This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

MagicFuse: Single Image Fusion for Visual and Semantic Reinforcement

Hao Zhang * Yanping Zha * Zizhuo Li Meiqi Gong Jiayi Ma †
Wuhan University, China
{zhpersonalbox, yanpingcha66, jyma2010}@gmail.com, {zizhuo_li, meiqigong}@whu.edu.cn

Abstract

This paper focuses on a highly practical scenario: how to continue benefiting from the advantages of multi-modal image fusion under harsh conditions when only visible imaging sensors are available. To achieve this goal, we propose a novel concept of single image fusion, which extends conventional data-level fusion to the knowledge level. Specifically, we develop MagicFuse, a novel single image fusion framework capable of deriving a comprehensive cross-spectral scene representation from a single low-quality visible image. MagicFuse first introduces an intra-spectral knowledge reinforcement branch and a cross-spectral knowledge generation branch based on the diffusion models. They mine scene information obscured in the visible spectrum and learn thermal radiation distribution patterns transferred to the infrared spectrum, respectively. Building on them, we design a multi-domain knowledge fusion branch that integrates the probabilistic noise from the diffusion streams of these two branches, from which a cross-spectral scene representation can be obtained through successive sampling. Then, we impose both visual and semantic constraints to ensure that this scene representation can satisfy human observation while supporting downstream semantic decision-making. Extensive experiments show that our MagicFuse achieves visual and semantic representation performance comparable to or even better than state-of-the-art fusion methods with multi-modal inputs, despite relying solely on a single degraded visible image. The code is publicly available at https://github.com/zhayanping/MagicFuse.

1. Introduction

Visible imaging sensors struggle to comprehensively capture scene content under harsh conditions, often yielding degraded appearances, *e.g.*, low-light, hazy, or noised. In this context, the infrared and visible image fusion (IVIF) technique has emerged. It leverages the complementary information from infrared modality to compensate for visible ones, enhancing the completeness of scene representation for human observation and machine perception [15, 19, 21, 24]. Owing to its excellent characteristics, this technique has been widely applied in various human-/machine-centered perception systems, *e.g.*, individual reconnaissance, intelligent transportation, and assisted driving.

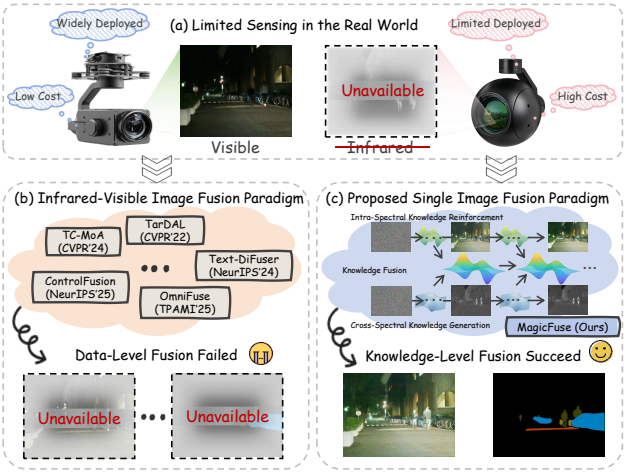


Figure 1. Under limited sensing conditions, existing image fusion methods fail completely, whereas our MagicFuse still achieves promising cross-spectral scene representation.

In recent years, deep learning has greatly advanced IVIF technology, continuously pushing its adaptability to complex real-world environments to new heights. Typically, TarDAL [8] employs a detection-driven dual adversarial learning architecture to produce salient target-aware fused images, which are particularly effective in enhancing the accuracy of semantic decision tasks under complex environments. OmniFuse [23] employs the diffusion model to perform degradation removal and information fusion in the latent space, enabling visually clear results even under challenging environments with composite degradations. Cleverly, Text-DiFuse [22] introduces a text-modulated robust image fusion model. Guided by user instructions, it enhances both the visual quality and semantic expressiveness of fused images under harsh imaging conditions.

*Equal Contribution
†Corresponding author

Admittedly, these IVIF methods offer attractive advantages. They generally rely on a strict prerequisite: paired visible and infrared images must be simultaneously available [7, 14, 18]. However, a noteworthy reality is that in most real-world scenarios, only visible imaging sensors are available, while thermal infrared sensors are absent due to their high cost, as shown in Fig. 1 (a). *Such a data-level absence fundamentally invalidates all existing IVIF methods*, as shown in Fig. 1 (b). Although image restoration techniques [1, 10] can be employed to recover as much scene information as possible from the visible spectrum, the prior knowledge within a single spectrum remains limited. When degradation types are diverse and coupled, existing restoration methods often fail to achieve satisfactory results.

It is natural to pose the following bold idea: *could we continue to enjoy the benefits of IVIF in harsh environments with only visible imaging sensors*? This is highly attractive, as it allows for a comprehensive cross-modal scene representation from a single low-quality visible image alone. This motivates our investigation into a previously unexamined research direction: **Single Image Fusion** (SIF). Before proceeding, we must raise two fundamental questions of SIF. (1) To overcome the representational limits of visible images, where does the new knowledge beyond the visible spectrum originate? (2) Since data-level fusion is infeasible due to the absence of infrared data, how can we achieve a leap from data-level to knowledge-level fusion?

Driven by these two fundamental questions, we propose MagicFuse, a single image fusion framework that infers a joint visible–infrared scene representation from a single low-quality visible image, enhancing both visual quality and semantic decision accuracy, as shown in Fig. 1 (c). First, powerful generative models may help us address the first question, as they are capable of producing new knowledge through learning from large-scale data. Accordingly, we design an intra-spectral knowledge reinforcement (IKR) branch and a cross-spectral knowledge generation (CKG) branch based on the diffusion model. IKR is dedicated to enhancing scene information from low-quality visible images, with the goal of maximizing knowledge mining within the visible spectrum. In contrast, CKG aims to learn thermal radiation distribution patterns from large-scale paired multi-modal images, thereby generating new knowledge in the infrared spectrum from the visible one. Second, to answer the second question, we consider that the estimated noise at each timestep in the diffusion sampling process inherently encodes valuable knowledge, which can be exploited as a medium for knowledge fusion. Therefore, we develop a multi-domain knowledge fusion (MKF) branch to integrate the probabilistic noise from diffusion streams of IKR and CKG branches. By performing continuous sampling from a shared starting point (*i.e.*, standard Gaussian noise), the MKF branch finally generates a visual image with cross-spectral representation capability, called Magic Image (**MagImg**). To further enhance the compatibility of MagImg for downstream semantic decision tasks, we introduce an auxiliary segmentation head embedded into the MKF branch. It aligns the fused features with semantic labels, enriching MagImg with refined structural semantics. Therefore, our MagicFuse constructs enhanced cross-spectral scene representations from a single low-quality visible image by mining new knowledge and performing knowledge fusion, improving both visual and semantic perception of degraded scenarios.

