

FlashReFocus: Interactive Image Refocusing in Seconds

Ching-Heng Cheng^{1*} Chia-Ming Lee^{2,3*} Ming-Ching Chang² Xin Li² Yu-Lun Liu³ Chih-Chung Hsu^{1,3†}

¹National Cheng Kung University ²State University of New York at Albany ³National Yang Ming Chiao Tung University

<https://ming0531.github.io/FlashReFocus/>

Abstract

Single-image refocusing changes the focal plane and depth of field of a photograph after capture. Users try many focal points and blur strengths on one image, so each edit should be fast and change only the blur; yet recent generative methods rerun large diffusion models for every edit. The two parts of the task differ: deblurring must recover lost detail, whereas rendering new blur is determined by depth and the focal plane. We therefore propose *FlashReFocus*, which runs diffusion once to restore an all-in-focus image and renders every edit with a lightweight network. For restoration, a one-step diffusion deblurrer is paired with a decoder refiner that recovers fine structures lost in VAE decoding. For rendering, a network conditioned on a defocus map from depth and the chosen focal point uses defocus-bias attention, which lets pixels with similar defocus exchange information regardless of their image distance; focal point and blur strength are controlled directly, and in-focus content is preserved. Editing the map also yields tilt-shift and two focal planes without retraining. We also introduce *3CReal*, a three-camera benchmark of 89 pairs. *FlashReFocus* achieves the best perceptual fidelity for zero-shot bokeh rendering on *EBB!* and *3CReal* and refocuses $148.3\times$ faster than *GenRefocus*, taking about half a second for each additional edit.

1. Introduction

Depth of field determines which parts of a photograph appear sharp and is fixed at capture. Single-image refocusing changes the focal plane and the blur strength afterwards, which requires recovering an all-in-focus (AiF) image and re-rendering blur around a new focal plane [40].

Refocusing is inherently interactive: a user tries several focal points and blur strengths on the same photograph, so each edit should be fast and should change the blur and nothing else. Its two sub-problems, however, differ in nature. Recovering the AiF image is ill-posed because defocus destroys high-frequency detail; regression-based re-

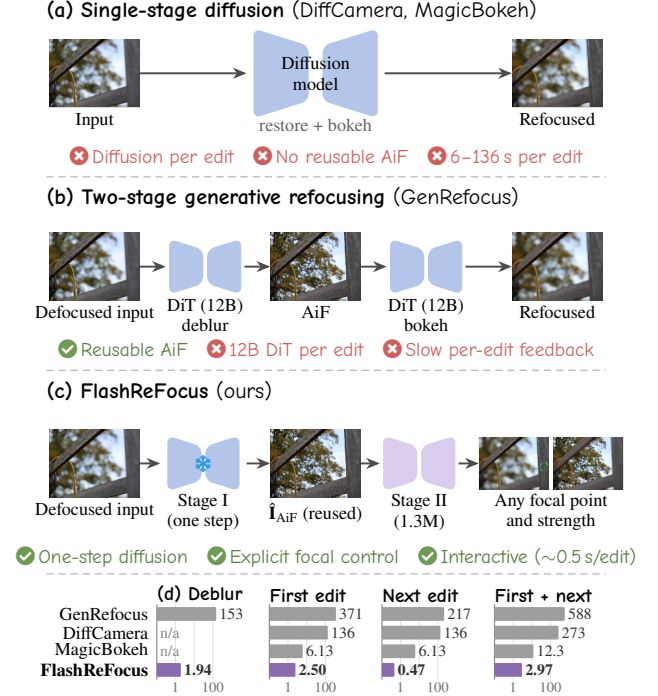


Figure 1. **Refocusing paradigms.** (a) Single-stage diffusion and (b) two-stage generative refocusing run a large diffusion model for every focal edit. (c) FlashReFocus restores the AiF image once and renders each new focal setting with a 1.3M-parameter network. (d) Seconds at 1536×1536 for deblurring, the first edit, each next edit, and both edits together.

storers [24, 71] over-smooth it [7], whereas diffusion priors [18, 45] synthesize realistic detail. Rendering new blur, in contrast, is largely determined by scene depth and the focal plane, and calls for precise, repeatable control rather than generation. Recent generative methods nevertheless rely on large diffusion models for both. DiffCamera [62] refocuses end to end with a single diffusion transformer, and MagicBokeh [53] fuses super-resolution and bokeh rendering into one step, arguing that two-stage pipelines accumulate errors and are inefficient. Such single-stage designs leave no reusable AiF image, so every edit reruns diffusion. GenRefocus [35] separates the two stages, but renders bokeh with a second 12B-parameter diffusion transformer, so each focal edit still requires a full generative pass that

*Equal contribution.

†Corresponding author: chihchung@nycu.edu.tw

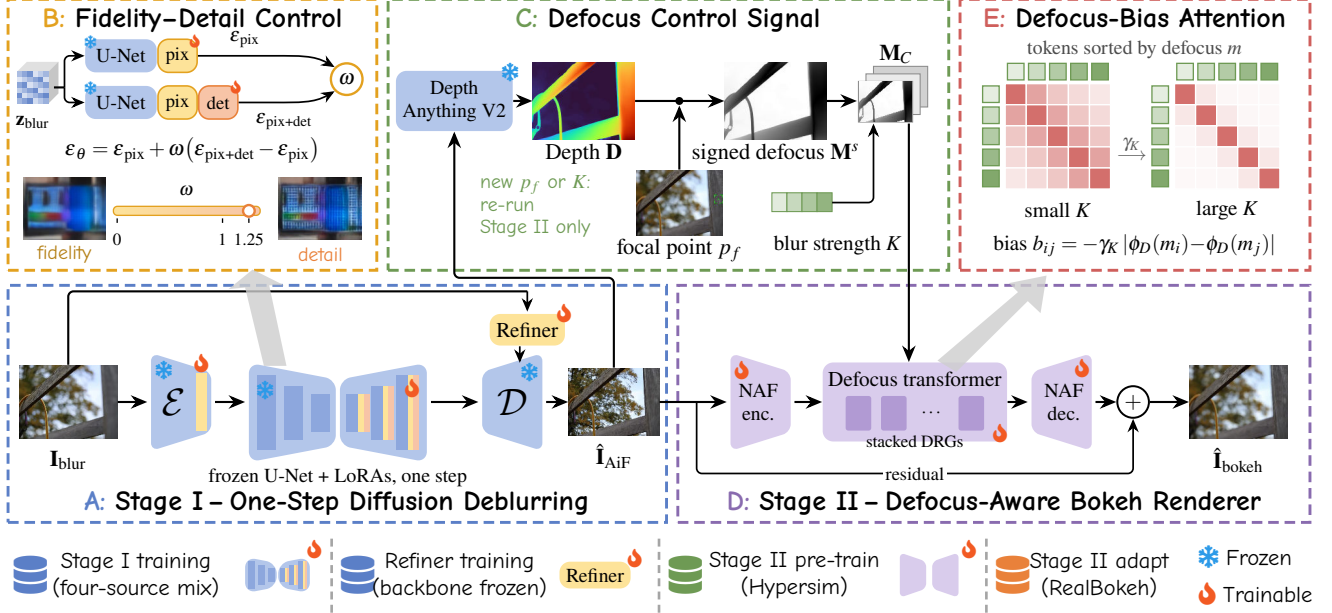


Figure 2. **Overview of FlashReFocus.** (A) Stage I restores an all-in-focus image $\hat{\mathbf{I}}_{\text{AiF}}$ in one step with a frozen latent diffusion U-Net and an optional decoder refiner; (B) ω mixes the pixel and detail LoRA passes to trade fidelity for detail. (C) Stage II computes a defocus map from monocular depth and the selected focal point, and (D) renders bokeh in a single feed-forward pass with (E) defocus-bias attention, which focuses attention on tokens of similar defocus, more sharply for a larger strength K . The bottom row shows the training phases.

also regenerates the regions meant to stay sharp and takes minutes per edit at 1536×1536 , as compared in Fig. 1.

We present FlashReFocus, shown in Fig. 2, whose design follows this asymmetry: diffusion is spent once, on restoration, and every edit is rendered explicitly. Stage I adapts a one-step diffusion restorer [56] to defocus; its two LoRA branches, one for pixel fidelity and one for detail, are blended by a single weight ω . We train it on mixed real and synthetic defocus pairs and add a decoder refiner that reinjects input features to recover structure lost in VAE decoding. Stage II is a 1.3M-parameter feed-forward renderer conditioned on a depth-derived defocus map, the focal point, and the blur strength. Its defocus-bias attention relates pixels by their defocus rather than by their image distance as in aperture-aware attention [49]. Because the renderer only adds a residual to the restored image, in-focus regions stay nearly unchanged across edits; because the defocus map is explicit, editing it yields tilt-shift and two focal planes without retraining. Diffusion thus runs once per photograph, not once per edit, and no second generative pass compounds restoration errors.

Controllable rendering also needs data with diverse focal settings. Real pairs [21, 49] capture lens-specific bokeh but few focal planes; synthetic pairs [37] offer any focal depth but simplified optics. We therefore pre-train on synthetic multi-focal scenes and adapt to real pairs. EBB! [21] uses one camera, so we collect 3CReal: 89 aperture pairs from three unseen cameras. Our contributions are:

- We propose FlashReFocus, which restores a photograph once with one-step diffusion and renders each edit with a

lightweight network, so refocusing is interactive and in-focus content stays consistent across edits.

- We adapt one-step diffusion to defocus deburring with mixed real and synthetic supervision and a decoder refiner that restores structure lost in VAE decoding.
- We design a lightweight renderer with defocus-bias attention and synthetic-to-real training, whose explicit defocus map gives focal, strength, and depth-editing control.
- We introduce 3CReal, a benchmark for bokeh rendering under unseen optics. FlashReFocus achieves the best perceptual fidelity on EBB! and 3CReal while refocusing $148.3\times$ faster than GenRefocus, at 0.47 s per further edit.

2. Related Work

Image restoration and defocus deburring. Convolutional and Transformer restorers optimize reconstruction losses [10, 29, 71] and achieve high PSNR and SSIM; the perception-distortion trade-off [7] motivates generative priors for perceptual quality. Defocus blur differs from motion or uniform blur in that its point-spread function varies with scene depth and focal distance [6, 72]. Classical methods estimate a spatially varying kernel and deconvolve [25, 74, 77], while learned models predict kernels or adaptive filters or exploit dual-pixel cues [1, 2, 24, 41, 55]. DRBNet trains on light-field-generated pairs and adapts to real images [47], and recent work models the image formation process explicitly [75].

Diffusion models provide a learned prior over natural images for ill-posed inverse problems [18, 45]. Restoration methods condition on the degraded image, predict residuals, or fuse latent priors [13, 32, 67], and recent defocus meth-

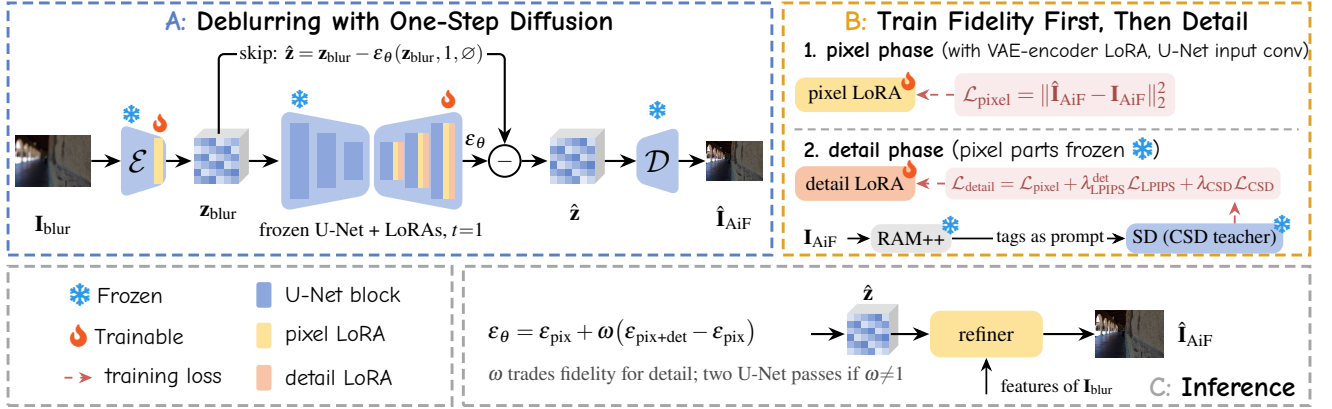


Figure 3. **Stage I deblurs in one step with separate fidelity and detail LoRAs.** (A) A frozen U-Net with both LoRAs restores the input latent. (B) The pixel LoRA, a VAE-encoder LoRA, and the input convolution are trained with $\mathcal{L}_{\text{pixel}}$ and then frozen while the detail LoRA is trained with LPIPS and CSD. (C) At inference, ω weights the detail branch and the refiner decodes $\hat{\mathbf{I}}_{\text{AiF}}$.

ods combine diffusion with Transformer features or fidelity-aware guidance [9, 26–28]. Diffusion priors can hallucinate texture [4], and VAE compression loses fine structures, which motivates our detail LoRA and decoder refiner.

Bokeh rendering and refocusing. Classical renderers convolve depth-dependent kernels, trace rays, or scatter in image space [23, 39, 40, 68], and layered or light-field inputs add scene information [36, 59]. Learned renderers fit camera-specific appearance from data. DeepLens predicts depth and renders from a single image [61]; PyNET-Bokeh learns DSLR-like blur from EBB! pairs [21]; BokehMe refines analytical scattering with a network [37]; MPIB renders inpainted multiplane images [38]; Bokehlicious adds a strength control [49]; and Dr.Bokeh models occlusion explicitly [51]. BokehDiff, MagicBokeh, and DiffCamera bring diffusion to bokeh synthesis [53, 62, 76], including depth-free rendering [52], and Any-to-Bokeh extends one-step diffusion bokeh to video with multiplane conditioning [69], while Bokeh Diffusion controls the blur level of text-to-image generation but not the focal plane [15]. Control signals range from scalar strength to masks, depth maps, and multiplane images; ours is a depth-derived signed defocus map with a global strength.

Refocusing combines restoration and rendering. RefocusGAN estimates an AiF radiance image before applying focal blur [48]; DC² uses a dual-camera rig [5]; and GenRefocus trains two diffusion transformers, DeblurNet and BokehNet, for the two operations with semi-supervised learning [35]. For text-to-image generation, EXIF-conditioned defocus control likewise generates an AiF image, its depth, and a focus distance before applying a differentiable lens blur [54]. FlashReFocus applies this decomposition to real photographs and replaces generative rendering with a lightweight feed-forward renderer that reuses the AiF image across edits.

3. Proposed Method

Preliminaries. Let $\mathbf{I}_{\text{blur}}$ be a photograph whose defocus is unknown and varies across the frame. We first estimate an AiF image and then render the requested depth of field:

$$\hat{\mathbf{I}}_{\text{AiF}} = \mathcal{F}_{\theta}(\mathbf{I}_{\text{blur}}), \quad \hat{\mathbf{I}}_{\text{bokeh}} = \mathcal{R}_{\psi}(\hat{\mathbf{I}}_{\text{AiF}}, \mathbf{M}_C), \quad (1)$$

where $\mathbf{M}_C$ collects the control signals defined in Sec. 3.2. Under a thin-lens model, a point at depth Z imaged with the lens focused at depth Z_f spreads into a circle of confusion whose diameter is proportional to

$$c(Z) \propto a \left| Z^{-1} - Z_f^{-1} \right|, \quad (2)$$

where a is the aperture diameter. We therefore condition the renderer on a spatial defocus map derived from depth, and use a diffusion prior for the ill-posed inverse map $\mathcal{F}_{\theta}$. The explicit AiF intermediate lets each stage be trained separately and is reused across focal settings.

3.1. Stage I: One-Step Diffusion Deblurring

Figure 3 illustrates Stage I. Defocus removes detail that must be re-synthesized, so we keep a diffusion prior but apply it in a single step. Following PiSA-SR [56], a frozen Stable Diffusion U-Net ϵ_{θ} maps the VAE latent $\mathbf{z}_{\text{blur}} = \mathcal{E}(\mathbf{I}_{\text{blur}})$ directly to that of the sharp image,

$$\hat{\mathbf{z}} = \mathbf{z}_{\text{blur}} - \epsilon_{\theta}(\mathbf{z}_{\text{blur}}, 1, \emptyset), \quad \hat{\mathbf{I}}_{\text{AiF}} = \mathcal{D}(\hat{\mathbf{z}}), \quad (3)$$

with VAE decoder $\mathcal{D}$, timestep 1, and an empty prompt, so inference needs no tagger, caption, or iterative sampling.

Pixel and detail LoRAs. Pixel losses over-smooth and generative losses hallucinate [7], so we separate fidelity from detail with two LoRA branches [19] trained in turn. In the pixel phase, a rank-16 pixel LoRA, together with a VAE-encoder LoRA and the U-Net input convolution, learns to remove blur with $\mathcal{L}_{\text{pixel}} = \|\hat{\mathbf{I}}_{\text{AiF}} - \mathbf{I}_{\text{AiF}}\|_2^2$ to the sharp target. These are then frozen, and a rank-4 detail LoRA (the semantic LoRA of PiSA-SR) learns the remaining detail with

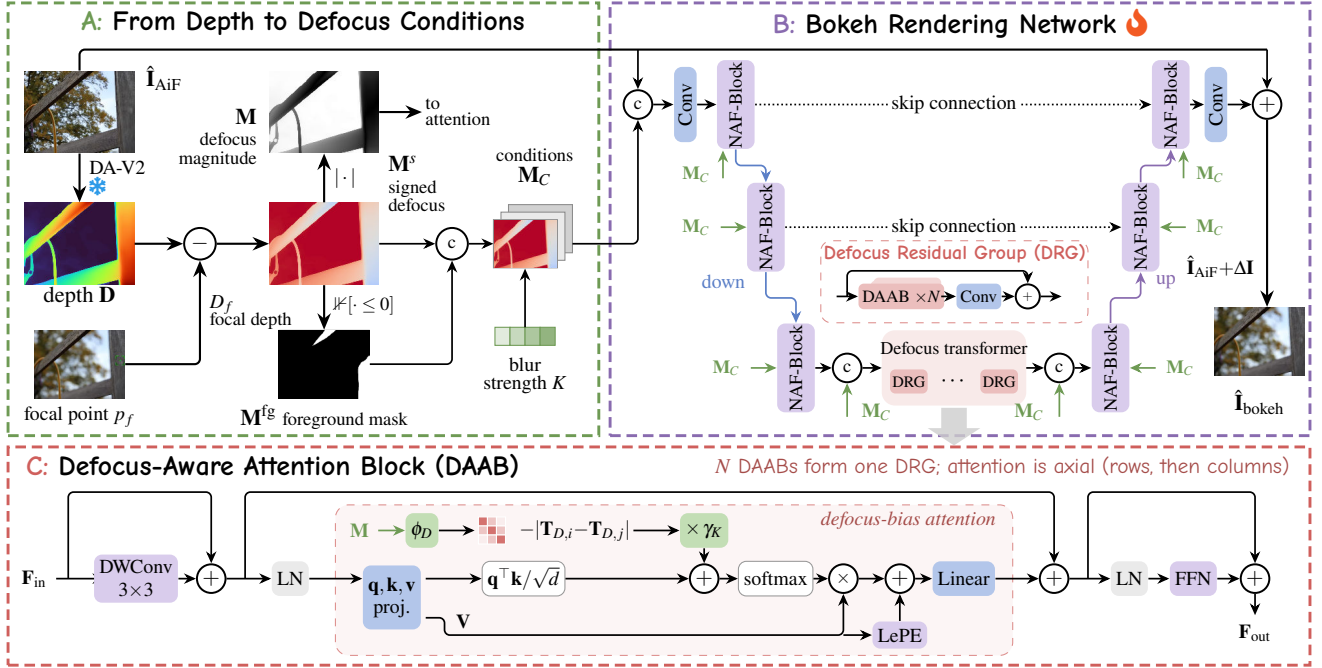


Figure 4. **Stage II renders bokeh from explicit defocus with a lightweight network.** (A) Depth, focal point, and strength form the condition stack M_C . (B) A NAF encoder-decoder predicts a residual from $\hat{I}_{AiF}$ and M_C ; its bottleneck stacks defocus residual groups (DRGs) of N defocus-aware attention blocks (DAABs). (C) Each DAAB biases attention by strength-scaled defocus differences.

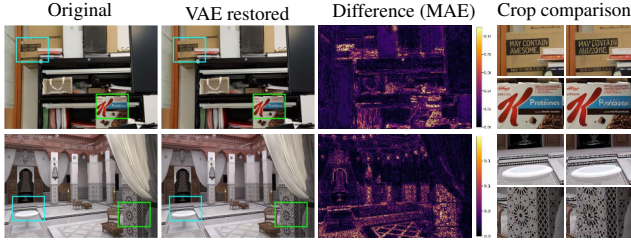


Figure 5. **What the VAE decoder loses.** Encoding and decoding with the original VAE weaken text and texture-rich regions, as the error maps and enlarged crops show, motivating our refiner.

$$\mathcal{L}_{\text{detail}} = \mathcal{L}_{\text{pixel}} + \lambda_{\text{LPIPS}}^{\text{det}} \mathcal{L}_{\text{LPIPS}} + \lambda_{\text{CSD}} \mathcal{L}_{\text{CSD}}, \quad (4)$$

with LPIPS [73] and weights λ . Classifier score distillation (CSD) [70] pulls outputs toward sharp natural images along the guidance direction of the frozen Stable Diffusion model, prompted by RAM++ tags of the target [20] with defocus terms in the negative prompt. With ϵ_{pix} and $\epsilon_{\text{pix+det}}$ the U-Net predictions with the pixel LoRA alone and with both LoRAs, ϵ_{θ} in Eq. (3) becomes

$$\epsilon_{\theta} = \epsilon_{\text{pix}} + \omega(\epsilon_{\text{pix+det}} - \epsilon_{\text{pix}}), \quad (5)$$

where the detail weight ω trades fidelity ($\omega = 0$) for detail. We use $\omega = 1.25$, selected on held-out data; any $\omega \neq 1$ costs a second U-Net pass.

Mixed supervision. Accurately paired real defocus data are scarce, and existing sets trade scale, realism, and blur strength [50]. We therefore draw pairs from four complementary sources. (A) DPDD [1] gives real lens blur, but from one camera at a mild $f/4$. (B) RealBokeh [49], reversed into deblurring pairs as in RealDefocus [50], adds

strong real blur ($f/2.0-f/20$); as its $f/22$ targets can retain slight defocus, we keep the 1,000 scenes with the sharpest targets (4,179 pairs). (C) LFDof [46] adds pixel-aligned pairs, but its blur is rendered from light fields. (D) Synthetically blurred DIV2K [3] images add exact pixel correspondence and spatially varying blur of known shape:

$$\mathbf{I}_{\text{blur}}(p) = \sum_{k=1}^3 w_k(p) (\mathcal{K}_{r_k} * \mathbf{I}_{AiF})(p), \quad (6)$$

where $\mathcal{K}_r$ is a disk, Gaussian, or motion kernel of radius r , chosen per image to diversify kernel shapes, and the weights w_k interpolate three blur radii along a random gradient map $\mathbf{M}_s$. Sources are drawn with probabilities 0.3, 0.2, 0.3, 0.2. **The decoder refiner.** To recover text and thin edges lost in VAE decoding, a decoder-side *refiner* fuses multi-scale encoder features of $\mathbf{I}_{\text{blur}}$, guided by a Laplacian high-frequency map of the input, into the VAE decoder at matching scales. The fusion modules and the decoder are trained on Stage-I latents with

$$\mathcal{L}_{\text{ref}} = \|\hat{\mathbf{I}}_{\text{ref}} - \mathbf{I}_{AiF}\|_2^2 + \lambda_{\text{struct}} \mathcal{L}_{\text{struct}} + \lambda_{\text{LPIPS}}^{\text{ref}} \mathcal{L}_{\text{LPIPS}}, \quad (7)$$

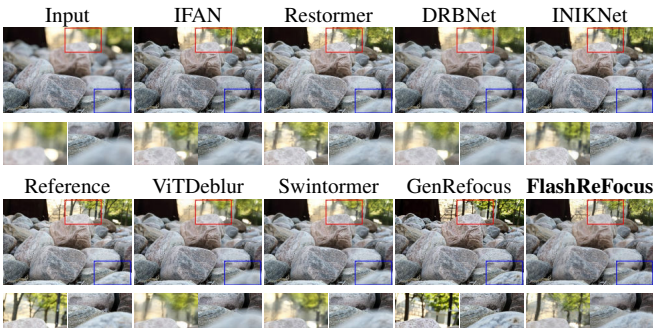
where $\hat{\mathbf{I}}_{\text{ref}}$ is the refined output and $\mathcal{L}_{\text{struct}}$ is the ℓ_1 error after a per-image affine color fit, which ignores global color shifts and penalizes structural errors. The refiner is optional and trades perceptual sharpness for fidelity (Sec. 4.1).

