BEYOND SHARP COVERAGE: FOCUS-AWARE CONTENT ROUTING FOR GENERALIZED MULTI-FOCUS IMAGE FUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-focus image fusion (MFIF) combines differently focused images into an all-in-focus result. Conventional methods rely on a sharp-coverage assumption, where each scene region is sufficiently observed in at least one input. In practice, however, limited focal-plane sampling and defocus spread can leave regions blurred in every input, while small viewpoint differences further complicate transferring observed content. We formulate these challenging cases as **Generalized MFIF**, where source selection alone is insufficient. We then introduce two benchmarks (CommonBlurGeometry and RealScene-68) and a region-wise evaluation protocol. The proposed **Focus-Aware Content Routing** framework uses continuous focus estimates to route each region according to the availability of reliable sharp observations: *it preserves what is reliably observed and restores only where observations are insufficient*. A focus-aware router determines the spatial contribution of the source observations and a restored candidate predicted by a one-step pixel-space restorer, while a contextual refiner corrects local inconsistencies in the composite. Our method achieves 32.64 dB on CommonBlurGeometry, exceeding the strongest external baseline by 3.86 dB, leads regional evaluation on RealScene-68, and reaches 42.93 dB on RealMFF. Inference takes 0.190 s at 625×433 and 2.869 s at 4104×2736 per image.

1 INTRODUCTION

Multi-focus image fusion (MFIF) combines differently focused images of the same scene into an all-in-focus result, overcoming the limited depth of field of a single capture (Zhang et al., 2020a; Liu et al., 2020). Methods based on focus detection, decision-map refinement, and learned feature fusion differ in implementation, but generally seek to identify and combine the sharper observations available in the inputs (Liu et al., 2017; Zhang et al., 2020b; Li et al., 2020; Ma et al., 2022a;b; Hu et al., 2023; Liu et al., 2024a).

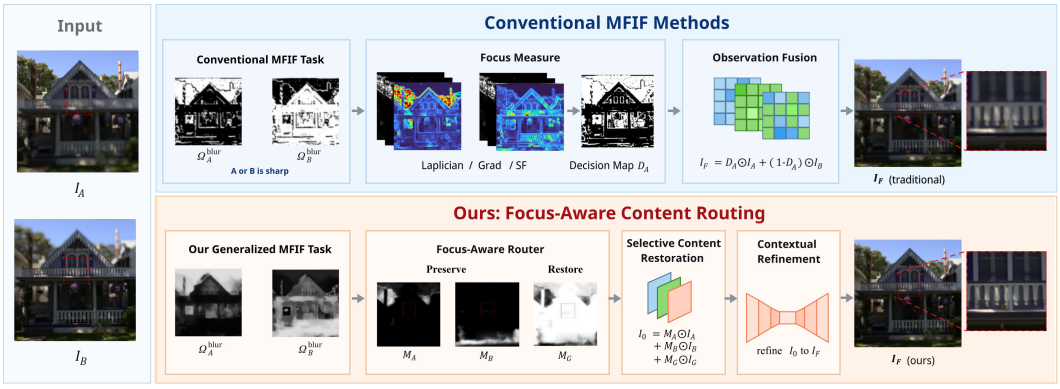

Figure 1: From complementary-focus selection to focus-aware routing. Conventional MFIF combines observed sharp regions but leaves jointly blurred content unresolved. Our method preserves reliable observations and incorporates a restored candidate where neither input is sharp.

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

Complementary focus coverage is not guaranteed in practice. A finite set of focal planes, defocus spread, and gradual transitions at depth boundaries may leave the same scene region blurred in both captures (Xu et al., 2020; Xie et al., 2025; Zhang et al., 2026). Small viewpoint changes also make corresponding source content harder to transfer, as shown in Figure 1. Even with a correct focus decision, selecting the less blurred input cannot provide detail that neither image recorded sharply.

We study this setting as **Generalized MFIF**, where the input images may exhibit incomplete sharp coverage and mild spatial displacement. Conventional MFIF with registered, complementary-focus images is a special case. In the generalized setting, however, the output must not only preserve available sharp observations, but also restore details that are not sharply recorded in either input.

This changes the central decision from which input to select to *whether a sharp observation is available*. When reliable sharp evidence exists, preserve the observed content directly; when it does not, source selection alone is insufficient, and the missing detail must be restored. This naturally reformulates Generalized MFIF as a **focus-aware content routing** problem: regions with reliable sharp evidence are routed to an observation-preserving fusion process, whereas regions lacking sufficient sharp evidence are routed to a restoration process.

Based on this formulation, we propose the **Focus-Aware Content Routing** framework, as shown in Figure 1. After aligning the input pair, a focus-aware router converts continuous focus estimates into spatial routing weights, which determine the contribution of the two source observations and a restored candidate at each location. According to the routing weights, selective content restoration directly preserves reliable RGB observations while selectively incorporating content from a one-step pixel-space restorer where observations are insufficient. Finally, contextual refinement corrects local inconsistencies and harmonizes the routed composite. The resulting framework follows a simple principle: *preserve what is reliably observed and restore only where observation is insufficient*.

This generalized setting also calls for evaluating preservation and restoration separately. We therefore establish a benchmark with complementary synthetic and real data: CommonBlurGeometry provides controllable defocus, small viewpoint changes, jointly blurred regions, and all-in-focus targets, while RealScene-68 comprises real focus pairs. A region-wise protocol separately evaluates source fidelity in reliably observed regions and restoration quality in common-blur regions.

In summary, our main contributions are as follows:

- We study **Generalized MFIF** with incomplete sharp coverage and mild spatial displacement, including registered complementary-focus pairs as a special case.
- We establish a benchmark comprising the controlled CommonBlurGeometry dataset, real-focus pairs in RealScene-68, and a region-wise protocol for assessing source fidelity and restoration quality separately.
- We introduce the **Focus-Aware Content Routing** framework, which uses a focus-aware router to allocate direct source preservation and an image recovered by a one-step pixel-space restorer, followed by the contextual refinement.
- Our method achieves the best results on the principal full-reference and region-wise measures of the proposed benchmark, while maintaining strong performance on conventional RealMFF.