In summary, our contributions are as follows:

- We propose MagicFuse, a novel single image fusion framework capable of deriving a comprehensive cross-spectral scene representation from a single low-quality visible image. To our knowledge, this is the first proposal of SIF, offering significant potential to broaden the practical scope of image fusion.
- We leverage IKR and CKG for intra-spectral knowledge reinforcement and cross-spectral knowledge generation, and integrate them with MKF to bridge data-level to knowledge-level fusion, thereby addressing two fundamental questions of SIF.
- We construct visually and semantically coupled scene representation constraints, enabling our MagicFuse to simultaneously preserve both visual and semantic perception capabilities for degraded scenarios.
- Extensive experiments confirm our MagicFuse's reliability and practicality, achieving highly competitive performance in both visual and semantic perception.

2. Methodology

2.1. Problem Formulation

Our MagicFuse aims to obtain a cross-spectral scene representation through knowledge fusion using only single degraded visible images. The overall framework is presented in Fig. 2. Two types of knowledge are considered. Intra-spectral knowledge captures color, texture, and other details obscured by degradations, reflecting inherent visible-spectrum characteristics. We adopt image restoration as its task carrier, leveraging a diffusion model to recover weakened or lost information, thereby reinforcing the visible-spectral knowledge. Cross-spectral knowledge learns thermal radiation patterns from the visible spectrum, inferring object-level radiation intensity and energy distribution. This is achieved via visible-to-infrared translation, where a diffusion model progressively infers latent infrared representations, generating new cross-spectral knowledge. Thus, the SIF problem can be formulated as a knowledge fusion task:

$$\begin{aligned} &\min_{\boldsymbol{\omega}^{\mathtt{f}}} \mathbb{E}_{\boldsymbol{z}_T\sim\mathcal{N}(0,\boldsymbol{I})}\left[\mathcal{L}^{\mathtt{f}}\left(\mathcal{F}(\boldsymbol{k}^{\psi},\boldsymbol{k}^{\phi};\boldsymbol{\omega}^{\mathtt{f}})\right)\right], \\ &s.t.\ \boldsymbol{k}^{\psi}\in\Psi(\boldsymbol{\mathcal{I}},\boldsymbol{z}_T;\boldsymbol{\omega}^{\psi}),\ \boldsymbol{k}^{\phi}\in\Phi(\boldsymbol{\mathcal{I}},\boldsymbol{z}_T;\boldsymbol{\omega}^{\phi}). \end{aligned} \tag{1}$$

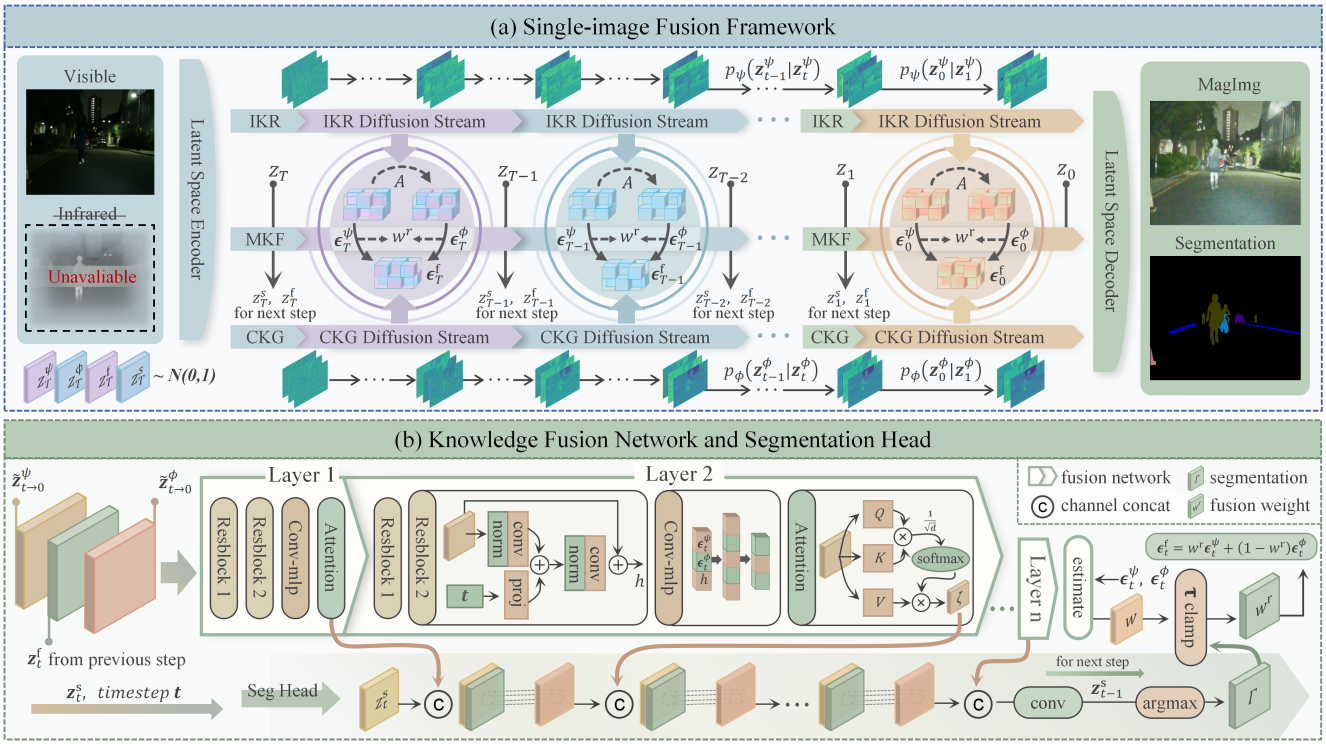


Figure 2. Overall framework of our proposed MagicFuse.

where $\mathcal{L}^{\mathtt{f}}$ denotes the fusion loss, and $\mathcal{F}(\cdot;\boldsymbol{\omega}^{\mathtt{f}})$ represents the knowledge fusion network in the MKF branch parameterized by $\boldsymbol{\omega}^{\mathtt{f}}$. $\boldsymbol{\mathcal{I}} \in \mathbb{R}^{H\times W\times 3}$ denotes the input degraded visible image. $\Psi(\cdot;\boldsymbol{\omega}^{\psi})$ and $\Phi(\cdot;\boldsymbol{\omega}^{\phi})$ indicate two diffusion streams for visible image restoration and visible-to-infrared image translation, with learnable parameters $\boldsymbol{\omega}^{\psi}$ and $\boldsymbol{\omega}^{\phi}$, respectively. $\boldsymbol{k}^{\psi}$ and $\boldsymbol{k}^{\phi}$ correspond to the above two types of knowledge produced by Ψ and Φ at different timesteps. For a total of T timesteps, $\boldsymbol{k}^{\psi} \in \{k_1^{\psi}, k_2^{\psi}, \ldots, k_T^{\psi}\}$ and $\boldsymbol{k}^{\phi} \in \{k_1^{\phi}, k_2^{\phi}, \ldots, k_T^{\phi}\}$. $\boldsymbol{z}_T$ is the noise sampled from a standard Gaussian distribution, serving as the shared starting point of the sampling process in Ψ and Φ.

