GeoMFF: Reconstructing a Latent Sharp 3D Scene for Misaligned Multi-Focus Image Fusion

Anonymous authors
Paper under double-blind review

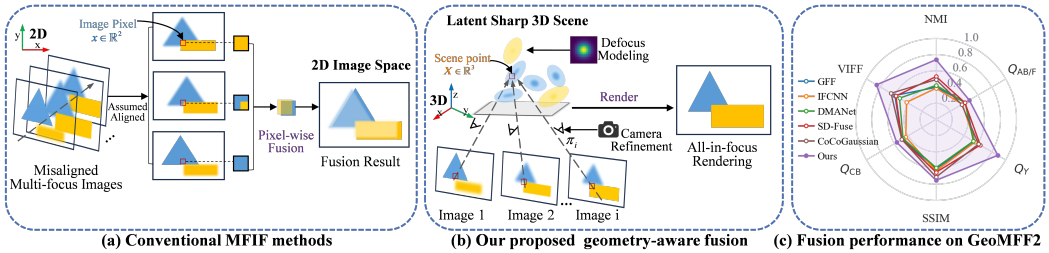

Figure 1: Conventional MFIF methods fuse images by combining content at corresponding pixel locations in 2D image space, making them susceptible to artifacts under misalignment. GeoMFF instead formulates misaligned MFIF as reconstructing a latent sharp 3D scene from observations with varying camera poses and defocus. By jointly optimizing scene representation, camera poses, and defocus modeling, it produces structurally consistent all-in-focus results.

Abstract

Multi-focus image fusion (MFIF) aims to integrate focused information from different focal planes into an all-in-focus image. However, existing methods typically perform focus estimation and fusion in the 2D image space, assuming pixel-wise correspondence among source images. They therefore struggle to handle spatial misalignment caused by camera motion and depth-dependent parallax during practical acquisition. To address this issue, we are the first to formulate misaligned MFIF as the recovery of a latent sharp 3D scene from multiple defocused observations, and propose GeoMFF, a geometry-aware fusion framework. GeoMFF initializes a 3D Gaussian scene using camera parameters and point clouds predicted by VGGT, and jointly optimizes the scene representation, camera poses, and a view-dependent spatially varying defocus model to aggregate focused information from different focal planes in 3D space. Within this framework, the Gaussian-Initialized Mixture Blur Network (GMBN) compactly models spatially varying defocus by adaptively combining a shared set of learnable blur kernels. We further construct a real-world misaligned multi-focus image dataset comprising two-image pairs and seven-image sequences. Experiments on our dataset and multiple public benchmarks demonstrate that our method achieves state-of-the-art performance on misaligned image pairs, remains highly competitive on aligned benchmarks, and extends effectively to multi-focus image sequences.

1 Introduction

Optical imaging systems have a limited depth of field, so a single image often contains both focused and defocused regions. Defocus blur removes fine details, weakens object boundaries, and makes scene interpretation less reliable. Multi-focus image fusion (MFIF) addresses this limitation by fusing images captured at different focal settings into an all-in-focus image. Ideally, the fused result preserves sharp structures from all source images and provides a more complete visual description of the scene, which can benefit downstream applications such as target detection Zhu et al. (2023); Yi et al. (2025), segmentation Anand et al. (2023), and recognition Liang et al. (2019).

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.

Most existing MFIF methods assume that the source images are precisely aligned (Basu et al., 2025; Wang et al., 2026b; Lin et al., 2025; Zhang et al., 2021b; Ouyang et al., 2026) or contain only slight mis-registration (Liang et al., 2019; Liu et al., 2015; Zhou et al., 2014). In real applications, however, source images are often not perfectly aligned. Slight camera shake and defocus spread effect (DSE) can break pixel-level correspondence across images, especially in hand-held acquisition. This problem becomes more severe when the input consists of multi-focus image sequences rather than a single image pair. Under such conditions, the same pixel coordinate may correspond to different scene points in different images, making conventional 2D image fusion unreliable. The resulting mismatch can lead to structural discontinuities and duplicated edges, as shown in Fig. 4 and Fig. 5.

Misalignment has become an important factor limiting the visual quality of MFIF in practical applications. A straightforward solution is to register the source images and then fuse them using an existing MFIF method (Liu & Yu, 2015; Laliberté et al., 2003). However, this registration–fusion pipeline has three limitations. First, the defocus spread effect (DSE) reduces the reliability of feature matching, while depth-dependent parallax caused by camera motion produces spatially non-uniform image displacements, making accurate registration more difficult. Second, conventional MFIF methods primarily based on sharpness estimation and content selection typically assume aligned inputs and have limited ability to correct residual geometric misalignment, making them prone to fusion artifacts. Third, the two stages operate independently, allowing registration errors to propagate to fusion without a mechanism for using the fusion objective to refine registration.

These limitations indicate that performing registration and fusion in 2D image space makes it difficult to achieve both accurate geometric alignment and reliable multi-focus fusion. This motivates us to consider a unified 3D scene representation that brings geometric alignment and multi-focus observation modeling into a common framework. Through scene geometry and camera projection, the 3D representation describes spatial correspondences between source images, providing a basis for jointly addressing their geometric misalignment and observations under different focal settings. Building on this idea, we propose GeoMFF (Geometry-aware Multi-Focus Fusion), a framework that uses 3DGS Kerbl et al. (2023) to represent a latent sharp 3D scene. First, we use VGGT Wang et al. (2025a) to predict initial camera parameters and point maps, then process the point maps into a point cloud for initializing the 3D Gaussians. We subsequently refine the VGGT-predicted poses through camera pose optimization to improve projection consistency between the 3D scene and the source images. Meanwhile, we introduce a Gaussian-Initialized Mixture Blur Network (GMBN), which represents spatially varying blur kernels as weighted combinations of learnable kernel atoms and reblurs the sharp renderings to explain the source images under different focal settings. Combining a reblur consistency constraint with depth and blur mask regularization, we jointly optimize Gaussian attributes, camera poses, and GMBN parameters. After optimization, we bypass GMBN and directly render an all-in-focus fused image from the target camera. Our main contributions are summarized as follows:

- To the best of our knowledge, we are the first to formulate misaligned MFIF as the recovery of a latent sharp 3D scene from images captured under different camera poses and focal settings. Based on this formulation, we propose GeoMFF, a geometry-aware fusion framework that integrates geometric alignment and multi-focus observation modeling into a unified 3D reconstruction process to aggregate complementary sharp information.

- We introduce a camera pose optimization strategy based on residual rigid transformations to refine initial poses estimated under defocus. This optimization reduces the structural misalignment between scene projections and source images, improving boundary consistency and detail preservation.

- We introduce GMBN to model the spatially varying defocus blur in multi-focus source images. GMBN represents local blur through spatially adaptive combinations of shared learnable kernel atoms, retaining local modeling capacity while constraining the blur representation through parameter sharing.

- We construct GeoMFF, a real-world misaligned multi-focus image dataset containing both two-source and multi-source inputs, and conduct extensive experiments on this dataset and several public benchmarks. The results demonstrate that our method achieves leading fusion performance under misalignment and extends to multiple source images.