2 Related Work

2.1 Conventional Multi-Focus Image Fusion

Most MFIF methods determine which input is focused at each location and combine the selected content. Decision-map approaches estimate focus regions through patch classification (Liu et al., 2017) or paired regression (Li et al., 2020). Feature-based methods instead extract representations from both inputs and fuse them before reconstruction, as in IFCNN (Zhang et al., 2020b). Later work jointly optimizes focus estimation and reconstruction (Ma et al., 2022a), uses multi-scale supervision (Zhao et al., 2019), or models interactions across image regions with attention (Ma et al., 2022b). Focus-relation modeling and invertible representations further address focus discrimination and detail preservation (Hu et al., 2023; Liu et al., 2024a; Zhai et al., 2025).

The focus boundary in a real capture is rarely a clean binary transition. MFFW documents defocus spread near such boundaries (Xu et al., 2020), while boundary-aware fusion and CSNet use boundary modeling, sharpness comparison, and similarity-based selection to improve decisions in transition regions (Ma et al., 2019; Zhang et al., 2026). These methods refine how available observations are selected or combined. When both observations are defocused, the missing sharp detail remains beyond the scope of source selection.

Spatial displacement introduces a separate challenge. DSAF-Net, MA-MFIF, and MVMFI establish correspondence with optical flow, homographies, or multi-view modeling (Chen et al., 2024; Zhao et al., 2024; Li et al., 2024b). We likewise align the pair before comparing focus responses and retain the valid sampling region. Our restoration branch addresses the remaining locations without a sharp observation.

2.2 Generative Multi-Focus Image Fusion

Generative MFIF can estimate detail that is not sharply recorded in the inputs. FusionDiff formulates fusion as diffusion-based denoising (Li et al., 2024a), and ReDiffuse introduces rotation-equivariant constraints to improve directional consistency (Li et al., 2026). GMFF explicitly considers incomplete focal-plane coverage and applies generative restoration to an initial deterministic fusion (Xie et al., 2025). These methods bring missing-detail recovery into a task previously dominated by source selection.

FusionDiff and ReDiffuse generate complete fused images, whereas GMFF restores the full output of a deterministic fusion stage (Li et al., 2024a; 2026; Xie et al., 2025). In all three cases, generative processing can affect regions already observed sharply. Our focus-aware routing instead uses the estimated availability of sharp observations to preserve source pixels directly and control where a restoration candidate contributes.

2.3 One-Step Image Restoration

The cost of iterative sampling motivates one-step conditional generation. Adversarial Diffusion Distillation reduces the sampling steps of diffusion models (Sauer et al., 2024); pix2pix-Turbo applies one-step latent diffusion to paired translation with VAE skip connections (Parmar et al., 2024). PixRestore instead models restoration on RGB image patches, combining conditional flow matching (Lipman et al., 2022) with hierarchical visual features (Sun et al., 2026). Pixel-space prediction is relevant to MFIF because fine observed structure must remain aligned with the input images.

We condition a one-step pixel-space restorer on the focus pair and predict a full-resolution candidate. The complete prediction provides context across the image, while focus-aware weights emphasize missing-detail regions during training and determine where the candidate contributes to fusion.

3 Generalized MFIF Benchmark

Generalized MFIF requires evaluating both preservation of available observations and recovery where no input is sharp. Our benchmark combines controlled data, real focus captures, and a region-wise protocol to assess these capabilities separately.

CommonBlurGeometry contains 4,000 training pairs and 200 validation pairs. Starting from an all-in-focus image, we estimate depth and synthesize a small viewpoint change with SHARP (Mescheder et al., 2026). We independently sample the focal plane and defocus strength for two captures and render their defocus with BokehMe (Peng et al., 2022). The sampled focus settings and depth information also provide continuous focus supervision. The resulting pairs include complementary-focus regions, gradual transitions, and regions blurred in both inputs (Figure 2).

RealScene-68 contains 68 real focus pairs with locally shared blur: 36 from ZEDD (Zuo et al., 2026) and 32 captured by us. Each pair was acquired at a fixed aperture with different focus distances. A small-aperture image, which has a larger depth of field, serves as an approximate reference for regional restoration measures.

Evaluation covers whole-image reconstruction, transfer of observed source information, and recovery in jointly blurred regions. A frozen focus estimator defines the same regional partition for every

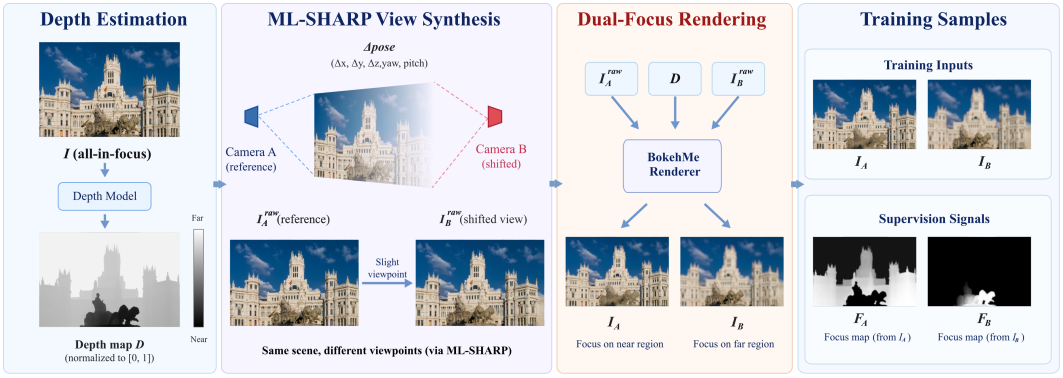

Figure 2: Construction of CommonBlurGeometry. Depth and a small viewpoint change are derived from an all-in-focus image, followed by rendering two focus settings and their continuous focus supervision.

method on a given input pair. The resulting measures assess source fidelity and restoration quality separately. Metric definitions and preprocessing are detailed in Appendices B and A.

