G2Fusion: Geometric-to-Generative Image Fusion via Registration-Restoration Evolution

Hao Zhang[1,2,*] **Douyu Wu**[1,*] **Han Xu**[3], **Linfeng Tang**[4,†] **Qiwen Jin**[5], **Jiayi Ma**[1,4,†]

[1]Electronic Information School, Wuhan University, Wuhan, China
[2]Suzhou Institute of Wuhan University, Suzhou, China
[3]School of Automation, Southeast University, Nanjing, China
[4]School of Robotics, Wuhan University, Wuhan, China
[5]School of Information Engineering, Wuhan University of Technology, Wuhan, China

{zhpersonalbox,linfeng0419,jyma2010}@gmail.com,
wudouyu@whu.edu.cn, xu_han@seu.edu.cn, qiwenjin@whut.edu.cn

Abstract

Real-world multi-modal image fusion faces two fundamental challenges: spatial misalignment and complex degradations induced by heterogeneous imaging sensors. These factors are intrinsically coupled, but most existing fusion methods typically address only one of them, inevitably leading to failure under the other challenge. In this paper, we propose G2Fusion, the first fusion framework that simultaneously addresses image registration and information restoration. By developing a novel geometric-to-generative paradigm, it can directly produce high-quality fused images from unregistered and degraded inputs captured by heterogeneous imaging sensors. This framework consists of two key modules: a flow-based geometric deformation reduction module (Flow-GDR) and a DiT-based generative fusion module (DiT-GF). The former reduces large non-rigid discrepancies through dense flow estimation. The latter performs generative fusion, progressively refining residual misalignments while restoring degraded content through iterative denoising. To enable effective interaction among registration, restoration, and fusion, we design two complementary mechanisms. On the one hand, a target distribution mining strategy is introduced to construct a joint objective distribution from registration, restoration, and fusion, effectively guiding the optimization of DiT-GF. On the other hand, we develop a mutual promotion mechanism that establishes a closed-loop interaction between Flow-GDR and DiT-GF by re-estimating the residual deformation between the fused output and the infrared reference. Extensive experiments demonstrate that G2Fusion consistently outperforms state-of-the-art methods in terms of registration, restoration, and fusion. The code is publicly available at https://github.com/yuwu09/G2Fusion.

1 Introduction

Due to the inherent limitations of imaging physics and sensor characteristics, single-modal images are often insufficient to comprehensively describe complex real-world scenes, thereby restricting their effectiveness in downstream tasks. Consequently, multi-modal image fusion [1, 2] has emerged as a fundamental technique aimed at integrating complementary information from disparate modalities into a unified representation. In particular, infrared-visible image fusion (IVIF) has been widely applied in scenarios [3, 4, 5, 6] such as autonomous driving, intelligent surveillance, and low-light perception, where it enhances both human visual experience and machine perception capabilities.

[*]Equal Contribution
[†]Corresponding author

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

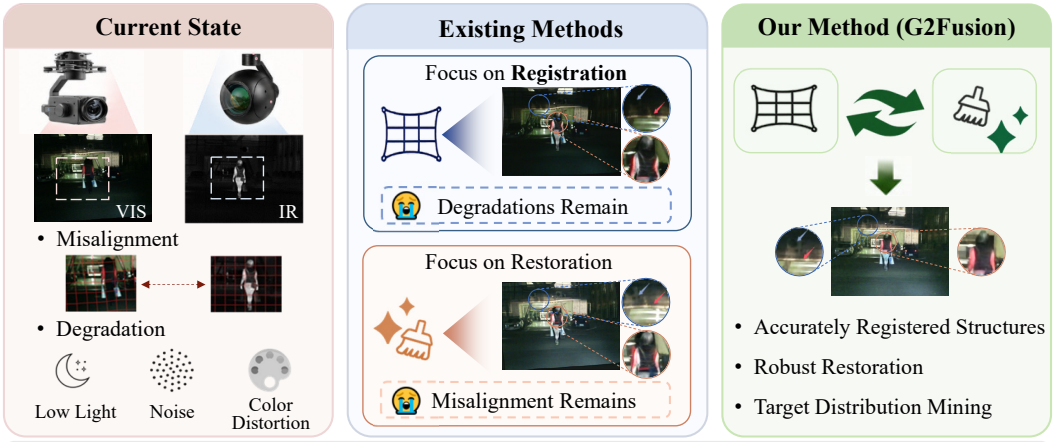

Figure 1: Current state of image fusion and comparison of existing paradigms with G2Fusion.

However, images acquired from heterogeneous sensors in real-world scenarios typically suffer from two tightly coupled issues: spatial misalignment and complex degradations. Misalignment, caused by viewpoint and sensor discrepancies, introduces geometric inconsistencies that hinder reliable cross-modal correspondence. Meanwhile, degradations such as noise, low illumination, and color distortion severely impair image quality, further complicating the fusion process. As a result, spatial misalignment and complex degradations have long been regarded as two major bottlenecks limiting the performance of multi-modal image fusion.

Recent efforts have attempted to alleviate these challenges from two complementary perspectives. On the one hand, joint registration–fusion frameworks, such as RFNet [7], MURF [8], C2RF [9], and MulFS-CAP [10], incorporate cross-task interactions or consistency constraints to improve alignment. On the other hand, restoration–fusion methods like DIVFusion [11], Text-DiFuse [12], and DAFusion [13], suppress degradations and enhance cross-modal representations using explicit priors or powerful generative models. However, joint registration–fusion methods struggle with complex degradations, while restoration–fusion methods fail to handle spatial misalignment. Consequently, both paradigms become fragile under the coexistence of misalignment and degradation, often suffering severe performance degradation or even failure in real-world scenarios, as illustrated in Fig. 1.

More fundamentally, spatial misalignment and complex degradations are intrinsically coupled during the fusion process. Misalignment can introduce artifacts such as ghosting in the fused results, thereby amplifying perceived degradations; conversely, severe degradations may mislead cross-modal correspondence estimation and further compromise registration. Under such coupled conditions, addressing these two factors independently from either a registration-driven or a restoration-driven perspective is inherently insufficient. Therefore, it is crucial to develop a framework in which registration and restoration co-evolve within the image fusion pipeline.

This observation calls for a more expressive modeling paradigm. Fortunately, the Diffusion Transformer (DiT) [14] provides a promising modeling paradigm, whose iterative denoising process and global attention mechanism make it potentially suitable for jointly reasoning about spatial structures and degradations across modalities. Nevertheless, adapting DiT to multi-modal fusion is non-trivial, as it faces two foundational predicaments. First, **the ambiguity of the target distribution**: diffusion models rely on a well-defined clean data distribution, whereas in multi-modal fusion, the desired distribution is inherently ill-defined, as no explicit distribution simultaneously captures accurate registration, faithful fusion, and degradation removal. This renders the learning objective underdetermined and the optimization ill-posed. Second, **the absence of a coordinated evolution mechanism**: registration and restoration are tightly coupled yet governed by distinct principles—geometric transformation and content recovery. Rather than an independent or naive combination, they should co-evolve within the generative trajectory. However, enabling such registration–restoration evolution is challenging; without explicit coordination, errors accumulate as misalignment propagates and degradations persist, leading to suboptimal results. Designing a process that enables progressive and mutually reinforcing evolution between them remains critical.

Motivated by the above observations and to address the two foundational predicaments in adapting DiT to unregistered and degraded multi-modal image fusion, we propose G2Fusion, a geometric-to-generative image fusion framework that bridges explicit geometric registration and generative fusion through a registration-restoration evolution paradigm. Specifically, we first introduce a flow-based geometric deformation reduction module that directly alleviates non-rigid geometric deformations, reducing large-scale discrepancies as an initial registration stage. The remaining misalignments are then incorporated into the DiT, which jointly models the registration, fusion, and restoration processes in the latent space. To resolve the ambiguity of the target distribution, we develop a target distribution mining strategy. This strategy incorporates a two-stage retrieval mechanism to progressively synthesize reliable pseudo-targets from a diverse pool of restoration and fusion candidates. By approximating an explicit target distribution, we ensure that the learning objective simultaneously accounts for registration consistency, fusion fidelity, and restoration quality. To enable coordinated evolution, we propose a mutual promotion mechanism for closed-loop optimization. Based on the insight that a well-fused image should exhibit minimal residual deformation with respect to the infrared reference, we reapply registration to the output to create a feedback loop. This interaction allows fusion quality to provide self-supervision for registration refinement. By enabling mutual reinforcement, this mechanism prevents the error accumulation inherent in cascaded pipelines and ensures global structural coherence throughout the generative process. As a result, G2Fusion can directly transform unregistered and degraded multi-modal inputs into geometrically coherent, visually restored, and information-rich fused images.