3.2. Stage II: Defocus-Aware Bokeh Rendering

As shown in Fig. 4, the renderer edits the restored AiF image under a depth-derived control signal, with defocus-bias attention and synthetic-to-real training.

Table 1. **Defocus deblurring on DPDD and RealDOF**. Darker green marks better results (best, second, and third per dataset). FlashReFocus is shown without the refiner at $\omega = 1$ and with the refiner at the default $\omega = 1.25$.

Methods	Full-reference Metrics				No-reference IQA Metrics					Distribution
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	DISTS $\downarrow$	NIQE $\downarrow$	MUSIQ $\uparrow$	MANIQA $\uparrow$	BRISQUE $\downarrow$	CLIPQA $\uparrow$	FID $\downarrow$
<i>DPDD</i> [1]										
IFAN (CVPR'21) [24]	25.3637	0.7953	0.2158	0.1331	4.7407	59.3878	0.3522	34.7442	0.4026	40.4626
Restormer* (CVPR'22) [71]	25.9768	0.8157	0.1762	0.1204	4.9868	61.7751	0.3887	39.3745	0.4405	32.0235
DRBNet (CVPR'22) [47]	25.7217	0.7978	0.1820	0.1063	4.5813	59.6612	0.3576	22.9788	0.4373	37.8632
INIKNet (ICCV'23) [42]	26.1000	0.8102	0.1838	0.1234	4.5837	59.3054	0.3505	32.9875	0.4092	35.4650
ViTDeblur (TCL'24) [30]	26.1209	0.8187	0.2009	0.1303	4.7617	60.7682	0.3660	34.4846	0.4037	38.8199
Swintormer (ICASSP'25) [9]	25.8700	0.8201	0.1911	0.1228	4.6015	60.7700	0.3661	37.3643	0.4462	33.1271
GenRefocus (arXiv'25) [35]	23.9342	0.7230	0.1686	0.0843	4.1277	63.4728	0.4449	22.5078	0.5290	31.0665
FlashReFocus	24.6061	0.7347	0.1718	0.0980	4.6871	60.3919	0.3724	21.3672	0.5051	36.1772
FlashReFocus + VAE Refiner	25.3699	0.7814	0.1557	0.0870	4.5208	59.6043	0.3576	19.4685	0.4936	33.1274
<i>RealDOF</i>										
IFAN (CVPR'21) [24]	24.7069	0.7611	0.3040	0.1687	6.0264	33.0571	0.2145	49.3803	0.2293	32.2959
Restormer* (CVPR'22) [71]	25.0813	0.7791	0.2863	0.1573	6.0189	37.4911	0.2420	52.7651	0.2899	27.5180
DRBNet (CVPR'22) [47]	25.7426	0.7818	0.2551	0.1312	5.2477	36.3471	0.2236	38.9563	0.2277	28.0395
INIKNet (ICCV'23) [42]	25.3061	0.7762	0.2866	0.1622	5.7921	33.3927	0.2151	47.5150	0.2896	29.1291
ViTDeblur (TCL'24) [30]	25.1551	0.7793	0.2950	0.1580	5.8195	34.8663	0.2321	49.5526	0.2771	32.3875
Swintormer (ICASSP'25) [9]	25.0624	0.7874	0.2828	0.1545	5.7826	34.7316	0.2245	49.8921	0.3166	30.2399
GenRefocus (arXiv'25) [35]	23.4872	0.7076	0.2447	0.1148	4.5444	42.4468	0.2722	35.6132	0.2912	24.2483
FlashReFocus	23.7002	0.7095	0.2527	0.1227	4.7525	44.2995	0.2601	36.0310	0.3342	30.3751
FlashReFocus + VAE Refiner	24.2544	0.7373	0.2437	0.1187	5.0235	38.8832	0.2271	34.6319	0.2894	28.8223



(a) DPDD



(b) RealDOF

Figure 6. **Qualitative deblurring** on (a) DPDD and (b) RealDOF. Each result is shown with two zoomed crops.

The control signal. We encode blur magnitude and position relative to the focal plane explicitly. Depth Anything V2 [65] $\mathcal{P}$ predicts relative inverse depth; min–max normalizing it per image and inverting gives $\mathbf{D} = 1 - \text{norm}(\mathcal{P}(\hat{\mathbf{I}}_{\text{AiF}})) \in [0, 1]$ (1 is farthest). Depth is estimated on the restored image, as the defocused input lacks the edges monocular depth relies on. A user-selected pixel p_f sets $D_f = \mathbf{D}(p_f)$. We define

$$\mathbf{M}^s(p) = \mathbf{D}(p) - D_f, \quad \mathbf{M} = |\mathbf{M}^s|, \quad (8)$$

As $\mathbf{D}$ is affine in inverse depth, $\mathbf{M} \propto |Z^{-1} - Z_f^{-1}|$ gives relative blur magnitude, and the sign of $\mathbf{M}^s$ separates the two sides of the focal plane. With global blur strength $K \in [0, 1]$, the condition stack is

$$\mathbf{M}_C = [\mathbf{M}^s, \mathbf{M}^{\text{fg}}, K\mathbf{1}], \quad \mathbf{M}^{\text{fg}} = \mathbb{I}[\mathbf{M}^s \leq 0]. \quad (9)$$

$\mathbf{M}^{\text{fg}}$ marks the focal plane and everything in front of it. The product $K\mathbf{M}$ thus mirrors the circle of confusion in Eq. (2), with K playing the role of the aperture a .

Defocus-bias attention. Bokehlicious [49] biases attention by the image-plane distance between tokens, with a range

set by one global aperture value regardless of depth. We instead derive the bias from defocus differences:

$$A_{ij} = \text{softmax}_j(\mathbf{q}_i^T \mathbf{k}_j / \sqrt{d} - \gamma_K |\phi_D(m_i) - \phi_D(m_j)|),$$

where $\mathbf{q}_i, \mathbf{k}_j$ are query and key of dimension d , m_i is $\mathbf{M}$ at token i , ϕ_D a learned embedding, and $\gamma_K > 0$ a scale set by K . Tokens at similar defocus thus exchange information regardless of image distance, so attention follows the selected focal plane. The encoder–transformer–decoder uses NAFNet blocks [10], with attention at the lowest resolution. Each block concatenates $\mathbf{M}_C$ with image features and fuses them through a depthwise separable convolution. The network predicts a residual, $\hat{\mathbf{I}}_{\text{bokeh}} = \hat{\mathbf{I}}_{\text{AiF}} + \Delta\mathbf{I}$, which stays near zero in focus and so keeps in-focus regions intact.

Two-phase training. Synthetic data provide focal diversity, while real pairs capture lens-specific appearance. We pre-train on Hypersim [44] RGB-D scenes by sampling focal depths $D_f^{(i)}$ and strengths $K^{(j)}$ and rendering targets with BokehMe [37], $\mathbf{I}_{\text{bokeh}}^{i,j} = \mathcal{G}(\mathbf{I}_{\text{AiF}}, \mathbf{D}, D_f^{(i)}, K^{(j)})$. We

Table 2. **Deblurring inference cost** at 1536×1536 .

Method	Params↓ (M)	Time↓ (s)	TFLOPs↓	Mem.↓ (GB)
IFAN	10.48	0.09	1.85	4.37
Restormer	26.13	1.62	11.15	20.01
DRBNet	11.69	0.17	3.19	7.02
INIKNet	1.98	0.60	5.52	9.28
ViTDeblur	121.60	5.65	42.78	1.60
Swintormer	194.12	321.57	10,179	4.04
GenRefocus	17,102.40	153.67	17,298	39.11
FlashReFocus ($\omega=1$)	1,290.32	0.83	53.9	9.43
FlashReFocus ($\omega=1$) + Refiner	1,363.24	1.86	95.7	17.87

Figure 7. **Bokeh rendering on EBB! Val176.**

then adapt on RealBokeh, using a larger learning rate for defocus-attention parameters than for the encoder and decoder so that the renderer adopts real lens appearance while retaining multi-focal control. Training minimizes

$$\mathcal{L}_{\text{ren}} = \lambda_r \|\hat{\mathbf{I}}_{\text{bokeh}} - \mathbf{I}_{\text{bokeh}}\|_1 + \lambda_{\text{LPIPS}}^{\text{ren}} \mathcal{L}_{\text{LPIPS}}, \quad (11)$$

with the synthetic or real target $\mathbf{I}_{\text{bokeh}}$. At inference, the renderer takes $\hat{\mathbf{I}}_{\text{AiF}}$ with any focal point and strength, so each new focal setting requires a single forward pass. Further details and results are provided in the supplementary material.

4. Experiments

Datasets and baselines. We evaluate deblurring on DPDD [1] and RealDOF [24], and bokeh rendering on EBB! Val176 [21] and 3CReal. 3CReal is our new benchmark of 89 aperture pairs from three camera systems unseen in training. We compare with seven deblurring and eight bokeh-rendering methods (Tables 1 and 3).

Evaluation protocol. Baselines use released weights or are retrained with official settings; Restormer is retrained on DPDD. Deblurring is evaluated on aligned DPDD regions and RealDOF outdoor scenes. For bokeh rendering, all methods receive the same AiF input and focal point. As strength scales differ, each method renders a range of strengths and is scored on its output with the highest SSIM to the target. Methods that cannot take a focal point use their official focus strategy (#). Stage-I settings (the refiner and ω) are chosen by the lowest rank sum over PSNR, LPIPS, DISTS, and FID on held-out data used neither for training nor in Table 1: the DPDD validation split, 50 LFDOF test scenes, and 60 RealBokeh validation pairs at $f/2.0$.

Training details. Stage I adapts Stable Diffusion 2.1 [45] following PiSA-SR, starting from its released LoRAs, on two NVIDIA RTX PRO 6000 GPUs with effective batch size 16, learning rate 5×10^{-5} , and 500 warm-up iterations. The pixel phase and the detail phase each run 20K iterations, with $\lambda_{\text{LPIPS}}^{\text{det}} = 2$, $\lambda_{\text{CSD}} = 1$, and CSD guidance 7.5; the refiner is then fine-tuned for 15K iterations on Stage-I latents. Inference processes the full image without tiling in a single step with $\omega = 1.25$.

Stage II is trained on one RTX 4090. Synthetic pre-training uses three Hypersim epochs, batch size 4, and learning rate 5×10^{-4} . Five RealBokeh adaptation epochs follow, at 5×10^{-4} for the defocus-attention parameters and 5×10^{-5} for the rest of the network. Both stages use 512×512 crops, random scale, flip, and right-angle rotation. Runtime and memory are measured after warm-up.

Metrics. For deblurring, we report pixel fidelity (PSNR, SSIM [64]), perceptual distance (LPIPS [73], DISTS [11]), no-reference quality (NIQE [34], MUSIQ [22], MANIQA [66], BRISQUE [33], CLIPQA [60]), and distributional distance (FID [16]). For bokeh rendering, we additionally report MS-SSIM [63], LPIPS with VGG features, CLIP-I [43], the no-reference CLIPQA, MUSIQ, and MANIQA, and LVCorr [15], the correlation between the strength setting and the rendered blur.

4.1. Defocus Deblurring

Comparison with prior work. As shown in Table 1, FlashReFocus with the refiner achieves the best LPIPS and BRISQUE and the second-best DISTS on both DPDD and RealDOF, and without the refiner it leads MUSIQ and CLIPQA on RealDOF. GenRefocus retains the best FID, and regression-based restorers the highest PSNR, consistent with their pixel-wise objectives. As listed in Table 2, Stage I with the refiner needs 96 TFLOPs at 1536×1536 and $\omega=1$, $181 \times$ fewer than the GenRefocus deblurring stage.

Figure 6 shows sharper vegetation, stone, and fine textures than with deterministic baselines.

Ablation. Table 4 reports cumulative ablations, starting from the official PiSA-SR recipe (4K pixel iterations at rank 4) trained on our data. Our longer, rank-16 pixel phase with the VAE-encoder LoRA and input convolution lowers LPIPS on both datasets and raises RealDOF MUSIQ from 37.94 to 44.30. The refiner favors fidelity, raising PSNR by 0.90 dB and 0.84 dB, and better preserves text-like edges and repeated structures; $\omega = 1.25$ then restores detail, as shown in Fig. 11, and gives the lowest LPIPS on both datasets at a small PSNR cost.

4.2. Controllable Bokeh Rendering

Comparison with prior work. As shown in Table 3, FlashReFocus achieves the best MS-SSIM, LPIPS, LPIPS_{VGG}, DISTS, and CLIP-I on both EBB! and 3CReal

Table 3. **Zero-shot bokeh rendering on EBB! Val176 and 3CReal.** Darker green marks the best, second, and third results per dataset. LVCorr, which measures strength controllability, is measured on LFDof (top) and 3CReal (bottom). # Selects its own focal region.

Methods	Full-reference Metrics							No-reference IQA Metrics			Control
	PSNR↑	SSIM↑	MS-SSIM↑	LPIPS↓	LPIPS _{VGG} ↓	DISTS↓	CLIP-I↑	CLIP-IQA↑	MUSIQ↑	MANIQA↑	LVCorr↑
<i>EBB! Val176 [21]</i>											
DeepLens (TOG'18) [61]	23.42	0.8707	0.8716	0.2254	0.3092	0.1157	0.9562	0.2903	49.72	0.2830	0.8866
MPIB (ECCV'22) [38]	24.10	0.8860	0.8922	0.2172	0.3089	0.1309	0.9630	0.4046	51.47	0.3149	0.6056
BokehMe (CVPR'22) [37]	24.52	0.8929	0.9007	0.2020	0.2968	0.1235	0.9696	0.4063	50.52	0.3136	0.7879
Dr.Bokeh (CVPR'24) [51]	24.13	0.8856	0.8904	0.2187	0.3114	0.1357	0.9644	0.3793	47.74	0.3026	0.6905
Bokehlicious [#] (ICCV'25) [49]	23.74	0.8561	0.8723	0.2354	0.3051	0.1585	0.9398	0.2656	28.47	0.2595	0.8473
BokehDiff [#] (ICCV'25) [76]	23.87	0.8568	0.8745	0.2137	0.3110	0.1182	0.9636	0.3981	55.34	0.3240	0.8281
GenRefocus (arXiv'25) [35]	23.80	0.8680	0.8815	0.2025	0.2860	0.1111	0.9644	0.3120	47.60	0.2880	0.6898
MagicBokeh (CVPR'26) [53]	23.14	0.8441	0.8629	0.2398	0.3334	0.1445	0.9331	0.4152	57.40	0.3241	0.7987
FlashReFocus	24.52	0.8917	0.9032	0.1805	0.2671	0.1039	0.9725	0.3534	50.86	0.3067	0.8570
<i>3CReal</i>											
DeepLens (TOG'18) [61]	25.97	0.8500	0.8933	0.2865	0.3539	0.1557	0.9626	0.3055	41.22	0.2794	0.8746
MPIB (ECCV'22) [38]	27.10	0.8759	0.9269	0.2643	0.3511	0.1631	0.9738	0.4180	43.42	0.3077	0.6367
BokehMe (CVPR'22) [37]	27.58	0.8835	0.9342	0.2471	0.3389	0.1569	0.9761	0.4183	43.52	0.3094	0.7768
Dr.Bokeh (CVPR'24) [51]	26.64	0.8713	0.9202	0.2830	0.3652	0.1772	0.9682	0.4032	39.44	0.3009	0.7137
Bokehlicious [#] (ICCV'25) [49]	26.72	0.8528	0.9090	0.2836	0.3298	0.1768	0.9630	0.3072	29.90	0.2833	0.6677
BokehDiff [#] (ICCV'25) [76]	25.80	0.8115	0.8610	0.3112	0.3791	0.1589	0.9566	0.4085	46.97	0.3102	0.7972
GenRefocus (arXiv'25) [35]	26.38	0.8528	0.9218	0.2426	0.3185	0.1493	0.9715	0.2928	37.66	0.2893	0.6658
MagicBokeh (CVPR'26) [53]	25.64	0.8329	0.8957	0.2543	0.3506	0.1553	0.9430	0.3920	48.52	0.3190	0.7589
FlashReFocus	27.98	0.8855	0.9375	0.2178	0.3004	0.1368	0.9807	0.3482	43.06	0.3043	0.8391

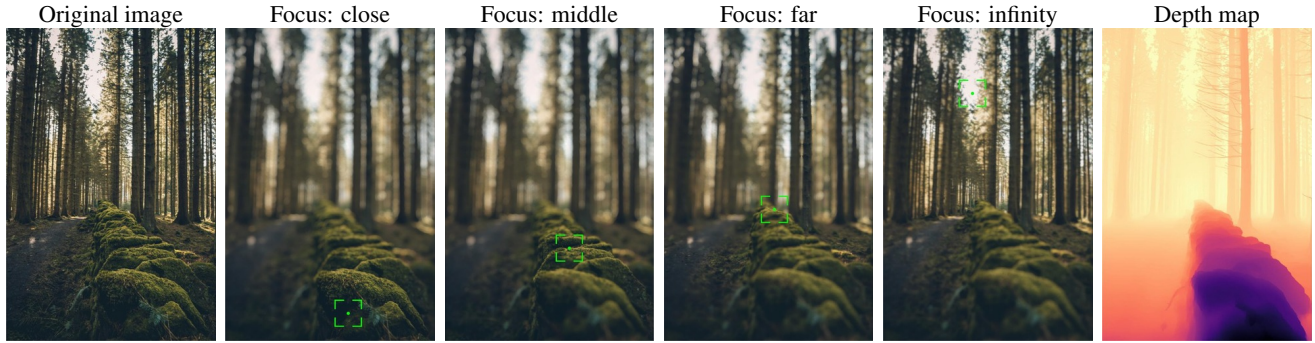


Figure 8. **Focal control.** One all-in-focus input is rendered with the focal point (green marker) moving from the nearest stones to the distant trees; the rightmost panel shows the estimated depth. The sharp region moves with the selected depth while the rest blurs.

Table 4. **Deblurring ablation on DPDD and RealDOF.** Starting from the official PiSA-SR recipe, our pixel phase (Pix.), the decoder refiner (Ref.), and $\omega = 1.25$ are added in turn.

Pix.	Ref.	ω	DPDD			RealDOF		
			PSNR↑	LPIPS↓	MUSIQ↑	PSNR↑	LPIPS↓	MUSIQ↑
×	×	×	24.57	.177	60.08	23.77	.260	37.94
✓	×	×	24.61	.172	60.39	23.70	.253	44.30
✓	✓	×	25.51	.167	59.07	24.54	.248	37.35
✓	✓	✓	25.37	.156	59.60	24.25	.244	38.88

and the best PSNR and SSIM on 3CReal; on EBB!, its PSNR ties BokehMe. On EBB!, it improves LPIPS from 0.202 to 0.181 and DISTS from 0.111 to 0.104 over the next-best method; on 3CReal, it leads PSNR by 0.40 dB and improves LPIPS from 0.243 to 0.218. Its high LVCorr (0.857, 0.839) shows that the blur follows the strength control. The gains on 3CReal, from three camera systems unseen in training, indicate generalization to new optics. In Fig. 7, FlashReFocus keeps the wooden blocks sharp while blurring the railing and foliage to a degree close to the ref-

Table 5. **Bokeh ablation on 3CReal.** The upper block compares training strategies for the renderer, and the lower block compares attention biases with the training data and schedule held fixed.

Variant	PSNR↑	SSIM↑	MS-SSIM↑	LPIPS↓	DISTS↓	CLIP-I↑
RealBokeh only	27.284	.864	.917	.251	.139	.973
Hypersim only	27.460	.882	.932	.251	.164	.976
Uniform fine-tuning	27.938	.885	.936	.225	.142	.980
Differential LR	27.982	.885	.937	.218	.137	.981
Relative-position attn.	27.898	.884	.936	.223	.139	.980
Mixed attention	27.956	.885	.937	.222	.139	.981
Defocus-bias attention	27.982	.885	.937	.218	.137	.981

erence, whereas GenRefocus leaves the foliage nearly sharp and BokehDiff blurs it more than the reference.

Controllability. Figure 8 moves the focal point along a moss-covered wall from the nearest stones to the distant trees. The sharp region follows the selected depth, and the blur of the other regions grows with their distance from it, consistent with the estimated depth map. With the focal point fixed, the strength acts as an aperture, as Fig. 10a shows: when it rises from $1\times$ to $3\times$, the background sharp-

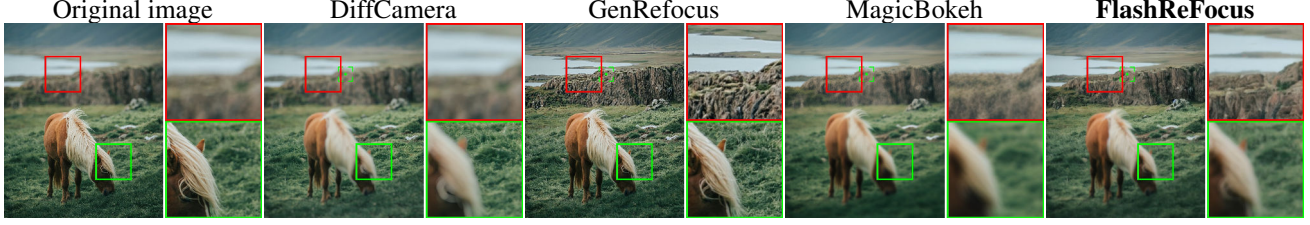


Figure 9. **Refocusing to the background.** The focal point (green marker) moves to the defocused cliff. Only FlashReFocus both restores the cliff (red box) and blurs the previously focused horse (green box).

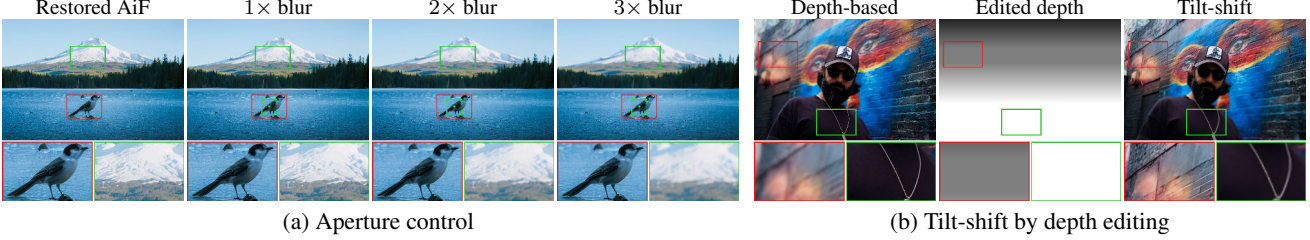


Figure 10. **Controls beyond the focal point.** (a) With the focal point fixed (green marker), raising the strength from $1\times$ to $3\times$ enlarges the background blur (green box) as a wider aperture would, while the subject stays sharp (red box). (b) Replacing the estimated depth with a vertical ramp keeps the whole band at face height sharp, including the receding wall (red box), and blurs the shirt below (green box), a tilt-shift look that the physical depth cannot produce.

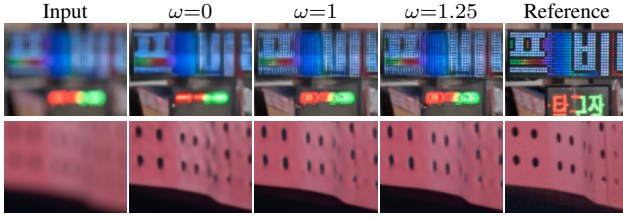


Figure 11. **Detail weight control.** RealDOF crops restored with the refiner as ω grows from 0, the pixel LoRA alone, to 1.25; the LED grid and the plate holes gain detail.

ness, measured as Laplacian variance, falls from 237 to 29 and 11, while the subject keeps 65% of its sharpness. Because the renderer reads an explicit depth map, editing that map enables effects that a physical depth cannot produce. As shown in Fig. 10b, replacing it with a vertical ramp keeps the whole band at face height sharp, raising the Laplacian variance of the receding wall from 7 to 321, and blurs the shirt below from 362 to 24, yielding a tilt-shift look.