4 METHOD

4.1 OVERVIEW

Given two differently focused images $I_A, I_B \in [-1, 1]^{3\times H\times W}$ of the same scene, we seek an all-in-focus output. The key decision at each location is whether a sharp observation can be used directly or a restored candidate is needed. An all-in-focus target I^* is available during training.

As illustrated in Figure 3, **Focus-Aware Content Routing** comprises three modules. The *Focus-Aware Router* aligns the inputs, estimates their local focus, and assigns weights to the source observations and the restoration candidate. *Selective Content Restoration* predicts that candidate and combines it with the observed pixels according to these weights. *Contextual Refinement* corrects local inconsistencies in the composite. The following subsections describe the modules in this order.

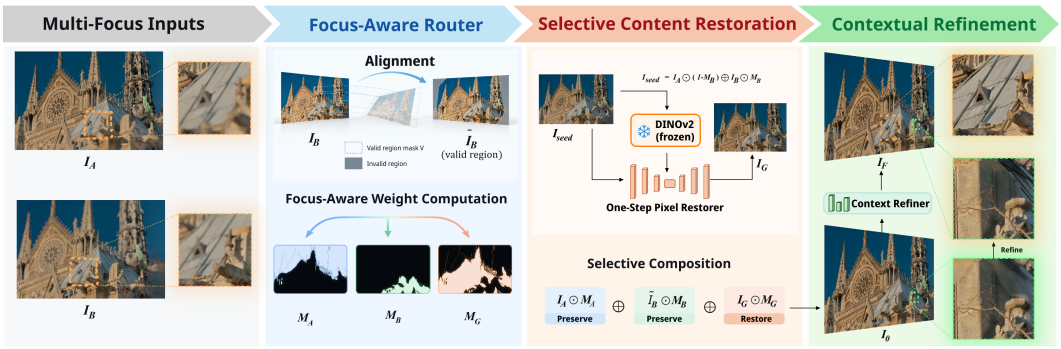

Figure 3: Focus-Aware Content Routing for Generalized MFIF. After aligning the input pair, continuous focus estimates produce weights for preserving observations from A and B and for using a restored candidate. A one-step pixel-space restorer predicts the candidate; weighted composition forms the initial fusion, which a contextual refiner turns into the final result.

4.2 FOCUS-AWARE ROUTER

The Focus-Aware Router assigns source and restoration weights using focus estimates at corresponding locations. For pairs with small spatial displacement, we take I_A as the output canvas and warp I_B into it. LoFTR matches (Sun et al., 2021) and a RANSAC-estimated homography (Fischler & Bolles, 1981) yield the aligned observation and its valid-sampling mask:

$$\widetilde{I}_B = \mathcal{W}(I_B, H_{B\to A}), \qquad V = \mathcal{W}(\mathbf{1}, H_{B\to A}), \tag{1}$$

where $V \in \{0,1\}^{H\times W}$ equals one where the warped B observation is valid. For an already registered pair, $\widetilde{I}_B = I_B$ and $V = \mathbf{1}$.

We then estimate the focus of both observations with a lightweight U-Net (Ronneberger et al., 2015). Its input X_F concatenates the aligned images, their absolute RGB difference, and their grayscale difference; the network returns two continuous focus maps:

$$X_F = [I_A; \widetilde{I}_B; |I_A - \widetilde{I}_B|; g(I_A) - g(\widetilde{I}_B)], \quad (F_A, F_B) = \mathcal{E}_F(X_F). \tag{2}$$

Here, g denotes grayscale conversion and $F_A, F_B \in [0,1]^{H\times W}$. A larger response indicates that the corresponding input is closer to the sharp state; two low responses indicate that a sharp observation is unavailable. The focus-estimation objective is given in Appendix C.

The focus maps determine how much each source contributes at every pixel. We first allocate weight to A according to F_A. Aligned B receives weight in the remaining area according to F_B and the valid-sampling mask. The residual is allocated to the restoration candidate:

$$\begin{cases} M_A = F_A, \\ M_B = (1 - M_A) \odot V \odot F_B, \\ M_G = 1 - M_A - M_B. \end{cases} \tag{3}$$

The weights sum to one at each pixel. A sharp input receives a high preservation weight, whereas M_G grows when both focus responses are low. Selective Content Restoration uses these weights to construct its conditioning image, train the restorer, and compose the candidate with the original observations.

4.3 Selective Content Restoration

Given the routing weights, Selective Content Restoration predicts a sharp candidate from the pair's scene context and incorporates it where restoration is needed. We form its dense conditioning image from A and aligned B, using the preservation weight of B:

$$I_{\text{seed}} = I_A \odot (1 - M_B) + \widetilde{I}_B \odot M_B. \tag{4}$$

Inspired by one-step pixel-space restoration (Sun et al., 2026), we initialize a conditional pixel transformer $\mathcal{G}_\theta$ from restoration pretraining. Its input concatenates Gaussian noise with the seed image, while a frozen DINOv2 encoder (Oquab et al., 2023) supplies multi-level conditioning features. One forward pass produces a full-resolution sharp candidate:

$$I_G = \mathcal{G}_\theta([I_{\text{seed}}; \epsilon], \Phi(I_{\text{seed}})), \qquad \epsilon \sim \mathcal{N}(0, \mathbf{I}), \tag{5}$$

where Φ denotes the frozen visual encoder. The candidate covers the full image to provide context for local recovery. Composition then combines the two source observations and the candidate with their respective weights:

$$I_0 = M_A \odot I_A + M_B \odot \widetilde{I}_B + M_G \odot I_G. \tag{6}$$

The restoration candidate must recover detail where both observations are blurred while remaining consistent with the surrounding scene. We therefore weight its squared error against the all-in-focus target by M_G, with a weaker whole-image term:

$$\mathcal{L}_G = \frac{\sum_{\mathbf{x}} M_G(\mathbf{x}) \|I_G(\mathbf{x}) - I^*(\mathbf{x})\|_2^2}{3\sum_{\mathbf{x}} M_G(\mathbf{x}) + \varepsilon} + \lambda_{\text{all}} \frac{\sum_{\mathbf{x}} \|I_G(\mathbf{x}) - I^*(\mathbf{x})\|_2^2}{3HW}. \tag{7}$$

The factor of three averages over RGB channels. The whole-image term also supervises the candidate outside high-demand regions. We additionally supervise the composed image with pixel and gradient losses against the all-in-focus target, so the preserved observations and restored content are optimized together. A patch-boundary gradient term discourages discontinuities between adjacent predictions; the complete objective is given in Appendix C. The resulting I_0 can still contain local errors where content from different sources meets.