In summary, our contributions are as follows:

- We propose G2Fusion, a geometric-to-generative image fusion framework that integrates geometric registration and generative fusion through a registration-restoration evolution paradigm. To our knowledge, this is the first image fusion model to combine registration and restoration tasks, enabling robust fusion directly from raw sensor inputs.
- We introduce a target distribution mining strategy to progressively approximate a reliable distribution from diverse candidates. This distribution-driven learning provides stable optimization, jointly constraining registration accuracy, fusion quality, and restoration fidelity.
- A mutual promotion mechanism is developed for closed-loop refinement. By enforcing the fused output to inherit the infrared reference structure and feeding it back into registration, the framework enables progressive evolution, reducing error accumulation and improving structural coherence.

2 Methodology

2.1 Problem Statement and Modeling

We formally define the problem of this work as follows: achieving infrared-visible image fusion directly from raw inputs. Given a warped visible image I_{vi}^{warp} and a reference infrared image I_{ir}, our goal is to address both geometric misalignments and complex degradations through a collaborative evolution process to generate a high-fidelity fused image I_f. As illustrated in Fig. 2, G2Fusion integrates geometric deformation reduction and generative fusion through mutual promotion.

Concretely, we begin by estimating a dense deformation field that aligns the visible image with the infrared reference. This is done by calculating the deformation field through a flow-based module: $I_{vi}^{reg} = \mathcal{W}(I_{vi}^{warp}, \mathcal{G}(I_{vi}^{warp}, I_{ir}))$, where $\mathcal{G}(I_{vi}^{warp}, I_{ir})$ represents the pixel-wise displacement field, and $\mathcal{W}(\cdot)$ represents a differentiable warping operator implemented via bilinear sampling. Based on this deformation, the registered visible image is obtained through a differentiable spatial transformation, alleviating large-scale non-rigid discrepancies.

We further model fusion as a conditional generative process. Specifically, we adopt a DiT-based generative fusion module $\mathcal{F}$ to perform modality integration in a shared latent space Φ: $I_f = \mathcal{F}(\Phi_t \mid I_{vi}^{reg}, I_{ir}, t)$, where Φ_t denotes the noisy latent variable at timestep t. In addition, we design a target distribution mining strategy to provide pseudo-targets for training $\mathcal{F}$ based on both source fidelity and intrinsic fusion quality.

Importantly, instead of treating registration and fusion as independent stages, we couple them through a mutual promotion mechanism. The fused image I_f acts as a structural prior to guide geometric registration. The matcher and fusion networks are alternately optimized as: $\theta_{\mathcal{G}}, \theta_{\mathcal{F}} \leftarrow \arg\min \mathcal{L}_{total}(\theta_{\mathcal{G}}, \theta_{\mathcal{F}})$, where $\mathcal{L}_{total}$ combines registration and fusion losses. This closed-loop

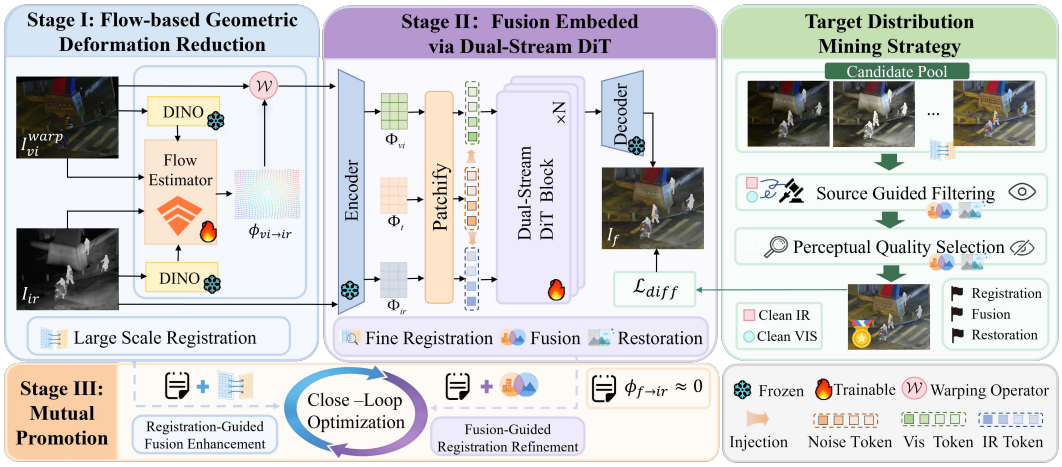

Figure 2: The overall framework of our G2Fusion.

optimization enables registration to enhance fusion, while the fused output further refines registration, driving both modules toward a coherent, high-quality representation.

2.2 Flow-based Geometric Deformation Reduction

Given a warped visible image I_{vi}^{warp} and an infrared image I_{ir}, we estimate a dense deformation field $\phi_{vi\to ir}$ to produce the aligned image I_{vi}^{reg}. To handle severe cross-modal discrepancies, we adopt a hybrid feature extraction scheme that combines global semantic priors with local structural details. Specifically, a DINOv3 [15] encoder $\mathcal{D}(\cdot)$ extracts modality-invariant features at the largest scale L, with its last transformer blocks fine-tuned on infrared-visible data to enhance cross-modal consistency. For finer scales $l < L$, a pyramid extractor $\mathcal{P}(\cdot)$ captures multi-scale structural information, forming hierarchical features as follows:

$$f^l = \begin{cases} \mathcal{D}(I), & l = L, \\ \mathcal{P}^l(I), & l < L. \end{cases} \tag{1}$$

The deformation field is estimated in a hierarchical, multi-scale manner. At the lowest resolution, an initial flow $\phi_{vi\to ir}^L = \mathcal{M}^L(f_{vi}^L, f_{ir}^L)$ is predicted and then refined at higher resolutions via

$$\phi_{vi\to ir}^l = \uparrow(\phi_{vi\to ir}^{l+1}) + \mathcal{M}^l\big(\mathcal{W}(f_{vi}^l, \uparrow(\phi_{vi\to ir}^{l+1})), f_{ir}^l\big), \tag{2}$$

where $\uparrow(\cdot)$ denotes upsampling. Each motion module $\mathcal{M}^l$ models local discrepancies instead of standard correlations for robustness to cross-modal intensity variations.

During training, the predicted deformation is directly supervised at both flow and image levels to ensure geometric accuracy and structural consistency. Specifically, a flow constraint is imposed:

$$\mathcal{L}_{flow} = \frac{1}{N}\sum_i w_i \left\|\phi_{vi\to ir}(i) - \phi^{gt}(i)\right\|_2^2, \tag{3}$$

where w_i emphasizes structural regions via gradient-based weighting. Meanwhile, the warped image I_{vi}^{reg} is regularized to preserve appearance consistency with the original visible image:

$$\mathcal{L}_{image} = \lambda_1\left(\|I_{vi}^{reg} - I_{vi}\|_1 + \|I_{vi}^{reg} - I_{vi}\|_2^2\right) + \lambda_2\mathcal{L}_{grad}(I_{vi}^{reg}, I_{vi}), \tag{4}$$

where $\mathcal{L}_{grad}$ enforces gradient loss. The overall objective $\mathcal{L}_{reg} = \mathcal{L}_{flow} + \mathcal{L}_{image}$ jointly guides the model to produce accurate and structurally coherent registration.

2.3 DiT-based Generative Fusion

2.3.1 Fusion Embedded via Dual-Stream DiT

To mitigate unpredictable real-world degradations Ω, we shift the fusion process from the unreliable pixel domain to a robust shared latent space. Specifically, the registered visible image I_{vi}^{reg} and infrared image I_{ir} are first encoded into latent representations via a shared deterministic encoder $E(\cdot)$,

yielding $\Phi_{vi} = E(I_{vi}^{reg} \,|\, \Omega)$ and $\Phi_{ir} = E(I_{ir} \,|\, \Omega)$, where a deterministic autoencoder is adopted to avoid the ambiguity and blurriness introduced by stochastic latent sampling.