Ablation. In Table 5, adapting the synthetic pre-training to real pairs raises PSNR to 27.938 dB, and differential learning rates add 0.043 dB and lower LPIPS to 0.218. Defocus-bias attention outperforms the relative-position attention of Bokehlicious by 0.084 dB PSNR and 0.005 LPIPS. Other depth models change PSNR by ≤ 0.11 dB (App. G).

4.3. Complete Refocusing and Efficiency

Cost of one refocus. As reported in Table 6, a complete refocus at 1536×1536 takes 2.50 s and 16.94 GB of peak memory, $54.5\times$ and $148.3\times$ faster than DiffCamera and GenRefocus with $161\times$ and $350\times$ fewer FLOPs. Further edits reuse $\hat{\mathbf{I}}_{\text{AiF}}$ and its depth and run only the renderer, 0.47 s each; the one-step MagicBokeh, without restoration,

needs 6.13 s and $133\times$ more FLOPs per edit. In-focus regions stay nearly unchanged across edits, at 44.4 dB PSNR to the restored image on 3CReal against 29.0 dB for GenRefocus. In Fig. 9, where the focus moves to a defocused background, only FlashReFocus restores the new focus and blurs the former one: DiffCamera and MagicBokeh leave the cliff blurred; GenRefocus keeps the horse sharp. With light-field ground truth, it leads PSNR by 2.16 dB (App. H).

Table 6. **Refocusing inference cost** at 1536×1536 . *First*: restoration and first rendering; *New focus*: each additional focal edit. $\dagger$ No restoration; renders directly from the input.

Method	Params $\downarrow$ (B)	Time (s) $\downarrow$		TFLOPs $\downarrow$		Mem. $\downarrow$ (GB)
		First	New focus	First	New focus	
DiffCamera	12.33	136.35	136.35	19,363	19,363	28.37
GenRefocus	18.30	370.65	217.32	42,031	24,763	42.82
BokehDiff $\dagger$	3.83	—	3.43	—	73.4	11.82
MagicBokeh $\dagger$	1.73	—	6.13	—	113	5.29
FlashReFocus	1.70	2.50	0.47	120	0.85	16.94

5. Conclusion

We presented FlashReFocus, a two-stage framework for interactive single-image refocusing. Stage I restores an all-in-focus image once with one-step diffusion, and a 1.3M-parameter renderer renders each focal setting in 0.47 s. It achieves the best perceptual bokeh fidelity on EBB! Val176 and 3CReal, and its explicit defocus map enables aperture control, tilt-shift, and two focal planes.

Limitations. Blur follows monocular relative depth, so depth errors misplace it, and background occluded by a blurred foreground edge cannot be revealed. The aperture shape is not controllable, and Stage I can add unsupported texture or leave heavily blurred regions soft (Apps. F and I).

References

- [1] Abdullah Abuolaim and Michael S Brown. Defocus deblurring using dual-pixel data. In *European conference on computer vision*, pages 111–126. Springer, 2020. 2, 4, 5, 6, 13
- [2] Abdullah Abuolaim, Mauricio Delbracio, Damien Kelly, Michael S Brown, and Peyman Milanfar. Learning to reduce defocus blur by realistically modeling dual-pixel data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2289–2298, 2021. 2
- [3] Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 126–135, 2017. 4, 13
- [4] Sumukh K. Aithal, Pratyush Maini, Zachary C. Lipton, and J. Zico Kolter. Understanding hallucinations in diffusion models through mode interpolation. In *Advances in Neural Information Processing Systems*, pages 134614–134644, 2024. 3
- [5] Hadi Alzayer, Abdullah Abuolaim, Leung Chun Chan, Yang Yang, Ying Chen Lou, Jia-Bin Huang, and Abhishek Kar. DC²: Dual-camera defocus control by learning to refocus. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21488–21497, 2023. 3
- [6] Sajjad Amrollahi Biyouki and Hoon Hwangbo. A comprehensive survey on deep neural image deblurring. *arXiv preprint arXiv:2310.04719*, 2023. 2
- [7] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6228–6237, 2018. 1, 2, 3
- [8] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. In *International Conference on Learning Representations*, 2025. 20
- [9] Kang Chen and Yuanjie Liu. Efficient defocus deblurring networks based on diffusion models. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025. 3, 5, 17
- [10] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *European conference on computer vision*, pages 17–33. Springer, 2022. 2, 5
- [11] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2022. 6, 17
- [12] Georgios D. Evangelidis and Emmanouil Z. Psarakis. Parametric image alignment using enhanced correlation coefficient maximization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(10):1858–1865, 2008. 16
- [13] Haoxuan Feng, Haohui Zhou, Tian Ye, Sixiang Chen, and Lei Zhu. Residual diffusion deblurring model for single image defocus deblurring. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2960–2968, 2025. 2
- [14] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 16
- [15] Armando Fortes, Tianyi Wei, Shangchen Zhou, and Xingang Pan. Bokeh Diffusion: Defocus blur control in text-to-image diffusion models. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pages 1–11, 2025. 3, 6, 17
- [16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017. 6, 18
- [17] Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. Axial attention in multidimensional transformers. *arXiv preprint arXiv:1912.12180*, 2019. 14
- [18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851, 2020. 1, 2
- [19] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 3
- [20] Xinyu Huang, Yi-Jie Huang, Youcai Zhang, Weiwei Tian, Rui Feng, Yuejie Zhang, Yanchun Xie, Yaqian Li, and Lei Zhang. Open-set image tagging with multi-grained text supervision. *arXiv preprint arXiv:2310.15200*, 2023. 4, 13
- [21] 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*, pages 418–419, 2020. 2, 3, 6, 7, 15, 16
- [22] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5148–5157, 2021. 6, 18
- [23] Craig Kolb, Don Mitchell, and Pat Hanrahan. A realistic camera model for computer graphics. In *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques*, page 317–324, New York, NY, USA, 1995. Association for Computing Machinery. 3
- [24] Junyong Lee, Hyeonseok Son, Jaesung Rim, Sunghyun Cho, and Seungyong Lee. Iterative filter adaptive network for single image defocus deblurring. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2034–2042, 2021. 1, 2, 5, 6
- [25] Anat Levin, Rob Fergus, Frédo Durand, and William T Freeman. Image and depth from a conventional camera with a coded aperture. *ACM transactions on graphics (TOG)*, 26(3):70–es, 2007. 2
- [26] Xiaopan Li, Shiqian Wu, Qile Zhu, Shoulie Xie, and Sos Agaian. Single image defocus deblurring via multimodal-guided diffusion and depth-aware fusion. *Pattern Recognition*, 171:112133, 2026. 3

- [27] Yuhao Li, Haoran Fang, Xiang Lei, Qi Wang, Gang Hu, Jiaqing Dong, Zilong Li, Jiabin Lin, Qiegen Liu, and Xianlin Song. Real-world defocus deblurring via score-based diffusion models. *Scientific Reports*, 15(1):22942, 2025.
- [28] Hanyan Liang, Shuyao Chai, Xixuan Zhao, and Jiangming Kan. Swin-Diff: A single defocus image deblurring network based on diffusion model. *Complex & Intelligent Systems*, 11(3):170, 2025. 3
- [29] Jingyun Liang, Jiezhong Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. SwinIR: Image restoration using Swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 1833–1844, 2021. 2
- [30] Pengwei Liang, Junjun Jiang, Xianming Liu, and Jiayi Ma. Decoupling image deblurring into twofold: A hierarchical model for defocus deblurring. *IEEE Transactions on Computational Imaging*, 10:1207–1220, 2024. 5
- [31] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 16
- [32] Ziwei Luo, Fredrik K. Gustafsson, Zheng Zhao, Jens Sjöblund, and Thomas B. Schön. Refusion: Enabling large-size realistic image restoration with latent-space diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1680–1691, 2023. 2
- [33] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012. 6, 18
- [34] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a completely blind image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013. 6, 18
- [35] Chun-Wei Tuan Mu, Cheng-De Fan, Jia-Bin Huang, and Yu-Lun Liu. Generative refocusing: Flexible defocus control from a single image. *arXiv preprint arXiv:2512.16923*, 2025. 1, 3, 5, 7, 17, 21, 22, 23, 25
- [36] Ren Ng, Marc Levoy, Mathieu Brédif, Gene Duval, Mark Horowitz, and Pat Hanrahan. Light field photography with a hand-held plenoptic camera. Technical Report CSTR 2005-02, Stanford University Computer Science, 2005. 3
- [37] 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*, pages 16283–16292, 2022. 2, 3, 5, 7, 15, 25
- [38] 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*, pages 590–607. Springer, 2022. 3, 7
- [39] Matt Pharr, Wenzel Jakob, and Greg Humphreys. *Physically Based Rendering: From Theory to Implementation*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition, 2016. 3
- [40] Michael Potmesil and Indranil Chakravarty. A lens and aperture camera model for synthetic image generation. *ACM SIGGRAPH Computer Graphics*, 15(3):297–305, 1981. 1, 3
- [41] Yuhui Quan, Zicong Wu, and Hui Ji. Gaussian kernel mixture network for single image defocus deblurring. In *Advances in Neural Information Processing Systems*, pages 20812–20824, 2021. 2
- [42] Yuhui Quan, Xin Yao, and Hui Ji. Single image defocus deblurring via implicit neural inverse kernels. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12600–12610, 2023. 5
- [43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021. 6, 17
- [44] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10912–10922, 2021. 5, 15, 19
- [45] 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*, pages 10684–10695, 2022. 1, 2, 6, 13
- [46] Lingyan Ruan, Bin Chen, Jizhou Li, and Miu-Ling Lam. AIFNet: All-in-focus image restoration network using a light field-based dataset. *IEEE Transactions on Computational Imaging*, 7:675–688, 2021. 4, 13, 17, 21
- [47] Lingyan Ruan, Bin Chen, Jizhou Li, and Miuling Lam. Learning to deblur using light field generated and real defocus images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16304–16313, 2022. 2, 5
- [48] Parikshit Sakurikar, Ishit Mehta, Vineeth N. Balasubramanian, and P. J. Narayanan. RefocusGAN: Scene refocusing using a single image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 497–512, 2018. 3
- [49] Tim Seizinger, Florin-Alexandru Vasluianu, Marcos V Conde, Zongwei Wu, and Radu Timofte. Bokehlicious: Photorealistic bokeh rendering with controllable apertures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8908–8917, 2025. 2, 3, 4, 5, 7, 13, 15, 16, 25
- [50] Tim Seizinger, Zhuyun Zhou, and Radu Timofte. The RealDefocus benchmark for defocus deblurring. In *IEEE International Conference on Image Processing (ICIP)*, 2026. 4, 15, 16
- [51] 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*, pages 4515–4525, 2024. 3, 7

- [52] Linxiao Shi and Shifeng Chen. Bokeh rendering via diffusion models without depth priors. In *2025 International Conference on Machine Learning, Computational Intelligence and Pattern Recognition (MLCIPR)*, pages 117–120. IEEE, 2025. 3
- [53] Linxiao Shi, Siming Zheng, Zerong Wang, Hao Zhang, Jinwei Chen, Bo Li, Shifeng Chen, and Peng-Tao Jiang. Towards photorealistic and efficient bokeh rendering via diffusion framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. 1, 3, 7, 22, 25
- [54] Ayush Shrivastava, Connelly Barnes, Xuaner Zhang, Lingzhi Zhang, Andrew Owens, Sohrab Amirghodsi, and Eli Shechtman. Fine-grained defocus blur control for generative image models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4335–4344, 2026. 3
- [55] Hyeonseoek Son, Junyong Lee, Sunghyun Cho, and Seungyong Lee. Single image defocus deblurring using kernel-sharing parallel atrous convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2642–2650, 2021. 2
- [56] Lingchen Sun, Rongyuan Wu, Zhiyuan Ma, Shuaizheng Liu, Qiaosi Yi, and Lei Zhang. Pixel-level and semantic-level adjustable super-resolution: A dual-LoRA approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2333–2343, 2025. 2, 3, 13
- [57] Nisarg K. Trivedi, Vinayak A. Belludi, Li-Yun Wang, Pardis Taghavi, and Dante Lok. MODEST: Multi-optics depth-of-field stereo dataset, 2025. 17
- [58] Unsplash. Unsplash. <https://unsplash.com>, 2026. Photographs used under the Unsplash License, accessed 2026. 17, 26
- [59] Neal Wadhwa, Rahul Garg, David E Jacobs, Bryan E Feldman, Nori Kanazawa, Robert Carroll, Yair Movshovitz-Attias, Jonathan T Barron, Yael Pritch, and Marc Levoy. Synthetic depth-of-field with a single-camera mobile phone. *ACM Transactions on Graphics (ToG)*, 37(4):1–13, 2018. 3
- [60] Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. Exploring CLIP for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2555–2563, 2023. 6, 18
- [61] 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. *ACM Transactions on Graphics*, 37(6):1–11, 2018. 3, 7, 25
- [62] 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*, pages 1–10, 2025. 1, 3, 17, 22
- [63] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers*, pages 1398–1402, 2003. 6, 17
- [64] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 6, 17
- [65] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything V2. In *Advances in Neural Information Processing Systems*, pages 21875–21911, 2024. 5, 20
- [66] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. MANIQA: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1191–1200, 2022. 6, 18
- [67] Tao Yang, Rongyuan Wu, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. In *European conference on computer vision*, pages 74–91. Springer, 2024. 2
- [68] Yang Yang, Haiting Lin, Zhan Yu, Sylvain Paris, and Jingyi Yu. Virtual DSLR: High quality dynamic depth-of-field synthesis on mobile platforms. *Electronic Imaging*, 28(18):1–9, 2016. 3
- [69] Yang Yang, Siming Zheng, Qirui Yang, Jinwei Chen, Boxi Wu, Xiaofei He, Deng Cai, Bo Li, and Peng-Tao Jiang. Any-to-Bokeh: Arbitrary-subject video refocusing with video diffusion model. In *International Conference on Learning Representations*, 2026. 3
- [70] Xin Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Song-Hai Zhang, and Xiaojuan Qi. Text-to-3D with classifier score distillation. In *International Conference on Learning Representations*, 2024. 4, 13
- [71] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5728–5739, 2022. 1, 2, 5
- [72] Kaihao Zhang, Wenqi Ren, Wenhan Luo, Wei-Sheng Lai, Björn Stenger, Ming-Hsuan Yang, and Hongdong Li. Deep image deblurring: A survey. *International Journal of Computer Vision*, 130(9):2103–2130, 2022. 2
- [73] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 4, 6, 17
- [74] Xinxin Zhang, Ronggang Wang, Xiubao Jiang, Wenmin Wang, and Wen Gao. Spatially variant defocus blur map estimation and deblurring from a single image. *Journal of Visual Communication and Image Representation*, 35:257–264, 2016. 2
- [75] Zhengbo Zhang, Lin Geng Foo, Hossein Rahmani, Jun Liu, and De Wen Soh. Performing defocus deblurring by modeling its formation process. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5791–5801, 2025. 2
- [76] 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*, pages 9508–9518, 2025. 3, 7, 25

- [77] Shaojie Zhuo and Terence Sim. Defocus map estimation from a single image. *Pattern Recognition*, 44(9):1852–1858, 2011. [2](#)

FlashReFocus: Interactive Image Refocusing in Seconds

Supplementary Material

Contents

A	Defocus Image Formation	13
B	Restoration Implementation Details	13
C	Rendering Implementation Details	14
D	Evaluation Data	15
E	Experiment Settings	17
F	Additional Deblurring Results	20
G	Additional Bokeh Results	20
H	Refocusing and Editing	21
I	Limitations and Broader Impact	25

Figures, tables, and equations are numbered per section.

A. Defocus Image Formation

This section summarizes the thin-lens and disk-blur approximations behind the defocus control signal and the synthetic supervision. For focal length f , object distance Z , and image distance v ,

$$\frac{1}{f} = \frac{1}{Z} + \frac{1}{v}. \quad (\text{A.1})$$

When the camera focuses at Z_f , the sensor lies at v_f , where $1/f = 1/Z_f + 1/v_f$. A point at another depth forms a circle of confusion (CoC), approximated by

$$c(Z) \propto a \left| \frac{1}{Z} - \frac{1}{Z_f} \right|, \quad r(\mathbf{x}) \approx \kappa c(Z(\mathbf{x})), \quad (\text{A.2})$$

where a is the aperture diameter, $Z(\mathbf{x})$ the depth at pixel $\mathbf{x}$, and κ converts CoC size to a pixel radius. The inverse-depth dependence makes equal depth offsets on opposite sides of the focal plane produce different blur sizes. Under the disk approximation,

$$\mathbf{I}_{\text{blur}}(\mathbf{x}) = \sum_{\mathbf{u} \in \Omega(\mathbf{x})} \mathcal{K}_{r(\mathbf{x})}(\mathbf{u}) \mathbf{I}_{\text{AiF}}(\mathbf{x} - \mathbf{u}), \quad (\text{A.3})$$

where $\mathcal{K}_{r(\mathbf{x})}$ is a disk kernel and $\Omega(\mathbf{x})$ its support. This approximation omits aperture shape, lens aberrations, occlusion, and camera processing.

B. Restoration Implementation Details

B.1. Conditioning and Data Preparation

Stage I follows PiSA-SR [56] on Stable Diffusion 2.1 [45]. At inference the U-Net runs at timestep 1 with an empty prompt and the VAE encoder uses its posterior mode, so an input always yields the same all-in-focus image, which every later edit reuses. LoRAs adapt the convolution, attention, and feed-forward projections of the U-Net and the convolution and attention layers of the VAE encoder. Training

starts from the released PiSA-SR LoRAs: the rank-4 pixel LoRA is embedded exactly into the first ranks of the rank-16 one, and every rank keeps the released scaling. CSD [70] uses the frozen model as teacher with guidance scale 7.5 and timesteps from 20 to 500. Its prompt is formed from the 50 highest-scoring RAM++ tags [20] of the sharp target, and its negative prompt adds bokeh, shallow depth of field, out of focus, and defocus blur to the official one. Half of the training samples give the student an empty prompt. With $\omega \neq 1$, the adjustable mode of PiSA-SR computes ϵ_{pix} and $\epsilon_{\text{pix+det}}$ with two U-Net passes.

Figure B.1 summarizes the mixed data sources: 350 training pairs of DPDD [1], 4,179 filtered pairs of RealBokeh [49], 11,261 training pairs of LFDOF [46], and 639 sharp DIV2K [3] images for synthetic blur, 16,429 samples in total. RealBokeh scenes are ranked by the Laplacian variance of their small-aperture ($f/22$) images, and the top 1,000 scenes are kept. DIV2K images with visible defocus are manually excluded, and the central 60% of each remaining image is used to generate synthetic pairs with Eq. 6 of the main paper. For each image, one kernel type is drawn uniformly from disk, Gaussian, and motion, which improves robustness to non-ideal real kernels; the image is blurred at radii r_k of 1, 5.5, and 10 pixels, and the three results are blended per pixel with weights w_k derived from $\mathbf{M}_s$, a ramp in one of eight directions (vertical, horizontal, and diagonal ramps in either sense, or radial from or toward the center) or a uniform map, drawn at random. Figure B.1(E) lists the sampling probability of each source.

B.2. High-Frequency Decoder Guidance

Figure 5 of the main paper isolates reconstruction errors introduced by VAE encoding and decoding. Figure B.2 shows the decoder guidance used to reduce these errors.

The high-frequency input is computed with a single-level Laplacian pyramid:

$$\mathbf{H} = |\mathbf{I}_{\text{blur}} - \uparrow_2 (\downarrow_2 (G_\sigma * \mathbf{I}_{\text{blur}}))|, \quad (\text{B.1})$$

where G_σ is a Gaussian kernel, $*$ denotes convolution, and $\downarrow_2$ and $\uparrow_2$ denote resampling by a factor of two. The degraded image and frequency map are concatenated and encoded as

$$\mathbf{F}_{\text{prior}} = \phi_{\text{prior}}([\mathbf{I}_{\text{blur}}, \mathbf{H}]), \quad (\text{B.2})$$

where ϕ_{prior} is a lightweight CNN and $\mathbf{F}_{\text{prior}}$ is fused with intermediate VAE decoder features. The refiner’s fidelity-perception trade-off is evaluated in the main paper, with further examples in Appendix F.

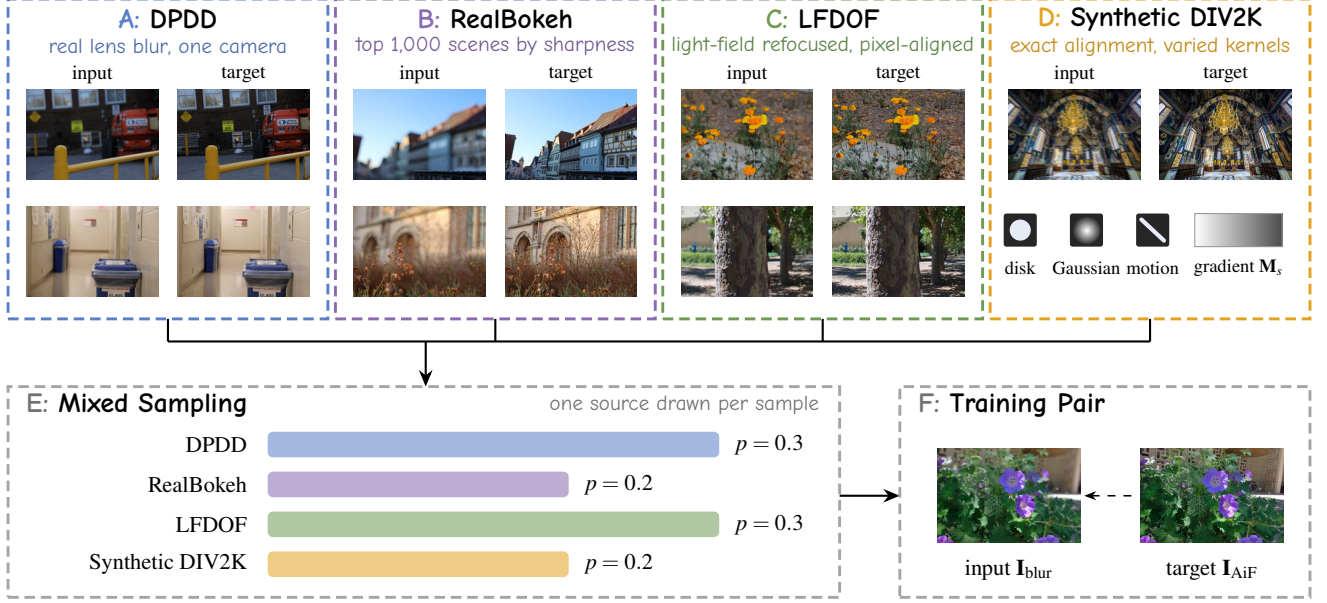


Figure B.1. **Mixed deblurring data.** (A) DPDD provides real lens blur from a single camera. (B) RealBokeh pairs, reversed into deblurring pairs and filtered to the 1,000 scenes with the sharpest $f/22$ targets, add strong real blur. (C) LFDOF adds pixel-aligned pairs refocused from light fields. (D) Synthetic DIV2K pairs blend disk, Gaussian, or motion kernels along a random gradient map $\mathbf{M}_s$, giving exact alignment and spatially varying blur. (E) One source is drawn per sample with the listed probability, (F) giving a pair $(\mathbf{I}_{\text{blur}}, \mathbf{I}_{\text{AiF}})$.

