style awareness, resulting in outputs that do not match the specified lens style. After fine-tuning on our MultiLens dataset, these models begin to capture certain lens style characteristics. However, thanks to the proposed Dual Path Controller with its disentangled continuous-discrete conditioning mechanism, our method achieves the most consistent and visually superior results.

Figure 5 compares other lens styles with diffusion-based image editing methods. Since UltraEdit and SuperEdit lack lens-aware related knowledge, they produce outputs that are semantically inaccurate with respect to the editing instructions. After fine-tuning on our MultiLens dataset, these models begin to capture certain lens style characteristics. However, they either produce less physically consistent bokeh patterns (e.g., the hexagon bokeh in the first row of Figure 5) or over-produces starbursts (e.g., the sky in the fourth row of Figure 5). Our method still achieves the most consistent and visually superior results, benefiting from the proposed Dual Path Controller with its disentangled continuous modulation and discrete cross-attention mechanism.

4.3 User Study

Since no publicly available benchmark provides real paired images across diverse lens styles for perceptual evaluation, we conduct a user study to assess aesthetic quality and style fidelity. We collect 32 all-in-focus images from Unsplash [1], a photography platform containing real-world captured images. The selected images cover diverse content categories, including portraits, animals, indoor and outdoor scenes, as well as daytime and nighttime settings. Each image is ren-



Fig. 6: User study visual comparisons under focus distance 0.70 and blur intensity 0.80 with hexagonal aperture style. Our method produces hexagonal bokeh patterns that are geometrically consistent and physically plausible, whereas SuperEdit* [18] generates hexagons with overly sharp corner transitions.

Table 4: Ablation study on the MultiLens dataset. We evaluate the contribution of each core component in LensStyle.

Variant	Flow	COPM	DLSC	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
w/o Flow (Direct Regression)		✓	✓	21.87	0.6822	0.2433
w/o OPM	✓		✓	22.14	0.7034	0.2126
w/o SCA	✓	✓		22.06	0.6951	0.2195
Full Model	✓	✓	✓	23.38	0.7389	0.1718

dered with different methods to produce corresponding lens-style outputs under matched style prompts and optical parameters.

We perform a pairwise A/B comparison protocol. For each trial, participants are shown two rendered images generated from the same input image and style configuration. One image is always produced by our method, while the other is randomly selected from a competing approach. The presentation order is randomized to avoid bias. Participants are asked to choose the one that (1) exhibits stronger aesthetic quality, (2) better reflects the intended lens style, and (3) maintains higher consistency with the input image content.

A total of 58 participants took part in the study. As summarized in Table 3 and illustrated in Figure 6, our method is consistently preferred over competing approaches. The results demonstrate that LensStyle achieves superior perceptual realism, stylistic fidelity, and structural consistency, validating its effectiveness in multi-style lens effect rendering.

4.4 Ablation Study

We conduct ablation experiments to analyze the contribution of each key component in LensStyle, including the continuous optical modulation, discrete style conditioning, and flow-based transport.

Continuous Optical Parameter Modulation(COPM). To validate the effectiveness of adaptive feature modulation, we replace the $\gamma(\mathbf{p}), \beta(\mathbf{p})$ modulation with simple concatenation of parameter embeddings to the input features.

This modification significantly weakens smooth parameter interpolation. In particular, blur magnitude transitions become less monotonic when sweeping focus distance or blur strength. These results indicate that affine feature modulation provides more stable and interpretable continuous control than naive conditioning.

Discrete Lens Style Cross-Attention(DLSC). We further remove the cross-attention branch and inject style embeddings only through global concatenation. This leads to weaker aperture-specific characteristics, especially for complex styles such as donut bokeh and starburst diffraction. This demonstrates that cross-attention effectively governs spatially structured lens-style semantics.

Flow-Based Transport vs. Direct Regression. To examine the impact of the flow formulation, we replace the flow matching objective with a direct latent regression loss:

$$\mathcal{L}_{reg} = \|\hat{z}_L - z_L\|_2^2.$$

This variant produces less stable structural preservation and more over-smoothed outputs. Flow-based transport better maintains input content consistency while allowing large stylistic transformations, suggesting that learning a velocity field provides a more effective mapping between the two domains.