During training, the pseudo-target I^* mined in Sec. 2.3.2 is encoded as Φ_0. This latent serves as the clean sample for conditional diffusion, with Φ_{ir} and Φ_{vi} providing the modality-specific conditions. The forward process is defined by Gaussian perturbation:

$$\Phi_t = \sqrt{\bar{\alpha}_t}\,\Phi_0 + \sqrt{1-\bar{\alpha}_t}\,\epsilon, \qquad \epsilon \sim \mathcal{N}(0, \mathbf{I}), \tag{5}$$

where Φ_t denotes the noisy latent at timestep t. The reverse denoising process is parameterized by DiT: $\epsilon_\theta = \mathcal{F}_{\mathrm{DiT}}(\Phi_t, t, \Phi_{ir}, \Phi_{vi})$, and optimized using the standard noise prediction objective:

$$\mathcal{L}_{diff} = \mathbb{E}\big[\|\epsilon - \epsilon_\theta\|_2^2\big]. \tag{6}$$

The noisy latent and modality-specific latents are converted into token sequences via patch embedding, i.e., $x_t^0 = \mathrm{PE}(\Phi_t)$, $x_{ir}^0 = \mathrm{PE}(\Phi_{ir})$, and $x_{vi}^0 = \mathrm{PE}(\Phi_{vi})$. To explicitly condition denoising on multi-modal observations, noisy tokens are injected into both streams through learnable projections: $\tilde{x}_{ir}^0 = x_{ir}^0 + \mathbf{W}_{ir} x_t^0$ and $\tilde{x}_{vi}^0 = x_{vi}^0 + \mathbf{W}_{vi} x_t^0$.

Each dual-stream block performs bidirectional cross-attention, enabling infrared and visible streams to exchange complementary structural cues and compensate for spatial misalignment in feature space. To preserve multi-scale details under registration uncertainty, we further adopt a U-shaped hierarchy [16] with a quad-interaction skip-merge module:

$$x_{merge} = \mathrm{MLP}\big([x_{up}, x_{skip}, x_{up} - x_{skip}, x_{up} \odot x_{skip}]\big), \tag{7}$$

where the difference term captures local spatial discrepancies, while the Hadamard product enhances feature co-occurrence consistency.

A spatially adaptive gating module dynamically integrates both streams:

$$g = \sigma(\mathrm{MLP}([x_{ir}, x_{vi}])), \qquad x_f = g \odot x_{ir} + (1-g) \odot x_{vi}, \tag{8}$$

yielding the final fused latent $\hat{\Phi}_0$, which is finally decoded into the fused image via $I_f = D(\hat{\Phi}_0)$.

2.3.2 Target Distribution Mining for Registration-Restoration-Fusion

To address the ambiguity of the target distribution, we propose a target distribution mining strategy that replaces explicit supervision with an empirical fusion prior derived from state-of-the-art fusion and restoration outputs. Since existing multi-modal fusion datasets are typically pre-aligned, the candidate pool $\mathcal{C} = \{I^{(1)}, \ldots, I^{(K)}\}$ provides structurally coherent fusion references and serves as an empirical approximation of the aligned fusion manifold. All candidate methods use the same aligned infrared-visible pair at a common resolution, without candidate-specific registration or additional degradation removal. For each pair, we then min–max normalize each metric m across the K resulting candidates.

We adopt a two-stage selection strategy to identify the optimal pseudo-target. In the first stage, the source fidelity score $\mathcal{S}_1(\cdot)$ filters candidates based on how well they preserve information from both source modalities. This reference-based assessment ensures that the selected subset preserves the essential features of the clean source images:

$$\mathcal{C}_{top} = \mathrm{TopN}_{I \in \mathcal{C}}\ \mathcal{S}_1(I). \tag{9}$$

We set $N = 5$ in our implementation. In the second stage, we introduce an unsupervised perceptual score $\mathcal{S}_2(\cdot)$ to evaluate the intrinsic quality of the candidates. By utilizing this no-reference metric, the final pseudo-target I^* is determined by maximizing the visual fidelity within the filtered subset:

$$I^* = \arg\max_{I \in \mathcal{C}_{top}} \mathcal{S}_2(I). \tag{10}$$

From a probabilistic perspective, this targeted search process can be viewed as an informed sampling from the empirical fusion manifold. The two-stage selection effectively prunes this distribution to a high-quality subset $\mathcal{P}_{ts}$:

$$\mathcal{P}_{ts} = \{I \in \mathcal{C} \mid I \in \mathrm{TopN}_{J \in \mathcal{C}}\, \mathcal{S}_1(J),\ \mathcal{S}_2(I) \geq \tau\}, \tag{11}$$

where τ controls the truncation level of the empirical prior based on perceptual quality. The final pseudo-target I^* can thus be interpreted as the optimal sample drawn from $\mathcal{P}_{ts}$. Beyond providing a reliable distribution, the selected target serves an additional role by offering a complete and visually coherent image for diffusion training, including regions partially missing due to registration-induced misalignment. This enables the diffusion model to learn implicit completion of incomplete regions, further improving structural continuity and robustness in real-world fusion scenarios.

2.4 Mutual Promotion Mechanism

While the registration and fusion modules achieve their respective goals, their cascade connection lacks end-to-end feedback. To address this, we propose a mutual promotion mechanism, based on the observation that once the visible image is properly aligned to the infrared image and the fusion model integrates both modalities, the fused image should inherit the geometry of the infrared target. Therefore, matching the fused image to the infrared image should ideally produce a near-zero deformation field. To enforce this geometric consistency, we apply the registration module again between the fused image and the infrared image:

$$\phi_{f\to ir} = \mathcal{G}(I_f, I_{ir}), \tag{12}$$

where $\phi_{f\to ir}$ denotes the deformation from the fused image to the infrared image, and ideally, this deformation should vanish, meaning that $\phi_{f\to ir} \approx 0$.

Based on this observation, we introduce a fusion-induced registration loss:

$$\mathcal{L}_{mp} = \mathcal{L}_{flow}(I_f, I_{ir}, \phi_{f\to ir}, 0), \tag{13}$$

which encourages the fused image to be geometrically consistent with the infrared reference.

During optimization, the two branches are updated alternately. The registration branch is optimized with both the registration loss and the mutual-promotion term, where $\mathcal{L}_{match} = \mathcal{L}_{reg} + \lambda_m \mathcal{L}_{mp}$. The fusion branch is optimized as $\mathcal{L}_{fusion} = \mathcal{L}_{diff} + \lambda_f \mathcal{L}_{mp}$. These formulations establish a closed loop between registration and fusion. Improved registration provides better-aligned inputs for fusion, while improved fusion produces geometrically consistent outputs that serve as an additional self-supervisory signal for registration. As a result, the two branches promote each other during training.

3 Experiments

3.1 Experimental Configurations

We construct a mixed training set of 2,500 image pairs from MSRS [17], M^{3}FD [18], LLVIP [19], and RoadScene [20], covering diverse scenes and degradations. The candidate pool includes ControlFusion [21], Text-IF [22], TG-ECNet [23], DAFusion [13], URFusion [24], OmniFuse [25], SDSFusion [26], DIVFusion [11], Text-DiFuse [12], and DCEvo [27]. The two-stage mining strategy ranks candidates first with SSIM [28], PSNR [29], and VIF [30], and then refines the selection using SD [31], SF [32], AG [33], and EN [34] to obtain the final pseudo-targets. The Flow-GDR and the DiT-GF are first trained separately, and both are optimized using AdamW with a learning rate of 5×10^{-5} and cosine scheduling. For the mutual promotion mechanism, a reduced learning rate of 5×10^{-7} is used with mutual and registration losses weighted by $\lambda_m = 0.01$, $\lambda_f = 0.005$.

To simulate realistic misalignment, visible images are randomly rotated within $[-5°, 5°]$, translated by $\pm 5\%$ of image size, and subjected to elastic deformations via smoothed Gaussian noise. We evaluate the trained model separately on the four datasets and compare it against three categories of methods: (1) fusion-only methods: CrossFuse [35], SHIP [36], DCEvo [27]; (2) registration-based: C2RF [9], SuperFusion [37], CPMFusion [38]; (3) restoration-based: ControlFusion [21], Text-IF [22], TG-ECNet [23]. Fusion-only methods are augmented with registration and restoration via MINIMA [39] and InstructIR [40] for fairness, while registration-based and restoration-based methods receive missing components. In contrast, G2Fusion operates on raw inputs without preprocessing. Performance is evaluated using MI [41], EN [34], CC [42], SCD [43], VIF [30], PSNR [29], and SSIM [28]. All experiments are conducted on NVIDIA TITAN RTX GPUs and an AMD EPYC 7H12 64-core CPU.