2 Related Work

Multi-focus image fusion. Existing MFIF methods can be broadly categorized as traditional and deep learning-based approaches. Traditional methods operate either in the transform domain, including multi-scale decomposition, sparse representation, and gradient-domain fusion (Shreyamsha Kumar, 2013; Li & Jiang, 2020; Zhang et al., 2018; Li et al., 2021; Fu et al., 2020), or in the spatial domain, where focus is estimated at the block, region, or pixel level (Li et al., 2001; De & Chanda, 2013; Tian et al., 2018; Li et al., 2006; Yin et al., 2019; Zhang et al., 2021b; Aymaz & Köse, 2019; Chen et al., 2022). Deep learning methods learn focus decisions and image reconstruction from data. Early supervised methods predict focus maps or fused images using CNN-based models and regression learning (Liu et al., 2017; Li et al., 2020). Other methods perform feature extraction, fusion, and reconstruction in deep feature spaces (Zhang et al., 2020; Jung et al., 2020; Zhang & Ma, 2021; Xu et al., 2022; Ma et al., 2022b). Zheng et al. (2025) further propose MCCSR-Net by unfolding coupled convolutional sparse representation into a learnable network. To reduce the reliance on paired data and focus labels, Zhang et al. (2021a) introduce MFF-GAN with adversarial and gradient constraints, while Hu et al. (2023) propose ZMFF based on per-instance deep image prior optimization. Diffusion models have also been introduced for general image fusion and MFIF (Zhao et al., 2023; Li et al., 2024). Recent methods further model defocus and degradation explicitly: Quan et al. (2025) propose DMANet for spatially varying defocus blur, Ouyang et al. (2026) use multi-expert textual constraints for degradation-aware fusion, and Wang et al. (2026a) introduce structural priors to improve focus-boundary and local-detail modeling. Traditional and deep fusion methods generally assume aligned input images and operate in a shared 2D coordinate system, making them insufficiently adapted to misaligned inputs.

Misaligned multi-focus image fusion. Only a few existing methods explicitly address this issue, primarily in the 2D image space. Liu & Yu (2015) adopt a registration-then-fusion pipeline based on feature matching and wavelet fusion. Zhang & Levine (2016) propose MRSR, which incorporates spatial context into robust sparse representation, while Liang et al. (2019) jointly perform registration and fusion for microscopic image sequences. Li et al. (2013) combine morphological filtering and image matting to extract focused regions from moderately misaligned inputs. Recent learning-based methods incorporate alignment into the fusion process. Ji et al. (2023) perform homography-based coarse alignment and use deformable convolution to compensate for residual displacement. Chen et al. (2024) propose DSAF-Net, which estimates defocus maps and optical flow to warp source images to a common view before coarse-to-fine aggregation. In contrast, Zhu et al. (2025) directly fuse misaligned industrial images using a stage-wise strategy without explicit registration. Several conventional methods also exhibit robustness to limited misalignment through local or multi-scale processing (De & Chanda, 2013; Zhou et al., 2014; Ma et al., 2019; Liu et al., 2015; Zhang et al., 2018). These methods reduce the sensitivity of focus estimation to small spatial displacement. This limitation motivates our geometry-aware formulation, which jointly recovers camera poses and a shared 3D scene while modeling spatially varying defocus.

Defocus-aware 3D reconstruction. Representative 3D reconstruction methods, such as NeRF (Mildenhall et al., 2021) and 3DGS (Kerbl et al., 2023), typically assume sharp input images. To handle defocused observations, defocus-aware 3D reconstruction methods explicitly incorporate image degradation into scene optimization and model defocus formation using deformable blur kernels, circles of confusion (CoC), or thin-lens camera models (Ma et al., 2022c; Wu et al., 2022; Kim et al., 2024; Peng et al., 2024). The methods most closely related to ours are CoCoGaussian and DOF-GS, which incorporate physically based CoC models into 3DGS for scene reconstruction from defocused images (Lee et al., 2025; Wang et al., 2025b). However, these methods typically target conventional multi-view inputs with sufficient parallax and rely on reliable camera and geometric initialization. In contrast, MFIF images exhibit limited viewpoint variations and provide insufficient parallax for stable reconstruction, while severe defocus further reduces the reliability of pose estimation. We therefore adopt learned geometric initialization and jointly optimize the Gaussian scene, camera poses, and a spatially varying defocus model.

3 Method

3.1 Multi-Focus Fusion as Latent Sharp 3D Scene Recovery

Conventional MFIF methods typically perform focus estimation and content fusion in the 2D image space, producing all-in-focus results by selecting and combining information at the same pixel coordinates or corresponding local neighborhoods across source images. This process relies on

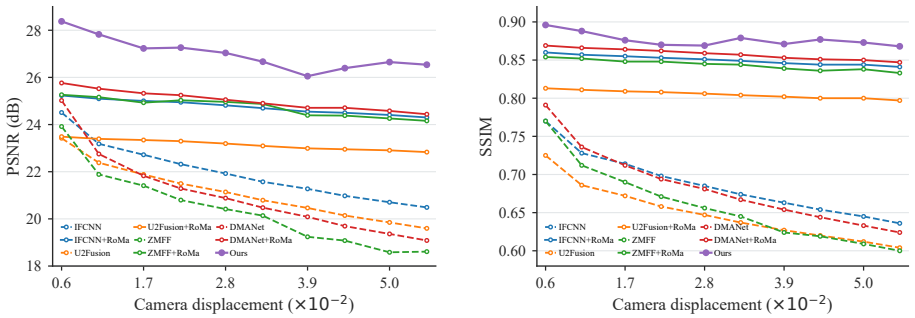

Figure 2: Sensitivity of MFIF methods to input misalignment on synthetic data. We vary camera translation while keeping rotation fixed to precisely control spatial misalignment.

spatial alignment between the source images, such that the same coordinate corresponds to the same physical scene location. However, in misaligned multi-focus images, camera pose variations cause the same scene point to project to different image locations. In practice, few MFIF methods explicitly address misalignment, and most still assume aligned source images. Pre-registration can reduce spatial misalignment, but defocus weakens cross-image correspondence estimation and leaves residual errors. Moreover, it may force pixels into incorrect correspondences, introducing artifacts. As shown in Fig. 2, RoMa Edstedt et al. (2024) alleviates this issue, but PSNR still decreases markedly as displacement increases.

Although the source images no longer share consistent 2D pixel coordinates, they are still captured from the same static 3D scene under different camera poses and focus settings. For a scene point $\mathbf{X} \in \mathbb{R}^3$ visible in both views, its ideal correspondence satisfies

$$\phi_{j\to i}\left(\mathcal{P}(\mathbf{X};\pi_j)\right) = \mathcal{P}(\mathbf{X};\pi_i), \tag{1}$$

where $\phi_{j\to i}$ denotes the 2D correspondence mapping from the j-th image to the i-th image. π_i and π_j are the camera parameters of the source images. $\mathcal{P}$ denotes the camera projection operator. This indicates that cross-view geometric correspondences can be jointly determined by the shared scene structure and camera parameters.

We therefore formulate misaligned MFIF as the recovery of a latent sharp 3D scene from multiple defocused observations, avoiding the errors and artifacts introduced by explicit image registration. Given a set of multi-focus source images $\{I_i\}_{i=1}^{N}$, we use a latent sharp 3D scene $\mathcal{G}$ to represent the shared scene content, and model the i-th input image I_i as an observation of this scene under camera parameters π_i and a defocus blur operator $\mathcal{B}_i$:

$$I_i \approx \mathcal{B}_i\left(\mathcal{R}(\mathcal{G},\pi_i)\right), \tag{2}$$

where $\mathcal{R}(\cdot)$ is the rendering process of the latent 3D scene, and $\pi_i = (\mathbf{K}_i, \mathbf{T}_i)$ denotes the camera model of the i-th image, where $\mathbf{K}_i$ and $\mathbf{T}_i$ are the intrinsic and extrinsic camera parameters, respectively. $\mathcal{B}_i$ denotes a view-dependent spatially varying defocus blur model. This observation model relates the scene content across different source images through scene geometry and camera projection, while using the defocus operator to account for imaging differences under different focus settings. The complementary sharp information contained in multiple source images jointly constrains the latent scene $\mathcal{G}$. After optimization, the all-in-focus fusion result is directly rendered from a target camera, without applying the defocus operator.