4.4 Contextual Refinement

The initial fusion may contain differences in color, sharpness, or boundary continuity where its sources meet. Contextual Refinement receives the inputs, restored candidate, initial fusion, and allocation weights, giving its residual network both image context and source information. With $C_R = [I_A; \widetilde{I}_B; I_G; I_0; M_A; M_B; M_G]$, the final output is

$$I_F = \mathrm{clip}\left(I_0 + \rho \tanh\left(\frac{\mathcal{H}_\psi(C_R)}{\rho}\right), -1, 1\right). \tag{8}$$

The network $\mathcal{H}_\psi$ comprises a convolutional stem, six residual blocks, and an RGB head; ρ bounds the correction magnitude. Pixel and gradient losses on the final output train it to correct residual color, sharpness, and boundary errors. The objective and hyperparameters are specified in Appendix C.

Training and inference. The focus estimator, pixel-space restorer, and contextual refiner are trained sequentially. Inference performs alignment, focus estimation, and weight allocation, followed by one forward pass through the restorer and one through the refiner. Architecture settings and training details are provided in Appendix C.

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate Generalized MFIF on CommonBlurGeometry and RealScene-68 (Section 3), and use RealMFF (Zhang et al., 2020a) to assess performance under conventional complementary-focus conditions. Dataset splits and preprocessing are detailed in Appendix A.

For pairs with spatial displacement, we align the second input to the first before evaluation and provide the same aligned pair to every compared method. The target image is not used for this pairwise alignment.

Compared methods. We compare with DSIFT (Liu et al., 2015), IFCNN (Zhang et al., 2020b), SwinFusion (Ma et al., 2022b), ZMFF (Hu et al., 2023), FLUX2.0 (Labs, 2025), and ReDiffuse (Li et al., 2026).

5.2 Evaluation Metrics

We report PSNR, SSIM (Wang et al., 2004), MS-SSIM (Wang et al., 2003), and LPIPS (Zhang et al., 2018) where an all-in-focus target is available. We also report three source-referenced fusion measures: $Q_{AB/F}$ assesses edge transfer (Xydeas & Petrović, 2000), Q_{CB} assesses perceptually weighted contrast transfer (Chen & Blum, 2009), and Q_{VIFF} assesses visual information transferred from the inputs (Han et al., 2013). These measures characterize the use of observed source information; recovery of detail absent from both inputs requires a separate assessment.

Our regional protocol uses the weights M_A, M_B, M_G defined in Section 4. Each pixel is assigned to its largest weight, yielding source-supported regions Ω_A, Ω_B and a restoration region Ω_G. We erode the region boundaries to exclude mixed pixels. The frozen focus estimator computes this partition once per input pair, and every compared method is evaluated on the same pixels.

On datasets with an all-in-focus target, P_A, P_B, P_G denote PSNR within the respective regions. To measure whether restoration improves on source selection, we compare the fused result with the better input in the same restoration region:

$$\Delta P_G = P_G(I_F, I^*) - \max_{S \in \{I_A, \widetilde{I}_B\}} P_G(S, I^*). \tag{9}$$

The restoration success rate in Ω_G (G-RSR) measures how often local recovery succeeds, rather than only its mean gain. Among patches dominated by Ω_G, a patch counts as successful when its PSNR exceeds the better input by at least 0.2 dB and its LPIPS is lower than both inputs. On RealScene-68, source-region SSIM measures fidelity to the corresponding input, while the small-aperture image serves as an approximate reference for restoration-region measures. Patch selection and implementation details are given in Appendix B.

Table 1: Full-reference results. Best and second-best values are shown in bold and italics.

	CommonBlurGeometry				RealMFF			
Method	PSNR ↑	SSIM ↑	LPIPS ↓	MS-SSIM ↑	PSNR ↑	SSIM ↑	LPIPS ↓	MS-SSIM ↑
DSIFT	*28.7797*	*0.8646*	0.2021	0.9349	*38.1956*	*0.9796*	0.01975	*0.99583*
IFCNN	27.3598	0.8464	0.2101	0.9179	36.8310	0.9690	*0.00933*	0.99484
SwinFusion	27.4658	0.8378	0.2371	0.9132	36.4101	0.9696	0.01974	0.99301
ZMFF	26.8054	0.8424	0.2385	0.9163	33.3630	0.9565	0.03816	0.98986
FLUX2.0	28.1246	0.8556	**0.1284**	*0.9464*	29.6713	0.9006	0.07990	0.97477
ReDiffuse	24.8152	0.8226	0.2245	0.9183	29.6909	0.9443	0.02039	0.99018
Ours	**32.6422**	**0.9098**	*0.1391*	**0.9669**	**42.9280**	**0.99209**	**0.00643**	**0.99837**

Table 2: Region-wise evaluation with ground truth. P_A, P_B, and P_G are regional PSNR values; ΔP_G is the gain over the better input in Ω_G.