By solving Eq. (1), we can obtain a pleasing MagImg that enables cross-spectral visual representation of degraded scenarios. However, recent studies [12] generally agree that image fusion should not only focus on visual quality for human perception but also ensure semantic attributes for machine understanding. To achieve this goal, we embed a segmentation head in the MKF branch to enforce alignment between the fused features and semantic labels, thereby enriching the semantic information in the knowledge fusion process. Consequently, Eq. (1) can be revised as follows:

$$\min_{\boldsymbol{\omega}^{\mathtt{f}},\boldsymbol{\omega}^{\mathtt{s}}} \left[\mathcal{L}^{\mathtt{f}}\left(\mathcal{F}(\boldsymbol{k}^{\psi},\boldsymbol{k}^{\phi};\boldsymbol{\omega}^{\mathtt{f}})\right) + \mathcal{L}^{\mathtt{s}}\left(\mathcal{S}(\boldsymbol{\zeta};\boldsymbol{\omega}^{\mathtt{s}})\right)\right], \quad (2)$$

where $\mathcal{S}(\cdot;\boldsymbol{\omega}^{\mathtt{s}})$ represents the segmentation head parameterized by $\boldsymbol{\omega}^{\mathtt{s}}$. $\boldsymbol{\zeta}$ denotes the attention features extracted from fusion network $\mathcal{F}$, which serve as inputs to the segmentation head $\mathcal{S}$, ensuring effective coupling of visual and semantic features. Solving Eq. (2) enables knowledge fusion, obtaining a cross-spectral scene representation that preserves both visual and semantic performance from a single low-quality visible image.

2.2. Knowledge Reinforcement and Generation

In the IKR and CKG branches, we employ the latent diffusion models (LDMs) Ψ and Φ to reinforce intra-spectral knowledge and generate cross-spectral knowledge. Note that Ψ and Φ share the same basic LDM design but differ in parameter capacity. This choice is motivated by the differing difficulty levels of the image restoration and visible-to-infrared image translation tasks. Below, we uniformly describe their architectures.

Autoencoder Mapping. We introduce lightweight autoencoders for Ψ and Φ, which incorporate InstanceNorm to guide the network in learning content-centric features, promoting a smooth and continuous manifold structure in the latent space. This resulting architecture facilitates more stable and efficient diffusion model training and better suits the requirements of image restoration and visible-to-infrared image translation tasks. Uniformly, the encoding and decoding processes are formalized as: $\boldsymbol{z} = \boldsymbol{\mathcal{E}}(\boldsymbol{\mathcal{I}}), \hat{\boldsymbol{\mathcal{I}}} = \boldsymbol{\mathcal{D}}(\boldsymbol{z})$.

Diffusion Estimation. We employ a latent diffusion model built upon a U-Net backbone. The model operates by grad-

ually corrupting the latent representation $\boldsymbol{z}_0$ of the target image into Gaussian noise over T timesteps (forward process) and then learning to reverse this process to reconstruct $\boldsymbol{z}_0$ from pure noise $\boldsymbol{z}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ (reverse process). Instead of modeling the complex data distribution directly, the network is trained to predict the added noise $\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t)$ via a simplified variational objective [4].

To adapt this framework for image restoration and visible-to-infrared translation, we introduce task-specific conditioning: the degraded visible image serves as the condition $\boldsymbol{c}$, concatenated with the noisy latent $\boldsymbol{z}_t$ along the channel dimension. Consequently, the denoiser predicts the noise component as $\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, \boldsymbol{c}, t)$. During inference, we adopt the accelerated deterministic sampling strategy from DDIM [11] to efficiently recover $\boldsymbol{z}_0$ from $\boldsymbol{z}_T$. Here, $\boldsymbol{z}_0$ denotes the latent representation of the clean visible or infrared image depending on the task.

Image restoration and visible-to-infrared translation exhibit fundamentally different learning objectives: the former focuses on recovering intra-spectral details, while the latter requires modeling cross-spectral semantic correspondence. To explicitly capture these complementary properties, we instantiate two specialized diffusion branches built upon the same generative framework. Specifically, the IKR and CKG branches are designed to learn *intra-spectral reinforced knowledge* ($\boldsymbol{k}^\psi$) and *cross-spectral generated knowledge* ($\boldsymbol{k}^\phi$), respectively. This is achieved by training two separate LDMs (Ψ and Φ) on paired degraded-clean and visible-infrared datasets. The acquired knowledge is implicitly embedded in the noise predicted during the denoising process: $\boldsymbol{\epsilon}_t^\psi = \boldsymbol{\epsilon}_\theta^\psi(\boldsymbol{z}_t^\psi, \boldsymbol{c}, t)$, $\boldsymbol{\epsilon}_t^\phi = \boldsymbol{\epsilon}_\theta^\phi(\boldsymbol{z}_t^\phi, \boldsymbol{c}, t)$.

2.3. Knowledge Fusion

To fuse knowledge $\boldsymbol{k}^\psi$ and $\boldsymbol{k}^\phi$, we construct an MKF branch aligned with the diffusion streams of the IKR and CKG branches. As shown in Fig. 2 (b), a knowledge fusion network $\mathcal{F}$ is designed to generate the weighting coefficients, combining the noise estimated by denoisers in Ψ and Φ:

$$\boldsymbol{\epsilon}_t^{\mathtt{f}} = \boldsymbol{w}\boldsymbol{\epsilon}_t^\psi + (1 - \boldsymbol{w})\boldsymbol{\epsilon}_t^\phi. \tag{3}$$

The rationality of the weighting coefficient is crucial, as it determines whether the knowledge $\boldsymbol{k}^\psi$ and $\boldsymbol{k}^\phi$ are balanced in the probabilistic space. We consider the generation of $\boldsymbol{w}$ from two perspectives. First, the knowledge quality of the diffusion streams in the IKR and CKG branches needs to be evaluated, which can be achieved by comparing their respective estimated representations $\widetilde{\boldsymbol{z}}_{t\to0}^\psi$ and $\widetilde{\boldsymbol{z}}_{t\to0}^\phi$. Second, multi-domain noise $\boldsymbol{\epsilon}_t^\psi$ and $\boldsymbol{\epsilon}_t^\phi$ should be analyzed, since they are the objects of weight aggregation. Besides, to satisfy the iterative nature of the diffusion stream, the current-step sampled representation $\boldsymbol{z}_t^{\mathtt{f}}$ in the MKF branch also needs to participate in the weighting estimation. Therefore, the computation of $\boldsymbol{w}$ in Eq. (3) can be formalized as:

$$\boldsymbol{w} = \mathcal{F}(\widetilde{\boldsymbol{z}}_{t\to0}^\psi, \widetilde{\boldsymbol{z}}_{t\to0}^\phi, \boldsymbol{z}_t^{\mathtt{f}}, \boldsymbol{\epsilon}_t^\psi, \boldsymbol{\epsilon}_t^\phi, t), \tag{4}$$

where both $\widetilde{\boldsymbol{z}}_{t\to0}^\psi$ and $\widetilde{\boldsymbol{z}}_{t\to0}^\phi$ are calculated according to $\widetilde{\boldsymbol{z}}_{t\to0} = \frac{\boldsymbol{z}_t - \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}_t}{\sqrt{\bar{\alpha}_t}}$. Notably, $\widetilde{\boldsymbol{z}}_{t\to0}$ is not the result of full sampling from the diffusion model, but rather a single-step deterministic estimate of the initial representation.

