GAUSSIAN PRIMITIVE FUSION FOR ANY-RESOLUTION MULTIMODAL IMAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal image fusion integrates complementary observations from different modalities into a unified representation. Existing methods typically estimate modality contributions on a common discrete grid. When source modalities have different native resolutions, constructing such a grid generally involves resampling, which may introduce artifacts and affect subsequent fusion decisions. To address this problem, we propose Gaussian Primitive Fusion (GPF), which reformulates heterogeneous-resolution fusion in a shared continuous Gaussian space. Instead of first converting source images to a common resolution, GPF directly fits their native-resolution observations with shared Gaussian geometry and modality-specific content representations, and learns modality contributions within the resulting continuous representation. Once optimized, the fused representation is rendered at the finest native resolution without resampling the sources. With heterogeneous-resolution inputs, GPF produces clearer target contours and finer textures, with less blurring and blocking artifacts than resample-then-fuse pipelines. Quantitatively, it achieves the best results on multiple infrared-visible benchmarks against state-of-the-art fusion methods equipped with interpolation or super-resolution frontends.

1 INTRODUCTION

Multimodal image fusion integrates complementary observations into a single image to provide a more complete scene representation, with applications in nighttime surveillance, autonomous driving, and medical imaging (Ma et al., 2019; Liu et al., 2022; James & Dasarathy, 2014). Effective fusion should preserve useful source information while estimating the relative contribution of each modality. Most existing multimodal fusion methods achieve this by comparing and integrating pixels or features at corresponding locations (Li et al., 2013; Li & Wu, 2018; Xu et al., 2022; Ma et al., 2022; Zhao et al., 2023), implicitly assuming that the source images are aligned on a common sampling grid. However, variations in sensor configurations often produce observations with unequal native resolutions (Ma et al., 2020; Li et al., 2021). When input resolutions differ, a common solution is to upsample the lower-resolution source or resample both sources to a common size (Xiao et al., 2022; Jie et al., 2025). This process estimates values at unobserved locations and may alter edges and fine structures (Keys, 1981; Thévenaz et al., 2000). Such changes propagate into feature extraction and cross-modal interaction, affecting both the structures available for fusion and the evidence used to estimate modality contributions.

Continuous representations decouple image content modeling from fixed discrete sampling grids (Sitzmann et al., 2020; Chen et al., 2021; Sun et al., 2025a). Each modality can be represented over a shared continuous scene domain while remaining directly constrained by its native-resolution observations (McGinnis et al., 2023). This enables cross-modal fusion to be learned directly in the continuous domain, without first resampling the sources to a common grid (Tang et al., 2021). To model this shared domain, we employ a common set of two-dimensional Gaussian primitives. Their learnable centers, scales, and orientations adapt the representation to local source structures, while modality-specific attributes preserve the content of each source (Kerbl et al., 2023; Zhang et al., 2024). We further associate multiscale fusion coefficients with the same primitives (Burt & Adelson, 1983), allowing source representation and modality contribution estimation to share a unified spatial support. The output is rendered at the finest native resolution, so no source is resampled to a common grid before fusion (Figure 1).

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.

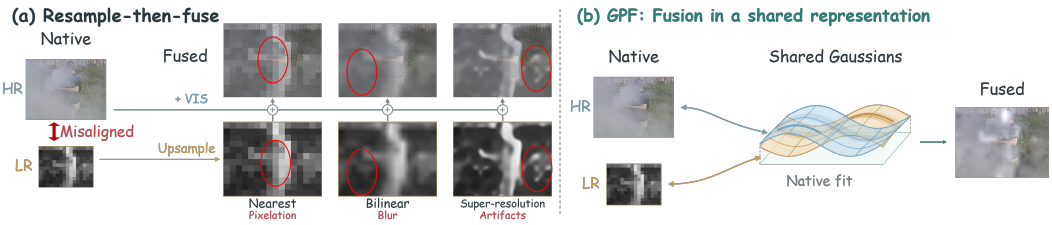

Figure 1: Two paradigms for heterogeneous-resolution image fusion. Grid-based pipelines resample sources before fusion. GPF fits a shared Gaussian representation from native observations and learns contributions on its primitives.

Building on this formulation, we introduce *Gaussian Primitive Fusion* (GPF), which adopts a two-stage optimization strategy. GPF first fits a shared Gaussian representation directly to the native-resolution observations, jointly learning common spatial geometry and modality-specific source content. It then freezes the fitted source representation and optimizes multiscale modality contributions on the same Gaussian supports. This separation preserves the source content established during native fitting while allowing fusion decisions to adapt across different detail levels. The resulting fused representation remains continuous and can be rendered on target grids without refitting. Our contributions are summarized as follows:

1. We formulate heterogeneous-resolution fusion through continuous representations constrained directly by native observations, avoiding source resizing before contribution learning.
2. We introduce shared Gaussian primitives that co-represent modality content and multiscale fusion decisions on adaptive spatial supports fitted jointly to native observations.
3. We develop a two-stage fitting-and-fusion strategy that preserves native-resolution source content during contribution learning and supports flexible-grid readout without refitting.

2 Related Work

Grid-based and resolution-aware fusion. DenseFuse combines encoded source features (Li & Wu, 2018), and U2Fusion learns information-preserving fusion across tasks (Xu et al., 2022). SwinFusion models cross-domain interactions (Ma et al., 2022), while CDDFuse separates shared and modality-specific features through correlation-guided decomposition (Zhao et al., 2023). These approaches organize fusion on discrete feature grids. Resolution-aware methods learn resolution conversion in trained networks: DDcGAN adversarially matches the downsampled fused image to the low-resolution source (Ma et al., 2020), Li et al. (2021) meta-learn arbitrary-resolution upscaling, and FS-Diff couples fusion with super-resolution via diffusion (Jie et al., 2025). GPF instead fits shared Gaussians to each input's native observations, without dataset-level training.

Continuous image representations. Coordinate-based models represent images as continuous functions. SIREN uses periodic activations to capture signals and derivatives (Sitzmann et al., 2020), and LIIF combines coordinates with local features for arbitrary-scale readout (Chen et al., 2021). INR-Fuse represents the fused image with a coordinate network and fits it to both sources through spatially uniform intensity and gradient consistency losses (Sun et al., 2025a). Where the sources conflict, these squared terms favor averaged responses, which can weaken target contrast and fine details. Its multi-resolution formulation aligns fused outputs with each source without a pixel-footprint observation operator. GPF instead fits modality-specific source fields through native observation operators, preserving each modality's observed content before learning fusion contributions.

Gaussian representations and multimodal fusion. Three-dimensional Gaussian splatting enables differentiable scene rendering with anisotropic primitives (Kerbl et al., 2023), while GaussianImage applies two-dimensional Gaussians to image representation and compression (Zhang et al., 2024). IVGF reconstructs an infrared–visible Gaussian scene for free-viewpoint fusion and uses cross-modal opacity adjustment (Yang et al., 2026). GPF focuses on registered two-dimensional observations with heterogeneous resolutions. Its shared geometry supports independent source content and multiscale contributions, connecting native-observation fitting directly to the spatial units used for fusion.

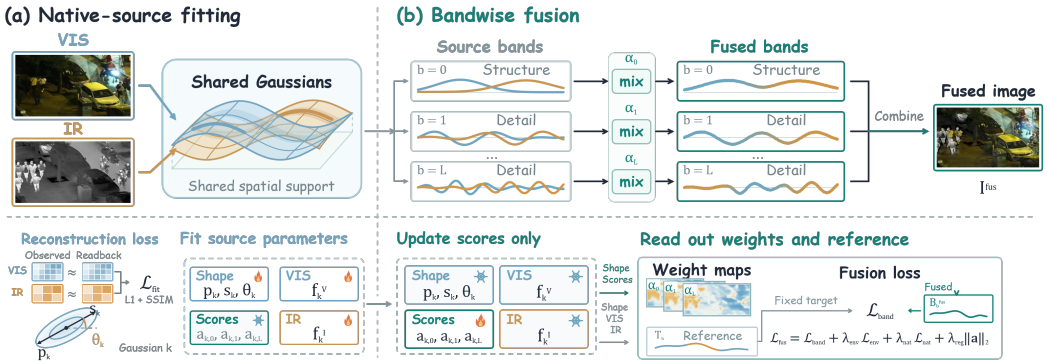

Figure 2: Overview of GPF. Native-source fitting learns shared Gaussian geometry and separate modality attributes from reconstruction at native resolutions. Bandwise fusion freezes the source fields and updates only primitive scores. Source components and scores share an averaging hierarchy.

3 Gaussian Primitive Fusion

Given two registered observations $Y_m \in \mathbb{R}^{H_m \times W_m \times d_m}$, $m \in \{1, 2\}$, GPF learns source content and contributions over the common domain $\Omega = [0, 1]^2$. The native grids $C_m = (H_m, W_m)$ may differ. Figure 2 illustrates the two stages: joint source fitting followed by contribution learning, with both stages operating on the same adaptive Gaussian spatial supports.

3.1 Shared Gaussian representation

We represent the observations with K two-dimensional Gaussian primitives. Each primitive stores a center p_k, positive-definite covariance Σ_k, independent source attributes $\mathbf{f}_k^{(m)} \in \mathbb{R}^{d_m}$, and one scalar coefficient per fusion band:

$$\mathcal{P}_k = \big(p_k, \Sigma_k, \mathbf{f}_k^{(1)}, \mathbf{f}_k^{(2)}, \{a_{k,b}\}_{b=0}^{L}\big), \qquad k = 1, \dots, K. \tag{1}$$

The covariance is parameterized as $\Sigma_k = R(\theta_k)\,\mathrm{diag}(s_{k,1}^2, s_{k,2}^2)R(\theta_k)^\top$, with positive axis scales. Centers, extents, and orientations are shared across modalities. Theorem A.1 (Appendix A) shows that a sufficiently large shared Gaussian dictionary reproduces the native observations of both modalities exactly, and that Gaussian envelopes minimize covariance-normalized sensitivity to center displacement. The Gaussian kernels define a normalized spatial basis for evaluating the fields at any coordinate $x \in \Omega$:

$$g_k(x) = \exp\!\big[-\tfrac{1}{2}(x - p_k)^\top \Sigma_k^{-1}(x - p_k)\big], \qquad \phi_k(x) = \frac{g_k(x)}{\epsilon + \sum_j g_j(x)}, \quad \epsilon > 0. \tag{2}$$

Source fields and bandwise score fields are then defined on the same basis:

$$\widehat{f}_m(x) = \sum_{k=1}^{K} \phi_k(x)\mathbf{f}_k^{(m)}, \qquad q_b(x) = \sum_{k=1}^{K} \phi_k(x) a_{k,b}, \quad b = 0, \dots, L. \tag{3}$$

Independent attributes retain modality-specific values, while shared geometry determines where these values are expressed. Attaching scores to the same primitives associates contribution control with those fitted spatial regions. Separate coefficients across bands allow a region to receive different source contributions for coarse structure and fine details.

3.2 Fitting native-resolution observations

Each source field is supervised through an observation operator $\mathcal{A}_{m,C}$ that predicts the original input grid, with $C = (C_1, C_2)$. Numerically, let $Q_\star = (\max_m H_m, \max_m W_m)$ be the fitting grid and $\mathcal{R}_{Q_\star}$ evaluate a field at its pixel centers. Area-overlap averaging D_m aggregates these predictions over the spatial footprint of each pixel in the corresponding native observation:

$$\widehat{Y}_m = \mathcal{A}_{m,C}\widehat{f}_m, \qquad \mathcal{A}_{m,C} = D_m \mathcal{R}_{Q_\star}, \qquad (D_m)_{P,R} = \frac{|P \cap R|}{|P|}. \tag{4}$$

Here P and R are native and fitting-grid cells. This construction accommodates noninteger resolution ratios and compares predictions directly with unchanged input arrays. The first stage jointly optimizes shared geometry $\mathcal{G} = \{p_k, \Sigma_k\}_k$ and modality attributes $\mathbf{F}_m = \{\mathbf{f}_k^{(m)}\}_k$:

$$\mathcal{L}_{\text{fit}} = \sum_{m=1}^{2} \lambda_m \Big[\text{MAE}(\widehat{Y}_m, Y_m) + \eta\big(1 - \text{SSIM}(\widehat{Y}_m, Y_m)\big) \Big]. \tag{5}$$

Mean absolute error preserves source values, and structural similarity encourages local agreement (Wang et al., 2004). Both reconstruction losses update the common geometry; each also updates its own modality attributes. After fitting, we freeze $\mathcal{G}$ and $\mathbf{F}_m$ so contribution learning operates on fixed source content fitted jointly to the native observations.