C. Rendering Implementation Details

C.1. Defocus-Bias Attention

We detail the attention computation of Sec. 3.2. For the feature $\mathbf{f}_i$ of token i in an intermediate feature map,

$$\mathbf{q}_i = W_q \mathbf{f}_i, \quad \mathbf{k}_i = W_k \mathbf{f}_i, \quad \mathbf{v}_i = W_v \mathbf{f}_i, \quad (\text{C.1})$$

where W_q , W_k , and W_v are learned projections, applied per head; we omit the head index below. The embedding ϕ_D encodes the defocus magnitude $\mathbf{M} \in [0, 1]$ with fixed logarithmic features at several scales and maps them to one value per head with a learned linear layer; these defocus tokens define a head-wise pairwise bias:

$$\mathbf{T}_D = \phi_D(\mathbf{M}), \quad B_{ij}^{\text{dm}} = -|\mathbf{T}_{D,i} - \mathbf{T}_{D,j}|, \quad (\text{C.2})$$

where $\mathbf{T}_{D,i}$ and $\mathbf{T}_{D,j}$ are tokens at positions i and j . Attention is then

$$A_{ij} = \text{Softmax}_j \left(\frac{\mathbf{q}_i^T \mathbf{k}_j}{\sqrt{d}} + \gamma_b \alpha B_{ij}^{\text{dm}} \right), \quad (\text{C.3})$$

where d is the key dimension, α is a learned head-wise scale, and $\gamma_b = \exp(\text{clip}(w_b K + c_b, -4, 4))$, with learned head-wise w_b and c_b , is a positive scale computed from the global blur strength K ; the scale γ_K of the main paper is $\gamma_b \alpha$. As in Bokehlicious, attention is axial [17]: it is applied along each row and then along each column of the feature map, so j ranges over one row or column. The output adds

a locally enhanced positional encoding (LePE), a depthwise convolution over the value map $\mathbf{V} = \{\mathbf{v}_j\}$, before the output projection W_o :

$$\mathbf{f}'_i = W_o \left(\sum_j A_{ij} \mathbf{v}_j + \text{LePE}(\mathbf{V})_i \right). \quad (\text{C.4})$$

Attention visualization. For a query token q , the axial attention gives token (h', w') the effective weight $A_{q,h'}^{\text{col}} A_{h',w'}^{\text{row}}$, the column weight of row h' times the row weight within that row. Figure C.1 compares these weights with those of the same blocks with the bias term removed, for the same queries, keys, and values, averaged over the nine DAABs and three heads. With the bias, a query in the background attends more to the background and less to the in-focus subject, and a query on the subject attends more to the subject; the bias thus keeps tokens on one side of a depth edge from mixing with the other, where blur must change abruptly. Over the 89 3CReal images, resized to a long side of 1,024 and rendered at the annotated focal point with $K = 0.5$, the share of attention on tokens whose defocus differs from the query’s by less than 0.1 rises from 0.40 without the bias to 0.64 with it (16 queries per image).

Preservation of in-focus regions. The renderer adds a residual $\Delta \mathbf{I}$ to the restored image, and Figure C.2 shows where this residual is nonzero. The in-focus subject stays dark at every strength, while the residual elsewhere grows with the defocus magnitude $\mathbf{M}$ and with K . Over the 89

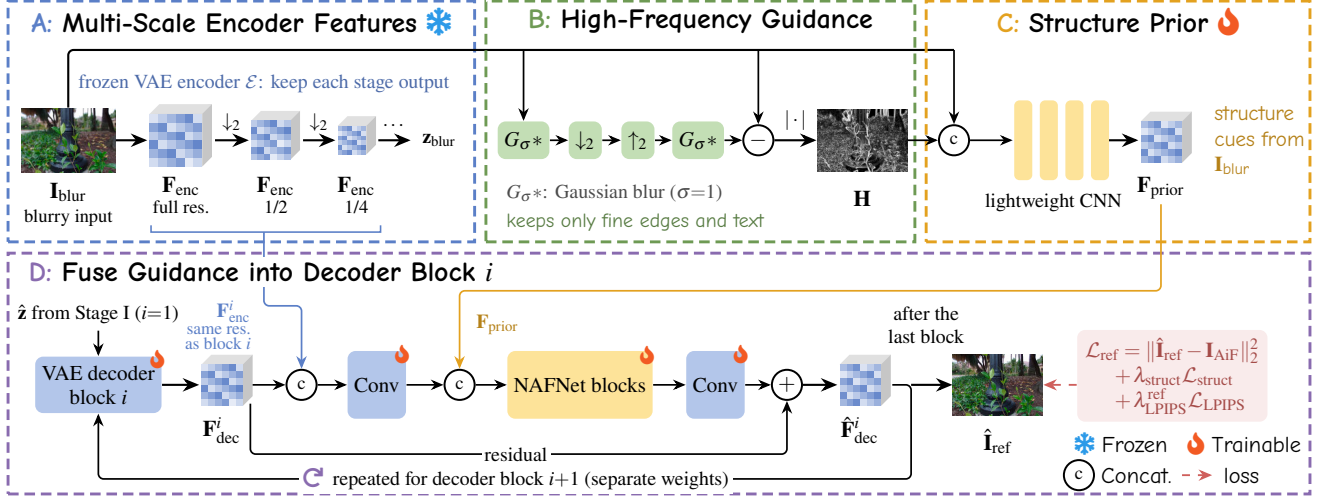


Figure B.2. **Input-guided decoder refiner.** (A) Multi-scale encoder features, each matched to the decoder block of the same resolution, (B) a high-frequency map, and (C) structure features of $\mathbf{I}_{\text{blur}}$ are (D) fused into each block of the decoder, which decodes the Stage-I latent $\hat{\mathbf{z}}$. The fusion modules and the decoder are trained with $\mathcal{L}_{\text{ref}}$ while Stage I stays frozen. $\mathbf{H}$ is contrast-enhanced for display.

3CReal images, rendered as above, the mean $|\Delta \mathbf{I}|$ is 0.7–1.8/255 where $\mathbf{M} < 0.05$ for K from 0.25 to 1, against 6–17/255 at the largest defocus, so in-focus content is left nearly unchanged. The learned conditioning agrees with this: the embedding ϕ_D rises fastest near the focal plane, reaching half of its range at $\mathbf{M} \approx 0.30$ (median over the 22 of 27 heads with an embedding range above one) instead of 0.5 for a linear map, and the scale γ_K grows with K most in the first DAAB of each DRG, by up to $2\times$ at $K = 1$.

Depth from the restored image. The renderer estimates depth on the restored image rather than on the input (Sec. 3.2). On the 89 shallow-depth-of-field photographs of 3CReal, whose deep-depth-of-field captures of the same views give a reference depth, the mean absolute difference to that reference falls from 0.048 with depth estimated on the input to 0.027 with depth estimated on the restored image, and is lower for 74 of the 89 views. Figure C.3 shows that blurred backgrounds lose the thin structures that monocular depth needs.

C.2. Synthetic Multi-Focal Supervision

Figure C.4 illustrates synthetic pre-training followed by real-data adaptation. For each RGB-D scene, we sample focal depths and blur strengths independently to create multiple targets:

$$\mathbf{I}_{\text{bokeh}}^{i,j} = \mathcal{G}(\mathbf{I}_{\text{AiF}}, \mathbf{D}, D_f^{(i)}, K^{(j)}), \quad (\text{C.5})$$

where $\mathcal{G}$ is the BokehMe renderer [37], $\mathbf{D} = 1 - \min(Z^{-1}, 1)$ computed from the ground-truth depth Z in meters, $D_f^{(i)}$ a sampled focal depth, and $K^{(j)}$ a sampled blur strength, with BokehMe blur parameter $100K^{(j)}$.

Pre-training uses 59,543 RGB-D frames from 367 HyperSim [44] scenes, each paired with one target rendered at a random focal depth and strength, and adaptation uses the 20,554 pairs of the 3,960 RealBokeh training scenes, with $\mathbf{D}$ from Depth Anything V2 as at inference and, following Bokehlicious, $K = 2/N$ for a target at f/N . Optimizer settings are listed in Appendix E.

D. Evaluation Data

D.1. 3CReal Benchmark

We collect 3CReal to evaluate transfer to unseen camera and lens systems (Fig. D.1). Real bokeh benchmarks pair a small- and a large-aperture photograph of the same scene, but each existing one is captured with a single camera and lens (Table D.1): EBB! [21] with one DSLR and a 50 mm lens, and RealBokeh [49, 50] with one Canon EOS R6 II and a 28–70 mm lens, whose training split is also used to adapt our renderer. A method can therefore fit the appearance of one lens without this being measured. 3CReal instead spans three camera and lens systems that no compared method has seen during training. Scenes around a university campus are captured from a tripod at two aperture settings: a small-aperture exposure provides a deep-depth-of-field input, and a large-aperture exposure provides the shallow-depth-of-field target. Automatic ISO and shutter speed reduce exposure differences.

The three systems are a Canon EOS 750D with a Canon EF-S 18–55 mm f/3.5–5.6 IS STM lens, a FUJIFILM X-E5 with a Viltrox AF 35 mm F1.7 Air lens for FUJIFILM X-mount, and a FUJIFILM X100VI with its built-in Fujinon 23 mm F2 lens, contributing 31, 31, and 27 pairs for a

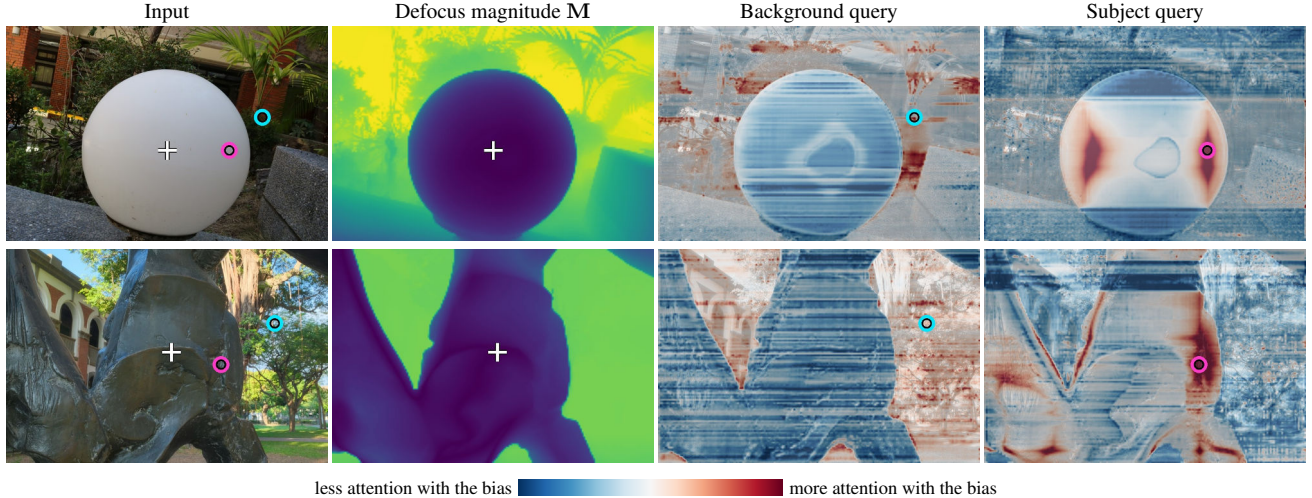


Figure C.1. **Effect of the defocus bias on attention.** Ratio of the effective attention weights of each query (circle) with and without the bias, on a log scale from $0.1\times$ (blue) to $10\times$ (red), averaged over all attention blocks; white cross: focal point. With the bias, attention concentrates at the query’s defocus level: background queries leave the in-focus subject, and subject queries leave the background.

Table D.1. **Real bokeh benchmarks.** Each pair is a small- and large-aperture photograph of one scene. EBB! uses a Canon 7D with a 50 mm lens, RealBokeh a Canon EOS R6 II with a 28–70 mm lens (statistics from [50]). EBB! Val176 is our filtered validation split (Appendix D.2), *Registration* is how pairs are aligned, and *Seen* marks optics present in the renderer’s training data.

Benchmark	Pairs	Systems	Registration	Seen
EBB! Val176 [21]	176	1	SIFT+RANSAC	No
RealBokeh [49]	1,250	1	Tripod	Yes
3CReal (ours)	89	3	Tripod+ECC	No

total of 89. Deep-DoF inputs are taken at $f/16$ (Fujifilm) or $f/22$ – $f/36$ (Canon) and shallow-DoF targets at $f/1.7$ – $f/5.6$. Captures use a 3:2 aspect ratio, including three portrait pairs, except for eight X100VI pairs captured at 16:9. Residual shifts from tripod vibration, scene motion, or manual operation are corrected by Enhanced Correlation Coefficient alignment [12], after which the shared valid region is cropped and black padding removed, using the same cropping rule within each camera system. Table D.2 lists the resulting sizes. The scenes cover indoor and outdoor layouts, specular highlights, thin structures, and depth boundaries. None of these images is used for training.

All aligned crops are manually inspected. Residual exposure differences, vignetting, independently moving objects, and aperture-dependent occlusion remain: ECC corrects global geometric shifts but cannot remove these effects. Full-reference scores should therefore be interpreted alongside visual comparisons. Figure D.2 provides additional paired examples.

Table D.2. **3CReal composition.** Pairs, EXIF apertures of the deep-/shallow-DoF exposures, and post-processed size per camera and lens system. *Eight 16:9 X100VI pairs lack EXIF metadata, so their apertures are unknown.

System	Pairs	Deep / shallow DoF	Post-processed size
750D / 18–55 mm	31	$f/22$ – $f/36$ / $f/3.5$ – 5.6	2966×1974
X-E5 / 35 mm	31	$f/16$ / $f/1.7$ – 5.6	1900×1256
X100VI / 23 mm	27	$f/16$ / $f/2.0$ *	$1916 \times 1272, 2032 \times 1135$
Total	89	–	–

D.2. Other Evaluation Data

EBB! Val176. No evaluated method is trained on the EBB! training set, and because the official EBB! test set is not publicly available, we use its validation split. EBB! pairs are captured in the wild with a single Canon 7D DSLR and a 50 mm $f/1.8$ lens, the sharp image at $f/16$ and the bokeh image at $f/1.8$, in automatic mode [21]. The two exposures are not aligned exactly; the authors register them with SIFT and RANSAC, crop them to their intersection, and down-scale them to a height of 1,024 pixels, so pairs may still contain viewpoint changes, exposure differences, black borders, and pixel misalignment. We detect mismatched pairs using SIFT feature matching [31] with RANSAC-based geometric verification [14] then remove visually invalid pairs, retaining **176 pairs (EBB! Val176)**.

RealBokeh. We use the $f/22$ to $f/2.0$ pairs of 219 of the 220 validation scenes [49] at 2000×1500 , with annotated focal points (Table G.1); the edit-consistency study (Appendix H) uses $f/22$ captures of the test split. Both are disjoint from the training pairs that adapt our renderer but share their camera and lens.

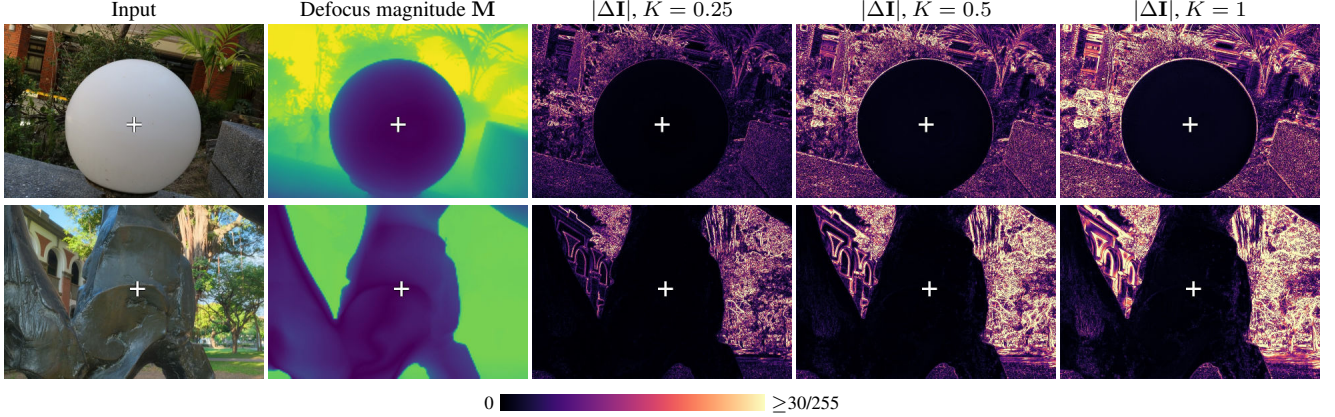


Figure C.2. **Preservation of in-focus regions.** How much the renderer changes each pixel of the restored image, $|\Delta \mathbf{I}| = |\hat{\mathbf{I}}_{\text{bokeh}} - \hat{\mathbf{I}}_{\text{AiF}}|$, when focusing at the white cross with strength $K = 0.25, 0.5$, and 1 ; $\mathbf{M}$ is the defocus magnitude given to the renderer. Pixels on the in-focus subject are left unchanged at every strength, and the change elsewhere grows with $\mathbf{M}$ and K .

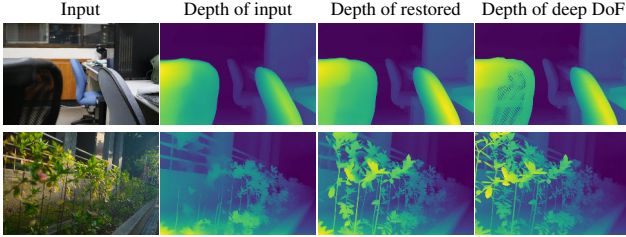


Figure C.3. **Depth source.** DA-V2 on a shallow-depth-of-field photograph of 3CReal, on its restoration by Stage I, and on the deep-depth-of-field capture of the same view (near is bright).

MODEST. MODEST [57] photographs each scene at several f -numbers. From its camera-A 70 mm captures, we pair an $f/22$ input with $f/16, f/9, f/5$, and $f/2.8$ targets of the same viewpoint, resized to a long side of 1536. We keep only aperture series whose targets are aligned with the input and show the same scene content, which gives 17 series from six scenes, and place the focal point at the sharpest region of the $f/2.8$ image.

In-the-wild images. Unsplash [58] photographs have no paired refocused references and are used only for qualitative comparisons, together with our own captures from an iPhone 15 and a Canon EOS 750D.

E. Experiment Settings

Baselines. Baselines use released weights and official inference settings, except Swintormer [9] and Restormer, which are retrained on DPDD with their official code and settings. Deblurring outputs are evaluated at the reference resolution unless otherwise specified.

Bokeh rendering. Focal points are manually annotated for EBB! Val176, 3CReal, and RealBokeh and shared across

methods that support this input; auxiliary depth, disparity, or masks follow official inference settings. Because strength scales differ across methods, each method renders a fixed set of its own strengths and the output with the highest SSIM to the target is evaluated.

End-to-end refocusing. For the qualitative comparisons with the DiT-based DiffCamera [62] and GenRefocus [35], the shorter side of every input is resized to 768 pixels with the aspect ratio kept, and all methods receive the same resized input. DiffCamera requires square inputs, so rectangular images are reflection-padded to a square and the padding is removed after inference.

Efficiency. Cost is measured at 1536×1536 on one NVIDIA RTX PRO 6000 GPU under each method’s official inference settings, including patch- or tile-based inference where applicable. Parameters count every model a method loads, and FLOPs follow PyTorch’s FLOP counter (one multiply-add equals two FLOPs; element-wise operations excluded); for diffusion-based methods the cost covers denoising and VAE encoding and decoding. Deblurring time (Table 2) excludes disk I/O and is averaged over repeated runs after warm-up. In Table 6, *New focus* is each additional focal edit on the same image: FlashReFocus estimates the depth of the restored image once and reuses it, and its memory is the joint peak of both stages.

Metrics. We use PSNR, SSIM [64], LPIPS [73], and DISTS [11] for both stages. Bokeh rendering and refocusing additionally use MS-SSIM [63], LPIPS with VGG features, and CLIP-I, the cosine similarity between CLIP ViT-L/14 [43] image embeddings of output and reference, following GenRefocus [35]. Blur-strength controllability is measured by LVCorr [15] with the protocol of GenRefocus [35]: for each all-in-focus image and a fixed focal point, using the 50 LFDof [46] test scenes and the 89

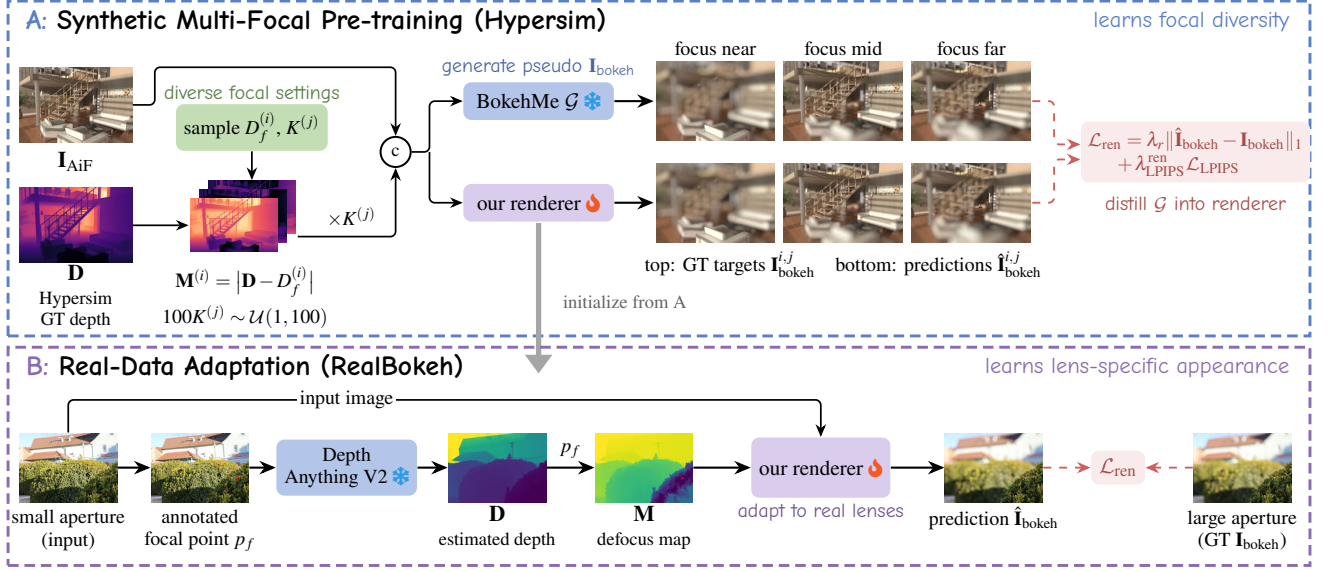


Figure C.4. **Two-phase renderer training.** (A) Pre-training on synthetic multi-focal Hypersim scenes, where each sampled focal depth $D_f^{(i)}$ and strength $K^{(j)}$ gives a new aligned target rendered by BokehMe. (B) Fine-tuning on RealBokeh pairs, with a $10\times$ larger learning rate for the defocus-attention parameters than for the rest of the renderer.

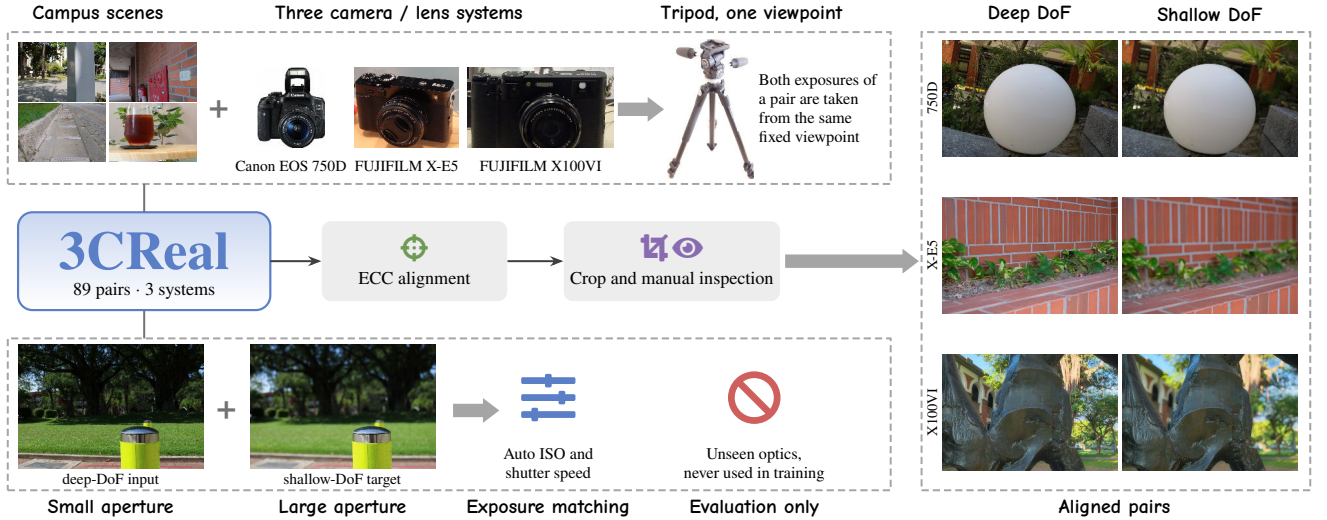


Figure D.1. **3CReal collection.** Campus scenes are photographed with three camera and lens systems from a fixed tripod (top), each at a small and a large aperture with automatic exposure (bottom). The pairs are aligned with ECC, cropped, and inspected (middle), giving 89 aligned pairs (right). Camera and tripod photos from Wikimedia Commons: Canon EOS 750D by Badhon7010 (CC BY-SA 4.0), tripod by Andrés Nieto Porras (CC BY-SA 2.0), FUJIFILM cameras CC0.