To this end, we propose GeoMFF, a geometry-aware multi-focus image fusion framework. As illustrated in Fig. 3, GeoMFF uses 3DGS as its 3D representation and comprises geometry initialization, camera pose refinement, view-dependent defocus modeling (GMBN), and joint optimization. We use VGGT to predict initial camera parameters and point maps from multi-focus source images. The confidence of the point maps is adjusted according to the blur level, and the processed point maps are converted into a point cloud for initializing the 3D Gaussian scene. The camera parameters predicted by VGGT are used only as initial estimates. During joint optimization, we refine the camera poses through residual rigid transformations and render latent sharp images and depth maps from the 3D Gaussian scene. GMBN models a view-dependent spatially varying defocus operator based on the rendered results, producing reblurred observations corresponding to the source images. By combining reblur consistency constraints with depth regularization and blur-mask regularization, GeoMFF jointly optimizes the 3D Gaussian attributes, camera poses, and GMBN parameters.

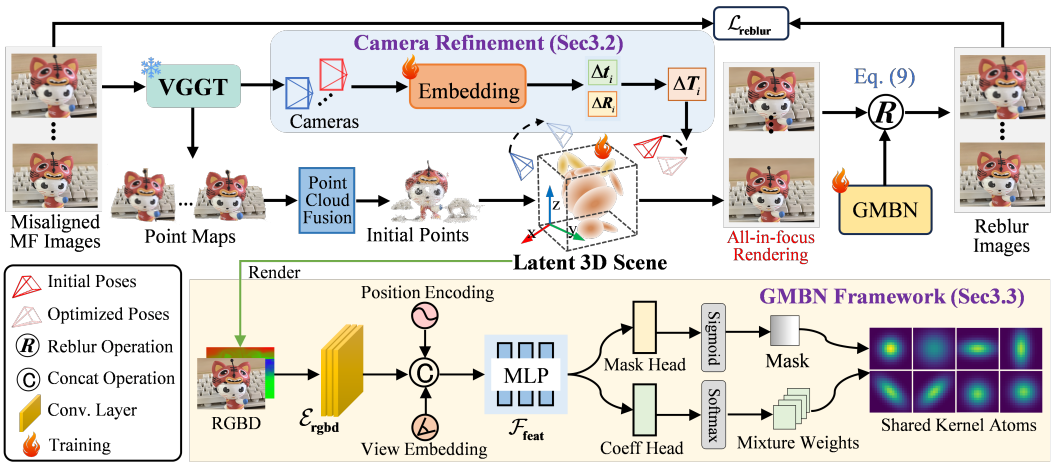

Figure 3: Overview of the GeoMFF framework. VGGT predicts camera parameters and point maps from misaligned multi-focus inputs to initialize a 3D Gaussian scene. The scene, camera poses, and GMBN are jointly optimized. GMBN represents spatially varying blur kernels through shared learnable kernel atoms. At inference, an all-in-focus image is rendered from the optimized 3D scene.

3.2 Geometry Initialization and Camera Parameter Refinement

We reconstruct the latent sharp scene $\mathcal{G}$ using 3DGS, which represents scenes as a set of optimizable 3D Gaussian ellipsoids. Constructing $\mathcal{G}$ requires an initial point cloud and camera parameters. For misaligned multi-focus images, the source images typically exhibit only small viewpoint variations, while defocus degradation weakens local textures and object boundaries, making it difficult for SfM-based methods (Snavely et al., 2006; Wu, 2013; Schönberger & Frahm, 2016; Liu et al., 2024; Wang et al., 2024) to obtain stable initialization. To this end, we adopt VGGT Wang et al. (2025a), a feed-forward visual geometry model, to predict initial camera parameters $\{\pi_i^0\}_{i=1}^N$ and corresponding point maps $\{\hat{\mathbf{P}}_i\}_{i=1}^N$ from the multi-focus source images:

$$\left(\{\pi_i^0\}_{i=1}^N, \{\hat{\mathbf{P}}_i\}_{i=1}^N\right) = \mathcal{F}_{\text{VGGT}}\left(\{I_i\}_{i=1}^N\right), \tag{3}$$

To avoid a sparse initial point cloud, we use a lower confidence threshold $\theta = 0.5$ to retain predicted points from the VGGT point maps and fuse the retained points from all source images into a unified initial point cloud $\mathcal{P}^0$, which is used to initialize the 3D Gaussian scene.

Although VGGT provides usable camera parameters, its predicted poses may still contain errors due to weak parallax and defocus degradation and therefore cannot be treated as fixed camera calibration results. As shown in Eq. 2, camera poses determine the projection of the latent sharp scene. Pose errors can introduce structural inconsistencies between rendered results and source images and may be incorrectly absorbed by the 3D Gaussian scene or the defocus model, thereby affecting multi-focus information aggregation. We therefore refine the initial camera poses using residual rigid transformations. Specifically, we fix the camera intrinsics $\mathbf{K}_i$ predicted by VGGT and optimize the camera-to-world poses $\mathbf{T}_i$. For the i-th source image, let $\mathbf{T}_i^0 \in SE(3)$ denote the initial camera-to-world pose predicted by VGGT. We introduce a learnable residual rigid transformation $\Delta\mathbf{T}_i \in SE(3)$ and represent the optimized pose as

$$\mathbf{T}_i = \mathbf{T}_i^0 \Delta\mathbf{T}_i, \quad \text{where} \quad \Delta\mathbf{T}_i = \begin{bmatrix} \Delta\mathbf{R}_i & \Delta\mathbf{t}_i \\ \mathbf{0}^\top & 1 \end{bmatrix}, \tag{4}$$

where $\Delta\mathbf{R}_i \in SO(3)$ and $\Delta\mathbf{t}_i \in \mathbb{R}^3$ denote the rotation and translation residuals, respectively. The residual transformation is initialized as the identity, allowing optimization to start from the feed-forward VGGT pose while ensuring that the updated pose remains a valid rigid transformation. Accordingly, the camera parameters used in subsequent rendering are denoted as $\pi_i = (\mathbf{K}_i, \mathbf{T}_i)$, where $\mathbf{T}_i$ is the optimized camera-to-world pose.

3.3 Defocus Modeling with Gaussian-Initialized Kernel Mixtures

According to Eq. (2), the i-th source image I_i can be regarded as an observation of the latent sharp rendering $R_i = \mathcal{R}(\mathcal{G}, \pi_i)$ degraded by the defocus operator $\mathcal{B}_i$. Following the formulation of

defocus degradation using a point spread function (Quan et al., 2025; Aslantas & Toprak, 2017), we approximate this process as a local reblurring operation with spatially varying blur kernels:

$$\tilde{I}_i(\mathbf{x}) = \sum_{\mathbf{u}\in\Omega} R_i(\mathbf{x}-\mathbf{u}) D_i(\mathbf{x},\mathbf{u}), \tag{5}$$

where $\mathbf{x}$ denotes the pixel location, $\mathbf{u}$ is the offset within the local window Ω, $D_i(\mathbf{x},\mathbf{u})$ is the blur kernel at location $\mathbf{x}$, and $\tilde{I}_i$ is the locally reblurred result.