3.3 Multiscale contributions on shared primitives

We assign contributions at several detail levels using a Laplacian-style decomposition (Burt & Adelson, 1983). On an evaluation grid Q, define $\Phi_Q[p,k] = \phi_k(x_p)$ and scalar source images $J_0^{(m)} = \chi_m(\Phi_Q \mathbf{F}_m)$, where χ_m is a fixed linear intensity map. Let $Q_0 = Q$, and ceiling-halve each dimension to form $Q_{\ell+1}$. Successive area-averaging maps P_ℓ define $E_0 = I$, $E_{\ell+1} = P_\ell E_\ell$, and $J_\ell^{(m)} = E_\ell J_0^{(m)}$. Bilinear lifts U_ℓ between adjacent levels give the following source components:

$$B_0^{(m)} = J_L^{(m)}, \qquad B_b^{(m)} = J_{L-b}^{(m)} - U_{L-b} J_{L-b+1}^{(m)}, \quad b = 1, \dots, L. \tag{6}$$

The same hierarchy evaluates contribution scores at each band's spatial scale. Writing $a_b = (a_{1,b}, \dots, a_{K,b})^\top$, we average the score field before applying the sigmoid and combine the corresponding source components:

$$\begin{aligned} W_{b,Q} &= E_{L-b}\Phi_Q, & \alpha_{b,Q} &= \sigma(W_{b,Q} a_b), \\ B_b^{\text{fus}} &= (1 - \alpha_{b,Q}) \odot B_b^{(1)} + \alpha_{b,Q} \odot B_b^{(2)}, & \sigma(t) &= (1 + e^{-t})^{-1}. \end{aligned} \tag{7}$$

Thus source structure and contribution scores use the same geometry and averaging operators. Starting with $\widetilde{J}_L = B_0^{\text{fus}}$, synthesis proceeds from coarse to fine:

$$\widetilde{J}_\ell = B_{L-\ell}^{\text{fus}} + U_\ell \widetilde{J}_{\ell+1}, \quad \ell = L-1, \dots, 0, \qquad J_Q^{\text{fus}} = \widetilde{J}_0. \tag{8}$$

The recurrence combines broad image structure with progressively finer residual content. All spatially varying decisions originate from the coefficients attached to the fitted primitives, and the same readout is used for fitting-grid supervision and requested output grids. In practice, the renderer treats each source attribute and score as a channel. Successive averaging then evaluates $W_{b,Q} a_b$ without explicitly constructing the dense basis matrix, and backpropagation updates the primitive scores.

3.4 Learning source contributions

Stage two optimizes only the concatenated coefficients $a = \{a_{k,b}\}_{k,b}$ on the fixed fitting grid $Q_\star$. The frozen source fields provide all references. Let $G_b^{(m)}(p) = \|\nabla B_b^{(m)}(p)\|_2$ be local gradient magnitude. References select the stronger source response, with ties assigned to source 1:

$$i_b(p) = \arg\max_{m \in \{1,2\}} G_b^{(m)}(p), \qquad T_b(p) = B_b^{(i_b(p))}(p), \qquad \mathcal{L}_{\text{band}} = \sum_{b=0}^{L} \|B_b^{\text{fus}} - T_b\|_2^2. \tag{9}$$

This reference encourages each band to retain locally informative source content through the shared Gaussian supports. A complementary envelope term penalizes attenuation of strong source gradients:

$$\mathcal{L}_{\text{env}} = \sum_{b=0}^{L} \mathbb{E}_p \left[\max_m G_b^{(m)}(p) - \|\nabla B_b^{\text{fus}}(p)\|_2 \right]_+, \qquad [t]_+ = \max(t, 0). \tag{10}$$

To preserve detail at the original observation scales, we also compare native readbacks of the fused image with the frozen native source reconstructions $\overline{J}_m = D_m J_0^{(m)}$. For source and fused derivatives t and v, we use a one-sided penalty on gradient attenuation:

$$\psi_{\epsilon_n}(t, v) = \left[|t + \epsilon_n \operatorname{sign}(t)| - |v + \epsilon_n \operatorname{sign}(t)| \right]_+, \qquad \operatorname{sign}(0) = 0. \tag{11}$$

The small fixed offset is oriented by the source derivative. The native-resolution objective is

$$\mathcal{L}_{\mathrm{nat}} = \tfrac{1}{2} \sum_{m=1}^{2} \mathbb{E}_{p,d}\, \psi_{\epsilon_n}\left(\nabla_d \overline{J}_m(p), \nabla_d (D_m J^{\mathrm{fus}}_{Q_\star})(p)\right), \tag{12}$$

where d indexes horizontal and vertical differences. Gradients pass through fused readback to the coefficients, while native references remain fixed. Combining these terms yields

$$\mathcal{L}_{\mathrm{fus}} = \mathcal{L}_{\mathrm{band}} + \lambda_{\mathrm{env}}\mathcal{L}_{\mathrm{env}} + \lambda_{\mathrm{nat}}\mathcal{L}_{\mathrm{nat}} + \lambda_{\mathrm{reg}}\|a\|_2. \tag{13}$$

The band and envelope terms retain complementary structure across scales, the native term preserves source responses on their original grids, and the unsquared coefficient norm controls score magnitude. The source representation remains unchanged throughout this stage.

3.5 Optimization and output rendering

We optimize both stages with Adam (Kingma & Ba, 2015). The default setting uses $L = 2$, $K = 25{,}000$, $\epsilon = \epsilon_n = 10^{-6}$, $\eta = 0.2$, $\lambda_m = 1$, and $(\lambda_{\mathrm{env}}, \lambda_{\mathrm{nat}}, \lambda_{\mathrm{reg}}) = (0.1, 0.1, 0.01)$. Source fitting uses 1000 steps at learning rate 10^{-2}, followed by 200 score updates at 5×10^{-3}. The appendix specifies intensity conversion, sampling, and loss averaging conventions.

Score coefficients are initialized to zero, giving equal source contributions at every band. The fitted source bands, gradient-selected references, and native readbacks are cached once. Each subsequent iteration updates only the $K(L+1)$ coefficients, with loss gradients passing through Gaussian readout, scale averaging, source mixing, and coarse-to-fine synthesis.

After learning, we store $\mathcal{G}$, $\mathbf{F}_1$, $\mathbf{F}_2$, and a. A requested output grid Q determines where Gaussian splatting evaluates the source and score channels. Their common hierarchy then yields weights, fused bands, and the output through Eqs. (6)–(8). Output requests reuse all learned parameters and leave the native fitting configuration unchanged. Fusion relationships are therefore learned from the original-resolution observations before the target rendering grid is selected.

4 Experiments

4.1 Experimental setup

Datasets and resolution settings. We evaluate infrared-visible fusion on LLVIP, MSRS, M3FD, RoadScene and TNO (Jia et al., 2021; Tang et al., 2022b; Liu et al., 2022; Xu et al., 2020; Toet, 2017), covering low-light pedestrians, urban roads, natural backgrounds and surveillance scenes. PET–MRI (Ma et al., 2020) adds medical pairs with natively mismatched resolutions ($\times 4$), requiring no synthetic degradation. We construct asymmetric-resolution inputs by degrading either the infrared or visible modality while keeping the other at its native resolution. Directional experiments independently vary the horizontal and vertical sampling rates, including complementary VIS-x/IR-y degradation. Observations are generated using area resampling and antialiased bicubic degradation, covering different pixel-aggregation and smoothing characteristics.

Compared methods and reconstruction frontends. We compare 16 image-fusion methods (Zhao et al., 2020; Tang et al., 2022a; Zhao et al., 2023; Liu et al., 2024; Song et al., 2026; Lei et al., 2025; Bai et al., 2025; Xu et al., 2025; Sun et al., 2025b; Ma et al., 2026; Yi et al., 2025; Hu et al., 2025; Li et al., 2025; Cheng et al., 2025; Chen et al., 2026; Tang et al., 2026) under heterogeneous-resolution inputs. Since existing fusion methods generally require inputs on a common grid, we pair them with interpolation-based reconstruction (Keys, 1981) or learned super-resolution frontends, including HAT (Chen et al., 2023), SwinIR (Liang et al., 2021), and LIIF (Chen et al., 2021). For each fusion method, we report the strongest evaluated frontend configuration, with the corresponding reconstruction method indicated in the Frontend column. In contrast, GPF operates directly on the native-resolution observations and performs fusion within the fitted continuous representation, without an external resolution-conversion frontend.

Evaluation metrics. We report edge preservation Q_{abf} (Xydeas & Petrović, 2000), visual information fidelity (Sheikh & Bovik, 2006), and structural similarity (Wang et al., 2004). Following the source-wise summation used in fusion evaluation (Zhao et al., 2023), we use $\mathrm{VIF}_{\Sigma} =$

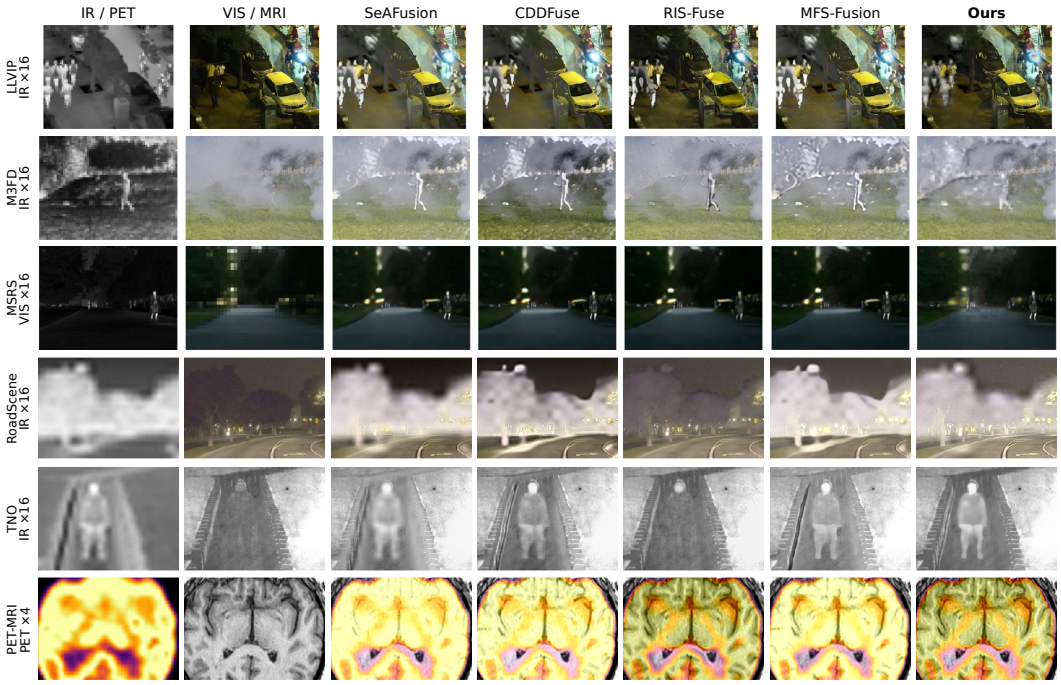

Figure 3: Qualitative comparison under heterogeneous-resolution inputs.

Table 1: Results with infrared downsampled ($\times 16$) and on native PET–MRI. Frontend identifies each method's best reconstruction. Bold/underline: best/second-best.