3.2 Comparative Experiments

3.2.1 Fusion Evaluation

As shown in Fig. 3, G2Fusion produces more accurate registration, clearer details, and more coherent fusion results, particularly in challenging regions with severe distortions and structural inconsistencies. The fused images preserve salient thermal targets while maintaining richer visible textures, demonstrating superior visual quality in complex scenes. As reported in Table 1, G2Fusion achieves



Figure 3: Qualitative comparison results on MSRS and M^3FD datasets. Magenta and green edges denote gradients from the infrared reference and the visible/fused results, respectively, for registration visualization. "*" and "†" indicate that preprocessing-based restoration enhancement and registration enhancement are applied to the corresponding methods for comparison, respectively.

leading performance on most metrics across the four datasets, showing strong robustness and generalization. This is mainly because existing methods typically optimize registration, restoration, and fusion separately, causing objective inconsistency and error accumulation. In contrast, G2Fusion enables registration, degradation removal, and fusion to mutually enhance each other.

3.2.2 Registration Evaluation

To quantify registration accuracy throughout G2Fusion, we manually annotate five corresponding landmarks in each of 100 scenes, yielding 500 correspondences, as shown in Fig. 4. Using the infrared image as the geometric reference, we evaluate registration errors before registration, after Flow-GDR, and in the final fused output. Registration accuracy is measured using root mean square error (RMSE), maximum alignment error (MaxAE), and median Euclidean error (MEE), computed from the Euclidean distances between corresponding landmarks. As reported in Table 2, Flow-GDR substantially reduces all three registration error metrics, while the final fused output achieves further reductions, particularly in MEE. These results support the complementary roles of Flow-GDR in explicit geometric correction and DiT-GF in implicitly refining residual registration errors during generative fusion.

Table 1: Quantitative comparison on MSRS, M^{3}FD, LLVIP, and RoadScene datasets.

Type	Methods	MSRS							M^{3}FD						
		MI↑	EN↑	CC↑	SCD↑	VIF↑	PSNR↑	SSIM↑	MI↑	EN↑	CC↑	SCD↑	VIF↑	PSNR↑	SSIM↑
/	CrossFuse	1.863	7.172	0.413	0.645	0.310	58.430	0.596	2.460	6.648	0.307	0.482	0.405	59.085	0.645
/	DCEvo	2.045	**7.464**	0.477	0.950	0.379	57.863	0.623	2.768	6.904	0.449	0.950	0.540	59.537	0.687
/	SHIP	1.978	7.405	0.475	0.930	0.366	57.757	0.617	2.665	6.873	0.423	0.827	0.510	59.567	0.669
Reg.	C2RF	1.770	7.349	0.482	0.982	0.335	58.609	0.612	2.846	6.874	0.499	1.079	0.560	60.997	0.718
Reg.	CPMFusion	1.897	7.371	0.462	0.907	0.278	58.098	0.557	3.215	6.760	0.467	0.946	0.439	60.496	0.649
Reg.	SuperFusion	2.213	7.374	0.503	1.023	0.426	58.312	0.658	2.909	6.733	0.488	0.995	0.512	60.518	0.696
De.	ControlFusion	2.173	7.166	0.519	1.084	0.448	59.999	0.652	2.919	6.949	0.468	0.976	0.638	60.593	0.711
De.	TG-ECNet	1.794	6.905	0.485	0.910	0.328	60.176	0.603	2.317	6.949	0.420	0.764	0.507	61.685	0.589
De.	Text-IF	2.057	7.171	0.499	1.001	0.429	58.747	0.675	2.762	6.742	0.432	0.864	0.643	60.693	0.746
De.&Reg.	Ours	**2.218**	7.173	**0.543**	**1.178**	**0.481**	**60.493**	**0.676**	**3.227**	**6.956**	**0.539**	**1.263**	**0.748**	**63.390**	**0.747**

Type	Methods	LLVIP							RoadScene						
		MI↑	EN↑	CC↑	SCD↑	VIF↑	PSNR↑	SSIM↑	MI↑	EN↑	CC↑	SCD↑	VIF↑	PSNR↑	SSIM↑
/	CrossFuse	1.335	7.040	0.393	0.446	0.338	61.159	0.486	2.490	6.579	0.355	0.295	0.395	61.110	0.615
/	DCEvo	1.625	**7.467**	0.509	0.932	0.430	60.494	0.557	2.592	6.974	0.544	0.865	0.550	59.930	0.780
/	SHIP	1.527	7.407	0.488	0.842	0.404	60.439	0.517	2.614	6.939	0.514	0.710	0.530	60.149	0.749
Reg.	C2RF	1.682	7.385	0.529	0.994	0.389	61.527	0.556	3.048	7.146	0.634	1.200	0.722	62.168	**0.884**
Reg.	CPMFusion	2.074	7.319	0.483	0.826	0.302	60.797	0.470	**3.664**	7.046	0.567	0.926	0.469	61.677	0.633
Reg.	SuperFusion	1.671	7.316	0.521	0.932	0.381	61.103	0.532	2.866	6.962	0.604	1.009	0.538	61.576	0.742
De.	ControlFusion	2.114	7.321	0.554	1.053	**0.542**	63.217	0.599	3.165	7.296	0.595	1.128	0.740	61.982	0.831
De.	TG-ECNet	1.787	7.094	0.510	0.827	0.432	62.928	0.526	2.732	7.077	0.547	0.900	0.576	62.193	0.656
De.	Text-IF	1.613	7.298	0.521	0.898	0.488	62.255	0.599	2.920	7.212	0.576	1.035	0.765	62.422	0.857
De.&Reg.	Ours	**2.115**	7.104	**0.577**	**1.054**	0.528	**64.000**	**0.599**	3.167	**7.315**	**0.646**	**1.362**	**0.768**	**63.864**	0.781



Figure 4: Visualization of point landmark correspondences.

Table 2: Registration evaluation using point landmarks.

Evaluation stage	RMSE ↓	MaxAE ↓	MEE ↓
Unregistered	13.183	19.258	11.083
After Flow-GDR	6.406	10.640	4.074
Fused output	**5.867**	**10.505**	**2.790**

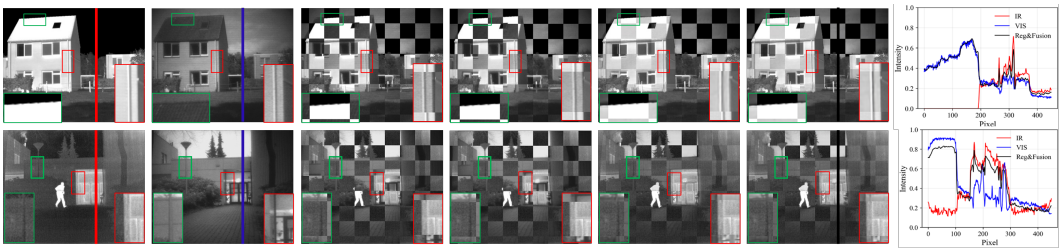

Figure 5: Real-world generalization results on the TNO dataset.

3.3 Real-World Generalization

We evaluate G2Fusion's generalization under real-world misalignment on the TNO [44] dataset. As shown in Fig. 5, despite being trained solely on simulated misalignments, G2Fusion effectively corrects the noticeable deformations and artifacts in real-world image pairs, as evidenced by the checkerboard visualizations, intensity profiles, and fused results. While DiT-GF can handle mild misalignments, its performance degrades under larger deformations. In contrast, the full G2Fusion achieves more accurate alignment and cleaner fusion results, demonstrating stronger robustness in real-world scenarios.

3.4 High-Level Semantic Tasks Verification

To further evaluate the effectiveness of our fusion method on downstream tasks, we conduct experiments on object detection and semantic segmentation. For object detection, YOLOv5 [3] is adopted as the baseline, while SegNeXt [45] is used as the baseline for semantic segmentation. For each fusion

[3] https://github.com/ultralytics/yolov5

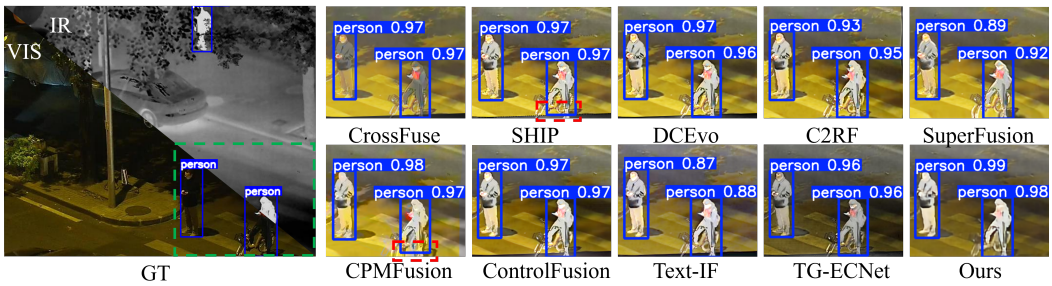

Figure 6: Visual verification on object detection task.