3CReal inputs resized to a long side of 1024, a method renders ten strengths evenly spanning its own range, and LVCorr is the Pearson correlation between the strength and the rendered blur, measured as the negative Laplacian variance, averaged over scenes; for GenRefocus, whose ten renderings per scene each require a full diffusion pass, we use the first 20 scenes of each set. No-reference quality

is measured by NIQE [34], MUSIQ [22], MANIQA [66], BRISQUE [33], and CLIPQA [60] for deblurring and by CLIPQA, MUSIQ, and MANIQA for bokeh rendering; deblurring also reports FID [16].

Implementation details. The framework is implemented in PyTorch and trained in half precision. Stage I uses a constant learning rate after warm-up; the pixel phase optimizes

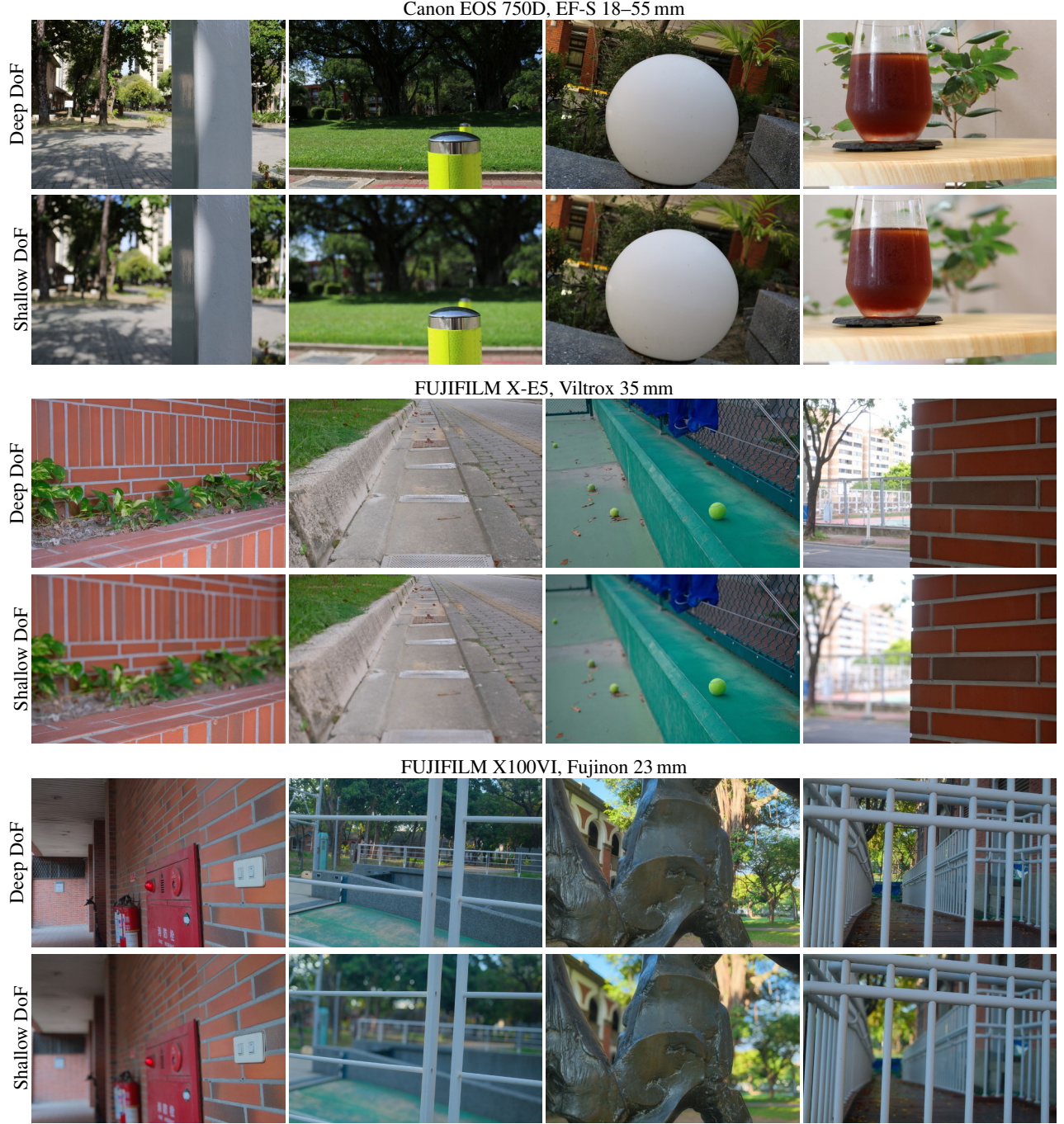


Figure D.2. **More 3CReal pairs.** Four aligned pairs per camera system (Canon EOS 750D with an EF-S 18–55 mm lens, FUJIFILM X-E5 with a Viltrox 35 mm lens, and FUJIFILM X100VI with its 23 mm lens), deep-depth-of-field input above shallow-depth-of-field target. Each pair differs only in aperture; the target shows that lens’s blur on thin structures, depth boundaries, and the background.

the pixel LoRA, the VAE-encoder LoRA, and the U-Net input convolution, and the detail phase optimizes only the detail LoRA. In the refiner phase Stage I is frozen, and the fusion modules and decoder of the refiner are updated in bf16 on 512×512 crops with batch size 4, learning rate

5×10^{-5} with cosine decay, 15K iterations, $\lambda_{\text{struct}} = 0.2$, and $\lambda_{\text{LPIPS}}^{\text{ref}} = 0.1$. Renderer pre-training on Hypersim [44] and adaptation on RealBokeh use cosine decay and $\lambda_r = 1$, $\lambda_{\text{LPIPS}}^{\text{ren}} = 0.2$; in both, the random scale factor is drawn from $[0.4, 1.0]$, flips are horizontal, and rotations are by a random

multiple of 90° . At inference the renderer estimates depth from the restored image with Depth Anything V2 [65].

F. Additional Deblurring Results

Qualitative comparisons. Figures F.1 and F.2 add a DPDD scene with rocks, grass, and leaves and a RealDOF scene with dense structure. Natural textures become sharper than with the deterministic baselines, with fewer content changes than GenRefocus introduces, while small characters and repeated man-made patterns remain susceptible to incorrect reconstruction. Figure F.6 deblurs shallow-depth-of-field photographs of 3CReal, whose blur comes from real lenses that Stage I has not seen, next to the deep-depth-of-field captures of the same views.

Failure cases. Figure F.3 shows that heavily defocused foreground regions may become over-smoothed or exhibit ghosting when the input provides insufficient structure. Textureless regions may also produce web-like patterns, because the diffusion prior can hallucinate weak high-frequency content; the detail-weight study below shows how this grows with ω .

Ablation. Table F.1 reports all metrics for the cumulative deblurring ablation of the main paper.

Detail weight. Table F.2 varies ω on the held-out sets. PSNR peaks at $\omega = 0.5$ – 0.75 , whereas LPIPS, DISTs, and FID improve up to $\omega = 1.25$ – 1.5 . Beyond it, all metrics degrade as the extrapolated detail branch hallucinates texture. We use $\omega = 1.25$, which has the lowest rank sum.

Figure F.4 shows where the detail branch acts: the difference between the outputs at $\omega = 1.25$ and $\omega = 0$ follows bark, foliage, and print and is small on smooth surfaces. Over the 76 DPDD test images it averages 4.3/255; 32.5% of it falls on the 20% of pixels with the most texture in the sharp target, where the error to the target decreases from 18.09/255 to 17.66/255, while the error on the smoothest half stays at 7.3/255.

Decoder refiner. Figure F.5 shows that text regions relatively clear in the input can still be distorted by latent encoding and decoding, and that the decoder-side high-frequency injection recovers them; severely blurred characters remain limited because their structure cannot be inferred reliably.

G. Additional Bokeh Results

RealBokeh validation. RealBokeh validation shares its cameras and lenses with the RealBokeh training split, on which both Bokehlicious and our renderer are trained, so Table G.1 measures in-distribution rendering at the original resolution. Bokehlicious, which is conditioned on the target aperture, obtains the best full-reference scores. FlashReFocus follows on PSNR, MS-SSIM, LPIPS, DISTs, and CLIP-I and has the second-best MUSIQ, and on the ten

pairs evaluated for GenRefocus it is better on every metric (PSNR 25.46 vs. 23.38 dB, MUSIQ 47.61 vs. 38.79).

Real apertures. LVCorr measures whether blur grows with the strength setting, but not whether the selected strength follows a real aperture. On MODEST, the strength selected for FlashReFocus rises with the aperture in 16 of the 17 aperture series, with the highest rank correlation to $1/N$ ($\rho = 0.985$, Table G.2), on par with the physically based BokehMe (0.979) and above the diffusion-based MagicBokeh (0.903), whose smallest setting already blurs more than $f/16$. Its SSIM matches BokehMe at small apertures and is the highest at $f/5$ and $f/2.8$. The strength of FlashReFocus is monotonic but not linear in $1/N$ (Fig. G.1), unlike the blur-radius parameters of BokehMe and DeepLens, so mapping f-numbers to strengths needs a per-lens calibration curve rather than a single scale.

Qualitative comparisons. Figure G.3 adds an EBB! Val176 case, Figs. G.4 and G.5 two 3CReal cases, and Figs. G.6 and G.7 two RealBokeh validation pairs, where GenRefocus is included; together they illustrate subject preservation, highlight appearance, and boundary transitions on real captures. Figure G.8 places FlashReFocus renders next to real captures of two MODEST aperture series.

Ablation. Table G.3 extends the main-paper ablation with LPIPS-VGG and four-decimal precision. The mixed (relative-to-defocus) variant uses relative-position attention in the first half of the attention blocks and defocus-bias attention in the second half; it ties defocus-bias attention on MS-SSIM and is marginally higher on CLIP-I (0.9811 vs. 0.9807), while defocus-bias attention throughout is best on all five other metrics. Figure G.2 illustrates the differences in highlight halos and low-light blur.

Depth source. The renderer is trained with Depth Anything V2 Large (DA-V2-L) depth, but it reads depth only through the normalized map $\mathbf{D}$, so any monocular estimator can be used at inference without retraining. Table G.4 renders the same inputs with the smaller DA-V2 models and with Depth Pro [8], the depth estimator of GenRefocus, under the protocol of Table 3. Across a $38\times$ range of depth-model size, PSNR changes by at most 0.21 dB and LPIPS by at most 0.009: weaker depth from DA-V2-S costs 0.11 dB on 3CReal, and Depth Pro gains 0.10 dB and improves LPIPS and DISTs on both benchmarks. Errors of the depth estimator therefore affect the rendering only mildly, and a stronger estimator can be swapped in at the cost of a slower first edit (Depth Pro takes about $1.6\times$ as long as DA-V2-L at 1536×1536 and has $2.8\times$ its parameters).

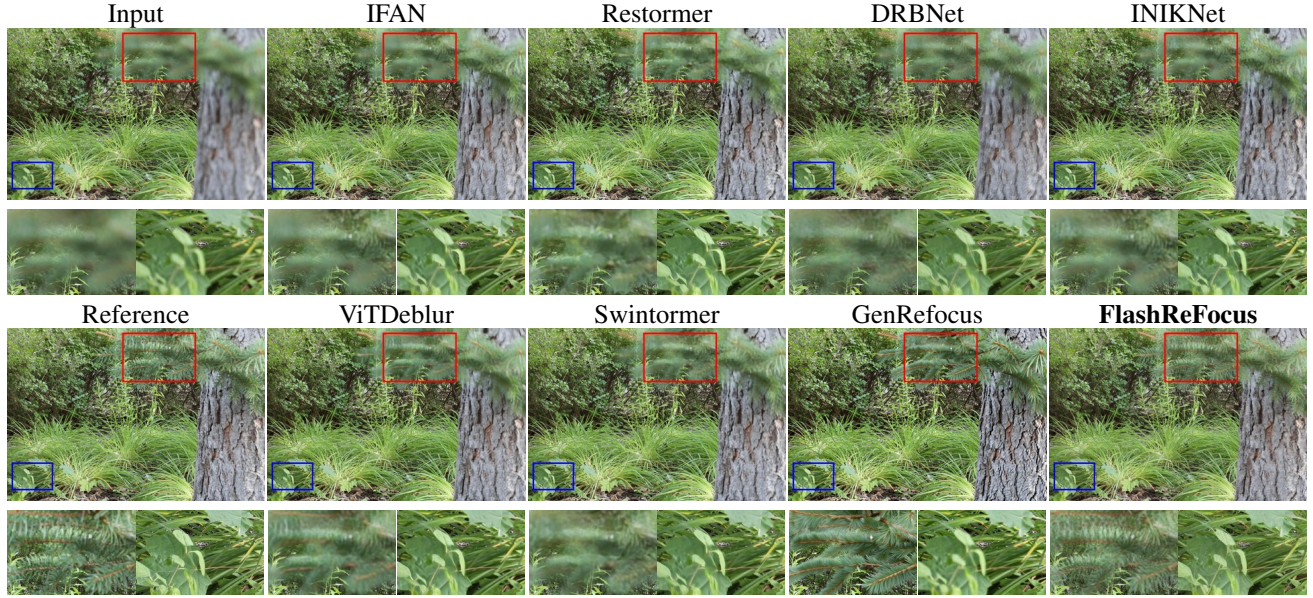


Figure F.1. **DPDD, texture-rich scene.** Deblurring of a DPDD test scene with grass, leaves, and tree bark by the compared methods, next to the reference; two crops (red and blue boxes) are enlarged below each result. The deterministic baselines leave the blades and the bark soft, while FlashReFocus recovers edges and texture closer to the reference.



Figure F.2. **RealDOF, dense structure.** A RealDOF outdoor scene with shop signs, window frames, and foliage, as in Fig. F.1. Most methods recover large letters and frames; crops show where they differ on fine text and repeated structures, the scene’s hardest parts.

Table G.4. **Depth source** for the same renderer, without retraining. Parameters in millions; the renderer was trained with DA-V2-L.

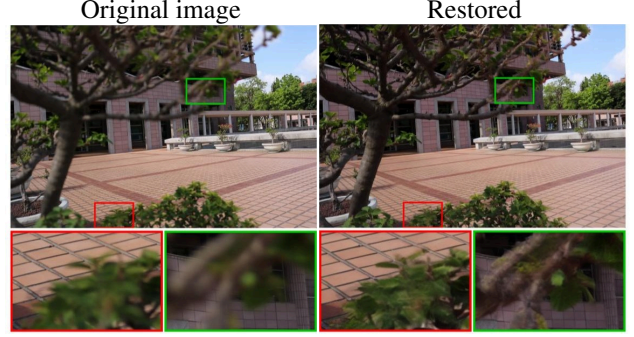
Depth	Params	3CReal			EBB! Val176		
		PSNR↑	LPIPS↓	DISTS↓	PSNR↑	LPIPS↓	DISTS↓
DA-V2-S	25	27.87	.2232	.1390	24.49	.1804	.1026
DA-V2-B	98	27.99	.2198	.1373	24.51	.1778	.1007
DA-V2-L (ours)	335	27.98	.2181	.1372	24.52	.1780	.1019
Depth Pro	952	28.08	.2145	.1331	24.53	.1746	.0983

H. Refocusing and Editing

Complete refocusing on LFDof. Light fields give exact refocusing ground truth. Following the LF-Refocus protocol of GenRefocus [35], we take the 50 LFDof test scenes [46] at the largest aperture and, for each, use one de-focused focal slice as input and another slice of the same



(a) Grid and colour-shift artefacts



(b) Oversmoothed and ghosting artefacts

Figure F.3. **Failure cases.** (a) Grid-like texture and a color shift in a low-texture region, where the diffusion prior adds unsupported high-frequency content. (b) Over-smoothing and ghosting of a severely blurred foreground subject, whose structure the input lacks.

Table F.1. **Deblurring ablation with all metrics.** Starting from the official PiSA-SR recipe trained on our data, our pixel phase (Pix.), the decoder refiner (Ref.), and $\omega = 1.25$ are added in turn; bold marks the best value per dataset.

Dataset	Setting	Pix.	Ref.	ω	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	DISTS $\downarrow$	FID $\downarrow$	NIQE $\downarrow$	MUSIQ $\uparrow$	MANIQA $\uparrow$	BRISQUE $\downarrow$	CLIPQA $\uparrow$
DPDD	Official PiSA-SR	$\times$	$\times$	$\times$	24.57	0.730	0.177	0.097	39.70	4.87	60.08	0.374	20.60	0.525
	+ our pixel phase	$\checkmark$	$\times$	$\times$	24.61	0.735	0.172	0.098	36.18	4.69	60.39	0.372	21.37	0.505
	+ refiner	$\checkmark$	$\checkmark$	$\times$	25.51	0.790	0.167	0.093	34.74	4.70	59.07	0.353	23.26	0.477
	+ $\omega = 1.25$	$\checkmark$	$\checkmark$	$\checkmark$	25.37	0.781	0.156	0.087	33.13	4.52	59.60	0.358	19.47	0.494
RealDOF	Official PiSA-SR	$\times$	$\times$	$\times$	23.77	0.710	0.260	0.122	32.94	5.01	37.94	0.237	38.02	0.320
	+ our pixel phase	$\checkmark$	$\times$	$\times$	23.70	0.709	0.253	0.123	30.38	4.75	44.30	0.260	36.03	0.334
	+ refiner	$\checkmark$	$\checkmark$	$\times$	24.54	0.752	0.248	0.120	28.44	5.33	37.35	0.223	37.77	0.277
	+ $\omega = 1.25$	$\checkmark$	$\checkmark$	$\checkmark$	24.25	0.737	0.244	0.119	28.82	5.02	38.88	0.227	34.63	0.289

light field as target, with a focal point annotated where the target is sharp. Every method refocuses the input at that point, and its strength (or blur level) is searched per image for the highest SSIM to the target. Table H.1 shows that FlashReFocus is best on all full-reference metrics, and Figs. H.1 and H.6 show four scenes; the ablations show both stages contribute. Two caveats apply: Stage I is trained on LFDOF training scenes, so its input domain is familiar, and these 50 test scenes are in the held-out data used to select the refiner and ω (Sec. 4), although the renderer and the evaluation targets were not used for any choice.

Table H.1. **Complete refocusing on LFDOF** (50 test scenes, input and target are two focal slices of one light field). Strength searched per image for the highest SSIM.

Method	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	DISTS $\downarrow$	CLIP-I $\uparrow$
Input	25.97	.779	.319	.152	.945
DiffCamera [62]	27.23	.824	.204	.124	.955
GenRefocus [35]	26.07	.840	.173	.101	.961
MagicBokeh [53]	26.36	.799	.234	.160	.921
FlashReFocus w/o Stage I	28.46	.848	.227	.142	.953
FlashReFocus w/o refiner	27.40	.814	.172	.110	.962
FlashReFocus	29.39	.867	.144	.096	.970

Refocusing examples. Figure H.3 compares FlashReFocus with DiffCamera, GenRefocus, and MagicBokeh [53] on two Unsplash images. The newly selected focal regions, the cliff in the first image and the foreground rock in the second, are out of focus in the input. MagicBokeh does not restore the input, so these regions remain blurred, whereas FlashReFocus recovers them before rendering. Figure H.4 follows complete refocusing on photographs from a Canon EOS 750D and an iPhone 15 stage by stage: the input, the restored all-in-focus image, and the rendered result. Figure H.5 moves the focus of three Unsplash images from the foreground subject to the background; enlarged crops show detail in the new focal region and the transition into the blurred foreground. In its third image the face behind the globe is blurred in the input, and FlashReFocus recovers it less sharply than GenRefocus and MagicBokeh, which synthesize new facial detail. Figure H.2 applies the tilt-shift edit of Fig. 10b to a portrait. Figure H.11 refocuses four Unsplash portraits from the face to the background, each from one restored image. Figure H.12 varies the strength for two of them with the focus on the face. Figure H.14 moves the focus within a flower bed that surrounds the subject.

Consistency across focal edits. Interactive refocusing moves the focal point many times on the same image, so each edit should change the blur and nothing else. We place

Table F.2. **Detail weight ω on the held-out sets.** Stage I with the refiner; bold marks the best value per column. The rank sum over PSNR, LPIPS, DISTS, and FID on all three sets selects $\omega = 1.25$.

ω	DPDD val				LFDOF test				RealBokeh val60				Rank sum $\downarrow$
	PSNR $\uparrow$	LPIPS $\downarrow$	DISTS $\downarrow$	FID $\downarrow$	PSNR $\uparrow$	LPIPS $\downarrow$	DISTS $\downarrow$	FID $\downarrow$	PSNR $\uparrow$	LPIPS $\downarrow$	DISTS $\downarrow$	FID $\downarrow$	
0	25.08	0.288	0.149	38.37	28.90	0.187	0.107	26.17	21.42	0.460	0.231	109.53	75
0.5	25.27	0.234	0.121	34.75	29.22	0.144	0.084	20.07	21.54	0.413	0.203	101.57	59
0.75	25.27	0.205	0.106	32.19	29.27	0.125	0.073	17.95	21.52	0.377	0.184	93.82	48
1	25.20	0.182	0.094	30.95	29.21	0.111	0.065	16.75	21.41	0.342	0.166	87.97	39
1.25	25.02	0.169	0.087	30.37	28.98	0.104	0.062	16.27	21.14	0.318	0.151	85.48	29
1.5	24.65	0.170	0.086	30.34	28.54	0.105	0.063	16.63	20.65	0.308	0.144	87.91	32
1.75	24.02	0.185	0.094	32.05	27.88	0.113	0.069	18.75	20.00	0.316	0.146	93.28	54

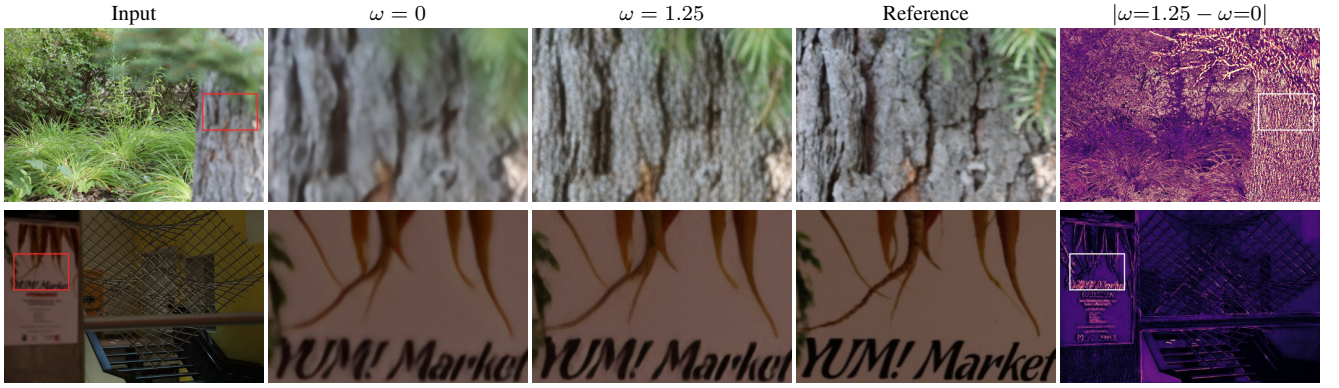


Figure F.4. **Detail branch of Stage I.** Outputs on DPDD of the pixel branch alone ($\omega = 0$) and with the detail branch ($\omega = 1.25$) in the red box, the reference, and the absolute difference between the two outputs over the whole image, from 0 (dark) to 20/255 or more (bright).

eight focal points at depth quantiles from far to near over three 3CReal and three RealBokeh test inputs at a long side of 1024, and compare with GenRefocus [35] on two measures. *In-focus fidelity* is the PSNR and LPIPS between a result and the all-in-focus input over pixels whose depth lies within 0.04 of the focal depth. *Edit consistency* is the PSNR between consecutive results over pixels whose defocus does not change between the two settings, i.e. whose distances to the two focal depths differ by less than 0.02. Both use the depth of our renderer.