Considering that defocus blur typically manifests as the smooth spread of local image content, we design a Gaussian-initialized mixture blur network (GMBN) to model view-dependent spatially varying defocus degradation. GMBN maintains M shared learnable kernel atoms $\{A_m\}_{m=1}^{M}$ and constructs local blur kernels using positional mixture weights:

$$D_i(\mathbf{x},\mathbf{u}) = \sum_{m=1}^{M} w_{i,m}(\mathbf{x}) A_m(\mathbf{u}), \tag{6}$$

where $A_m(\mathbf{u})$ denotes the response of the m-th kernel atom at offset $\mathbf{u}$, and $w_{i,m}(\mathbf{x})$ denotes its mixture weight at location $\mathbf{x}$ in the i-th image. We initialize the kernel atoms using anisotropic Gaussian kernels with different initial scales, orientations, and spatial offsets, providing diverse initial blur patterns for constructing local kernels:

$$A_m^{(0)}(\mathbf{u}) = \frac{1}{C_m} \exp\left[-\frac{1}{2}(\mathbf{u}-\boldsymbol{\mu}_m)^\top \left(\boldsymbol{\Sigma}_m^{(0)}\right)^{-1} (\mathbf{u}-\boldsymbol{\mu}_m)\right], \tag{7}$$

where C_m is a normalization constant ensuring that the kernel responses sum to 1 over the discrete window Ω, $\boldsymbol{\mu}_m$ controls the initial spatial offset, and $\boldsymbol{\Sigma}_m^{(0)}$ controls the initial scale, anisotropy, and orientation.

GMBN adopts a lightweight convolutional architecture, as shown in Fig. 3. For the i-th source image, it takes the latent sharp rendering R_i, the rendered depth map Z_i as inputs. A convolution encoder $\mathcal{E}_{\text{rgbd}}$ first extracts joint RGB-D features from R_i and Z_i, which are then concatenated with the view embedding e_i and positional encoding $\gamma(\mathbf{x})$ to obtain conditional features $\mathcal{F}_{\text{feat}}$, which can be defined as:

$$\mathcal{F}_{\text{feat}}(\mathbf{x}) = \left[\mathcal{E}_{\text{rgbd}}([R_i, Z_i])(\mathbf{x}), e_i, \gamma(\mathbf{x})\right]. \tag{8}$$

Based on $\mathcal{F}_{\text{feat}}$, two parallel prediction heads output the mixture weights of the kernel atoms and a spatially varying blur mask, respectively. The coefficient head predicts the mixture weights $w_{i,m}(\mathbf{x})$ for the M shared kernel atoms. Since multi-focus source images contain both sharp and defocus regions, we further introduce a blur mask $\hat{M}_i$ to control the contribution of the reblurring branch at different locations and avoid unnecessary smoothing in sharp regions. Finally, the blur mask blends the locally reblurred result with the latent sharp rendering to obtain the predicted observation:

$$I_i^{\text{reblur}}(\mathbf{x}) = \hat{M}_i(\mathbf{x})\tilde{I}_i(\mathbf{x}) + \left(1-\hat{M}_i(\mathbf{x})\right) R_i(\mathbf{x}). \tag{9}$$

3.4 Optimization Objective

We jointly optimize the 3D Gaussian scene, camera parameters, and GMBN using a reblur reconstruction loss. Following 3DGS, we combine an ℓ_1 term with an SSIM term between the predicted reblurred image and the corresponding multi-focus source image:

$$\mathcal{L}_{\text{reblur}} = (1-\lambda_{\text{ssim}})\left\|I_i^{\text{reblur}} - I_i\right\|_1 + \lambda_{\text{ssim}}\left(1-\text{SSIM}(I_i^{\text{reblur}}, I_i)\right), \tag{10}$$

where $\lambda_{\text{ssim}} = 0.2$, consistent with 3DGS. Severe defocus weakens local textures and edges, making geometry estimation based on photometric consistency more susceptible to ambiguity. We therefore apply a total-variation regularizer to the rendered depth maps Z_i as follows, encouraging local depth smoothness and mitigating irregular depth variations in the scene geometry.

$$\mathcal{L}_{\text{depth}} = \|\nabla_x Z_i\|_1 + \|\nabla_y Z_i\|_1. \tag{11}$$

To reduce unnecessary reblurring contributions, we penalize the mean value of the predicted blur mask $\hat{M}_i$ during joint optimization. Together with the reconstruction loss, this regularization encourages the model to fit the defocused observations while reducing excessive reliance on the reblurring operation. The overall optimization objective is

$$\mathcal{L} = \mathcal{L}_{\text{reblur}} + \lambda_{\text{depth}}\mathcal{L}_{\text{depth}} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}}, \tag{12}$$

where $\mathcal{L}_{\text{mask}}$ is the mean value of the predicted blur mask. The coefficients λ_{depth} and λ_{mask} control the strengths of the depth regularization and mask sparsity constraints, respectively.

4 EXPERIMENTS

We conduct comprehensive experiments to evaluate the proposed method. We first describe the experimental setup, then compare our method with state-of-the-art approaches, and finally conduct ablation studies to analyze the key components.

4.1 EXPERIMENTAL SETUP

Implementation. Our method is implemented in PyTorch Paszke et al. (2019) using the gsplat Ye et al. (2026) framework. The Gaussian parameters, camera-pose residuals, and GMBN are optimized using the Adam optimizer. The learning rates of the Gaussian parameters follow the original 3DGS settings. The camera-pose residuals are initialized to zero and optimized with a learning rate of 1×10^{-5}, while the learning rate of GMBN is set to 5×10^{-4}. The remaining Adam hyperparameters use the PyTorch default settings. The initial Gaussians are generated from the point cloud predicted by VGGT, and each per-pixel blur kernel consists of 8 Gaussian basis kernels. Following Peng et al. (2024), we adopt a progressive strategy that gradually increases the rendering resolution and expands the per-pixel blur kernel size from 5×5 to 9×9 and 17×17. The loss weights λ_{depth} and λ_{mask} are both set to 0.001. The optimization runs for 22K iterations on GeoMFF and 15K iterations on the other datasets. All experiments are conducted on an NVIDIA RTX 5090 GPU.

Datasets. Evaluating MFIF on aligned images does not fully reflect the combined challenges of camera motion and focal plane changes encountered during real-world capture. To this end, we construct GeoMFF, a real-world multi-focus image dataset that preserves spatial misalignment and defocus variations for evaluating misaligned MFIF. GeoMFF comprises 25 multi-focus image sequences containing 115 source images, captured using smartphones and digital cameras. GeoMFF2 contains 12 two-image pairs, whereas GeoMFF7 contains 13 seven-image sequences, supporting performance evaluation on image pairs and image sequences, respectively. Detailed settings are provided in the **supplementary material**. In addition, we select 10 grayscale multi-focus image pairs with slight spatial misalignment from the public Classic dataset Ma et al. (2022a) to evaluate the ability of different methods to handle mildly misregistered inputs. To further assess our method under conventional MFIF settings, we conduct experiments on two commonly used aligned datasets: Lytro Nejati et al. (2015) and MFFW Xu et al. (2020).

Baselines and Evaluation Metrics. To evaluate our method, we compare it with 10 representative MFIF methods, including GFF Li et al. (2013), IFCNN Zhang et al. (2020), DRPL Li et al. (2020), MFF-GAN Zhang et al. (2021a), U2Fusion Xu et al. (2022), SwinFusion Ma et al. (2022b), ZMFF Hu et al. (2023), DMANet Quan et al. (2025), Text-MFF Ouyang et al. (2026), and SD-Fuse Wang et al. (2026a). These baselines span traditional, supervised, unsupervised, zero-shot, and degradation-aware fusion paradigms. We further compare with CoCoGaussian Lee et al. (2025), a defocus-aware 3D reconstruction method closely related to our formulation, and evaluate a pre-registration setting in which RoMa Edstedt et al. (2024) is applied to the source images before fusion.