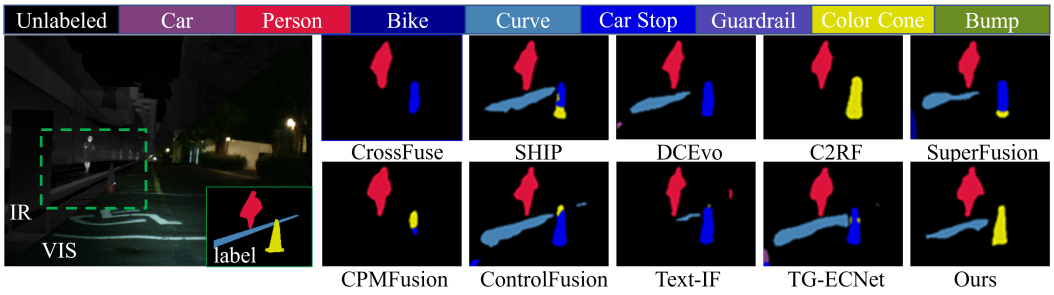

Figure 7: Visual verification on the semantic segmentation task.

Table 3: Quantitative comparison on object detection and semantic segmentation.

	Detection					Segmentation								
	Prec.	Rec.	AP@.5	AP@.75	AP@.5:.95	Car	Person	Bike	Curve	Car Stop	Guar.	Cone	Bump	mIoU
CrossFuse	0.909	0.799	0.879	0.460	0.494	0.874	0.667	0.629	0.351	0.430	0.065	0.459	0.455	0.545
SHIP	0.949	0.848	0.915	0.597	0.565	0.870	0.714	0.619	0.415	0.397	0.055	0.468	0.406	0.547
DCEvo	0.966	0.837	0.920	0.569	0.553	0.877	0.691	0.612	0.424	0.443	0.026	0.460	0.434	0.550
C2RF	0.936	0.830	0.914	0.595	0.549	0.854	0.692	0.613	0.353	0.408	0.113	0.442	0.487	0.549
SuperFusion	0.920	0.845	0.916	0.587	0.551	0.876	0.702	0.614	0.390	0.445	0.102	0.467	0.411	0.554
CPMFusion	0.907	0.851	0.904	0.585	0.549	0.838	0.653	0.583	0.329	0.414	0.001	0.433	0.353	0.509
ControlFusion	0.916	0.873	0.914	0.617	0.572	0.875	0.717	0.627	0.472	0.349	0.000	0.492	0.536	0.560
Text-IF	0.930	0.920	0.956	0.586	0.556	0.882	0.714	0.603	0.405	0.322	0.076	0.483	0.515	0.553
TG-ECNet	0.915	0.854	0.906	0.589	0.545	0.877	0.714	0.613	0.441	0.371	0.079	0.495	0.429	0.555
Ours	0.921	0.865	0.916	0.618	**0.574**	0.865	0.694	0.597	0.433	0.389	0.069	0.453	0.581	**0.562**

method, we generate its fused training and test sets using the same data splits, and train separate YOLOv5 and SegNeXt models with identical architectures and training settings within each task. The quantitative results are reported in Table 3, where G2Fusion achieves the highest performance in terms of AP@0.5:0.95 and mIoU. As shown in Fig. 6, in object detection, some existing methods suffer from degraded accuracy due to black borders introduced by registration (highlighted by red dashed boxes), which interfere with accurate localization. In contrast, G2Fusion can reconstruct complete and well-aligned images, leading to the highest average detection accuracy. Similarly, in semantic segmentation shown in Fig. 7, our approach consistently outperforms competing methods, demonstrating superior structural consistency and semantic preservation.

3.5 Ablation Study

To verify the effectiveness of each component, we conduct detailed ablation studies involving eight variants: **I**: replacing flow estimation with affine transformation estimation; **II**: removing DINO fine-tuning; **III**: removing the Flow-GDR; **IV**: replacing target distribution mining with random selection from the candidate pool; **V**: merging the two-stage target distribution mining strategy into a single-stage scheme; **VI**: removing the dual-stream mechanism by directly concatenating infrared and visible inputs; **VII**: omitting the entire mutual promotion optimization stage; and **VIII**: our full model. As shown in Fig. 8, the first row presents the registration-only results and the second row shows the fusion results, where removing any component noticeably affects registration or fusion quality. Specifically, variant **I** is limited in modeling non-rigid deformation, while variants **II** and **III** reduce cross-modal matching accuracy. Variants **IV** and **V** introduce less reliable pseudo-targets, leading to degraded fusion results. Variant **VI** weakens modality-specific interaction, and variant

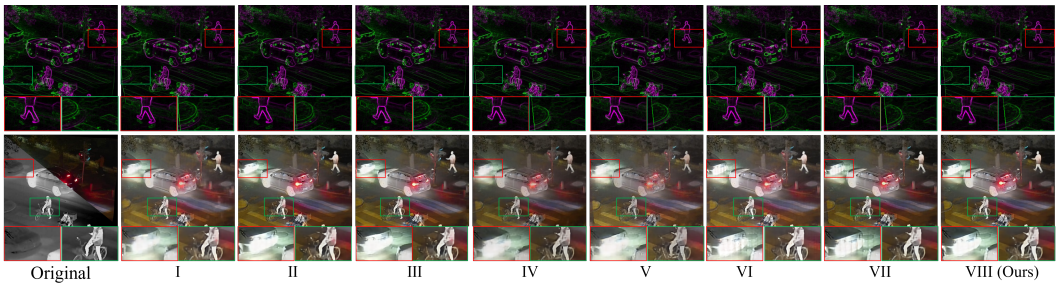

Figure 8: Visual results of ablation studies.

Table 4: Quantitative results of ablation studies.

	Registration							Fusion						
	LNCC↑	MI↑	CC↑	SSIM↑	MSE↓	MEE↓	MAE↓	MI↑	EN↑	CC↑	SCD↑	VIF↑	PSNR↑	SSIM↑
I	0.199	1.507	0.771	0.450	0.039	0.123	0.150	2.124	6.576	0.561	1.207	0.499	63.190	0.709
II	0.206	1.513	0.758	0.449	0.040	0.124	0.151	2.164	6.656	0.552	1.151	0.484	62.931	0.698
III	0.072	1.110	0.623	0.379	0.047	0.145	0.166	1.913	6.648	0.517	1.031	0.407	62.737	0.658
IV	0.317	1.760	0.769	0.477	0.041	0.116	0.151	1.981	6.080	0.569	1.149	0.466	63.740	0.723
V	0.319	1.765	0.772	0.475	0.040	0.116	0.151	1.820	6.434	**0.582**	1.178	0.500	62.682	0.698
VI	0.298	1.738	**0.788**	0.475	0.039	0.116	0.149	1.953	**6.768**	0.548	1.126	0.497	62.842	0.713
VII	0.305	1.745	0.772	0.473	0.042	0.117	0.150	2.248	6.616	0.571	1.198	0.537	63.173	0.719
Ours	**0.319**	**1.770**	0.775	**0.485**	**0.038**	**0.115**	**0.130**	**2.308**	6.651	0.572	**1.223**	**0.548**	**63.804**	**0.732**

Table 5: Model complexity and inference efficiency of G2Fusion.

Module	Params (M)	FLOPs (G)	Runtime (s)
Flow-GDR	88.792	503.921	0.106
DiT-GF	688.054	8916.685	1.958
Full Model	776.847	9420.606	2.064

VII reduces the closed-loop promotion between registration and fusion. For the registration-related evaluation, normalized mean squared error (MSE), normalized mean absolute error (MAE), and mean edge magnitude error (MEE) are computed between the output and the visible image. The quantitative results in Table 4 further confirm the effectiveness of each component in improving overall performance.

3.6 Complexity and Efficiency

Table 5 reports the complexity and inference efficiency of G2Fusion on 640×480 images. DiT-GF dominates the computational cost due to iterative latent denoising, while Flow-GDR adds limited overhead. Peak GPU memory usage is approximately 7.1 GB. Future acceleration can focus on fewer denoising steps and more efficient DiT architectures.