The two methods’ out-of-focus blur is equalized, measured as the mean absolute Laplacian over pixels more than 0.25 from the focal depth relative to the input, in two settings. In setting (i), GenRefocus runs at its default blur parameter $K_G = 15$ and FlashReFocus’s strength K is matched to it, per focal setting for fidelity and per scene for consistency; blur beyond the trained range is obtained by extrapolating the strength up to 2 or applying the renderer twice. In setting (ii), FlashReFocus runs at its default strength $K = 0.5$ and K_G is set per scene to the same blur ($K_G = 8.5\text{--}14.19$ on 3CReal and $5.25\text{--}7.69$ on RealBokeh).

In both settings FlashReFocus preserves in-focus regions far better (Table H.2): 44.44 dB against 28.96 dB on 3CReal

in setting (i). The residual renderer only adds blur to the restored image, whereas GenRefocus regenerates the whole image at every edit, so its in-focus regions drift from the input and pixels whose blur should not change still vary; between consecutive edits they change 2.5 to 3.6 times more than with FlashReFocus. An edit takes 0.10 s against 77 s. Figure H.7 shows four of the eight settings for two scenes of each dataset under setting (i), and Fig. H.8 one scene of each under setting (ii).

Two focal planes. A physical lens has one focal plane, so two subjects at different depths cannot both be sharp while the space between them is blurred. With an explicit depth map, FlashReFocus obtains this without retraining. With the focal depths D_A and D_B of two selected points, the renderer receives the edited depth $D' = D_A + \min(|D - D_A|, |D - D_B|)$ and the focal depth D_A , so each pixel is blurred by its distance to the nearer plane. Figure H.9 shows three scenes, the third from 3CReal. Measured as the mean absolute Laplacian relative to the input, the two planes keep 84%/75%, 81%/82% and 81%/66% of their sharpness with the edited depth, whereas focusing on one plane leaves the other at 9–21%; regions far from both planes keep 8–

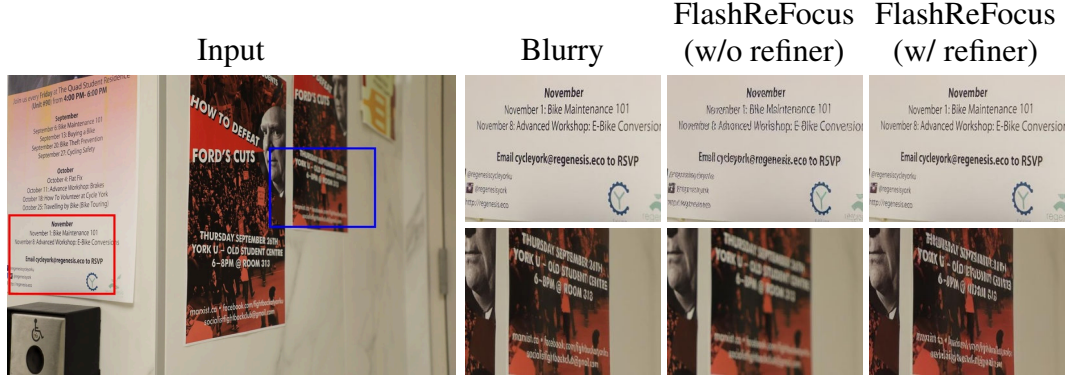


Figure F.5. **Decoder refiner on text.** Crops of printed text in the blurred input and in the outputs of FlashReFocus decoded without and with the refiner. Letters that are legible in the input are distorted by plain VAE decoding and recovered by the refiner, which reinjects the input’s high-frequency structure; severely blurred characters stay hard to recover.

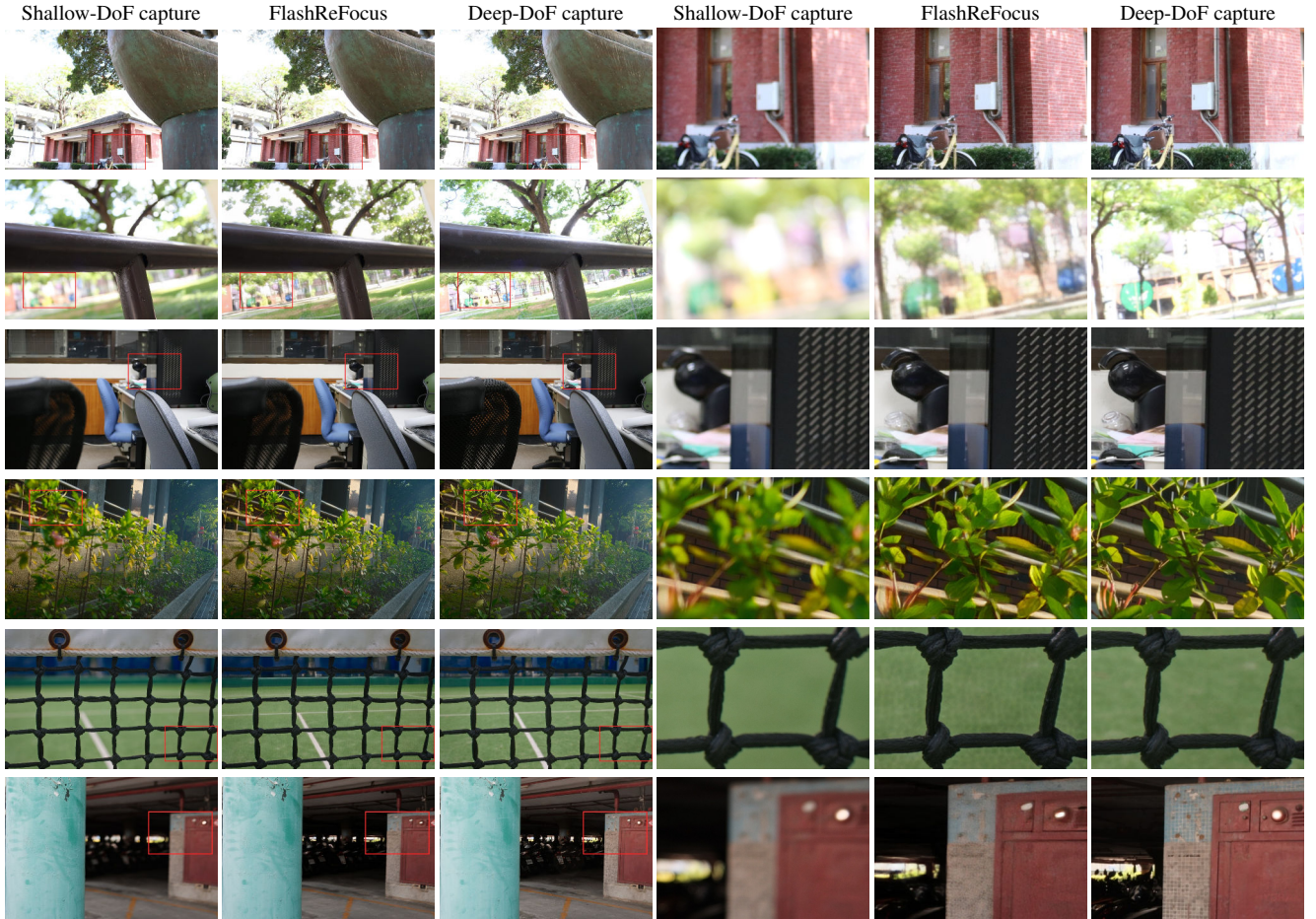


Figure F.6. **Deblurring real lens blur.** Shallow-depth-of-field photographs of 3CReal from three unseen cameras and lenses, restored by Stage I, next to the deep-depth-of-field capture of the same view; the red box is enlarged on the right.

Table G.1. **Bokeh rendering on RealBokeh validation.** 219 pairs from $f/22$ to $f/2.0$ at the original 2000×1500 resolution, with the annotated focal points and the protocol of Table 3. Darker green marks the best, second, and third results. GenRefocus is evaluated on the first 10 pairs (bottom) because of its cost at this resolution. [#] Selects its own focal region. [†] Trained on RealBokeh pairs (Bokehlicious and FlashReFocus: the training split; GenRefocus does not state the split).

Methods	Full-reference Metrics							No-reference IQA Metrics		
	PSNR $\uparrow$	SSIM $\uparrow$	MS-SSIM $\uparrow$	LPIPS $\downarrow$	LPIPS _{VGG} $\downarrow$	DISTS $\downarrow$	CLIP-I $\uparrow$	CLIP-IQA $\uparrow$	MUSIQ $\uparrow$	MANIQA $\uparrow$
<i>All 219 pairs</i>										
DeepLens (TOG'18) [61]	22.86	0.8297	0.8011	0.3245	0.3792	0.1698	0.9416	0.3195	43.81	0.2888
BokehMe (CVPR'22) [37]	25.68	0.9027	0.9010	0.2275	0.3295	0.1584	0.9627	0.4572	43.22	0.3362
Bokehlicious ^{#†} (ICCV'25) [49]	29.81	0.9320	0.9568	0.1086	0.1802	0.0697	0.9869	0.3604	42.83	0.3225
BokehDiff [#] (ICCV'25) [76]	22.92	0.8075	0.8038	0.3312	0.3855	0.1745	0.9401	0.4274	49.77	0.3378
MagicBokeh (CVPR'26) [53]	22.25	0.8267	0.8229	0.2730	0.3578	0.1751	0.9342	0.4026	45.54	0.3016
FlashReFocus [†]	26.38	0.8973	0.9042	0.1932	0.2911	0.1178	0.9701	0.3967	48.11	0.3270
<i>First 10 pairs</i>										
GenRefocus [†] (arXiv'25) [35]	23.38	0.8301	0.8516	0.2641	0.3618	0.1830	0.9504	0.2800	38.79	0.3034
FlashReFocus [†]	25.46	0.8623	0.8705	0.2462	0.3389	0.1495	0.9744	0.3990	47.61	0.3157

Table G.2. **Real apertures on MODEST.** 17 aperture series (camera A, 70 mm), $f/22$ input and $f/16$, $f/9$, $f/5$, $f/2.8$ targets. Each method's strength is searched per image as in Table 3. Monotonic: series whose best strength never decreases as the aperture opens; ρ : mean Spearman correlation between best strength and $1/N$ per series.

Method	Monotonic $\uparrow$	$\rho \uparrow$	SSIM $\uparrow$			
			$f/16$	$f/9$	$f/5$	$f/2.8$
DeepLens [61]	16/17	0.917	0.9419	0.9070	0.9007	0.8832
BokehMe [37]	16/17	0.979	0.9478	0.9168	0.9129	0.8975
MagicBokeh [53]	14/17	0.903	0.8547	0.8548	0.8729	0.8825
FlashReFocus	16/17	0.985	0.9464	0.9159	0.9141	0.9041

11%. Figure H.13 applies the same edit to two portraits of Fig. H.11, keeping the face and a distant region sharp.

Controls on a portrait. Figure H.10 applies the controls of FlashReFocus to one Unsplash portrait with three depth layers: the can held towards the camera, the face, and the lake behind. All nine results reuse one restored image and depth, one renderer pass each. Moving the focal point blurs whichever layers are away from it, raising the strength with the focus on the face progressively blurs the can, and the depth edit of Appendix H keeps the can and the face sharp while the lake stays blurred.

I. Limitations and Broader Impact

Depth. The renderer reads the relative depth of a monocular model, so its errors become blur errors: a misestimated thin structure, glass surface, or reflection is blurred as if

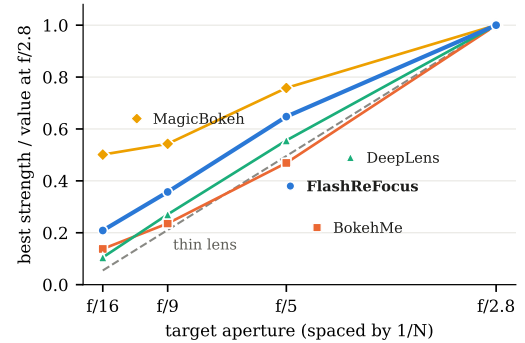


Figure G.1. **Blur strength against aperture on MODEST.** Strength selected for each method, averaged over the 17 aperture series and normalized at $f/2.8$. The dashed line is the thin-lens blur relative to the $f/22$ input.

it lay at the wrong depth. Replacing DA-V2-L by Depth Pro or by smaller DA-V2 models changes PSNR by at most 0.11 dB on the benchmarks (Table G.4), but these averages do not bound the error on a single image with such content.

Occlusion. A real large aperture sees slightly around a foreground edge, so the background behind it shows through the blurred edge. FlashReFocus renders from one restored image that contains no occluded content, so it can only spread visible colors, and the background hidden behind the blurred rim of a near subject is not recovered.

Lens appearance. The renderer learns the blur of the lenses in its training data; it has no input for the aperture shape, which GenRefocus [35] can condition on, or for lens-specific effects such as cat-eye highlights. Strengths above 1 ($\dagger$ in Fig. H.10) are extrapolated.

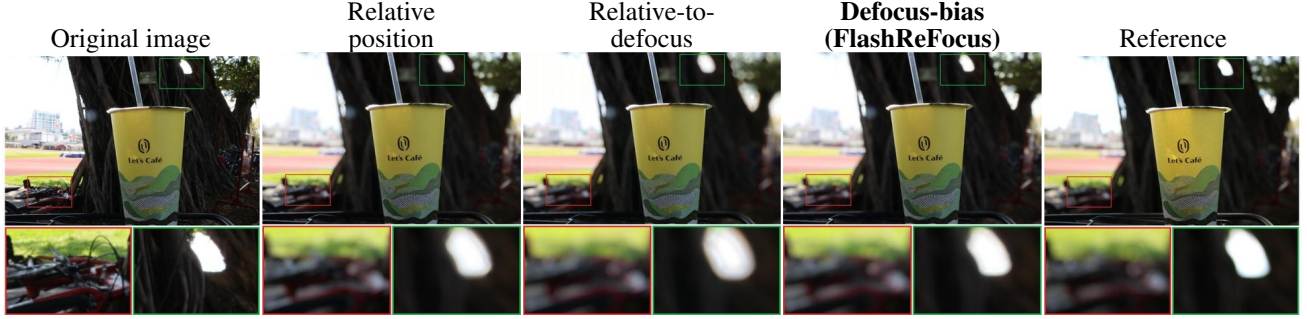


Figure G.2. **Attention design.** One 3CReal scene rendered by renderers that differ only in the attention: relative-position attention as in Bokehlicious, relative position modulated by defocus, and the defocus bias of FlashReFocus, next to the real shallow-depth-of-field capture. Enlarged crops are shown below each result.

Table G.3. **Bokeh ablation on 3CReal.** (a) Training strategy: renderers trained on RealBokeh only, on Hypersim only, pre-trained on Hypersim and fine-tuned on RealBokeh, and fine-tuned with a separate learning rate for the defocus parameters (ours). (b) Attention design under the same training. All variants are evaluated zero-shot on 3CReal with the protocol of Table 3.

(a) Training strategy

Training strategy	PSNR $\uparrow$	SSIM $\uparrow$	MS-SSIM $\uparrow$	LPIPS $\downarrow$	LPIPS _{VGG} $\downarrow$	DISTS $\downarrow$	CLIP-I $\uparrow$
RealBokeh only	27.2844	0.8642	0.9173	0.2505	0.3301	0.1389	0.9734
Hypersim only	27.4598	0.8815	0.9324	0.2513	0.3362	0.1637	0.9759
Full fine-tuning	27.9383	0.8850	0.9363	0.2254	0.3075	0.1416	0.9802
Diff. LR fine-tuning	27.9815	0.8855	0.9375	0.2178	0.3004	0.1368	0.9807

(b) Attention design

Attention design	PSNR $\uparrow$	SSIM $\uparrow$	MS-SSIM $\uparrow$	LPIPS $\downarrow$	LPIPS _{VGG} $\downarrow$	DISTS $\downarrow$	CLIP-I $\uparrow$
Relative position	27.8976	0.8838	0.9362	0.2232	0.3055	0.1386	0.9800
Relative-to-defocus	27.9555	0.8845	0.9375	0.2221	0.3049	0.1390	0.9811
Defocus bias (FlashReFocus)	27.9815	0.8855	0.9375	0.2178	0.3004	0.1368	0.9807

Restoration. Stage I can add texture that the input does not support or leave heavily blurred regions soft (Fig. F.3). When a face is strongly blurred in the input, FlashReFocus recovers it less sharply than generative refocusing methods, which synthesize new facial detail (Fig. H.5, third row).

Broader impact. Refocusing after capture can rescue photographs with a missed focus and gives casual photographers depth-of-field control without special optics. The restored detail, however, is partly generated by a diffusion prior, so a refocused image should not be treated as a record of what the camera captured, for example as evidence. The portraits in this appendix are Unsplash photographs used under the Unsplash License [58].

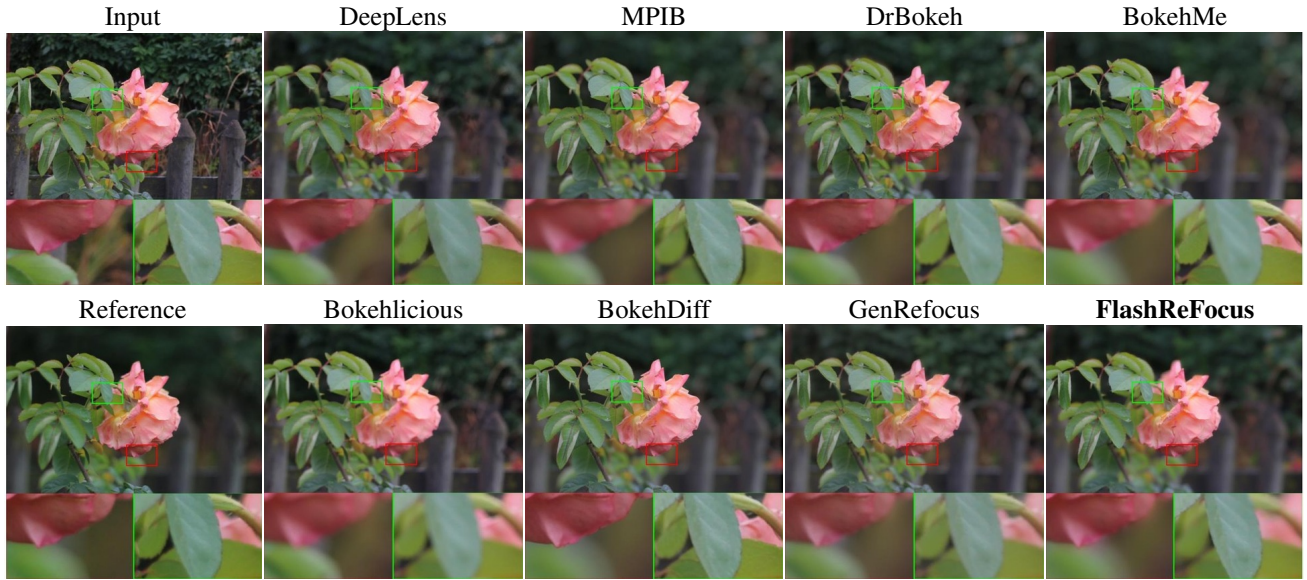


Figure G.3. **EBB! Val176, foreground focus.** A flower in front of foliage, rendered by each method from the same all-in-focus input and focal point, next to the EBB! reference. Crops below each result compare the edge of the in-focus flower and the blurred background.

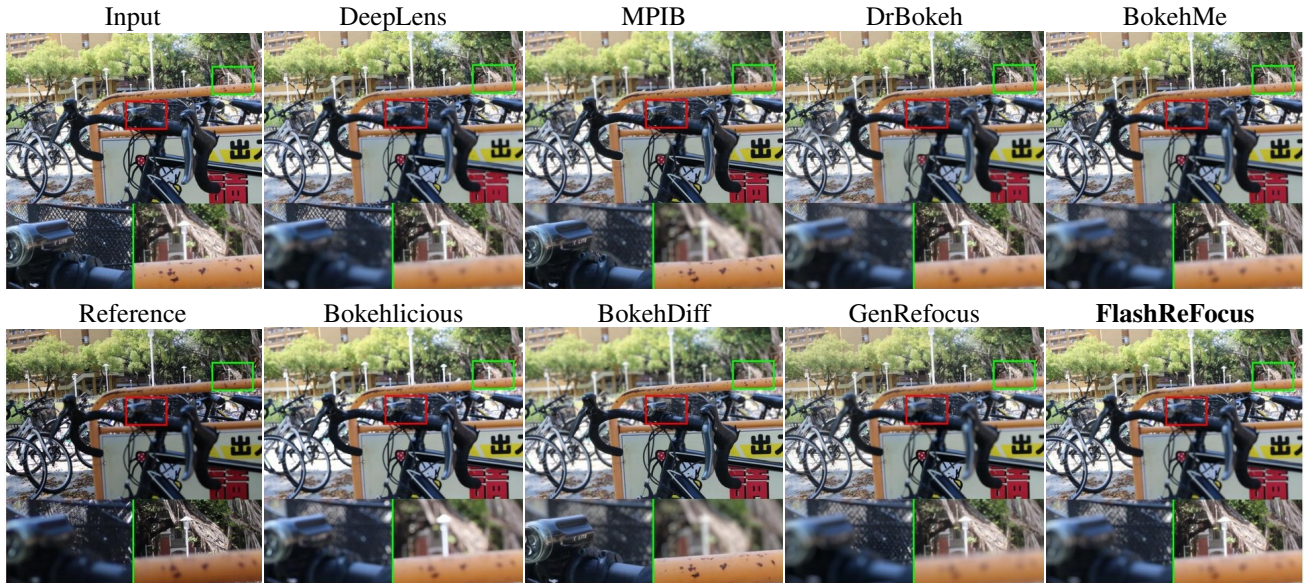


Figure G.4. **3CReal, unseen optics.** A street scene with bicycles from a 3CReal camera that no compared method has seen in training. All methods render the same input at the annotated focal point; the reference is the real capture, and insets enlarge the boxed regions.

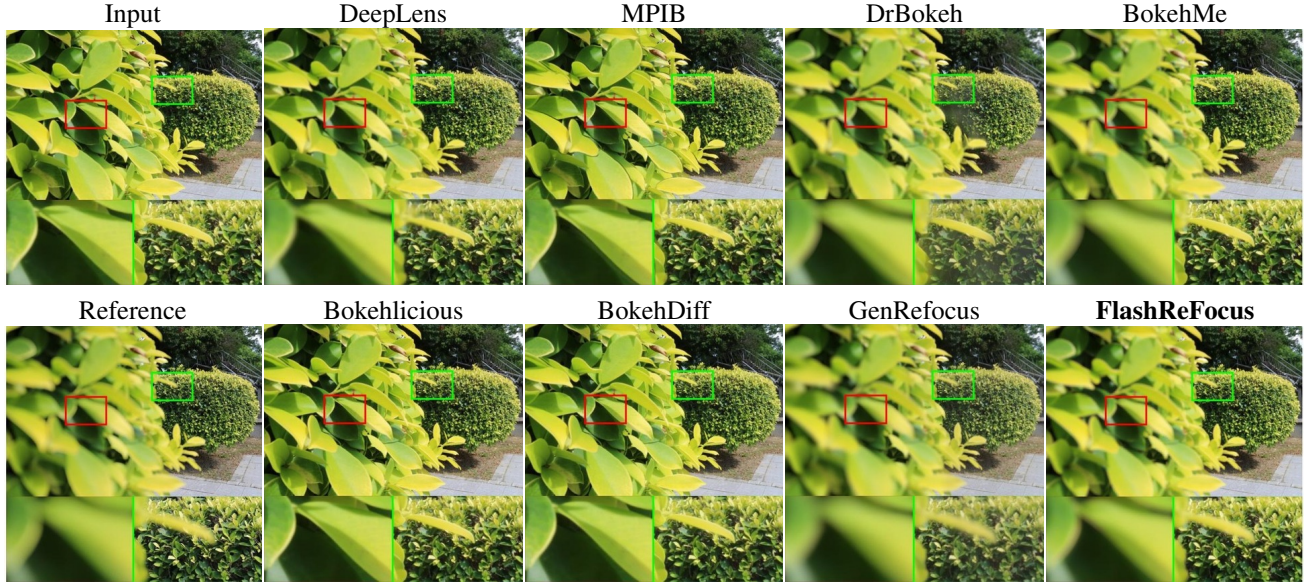


Figure G.5. **3CReal, lens appearance.** Leaves in front of a hedge, where the blur at depth boundaries and the shape of out-of-focus highlights depend on the lens. Each rendering is compared with the real shallow-depth-of-field capture; the insets enlarge the boundary between the in-focus leaves and the background.