Method	Frontend	LLVIP			MSRS			M3FD			RoadScene			PET–MRI			
		$Q_{abf}\uparrow$	$\text{VIF}_\Sigma\uparrow$	$S_\Sigma\uparrow$	$Q_{abf}\uparrow$	$\text{VIF}_\Sigma\uparrow$	$S_\Sigma\uparrow$	$Q_{abf}\uparrow$	$\text{VIF}_\Sigma\uparrow$	$S_\Sigma\uparrow$	$Q_{abf}\uparrow$	$\text{VIF}_\Sigma\uparrow$	$S_\Sigma\uparrow$	Frontend	$Q_{abf}\uparrow$	$\text{VIF}_\Sigma\uparrow$	$S_\Sigma\uparrow$
DIDFuse	Bicubic	0.250	0.366	0.900	0.223	0.307	0.701	0.494	0.594	1.241	0.312	0.465	1.178	Bicubic	0.460	0.412	1.168
SeAFusion	HAT	0.441	0.516	1.224	0.604	0.839	1.330	0.554	0.639	1.373	0.325	0.464	1.180	Lanczos	0.652	0.586	1.305
CDDFuse	HAT	0.388	0.502	1.239	0.604	0.904	1.318	0.580	0.747	1.384	0.352	0.588	1.150	Bilinear	0.643	0.592	1.293
AGMFusion	Bicubic	0.424	0.575	1.156	0.567	0.807	1.306	0.514	0.626	1.295	0.301	0.486	1.184	Lanczos	0.607	0.491	1.216
MFDFuse	Bicubic	0.385	0.510	1.253	0.612	0.897	1.337	0.582	0.744	1.384	0.344	0.608	1.124	Bilinear	0.646	0.630	1.000
MLFuse	Bicubic	0.378	0.522	1.242	0.556	0.718	1.283	0.541	0.621	<u>1.424</u>	0.361	0.600	1.201	Lanczos	0.651	0.597	0.816
TDFusion	LIIF	0.527	0.734	1.272	0.628	0.931	1.333	0.628	0.858	1.410	0.344	0.535	1.190	Lanczos	0.626	0.558	1.146
Deno-IF	Bicubic	0.347	0.517	1.073	0.542	0.768	1.332	0.505	0.578	1.412	0.349	0.560	1.201	Nearest	0.634	0.515	1.311
TG-ECNet	Bicubic	0.411	0.582	1.101	0.494	0.527	1.182	0.448	0.744	1.209	0.196	0.189	1.010	Lanczos	0.662	0.565	1.220
RIS-Fuse	Bicubic	<u>0.535</u>	<u>0.813</u>	<u>1.276</u>	<u>0.631</u>	0.946	1.343	<u>0.652</u>	<u>0.944</u>	1.401	<u>0.439</u>	<u>0.910</u>	<u>1.363</u>	Bilinear	0.660	<u>0.738</u>	1.018
LUT-Fuse	Bicubic	0.283	0.383	1.204	0.531	0.701	1.314	0.414	0.540	1.368	0.282	0.426	1.258	HAT	0.649	0.522	1.285
TITA	Bicubic	0.459	0.587	1.237	0.620	0.872	1.351	0.600	0.721	1.411	0.350	0.515	1.291	Bicubic	0.650	0.661	1.253
MaeFuse	Bicubic	0.183	0.408	1.201	0.502	0.671	1.314	0.373	0.452	1.382	0.336	0.520	1.291	Nearest	0.650	0.530	0.990
FDFuse	Bicubic	0.468	0.641	1.212	0.616	<u>0.948</u>	1.326	0.610	0.848	1.384	0.367	0.689	1.145	Bilinear	0.652	0.624	0.971
MFS-Fusion	LIIF	0.513	0.681	1.234	0.620	0.930	<u>1.352</u>	0.614	0.800	1.385	0.375	0.579	1.279	Bilinear	<u>0.687</u>	0.715	<u>1.319</u>
SETFusion	Bicubic	0.176	0.251	1.185	0.191	0.294	1.164	0.388	0.462	1.295	0.264	0.458	1.215	Swin2SR	0.302	0.479	0.914
GPF (Ours)	Native	**0.575**	**0.935**	**1.314**	**0.659**	**0.959**	**1.367**	**0.664**	**0.972**	**1.427**	**0.452**	**0.965**	**1.379**	Native	**0.690**	**0.742**	**1.329**

$\text{VIF}(I,F)+\text{VIF}(V,F)$ and $S_\Sigma = \text{SSIM}(I,F)+\text{SSIM}(V,F)$. Here I, V and F denote the degraded input sources and the fused luminance.

Implementation. Source fitting uses 1000 steps and 200 contribution updates. Experiments run on RTX 3090 GPUs. All methods in a qualitative row are displayed over the same spatial region.

4.2 Quantitative comparisons

Table 1 compares GPF with fusion methods under heterogeneous-resolution inputs. For each competing method, we report its strongest evaluated reconstruction configuration. GPF achieves consistently strong performance across the five benchmarks in terms of Q_{abf}, VIF_Σ, and S_Σ.

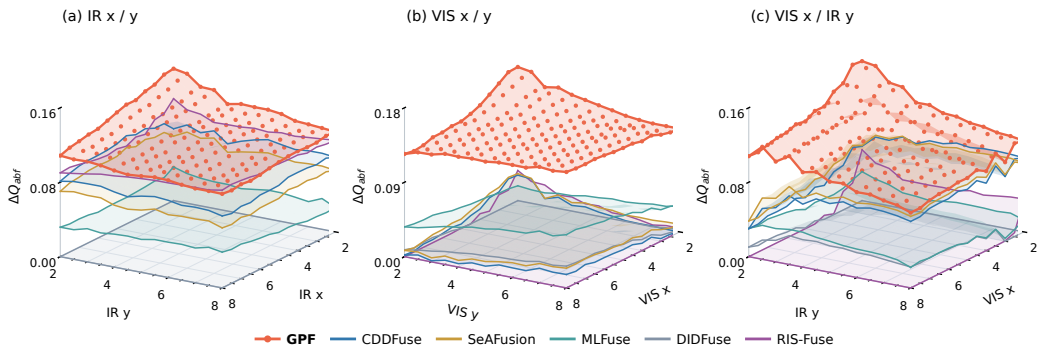

Figure 4: Fusion performance across heterogeneous resolution settings. ΔQ_{abf} subtracts the lowest displayed baseline mean at each sampling coordinate.

The gains are observed across diverse scenes, including pedestrian-focused low-light images, urban roads, and natural backgrounds. In particular, the higher Q_{abf} reflects improved preservation of source edges, while the gains in VIF_{Σ} and S_{Σ} indicate stronger retention of complementary visual and structural information from the two modalities. Further results across different input resolution configurations, including visible-modality degradation, are provided in the appendix.

4.3 Qualitative comparisons

Figure 3 presents representative results on infrared–visible datasets together with a PET–MRI example. On LLVIP, GPF preserves clear pedestrian contours while retaining visible background structures under strong local illumination. On M3FD, where the low-resolution infrared observation contains a small warm target, reconstruction-based pipelines tend to alter its apparent extent or contrast before fusion, whereas GPF retains the target response together with visible foliage and ground textures. On MSRS, despite the degraded visible observation, GPF preserves distant infrared pedestrians while maintaining surrounding illumination and background structure.

Similar behavior is observed on RoadScene and TNO, where GPF retains road markings, thermal vegetation, and coherent pedestrian boundaries with fewer visible artifacts. For PET–MRI fusion, the result preserves both the PET intensity distribution and MRI anatomical boundaries. These visual results are consistent with the quantitative gains in Table 1, showing that GPF preserves complementary source structures under heterogeneous-resolution inputs. Additional qualitative results are provided in the appendix.

4.4 Performance across heterogeneous resolution settings

We further evaluate whether the performance gains extend beyond the resolution configurations used in the main comparison. Figure 4 independently varies the horizontal and vertical sampling factors of the infrared or visible modality, and additionally considers a crossed VIS-x/IR-y setting in which the two modalities are degraded along different spatial directions. These settings extend the evaluation from fixed asymmetric degradation to a broader range of heterogeneous-resolution inputs.

Across all three settings, GPF maintains higher mean edge fidelity than the compared methods over a wide range of sampling configurations. The separation is particularly clear when the visible modality is degraded and remains evident in the crossed setting. These results show that the advantage of GPF is not restricted to a particular resolution ratio or degraded modality, but extends across diverse spatial sampling configurations.

4.5 Discussion

Impact of reconstruction frontends on fusion. Figure 5 shows that the performance of existing fusion methods changes noticeably with the choice of reconstruction frontend. For each fusion method, we compute the variation in Q_{abf} induced solely by changing the reconstruction frontend, and report scene-level values together with their range and mean. With the fusion method unchanged, this

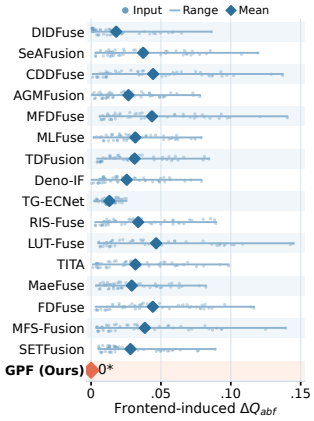

Figure 5: Impact of reconstruction frontends on fusion quality. 0^* denotes native GPF without external reconstruction frontend.



Figure 6: Qualitative comparison under crossed VIS-x/IR-y 16$\times$ degradation.

variation indicates that resolution reconstruction can alter the spatial structures and local responses presented to the fusion stage. In heterogeneous-resolution settings, the reconstruction frontend therefore becomes an additional factor that affects the final fusion result. In contrast, GPF directly fits each modality from its native-resolution observations and preserves the source structure within the learned continuous representation.

The source-fidelity results in Figure 7 further show that the adaptive Gaussian representation maintains strong reconstruction quality and native-grid consistency. Figure 6 provides a qualitative example under crossed VIS-x/IR-y degradation, where resample-then-fuse pipelines exhibit blurred contours and block-shaped artifacts, while GPF retains more coherent target boundaries and surrounding structures. These results indicate that GPF not only avoids the uncertainty introduced by an external reconstruction frontend, but also provides a faithful source representation for subsequent fusion.

Effect of adaptive Gaussian representation. We investigate whether the advantage of GPF stems from the use of a continuous representation alone or from the adaptive geometry of Gaussian primitives. We therefore compare adaptive Gaussians with linear grids, cubic B-splines (Unser, 1999), SIREN (Sitzmann et al., 2020), and fixed Gaussians under comparable representation capacity, identical native observations, and the same fusion rule. As shown in Figure 7, adaptive Gaussians achieve the strongest overall performance across source fidelity, native readback, and fusion quality. Compared with grid and spline representations, adaptive Gaussians provide content-dependent spatial supports instead of regular predefined ones. The comparison with SIREN shows that a continuous representation alone does not account for the observed improvement, while the gap between fixed and adaptive Gaussians further highlights the benefit of learning the primitive geometry. These results indicate that adaptive Gaussian geometry provides a more suitable spatial representation for preserving source structure and supporting subsequent fusion.

Ablations. Table 2 evaluates the main design choices in GPF. Replacing native observations with resized fitting targets reduces all three metrics, with Q_{abf} dropping from 0.575 to 0.529. We next vary the geometry used to represent the two modalities. Jointly fitted geometry performs better than IR-led, VIS-led, and independent alternatives; the largest degradation occurs when the two modalities use independent geometries.

We also compare different ways of parameterizing the fusion weights. Removing learned weights leads to a substantial performance drop. Sharing weights across scales improves over this baseline, but remains below the proposed band-specific primitive scores. Pixel-wise and matched-grid scores also give lower fusion quality, suggesting that associating the fusion weights with the fitted primitives provides a better match between source representation and contribution estimation.

Table 2: Design ablations.