	CommonBlurGeometry					RealMFF				
Method	P_A	P_B	P_G	ΔP_G	G-RSR	P_A	P_B	P_G	ΔP_G	G-RSR
DSIFT	*42.1563*	*39.1527*	*27.6336*	*1.3074*	0.0662	*38.5648*	*39.5090*	*42.4340*	*0.3417*	0.1172
IFCNN	33.9222	33.5853	26.5337	0.2075	0.0429	37.7980	36.8640	40.6867	-1.4056	*0.2344*
SwinFusion	34.7592	31.6377	26.6044	0.2782	0.0350	38.2512	35.3746	39.5390	-2.5533	0.0156
ZMFF	36.5743	32.0369	25.7359	-0.5903	0.0981	35.2947	32.1151	36.2802	-5.8121	0.0078
FLUX2.0	31.2125	32.0836	27.4595	1.1333	*0.2895*	30.9156	28.9788	31.9741	-10.1182	0.0000
ReDiffuse	28.0176	28.1607	24.2921	-2.0341	0.0293	29.8948	29.4164	29.9541	-12.1382	0.0234
Ours	**43.8366**	**40.0658**	**31.5224**	**5.1962**	**0.6675**	**43.3758**	**43.2414**	**45.7353**	**3.6430**	**0.5703**

5.3 Evaluation of Generalized and Conventional MFIF

Full-reference results. Table 1 reports full-reference results on CommonBlurGeometry and conventional RealMFF. On CommonBlurGeometry, our method ranks first in PSNR, SSIM, and MS-SSIM and second in LPIPS; on RealMFF, it ranks first in all four measures.

Our method exceeds DSIFT, the strongest external baseline in PSNR, by 3.8625 dB on CommonBlurGeometry and 4.7324 dB on RealMFF. On RealMFF, it also reduces LPIPS from IFCNN's 0.00933 to 0.00643.

Source-referenced results. The full source-referenced results for all three datasets are reported in Appendix E (Table 5). DSIFT and ReDiffuse score higher than our method on several measures. The example below and the region-wise results examine whether these rankings reflect recovery where both inputs are blurred.

In Figure 4(a), DSIFT has higher $Q_{AB/F}$ and Q_{CB} but retains blur shared by the inputs. Our result reaches 30.3980 dB global PSNR, gains 6.4106 dB over the better input in the restoration region, and achieves 68.94% G-RSR. In the ZEDD example of Figure 4(b), the lamp remains diffuse in the ReDiffuse result, while its edges are clearer in ours.

Region-wise results. Table 2 separates reconstruction quality in source-supported regions from that in the jointly blurred region. In Ω_G, our method exceeds the better input by 5.1962 dB on CommonBlurGeometry and 3.6430 dB on RealMFF; its corresponding G-RSR values are 66.75% and 57.03%. The source-region scores show that this recovery is accompanied by high fidelity where observations are available.

On RealScene-68 (Table 3), our method attains source-region SSIM values of 0.9807 and 0.9329. Against the approximate small-aperture reference, it also has the highest restoration-region PSNR and G-RSR, the lowest LPIPS, and a positive mean gain in Ω_G. Figure 4(b) shows a corresponding ZEDD example alongside ReDiffuse.

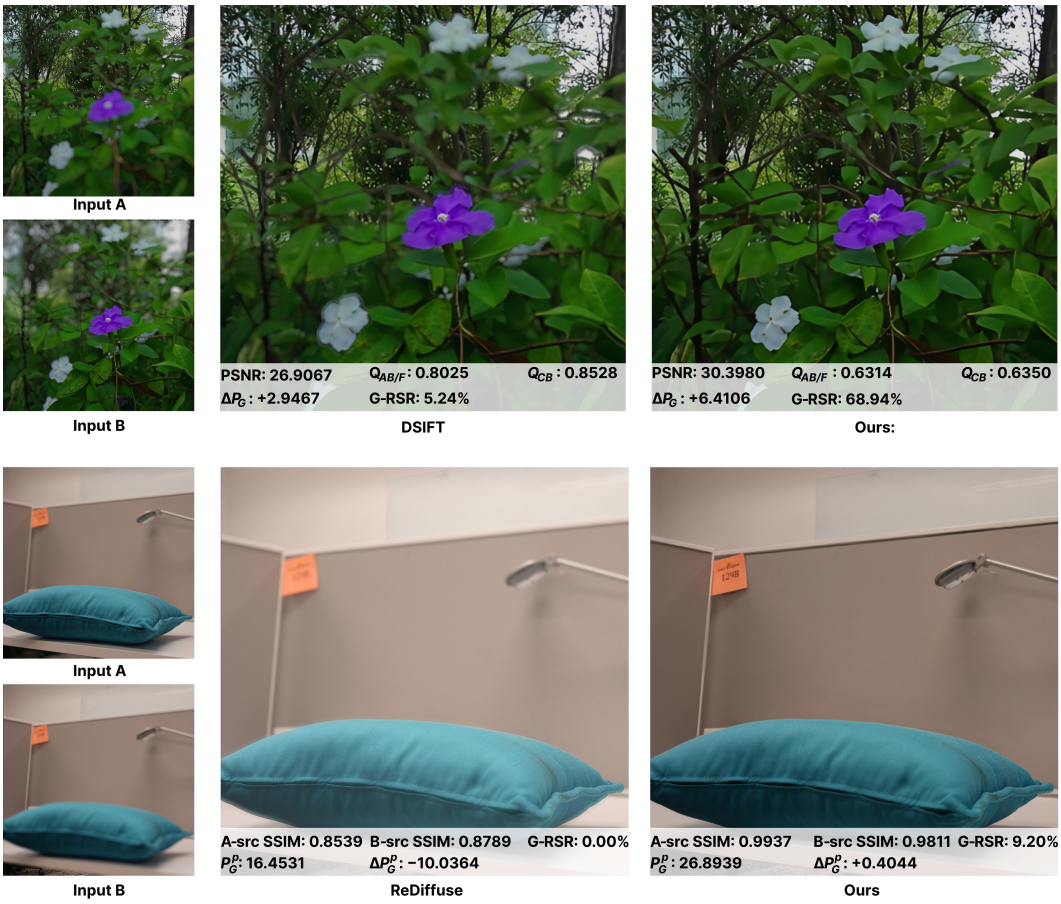

Figure 4: Enlarged comparisons with per-image metrics. (a) DSIFT and Ours on CommonBlurGeometry. (b) ReDiffuse and Ours on a ZEDD scene from RealScene-68. The metric values are included in the figure.