4 Conclusion

In this paper, we propose G2Fusion, a geometric-to-generative framework for robust infrared-visible image fusion under real-world misalignment and complex degradations. By coupling Flow-GDR with DiT-GF, G2Fusion reduces large non-rigid discrepancies and progressively models residual alignment, degradation suppression, and cross-modal fusion. A target distribution mining strategy provides reliable pseudo-targets to alleviate fusion-target ambiguity, while a mutual promotion mechanism builds a closed-loop interaction between registration and fusion via residual deformation estimation. Extensive experiments demonstrate that G2Fusion consistently outperforms state-of-the-art methods across diverse datasets and real-world scenarios.

Acknowledgments and Disclosure of Funding

This work was supported by the NSFC (62506268, 62676289), the Basic Research Program of Jiangsu (BK20250454), and the CPSF (GZB20250066, 2025M78150,2026T190426). The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University. The authors declare no competing interests.

References

[1] Hao Zhang, Han Xu, Xin Tian, Junjun Jiang, and Jiayi Ma. Image fusion meets deep learning: A survey and perspective. *Information Fusion*, 76:323–336, 2021.

[2] Rongchao Wang, Zhaofa Zhou, Shuhui Li, and Zhili Zhang. Advances and challenges in infrared-visible image fusion: A comprehensive review of techniques and applications. *Artificial Intelligence Review*, 59(1):18, 2026.

[3] Xingchen Zhang, Ping Ye, Henry Leung, Ke Gong, and Gang Xiao. Object fusion tracking based on visible and infrared images: A comprehensive review. *Information Fusion*, 63:166–187, 2020.

[4] Jinyuan Liu, Guanyao Wu, Zhu Liu, Di Wang, Zhiying Jiang, Long Ma, Wei Zhong, and Xin Fan. Infrared and visible image fusion: From data compatibility to task adaption. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):2349–2369, 2024.

[5] Wenda Zhao, Shigeng Xie, Fan Zhao, You He, and Huchuan Lu. Metafusion: Infrared and visible image fusion via meta-feature embedding from object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13955–13965, 2023.

[6] Shihua Zhang, Qiuhong Shen, Shizun Wang, Tianbo Pan, and Xinchao Wang. Make geometry matter for spatial reasoning. In *Proceedings of the European Conference on Computer Vision*, pages 231–249, 2026.

[7] Han Xu, Jiayi Ma, Jiteng Yuan, Zhuliang Le, and Wei Liu. Rfnet: Unsupervised network for mutually reinforcing multi-modal image registration and fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19647–19656, 2022.

[8] Han Xu, Jiteng Yuan, and Jiayi Ma. Murf: Mutually reinforcing multi-modal image registration and fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12148–12166, 2023.

[9] Linfeng Tang, Qinglong Yan, Xinyu Xiang, Leyuan Fang, and Jiayi Ma. C2rf: Bridging multi-modal image registration and fusion via commonality mining and contrastive learning. *International Journal of Computer Vision*, 133(8):5262–5280, 2025.

[10] Huafeng Li, Zengyi Yang, Yafei Zhang, Wei Jia, Zhengtao Yu, and Yu Liu. Mulfs-cap: Multimodal fusion-supervised cross-modality alignment perception for unregistered infrared-visible image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3673–3690, 2025.

[11] Linfeng Tang, Xinyu Xiang, Hao Zhang, Meiqi Gong, and Jiayi Ma. Divfusion: Darkness-free infrared and visible image fusion. *Information Fusion*, 91:477–493, 2023.

[12] Hao Zhang, Lei Cao, and Jiayi Ma. Text-difuse: An interactive multi-modal image fusion framework based on text-modulated diffusion model. *Advances in Neural Information Processing Systems*, 37:39552–39572, 2024.

[13] Xue Wang, Zheng Guan, Wenhua Qian, Jinde Cao, Runzhuo Ma, and Cong Bi. A degradation-aware guided fusion network for infrared and visible image. *Information Fusion*, 118:102931, 2025.

[14] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4172–4182, 2023.

[15] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

[16] Yuchuan Tian, Zhijun Tu, Hanting Chen, Jie Hu, Chao Xu, and Yunhe Wang. U-dits: Downsample tokens in u-shaped diffusion transformers. *Advances in Neural Information Processing Systems*, pages 51994–52013, 2024.

[17] Linfeng Tang, Jiteng Yuan, Hao Zhang, Xingyu Jiang, and Jiayi Ma. Piafusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*, 83-84:79–92, 2022.

[18] Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5802–5811, 2022.

[19] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3496–3504, 2021.

[20] Han Xu, Jiayi Ma, Zhuliang Le, Junjun Jiang, and Xiaojie Guo. Fusiondn: A unified densely connected network for image fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12484–12491, 2020.

[21] Linfeng Tang, Yeda Wang, Zhanchuan Cai, Junjun Jiang, and Jiayi Ma. Controlfusion: A controllable image fusion network with language-vision degradation prompts. *Advances in Neural Information Processing Systems*, pages 1–13, 2025.

[22] Xunpeng Yi, Han Xu, Hao Zhang, Linfeng Tang, and Jiayi Ma. Text-if: Leveraging semantic text guidance for degradation-aware and interactive image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27026–27035, 2024.

[23] Yiming Sun, Xin Li, Pengfei Zhu, Qinghua Hu, Dongwei Ren, Huiying Xu, and Xinzhong Zhu. Task-gated multi-expert collaboration network for degraded multi-modal image fusion. In *Proceedings of the International Conference on Machine Learning*, pages 57571–57586, 2025.

[24] Han Xu, Xunpeng Yi, Chen Lu, Guangcan Liu, and Jiayi Ma. Urfusion: Unsupervised unified degradation-robust image fusion network. *IEEE Transactions on Image Processing*, 34:5803–5818, 2025.

[25] Hao Zhang, Lei Cao, Xuhui Zuo, Zhenfeng Shao, and Jiayi Ma. Omnifuse: Composite degradation-robust image fusion with language-driven semantics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(9):7577–7595, 2025.

[26] Jun Chen, Liling Yang, Wei Yu, Wenping Gong, Zhanchuan Cai, and Jiayi Ma. Sdsfusion: A semantic-aware infrared and visible image fusion network for degraded scenes. *IEEE Transactions on Image Processing*, 34:3139–3153, 2025.

[27] Jinyuan Liu, Bowei Zhang, Qingyun Mei, Xingyuan Li, Yang Zou, Zhiying Jiang, Long Ma, Risheng Liu, and Xin Fan. Dcevo: Discriminative cross-dimensional evolutionary learning for infrared and visible image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2226–2235, 2025.

[28] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[29] Quan Huynh-Thu and Mohammed Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics Letters*, 44(13):800–801, 2008.

[30] 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, 2013.

[31] Yun-Jiang Rao. In-fibre bragg grating sensors. *Measurement Science and Technology*, 8(4):355–375, 1997.

[32] Ahmet M Eskicioglu and Paul S Fisher. Image quality measures and their performance. *IEEE Transactions on Communications*, 43(12):2959–2965, 1995.

[33] Guangmang Cui, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition. *Optics Communications*, 341:199–209, 2015.

[34] J Wesley Roberts, Jan A Van Aardt, and Fethi Babikker Ahmed. Assessment of image fusion procedures using entropy, image quality, and multispectral classification. *Journal of Applied Remote Sensing*, 2(1):023522, 2008.

[35] Hui Li and Xiao-Jun Wu. Crossfuse: A novel cross attention mechanism based infrared and visible image fusion approach. *Information Fusion*, 103:102147, 2024.

[36] Naishan Zheng, Man Zhou, Jie Huang, Junming Hou, Haoying Li, Yuan Xu, and Feng Zhao. Probing synergistic high-order interaction in infrared and visible image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26384–26395, 2024.

[37] Linfeng Tang, Yuxin Deng, Yong Ma, Jun Huang, and Jiayi Ma. Superfusion: A versatile image registration and fusion network with semantic awareness. *IEEE/CAA Journal of Automatica Sinica*, 9(12):2121–2137, 2022.

[38] Quanquan Xiao, Haiyan Jin, Haonan Su, Yuanlin Zhang, and Guanlei Sheng. Cpmfusion: A method for infrared and visible image registration and fusion using modal correlation. *Information Fusion*, 127:103924, 2026.

[39] Jiangwei Ren, Xingyu Jiang, Zizhuo Li, Dingkang Liang, Xin Zhou, and Xiang Bai. Minima: Modality invariant image matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23059–23068, 2025.