We employ 10 objective metrics: normalized mutual information (NMI), mean absolute error (MAE), standard deviation (SD), spatial frequency (SF), edge-based fusion quality ($Q_{\mathrm{AB/F}}$), Yang's structural similarity-based fusion metric (Q_Y), the sum of correlations of differences (SCD), structural similarity (SSIM), the Chen–Blum metric (Q_{CB}), and visual information fidelity for fusion (VIFF). Together, these metrics provide a comprehensive evaluation across information transfer, image features, image structural preservation, and perceptual quality. As conventional MFIF metrics assume spatially corresponding source images, we use RoMa to align each source-image pair for metric computation.

4.2 COMPARISON WITH STATE-OF-THE-ART METHODS

Results on GeoMFF2. As shown in Table 1, our method achieves the best performance on 9 metrics. Our method improves NMI by 39.2% and reduces MAE by 21.2% over the best competing results, demonstrating more effective integration of complementary focused information under spatial misalignment. It also improves $Q_{\mathrm{AB/F}}$, Q_Y, and VIFF over CoCoGaussian by 29.3%, 45.4%, and 32.3%, respectively, indicating better edge, structural, and perceptual preservation. As shown in Fig. 4, existing MFIF methods often produce ghosting, blurred boundaries, or structural loss. Although CoCoGaussian reconstructs a 3D scene, its CoC-based blur modeling alone can still yield blurred boundaries and missing details. By jointly refining camera poses and modeling spatially varying defocus with GMBN, GeoMFF produces clearer and more coherent structures, demonstrating the effectiveness of unified 3D modeling for real-world misaligned MFIF. More qualitative results and pre-registration setting results are provided in the **supplementary material**.

Table 1: Quantitative comparison on the GeoMFF2 dataset. The best and second-best results are highlighted in bold and underlined, respectively.

Method	NMI↑	MAE↓	SD↑	SF↑	$Q_{AB/F}$ ↑	Q_Y ↑	SCD↑	SSIM↑	Q_{CB} ↑	VIFF↑
GFF	0.3967	21.1806	40.8825	14.5341	<u>0.4174</u>	0.5565	-0.0842	0.6067	<u>0.4939</u>	0.5913
IFCNN	0.3852	21.4361	40.1252	15.2776	0.3755	0.5587	-0.0918	0.6357	0.4348	0.4237
DRPL	0.4067	24.5868	<u>43.4044</u>	**19.1370**	0.3581	0.4292	-0.1610	0.5267	0.4515	0.3712
U2Fusion	0.3395	23.2957	39.4495	12.3320	0.3257	<u>0.8211</u>	-0.1088	0.6329	0.4312	0.3728
MFF-GAN	0.4116	19.1077	39.8980	15.1098	0.3851	0.6099	-0.0193	0.6696	0.4576	0.4910
SwinFusion	0.3796	20.7216	38.6794	12.9396	0.3749	0.5405	-0.1182	0.6532	0.4298	0.4502
ZMFF	0.4014	22.6894	42.2939	<u>16.4376</u>	0.4125	0.5142	-0.1519	0.5651	0.4793	0.4891
DMANet	0.4205	21.3946	41.2652	15.3721	0.4163	0.5358	-0.1214	0.5959	0.4743	0.5249
SD-Fuse	<u>0.5294</u>	19.5266	40.5296	12.8797	0.4138	0.6346	-0.0814	0.6531	0.4907	0.5779
Text-MFF	0.3849	76.6190	19.8310	7.4052	0.1328	0.3718	-0.9121	0.5647	0.1874	0.3071
CoCoGaussian	0.4923	<u>18.4133</u>	42.5611	11.8773	0.3681	0.6091	<u>0.1595</u>	<u>0.7121</u>	0.4924	<u>0.6461</u>
Ours	**0.7371**	**14.5106**	**43.5676**	14.8173	**0.4758**	**0.8856**	**0.3671**	**0.7514**	**0.5657**	**0.8549**

Table 2: Ablation study on GeoMFF2. The best results are highlighted in bold.

Variant	NMI↑	MAE↓	SD↑	SF↑	$Q_{AB/F}$ ↑	Q_Y ↑	SCD↑	SSIM↑	Q_{CB} ↑	VIFF↑
w/o Camera Refinement	0.6709	17.7297	43.3206	14.5163	0.4268	0.8379	0.2693	0.7101	0.5375	0.7673
w/o Defocus Modeling	0.7054	**13.2665**	42.2837	11.5617	0.4671	0.8648	0.3030	**0.7921**	0.5341	0.8325
w/o Depth Regularization	0.7222	14.5474	43.3017	13.9788	0.4721	0.8676	0.3651	0.7394	0.5553	0.8476
w/o Mask Regularization	0.6567	14.9452	43.5139	14.4239	0.4396	0.8587	0.3514	0.7433	0.5363	0.7761
Free-form Blur Kernel	0.7178	14.5544	43.2706	13.8630	0.4699	0.8679	0.3669	0.7501	0.5618	0.8523
Full Model	**0.7371**	14.5106	**43.5676**	**14.8173**	**0.4758**	**0.8856**	**0.3671**	0.7514	**0.5657**	**0.8549**

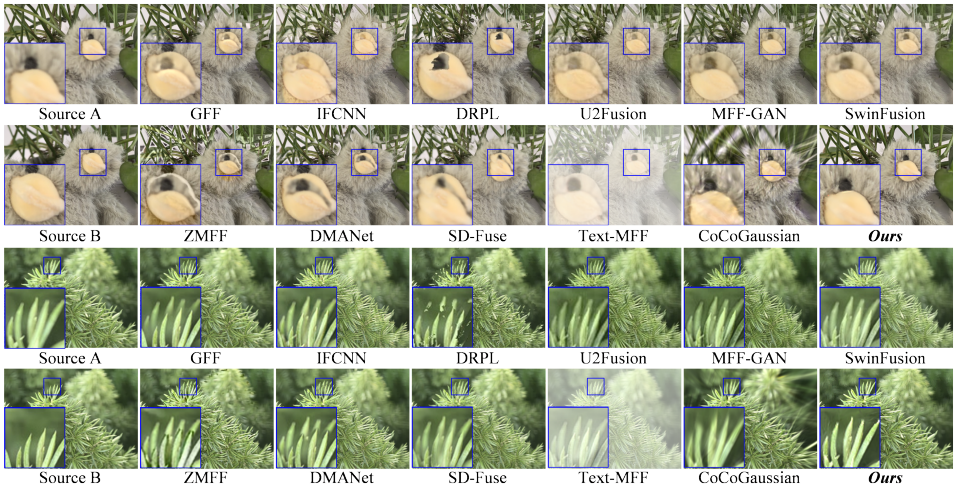

Figure 4: Qualitative comparison of fusion results on representative image pairs from GeoMFF2.

Results on GeoMFF7. Since existing MFIF metrics are primarily designed for image pairs, we conduct qualitative comparisons in Fig. 5. Among conventional MFIF methods, only IFCNN supports sequence inputs; however, it accumulates fusion errors and produces severe ghosting as the number of misaligned views increases. Although pre-registration partially alleviates ghosting and double-edge artifacts, residual registration errors still introduce abnormal textures and redundant structures inconsistent with the source images. In contrast, both CoCoGaussian and GeoMFF mitigate ghosting through a unified 3D representation. Our camera pose refinement and defocus modeling better integrate focused information, producing more coherent structures across depths and supporting multi-source misaligned MFIF.

We further evaluate on the **Classic**, **Lytro**, and **MFFW** datasets. The results show that our method is also effective on mildly misaligned and aligned inputs. See **supplementary material** for details.

4.3 Ablation Studies

We conduct ablation studies on geometry initialization, camera refinement, defocus modeling, mask regularization, and depth regularization. The results are shown in Table 2, and Fig. 6.