Configuration	Q_{abf} ↑	VIF_Σ ↑	S_Σ ↑
GPF (Ours)	**0.575**	**0.935**	**1.314**
A. Observation constraint			
Resize targets	0.529	0.875	1.240
B. Geometry supervision			
IR-led	0.539	0.907	1.291
VIS-led	0.483	0.894	1.275
Independent	0.426	0.810	1.258
C. Fusion weights			
Without learned weights	0.467	0.831	1.192
Shared across scales	0.548	0.910	1.275
Pixel scores	0.530	0.822	1.229
Matched grid scores	0.514	0.878	1.268

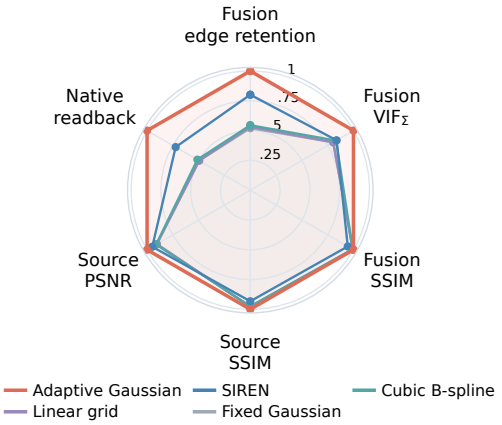

Figure 7: Normalized source and fusion fidelity.

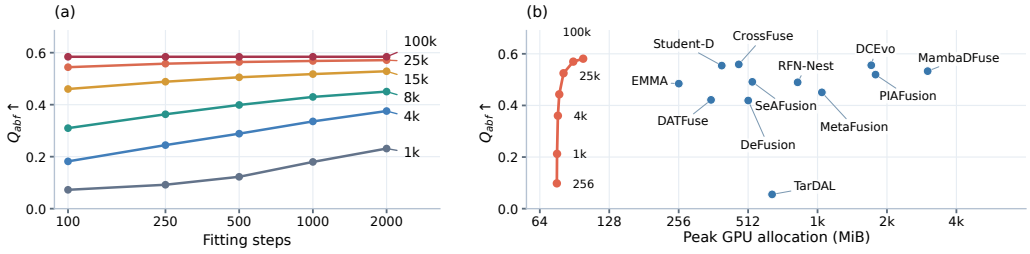

Figure 8: (a) Fitting steps at different Gaussian counts. (b) Peak GPU allocation versus fusion quality.

Gaussian capacity, fitting budget and memory. Figure 8(a) examines representation capacity together with the source-fitting budget. Increasing the Gaussian count raises the attainable fusion quality, while additional fitting is particularly helpful for smaller representations. The larger representations reach a stable region sooner, and their curves flatten as optimization continues. This trend links the practical fitting budget to the amount of spatial detail that the representation can retain: optimization improves the use of the available capacity, while additional capacity provides more freedom to represent that detail. Figure 8(b) relates this quality to peak GPU allocation. The Gaussian sweep moves toward stronger edge preservation with a modest increase in allocated memory. Its endpoint exceeds the displayed comparison methods in Q_{abf} on this scene, showing that a compact fitted representation can retain useful structural information without requiring a large allocation.

5 Conclusion

We presented Gaussian Primitive Fusion (GPF) for multimodal image fusion with heterogeneous-resolution inputs. Instead of reconstructing the source modalities onto a common grid before fusion, GPF directly fits their native-resolution observations with shared Gaussian geometry and modality-specific content, and learns multiscale fusion contributions on the same adaptive primitives. This formulation couples source representation and contribution estimation within a common continuous spatial support while keeping the original observation grids as the fitting constraints. Experiments on multiple infrared–visible benchmarks, together with PET–MRI examples, show that GPF preserves complementary source structures more effectively than resample-then-fuse pipelines. The advantage remains across diverse heterogeneous-resolution configurations and is less dependent on the choice of an external reconstruction frontend. These results suggest that native-observation-driven continuous fusion provides a practical alternative to common-grid preprocessing for multimodal images with unequal spatial resolutions.

AI USE STATEMENT

We used generative AI tools for language polishing and for preparing figures. The authors take responsibility for the final text, mathematical claims, experimental results, and figures.

REFERENCES

Haowen Bai, Jiangshe Zhang, Zixiang Zhao, Yichen Wu, Lilun Deng, Yukun Cui, Tao Feng, and Shuang Xu. Task-driven image fusion with learnable fusion loss. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7457–7468. IEEE, 2025.

Peter J Burt and Edward H Adelson. The Laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 31(4):532–540, 1983.

Tao Chen, Chuang Wang, Yuanpeng Zhang, Kaijian Xia, and Pengjiang Qian. MFS-Fusion: Mamba-integrated deep multi-modal image fusion framework with multi-scale Fourier enhancement and spatial calibration. *Expert Systems with Applications*, 299:130054, 2026.

Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22367–22377. IEEE, 2023.

Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8628–8638, 2021.

Muhang Cheng, Haiyan Huang, Xiangyu Liu, Hongwei Mo, Songling Wu, and Xiongbo Zhao. FDFuse: Infrared and visible image fusion based on feature decomposition. *IEEE Transactions on Instrumentation and Measurement*, 74:1–13, 2025.

Xingyu Hu, Junjun Jiang, Chenyang Wang, Kui Jiang, Xianming Liu, and Jiayi Ma. Balancing task-invariant interaction and task-specific adaptation for unified image fusion. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11262–11272. IEEE, 2025.

Alex Pappachen James and Belur V Dasarathy. Medical image fusion: A survey of the state of the art. *Information Fusion*, 19:4–19, 2014.

Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. LLVIP: A visible-infrared paired dataset for low-light vision. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 3489–3497. IEEE, 2021.

Yuchan Jie, Yushen Xu, Xiaosong Li, Fuqiang Zhou, Jianming Lv, and Huafeng Li. FS-Diff: Semantic guidance and clarity-aware simultaneous multimodal image fusion and super-resolution. *Information Fusion*, 121:103146, 2025.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian Splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):139:1–139:14, 2023.

Robert Keys. Cubic convolution interpolation for digital image processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(6):1153–1160, 1981.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Jia Lei, Jiawei Li, Jinyuan Liu, Bin Wang, Shihua Zhou, Qiang Zhang, Xiaopeng Wei, and Nikola K Kasabov. MLFuse: multi-scenario feature joint learning for multi-modality image fusion. *IEEE Transactions on Multimedia*, 27:3880–3894, 2025.

Huafeng Li, Yueliang Cen, Yu Liu, Xun Chen, and Zhengtao Yu. Different input resolutions and arbitrary output resolution: A meta learning-based deep framework for infrared and visible image fusion. *IEEE Transactions on Image Processing*, 30:4070–4083, 2021.

Hui Li and Xiao-Jun Wu. DenseFuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5):2614–2623, 2018.

Jiayang Li, Junjun Jiang, Pengwei Liang, Jiayi Ma, and Liqiang Nie. MaeFuse: Transferring omni features with pretrained masked autoencoders for infrared and visible image fusion via guided training. *IEEE Transactions on Image Processing*, 34:1340–1353, 2025.

Shutao Li, Xudong Kang, and Jianwen Hu. Image fusion with guided filtering. *IEEE Transactions on Image Processing*, 22(7):2864–2875, 2013.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. SwinIR: Image restoration using Swin Transformer. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 1833–1844. IEEE, 2021.

Tony Lindeberg. Generalized gaussian scale-space axiomatics comprising linear scale-space, affine scale-space and spatio-temporal scale-space. *Journal of Mathematical Imaging and Vision*, 40(1): 36–81, 2011. doi: 10.1007/s10851-010-0242-2.

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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5792–5801. IEEE, 2022.

Shenghao Liu, Xiaoxiong Lan, Wenyong Chen, Zhiyong Zhang, and Changzhen Qiu. AGMFusion: A real-time end-to-end infrared and visible image fusion network based on adaptive guidance module. *IEEE Sensors Journal*, 24(17):28338–28350, 2024.

Jiayi Ma, Yong Ma, and Chang Li. Infrared and visible image fusion methods and applications: A survey. *Information Fusion*, 45:153–178, 2019.

Jiayi Ma, Han Xu, Junjun Jiang, Xiaoguang Mei, and Xiao-Ping Zhang. DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Transactions on Image Processing*, 29:4980–4995, 2020.

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, 2022.

Siju Ma, Changsiyu Gong, Xiaofeng Fan, Yong Ma, and Chengjie Jiang. RIS-FUSION: Rethinking text-driven infrared and visible image fusion from the perspective of referring image segmentation. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 11902–11906. IEEE, 2026.

Sean McCurdy and Raghavendra Venkatraman. Quantitative stability for the heisenberg–pauli–weyl inequality. *arXiv preprint arXiv:2007.07111*, 2020. URL `https://arxiv.org/abs/2007.07111`.

Julian McGinnis, Suprosanna Shit, Hongwei Bran Li, Vasiliki Sideri-Lampretsa, Robert Graf, Maik Dannecker, Jiazhen Pan, Nil Stolt-Ansó, Mark Mühlau, Jan S Kirschke, et al. Single-subject multi-contrast MRI super-resolution via implicit neural representations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 173–183. Springer, 2023.

Hamid R Sheikh and Alan C Bovik. Image information and visual quality. *IEEE Transactions on Image Processing*, 15(2):430–444, 2006.

Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems*, 33:7462–7473, 2020.

Hao Song, Longkun He, Shan Zhao, Lijin Gao, and Chengjiang Zhou. MFDFuse: A multi-frequency decomposition-based network for infrared and visible image fusion. *Infrared Physics & Technology*, pp. 106442, 2026.

Shuchen Sun, Ligen Shi, Chang Liu, Lina Wu, and Jun Qiu. Infrared and visible image fusion based on implicit neural representations. In *2025 IEEE 15th International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER)*, pp. 78–83. IEEE, 2025a.

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 *International Conference on Machine Learning*, pp. 57571–57586. PMLR, 2025b.

Jiaxiang Tang, Xiaokang Chen, and Gang Zeng. Joint implicit image function for guided depth super-resolution. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 4390–4399, 2021.

Linfeng Tang, Jiteng Yuan, and Jiayi Ma. Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network. *Information Fusion*, 82: 28–42, 2022a.

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: 79–92, 2022b.

Wei Tang, Fazhi He, Lin Zhang, and Shengjie Zhao. SETFusion: A semantic transformer for infrared and visible image fusion. *Pattern Recognition*, 175:113130, 2026.

Philippe Thévenaz, Thierry Blu, and Michael Unser. Interpolation revisited. *IEEE Transactions on Medical Imaging*, 19(7):739–758, 2000.

Alexander Toet. The TNO multiband image data collection. *Data in Brief*, 15:249, 2017.

Michael Unser. Splines: A perfect fit for signal and image processing. *IEEE Signal Processing Magazine*, 16(6):22–38, 1999.

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.

Wanxin Xiao, Yafei Zhang, Hongbin Wang, Fan Li, and Hua Jin. Heterogeneous knowledge distillation for simultaneous infrared-visible image fusion and super-resolution. *IEEE Transactions on Instrumentation and Measurement*, 71:1–15, 2022.

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*, volume 34, pp. 12484–12491, 2020.

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.

Han Xu, Yuyang Li, Yunfei Deng, Jiayi Ma, and Guangcan Liu. Deno-IF: Unsupervised noisy visible and infrared image fusion method. In *Advances in Neural Information Processing Systems*, volume 38, pp. 141649–141671, 2025.

Costas S Xydeas and Vladimir Petrović. Objective image fusion performance measure. *Electronics Letters*, 36(4):308–309, 2000.

Chao Yang, Deshui Miao, Chao Tian, Guoqing Zhu, Yameng Gu, and Zhenyu He. Fusing in 3D: Free-viewpoint fusion rendering with a 3D infrared-visible scene representation. *arXiv preprint arXiv:2601.12697*, 2026.

Xunpeng Yi, Yibing Zhang, Xinyu Xiang, Qinglong Yan, Han Xu, and Jiayi Ma. LUT-Fuse: Towards extremely fast infrared and visible image fusion via distillation to learnable look-up tables. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14559–14568. IEEE, 2025.

