Mask-DiFuser: A Masked Diffusion Model for Unified Unsupervised Image Fusion

Linfeng Tang, Chunyu Li, and Jiayi Ma, *Senior Member, IEEE*

***Abstract*—The absence of ground truth (GT) in most fusion tasks poses significant challenges for model optimization, evaluation, and generalization. Existing fusion methods achieving complementary context aggregation predominantly rely on hand-crafted fusion rules and sophisticated loss functions, which introduce subjectivity and often fail to adapt to complex real-world scenarios. To address this challenge, we propose Mask-DiFuser, a novel fusion paradigm that ingeniously transforms the unsupervised image fusion task into a dual masked image reconstruction task by incorporating masked image modeling with a diffusion model, overcoming various issues arising from the absence of GT. In particular, we devise a dual masking scheme to simulate complementary information and employ a diffusion model to restore source images from two masked inputs, thereby aggregating complementary contexts. A content encoder with an attention parallel feature mixer is deployed to extract and integrate complementary features, offering local content guidance. Moreover, a semantic encoder is developed to supply global context which is integrated into the diffusion model via a cross-attention mechanism. During inference, Mask-DiFuser begins with a Gaussian distribution and iteratively denoises it conditioned on multi-source images to directly generate fused images. The masked diffusion model, learning priors from high-quality natural images, ensures that fusion results align more closely with human visual perception. Extensive experiments on several fusion tasks, including infrared-visible, medical, multi-exposure, and multi-focus image fusion, demonstrate that Mask-DiFuser significantly outshines SOTA fusion alternatives.**

***Index Terms*—Image fusion, unified model, masked image modeling, diffusion model.**

I. Introduction

SINGLE-MODAL (or setting) sensors capture scene information from a specific band or perspective, rendering it challenging to describe scenarios comprehensively. Thus, multi-source fusion has become the new favorite in machine vision, with image fusion receiving significant attention. Image fusion aims to aggregate complementary information from multiple images, facilitating visual perception and semantic decision. Generally, image fusion mainly involves multi-modal and digital photography fusion. Infrared and visible image fusion (IVIF) and medical image fusion (MIF) are representative tasks of multi-modal image fusion (MMIF). In digital photography image fusion, two typical tasks are multi-exposure image fusion (MEF) and multi-focus image fusion (MFF). The effective information aggregation and visually appealing results of image fusion make it widely applied in both military and civilian applications, like military detection, smart healthcare, assisted driving [1], and scene understanding [2], [3].

Advancements in deep learning have significantly propelled the development of image fusion. A series of sophisticated network architectures, such as auto-encoder (AE), convolutional neural network (CNN), generative adversarial network (GAN), Transformer, and diffusion model, have been employed in image fusion tasks, achieving excellent performance. In the early stages, researchers approached different fusion tasks as independent problems and developed several well-known algorithms, such as DeepFuse [4], DenseFuse [5], FusionGAN [6], EMFusion [7], and SeAFusion [8]. DeepFuse is the first to introduce deep learning into multi-exposure image fusion. DenseFuse, FusionGAN, and SeAFusion are proposed for infrared and visible image fusion. Specifically, DenseFuse pioneers an AE-based framework, where deep learning is explicitly responsible for both feature extraction and image reconstruction. FusionGAN defines image fusion as an adversarial game between the generator and discriminator, marking the first application of GANs in this domain. SeAFusion is the first semantic-aware framework, which introduces semantic guidance into image fusion. EMFusion is an unsupervised medical image fusion method that improves information preservation and complementarity through shallow and deep constraints. In addition to task-specific methods, researchers extensively explored the commonalities among various fusion tasks and modeled them in a unified framework. Some representative unified image fusion methods include IFCNN [9], PMGI [10], U2Fusion [11], DeFusion [12], and SwinFusion [13].

It is worth emphasizing that authentic fused images are typically not available in most image fusion tasks [14], which significantly impedes the advancement of the image fusion field. Due to the absence of authoritative supervision signals, researchers face significant challenges in model training, performance evaluation, and the interpretability and generalization of methods, making it difficult for existing approaches to achieve stable application in complex and dynamic real-world scenarios. Therefore, the image fusion community proposes various solutions to mitigate this challenge. On the one hand,

Received 12 February 2025; revised 14 August 2025; accepted 9 September 2025. Date of publication 12 September 2025; date of current version 3 December 2025. This work was supported by the National Natural Science Foundation of China under Grant 62276192. Recommended for acceptance by B. Rosenhahn. *(Linfeng Tang and Chunyu Li contributed equally to this work.) (Corresponding author: Jiayi Ma.)*

The authors are with Electronic Information School, Wuhan University, Wuhan 430072, China (e-mail: linfeng0419@gmail.com; licy0089@gmail.com; jyma2010@gmail.com).