Table 3: Region-wise evaluation on RealScene-68. A/B-src SSIM measures source fidelity; superscript p denotes a measure using the small-aperture image as an approximate reference.

Method	A-src SSIM	B-src SSIM	P_G^p	LPIPS^p ↓	ΔP_G^p	G-RSR^p
DSIFT	*0.9548*	0.8655	27.2231	0.2988	-0.3780	0.0096
IFCNN	0.9331	*0.8840*	26.9810	0.3056	-0.6201	0.0232
SwinFusion	0.9445	0.8818	*27.4242*	*0.2950*	*-0.1769*	0.0693
ZMFF	0.8172	0.8342	27.0946	0.3614	-0.5065	*0.1078*
FLUX2.0	0.7641	0.7765	24.5318	0.3354	-3.0693	0.0713
ReDiffuse	0.8761	0.8326	20.6510	0.3360	-6.9501	0.0160
Ours	**0.9807**	**0.9329**	**28.2910**	**0.2906**	**0.6899**	**0.2147**

Qualitative comparison. Figure 5 compares representative RealScene-68 and CommonBlurGeometry examples. Conventional fusion leaves residual blur where both inputs are defocused. Some generative results sharpen these regions but alter local structure or appearance. Our method recovers contours in jointly blurred areas while retaining the appearance of already sharp regions.

5.4 Ablation Studies

We evaluate source preservation, content restoration, and contextual refinement with three variants. **w/o Preservation** bypasses source-preserving composition and uses full-image restoration; **w/o Restoration** removes the restoration candidate and renormalizes the two source weights; **w/o Refiner** outputs I_0. Relative to these three variants, the complete model improves PSNR by 0.5246, 3.9607, and 0.1326 dB on CommonBlurGeometry, and by 7.3939, 5.4054, and 4.4506 dB on RealMFF, respectively (Table 4).

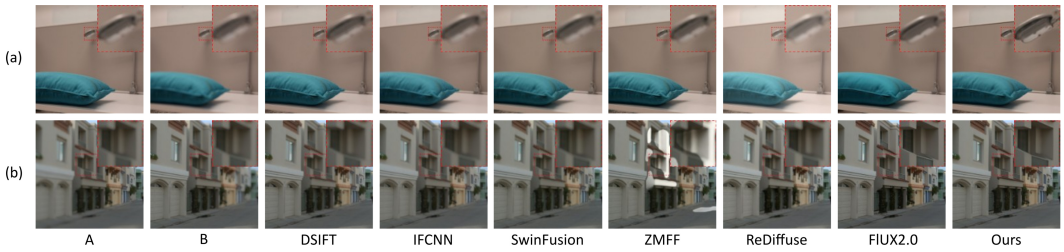

Figure 5: Qualitative comparison on RealScene-68 (a) and CommonBlurGeometry (b). Red dashed boxes indicate magnified regions.

Table 4: Ablation of source preservation, restoration, and refinement. Best and second-best values are shown in bold and italics.

	CommonBlurGeometry				RealMFF			
Variant	PSNR ↑	SSIM ↑	LPIPS ↓	MS-SSIM ↑	PSNR ↑	SSIM ↑	LPIPS ↓	MS-SSIM ↑
w/o Preservation	32.1176	0.8969	0.1643	0.9624	35.5341	0.95039	0.03165	0.99128
w/o Restoration	28.6815	0.8632	0.2166	0.9293	37.5226	0.97910	0.02359	0.99531
w/o Refiner	*32.5096*	*0.9070*	*0.1469*	*0.9659*	*38.4774*	*0.98006*	*0.01981*	*0.99605*
Ours	**32.6422**	**0.9098**	**0.1391**	**0.9669**	**42.9280**	**0.99209**	**0.00643**	**0.99837**

The size of each ablation effect depends on the dataset. Restoration contributes most on CommonBlurGeometry, where both inputs can lack a sharp observation. Preservation and refinement matter more on RealMFF, where sharp source content is more frequently available. The complete model outperforms all three ablations on both datasets.

The regional ablations sharpen this comparison. On CommonBlurGeometry, removing restoration lowers the restoration-region gain from 5.1962 to 1.0423 dB and G-RSR from 66.75% to 3.58%. On RealMFF, removing the refiner lowers the same measures from 3.6430 to -0.4365 dB and from 57.03% to 7.81%, respectively.

Figure 7 in the appendix complements the numerical ablations. Removing preservation changes fine texture already present in the inputs. Without restoration, jointly blurred regions remain soft; without refinement, local discontinuities remain near composition boundaries.

5.5 Discussion

The three evaluations answer different questions. Full-reference scores measure agreement with an all-in-focus target. Source-referenced metrics assess how much edge, contrast, or visual information from the inputs appears in the output. Regional scores distinguish fidelity to available sharp observations from recovery where neither input is sharp. The source-referenced rankings in Appendix Table 5 differ from the full-reference and regional results in Tables 1–3; Figure 4(a) shows why a high source-transfer score alone does not establish recovery of jointly blurred detail.

RealScene-68 uses the same predicted regional partition for every method and a small-aperture image as an approximate sharp reference. Focus-estimation, exposure, and alignment errors can affect its scores, so we interpret them alongside the visual comparisons.

Our per-image inference times are 0.762 s on CommonBlurGeometry, 0.190 s on RealMFF, and 2.869 s on RealScene-68; Appendix F reports the full efficiency comparison.

6 Conclusion

We studied Generalized MFIF under shared blur and mild spatial displacement with CommonBlurGeometry, RealScene-68, and regional evaluation. Focus-Aware Content Routing preserves reliable observations and incorporates restored content where sharp evidence is missing. It leads the principal full-reference and regional evaluations while performing strongly on conventional RealMFF. Focus-estimation errors and larger viewpoint changes remain directions for further work.