[40] Marcos V Conde, Gregor Geigle, and Radu Timofte. Instructir: High-quality image restoration following human instructions. In *Proceedings of the European Conference on Computer Vision*, pages 1–21, 2024.

[41] Claude E Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

[42] Gemma Piella. A general framework for multiresolution image fusion: From pixels to regions. *Information Fusion*, 4(4):259–280, 2003.

[43] Veysel Aslantas and Emre Bendes. A new image quality metric for image fusion: The sum of the correlations of differences. *AEU-International Journal of Electronics and Communications*, 69(12):1890–1896, 2015.

[44] Alexander Toet. The tno multiband image data collection. *Data in Brief*, 15:249–251, 2017.

[45] Meng-Hao Guo, Cheng-Ze Lu, Qibin Hou, Zhengning Liu, Ming-Ming Cheng, and Shi-Min Hu. Segnext: Rethinking convolutional attention design for semantic segmentation. *Advances in Neural Information Processing Systems*, pages 1140–1156, 2022.

A Supplemental Material

A.1 More Implementation Details

We provide additional implementation details that are omitted from the main text due to space limitations. G2Fusion is trained in three stages: Flow-GDR pre-training, DiT-GF pre-training, and mutual promotion optimization. For Flow-GDR, we adopt a progressive-resolution training strategy. Specifically, the module is first trained on 256×256 patches and then fine-tuned on 384×384 patches, enabling it to learn coarse geometric correspondences before refining higher-resolution structural details. For DiT-GF, we adopt a DiT-B/2 backbone and train it on 256×256 patches. Flow-GDR and DiT-GF are independently trained for 100k and 500k iterations, respectively. The batch size is set to 16 for all three stages. After the two modules are independently pre-trained, we perform the mutual promotion mechanism through cyclic alternating optimization. Each cycle contains 200 iterations, where the matcher is updated during the first 20 iterations and the diffusion model is updated during the remaining iterations. This stage establishes a closed-loop interaction between registration and fusion, allowing the registration and fusion branches to progressively refine each other while maintaining stable optimization. For target distribution mining, we use weighted scores in both selection stages. In the source-guided filtering stage, SSIM, PSNR, and VIF are combined with weights of 0.4, 0.3, and 0.3, respectively. In the perceptual quality selection stage, SD, SF, AG, and EN are combined with weights of 0.25, 0.3, 0.3, and 0.15, respectively. In the registration objective, the image-level loss weights are set to $\lambda_1 = 0.1$ and $\lambda_2 = 0.9$.

A.2 More Visual Results

We provide additional visual comparisons on the RoadScene and LLVIP datasets in Fig. 9. Compared with existing methods, G2Fusion produces more coherent fusion results under misalignment and degradation, while better preserving salient infrared targets and visible texture details. In challenging scenes with structural inconsistency and illumination variation, our method also suppresses artifacts more effectively and recovers clearer structural information.

A.3 Extended Application

To further verify the extensibility of G2Fusion beyond infrared-visible image fusion, we conduct additional experiments on two representative fusion scenarios: RGB-NIR image fusion and polarization image fusion. RGB-NIR image fusion aims to integrate the rich color and texture information of RGB images with the illumination-robust structural cues of near-infrared images. Polarization image fusion combines intensity information with polarization-aware cues derived from different polarization observations, which reflect material and surface properties. Although these tasks differ in imaging mechanisms, both require effective integration of complementary cross-modal information.

As shown in Fig. 10, for RGB-NIR image fusion, G2Fusion preserves the natural appearance and texture details from RGB images while incorporating structural information from NIR images. The fused results exhibit clearer scene structures and richer visual information. Fig. 11 presents the results of polarization image fusion. Our method effectively integrates polarization-aware cues and spatial details, producing fused images with enhanced informative regions and consistent visual quality. These results indicate that G2Fusion can be extended beyond infrared-visible fusion and has the potential to handle diverse multi-modal fusion tasks involving different imaging sensors.

A.4 Sensitivity to Mutual Promotion Weights

We evaluate the sensitivity of G2Fusion to the mutual promotion weights λ_m and λ_f, which control the geometric feedback applied to the registration and fusion branches, respectively:

$$\begin{aligned} \mathcal{L}_{\text{match}} &= \mathcal{L}_{\text{reg}} + \lambda_m \mathcal{L}_{\text{mp}}, \\ \mathcal{L}_{\text{fusion}} &= \mathcal{L}_{\text{diff}} + \lambda_f \mathcal{L}_{\text{mp}}. \end{aligned} \tag{14}$$

Starting from the default setting $(\lambda_m, \lambda_f) = (0.01, 0.005)$, we compare removing mutual promotion, halving or doubling both coefficients, and reversing their relative magnitudes.

As shown in Table 6, the default setting improves all seven metrics over the zero-weight setting and achieves the best results on most metrics. Halving or doubling both coefficients causes relatively



Figure 9: Qualitative comparison results on RoadScene and LLVIP datasets.

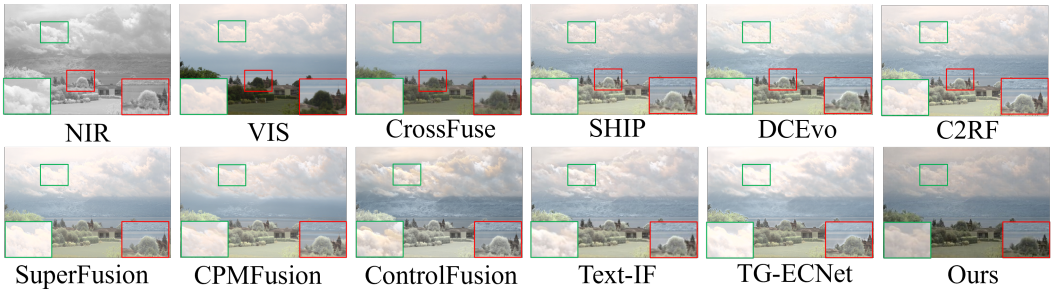

Figure 10: Visual results of extended RGB-NIR image fusion.

small changes, whereas reversing their ratio produces a larger overall decrease. These results suggest that the relative weighting of the two branches matters alongside the overall constraint strength, supporting the use of mutual promotion as an auxiliary geometric regularizer.

A.5 Sensitivity of Target Distribution Mining

We investigate the sensitivity of target distribution mining to metric weights, candidate pool composition, and selection order. The default strategy first filters candidates according to fidelity to the clean source images and then selects the pseudo-target according to intrinsic fusion quality.

Metric weights. The default weights are $(0.40, 0.30, 0.30)$ for $(\mathrm{SSIM}, \mathrm{VIF}, \mathrm{PSNR})$ in the first stage and $(0.25, 0.30, 0.30, 0.15)$ for $(\mathrm{SD}, \mathrm{SF}, \mathrm{AG}, \mathrm{EN})$ in the second stage. Each metric is normal-

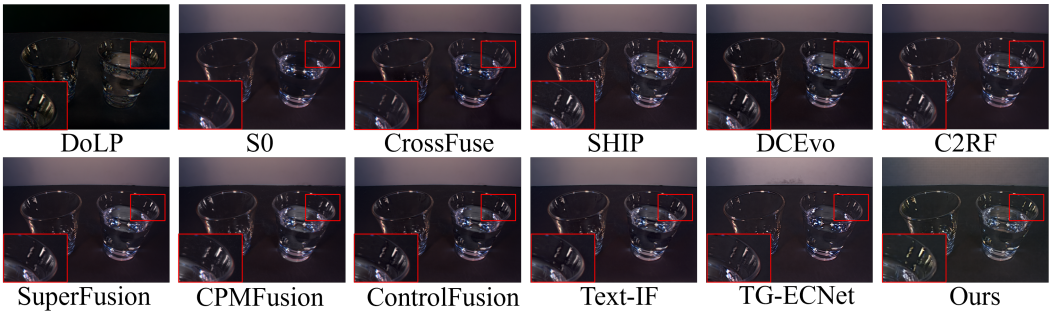

Figure 11: Visual results of extended polarization image fusion.

Table 6: Sensitivity to mutual promotion weights.