The source code is publicly available at https://github.com/Linfeng-Tang/Mask-DiFuser.

Digital Object Identifier 10.1109/TPAMI.2025.3609323

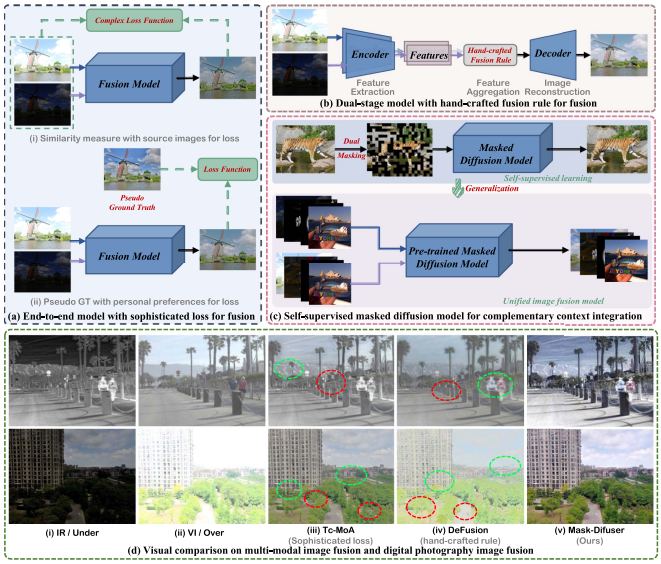


Fig. 1. Various solution workflow for overcoming the lack of ground truth in the field of image fusion and schematic illustration of different image fusion tasks.

researchers designed sophisticated loss functions [6], [13], [15], [16] to train an end-to-end fusion model that directly synthesizes fused images, as presented in Fig. 1(a). Initially, loss functions are constructed by measuring the similarity between fused and source images, typically relying on specific hand-crafted prior knowledge, which makes it challenging to adapt to complex and dynamic fusion scenarios. Furthermore, researchers attempted to manually synthesize ground truth (GT) for constructing loss functions [17], [18]. Nevertheless, manual GT construction is not only labor-intensive but also subject to personal preferences and is applicable only to specific fusion tasks, e.g., MEF and MFF. On the other hand, some approaches decompose image fusion into three processes, i.e., feature extraction, aggregation, and image reconstruction [5], [9], [19], as shown in Fig. 1(b). They usually train an auto-encoder via self-supervision for feature extraction and image reconstruction. Finally, hand-crafted fusion rules are utilized for feature aggregation. Although this solution alleviates the issue of lacking fusion GTs, the simplistic hand-crafted fusion rules, which rely on human-designed heuristics, lack a deep understanding of the intrinsic fusion patterns of the data itself, and thus show significant limitations when dealing with complex and changeable fusion scenarios. Additionally, due to the lack of GT and differing optimization objectives across various fusion tasks, most unified methods still necessitate training a specific model for each task [10], [13]. Although U2Fusion [11] and TC-MoA [20] achieve multi-task fusion in a single model through elastic weight consolidation and a mixture of expert mechanisms, respectively, they have yet to overcome the challenges associated with GT. As a consequence, the absence of GT has not only become a significant barrier to the advancement of the image fusion field, but also greatly limits the applicability and generalization of unified image fusion methods. To address this problem, we propose a novel perspective: all image fusion tasks can be unified under the goal of generating high-quality images with natural visual characteristics, such as clear textures, proper exposure, and high contrast, by effectively integrating complementary information from multiple source images to construct a complete scene representation. This goal aligns with human perceptual expectations and can serve as a unified standard for both multi-modal image fusion and digital photography image fusion.