AI USE STATEMENT

We used generative AI tools to assist with Chinese-to-English translation and manuscript language preparation, literature discovery and organization, feedback on research framing and experimental presentation, and limited code edits. The authors led the method design, data preparation, and experiments. They reviewed and verified the final text, citations, code, and reported results, and take responsibility for the full content of this paper.

REFERENCES

Peiming Chen, Jiaqin Jiang, Li Li, and Jian Yao. A defocus and similarity attention-based cascaded network for multi-focus and misaligned image fusion. *Information Fusion*, 103:102125, March 2024. ISSN 1566-2535. doi: 10.1016/j.inffus.2023.102125. URL http://dx.doi.org/10.1016/j.inffus.2023.102125.

Yin Chen and Rick S. Blum. A new automated quality assessment algorithm for image fusion. *Image and Vision Computing*, 27(10):1421–1432, September 2009. ISSN 0262-8856. doi: 10.1016/j.imavis.2007.12.002. URL http://dx.doi.org/10.1016/j.imavis.2007.12.002.

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, June 1981. ISSN 1557-7317. doi: 10.1145/358669.358692. URL http://dx.doi.org/10.1145/358669.358692.

Yu Han, Yunze Cai, Yin Cao, and Xiaoming Xu. A new image fusion performance metric based on visual information fidelity. *Information Fusion*, 14(2):127–135, April 2013. ISSN 1566-2535. doi: 10.1016/j.inffus.2011.08.002. URL http://dx.doi.org/10.1016/j.inffus.2011.08.002.

Xingyu Hu, Junjun Jiang, Xianming Liu, and Jiayi Ma. ZMFF: Zero-shot multi-focus image fusion. *Information Fusion*, 92:127–138, 2023. doi: 10.1016/j.inffus.2022.11.014.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

Bo Li, Tingting Bao, Lingling Zhang, Weiping Fu, Yaxian Wang, and Jun Liu. Rediffuse: Rotation equivariant diffusion model for multi-focus image fusion, 2026. URL https://arxiv.org/abs/2603.21129.

Jinxing Li, Xiaobao Guo, Guangming Lu, Bob Zhang, Yong Xu, Feng Wu, and David Zhang. Drpl: Deep regression pair learning for multi-focus image fusion. *IEEE Transactions on Image Processing*, 29:4816–4831, 2020. ISSN 1941-0042. doi: 10.1109/tip.2020.2976190. URL http://dx.doi.org/10.1109/TIP.2020.2976190.

Mining Li, Ronghao Pei, Tianyou Zheng, Yang Zhang, and Weiwei Fu. Fusiondiff: Multi-focus image fusion using denoising diffusion probabilistic models. *Expert Systems with Applications*, 238:121664, March 2024a. ISSN 0957-4174. doi: 10.1016/j.eswa.2023.121664. URL http://dx.doi.org/10.1016/j.eswa.2023.121664.

Zhilong Li, Kejun Wu, Junhao Liu, Qiong Liu, and You Yang. Multi-view multi-focus image fusion: A novel benchmark dataset and method. In *2024 IEEE International Conference on Image Processing (ICIP)*, pp. 200–206. IEEE, October 2024b. doi: 10.1109/icip51287.2024.10647581. URL http://dx.doi.org/10.1109/ICIP51287.2024.10647581.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2022. URL https://arxiv.org/abs/2210.02747.

Jinyang Liu, Shutao Li, Renwei Dian, and Ze Song. Focus relationship perception for unsupervised multi-focus image fusion. *IEEE Transactions on Multimedia*, 26:6155–6165, 2024a. ISSN 1941-0077. doi: 10.1109/tmm.2023.3347099. URL http://dx.doi.org/10.1109/TMM.2023.3347099.

Yu Liu, Shuping Liu, and Zengfu Wang. Multi-focus image fusion with dense sift. *Information Fusion*, 23:139–155, May 2015. ISSN 1566-2535. doi: 10.1016/j.inffus.2014.05.004. URL http://dx.doi.org/10.1016/j.inffus.2014.05.004.

Yu Liu, Xun Chen, Hu Peng, and Zengfu Wang. Multi-focus image fusion with a deep convolutional neural network. *Information Fusion*, 36:191–207, July 2017. ISSN 1566-2535. doi: 10.1016/j.inffus.2016.12.001. URL http://dx.doi.org/10.1016/j.inffus.2016.12.001.

Yu Liu, Lei Wang, Juan Cheng, Chang Li, and Xun Chen. Multi-focus image fusion: A survey of the state of the art. *Information Fusion*, 64:71–91, December 2020. ISSN 1566-2535. doi: 10.1016/j.inffus.2020.06.013. URL http://dx.doi.org/10.1016/j.inffus.2020.06.013.

Yu Liu, Zhengzheng Qi, Juan Cheng, and Xun Chen. Rethinking the effectiveness of objective evaluation metrics in multi-focus image fusion: A statistic-based approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5806–5819, August 2024b. ISSN 1939-3539. doi: 10.1109/tpami.2024.3367905. URL http://dx.doi.org/10.1109/TPAMI.2024.3367905.

Boyuan Ma, Xiang Yin, Di Wu, Haokai Shen, Xiaojuan Ban, and Yu Wang. End-to-end learning for simultaneously generating decision map and multi-focus image fusion result. *Neurocomputing*, 470:204–216, January 2022a. ISSN 0925-2312. doi: 10.1016/j.neucom.2021.10.115. URL http://dx.doi.org/10.1016/j.neucom.2021.10.115.

Haoyu Ma, Juncheng Zhang, Shaojun Liu, and Qingmin Liao. Boundary aware multi-focus image fusion using deep neural network. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1150–1155. IEEE, July 2019. doi: 10.1109/icme.2019.00201. URL http://dx.doi.org/10.1109/ICME.2019.00201.

Jiayi Ma, Linfeng Tang, Fan Fan, Jun Huang, Xiaoguang Mei, and Yong Ma. Swinfusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA Journal of Automatica Sinica*, 9(7):1200–1217, 2022b. doi: 10.1109/JAS.2022.105686.