Setting	λ_m	λ_f	MI↑	EN↑	CC↑	SCD↑	VIF↑	PSNR↑	SSIM↑
Zero weights	0	0	2.250	6.632	0.566	1.199	0.541	63.143	0.725
Half weights	0.005	0.0025	2.266	**6.652**	**0.572**	1.201	0.543	63.137	0.730
Default	0.010	0.005	**2.308**	6.651	**0.572**	**1.223**	**0.548**	**63.804**	**0.732**
Double weights	0.020	0.010	2.297	6.638	0.567	1.218	0.539	63.103	0.728
Reversed ratio	0.005	0.010	2.189	6.612	0.553	1.154	0.515	62.945	0.721

Table 7: Sensitivity to metric weights in target distribution mining.

Weight setting	MI↑	EN↑	CC↑	SCD↑	VIF↑	PSNR↑	SSIM↑
Default	**2.308**	**6.651**	**0.572**	**1.223**	**0.548**	**63.804**	**0.732**
Equal weights	2.109	6.455	0.561	1.118	0.541	62.899	0.726
Perturbation ($\pm 20\%$)	2.095	6.513	0.554	1.192	0.546	63.113	0.728

ized before aggregation, and the weights remain fixed across images. We compare these settings with equal weights within each stage and perturbations of $\pm 20\%$ around the default weights. For each alternative setting, we regenerate the pseudo-targets and retrain the complete model while keeping all other settings unchanged.

Table 7 shows that the default weights achieve the best results on all seven metrics. The alternative settings retain similar VIF, PSNR, and SSIM scores, although MI decreases more noticeably. This indicates that the metric weights influence the quality of the mined targets, while the magnitude of performance degradation varies across metrics.

Candidate pool composition. The default pool contains outputs from ten fusion methods. We evaluate two alternatives: randomly retaining five methods and excluding the method whose outputs are most frequently selected by the default strategy. As reported in Table 8, the full pool achieves the best results on most metrics. Reducing candidate diversity or removing the most frequently selected method lowers several scores, while retaining much of the overall performance. These results suggest that candidate diversity benefits target mining and that the observed performance is not solely attributable to one candidate method.

Selection order. We further compare the proposed strategy with random target selection, a single selection stage, and a reversed order of the two stages. Table 9 shows that the proposed strategy achieves the best results on six of the seven metrics. The lower overall performance of the reversed order supports filtering for source fidelity before selecting for fusion quality. Together, these comparisons support the proposed organization of target distribution mining.

A.6 Sampling Efficiency

We evaluate the sampling efficiency of DiT-GF using 2, 5, 10, 20, 25, 50, and 100 sampling steps. All settings use the same checkpoint and batch size, with runtime measured on an NVIDIA TITAN RTX at an input resolution of 640×480. As shown in Table 10, most quality metrics improve rapidly at low step counts, while additional sampling yields limited or metric-dependent gains and steadily

Table 8: Sensitivity to candidate pool composition.

Candidate pool	MI↑	EN↑	CC↑	SCD↑	VIF↑	PSNR↑	SSIM↑
Full pool	**2.308**	**6.651**	0.572	**1.223**	**0.548**	**63.804**	**0.732**
Random five methods	2.267	6.432	**0.583**	1.175	0.501	62.989	0.724
Remove dominant candidate	2.256	6.456	0.567	1.188	0.513	63.019	0.722

Table 9: Comparison of target selection strategies.

Selection strategy	MI↑	EN↑	CC↑	SCD↑	VIF↑	PSNR↑	SSIM↑
Random target	1.981	6.080	0.569	1.149	0.466	63.740	0.723
Single stage	1.820	6.434	**0.582**	1.178	0.500	62.682	0.698
Reversed order	1.833	6.319	0.569	1.165	0.488	63.012	0.723
Proposed strategy	**2.308**	**6.651**	0.572	**1.223**	**0.548**	**63.804**	**0.732**

Table 10: Effect of sampling steps on runtime and fusion quality.

Steps	Runtime (s) ↓	MI ↑	EN ↑	CC ↑	SCD ↑	VIF ↑	PSNR ↑	SSIM ↑
2	0.649	1.410	5.467	0.527	0.697	0.307	60.131	0.675
5	0.867	2.225	6.686	0.572	1.179	0.520	63.435	0.740
10	1.232	2.319	6.813	0.571	1.210	0.541	63.345	0.737
20	1.958	2.308	6.651	0.572	1.223	0.548	63.804	0.732
25	2.334	2.327	6.862	0.572	1.223	0.548	63.084	0.730
50	4.138	2.312	6.883	0.572	1.228	0.547	62.979	0.728
100	7.949	2.293	6.894	0.572	1.230	0.545	62.930	0.725

increases runtime. We therefore adopt 20 steps as the default setting to balance fusion quality and efficiency, requiring 1.958 seconds per image for DiT-GF.

A.7 Limitation

Although G2Fusion achieves robust fusion performance under real-world misalignment and complex degradations, it still has several limitations. First, its computational cost remains relatively high due to the iterative denoising process in DiT-GF, which may limit its deployment in real-time or resource-constrained scenarios. Second, the proposed target distribution mining strategy depends on a candidate pool generated by existing fusion and restoration methods. Therefore, the quality, diversity, and coverage of the candidate pool may influence the resulting pseudo-target distribution. Future work will explore efficient diffusion sampling, lightweight generative architectures, and adaptive target distribution modeling to further improve efficiency and generalization.

A.8 Broader Impacts

This paper aims to address multi-modal image fusion under real-world misalignment and degraded conditions, producing high-quality fused results for both human and machine perception. Therefore, it may bring positive impacts in practical scenarios such as autonomous driving, surveillance, and low-visibility monitoring, where reliable fused images can help improve scene understanding and decision-making. For example, clearer infrared-visible fusion may help drivers better perceive road conditions at night or under poor weather. Since this work focuses on image registration, restoration, and fusion, and does not involve privacy-sensitive data or harmful content generation, we do not foresee significant negative societal impacts.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the paper's contributions. The proposed G2Fusion framework jointly addresses spatial misalignment and degradation under a geometric-to-generative paradigm, which is consistent with its key components, including flow-based geometric deformation reduction, DiT-based generative fusion, target distribution mining, and mutual promotion mechanism. These contributions are clearly stated and validated in the methodology and experiments.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of this work in the supplemental material. Specifically, the efficiency of G2Fusion is limited due to the iterative nature of the diffusion-based generative fusion module and multi-step flow estimation in the geometric deformation reduction module. Future work will explore strategies to accelerate diffusion sampling and optimize the computational efficiency of the framework while maintaining registration accuracy and fusion quality.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: This paper does not include formal theoretical results or theorems. The work focuses on a novel geometric-to-generative paradigm for multi-modal image fusion, with modules such as flow-based geometric deformation reduction, DiT-based generative fusion, target distribution mining, and mutual promotion.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide comprehensive experimental details to ensure reproducibility of the main results. This includes detailed descriptions in the Experimental Configurations section, as well as additional information in the supplementary material covering datasets, training details, and computational resources. All necessary code is also included in the supplementary material to enable full replication of our results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example

(a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All datasets used in this work are publicly available and properly cited. The supplementary material includes the complete code for our method along with a README file that provides detailed instructions for setup, data processing, and reproducing the main experimental results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide all relevant training and testing details in the Experimental Configurations section, including dataset splits, hyperparameters, optimizer choice, learning rates, scheduling, data augmentations, and loss weighting, ensuring that the results can be properly interpreted and reproduced.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.

- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The experimental results reported in this paper are the average of a large number of test results in the dataset. Therefore, they are statistically significant, being able to support and validate the contributions and claims of this paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: In the experimental configuration section, we provide the computing resources required to reproduce the experiments in this paper, including NVIDIA TITAN RTX GPUs and an AMD EPYC 7H12 64-core CPU.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

Justification: All data, code, and methodologies used in this work comply with the NeurIPS Code of Ethics. No personally identifiable information is used, and the experiments pose no ethical risks beyond standard research practices.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We discuss the potential societal impacts of this work in the supplementary material. Specifically, G2Fusion can improve human and machine perception in low-visibility or degraded imaging scenarios, benefiting applications such as autonomous driving and surveillance. No significant negative societal impacts are anticipated.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: G2Fusion poses no high-risk of misuse. The released code and datasets are standard publicly available resources, and the method itself does not generate content that could be misused.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets and methods used as assets in this work are publicly available and properly cited. We respect the licenses and terms of use of each dataset and model referenced in the paper.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The code for G2Fusion is publicly released, which includes detailed documentation and usage instructions. This ensures that the new assets are fully accessible and usable by other researchers.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: This work does not involve crowdsourcing experiments or human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: This work does not involve human subjects or crowdsourcing experiments.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: This paper does not use LLMs as part of the core methodology.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.