In this work, we endeavor to address a series of longstanding challenges in image fusion arising from the absence of ground truth (GT), including the optimization process converging to locally suboptimal solutions due to dependence on manually defined prior-based loss functions, as well as the limited adaptability of models trained for specific tasks to other fusion scenarios. Considering that self-supervised learning can efficiently mine latent features and structural information from massive amounts of unlabeled data [21], we ingeniously apply Masked Image Modeling (MIM) [22] into the field of image fusion to transform the unsupervised image fusion task into a self-supervised dual-masked image reconstruction process with GTs. Specifically, we apply a dual masking scheme by masking the image twice to generate two complementary inputs, where multiple masks selectively combine high-quality image content, solid color patch, and degraded versions. As depicted in Fig. 1(c), this double masking technique is the key to our solution, as it not only simulates the precious complementary backgrounds found in actual fusion scenarios but also provides reliable and task-agnostic supervision signals. Consequently, the source images can serve as the true ground truth, providing a reliable basis for the optimization and evaluation of the complementary information aggregation process. To achieve this, we first develop a novel dual masking scheme. Furthermore, we employ a masked diffusion model as the fusion network to capture generative priors of informative reference images, ensuring visually pleasing fusion results. To maximize the potential of the diffusion model, we introduce the content and semantic conditional branches to provide comprehensive local and global guidance, respectively. The content branch equipped with an attention parallel feature mixer integrates and provides local complementary context. The semantic branch deploys Transformer to extract global context and integrate it into the diffusion model via cross-attention. During inference, the pre-trained masked diffusion model serves as a unified and versatile fusion model, adaptively aggregating complementary information from multiple sources without fine-tuning, effectively addressing the GT challenge in image fusion through our innovative design and the superiority of MIM. Our major contributions are summarized as follows:

- We propose a self-supervised paradigm for unified image fusion based on masked image modeling and diffusion models, termed Mask-DiFuser, which learns complementary information aggregation via masked image restoration, alleviating the issue of lacking GT in image fusion.
- A novel dual masking scheme embedding degradation factors is devised to simulate valuable complementary information and potentially enhance the robustness of fusion models to interference.
- A masked diffusion model integrating local content and global semantic contexts is developed to learn high-quality

priors from natural images through self-supervision, enabling fused images that align more closely with human visual perception.
- Extensive experiments demonstrate that our model excels in infrared-visible, medical, multi-exposure, and multi-focus image fusion tasks, particularly in color preservation, contrast enhancement, and exposure adaptation.

II. Related Work

A. Image Fusion

Recently, image fusion has achieved notable advancements. On the one hand, task-specific approaches now extend beyond enhancing visual quality of fusion results [7], [23], [24], [25] to integrate fusion with upstream/downstream tasks. Some practical schemes, including joint registration and fusion [26], [27], [28], [29], joint enhancement and fusion [30], [31], [32], and semantic-driven fusion methods [8], [15], [33], are proposed. On the other hand, unified methods are gradually proliferating to increase general applicability across various scenarios. Zhang et al. modeled image fusion as texture and intensity proportional maintenance and introduced a squeeze-decomposition network to retain critical information during the fusion process [10], [34]. Similarly, Ma et al. unified multiple image fusion tasks as structure maintenance, texture preservation, and appropriate intensity control, employing the Swin Transformer to capture global contexts for sufficient information aggregation [13]. Due to the absence of GT, these methods require manually selecting appropriate hyperparameters and training separate models for different tasks. Therefore, Xu et al. utilized elastic weight consolidation to integrate parameters of various tasks within a single model via continuous learning [11]. Moreover, Zhu et al. introduced the mixture of experts mechanism to achieve multiple fusion tasks within a single framework [20]. Notably, these methods still depend on constructing loss functions based on the similarity between fused and source images, which may lead to suboptimal objectives.

B. Diffusion Model

Diffusion models (DMs) [35] have gained prominence in various computer vision domains, including text-to-image generation [36], text-to-3D generation [37], image manipulation [38], and image restoration [39], [40]. The image fusion community has also introduced DMs [41], [42], [43]. For instance, DifFusion [42] uses the DM as a powerful feature extractor while DDFM [41] integrates the pre-trained DM with score matching for image fusion. Tang et al. developed the degradation-robust DMs and diffusion prior combination module to aggregate high-quality priors from multiple modalities, effectively mitigating interference [32]. Note that training DMs typically requires objective data (GT) to construct the Markov chain. Yi et al. manually selected the best results of various fusion algorithms as GTs to train a specific diffusion model for image fusion [44], which may introduce subjective biases and restrict the performance of DMs. Consequently, the potential of diffusion models in image fusion remains underexplored due to the lack of GT.