2.4. Visual-Semantic Reinforcement

Visual-Semantic Coupling. We expect the guidance of knowledge fusion to encompass not only the visual aspect but also the semantic one. Therefore, a segmentation head $\mathcal{S}$ is embedded in the MKF branch. As illustrated in Fig. 2 (b), in addition to the implicit representation $\boldsymbol{z}_t^s$ from the previous step for segmentation, $\mathcal{S}$ primarily takes the attention features $\boldsymbol{\zeta}$ from the knowledge fusion network as input:

$$\Gamma = \arg\max(\underbrace{\mathcal{S}(\boldsymbol{z}_t^s, \zeta_1, \zeta_2, \ldots, \zeta_l)}_{\boldsymbol{z}_{t-1}^s \text{ for next step}}), \tag{5}$$

where Γ is the predicted segmentation map, and ζ_l denotes the attention map from the l-th layer of $\mathcal{F}$. Such an operation is designed to enhance the semantic quality of visual features under the guidance of segmentation optimization. Its rationality can be evidenced from [6], as sufficiently general visual features are capable of supporting semantic decisions. In turn, a radiation category map $\mathcal{M}$ derived from Γ is used to further refine the weighting coefficients:

$$\boldsymbol{w}^{\mathtt{r}} = \mathcal{M}\min(\boldsymbol{w}, \boldsymbol{\tau}) + (1 - \mathcal{M})\,\boldsymbol{w}, \tag{6}$$

where $\mathcal{M}$ indicates typical thermal categories (*e.g.*, pedestrians, cars), used to enhance the preservation of thermal objects during the fusion process. $\boldsymbol{\tau}$ serves as a hyperparameter to regulate the impact of the radiation category map.

Notably, the above enforced modulation not only incorporates segmentation into the knowledge fusion process to foster visual–semantic coupling but also resolves a key optimization challenge in knowledge fusion. More concretely, since the estimated noise $\boldsymbol{\epsilon}_t^\psi$ and $\boldsymbol{\epsilon}_t^\phi$ are derived from the same input $\{\boldsymbol{\mathcal{I}}, \boldsymbol{z}_T\}$, an inherent correlation exists between them: $\boldsymbol{\epsilon}^\phi = A\,\boldsymbol{\epsilon}^\psi$, where A is a transformation matrix. Eq. (3) can be reformulated as $\boldsymbol{\epsilon}_t^{\mathtt{f}} = (\boldsymbol{w} + (1-\boldsymbol{w})A)\boldsymbol{\epsilon}_t^\psi$. At this time, given that the input conditions in both Ψ and Φ are the visible images, the optimization tends to be completely biased towards $\boldsymbol{\epsilon}_t^\psi$, which is manifested in $\boldsymbol{w} \to 1$, leading to training collapse. The enforced modulation based on $\mathcal{M}$ in Eq. (6) effectively mitigates this problem, facilitating rational knowledge fusion. Therefore, Eq. (3) can be rewrite as: $\boldsymbol{\epsilon}_t^{\mathtt{f}} = \boldsymbol{w}^{\mathtt{r}}\boldsymbol{\epsilon}_t^\psi + (1 - \boldsymbol{w}^{\mathtt{r}})\boldsymbol{\epsilon}_t^\phi$.

Visual and Semantic Derivation. In the MKF branch, for any $t \in \{T, \cdots, 1\}$, we first perform continuous iterative

sampling from timestep T to t:

$$\begin{aligned} \boldsymbol{z}_{T-1}^{\mathtt{f}} &= \sqrt{\bar{\alpha}_{T-1}}\widetilde{\boldsymbol{z}}_{T\to 0}^{\mathtt{f}} + \sqrt{1-\bar{\alpha}_{T-1}}\boldsymbol{\epsilon}_{T}^{\mathtt{f}} \\ &\cdots \\ \boldsymbol{z}_{t}^{\mathtt{f}} &= \sqrt{\bar{\alpha}_{t}}\widetilde{\boldsymbol{z}}_{t+1\to 0}^{\mathtt{f}} + \sqrt{1-\bar{\alpha}_{t}}\boldsymbol{\epsilon}_{t+1}^{\mathtt{f}}. \end{aligned} \tag{7}$$

Then, we can perform a single-step deterministic estimate of the initial representation from $\boldsymbol{z}_t^{\mathtt{f}}$:

$$\widetilde{\boldsymbol{z}}_{t\to 0}^{\mathtt{f}} = \frac{\boldsymbol{z}_t^{\mathtt{f}} - \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}_t^{\mathtt{f}}}{\sqrt{\bar{\alpha}_t}}. \tag{8}$$

Finally, the fused representation is mapped back to the image space via the decoder to yield the MagImg:

$$I_{\text{MImg}} = \boldsymbol{\mathcal{D}}(\widetilde{\boldsymbol{z}}_{t\to 0}^{\mathtt{f}}). \tag{9}$$

High-quality cross-spectral visual representation of I_{MImg} is ensured by considering its contrast, texture, and color:

$$\begin{cases} \mathcal{L}_{\text{cont}} = \|I_{\text{MImg}}^{y} - \max(\mathcal{D}(\widetilde{\boldsymbol{z}}_{t\to 0}^{\psi})^{y}, \mathcal{D}(\widetilde{\boldsymbol{z}}_{t\to 0}^{\phi})^{y})\|, \\ \mathcal{L}_{\text{text}} = \|\nabla I_{\text{MImg}}^{y} - \max(\nabla\mathcal{D}(\widetilde{\boldsymbol{z}}_{t\to 0}^{\psi})^{y}, \nabla\mathcal{D}(\widetilde{\boldsymbol{z}}_{t\to 0}^{\phi})^{y})\|, \\ \mathcal{L}_{\text{color}} = \|I_{\text{MImg}}^{cbcr} - \mathcal{D}(\widetilde{\boldsymbol{z}}_{t\to 0}^{\psi})^{cbcr}\|, \end{cases} \tag{10}$$

where y and $cbcr$ denote the illuminance and chrominance components, and ∇ is the Sobel gradient operator. Therefore, the total visual regularization is summarized as $\mathcal{L}_{\text{visual}} = \lambda_1\mathcal{L}_{\text{cont}} + \lambda_2\mathcal{L}_{\text{text}} + \lambda_3\mathcal{L}_{\text{color}}$, $\lambda_{1\sim 3}$ are used to adjust the contribution weights of these loss components.