Geometry Initialization. MFIF source images exhibit defocus and limited viewpoint variation, making conventional SfM initialization unreliable. As shown in Fig. 6 (a), COLMAP typically

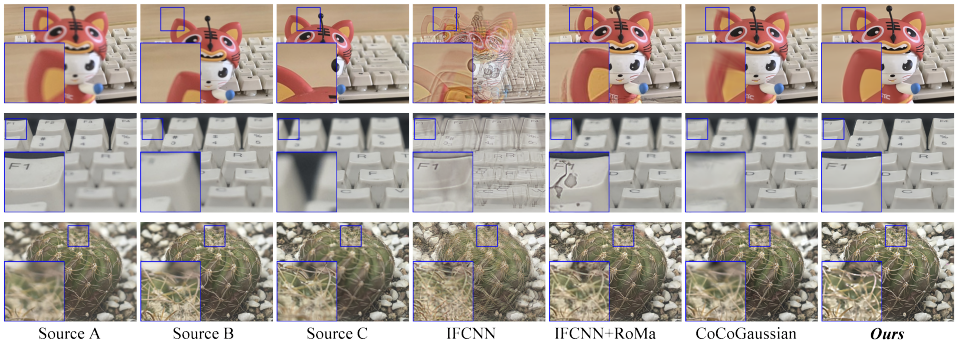

Figure 5: Qualitative comparison on representative seven-image sequences from GeoMFF7.

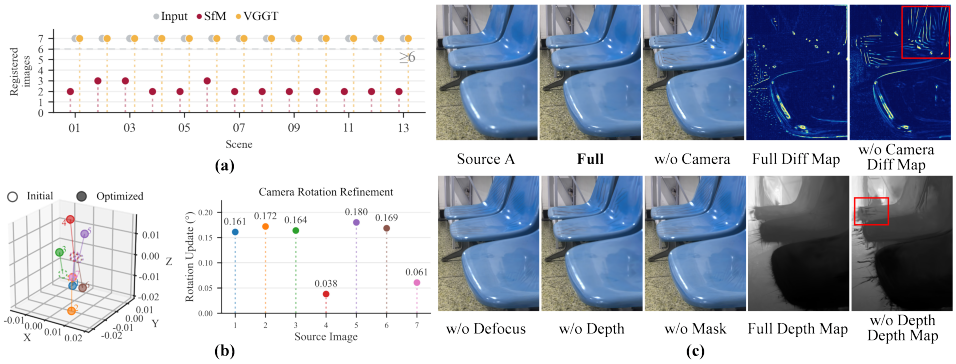

Figure 6: (a) Successful scene initialization. (b) Camera position and rotation updates after refinement. (c) Qualitative comparisons of component ablations.

initializes only 2 or 3 source images, whereas VGGT covers the complete set of 7 source images, enabling the subsequent joint optimization of the scene representation.

Camera Pose Refinement. Severe defocus introduces errors into the initial camera parameters. The full model achieves clear improvements in NMI, MAE, $Q_{\mathrm{AB/F}}$, and VIFF, indicating that camera pose optimization mitigates observation inconsistencies. Its effect is particularly evident on GeoMFF7, where small pose errors accumulate across multiple source images. Fig. 6 (b) shows the refined camera rotations and translations for a representative seven-image scene.

Defocus Modeling. Without GMBN, the reconstructed scene retains defocused regions instead of integrating complementary sharp information, which also explains its higher SSIM and lower MAE. Replacing the shared Gaussian-initialized kernel mixture with free-form kernels introduces excessive flexibility, making blur estimation less stable. Thus, explicit defocus modeling separates the latent sharp scene from defocused observations, while the shared kernel atoms improve fusion quality.

Regularization. Removing mask regularization degrades all evaluation metrics and causes noticeable blurring at object boundaries (Fig. 6 (c)). Mask regularization suppresses excessive activation of the blur branch in focused regions, avoiding unnecessary blur modeling. Removing $\mathcal{L}_{\mathrm{depth}}$ slightly degrades all metrics, indicating that depth regularization improves optimization stability. Without it, the depth map exhibits noticeable discontinuities and the fused result contains black ripple artifacts. More ablations are provided in the **supplementary material**.

5 Conclusion

We present GeoMFF, a geometry-aware framework that reconstructs a latent sharp 3D scene from misaligned, defocused multi-focus images. GeoMFF reformulates 2D fusion as the reconstruction of a latent sharp 3D scene from defocused observations, jointly optimizing the 3D representation, camera poses, and spatially varying defocus. GMBN models defocus degradation through adaptive combinations of shared learnable kernel atoms. We also construct a real-world misaligned multi-focus dataset with two-image pairs and seven-image sequences. Comprehensive experiments demonstrate the effectiveness of GeoMFF on misaligned image pairs and multi-focus image sequences.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools to assist with translation. We did not use generative AI tools to design research methods or experiments, provide feedback on research methods or experiments, or interpret results. The remaining required disclosure tasks, including generating synthetic datasets, developing theoretical models or conceptual frameworks, proposing mathematical claims, providing key elements for mathematical proofs, and assisting with proof writing, were not applicable to this work. Additionally, we used generative AI tools to edit software code, improve the readability of the manuscript, identify relevant literature, and format references. All AI-assisted translations and manuscript edits were reviewed and revised by the authors. AI-assisted code was verified and tested by the authors, and suggested literature and bibliographic information were checked against original or official sources. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Vatsala Anand, Sheifali Gupta, Deepika Koundal, and Karamjeet Singh. Fusion of u-net and cnn model for segmentation and classification of skin lesion from dermoscopy images. *Expert Systems with Applications*, 213:119230, 2023. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2022.119230. URL https://www.sciencedirect.com/science/article/pii/S0957417422022485.

Veysel Aslantas and Ahmet Nusret Toprak. Multi-focus image fusion based on optimal defocus estimation. *Computers & Electrical Engineering*, 62:302–318, 2017. ISSN 0045-7906. doi: https://doi.org/10.1016/j.compeleceng.2017.02.003. URL https://www.sciencedirect.com/science/article/pii/S0045790617302501.

Samet Aymaz and Cemal Köse. A novel image decomposition-based hybrid technique with super-resolution method for multi-focus image fusion. *Information Fusion*, 45:113–127, 2019. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2018.01.015. URL https://www.sciencedirect.com/science/article/pii/S1566253517304232.

Shatabdi Basu, Sunita Singhal, and Dilbag Singh. Multi-focus image fusion: A systematic literature review. *SN Comput. Sci.*, 6(2), February 2025. doi: 10.1007/s42979-025-03678-y. URL https://doi.org/10.1007/s42979-025-03678-y.

Jun Chen, Xuejiao Li, Linbo Luo, and Jiayi Ma. Multi-focus image fusion based on multi-scale gradients and image matting. *IEEE Transactions on Multimedia*, 24:655–667, 2022. doi: 10.1109/TMM.2021.3057493.

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, 2024. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2023.102125. URL https://www.sciencedirect.com/science/article/pii/S1566253523004414.

Ishita De and Bhabatosh Chanda. Multi-focus image fusion using a morphology-based focus measure in a quad-tree structure. *Information Fusion*, 14(2):136–146, 2013. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2012.01.007. URL https://www.sciencedirect.com/science/article/pii/S1566253512000085.

Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. Roma: Robust dense feature matching. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19790–19800, 2024. doi: 10.1109/CVPR52733.2024.01871.

Guan-Peng Fu, Shao-Hua Hong, Fu-Lin Li, and Lin Wang. A novel multi-focus image fusion method based on distributed compressed sensing. *Journal of Visual Communication and Image Representation*, 67:102760, 2020. ISSN 1047-3203. doi: https://doi.org/10.1016/j.jvcir.2020.102760. URL https://www.sciencedirect.com/science/article/pii/S1047320320300109.