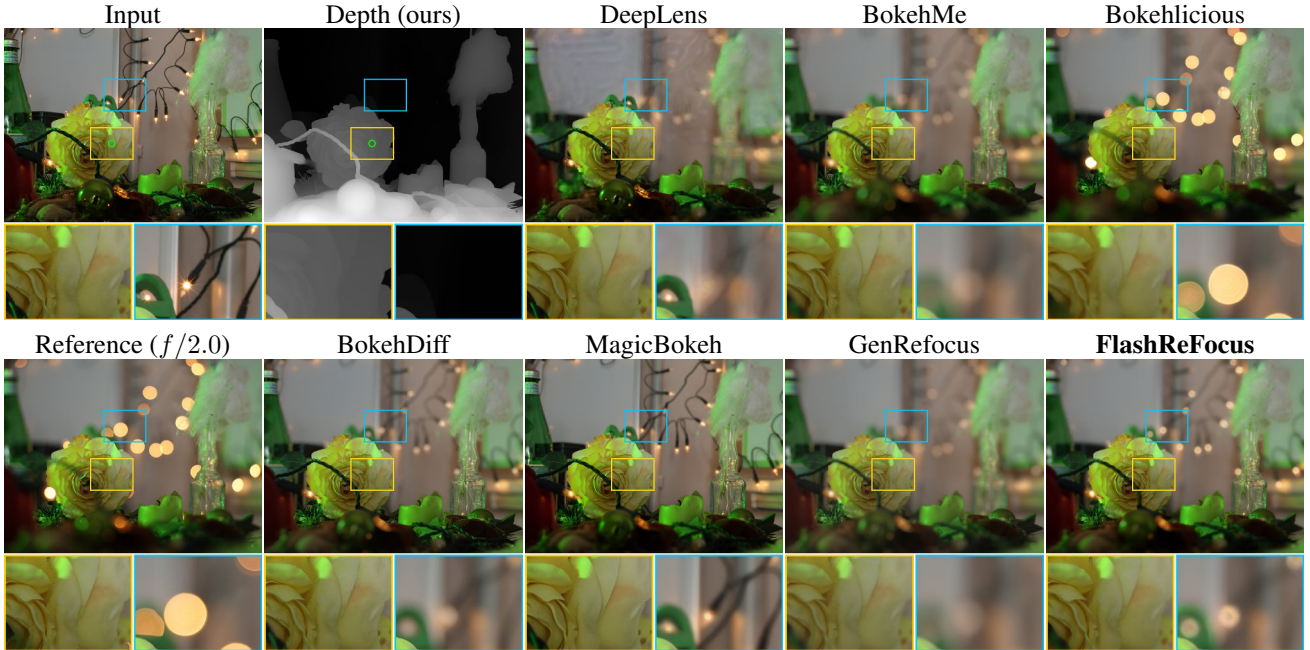


Figure G.6. **RealBokeh validation, point lights.** Focus on the flower (green marker); yellow box: in-focus crop, blue box: where the $f/2.0$ reference differs most from the input. Bokehlicious, trained on the same cameras and lenses, reproduces the large highlight discs of the reference; FlashReFocus renders smaller, softer discs, and GenRefocus leaves the background lights nearly sharp.

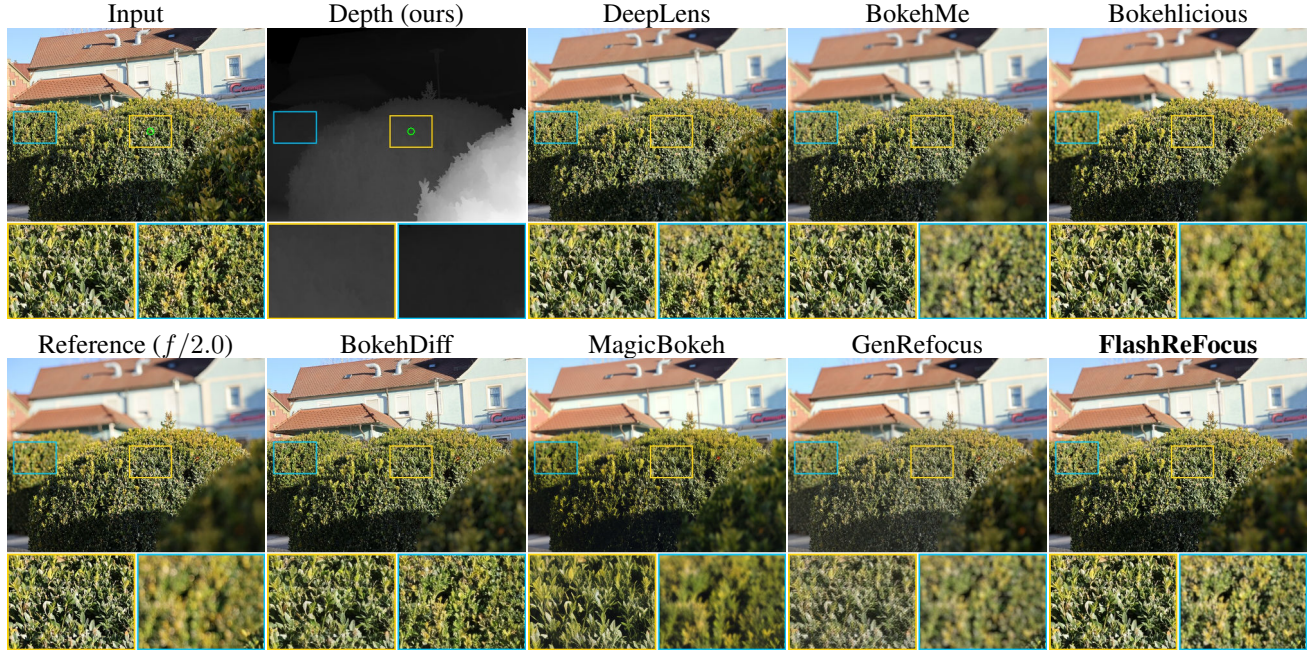


Figure G.7. **RealBokeh validation, foreground focus.** Focus on the hedge in front of a house: the hedge should stay sharp while the roof and the sky blur as in the $f/2.0$ reference. The depth of our renderer is shown next to the input; boxes are enlarged as in Fig. G.6.

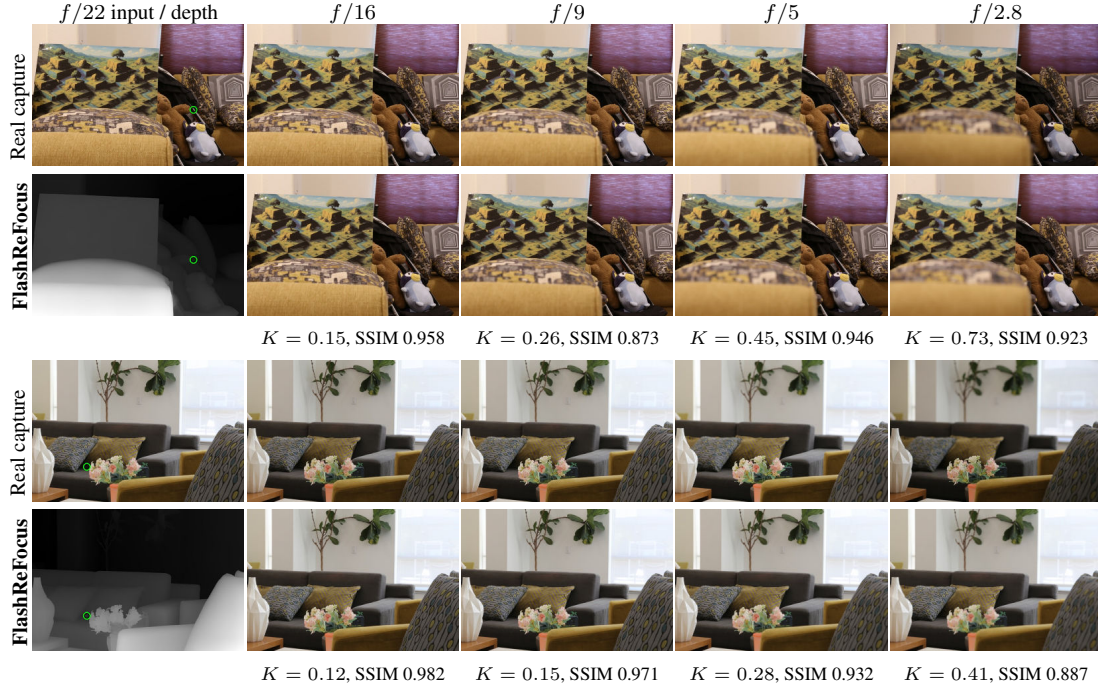


Figure G.8. **MODEST aperture series.** For two series, real captures of the same view (top) and FlashReFocus renders of the $f/22$ input (bottom) at the strength K selected for each aperture; the first column of each bottom row shows the depth of our renderer (near bright) and the focal point (green marker). The selected strength grows with the aperture; at $f/2.8$ the real lens still blurs the near foreground of the first series more strongly than FlashReFocus.

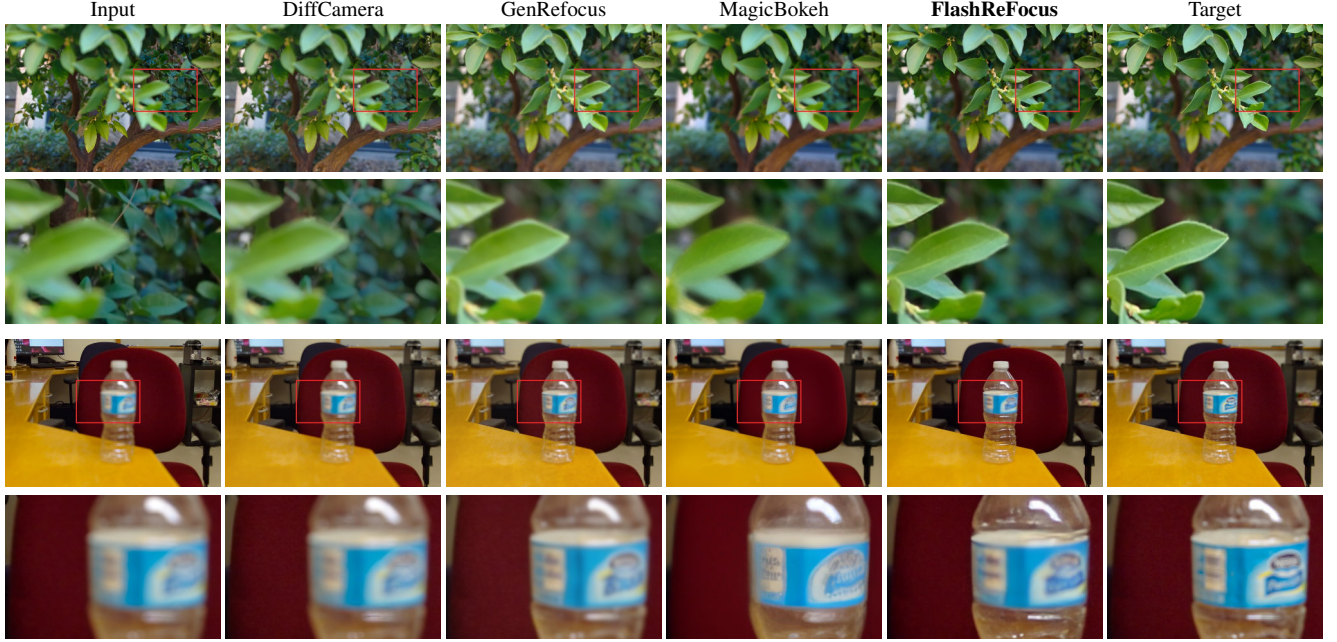


Figure H.1. **Complete refocusing on LFDOF.** Each input is one focal slice of a light field and the target another slice of the same light field, with the focus moved onto the foreground leaves (top) or the bottle (bottom); the red box is enlarged below each image. Only FlashReFocus brings the new focal region to the sharpness of the target, as the restored all-in-focus image keeps the detail that the renderer then leaves in focus; the generative baselines keep it soft.

Table H.2. **Consistency across focal edits.** Three scenes per dataset with eight focal points each. In-focus fidelity (PSNR, LPIPS) compares the in-focus region with the input; consistency (PSNR) compares consecutive edits where the blur should not change. In setting (i), FlashReFocus is matched to the blur of GenRefocus at $K_G = 15$; in setting (ii), GenRefocus is matched to FlashReFocus at $K = 0.5$.

Method	3CReal						RealBokeh test					
	Setting (i)			Setting (ii)			Setting (i)			Setting (ii)		
	In-focus			In-focus			In-focus			In-focus		
	PSNR $\uparrow$	LPIPS $\downarrow$	Consist. PSNR $\uparrow$	PSNR $\uparrow$	LPIPS $\downarrow$	Consist. PSNR $\uparrow$	PSNR $\uparrow$	LPIPS $\downarrow$	Consist. PSNR $\uparrow$	PSNR $\uparrow$	LPIPS $\downarrow$	Consist. PSNR $\uparrow$
GenRefocus	28.96	0.155	43.84	29.37	0.123	45.16	25.38	0.186	40.19	27.10	0.118	42.56
FlashReFocus	44.44	0.036	49.91	42.44	0.043	49.74	38.81	0.048	49.31	43.54	0.023	51.47



Figure H.2. **Tilt-shift on a portrait.** The depth given to the renderer is replaced by a vertical ramp around a chosen row, at strength $K = 1$. A horizontal band stays sharp regardless of scene depth: at the face (left) the distant city and the valley blur, and at the valley (right) the buildings and the road are sharp while the subject in front blurs, giving the miniature look of a tilted lens.

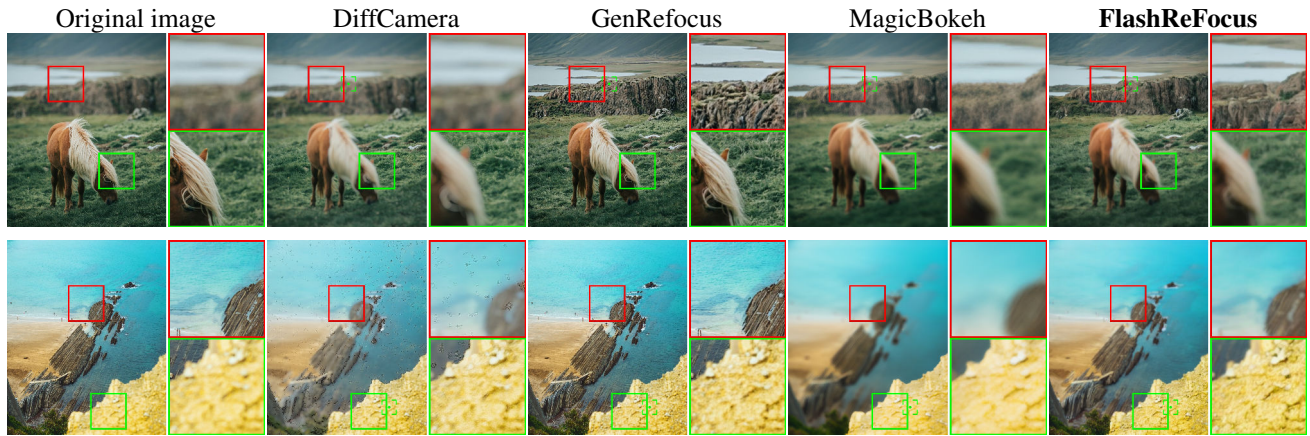


Figure H.3. **Refocusing on Unsplash images.** The focus moves to the green marker, onto a region that is out of focus in the input: the cliff in the first image and the foreground rock in the second. The red box enlarges the new focal region and the green box the former one; MagicBokeh does not restore the input, so its new focal region stays blurred.

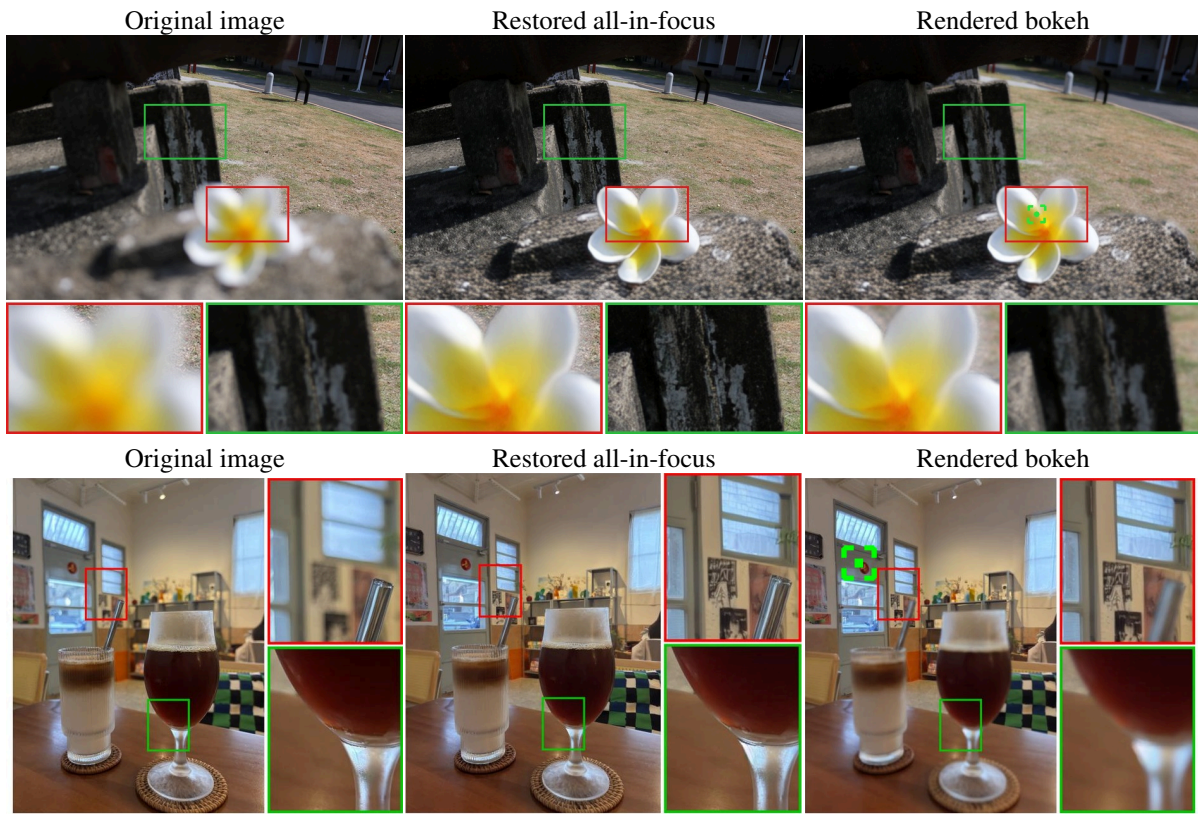


Figure H.4. **Refocusing real captures.** Canon EOS 750D (top) and iPhone 15 (bottom) photographs, each shown as input, the Stage I all-in-focus restoration, and the result rendered at the green focal point; the boxed regions are enlarged below or beside each image.

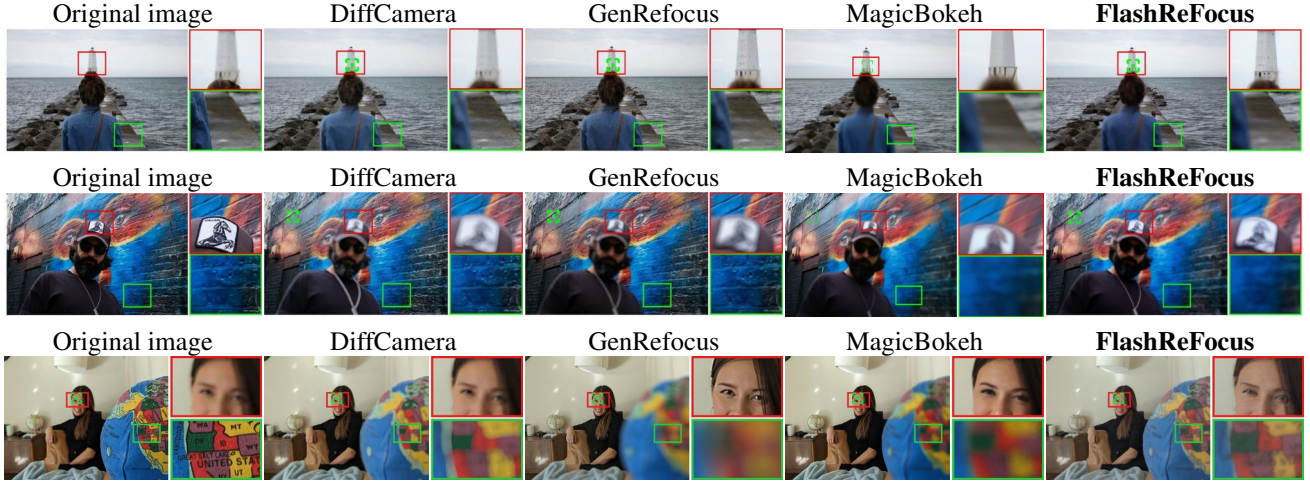


Figure H.5. **Refocusing to the background.** The focus moves from the foreground subject to the background of three Unsplash images (green marker); red box: new focal region, green box: former foreground, enlarged at right. Row 3 uses strength 0.8.

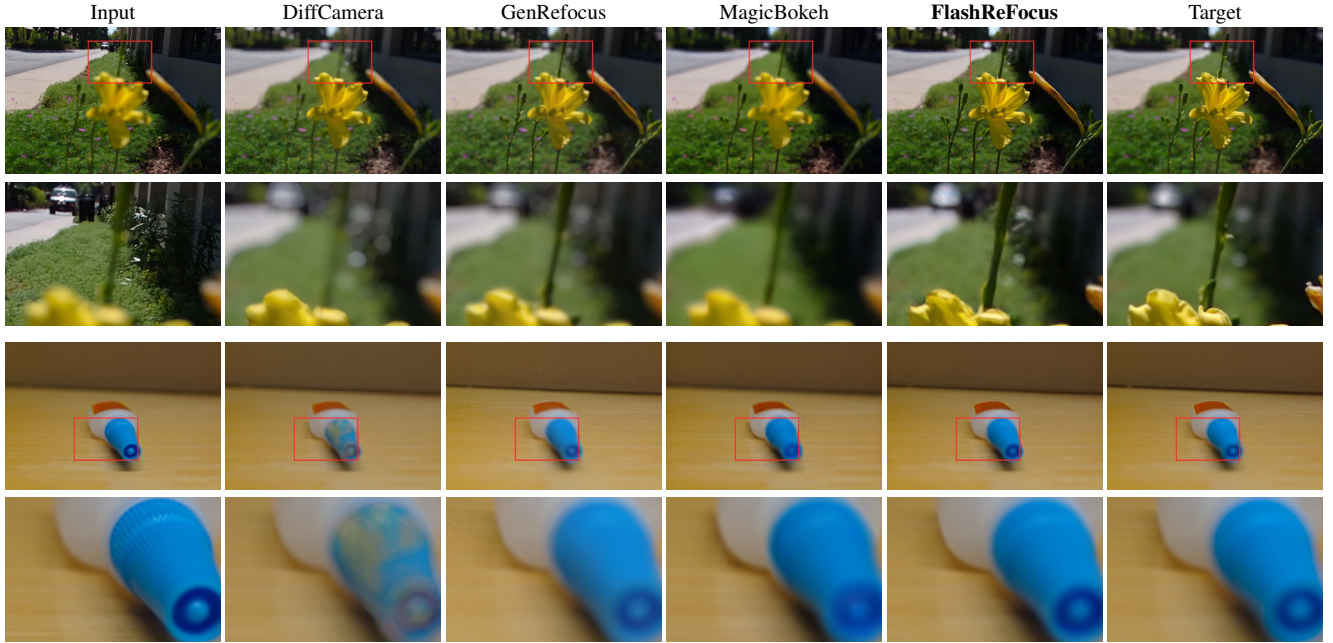


Figure H.6. **More complete refocusing on LFDOF**, as in Fig. H.1. The focus moves onto the flower stem (top) and the cap of the glue bottle (bottom); the red box is enlarged below each image. FlashReFocus renders the stem as sharp as the target, where the baselines leave it soft, and on the glue bottle DiffCamera adds a texture that is not in the target.