According to Eq. (5), the output Γ of the segmentation head can be obtained. We constrain it using a cross-entropy loss to ensure the rationality of captured semantic attributes:

$$\mathcal{L}_{\text{seg}} = -\sum_{r}\sum_{p}\hat{\Gamma}_p(r)\log\big(\Gamma_p(r)\big), \tag{11}$$

where r and p iterate over all semantic categories and pixel positions, respectively; $\hat{\Gamma}_p(r) \in \{0,1\}$ denotes the ground truth label, and $\Gamma_p(r)$ represents the predicted probability for pixel p belonging to class r. Finally, the joint optimization of $\mathcal{L}_{\text{visual}}$ and $\mathcal{L}_{\text{seg}}$ drives a dual enhancement of both visual fidelity and semantic consistency.

2.5. Optimization Scheme

We adopt a two-phase optimization strategy. At the first phase, the latent diffusion models Ψ and Φ in the IKR and CKG branches are optimized independently for intra-spectral knowledge reinforcement and cross-spectral knowledge generation. At the second phase, we freeze the parameters of Ψ and Φ, and proceed to train the knowledge fusion network $\mathcal{F}$ along with the segmentation head $\mathcal{S}$. The training signals for this phase are derived from the outputs of the dual diffusion streams of Ψ and Φ at each timestep, as outlined in Algorithm 1.

Algorithm 1 Optimization Scheme of our MagicFuse

Require: Training dataset $\mathcal{R}$; Diffusion steps T; Diffusion model Ψ in IKR; Diffusion model Φ in CKG; Autoencoder $\{\boldsymbol{\mathcal{E}}, \boldsymbol{\mathcal{D}}\}$; Knowledge fusion network $\mathcal{F}$; Segmentation head $\mathcal{S}$
Ensure: Optimized parameters $\boldsymbol{\omega}^{\mathtt{f}}, \boldsymbol{\omega}^{\mathtt{s}}$

```
Initialize ω^f, ω^s randomly
for epoch = 1 to N do
    for (I, Î) ∈ R do                          ▷ Î is segmentation label
        Sample z_T^{ψ,φ,f} ~ N(0, I)           ▷ z_T^ψ = z_T^φ = z_T^f
        Sample z_T^s ~ N(0, I)                 ▷ ∈ R^{B×N_class×H×W}
        Select i from [0, T] randomly
        for t = T downto i do                  ▷ Reverse process
            z_t^ψ, ε_t^ψ, z̃_{t→0}^ψ ← Ψ(I, z_t^ψ, t)
            z_t^φ, ε_t^φ, z̃_{t→0}^φ ← Φ(I, z_t^φ, t)
            Enable gradient for F(·), S(·) when t = i
            w^r ← F(z̃_{t→0}^ψ, z̃_{t→0}^φ, z_t^f, ε_t^ψ, ε_t^φ, t)
            z_t^s, Γ ← S(z_t^s, ζ)             ▷ ζ from F(·)
            ε_t^f ← ε_t^ψ, ε_t^φ, w^r
            z_t^f ← p(z_{t-1}^f | z_t^f, z̃_{t→0}^f, ε_t^f)
        end for
        {I_ψ, I_φ, I_MImg} ← D({z̃_{t→0}^ψ, z̃_{t→0}^φ, z̃_{t→0}^f})
        update ω^f, ω^s by L^f(I_ψ, I_φ, I_MImg), L^s(z_t^s, Î)
    end for
end for
```

Table 1. Implementation Details of MagicFuse. Steps(T/I) denotes the number of time steps for training/inference.

Module	Params(M)	FLOPs(G)	Optimizer	LR	BS	Steps(T/I)
$\boldsymbol{\Psi}$	39.65	50.74	Adam	1e-4	16	1000/25
$\boldsymbol{\Phi}$	517.67	773.7	Adam	1e-5	4	1000/25
$\boldsymbol{\mathcal{F}}$&$\boldsymbol{\mathcal{S}}$	2.74	83.60	SGD	1e-3	1	25/25

3. Experiments

3.1. Configuration

Datasets. Ψ and Φ in the IKR and CKG branches are trained on the combined set composed of MFNet [3], FMB [9], and LLVIP [5]. Specifically, we use $14,190$ degraded-clean visible image pairs to train Ψ for intra-spectral knowledge reinforcement, while $25,186$ degraded visible-clean infrared image pairs are used to train Φ for cross-spectral knowledge generation. Subsequently, we use $1,177$ degraded visible images and their corresponding segmentation labels from the MFNet [3] dataset for training $\mathcal{F}$ and $\mathcal{S}$ in the MKF branch. The final testing is performed on 392 degraded visible images from the MFNet [3] dataset.

Implementation Details. Ψ, Φ, $\mathcal{F}$, and $\mathcal{S}$ are all trained on four NVIDIA GeForce RTX 3090 GPUs with 24GB memory, and the AMD EPYC 7H12 64-core processor CPU. The training details are summarized in Table 1.

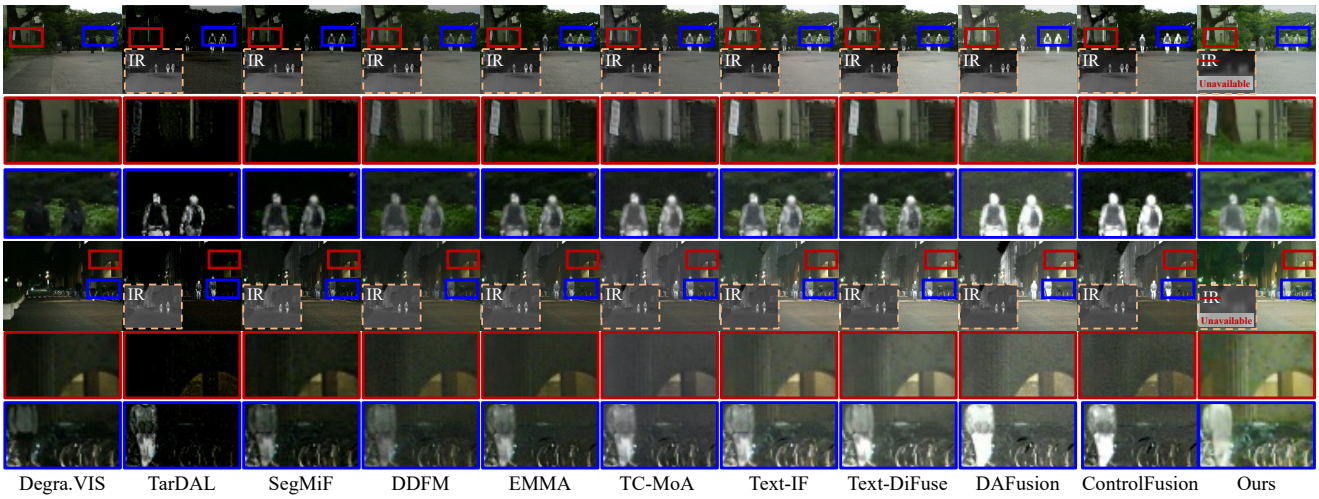


Figure 3. Qualitative visual representation comparison.