C. Masked Image Modeling

Inspired by the success of masked language modeling [45], [46] in natural language processing (NLP), Masked Image Modeling (MIM) [21], [47], [48] has recently emerged as a powerful self-supervised learning paradigm for visual representation learning. By masking a portion of an input image and predicting missing parts, MIM promotes an understanding of spatial and structural patterns, thereby enabling the learning of versatile visual representations. Pioneering works like Masked Autoencoders (MAE) [21] introduce an asymmetric encoder-decoder structure, where the encoder processes visible patches with high efficiency, and a lightweight decoder reconstructs the masked regions, effectively capturing semantic and structural features from incomplete inputs. Building on the foundation, BEiT [48] extends MIM with employing a tokenizer to convert image patches into discrete tokens, leveraging a pre-learned dictionary to guide the reconstruction, which establishes a closer connection between MIM and masked language modeling in NLP. In addition, SimMIM [22] simplifies the MIM framework by predicting raw pixel values for masked regions, avoiding complex tokenization and pre-learned dictionaries, achieving a balance between simplicity and scalability.

Although Masked Image Modeling (MIM) has achieved significant success in unimodal vision tasks, it has not been widely applied in the field of image fusion. Currently, DeFusion is the only work that has so far incorporated the idea of MIM, leveraging self-supervised learning and masking techniques for unified image fusion and demonstrating a certain level of fusion performance. However, it primarily focuses on improving the interpretability of image fusion by decomposing images into common and unique features, rather than directly addressing the absence of GT in image fusion. Moreover, DeFusion employs a simple masking scheme that simulates only the common and unique information of source images. It also limits fusion to complementary information in the Y channel and relies on hand-crafted color fusion rules, which restricts its overall performance. In this work, we overcome the lack of GT by combining masked image modeling with diffusion modeling and taking into account the presence of degradation factors in real scenes to capture a high-quality generative prior and produce fusion results that are more consistent with human visual perception.

III. Methodology

A. Problem Overview

The essence of image fusion is to integrate complementary and significant information from source images into a single fused image. However, it is difficult to explicitly define meaningful complementary information due to the absence of realistic fused images (*i.e.*, ground truth). To address this, we propose a special dual masking scheme to emulate useful complementary context, thus transforming unsupervised image fusion into a self-supervised masked image restoration task, as depicted in Fig. 2. Furthermore, we introduce a diffusion model with powerful generative and generalization capabilities, enabling the proposed solution to be applied to various image fusion

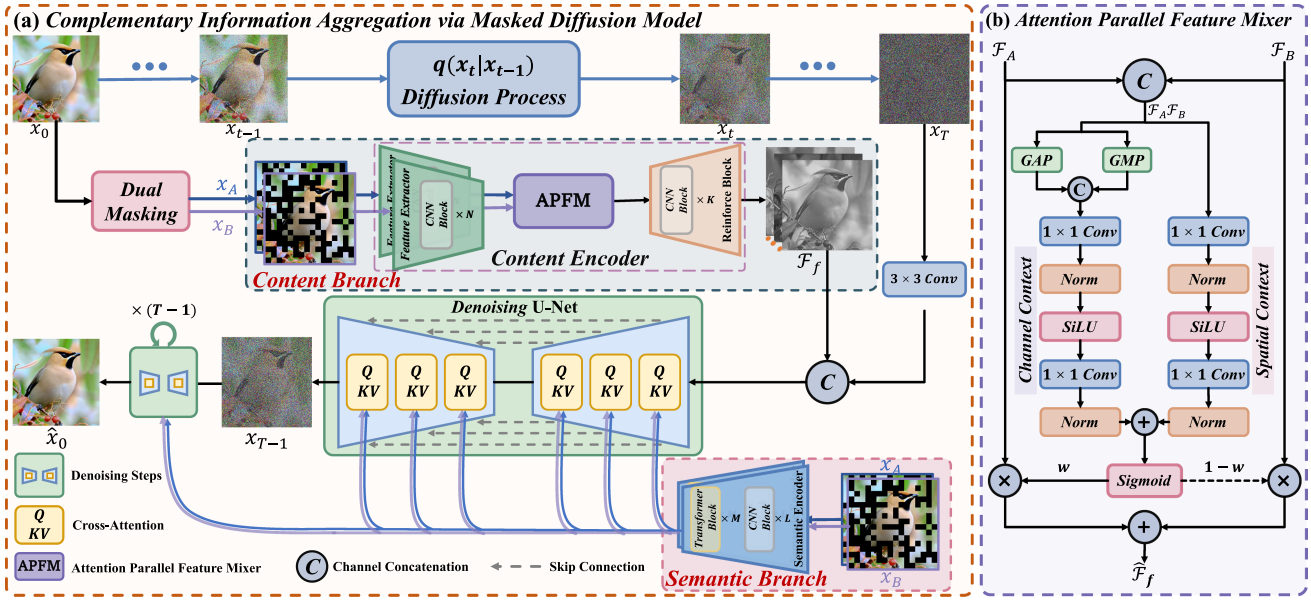


Fig. 2. The overall framework of Mask-DiFuser, which focuses on self-supervised training through masked image modeling.