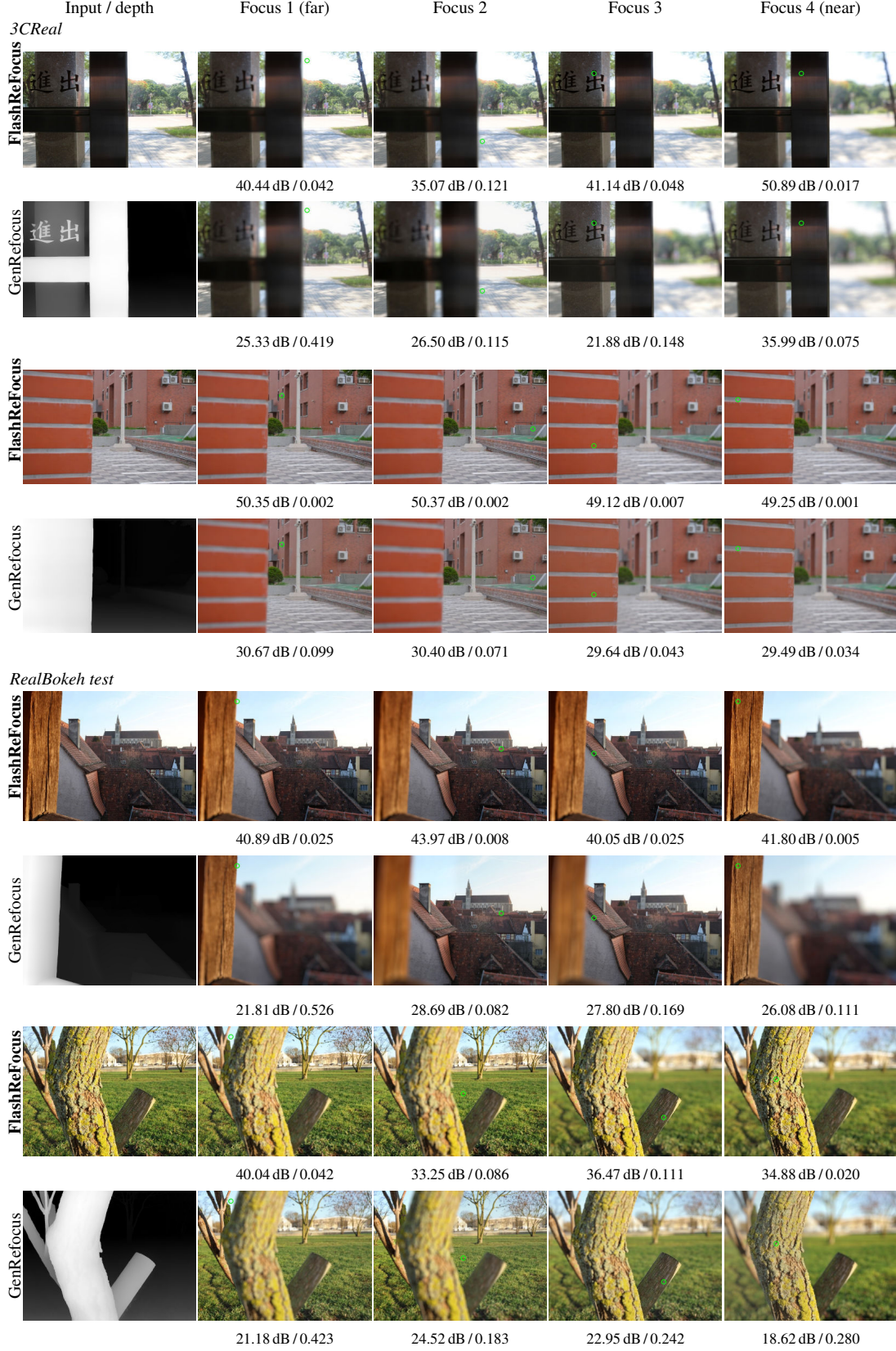


Figure H.7. **Moving the focal point on two 3CReal and two RealBokeh test scenes**, with the strength of FlashReFocus matched to GenRefocus ($K = 15$). The focal point (green marker) moves from far to near; for each scene the first column shows the all-in-focus input and the depth of our renderer (near bright). Both methods receive the same input and focal points. Below each result: PSNR and LPIPS of the in-focus region against the input, as in Table H.2.

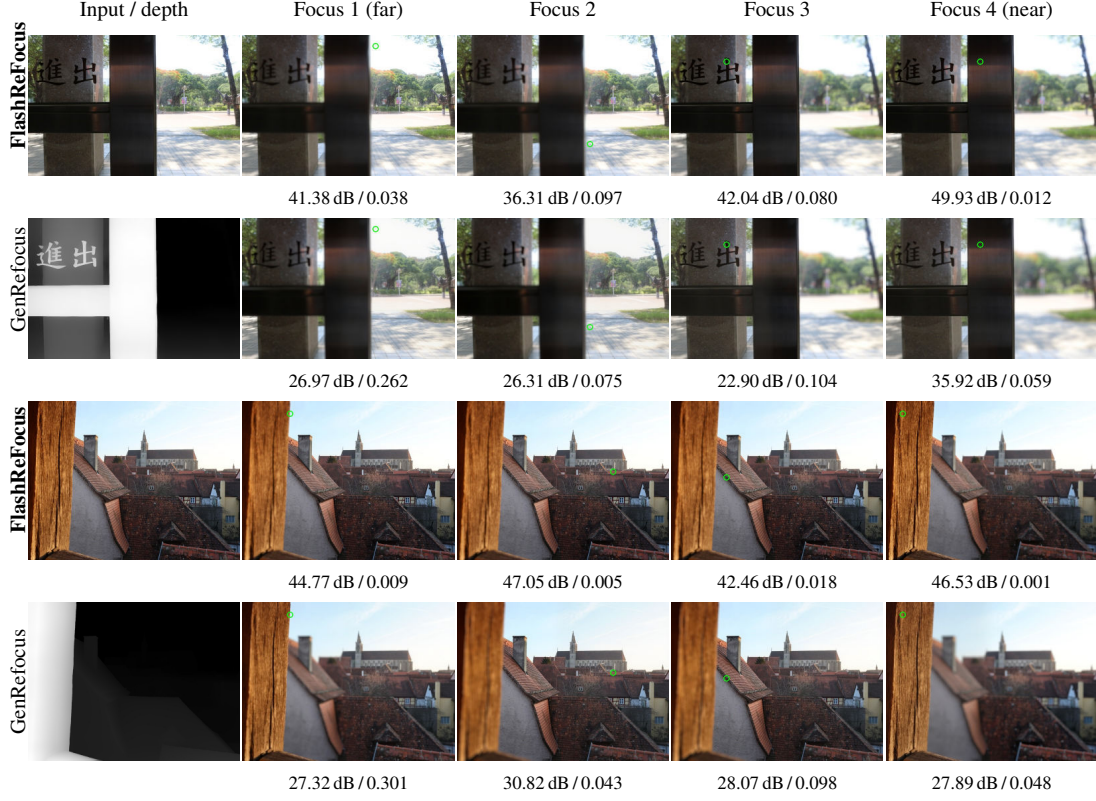


Figure H.8. **Moving the focal point with GenRefocus matched to FlashReFocus.** The first 3CReal (top) and RealBokeh test (bottom) scenes of Fig. H.7, with FlashReFocus at default strength 0.5 and GenRefocus at $K = 8.5$ and 5.25; layout and metrics as in Fig. H.7.

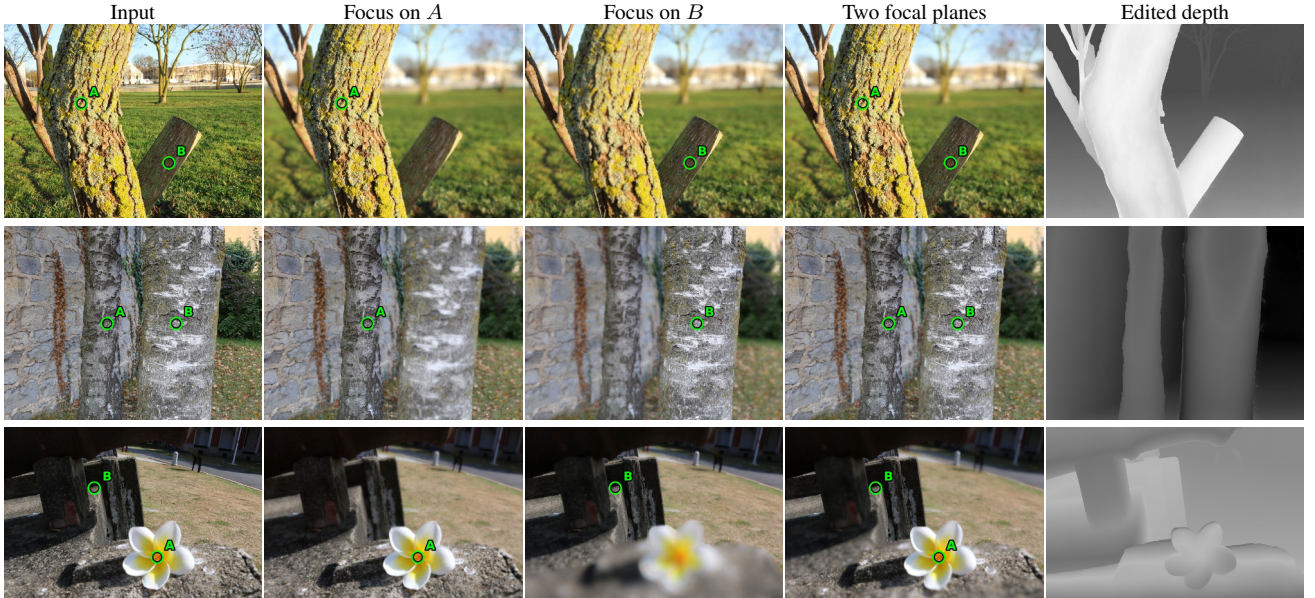


Figure H.9. **Two focal planes.** Focus points are marked A and B (row 1: trunk, stump; row 2: left, right birch; row 3: flower, bench). Focusing on A or on B blurs the other one; editing the depth to the distance from the nearer of the two planes keeps both sharp while the rest stays blurred. Rows 1–2: RealBokeh test; row 3: 3CReal. Strength 1.5.

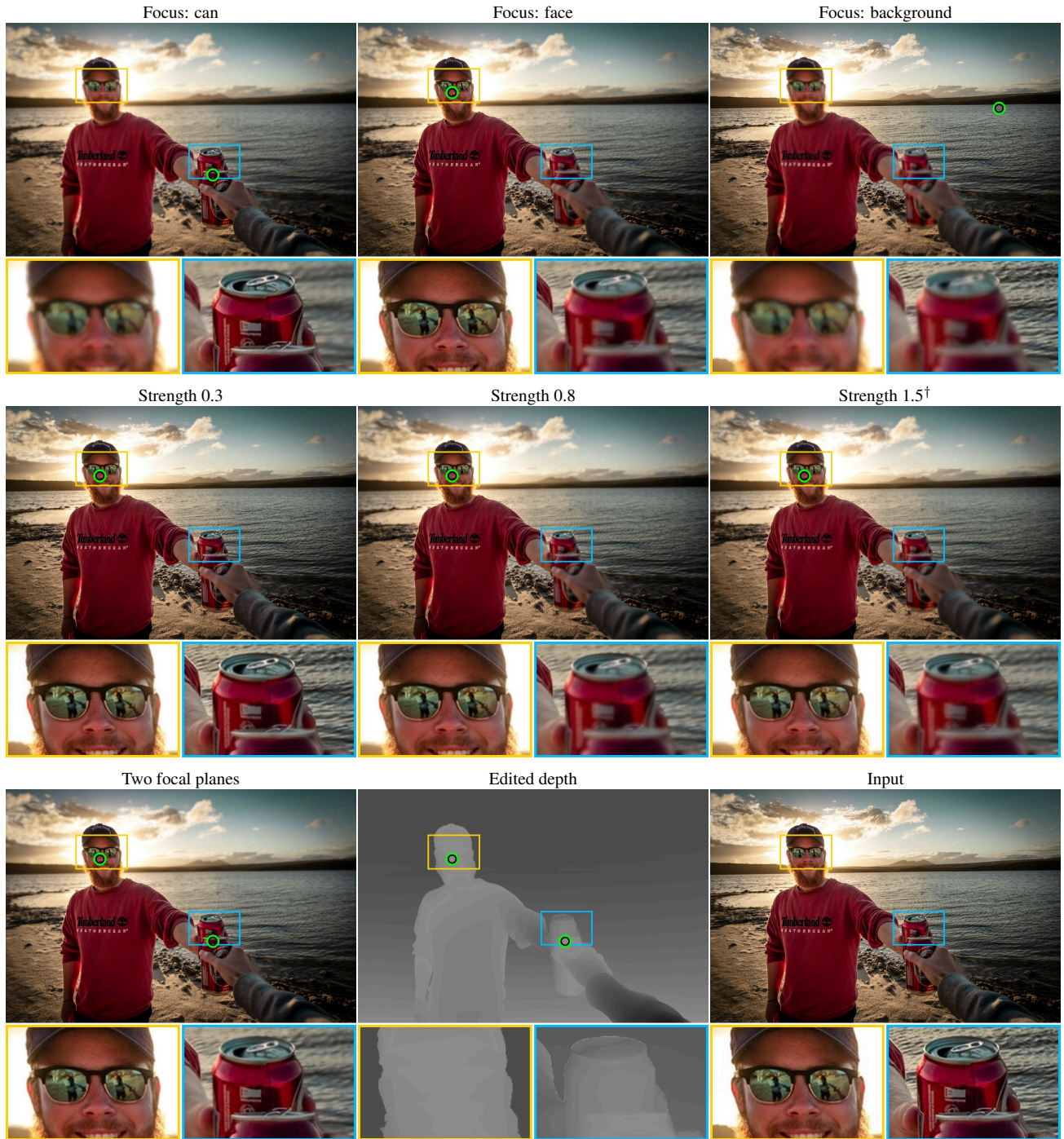


Figure H.10. **Controls on a portrait.** All results are rendered from one restored all-in-focus image. Top: the focal point (green marker) on the can, the face, and the background at strength 0.8. Middle: strength with the focus on the face. Bottom: two focal planes by depth editing (Appendix H), the edited depth, and the input. Boxes are enlarged below each result. [†]Beyond the trained range $[0, 1]$.

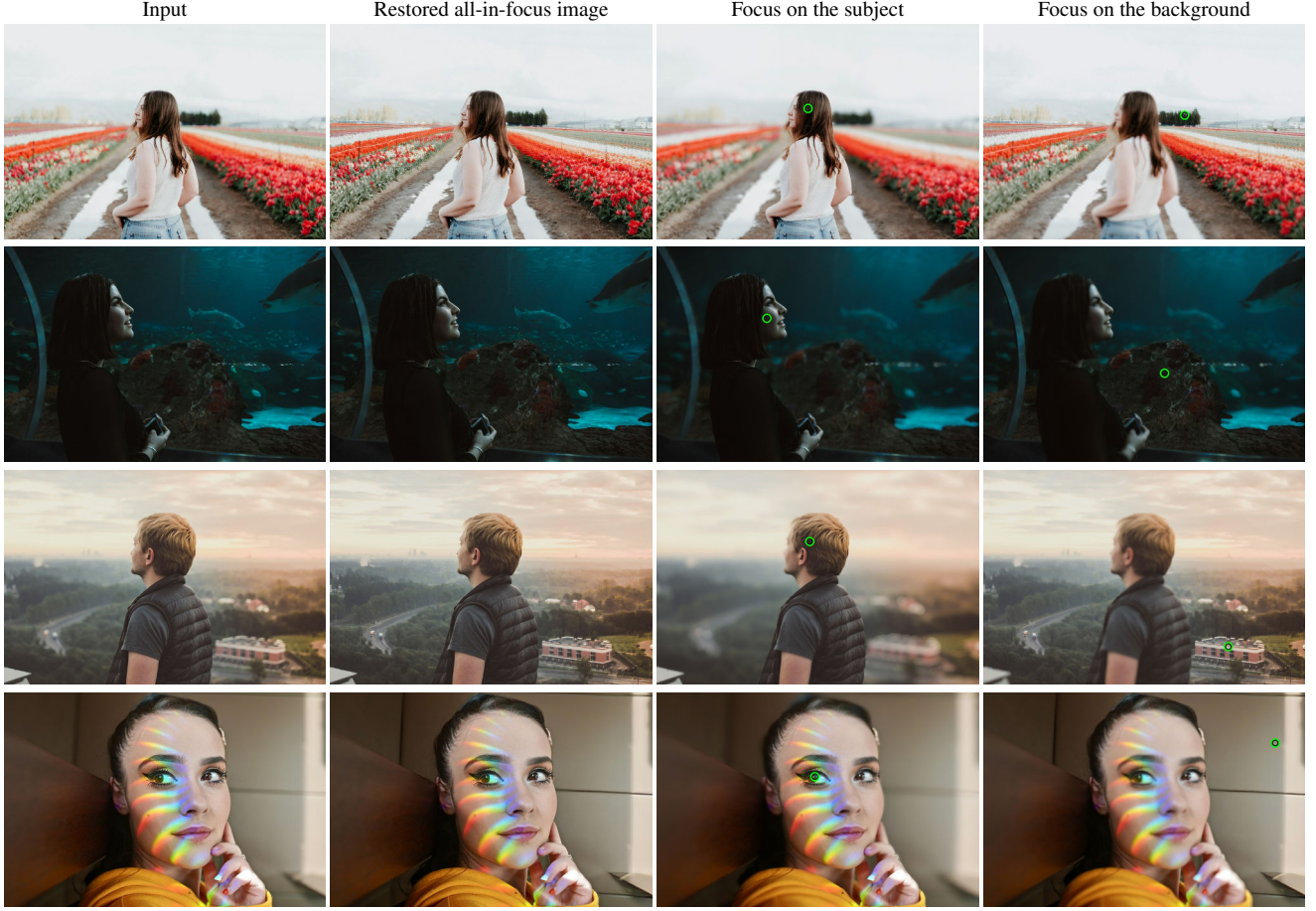


Figure H.11. **Refocusing portraits.** Four Unsplash portraits with a shallow depth of field. Stage I restores an all-in-focus image from each input, including its out-of-focus background, and the renderer then places the focus on the face or on the background (green marker) at strength $K = 0.8$. Focusing on the background brings back the trees, the aquarium rocks, the distant buildings, and the wall, while the subject is blurred according to its distance from the new focal plane; all edits of a portrait reuse one restored image.

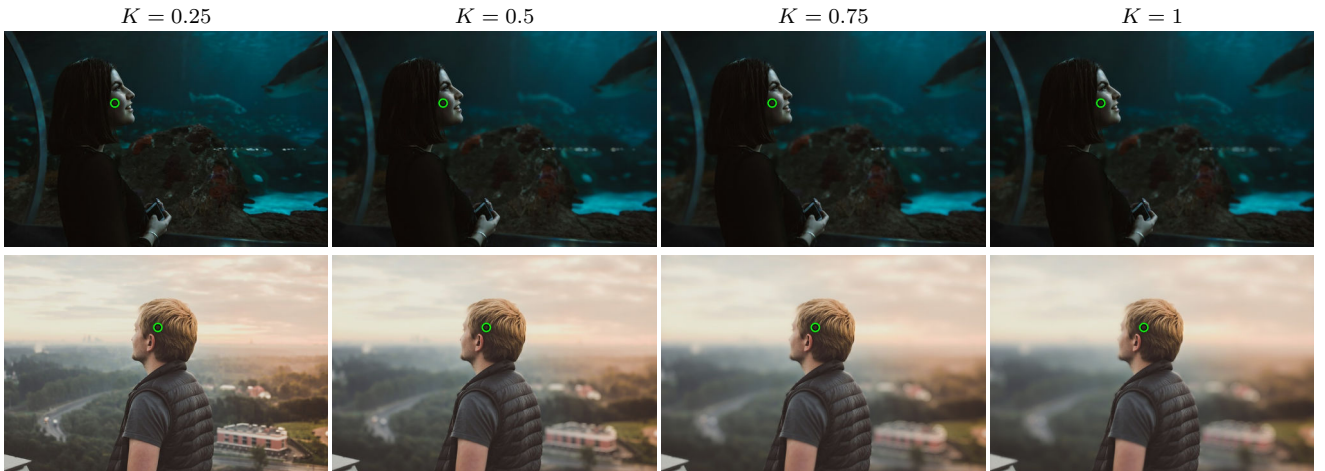


Figure H.12. **Strength on portraits.** The aquarium and city portraits of Fig. H.11 with the focus on the face (green marker) at strengths from $K = 0.25$ to $K = 1$. The face stays sharp at every strength while the background blur grows with K , as when the aperture of a lens is opened; each result is one forward pass of the renderer on the same restored image.

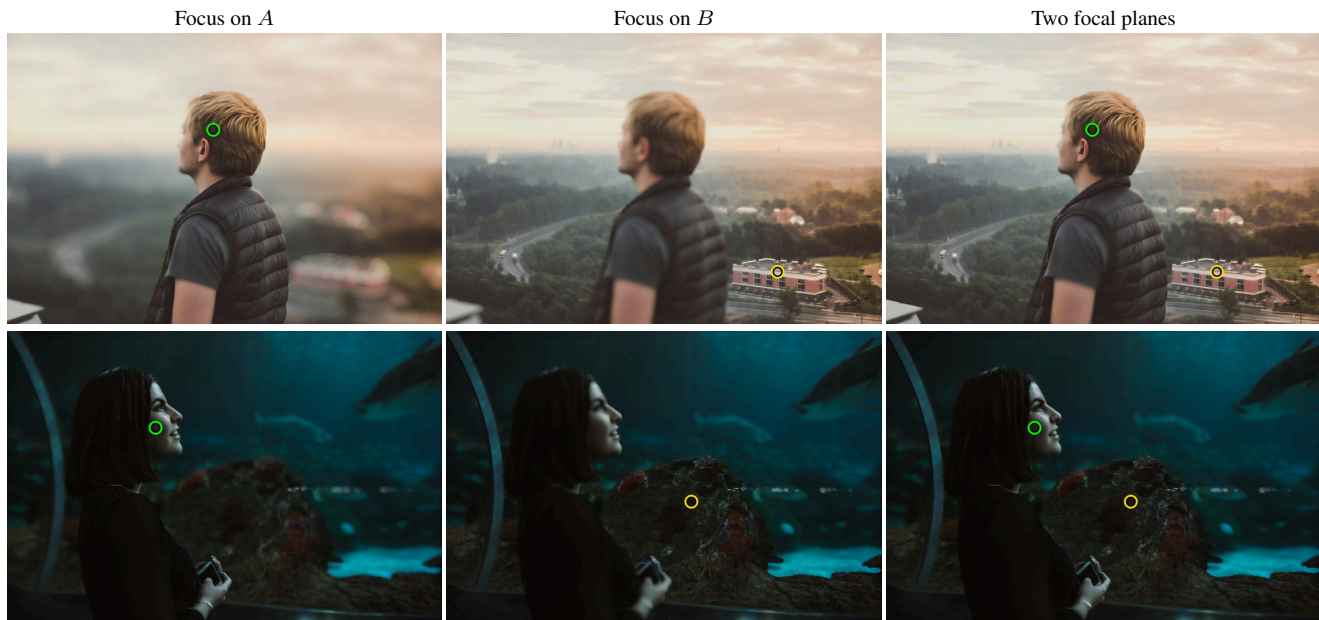


Figure H.13. **Two focal planes on portraits.** Focus on the face (A , green marker), on the background (B , yellow marker), and on both planes with the edited depth, at strength $K = 1$. With two planes, the face and the distant buildings or rocks are sharp at once while the space between them stays blurred, which a single lens cannot produce.



Figure H.14. **Refocusing within the foreground.** An Unsplash portrait whose flowers span a range of depths in front of and behind the subject. Moving the focal point from the near buds (green marker) to the flowers behind the subject (yellow marker), at $K = 0.8$, shifts the sharp region through the flower bed while the subject's dress and hands blur and sharpen with their distance to each plane.