Xingyu Hu, Junjun Jiang, Xianming Liu, and Jiayi Ma. Zmff: Zero-shot multi-focus image fusion. *Information Fusion*, 92:127–138, 2023. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2022.11.014. URL https://www.sciencedirect.com/science/article/pii/S1566253522002263.

Jiayu Ji, Feng Pan, Xuanyin Wang, Jixiang Tang, Ze'an Liu, and Bin Pu. An end-to-end anti-shaking multi-focus image fusion approach. *Image and Vision Computing*, 137:104788, 2023. ISSN 0262-8856. doi: https://doi.org/10.1016/j.imavis.2023.104788. URL https://www.sciencedirect.com/science/article/pii/S0262885623001622.

Hyungjoo Jung, Youngjung Kim, Hyunsung Jang, Namkoo Ha, and Kwanghoon Sohn. Unsupervised deep image fusion with structure tensor representations. *IEEE Transactions on Image Processing*, 29:3845–3858, 2020. doi: 10.1109/TIP.2020.2966075.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4), July 2023. ISSN 0730-0301. doi: 10.1145/3592433. URL https://doi.org/10.1145/3592433.

Min-Jung Kim, Gyojung Gu, and Jaegul Choo. Lensnerf: Rethinking volume rendering based on thin-lens camera model. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3170–3179, 2024. doi: 10.1109/WACV57701.2024.00315.

France Laliberté, Langis Gagnon, and Yunlong Sheng. Registration and fusion of retinal images-an evaluation study. *IEEE Transactions on Medical Imaging*, 22(5):661–673, 2003.

Jungho Lee, Suhwan Cho, Taeoh Kim, Ho-Deok Jang, Minhyeok Lee, Geonho Cha, Dongyoon Wee, Dogyoon Lee, and Sangyoun Lee. Cocogaussian: Leveraging circle of confusion for gaussian splatting from defocused images. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16101–16110, 2025. doi: 10.1109/CVPR52734.2025.01501.

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. doi: 10.1109/TIP.2020.2976190.

Min Li, Wei Cai, and Zheng Tan. A region-based multi-sensor image fusion scheme using pulse-coupled neural network. *Pattern Recognition Letters*, 27(16):1948–1956, 2006. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2006.05.004. URL https://www.sciencedirect.com/science/article/pii/S0167865506001565.

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, 2024. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2023.121664. URL https://www.sciencedirect.com/science/article/pii/S0957417423021668.

Shutao Li, James T. Kwok, and Yaonan Wang. Combination of images with diverse focuses using the spatial frequency. *Information Fusion*, 2(3):169–176, 2001. ISSN 1566-2535. doi: https://doi.org/10.1016/S1566-2535(01)00038-0. URL https://www.sciencedirect.com/science/article/pii/S1566253501000380.

Shutao Li, Xudong Kang, Jianwen Hu, and Bin Yang. Image matting for fusion of multi-focus images in dynamic scenes. *Information Fusion*, 14(2):147–162, 2013. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2011.07.001. URL https://www.sciencedirect.com/science/article/pii/S1566253511000510.

Xiaosong Li, Fuqiang Zhou, Haishu Tan, Yuanze Chen, and Wangxia Zuo. Multi-focus image fusion based on nonsubsampled contourlet transform and residual removal. *Signal Processing*, 184: 108062, 2021. ISSN 0165-1684. doi: https://doi.org/10.1016/j.sigpro.2021.108062. URL https://www.sciencedirect.com/science/article/pii/S0165168421001006.

Yanping Li and Shengming Jiang. Multi-focus image fusion using geometric algebra based discrete fourier transform. *IEEE Access*, 8:60019–60028, 2020. doi: 10.1109/ACCESS.2020.2981814.

Yixiong Liang, Yuan Mao, Zhihong Tang, Meng Yan, Yuqian Zhao, and Jianfeng Liu. Efficient misalignment-robust multi-focus microscopical images fusion. *Signal Processing*, 161:111–123, 2019. ISSN 0165-1684. doi: https://doi.org/10.1016/j.sigpro.2019.03.020. URL https://www.sciencedirect.com/science/article/pii/S0165168419301136.

Huangxing Lin, Yunlong Lin, Jingyuan Xia, Linyu Fan, Feifei Li, Yingying Wang, and Xinghao Ding. Fusion2void: Unsupervised multi-focus image fusion based on image inpainting. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(4):3328–3341, 2025. doi: 10.1109/TCSVT.2024.3502200.

Shaohui Liu, Yidan Gao, Tianyi Zhang, Rémi Pautrat, Johannes L. Schönberger, Viktor Larsson, and Marc Pollefeys. Robust incremental structure-from-motion with hybrid features. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XXXVI*, pp. 249–269, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-72763-4. doi: 10.1007/978-3-031-72764-1_15. URL https://doi.org/10.1007/978-3-031-72764-1_15.

Yan Liu and Feihong Yu. An automatic image fusion algorithm for unregistered multiply multi-focus images. *Optics Communications*, 341:101–113, 2015. ISSN 0030-4018. doi: https://doi.org/10.1016/j.optcom.2014.12.015. URL https://www.sciencedirect.com/science/article/pii/S0030401814011559.

Yu Liu, Shuping Liu, and Zengfu Wang. Multi-focus image fusion with dense sift. *Information Fusion*, 23:139–155, 2015. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2014.05.004. URL https://www.sciencedirect.com/science/article/pii/S1566253514000670.

Yu Liu, Xun Chen, Hu Peng, and Zengfu Wang. Multi-focus image fusion with a deep convolutional neural network. *Information Fusion*, 36:191–207, 2017. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2016.12.001. URL https://www.sciencedirect.com/science/article/pii/S1566253516302081.

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, 2022a. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2021.10.115. URL https://www.sciencedirect.com/science/article/pii/S0925231221016726.

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.

Jinlei Ma, Zhiqiang Zhou, Bo Wang, Lingjuan Miao, and Hua Zong. Multi-focus image fusion using boosted random walks-based algorithm with two-scale focus maps. *Neurocomputing*, 335:9–20, 2019. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2019.01.048. URL https://www.sciencedirect.com/science/article/pii/S0925231219300748.

Li Ma, Xiaoyu Li, Jing Liao, Qi Zhang, Xuan Wang, Jue Wang, and Pedro V. Sander. Deblur-nerf: Neural radiance fields from blurry images. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12851–12860, 2022c. doi: 10.1109/CVPR52688.2022.01252.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: representing scenes as neural radiance fields for view synthesis. *Commun. ACM*, 65(1):99–106, December 2021. ISSN 0001-0782. doi: 10.1145/3503250. URL https://doi.org/10.1145/3503250.

Mansour Nejati, Shadrokh Samavi, and Shahram Shirani. Multi-focus image fusion using dictionary-based sparse representation. *Information Fusion*, 25:72–84, 2015. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2014.10.004. URL https://www.sciencedirect.com/science/article/pii/S1566253514001213.

Yuncan Ouyang, Hao Zhai, Jinyuan Jiang, Hanyue Hu, and Xiaohang Li. Text-mff: Degradation multi-focus image fusion using multi expert text constraints. *Expert Systems with Applications*, 311: 131369, 2026. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2026.131369. URL https://www.sciencedirect.com/science/article/pii/S0957417426002824.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: an imperative style, high-performance deep learning library. 2019.