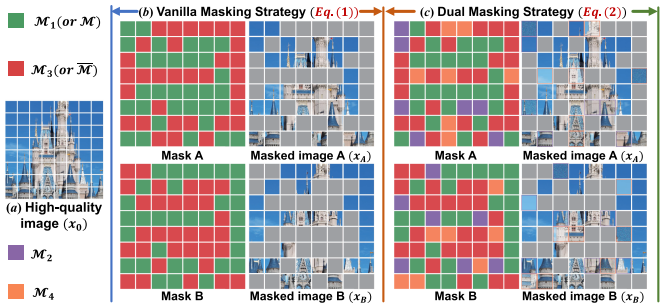


Fig. 3. Vanilla masking scheme vs. our dual masking scheme.

tasks. Ultimately, the masked diffusion model pre-trained with self-supervised learning serves as a unified and versatile generator for multiple fusion tasks without fine-tuning. Specifically, the diffusion model learns generative priors from high-quality images while integrating complementary information, thereby producing fusion results that are more consistent with human visual perception. Next, we elaborate on the key concepts of image fusion via the masked diffusion model, including the dual masking scheme, network architecture, and loss function, as well as the complete image fusion inference process.

III.-B. Image Fusion Via Masked Diffusion Model

1) Dual Masking: Given a high-quality source image $x_0 \in \mathbb{R}^{H\times W\times 3}$, the vanilla masking scheme typically constructs two complementary masked counterparts (*i.e.*, x_A and x_B) by applying a random binary mask $\mathcal{M}$ with a fixed masking ratio (e.g., 50%) sampled from a uniform distribution, as shown in Fig. 3(b). Mathematically, this process can be formulated as:

$$x_A = \mathcal{M}(x_0) + \bar{\mathcal{M}}(x_p); x_B = \bar{\mathcal{M}}(x_0) + \mathcal{M}(x_p), \quad (1)$$

where $\bar{\mathcal{M}}$ is the logical negation of $\mathcal{M}$, and $x_p \in \mathbb{R}^{H\times W\times 3}$ denotes a pure image with randomly assigned pixel values p. However, this simplistic simulation of complementary information often leads to mode collapse in subsequent fusion models, hindering effective information integration. Moreover, it overlooks common degradations in the actual imaging process, such as illumination degradation, noise, and blurring. To address these limitations, we propose a novel dual masking scheme, as illustrated in Fig. 3(c). The proposed dual masking scheme is defined as:

$$\begin{aligned} x_A &= \mathcal{M}_1^A(x_0)+\mathcal{M}_2^A(\mathcal{D}(x_0))+\mathcal{M}_3^A(x_p)+\mathcal{M}_4^A(\mathcal{D}(x_0)), \\ x_B &= \mathcal{M}_1^B(x_0)+\mathcal{M}_2^B(\mathcal{D}(x_0))+\mathcal{M}_3^B(x_p)+\mathcal{M}_4^B(\mathcal{D}(x_0)), \\ \text{s.t.} \quad & \begin{cases} \mathcal{M}_1^A \cup \mathcal{M}_2^A = \mathcal{M}; & \mathcal{M}_3^A \cup \mathcal{M}_4^A = \bar{\mathcal{M}}, \\ \mathcal{M}_1^B \cup \mathcal{M}_2^B = \bar{\mathcal{M}}; & \mathcal{M}_3^B \cup \mathcal{M}_4^B = \mathcal{M}, \end{cases} \end{aligned} \quad (2)$$

where $\mathcal{D}(\cdot)$ is a set of random degradation operations, including Gaussian blur, Gaussian noise, gamma correction, and amplitude-phase transformation [49]. This design serves dual purposes. On the one hand, $\mathcal{M}_2$ replaces portions of the non-masked region with degraded patches, *compelling the model to reconstruct high-quality content from degraded inputs, thereby enhancing its generative capabilities*. On the other hand, $\mathcal{M}_4$ introduces degraded patches into the masked region, *prompting the model to selectively extract and retain valuable content, thus improving its discriminative abilities*. Ultimately, by generating complementary images that comprehensively incorporate realistic degradations, the proposed dual masking scheme effectively simulates diverse scenarios encountered in real-world image fusion tasks. This approach significantly enhances the robustness and generalization of fusion models trained on synthetic data in a self-supervised manner.