Lars Mescheder, Wei Dong, Shiwei Li, Xuyang Bai, Marcel Santos, Peiyun Hu, Bruno Lecouat, Mingmin Zhen, Amaël Delaunoy, Tian Fang, Yanghai Tsin, Stephan R. Richter, and Vladlen Koltun. Sharp monocular view synthesis in less than a second. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/297fe652867e4897e9f1fe1cd715de19-Abstract-Conference.html.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. URL https://arxiv.org/abs/2304.07193.

Gaurav Parmar, Taesung Park, Srinivasa Narasimhan, and Jun-Yan Zhu. One-step image translation with text-to-image models, 2024. URL https://arxiv.org/abs/2403.12036.

Juewen Peng, Zhiguo Cao, Xianrui Luo, Hao Lu, Ke Xian, and Jianming Zhang. Bokehme: When neural rendering meets classical rendering. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16262–16271. IEEE, June 2022. doi: 10.1109/cvpr52688.2022.01580. URL http://dx.doi.org/10.1109/CVPR52688.2022.01580.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*, pp. 234–241. Springer International Publishing, 2015. ISBN 9783319245744. doi: 10.1007/978-3-319-24574-4_28. URL http://dx.doi.org/10.1007/978-3-319-24574-4_28.

Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1-4):259–268, November 1992. ISSN 0167-2789. doi: 10.1016/0167-2789(92)90242-f. URL http://dx.doi.org/10.1016/0167-2789(92)90242-F.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. *Adversarial Diffusion Distillation*, pp. 87–103. Springer Nature Switzerland, October 2024. ISBN 9783031730160. doi: 10.1007/978-3-031-73016-0_6. URL http://dx.doi.org/10.1007/978-3-031-73016-0_6.

Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8918–8927. IEEE, June 2021. doi: 10.1109/cvpr46437.2021.00881. URL http://dx.doi.org/10.1109/CVPR46437.2021.00881.

Lingchen Sun, Rongyuan Wu, Xiangtao Kong, Jixin Zhao, Qiaosi Yi, Yujing Sun, Shuaizheng Liu, Zhengqiang Zhang, and Lei Zhang. Pixrestore: Unified image restoration via pixel diffusion transformer, 2026. URL https://arxiv.org/abs/2608.16793.

Z. Wang, E.P. Simoncelli, and A.C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, ACSSC-03, pp. 1398–1402. IEEE, 2003. doi: 10.1109/acssc.2003.1292216. URL http://dx.doi.org/10.1109/ACSSC.2003.1292216.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, April 2004. ISSN 1941-0042. doi: 10.1109/tip.2003.819861. URL http://dx.doi.org/10.1109/TIP.2003.819861.

Xinzhe Xie, Buyu Guo, Bolin Li, Shuangyan He, Yanzhen Gu, Qingyan Jiang, and Peiliang Li. Generative multi-focus image fusion, 2025. URL https://arxiv.org/abs/2512.21495.

Shuang Xu, Xiaoli Wei, Chunxia Zhang, Junmin Liu, and Jiangshe Zhang. Mffw: A new dataset for multi-focus image fusion, 2020. URL https://arxiv.org/abs/2002.04780.

C.S. Xydeas and V. Petrović. Objective image fusion performance measure. *Electronics Letters*, 36(4):308–309, February 2000. ISSN 1350-911X. doi: 10.1049/el:20000267. URL http://dx.doi.org/10.1049/el:20000267.

Hao Zhai, Peng Chen, Nannan Luo, Qinyu Li, and Ping Yu. Cpfusion: A multi-focus image fusion method based on closed-loop regularization. *Image and Vision Computing*, 154:105399, February 2025. ISSN 0262-8856. doi: 10.1016/j.imavis.2024.105399. URL http://dx.doi.org/10.1016/j.imavis.2024.105399.

Juncheng Zhang, Qingmin Liao, Shaojun Liu, Haoyu Ma, Wenming Yang, and Jing-Hao Xue. Realmff: A large realistic multi-focus image dataset with ground truth. *Pattern Recognition Letters*, 138:370–377, October 2020a. ISSN 0167-8655. doi: 10.1016/j.patrec.2020.08.002. URL http://dx.doi.org/10.1016/j.patrec.2020.08.002.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595. IEEE, June 2018. doi: 10.1109/cvpr.2018.00068. URL http://dx.doi.org/10.1109/CVPR.2018.00068.

Yicheng Zhang, Haoyou Deng, Zhiqiang Li, Wenti Yin, Nong Sang, and Changxin Gao. Clarity contrast and similarity selection for multi-focus image fusion, 2026. URL https://arxiv.org/abs/2608.03252.

Yu Zhang, Yu Liu, Peng Sun, Han Yan, Xiaolin Zhao, and Li Zhang. Ifcnn: A general image fusion framework based on convolutional neural network. *Information Fusion*, 54:99–118, February 2020b. ISSN 1566-2535. doi: 10.1016/j.inffus.2019.07.011. URL http://dx.doi.org/10.1016/j.inffus.2019.07.011.

Baojun Zhao, Fei Luo, Joel Fuentes, Weichao Ding, and Chunhua Gu. Ma-mfif: When misaligned multi-focus image fusion meets deep homography estimation. *Multimedia Tools and Applications*, 84(12):10877–10898, May 2024. ISSN 1573-7721. doi: 10.1007/s11042-024-19385-4. URL http://dx.doi.org/10.1007/s11042-024-19385-4.

Wenda Zhao, Dong Wang, and Huchuan Lu. Multi-focus image fusion with a natural enhancement via a joint multi-level deeply supervised convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(4):1102–1115, April 2019. ISSN 1558-2205. doi: 10.1109/tcsvt.2018.2821177. URL http://dx.doi.org/10.1109/TCSVT.2018.2821177.

Yiming Zuo, Hongyu Wen, Venkat Subramanian, Patrick Chen, Karhan Kayan, Mario Bijelic, Felix Heide, and Jia Deng. Zero-shot depth from defocus. *arXiv preprint arXiv:2603.26658*, 2026. URL https://arxiv.org/abs/2603.26658.