Xinjie Zhang, Xingtong Ge, Tongda Xu, Dailan He, Yan Wang, Hongwei Qin, Guo Lu, Jing Geng, and Jun Zhang. GaussianImage: 1000 FPS image representation and compression by 2D Gaussian splatting. In *European Conference on Computer Vision*, pp. 327–345. Springer, 2024.

Zixiang Zhao, Shuang Xu, Chunxia Zhang, Junmin Liu, Jiangshe Zhang, and Pengfei Li. DIDFuse: Deep image decomposition for infrared and visible image fusion. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 970–976, 2020.

Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Shuang Xu, Zudi Lin, Radu Timofte, and Luc Van Gool. CDDFuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5906–5916. IEEE, 2023.

A GAUSSIAN REPRESENTATION AND STABILITY

This section establishes the approximation and geometric properties of the shared Gaussian representation. We use the normalization in the main paper,

$$g_k(x) = \exp\left[-\tfrac{1}{2}(x-p_k)^\top \Sigma_k^{-1}(x-p_k)\right], \qquad \phi_k(x) = \frac{g_k(x)}{\epsilon + \sum_j g_j(x)}, \qquad \epsilon > 0, \tag{A.1}$$

with positive-definite Σ_k and independent coefficients for each source channel. Shared geometry means that every channel uses the same p_k and Σ_k. Throughout, $\Omega = [0,1]^d$; the image case has $d = 2$.

A.1 STATEMENT

Theorem A.1 (Shared approximation and Gaussian displacement stability)**.** *The following properties hold.*

1. ***Shared representation.*** *For every fixed $\epsilon > 0$, there is a sequence of finite shared Gaussian dictionaries such that any finite collection of continuous source fields can be approximated uniformly on Ω by their linear combinations. The same dictionaries approximate every finite collection of $L^2(\Omega)$ fields in L^2.*

   For a finite family of bounded linear observation maps $\mathcal{H}_m : X \to \mathbb{R}^{s_m}$, where $X = C(\Omega)$ with the uniform norm or $X = L^2(\Omega)$ with its usual norm, a sufficiently large finite shared dictionary spans the complete response space of every $\mathcal{H}_m$. Thus each compatible observation vector $y_m \in \operatorname{range} \mathcal{H}_m$ has a representation satisfying $\mathcal{H}_m \widehat{f}_m = y_m$ with real-valued source coefficients.

2. ***Displacement stability.*** *Let $\psi \in H^1(\mathbb{R}^d)$ be real, with $\|\psi\|_2 = 1$ and finite second spatial moment. Write*

$$\mu = \int x\psi(x)^2\,dx, \qquad S = \int (x-\mu)(x-\mu)^\top \psi(x)^2\,dx, \qquad J = \int \nabla\psi \nabla\psi^\top\,dx. \tag{A.2}$$

   Then $S \succ 0$, and the covariance-normalized translation sensitivity is

$$\mathcal{L}(\psi) := \sup_{h \neq 0} \frac{\|\psi(\cdot - h) - \psi\|_2^2}{h^\top S^{-1} h} = \lambda_{\max}(S^{1/2} J S^{1/2}) \geq \frac{1}{4}. \tag{A.3}$$

   At fixed energy, center and covariance, equality holds precisely for

$$\psi(x) = \pm(2\pi)^{-d/4} (\det S)^{-1/4} \exp\left[-\tfrac{1}{4}(x-\mu)^\top S^{-1}(x-\mu)\right]. \tag{A.4}$$

   The positive sign gives the unique nonnegative minimizer.

The first part applies directly to the normalized fields in Eq. (A.1). The second part characterizes the unnormalized primitive envelope under a fixed spatial concentration. Below we prove both parts and give the corresponding perturbation bound for the normalized representation. The stability argument is a matrix form of the uncertainty principle; related quantitative results are studied by McCurdy & Venkatraman (2020).

A.2 CONSTRUCTIVE SHARED APPROXIMATION

Divide Ω into n^d cells $C_{n,k}$ of side $1/n$, with centers $p_{n,k}$. Set $\Sigma_{n,k} = n^{-1} I$, and write

$$s_n(x) = \sum_{k=1}^{n^d} \exp[-n\|x - p_{n,k}\|^2/2], \quad \phi_{n,k}(x) = \frac{\exp[-n\|x-p_{n,k}\|^2/2]}{\epsilon + s_n(x)}. \tag{A.5}$$

The Gaussian width $n^{-1/2}$ tends to zero while covering an increasing number of cell centers. For $n \geq 4$, each coordinate interval of radius $n^{-1/2}$ around x contains at least $\sqrt{n}/2$ lattice centers within $[0,1]$. Each corresponding kernel is at least $e^{-d/2}$, so

$$s_n(x) \geq c_d n^{d/2}, \qquad c_d = 2^{-d} e^{-d/2}, \qquad x \in \Omega. \tag{A.6}$$

This lower bound also controls the additive ϵ in the denominator.

For a scalar source define its cell averages $c_{n,k}(f) = n^d \int_{C_{n,k}} f(y)\,dy$, and the positive linear operator

$$T_n f(x) = \sum_k \phi_{n,k}(x) c_{n,k}(f) = \int_\Omega K_n(x,y) f(y)\,dy, \quad K_n(x,y) = n^d \sum_k \phi_{n,k}(x) \mathbf{1}_{C_{n,k}}(y). \tag{A.7}$$

Let $M = \|f\|_\infty$ and let ω_f denote the modulus of continuity of f. Splitting centers into $\|x - p_{n,k}\| \le r$ and its complement gives

$$\|T_n f - f\|_\infty \le \omega_f\left(r + \frac{\sqrt{d}}{2n}\right) + 2Mc_d^{-1} n^{d/2} e^{-nr^2/2} + M\epsilon c_d^{-1} n^{-d/2}. \tag{A.8}$$

The three terms account for local variation, the Gaussian tail and the normalization offset, respectively. Choosing $r_n = \sqrt{(d+4)\log n/n}$ makes the tail term $2Mc_d^{-1}n^{-2}$, while the other two terms also converge to zero. This proves uniform approximation with the same geometry for all continuous source channels.

For L^2 fields, $\int K_n(x,y)\,dy \le 1$. Moreover, if $y \in C_{n,k}$, Eq. (A.6) gives

$$\int_\Omega K_n(x,y)\,dx \le \frac{n^d}{c_d n^{d/2}} \int_{\mathbb{R}^d} e^{-n\|x - p_{n,k}\|^2/2}\,dx = \frac{(2\pi)^{d/2}}{c_d} =: C_d. \tag{A.9}$$

Cauchy–Schwarz with the positive kernel, followed by integration, yields $\|T_n f\|_2 \le \sqrt{C_d}\|f\|_2$. Given $f \in L^2$, choose continuous u with small $\|f - u\|_2$. Then

$$\|T_n f - f\|_2 \le (\sqrt{C_d} + 1)\|f - u\|_2 + \|T_n u - u\|_2. \tag{A.10}$$

Taking $n \to \infty$ and then $u \to f$ proves L^2 convergence. A finite sum over source channels establishes the shared multimodal result.

A.3 Finite native observations

Fix one observation map $\mathcal{H}$ of rank r. Identify its image with $\mathbb{R}^r$, and choose $u_1, \dots, u_r \in X$ whose responses are the coordinate vectors. Let $\widetilde{\mathcal{H}}$ denote this coordinate representation, and form

$$M_n = [\widetilde{\mathcal{H}} T_n u_1 \ \cdots \ \widetilde{\mathcal{H}} T_n u_r]. \tag{A.11}$$

Approximation in X and boundedness of $\mathcal{H}$ imply $M_n \to I_r$. Hence M_n is invertible for all sufficiently large n. For the coordinate vector $\widetilde{y}$ of any compatible observation y, set

$$\widehat{f}_n(y) = \sum_{j=1}^r (M_n^{-1}\widetilde{y})_j T_n u_j. \tag{A.12}$$

This is a linear combination of the same finite dictionary, and $\mathcal{H}\widehat{f}_n(y) = y$. Taking the largest threshold over the finite set of modalities gives one common dictionary. The geometry depends on the observation family, while the source coefficients depend on the observed values.

For an actual finite representation and any reference field f°, the residual obeys

$$\|\mathcal{H}\widehat{f} - y\| \le \|\mathcal{H}\|\,\|\widehat{f} - f^\circ\|_X + \|\mathcal{H}f^\circ - y\|. \tag{A.13}$$

Continuous pixel averages are bounded on L^2 when each pixel has positive area. The numerical operator in the main paper, $D_m\mathcal{R}_{Q_\star}$, is bounded on $C(\Omega)$: point evaluation has unit norm and each row of D_m is a nonnegative unit-sum average. Thus Eq. (A.12) covers both the continuous observation model and its finite quadrature implementation. In the latter case, using the maximum norm gives $\|D_m\mathcal{R}_{Q_\star}\| \le 1$.

A.4 PROOF OF DISPLACEMENT OPTIMALITY

If $v^\top S v = 0$ for a nonzero v, the probability density ψ^2 would be supported on the hyperplane $v^\top(x-\mu)=0$, a set of zero Lebesgue measure. This contradicts $\|\psi\|_2 = 1$, proving $S \succ 0$.

Integration by parts yields

$$\int (x-\mu)\psi(x)\nabla\psi(x)^\top \, dx = -\tfrac{1}{2} I. \tag{A.14}$$

For completeness, multiply by a smooth cutoff χ_R equal to one on the radius-R ball and zero outside the radius-$2R$ ball. The additional term contains $(x-\mu)\nabla\chi_R\psi^2$ and vanishes as $R \to \infty$; the remaining products are integrable because $(x-\mu)\psi$ and $\nabla\psi$ lie in L^2. Expanding the outer product of $v(x) = \nabla\psi(x) + \frac{1}{2}S^{-1}(x-\mu)\psi(x)$ now gives the exact identity

$$J - \tfrac{1}{4}S^{-1} = \int v(x)v(x)^\top \, dx \succeq 0. \tag{A.15}$$

For a translation h, the weak fundamental theorem of calculus gives

$$\psi(\cdot - h) - \psi = -\int_0^1 h^\top \nabla\psi(\cdot - th)\, dt. \tag{A.16}$$

Translation invariance of the L^2 norm and Minkowski's inequality imply $\|\psi(\cdot-h)-\psi\|_2^2 \le h^\top J h$. Conversely, for each direction u, the L^2 difference quotient as $h = tu \to 0$ converges to $-u^\top\nabla\psi$. Therefore the supremum of the ratio in Eq. (A.3) equals the largest generalized Rayleigh quotient of (J, S^{-1}). Equation (A.15) bounds it below by $1/4$.

If that largest eigenvalue is $1/4$, Eq. (A.15) forces $v = 0$ almost everywhere. Multiplying by $\exp[(x-\mu)^\top S^{-1}(x-\mu)/4]$ produces a function with zero weak gradient on connected $\mathbb{R}^d$, hence a constant. Unit energy determines Eq. (A.4). Conversely, that Gaussian satisfies

$$\|\psi(\cdot-h)-\psi\|_2^2 = 2\Big(1 - e^{-h^\top S^{-1}h/8}\Big), \tag{A.17}$$

whose ratio to $h^\top S^{-1} h$ has supremum $1/4$. This completes the proof of Theorem A.1.

A.5 NORMALIZED FIELDS AND SCALE TRANSFORMATIONS

For the peak-one kernel g_k in Eq. (A.1), the energy density $g_k^2/\|g_k\|_2^2$ has covariance $S_k = \Sigma_k/2$. The theorem therefore gives

$$\|g_k(\cdot - h_k) - g_k\|_2 \le \tfrac{1}{2}\|g_k\|_2 \sqrt{h_k^\top S_k^{-1} h_k}. \tag{A.18}$$

To propagate this bound through normalization, let $N = \sum_k c_k g_k$, $D = \epsilon + \sum_k g_k$, and similarly define N', D' for perturbed kernels. With $\|c_k\| \le M$ and $D' \ge d_- > 0$ on the observation domain,

$$\left\| \frac{N'}{D'} - \frac{N}{D} \right\|_{L^2(\Omega)} \le \frac{2M}{d_-} \sum_k \|g_k' - g_k\|_{L^2(\Omega)}. \tag{A.19}$$

Indeed, insert N/D' between the two fractions, use $\|N\| \le MD$, and bound $|D - D'|$ by $\sum_k |g_k - g_k'|$. This accounts for changes to both the kernel values and their common normalizer. Applying a bounded observation map multiplies the bound by its operator norm.