Overall, these ablation results validate the necessity of each design component in achieving controllable and realistic multi-style lens effect rendering.

5 Conclusion and Future Work

In this work, we introduced LensStyle, a unified framework for controllable multi-style lens effect rendering. Unlike conventional defocus rendering approaches that focus primarily on blur magnitude, LensStyle explicitly models lens aesthetics by jointly learning continuous optical parameter modulation and discrete lens-style conditioning within a flow-based generative architecture. By leveraging a physically grounded optical forward model for dataset synthesis, our method learns depth-consistent blur transitions while remaining depth-free during inference. The proposed Dual Path Controller disentangles continuous focus and blur control from aperture-specific stylistic characteristics, enabling interpretable and flexible manipulation of lens effects in a single model. Extensive quantitative evaluations and user studies demonstrate that LensStyle achieves improved realism, style fidelity, and controllability compared to existing lens rendering and image editing approaches.

Limitation and future work. While LensStyle effectively models diverse lens aesthetics and achieves controllable lens effect rendering, it cannot simulate bokeh effects of different DSLRs such as Fuji, Sony, Nikon and Canon. Future work will extend our model to lens of different brands.

References

1. Unsplash. <https://unsplash.com/> (2026)
2. Albergo, M.S., Vanden-Eijnden, E.: Building normalizing flows with stochastic interpolants. arXiv preprint arXiv:2209.15571 (2022)
3. 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)
4. Dao, Q., Phung, H., Nguyen, B., Tran, A.: Flow matching in latent space. arXiv preprint arXiv:2307.08698 (2023)
5. Fischer, J.S., Gui, M., Ma, P., Stracke, N., Baumann, S.A., Ommer, B.: Boosting latent diffusion with flow matching. arXiv preprint arXiv:2312.07360 (2023)
6. Fortes, A., Wei, T., Zhou, S., Pan, X.: Bokeh diffusion: Defocus blur control in text-to-image diffusion models. arXiv preprint arXiv:2503.08434 (2025)
7. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)
8. Geng, Z., Deng, M., Bai, X., Kolter, J.Z., He, K.: Mean flows for one-step generative modeling. arXiv preprint arXiv:2505.13447 (2025)
9. Geng, Z., Lu, Y., Wu, Z., Shechtman, E., Kolter, J.Z., He, K.: Improved mean flows: On the challenges of fastforward generative models. arXiv preprint arXiv:2512.02012 (2025)
10. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *Advances in neural information processing systems* **33**, 6840–6851 (2020)
11. Hu, T., Zhang, D.W., Mettes, P., Tang, M., Zhao, D., Snoek, C.G.: Latent space editing in transformer-based flow matching. In: AAAI (2024)
12. Huang, Y., Luo, X., Wang, Q., Shen, L., Li, J., Sun, H., Huang, Z., Jiang, W., Cao, Z.: Bokehflow: Depth-free controllable bokeh rendering via flow matching. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 5167–5175 (2026)
13. 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)
14. Kingma, D.P.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
15. Kingma, D.P., Welling, M., et al.: Auto-encoding variational bayes (2013)
16. Kulikov, V., Kleiner, M., Huberman-Spiegelglas, I., Michaeli, T.: Flowedit: Inversion-free text-based editing using pre-trained flow models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19721–19730 (2025)
17. Labs, B.F.: Flux. <https://github.com/black-forest-labs/flux> (2024)
18. Li, M., Gu, X., Chen, F., Xing, X., Wen, L., Chen, C., Zhu, S.: Superedit: Rectifying and facilitating supervision for instruction-based image editing. arXiv preprint arXiv:2505.02370 (2025)
19. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
20. Liu, D., Geng, H., Liu, T., Klette, R.: Star-effect simulation for photography. *Computers & Graphics* **61**, 19–28 (2016)
21. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022)