Table 2. Quantitative visual comparison. All competitors use infrared-visible pairs, while ours uses only single visible images.

Visual	TarDAL	SegMiF	DDFM	EMMA	TC-MoA	Text-IF	Text-DiFuse	DAFusion	ControlFusion	Ours
EN ↑	5.11	6.40	6.59	6.74	6.63	6.96	7.08	7.26	7.00	**7.29**
SSIM ↑	0.12	0.29	0.36	0.43	0.49	**0.52**	0.41	0.42	0.43	0.45
MI ↑	1.45	2.69	2.68	3.46	2.94	3.31	2.99	3.08	3.38	**4.13**
Qabf ↑	0.21	0.38	0.31	0.45	0.47	**0.59**	0.40	0.33	0.45	0.50
PSNR ↑	57.48	59.45	62.15	61.96	63.03	63.38	62.99	61.71	61.68	**63.49**

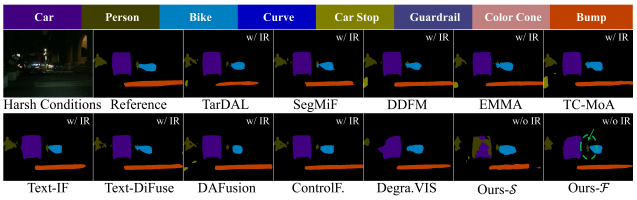


Figure 4. Qualitative semantic representation comparison.

3.2. Visual Representation Comparison

We first evaluate the effectiveness of knowledge fusion for cross-spectral visual representation. Nine state-of-the-art fusion methods are selected to compare with our MagicFuse, including TarDAL [8], SegMiF [9], DDFM [25], EMMA [26], TC-MoA [27], Text-IF [20], Text-DiFuse [22], DAFusion [16], and ControlFusion [13]. Notably, our MagicFuse produces the MagImg from a ***single degraded visible image***, whereas the competitors rely on ***multi-modal inputs***. Fig. 3 shows that our MagicFuse effectively integrates intra-spectral reinforced knowledge and cross-spectral generated knowledge. This is reflected in its ability not only to enhance the inherent scene information contained in degraded visible images, but also to reasonably highlight important objects in the scene based on the learned thermal radiation distribution. Table 2 further confirms its highly competitive performance, suggesting that *even with only a single degraded visible image, it can match or surpass multi-modal image fusion methods*.

Table 3. Quantitative semantic comparison. All competitors use infrared-visible pairs, while ours uses only single visible images.

Semantic	Car	Person	Bike	Curve	Stop	Guar.	Cone	Bump	mIoU ↑
TarDAL	85.19	69.39	70.49	15.89	45.15	34.83	51.07	32.13	55.73
SegMiF	86.75	67.61	72.41	25.95	46.88	52.13	56.79	54.32	**62.28**
DDFM	85.80	69.33	67.93	27.40	42.63	17.55	59.12	41.09	56.49
EMMA	86.03	70.42	71.41	24.73	51.17	50.05	56.60	49.96	61.98
TC-MoA	86.99	69.88	72.43	27.22	41.31	35.29	52.49	46.56	58.87
Text-IF	86.25	69.01	72.05	28.03	46.76	43.29	58.79	44.00	60.65
Text-DiFu.	83.24	68.84	70.93	25.06	44.13	43.25	56.56	49.91	59.93
DAFusion	85.94	70.41	72.48	24.23	40.85	50.46	56.82	37.08	59.54
ControlF.	86.99	70.50	71.95	25.66	42.69	29.32	53.77	57.39	59.55
Degra. VIS	81.55	56.35	70.89	30.82	40.03	3.89	61.47	44.51	54.09
Ours-$\mathcal{S}$	75.93	59.25	55.83	41.24	45.77	59.60	44.34	35.33	57.11
Ours-$\mathcal{F}$	83.80	61.24	70.62	35.28	54.09	49.42	62.69	45.09	62.19

3.3. Semantic Representation Comparison

Next, we evaluate the semantic representation capability of our method. We retrain SegFormer [17] separately on low-quality visible images, IKR-enhanced images, fused images from comparative methods, and MagImgs from MagicFuse, to measure their contained semantic quality. Besides, we also include the auxiliary segmentation head's output for comparison. As shown in Fig. 4, under harsh conditions, SegFormer fails to detect distant pedestrians in degraded visible images. In contrast, most fusion methods enable SegFormer to detect previously missed pedestrians by leveraging complementary infrared information. Remarkably, MagicFuse achieves the same without any infrared input, using its knowledge-level understanding that pedestrians exhibit strong thermal signatures, allowing ef-

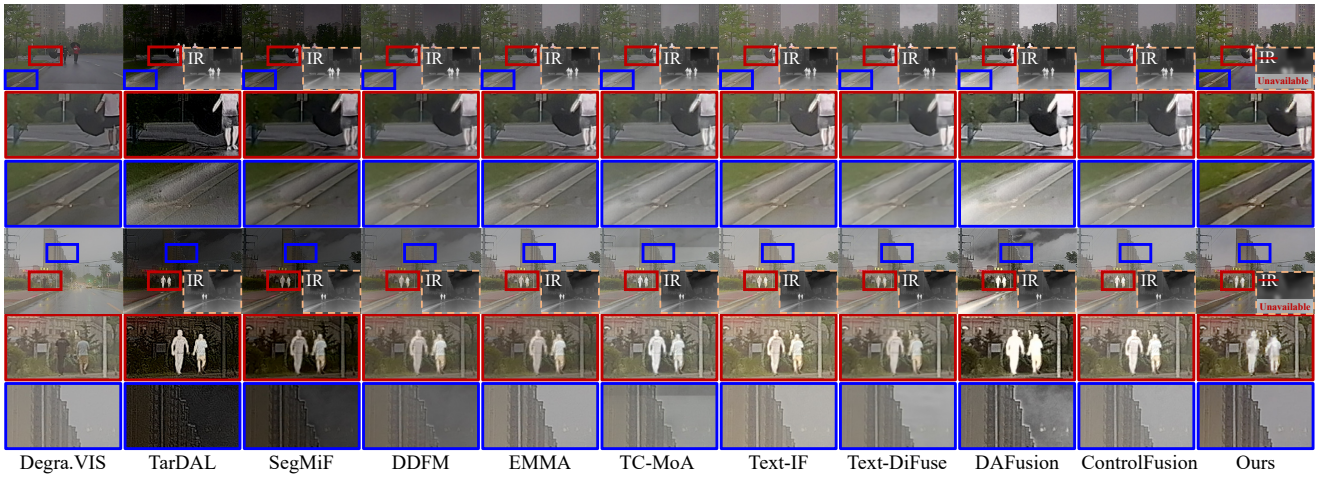


Figure 5. Qualitative visual representation generalization.