Gaussian kernels also admit exact scale and affine transformations (Lindeberg, 2011). If G_P is the unit-mass Gaussian with covariance P, Fourier multiplication gives

$$G_P * g_{\mu,\Sigma} = \sqrt{\frac{\det\Sigma}{\det(\Sigma+P)}}\, g_{\mu,\Sigma+P}. \tag{A.20}$$

Under an invertible coordinate map $x \mapsto Ax + b$, the center and covariance become $A\mu + b$ and $A\Sigma A^\top$. For a pixel average $\mathcal{O}$, the normalized field is evaluated in the order $\mathcal{O}(N/D)$. Its relation to separately averaged numerator and denominator is

$$\frac{\mathcal{O}N}{\mathcal{O}D} - \mathcal{O}(N/D) = \frac{\mathrm{Cov}_{\mathcal{O}}(D, N/D)}{\mathcal{O}D}. \tag{A.21}$$

The native area operator and the output-grid quadrature in the implementation retain this normalization-before-averaging order.

B IMPLEMENTATION DETAILS

Coordinates and initialization. Inputs are normalized to $[0, 1]$. Pixel (i, j) on an $H \times W$ grid has coordinate $((j+1/2)/W, (i+1/2)/H)$. Native observations are compared through nonnegative, unit-sum area-overlap weights. Gaussian centers start on a regular lattice, with axis scales $0.6/\sqrt{K}$ and zero rotation; source attributes are initialized by bilinear sampling. Visible intensity is $\chi_V(R, G, B) = 0.299R + 0.587G + 0.114B$.

Freezing and caching. At the beginning of contribution learning, we freeze all Gaussian centers, scales, rotations, and modality-specific source attributes. On the fixed fitting grid $Q_\star$, we evaluate the source fields and construct their averaging pyramids and band components once. For each band, the source gradient magnitudes determine the gradient-selected target, with ties assigned to source 1, and the pointwise gradient envelope. We also compute the native source readbacks $\overline{J}_m = D_m J_0^{(m)}$. The source reconstructions, pyramid levels, bands, targets, envelopes, and native readbacks are cached with gradient tracking disabled and reused throughout contribution learning.

The primitive score coefficients form a $K \times (L + 1)$ tensor initialized to zero. Consequently, the initial sigmoid weights are $1/2$ at every band and spatial location. At each iteration, the current score channels are rendered on $Q_\star$ and averaged to the corresponding band grids before applying the sigmoid. The resulting weights combine the cached source bands, and coarse-to-fine synthesis produces the current fused image. The band, envelope, and native-readback losses are then evaluated against the fixed cached references. Gradients propagate through score rendering, scale averaging, source mixing, synthesis, and fused readback to update only the score coefficients. With the default $L = 2$, this stage optimizes $3K$ coefficients while keeping the fitted source representation unchanged.

Numerical conventions. MAE and SSIM average over native pixels and channels. SSIM uses an 11-tap Gaussian window with standard deviation 1.5 and replicated boundaries. Pyramid levels ceiling-halve each dimension using area averaging; bilinear lifting uses half-pixel coordinates. The band loss sums squared residuals over pixels and bands. Envelope losses average within each band. Native gradient penalties average valid horizontal and vertical differences within each modality and then average the modalities. The score regularizer is the unsquared Euclidean norm.

Evaluation and hardware. We report edge fidelity Q_{abf}, $\mathrm{VIF}_\Sigma = \mathrm{VIF}(I, F) + \mathrm{VIF}(V, F)$, and $S_\Sigma = \mathrm{SSIM}(I, F) + \mathrm{SSIM}(V, F)$. For all metrics, low-resolution sources are bicubically resampled to the output grid, identically for all methods. Computation uses float32 on an NVIDIA RTX 3090, with PyTorch 2.4.1 and CUDA 12.1. Memory measurements report peak PyTorch allocated memory.

C ADDITIONAL COMPARISONS

Table C.1: IR×4. Column groups follow the main-table order; bold and underline mark the best and second-best scores.

Method	Frontend	LLVIP			MSRS			M3FD			RoadScene			TNO		
		$Q_{abf}\uparrow$	$\mathrm{VIF}_\Sigma\uparrow$	$S_\Sigma\uparrow$	$Q_{abf}\uparrow$	$\mathrm{VIF}_\Sigma\uparrow$	$S_\Sigma\uparrow$	$Q_{abf}\uparrow$	$\mathrm{VIF}_\Sigma\uparrow$	$S_\Sigma\uparrow$	$Q_{abf}\uparrow$	$\mathrm{VIF}_\Sigma\uparrow$	$S_\Sigma\uparrow$	$Q_{abf}\uparrow$	$\mathrm{VIF}_\Sigma\uparrow$	$S_\Sigma\uparrow$
DIDFuse	Bicubic	0.262	0.425	0.901	0.222	0.315	0.708	0.478	0.634	1.264	0.327	0.491	1.230	0.380	0.568	1.198
SeAFusion	HAT	0.537	0.737	1.303	0.623	0.885	1.348	0.561	0.656	1.396	0.373	0.513	1.266	0.441	0.615	1.289
CDDFuse	HAT	0.497	0.738	1.320	0.625	0.940	1.340	0.578	0.736	1.413	0.381	0.560	1.250	0.473	0.705	1.306
AGMFusion	Bicubic	0.488	0.747	1.229	0.557	0.808	1.324	0.497	0.640	1.313	0.303	0.481	1.238	0.359	0.519	1.237
MFDFuse	Bicubic	0.496	0.763	1.314	0.631	0.949	1.359	0.566	0.736	1.407	0.359	0.605	1.195	0.457	0.700	1.276
MLFuse	Bicubic	0.412	0.642	1.300	0.546	0.732	1.310	0.497	0.604	**1.463**	0.340	0.547	1.259	0.418	0.597	1.349
TDFusion	LIIF	0.585	0.807	**1.324**	0.626	0.918	1.357	0.605	0.764	1.440	0.364	0.517	1.277	0.455	0.633	1.317
Deno-IF	Bicubic	0.420	0.642	1.130	0.543	0.767	1.345	0.494	0.583	1.434	0.360	0.513	1.285	0.371	0.534	1.323
TG-ECNet	Bicubic	0.455	0.685	1.159	0.531	0.618	1.173	0.483	0.724	1.222	0.208	0.211	1.021	0.334	0.386	1.124
RIS-Fuse	Bicubic	**0.608**	0.876	1.313	0.648	0.967	1.355	0.652	0.901	1.403	0.434	0.738	1.354	**0.541**	0.834	1.333
LUT-Fuse	Bicubic	0.395	0.640	1.283	0.548	0.766	1.338	0.431	0.593	1.397	0.324	0.507	1.331	0.377	0.585	1.315
TITA	Bicubic	0.543	0.789	1.306	0.630	0.908	**1.371**	0.594	0.733	1.430	0.367	0.559	1.365	0.469	0.676	1.358
MaeFuse	Bicubic	0.258	0.565	1.256	0.489	0.686	1.332	0.355	0.500	1.416	0.317	0.482	1.338	0.321	0.484	1.335
FDFuse	Bicubic	0.562	0.821	1.278	0.629	0.959	1.338	0.616	0.829	1.398	0.410	0.616	1.265	0.510	0.788	1.298
MFS-Fusion	LIIF	0.599	0.843	1.303	0.632	0.929	1.367	0.621	0.780	1.405	0.396	0.517	1.329	0.506	0.683	1.343
SETFusion	Bicubic	0.224	0.441	1.242	0.200	0.362	1.189	0.379	0.508	1.315	0.259	0.438	1.250	0.282	0.447	1.211
GPF (Ours)	Native	0.568	**0.965**	1.319	**0.658**	**0.997**	1.368	**0.664**	**0.971**	1.428	**0.452**	**0.961**	**1.379**	0.537	**0.883**	**1.373**

Table C.2: IR×16. Column groups follow the main-table order; bold and underline mark the best and second-best scores.

Method	Frontend	LLVIP			MSRS			M3FD			RoadScene			TNO		
		$Q_{abf}\uparrow$	$\mathrm{VIF}_\Sigma\uparrow$	$S_\Sigma\uparrow$	$Q_{abf}\uparrow$	$\mathrm{VIF}_\Sigma\uparrow$	$S_\Sigma\uparrow$	$Q_{abf}\uparrow$	$\mathrm{VIF}_\Sigma\uparrow$	$S_\Sigma\uparrow$	$Q_{abf}\uparrow$	$\mathrm{VIF}_\Sigma\uparrow$	$S_\Sigma\uparrow$	$Q_{abf}\uparrow$	$\mathrm{VIF}_\Sigma\uparrow$	$S_\Sigma\uparrow$
DIDFuse	Bicubic	0.250	0.366	0.900	0.223	0.307	0.701	0.494	0.594	1.241	0.312	0.465	1.178	0.391	0.564	1.165
SeAFusion	HAT	0.441	0.516	1.224	0.604	0.839	1.330	0.554	0.639	1.373	0.325	0.464	1.180	0.418	0.544	1.236
CDDFuse	HAT	0.388	0.502	1.239	0.604	0.904	1.318	0.580	0.747	1.384	0.352	0.588	1.150	0.462	0.693	1.252
AGMFusion	Bicubic	0.424	0.575	1.156	0.567	0.807	1.306	0.514	0.626	1.295	0.301	0.486	1.184	0.366	0.501	1.192
MFDFuse	Bicubic	0.385	0.510	1.253	0.612	0.897	1.337	0.582	0.744	1.384	0.344	0.608	1.124	0.445	0.687	1.226
MLFuse	Bicubic	0.378	0.522	1.242	0.556	0.718	1.283	0.541	0.621	1.424	0.361	0.600	1.201	0.440	0.620	1.280
TDFusion	LIIF	0.527	0.734	1.272	0.628	0.931	1.333	0.628	0.858	1.410	0.344	0.535	1.190	0.466	0.645	1.263
Deno-IF	Bicubic	0.347	0.517	1.073	0.542	0.768	1.332	0.505	0.578	1.412	0.349	0.560	1.201	0.355	0.510	1.261
TG-ECNet	Bicubic	0.411	0.582	1.101	0.494	0.527	1.182	0.448	0.744	1.209	0.196	0.189	1.010	0.328	0.355	1.096
RIS-Fuse	Bicubic	0.535	0.813	1.276	0.631	0.946	1.343	0.652	0.944	1.401	0.439	0.910	1.363	0.528	0.865	1.322
LUT-Fuse	Bicubic	0.283	0.383	1.204	0.531	0.701	1.314	0.414	0.540	1.368	0.282	0.426	1.258	0.339	0.488	1.259
TITA	Bicubic	0.459	0.587	1.237	0.620	0.872	1.351	0.600	0.721	1.411	0.350	0.515	1.291	0.451	0.615	1.304
MaeFuse	Bicubic	0.183	0.408	1.201	0.502	0.671	1.314	0.373	0.452	1.382	0.336	0.520	1.291	0.316	0.452	1.284
FDFuse	Bicubic	0.468	0.641	1.212	0.616	0.948	1.326	0.610	0.848	1.384	0.367	0.689	1.145	0.485	0.793	1.250
MFS-Fusion	LIIF	0.513	0.681	1.234	0.620	0.930	1.352	0.614	0.800	1.385	0.375	0.579	1.279	0.503	0.712	1.287
SETFusion	Bicubic	0.176	0.251	1.185	0.191	0.294	1.164	0.388	0.462	1.295	0.264	0.458	1.215	0.273	0.389	1.169
GPF (Ours)	Native	**0.575**	**0.935**	**1.314**	**0.659**	**0.959**	**1.367**	**0.664**	**0.972**	**1.427**	**0.452**	**0.965**	**1.379**	**0.548**	**0.923**	**1.357**

Table C.3: VIS×4. Column groups follow the main-table order; bold and underline mark the best and second-best scores.