2) Masked Diffusion Model: As presented in Fig. 2(a), our masked diffusion model inherits the conditional diffusion model and integrates a dual masking scheme, which involves the forward diffusion and reverse denoising processes. Essentially, it learns to integrate complementary information from multiple source images while approximating the distribution of high-quality images with a Markov chain. The forward process involves a Markov chain that gradually adds Gaussian noise to the data distribution $x_0 \sim p(x_0)$, defined as:

$$\begin{aligned} q(x_{1:T}|x_0) &= \Pi_{t=1}^{T} q(x_t|x_{t-1}), \\ q\left(x_t \mid x_{t-1}\right) &= \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t \mathbf{I}), \end{aligned} \tag{3}$$

where $t \in [1, 2, \ldots, T]$ denotes the time step and β_t is the variance schedule of noise. Leveraging the properties of Gaussian distributions and reparameterized iterative derivation, we can reformulate the forward Markov process as:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}), \tag{4}$$

where $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$ and $\alpha_t = 1 - \beta_t$. Therefore, the t-step sample x_t can be directly derived from x_0 as:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t, \tag{5}$$

where $\epsilon_t \sim \mathcal{N}(0, \mathbf{I})$. If the diffusion time step T is sufficiently large, $\bar{\alpha}_T$ will approach 0 asymptotically. Therefore, at the end of the forward process, the distribution of x_T approximates a standard normal distribution $\mathcal{N}(0, \mathbf{I})$.

The reverse process starts with a pure Gaussian distribution and gradually removes the noise to recover the original data distribution conditioned on two complementary masked images x_A and x_B, which also obeys a Markov chain as:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, x_A, x_B, t), \sigma_t^2 \mathbf{I}). \tag{6}$$

Following previous works [32], [36], we employ a denoise network ϵ_θ to predict the noise at t-step, so as to approximate the mean μ, and σ_t^2 is a time-dependent constant. Particularly, the distribution parameters can be expressed as:

$$\begin{aligned} \mu_\theta &= \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, x_A, x_B, t)\right), \\ \sigma_t^2 &= \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t, \end{aligned} \tag{7}$$

where $\epsilon_\theta(x_t, x_A, x_B, t)$ represents the noise estimated by ϵ_θ conditioned on x_t, x_A, x_B, and t. Finally, using the reparameterization trick, x_{t-1} can be sampled as:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, x_A, x_B, t)\right) + \sigma_t z, \tag{8}$$

where $z \sim \mathcal{N}(0, \mathbf{I})$ is a standard Gaussian noise.

As shown in Fig. 2(a), our denoise network ϵ_θ involves three critical components, *i.e.*, content encoder $\mathcal{E}_\mathrm{c}$, semantic encoder $\mathcal{E}_\mathrm{s}$, and denoising U-Net $\mathcal{N}_\mathrm{d}$, where the first two aim to capture local and global conditional context. On the one hand, we introduce a content encoder that transforms complementary images into fused features, providing a comprehensive scene representation as conditioning. In this context, complementary information refers to high-quality content such as high-contrast regions, sharp structural edges, fine-grained textures, natural illumination, and saturated colors. Compared to raw pixels, feature-based conditioning better captures structural and abstract scene information, effectively mitigating the effects of exposure and modality differences, thereby facilitating the image generation process. The content branch consists of a feature extractor, an attention parallel feature mixer (APFM), and a reinforcement block, which takes x_A and x_B as inputs to generate the local content condition $\mathcal{F}_f \in \mathbb{R}^{H\times W\times 64}$, formulated as:

$$\mathcal{F}_f = \mathcal{E}_\mathrm{c}(x_A, x_B). \tag{9}$$

Initially, x_A and x_B are processed by the feature extractor to obtain $\mathcal{F}_A \in \mathbb{R}^{H\times W\times 64}$ and $\mathcal{F}_B \in \mathbb{R}^{H\times W\times 64}$. They are then fed into APFM, which performs coarse content fusion by integrating both channel-wise and spatial context to assess feature importance and generate fusion weights. The detailed workflow of APFM is depicted in Fig. 2(b). The fused content features are subsequently processed by the reinforcement block to reduce channel dimension and produce $\mathcal{F}_f$. Finally, $\mathcal{F}_f$ is concatenated with the feature projection of x_t along the channel to provide local content priors for the denoising U-Net.