Table 4. Quantitative visual generalization. All competitors use infrared-visible pairs, while ours uses only single visible images.

Visual	TarDAL	SegMiF	DDFM	EMMA	TC-MoA	Text-IF	Text-DiFuse	DAFusion	ControlFusion	Ours
EN↑	6.44	6.82	6.71	6.80	6.76	6.68	6.93	**7.47**	6.72	6.93
SSIM↑	0.19	0.39	0.46	0.48	0.46	0.44	0.35	0.45	**0.48**	0.40
MI↑	2.70	3.59	3.68	3.99	3.07	3.71	3.25	4.05	**4.23**	4.09
Qabf↑	0.34	0.60	0.41	0.66	0.68	0.68	0.47	0.56	**0.70**	0.57
PSNR↑	60.10	61.99	**65.37**	62.25	62.13	60.37	61.12	60.80	61.47	62.51

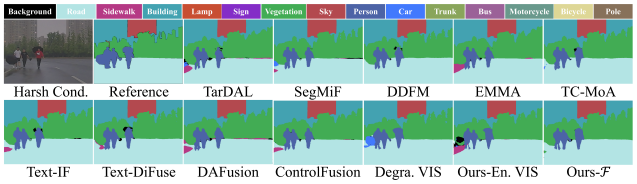


Figure 6. Qualitative semantic representation generalization.

Table 5. Quantitative semantic generalization. All competitors use infrared-visible pairs, while ours uses only single visible images.

Segmentation	Car	Person	Truck	Lamp	Sign	Buil.	Vege.	Road	mIoU↑
TarDAL	77.43	66.22	31.01	39.76	71.17	80.31	84.64	86.54	53.94
SegMiF	81.94	70.08	49.89	44.52	72.20	82.64	85.89	88.21	57.58
DDFM	78.60	65.81	43.18	36.72	73.53	82.54	85.73	88.07	56.57
EMMA	80.71	68.31	35.29	42.66	72.84	81.29	84.45	86.53	56.89
TC-MoA	79.92	66.78	49.28	41.83	71.08	81.13	85.70	88.44	**57.72**
Text-IF	81.66	66.73	22.63	44.14	71.20	80.37	85.23	88.87	56.10
Text-DiFu.	81.23	67.29	14.45	42.36	73.60	81.94	84.91	87.87	56.48
DAFusion	80.96	66.72	22.11	37.38	72.95	80.66	85.88	88.67	53.61
ControlF.	81.49	68.90	33.27	45.49	74.63	81.95	85.49	88.77	56.62
Degra. VIS	80.72	56.25	25.89	28.80	71.11	79.00	84.58	89.00	53.39
Ours-En. VIS	80.73	58.41	33.68	36.41	73.32	79.21	84.4	86.39	54.42
Ours-$\mathcal{F}$	81.63	58.84	34.61	36.71	72.50	80.03	84.67	87.99	55.41

fective emphasis in the MagImg. Table 3 shows that our method's average segmentation score surpasses most multi-modal fusion methods, second only to SegMiF.

3.4. Generalization of Knowledge Fusion

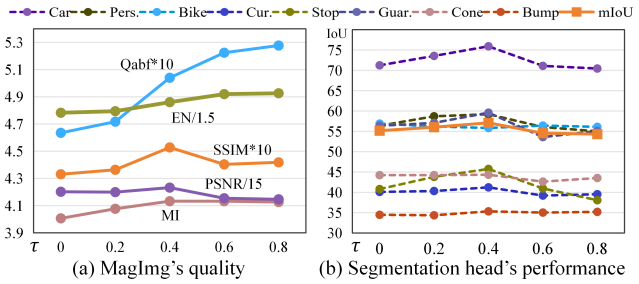
During the learning of knowledge fusion, MagicFuse is exposed only to the MFNet [3]. To verify the generalization of knowledge fusion, we directly apply the trained MagicFuse to the FMB dataset [9], containing 280 test images, to assess its visual and semantic representation performance.

Visual Representation Generalization. Fig. 5 presents qualitative results of the visual generalization. MagicFuse's output exhibits both the intra-spectral reinforced and the cross-spectral generated knowledge. Compared to degraded visible images, MagicFuse enhances scene structures, such as lane markings and building textures, while highlighting key objects like pedestrians via learned thermal patterns. Quantitative results in Table 4 show that MagicFuse achieves the second-best scores on most metrics, indicating that knowledge fusion enables our method to transfer easily to unseen data, providing reliable visual representations.

Semantic Representation Generalization. Fig. 6 shows the qualitative results for semantic representation generalization. The segmentation map derived from the MagImg generated by our MagicFuse is highly consistent with the reference. We also provide the quantitative results in Table 5. Although our method does not surpass fusion methods with multi-modal inputs, it still significantly outperforms both the original degraded and the enhanced visible images. These results indicate that semantic attributes propagated through the knowledge fusion process remain effective on unseen data, highlighting MagicFuse's potential for enhancing machine perception in extreme environments.

3.5. Ablation Studies

Effects of Hyperparameter τ. The hyperparameter τ is to regulate the influence of the radiation category map on



(a) MagImg's quality (b) Segmentation head's performance

Figure 7. Quantitative effects of hyperparameter τ.

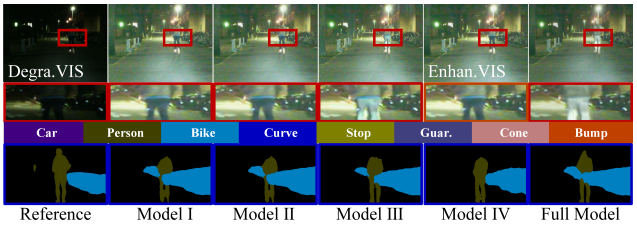


Figure 8. Quantitative results of ablations on key components.

knowledge fusion, which is also closely related to the degree of semantic attribute injection. We set τ to 0, 0.2, 0.4, 0.6, and 0.8, respectively, evaluating both changes of the MagImg's quality and the segmentation head's performance. The metric curves are presented in Fig. 7, where both MagImg's quality and segmentation head's accuracy reach their highest scores when $\tau = 0.4$. This indicates that while cross-spectral generated knowledge enhances scene representation, its influence should be regulated to prevent disruption of the original spectral information.