Method	Frontend	LLVIP			MSRS			M3FD			RoadScene			TNO		
		$Q_{abf}\uparrow$	$\mathrm{VIF}_{\Sigma}\uparrow$	$S_{\Sigma}\uparrow$	$Q_{abf}\uparrow$	$\mathrm{VIF}_{\Sigma}\uparrow$	$S_{\Sigma}\uparrow$	$Q_{abf}\uparrow$	$\mathrm{VIF}_{\Sigma}\uparrow$	$S_{\Sigma}\uparrow$	$Q_{abf}\uparrow$	$\mathrm{VIF}_{\Sigma}\uparrow$	$S_{\Sigma}\uparrow$	$Q_{abf}\uparrow$	$\mathrm{VIF}_{\Sigma}\uparrow$	$S_{\Sigma}\uparrow$
DIDFuse	Bicubic	0.126	0.339	0.862	0.120	0.266	0.703	0.248	0.541	1.212	0.460	0.598	1.239	0.303	0.530	1.164
SeAFusion	HAT	0.419	0.697	1.261	**0.415**	0.653	1.295	0.312	0.418	1.294	0.496	0.627	1.292	0.382	0.595	1.259
CDDFuse	HAT	0.440	0.733	1.278	0.409	0.672	1.292	0.292	0.395	1.308	0.493	0.630	1.312	0.378	0.582	1.277
AGMFusion	Bicubic	0.357	0.736	1.227	0.349	0.657	1.273	0.264	0.509	1.274	0.393	0.548	1.252	0.323	0.531	1.234
MFDFuse	Bicubic	0.395	0.805	1.280	0.356	0.644	1.301	0.256	0.468	1.322	0.489	0.682	1.276	0.370	0.648	1.252
MLFuse	Bicubic	0.292	0.577	1.288	0.294	0.572	1.287	0.243	0.458	1.372	0.433	0.574	1.341	0.323	0.529	1.319
TDFusion	LIIF	0.404	0.666	1.270	0.412	0.693	1.310	0.322	0.469	1.349	0.537	0.702	1.325	0.403	0.621	1.294
Deno-IF	Bicubic	0.337	0.678	1.274	0.299	0.591	1.291	0.244	0.477	1.374	0.434	0.570	1.336	0.300	0.507	1.300
TG-ECNet	Bicubic	0.288	0.463	1.163	0.278	0.405	1.178	0.223	0.314	1.228	0.201	0.228	1.081	0.227	0.321	1.137
RIS-Fuse	Bicubic	0.374	0.648	1.276	0.355	0.623	1.288	0.225	0.324	1.319	0.421	0.455	1.324	0.326	0.480	1.280
LUT-Fuse	Bicubic	0.362	0.718	1.269	0.332	0.622	1.283	0.208	0.405	1.335	0.443	0.592	1.345	0.340	0.560	1.287
TITA	Bicubic	0.384	0.760	1.281	0.326	0.601	1.318	0.231	0.429	1.376	0.507	0.670	**1.392**	0.355	0.598	**1.334**
MaeFuse	Bicubic	0.248	0.559	1.249	0.288	0.580	1.287	0.206	0.439	1.372	0.436	0.548	1.366	0.304	0.498	1.324
FDFuse	Bicubic	0.410	0.799	1.251	0.395	0.698	1.268	0.288	0.467	1.319	0.523	0.656	1.302	0.395	0.603	1.269
MFS-Fusion	LIIF	**0.448**	0.760	1.254	**0.415**	0.669	1.320	0.333	0.467	1.322	0.605	0.776	1.362	0.438	0.651	1.320
SETFusion	Bicubic	0.217	0.490	1.235	0.155	0.377	1.185	0.229	0.485	1.276	0.364	0.465	1.221	0.292	0.487	1.192
GPF (Ours)	Native	0.440	**1.011**	**1.318**	0.351	**0.968**	**1.339**	**0.375**	**0.956**	**1.420**	**0.651**	**0.973**	1.388	**0.487**	**0.904**	1.322

Table C.4: VIS×16. Column groups follow the main-table order; bold and underline mark the best and second-best scores.

Method	Frontend	LLVIP			MSRS			M3FD			RoadScene			TNO		
		$Q_{abf}\uparrow$	$\mathrm{VIF}_{\Sigma}\uparrow$	$S_{\Sigma}\uparrow$	$Q_{abf}\uparrow$	$\mathrm{VIF}_{\Sigma}\uparrow$	$S_{\Sigma}\uparrow$	$Q_{abf}\uparrow$	$\mathrm{VIF}_{\Sigma}\uparrow$	$S_{\Sigma}\uparrow$	$Q_{abf}\uparrow$	$\mathrm{VIF}_{\Sigma}\uparrow$	$S_{\Sigma}\uparrow$	$Q_{abf}\uparrow$	$\mathrm{VIF}_{\Sigma}\uparrow$	$S_{\Sigma}\uparrow$
DIDFuse	Bicubic	0.120	0.335	0.860	0.079	0.213	0.706	0.257	0.597	1.214	0.494	0.610	1.206	0.308	0.525	1.130
SeAFusion	HAT	0.372	0.719	1.251	0.260	0.436	1.167	0.270	0.480	1.306	0.497	0.628	1.237	0.361	0.583	1.212
CDDFuse	HAT	0.396	0.749	1.256	0.235	0.421	1.157	0.241	0.403	1.288	0.498	0.601	1.259	0.349	0.534	1.225
AGMFusion	Bicubic	0.366	0.791	1.224	0.246	0.537	1.154	0.281	0.613	1.279	0.423	0.565	1.215	0.337	0.572	1.198
MFDFuse	Bicubic	0.384	0.816	1.260	0.202	0.408	1.164	0.255	0.532	1.279	0.509	0.684	1.236	0.348	0.637	1.190
MLFuse	Bicubic	0.321	0.614	1.247	0.209	0.434	1.175	0.290	0.574	1.354	0.485	0.595	1.307	0.344	0.552	1.257
TDFusion	LIIF	0.388	0.759	1.282	0.286	0.580	1.207	0.327	0.686	1.362	0.565	0.752	1.280	0.412	0.672	1.258
Deno-IF	Bicubic	0.340	0.720	1.274	0.196	0.460	1.181	0.255	0.572	1.365	0.454	0.597	1.298	0.288	0.509	1.256
TG-ECNet	Bicubic	0.245	0.427	1.150	0.131	0.189	1.044	0.168	0.287	1.243	0.167	0.202	1.087	0.184	0.280	1.135
RIS-Fuse	Bicubic	0.357	0.629	1.268	0.216	0.406	1.164	0.186	0.237	1.301	0.383	0.325	1.265	0.268	0.346	1.223
LUT-Fuse	Bicubic	0.355	0.728	1.262	0.205	0.451	1.160	0.186	0.375	1.307	0.424	0.544	1.296	0.304	0.492	1.233
TITA	Bicubic	0.388	0.786	1.272	0.188	0.375	1.207	0.235	0.470	1.366	0.529	0.690	1.358	0.342	0.570	1.286
MaeFuse	Bicubic	0.249	0.550	1.249	0.199	0.451	1.209	0.208	0.476	1.368	0.480	0.568	1.333	0.308	0.514	1.290
FDFuse	Bicubic	0.405	0.850	1.244	0.277	0.580	1.148	0.290	0.544	1.320	0.537	0.667	1.266	0.376	0.593	1.222
MFS-Fusion	LIIF	0.416	0.843	1.272	0.273	0.501	1.210	0.311	0.619	1.356	0.615	0.846	1.331	0.436	0.721	1.291
SETFusion	Bicubic	0.215	0.504	1.242	0.130	0.329	1.174	0.215	0.537	1.273	0.366	0.422	1.176	0.288	0.491	1.178
GPF (Ours)	Native	**0.440**	**1.012**	**1.318**	**0.353**	**0.983**	**1.338**	**0.376**	**0.960**	**1.420**	**0.656**	**0.993**	**1.380**	**0.492**	**0.938**	**1.321**

C.1 QUALITATIVE COMPARISONS

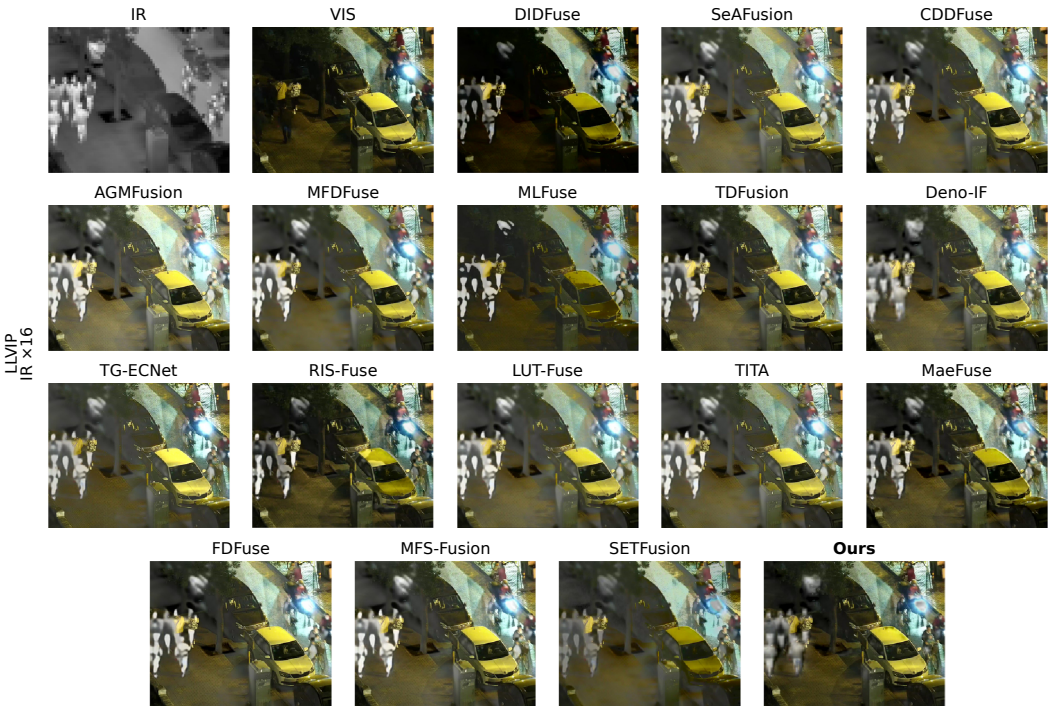

Figure C.1: **Scene 1: infrared** $\times 16$**.** Comparison of 17 fusion methods.