On the other hand, the semantic branch integrates CNNs and Transformer, which aims to provide global semantic priors. In detail, we first feed x_A and x_B into a shared semantic encoder to obtain semantic representations $\mathcal{S}_A \in \mathbb{R}^{H/4\times W/4\times 128}$ and $\mathcal{S}_B \in \mathbb{R}^{H/4\times W/4\times 128}$, which are expressed as:

$$\mathcal{S}_A = \mathcal{E}_\mathrm{s}(x_A), \quad \mathcal{S}_B = \mathcal{E}_\mathrm{s}(x_B). \tag{10}$$

Then, $\mathcal{S}_A$ and $\mathcal{S}_B$ are incorporated into the denoising U-Net through a cross-attention mechanism, offering global semantic guidance for image generation. The cross-attention mechanism is formulated as:

$$\text{Attention}\left(\varphi_i(x_t), \mathcal{S}\right) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V, \tag{11}$$

in which $Q = W_Q^{(i)} \cdot \varphi_i(x_t), K = W_K^{(i)} \cdot \mathcal{S}, V = W_V^{(i)} \cdot \mathcal{S}$, $\varphi_i(x_t)$ means the intermediate representation of the i-th layer in the denoising U-Net, $W_Q^{(i)}$, $W_K^{(i)}$, $W_V^{(i)}$ are learnable weight matrices, and d is a scaling factor. For multi-source information interaction, we concatenate $\text{Attention}(\varphi_i(x_t), \mathcal{S}_A)$ and $\text{Attention}(\varphi_i(x_t), \mathcal{S}_B)$ along the channel dimension. This branch ensures that coarse-grained scene context from multiple source images is fully leveraged, allowing the fused image to align with human perceptual expectations in terms of structure, illumination, and color fidelity. Ultimately, by integrating local content and global semantic information, the noise estimation process $\epsilon_\theta(x_t, x_A, x_B, t)$ can be implemented as $\mathcal{N}_\mathrm{d}(x_t, \mathcal{F}_f, \mathcal{S}_A, \mathcal{S}_B, t)$.

3) Loss Function: Following previous works [35], [36], we use a diffusion loss $\mathcal{L}_{diff}$ to train our masked diffusion model, defined as:

$$\mathcal{L}_{diff} = \|\epsilon_t - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t, x_A, x_B, t)\|_2, \tag{12}$$

where $\|\cdot\|_2$ denotes the l_2 norm. Although the masked diffusion model can approximate the $p(x_0|x_A, x_B)$ distribution by optimizing $\mathcal{L}_{diff}$ through self-supervised learning, we observe in practice that it often suffers from noticeable color

distortion. Inspired by [40], [50], we further introduce image-level consistency losses, including pixel loss $\mathcal{L}_{pix}$, structural similarity (SSIM) loss $\mathcal{L}_{ssim}$, perceptual loss $\mathcal{L}_{per}$, and color consistency loss $\mathcal{L}_{col}$. In particular, according to (5), we can approximate $\hat{x}_0^t$ based on the noise predicted by the denoise network ϵ_θ, expressed as:

$$\hat{x}_0^t = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t, x_A, x_B, t)\right). \quad (13)$$

Thus, the pixel loss $\mathcal{L}_{pix}$ is defined as the intensity difference between estimated $\hat{x}_0^t$ and ground truth x_0,

$$\mathcal{L}_{pix} = \|\hat{x}_0^t - x_0\|_2. \quad (14)$$

Moreover, the SSIM loss maintaining structural similarity, which is defined as:

$$\mathcal{L}_{ssim} = \frac{(2\mu_{\hat{x}_0^t}\mu_{x_0} + c_1)(2\sigma_{\hat{x}_0^t x_0} + c_2)}{(\mu_{\hat{x}_0^t}^2 + \mu_{x_0}^2 + c_1)(\sigma_{\hat{x}_0^t}^2 + \sigma_{x_0}^2 + c_2)}, \quad (15)$$

where $\mu_{\hat{x}_0^t}$ and μ_{x_0} are mean pixel values, $\sigma_{\hat{x}_0^t}^2$ and $\sigma_{x_0}^2$ are variances, $\sigma_{\hat{x}_0^t x_0}$ is the covariance, and c_1, c_2 are constants for numerical stability. The perceptual loss imposes constraints by extracting deep features from $\hat{x}_0^t$ and x_0 with the VGG network, formulated as:

$$\mathcal{L}_{per} = \sum_l \left\|\phi_{vgg}^l(\hat{x}_0^t) - \phi_{vgg}^l(x_0)\right\|_2, \quad (16)$$