22. Luo, X., Peng, J., Xian, K., Wu, Z., Cao, Z.: Defocus to focus: Photo-realistic bokeh rendering by fusing defocus and radiance priors. *Information Fusion* **89**, 320–335 (2023)
23. Luo, X., Peng, J., Zhao, W., Xian, K., Lu, H., Cao, Z.: Point-and-shoot all-in-focus photo synthesis from smartphone camera pair. *IEEE Transactions on Circuits and Systems for Video Technology* **33**(5), 2088–2101 (2022)
24. Luo, Y., Shi, M., Shen, L., Huang, Y., Ye, Z., Peng, J., Cao, Z.: Video bokeh rendering: Make casual videography cinematic. In: *Proceedings of the 32nd ACM International Conference on Multimedia*. pp. 7677–7685 (2024)
25. 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)
26. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: *International conference on machine learning*. pp. 8162–8171. PMLR (2021)
27. 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)
28. 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)
29. Qin, X., Wang, Z., Li, F., Chen, H., Pei, R., Li, W., Cao, X.: Camedit: Continuous camera parameter control for photorealistic image editing. In: *The Thirty-ninth Annual Conference on Neural Information Processing Systems* (2025)
30. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: *International conference on machine learning*. pp. 8748–8763. PmLR (2021)
31. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: *International conference on machine learning*. pp. 8821–8831. Pmlr (2021)
32. 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)
33. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dream-booth: Fine tuning text-to-image diffusion models for subject-driven generation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 22500–22510 (2023)
34. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems* **35**, 36479–36494 (2022)
35. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512* (2022)
36. Sauer, A., Lorenz, D., Blattmann, A., Rombach, R.: Adversarial diffusion distillation. In: *European Conference on Computer Vision*. pp. 87–103. Springer (2024)
37. 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)
38. 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)
39. Shrivastava, A., Barnes, C., Zhang, X., Zhang, L., Owens, A., Amirghodsi, S., Shechtman, E.: Fine-grained defocus blur control for generative image models. In: Proceedings of the Winter Conference on Applications of Computer Vision (WACV) (2026), <https://arxiv.org/abs/2510.06215>
 40. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (ICLR) (2015)
 41. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
 42. Voynov, A., Hertz, A., Arar, M., Fruchter, S., Cohen-Or, D.: Curved diffusion: A generative model with optical geometry control. In: European Conference on Computer Vision. pp. 149–164. Springer (2024)
 43. Wadhwa, N., Garg, R., Jacobs, D.E., Feldman, B.E., Kanazawa, N., Carroll, R., Movshovitz-Attias, Y., Barron, J.T., Pritch, Y., Levoy, M.: Synthetic depth-of-field with a single-camera mobile phone. ACM Transactions on Graphics (ToG) **37**(4), 1–13 (2018)
 44. 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)
 45. Wang, Y., Chen, X., Xu, X., Liu, Y., Zhao, H.: Diffcamera: Arbitrary refocusing on images. In: Proceedings of the SIGGRAPH Asia 2025 Conference Papers. SA Conference Papers '25, Association for Computing Machinery, New York, NY, USA (2025). <https://doi.org/10.1145/3757377.3763827>, <https://doi.org/10.1145/3757377.3763827>
 46. Xian, K., Shen, C., Cao, Z., Lu, H., Xiao, Y., Li, R., Luo, Z.: Monocular relative depth perception with web stereo data supervision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 311–320 (2018)
 47. Xiao, L., Kaplanyan, A., Fix, A., Chapman, M., Lanman, D.: Deepfocus: Learned image synthesis for computational display. In: ACM SIGGRAPH 2018 Talks, pp. 1–2 (2018)
 48. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. Advances in Neural Information Processing Systems **37**, 21875–21911 (2024)
 49. Yang, Y., Lin, H., Yu, Z., Paris, S., Yu, J.: Virtual dslr: High quality dynamic depth-of-field synthesis on mobile platforms. Electronic Imaging **28**, 1–9 (2016)
 50. Yuan, Y., Wang, X., Sheng, Y., Chennuri, P., Zhang, X., Chan, S.: Generative photography: Scene-consistent camera control for realistic text-to-image synthesis. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7920–7930 (2025)
 51. 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)
 52. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
 53. 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)

- 54. Zhao, H., Ma, X.S., Chen, L., Si, S., Wu, R., An, K., Yu, P., Zhang, M., Li, Q., Chang, B.: Ultraedit: Instruction-based fine-grained image editing at scale. *Advances in Neural Information Processing Systems* **37**, 3058–3093 (2024)
- 55. 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)