Cheng Peng, Yutao Tang, Yifan Zhou, Nengyu Wang, Xijun Liu, Deming Li, and Rama Chellappa. Bags: Blur agnostic gaussian splatting through multi-scale kernel modeling. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXX*, pp. 293–310, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-72988-1. doi: 10.1007/978-3-031-72989-8_17. URL https://doi.org/10.1007/978-3-031-72989-8_17.

Yuhui Quan, Xi Wan, Zitao Tang, Jinxiu Liang, and Hui Ji. Multi-focus image fusion via explicit defocus blur modelling. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025. ISBN 978-1-57735-897-8. doi: 10.1609/aaai.v39i6.32714. URL https://doi.org/10.1609/aaai.v39i6.32714.

Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4104–4113, 2016. doi: 10.1109/CVPR.2016.445.

BK Shreyamsha Kumar. Multifocus and multispectral image fusion based on pixel significance using discrete cosine harmonic wavelet transform. *Signal, Image and Video Processing*, 7(6):1125–1143, 2013.

Noah Snavely, Steven M. Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. volume 25, pp. 835–846. Association for Computing Machinery, New York, NY, USA, July 2006. doi: 10.1145/1141911.1141964. URL https://doi.org/10.1145/1141911.1141964.

Juanxiu Tian, Guocai Liu, and Jingguang Liu. Multi-focus image fusion based on edges and focused region extraction. *Optik*, 171:611–624, 2018. ISSN 0030-4026. doi: https://doi.org/10.1016/j.ijleo.2018.06.093. URL https://www.sciencedirect.com/science/article/pii/S003040261830901X.

Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. Vggsfm: Visual geometry grounded deep structure from motion. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21686–21697, 2024. doi: 10.1109/CVPR52733.2024.02049.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306, 2025a. doi: 10.1109/CVPR52734.2025.00499.

Yujie Wang, Praneeth Chakravarthula, and Baoquan Chen. Dof-gs: Adjustable depth-of-field 3d gaussian splatting for post-capture refocusing, defocus rendering and blur removal. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21297–21306, 2025b. doi: 10.1109/CVPR52734.2025.01984.

Zeyu Wang, Jiayu Wang, Haiyu Song, Pengjie Wang, Kedi Lyu, Wei Li, and Libo Zhao. Sdfuse: An image structure-driven model for multi-focus image fusion. *Information Fusion*, 129: 104058, 2026a. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2025.104058. URL https://www.sciencedirect.com/science/article/pii/S1566253525011200.

Zeyu Wang, Shuang Yu, Haoran Duan, Shidong Wang, Yang Long, and Ling Shao. Rethinking multi-focus image fusion: An input space optimization view. *IEEE Transactions on Image Processing*, 35:1321–1336, 2026b. doi: 10.1109/TIP.2026.3654370.

Changchang Wu. Towards linear-time incremental structure from motion. pp. 127–134, 2013. URL https://api.semanticscholar.org/CorpusID:5296119.

Zijin Wu, Xingyi Li, Juewen Peng, Hao Lu, Zhiguo Cao, and Weicai Zhong. Dof-nerf: Depth-of-field meets neural radiance fields. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, pp. 1718–1729, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392037. doi: 10.1145/3503161.3548088. URL https://doi.org/10.1145/3503161.3548088.

Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1): 502–518, 2022. doi: 10.1109/TPAMI.2020.3012548.

Shuang Xu, Xiao-Li Wei, Chunxia Zhang, Junmin Liu, and Jiangshe Zhang. Mffw: A new dataset for multi-focus image fusion. *ArXiv*, abs/2002.04780, 2020. URL https://api.semanticscholar.org/CorpusID:211082872.

Vickie Ye, Ruilong Li, Justin Kerr, Matias Turkulainen, Brent Yi, Zhuoyang Pan, Otto Seiskari, Jianbo Ye, Jeffrey Hu, Matthew Tancik, and Angjoo Kanazawa. gsplat: an open-source library for gaussian splatting. *J. Mach. Learn. Res.*, 26(1), September 2026. ISSN 1532-4435.

Xunpeng Yi, Yong Ma, Yansheng Li, Han Xu, and Jiayi Ma. Artificial intelligence facilitates information fusion for perception in complex environments. *The Innovation*, 6(4):100814, 2025. ISSN 2666-6758. doi: https://doi.org/10.1016/j.xinn.2025.100814. URL https://www.sciencedirect.com/science/article/pii/S2666675825000177.

Weiling Yin, Wenda Zhao, Di You, and Dong Wang. Local binary pattern metric-based multi-focus image fusion. *Optics & Laser Technology*, 110:62–68, 2019. ISSN 0030-3992. doi: https://doi.org/10.1016/j.optlastec.2018.07.045. URL https://www.sciencedirect.com/science/article/pii/S0030399218304596.

Hao Zhang and Jiayi Ma. Sdnet: A versatile squeeze-and-decomposition network for real-time image fusion. *International Journal of Computer Vision*, 129(10):2761–2785, 2021.

Hao Zhang, Zhuliang Le, Zhenfeng Shao, Han Xu, and Jiayi Ma. Mff-gan: An unsupervised generative adversarial network with adaptive and gradient joint constraints for multi-focus image fusion. *Information Fusion*, 66:40–53, 2021a. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2020.08.022. URL https://www.sciencedirect.com/science/article/pii/S1566253520303572.

Qiang Zhang and Martin D. Levine. Robust multi-focus image fusion using multi-task sparse representation and spatial context. *IEEE Transactions on Image Processing*, 25(5):2045–2058, 2016. doi: 10.1109/TIP.2016.2524212.

Qiang Zhang, Tao Shi, Fan Wang, Rick S. Blum, and Jungong Han. Robust sparse representation based multi-focus image fusion with dictionary construction and local spatial consistency. *Pattern Recognition*, 83:299–313, 2018. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2018.06.003. URL https://www.sciencedirect.com/science/article/pii/S0031320318302139.

Qiang Zhang, Fan Wang, Yongjiang Luo, and Jungong Han. Exploring a unified low rank representation for multi-focus image fusion. *Pattern Recognition*, 113:107752, 2021b. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2020.107752. URL https://www.sciencedirect.com/science/article/pii/S0031320320305550.

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, 2020. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2019.07.011. URL https://www.sciencedirect.com/science/article/pii/S1566253518305505.

Zixiang Zhao, Haowen Bai, Yuanzhi Zhu, Jiangshe Zhang, Shuang Xu, Yulun Zhang, Kai Zhang, Deyu Meng, Radu Timofte, and Luc Van Gool. Ddfm: Denoising diffusion model for multi-modality image fusion. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8048–8059, 2023. doi: 10.1109/ICCV51070.2023.00742.

Kecheng Zheng, Juan Cheng, and Yu Liu. Unfolding coupled convolutional sparse representation for multi-focus image fusion. *Information Fusion*, 118:102974, 2025. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2025.102974. URL https://www.sciencedirect.com/science/article/pii/S1566253525000478.

Zhiqiang Zhou, Sun Li, and Bo Wang. Multi-scale weighted gradient-based fusion for multi-focus images. *Information Fusion*, 20:60–72, 2014. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2013.11.005. URL https://www.sciencedirect.com/science/article/pii/S1566253513001462.

Dong Zhu, Xueqian Wang, Gang Li, and Xiao-Ping Zhang. Vessel detection via multi-order saliency-based fuzzy fusion of spaceborne and airborne sar images. *Information Fusion*, 89:473–485, 2023. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2022.08.022. URL https://www.sciencedirect.com/science/article/pii/S1566253522001142.

Wenjie Zhu, Xurong Chi, and Jingrun Chen. Stage-wise multi-focus fusion for numerous misaligned industrial images. *Neurocomputing*, 638:130062, 2025. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2025.130062. URL https://www.sciencedirect.com/science/article/pii/S0925231225007349.