Ablations on Key Components. Four model variants are derived for the ablation studies. *Model I*: remove $\widetilde{\boldsymbol{z}}^{\psi}_{t\to 0}$ and $\widetilde{\boldsymbol{z}}^{\phi}_{t\to 0}$ from Eq. (4) when estimating the fusion weight. *Model II*: remove the segmentation head, relying solely on visual guidance for knowledge fusion. *Model III*: aggregating the outputs of the IKR and CKG branches after the diffusion process instead of fusing noise at each timestep. *Model IV*: achieve scene representation based on visible images enhanced by the IKR branch. Qualitative results in Fig. 8 show that Models I, II, and IV lack cross-spectral representation capability, while Model III provides only limited capability. With all components enabled, MagicFuse produces the most accurate pedestrian contours. Quantitative results in Tables 6 and 7 confirm that removing any component degrades both visual and semantic performance.

3.6. Beyond Fusion Scenarios

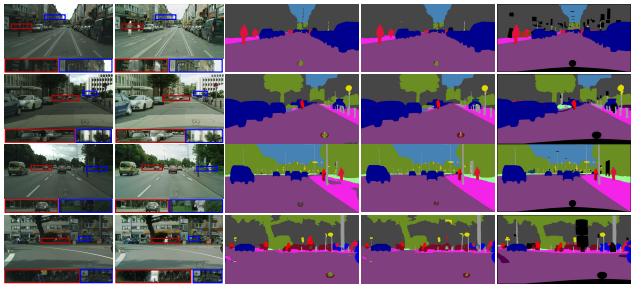
Previous experiments are implemented on the infrared-visible fusion datasets. Since MagicFuse requires only a single visible image, it can be applied to nearly any natural image. To demonstrate this, we evaluate it on the Cityscapes dataset [2]. Specifically, 490 visible images are enhanced

Table 6. Quantitative results of visual ablation studies.

Variants	EN↑	SSIM↑	MI↑	Qabf↑	PSNR↑
Model I	7.28	0.43	4.09	**0.53**	62.14
Model II	7.33	0.43	4.10	0.50	62.16
Model III	7.17	0.41	3.24	0.49	61.49
Model IV	**7.39**	0.44	4.08	0.50	62.11
Full Model	7.29	**0.45**	**4.13**	0.50	63.49

Table 7. Quantitative results of semantic ablation studies.

Variants	Car	Person	Bike	Curve	Car Stop	Guar.	Cone	Bump	mIoU↑
Model I	83.65	51.45	70.5	34.87	55.09	48.95	62.62	44.94	61.05
Model II	83.54	51.31	70.55	34.87	54.82	47.78	62.7	44.45	60.83
Model III	79.92	59.81	68.45	32.96	50.03	43.01	61.65	44.59	59.74
Model IV	85.99	69.06	70.79	25.49	42.98	7.15	56.61	58.93	57.18
Full Model	83.80	61.24	70.62	35.28	54.09	49.42	62.69	45.09	**62.19**



VIS MagImg VIS (mIoU: **65.80**) MagImg (mIoU: **66.91**) Reference

Figure 9. Applying our MagicFuse beyond fusion scenarios.

to generate MagImgs for retraining SegFormer, and evaluation is conducted on 154 MagImgs. As shown in Fig. 9, MagicFuse improves scene visibility and boosts semantic segmentation mIoU from 65.80 to 66.91, demonstrating its robust capability in enhancing scene representations.

4. Conclusion

This paper proposes MagicFuse, a single image fusion framework designed for robust scene representation under harsh conditions using only a degraded visible image. It uses an IKR branch to recover obscured visible information and a CKG branch to learn infrared-aligned thermal patterns. Then, an MKF branch is developed to integrate probabilistic noise from both branches into a unified diffusion stream, combining multi-domain knowledge. Guided by visual and semantic constraints, MagicFuse produces representations suitable for both human perception and semantic tasks. Experiments show that our method, relying solely on a single degraded visible image, achieves performance on par with or surpassing SOTA multi-modal fusion methods.

Acknowledgments

This work was supported by the NSFC under Grants 62506268 and 62276192, and by the CPSF under Grants GZB20250066 and 2025M781505.

References

[1] 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. 2

[2] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016. 8

[3] Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5108–5115, 2017. 5, 7

[4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 4

[5] 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. 5

[6] Chaehyun Kim, Heeseong Shin, Eunbeen Hong, Heeji Yoon, Anurag Arnab, Paul Hongsuck Seo, Sunghwan Hong, and Seungryong Kim. Seg4diff: Unveiling open-vocabulary segmentation in text-to-image diffusion transformers. *Advances in Neural Information Processing Systems*, 2025. 4

[7] 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. 2

[8] 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. 1, 6

[9] Jinyuan Liu, Zhu Liu, Guanyao Wu, Long Ma, Risheng Liu, Wei Zhong, Zhongxuan Luo, and Xin Fan. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8115–8124, 2023. 5, 6, 7

[10] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, and Jens Sjölund. Image restoration with mean-reverting stochastic differential equations. In *Proceedings of the International Conference on Machine Learning*, pages 23045–23066, 2023. 2

[11] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *Proceedings of the International Conference on Learning Representations*, pages 1–12, 2021. 4

[12] 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, 2022. 3

[13] 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*, 2025. 6

[14] 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*, 133:5262–5280, 2025. 2

[15] Linfeng Tang, Chunyu Li, and Jiayi Ma. Mask-difuser: A masked diffusion model for unified unsupervised image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(1):591–608, 2026. 1

[16] 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. 6

[17] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 6

[18] 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. 2

[19] Han Xu, Yuyang Li, Yunfei Deng, Jiayi Ma, and Guangcan Liu. Deno-if: Unsupervised noisy visible and infrared image fusion method. *Advances in Neural Information Processing Systems*, 2025. 1

[20] 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. 6

[21] 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. 1

[22] 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. 1, 6

[23] 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. 1

[24] Xingchen Zhang and Yiannis Demiris. Visible and infrared image fusion using deep learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10535–10554, 2023. 1

[25] Zixiang Zhao, Haowen Bai, Yuanzhi Zhu, Jiangshe Zhang, Shuang Xu, Yulun Zhang, Kai Zhang, Deyu Meng, Radu

Timofte, and Luc Van Gool. Ddfm: Denoising diffusion model for multi-modality image fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8082–8093, 2023. 6

[26] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Kai Zhang, Shuang Xu, Dongdong Chen, Radu Timofte, and Luc Van Gool. Equivariant multi-modality image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25912–25921, 2024. 6

[27] Pengfei Zhu, Yang Sun, Bing Cao, and Qinghua Hu. Task-customized mixture of adapters for general image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7099–7108, 2024. 6