C.1 QUALITATIVE COMPARISONS (CONTINUED)

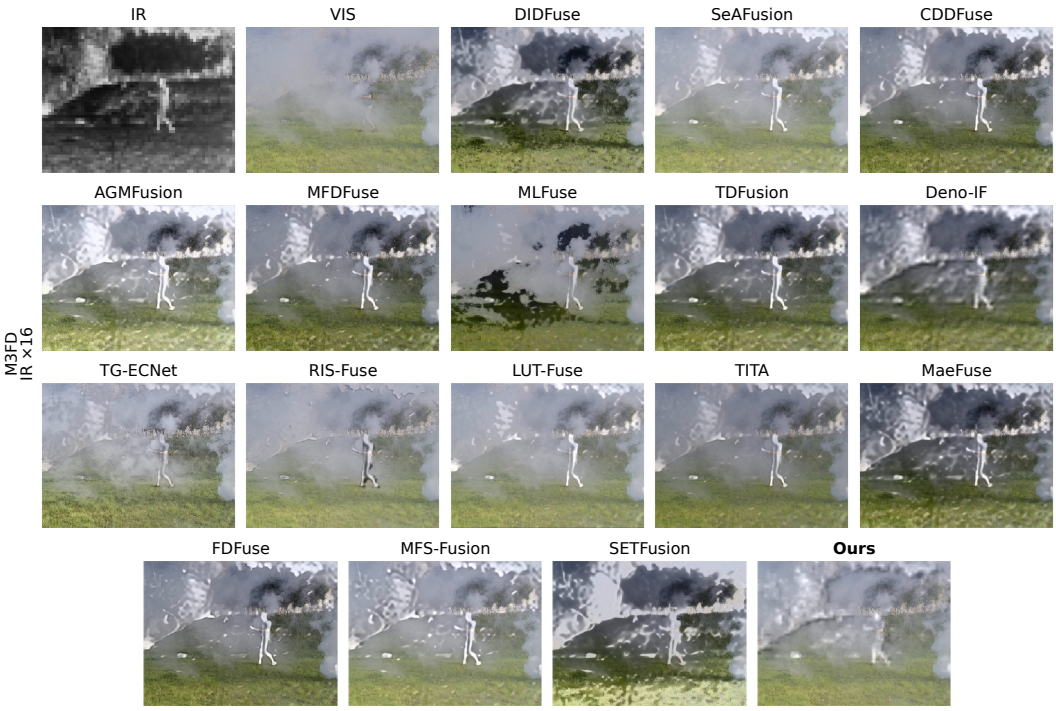

Figure C.2: **Scene 2: infrared** $\times 16$**.** Comparison of 17 fusion methods.

C.1 QUALITATIVE COMPARISONS (CONTINUED)

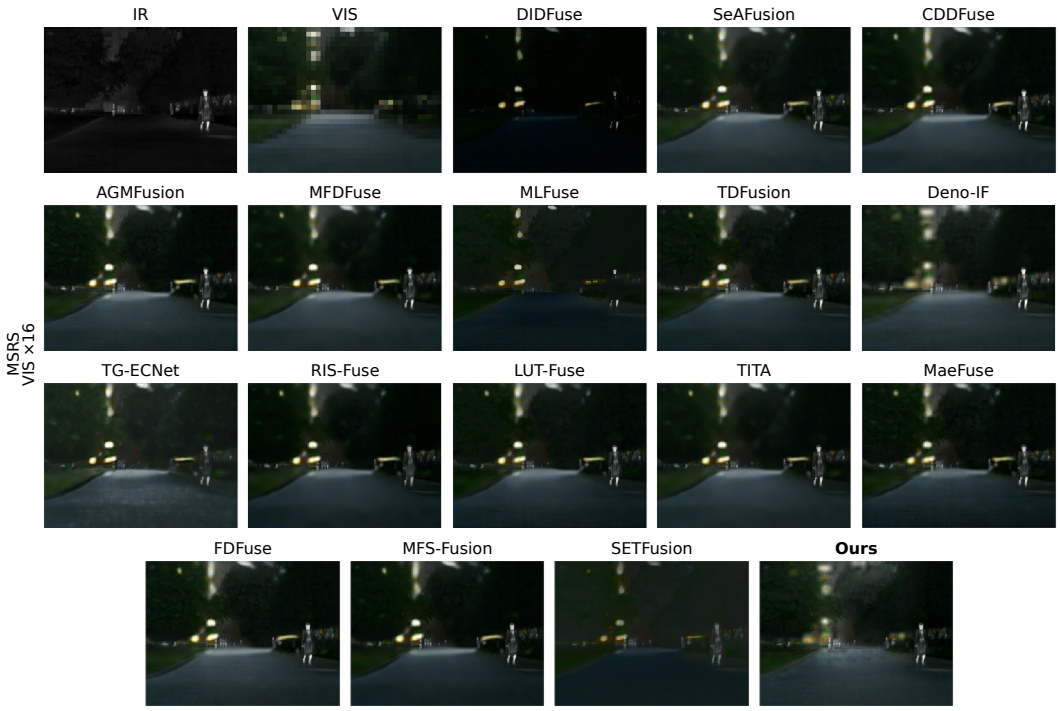

Figure C.3: **Scene 3: visible** $\times 16$**.** Comparison of 17 fusion methods.

C.1 QUALITATIVE COMPARISONS (CONTINUED)

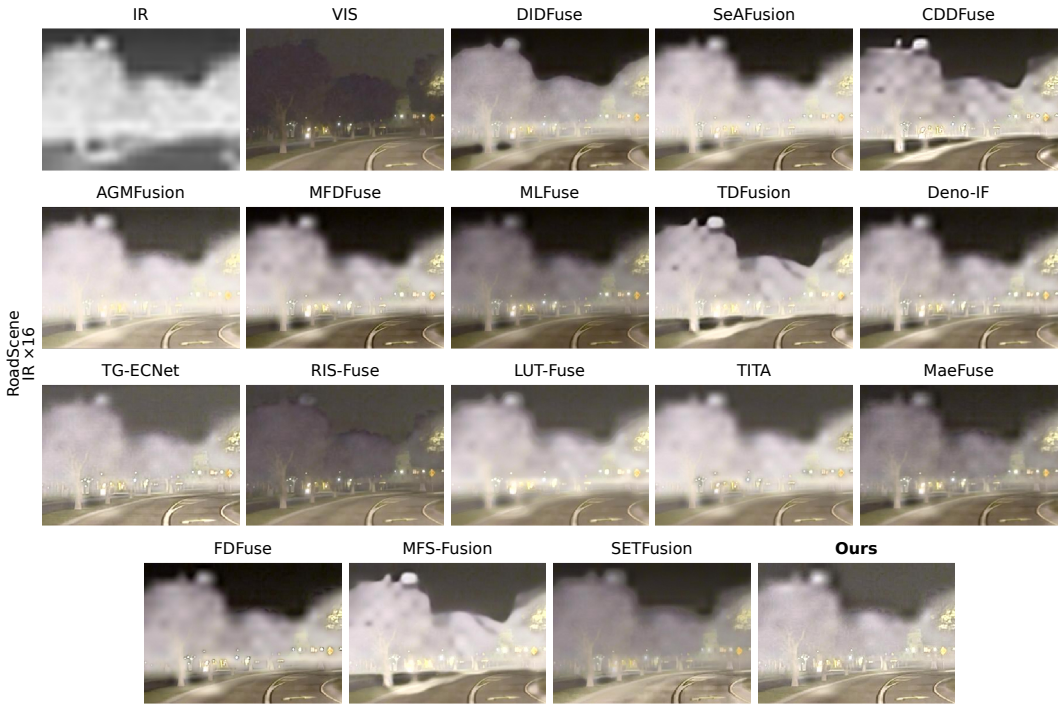

Figure C.4: **Scene 4: infrared** $\times 16$**.** Comparison of 17 fusion methods.

C.1 QUALITATIVE COMPARISONS (CONTINUED)

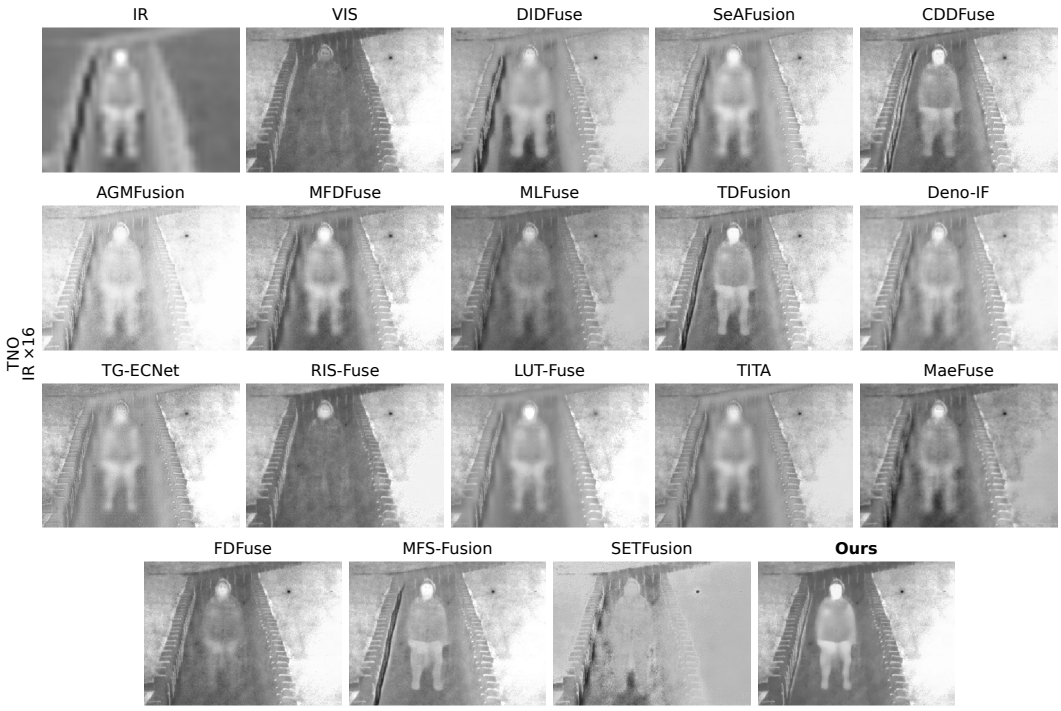

Figure C.5: **Scene 5: infrared** $\times 16$**.** Comparison of 17 fusion methods.

C.1 QUALITATIVE COMPARISONS (CONTINUED)

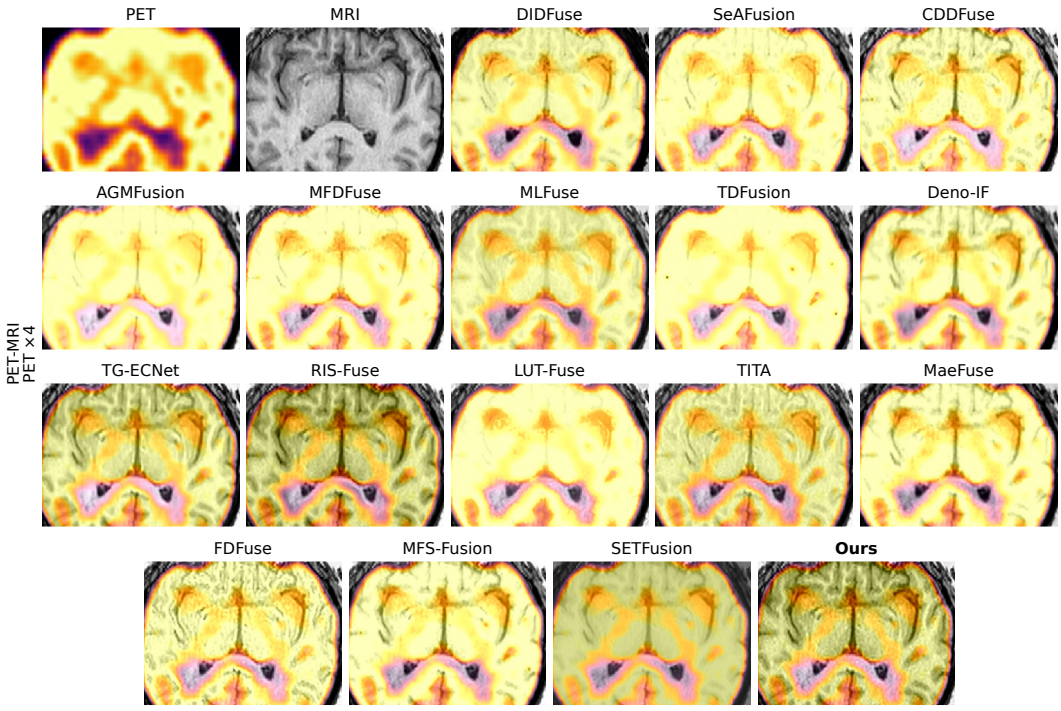

Figure C.6: **Scene 6: PET–MRI** ($\times 4$)**.** Comparison of 17 fusion methods.