where ϕ_{vgg}^l indicates the l-th convolutional layer of the VGG network. The color consistency loss explicitly minimizes the angle between the color vectors of $\hat{x}_0^t$ and x_0 to prevent color distortion, which is formulated as:

$$\mathcal{L}_{col} = \frac{1}{N}\sum_{i=1}^{N} \angle\left(\hat{x}_0^t(i), x_0(i)\right), \quad (17)$$

where $x(i)$ represents the i-th pixel of x, and $\angle(\cdot,\cdot)$ calculates the angle between two color vectors in the RGB color space. Finally, the total loss for training our masked diffusion model is the weighted sum of aforementioned losses:

$$\mathcal{L} = \mathcal{L}_{diff} + \lambda_1\mathcal{L}_{pix} + \lambda_2\mathcal{L}_{ssim} + \lambda_3\mathcal{L}_{per} + \lambda_4\mathcal{L}_{col}, \quad (18)$$

where λ_1, λ_2, λ_3, and λ_4 are parameters controlling the tradeoff of various losses. The overall training pipeline is summarized in Algorithm 1.

4) Inference: Once the training is complete, the masked diffusion model serves as a versatile and unified fusion network, which supports infrared-visible, medical, multi-exposure, and multi-focus image pairs as the inputs to directly generate the fused images without fine-tuning. As shown in Algorithm 1, complementary multi-source images are first fed into the content encoder and the semantic encoder to extract local content features and global semantic priors, respectively. The content features are then concatenated with the feature representation of the noise sample x_t along the channel dimension, serving as input to the denoising U-Net, while the semantic priors are incorporated into the U-Net via the cross-attention mechanism. Eventually, starting from a Gaussian distribution, our Mask-DiFuser performs iterative denoising according to (8), guided by local content and global semantic priors, to synthesize an informative fused image.

Algorithm 1: Training and Inference of Mask-DiFuser.

Tranining: Train content encoder $\mathcal{E}_c$, semantic encoder $\mathcal{E}_s$, and denoise U-Net $\mathcal{N}_d$ within ϵ_θ
Input: High-quality image x_0, noise schedule α_t
repeat
  Construct x_A and x_B from x_0 via Eq. (2);
  $t \sim \text{Uniform}(1, \cdots, T)$, $\epsilon_t \sim \mathcal{N}(0, \mathbf{I})$;
  $\mathcal{F}_f = \mathcal{E}_c(x_A, x_B)$, $\mathcal{S}_A = \mathcal{E}_s(x_A)$, $\mathcal{S}_B = \mathcal{E}_s(x_B)$;
  $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t$;
  $\hat{x}_0^t = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t, x_A, x_B, t)\right)$
    with $\epsilon_\theta(x_t, x_A, x_B, t) := \mathcal{N}_d(x_t, \mathcal{F}_f, \mathcal{S}_A, \mathcal{S}_B, t)$;
  Take a gradient descent step on $\nabla\mathcal{L}$;
until *converged*;
Inference: Sample the fused image x_f from multi-source images x_A and x_B
Input: Content encoder $\mathcal{E}_c$, semantic encoder $\mathcal{E}_s$, and denoise U-Net $\mathcal{N}_d$ within ϵ_θ, source images $\{x_A, x_B\}$ and noise schedule α_t
$\mathcal{F}_f = \mathcal{E}_c(x_A, x_B)$, $\mathcal{S}_A = \mathcal{E}_s(x_A)$, $\mathcal{S}_B = \mathcal{E}_s(x_B)$;
Initializing sample $x_T \sim \mathcal{N}(0, \mathbf{I})$;
for $t = T, \cdots, 1$ **do**
  Sample $z \sim \mathcal{N}(0, \mathbf{I})$ if $t > 1$, else $z = 0$;
  $x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, x_A, x_B, t)\right) + \sigma_t z$
    with $\epsilon_\theta(x_t, x_A, x_B, t) := \mathcal{N}_d(x_t, \mathcal{F}_f, \mathcal{S}_A, \mathcal{S}_B, t)$;
end
return x_0 as the fused image x_f

IV. Experiments

In this section, we will showcase the performance of Mask-DiFuser on various image fusion tasks, highlighting its remarkable effectiveness, robustness, and adaptability in handling complex fusion scenarios.

A. Experiment Details

1) Implementation Details: We utilize the high-quality (*normal exposure, clear, high contrast*) natural images from the DIV-2 K dataset [51] to train our Mask-DiFuser, rather than relying on typical image fusion datasets. The primary limitation of these datasets lies in the fact that *they fail to provide high-quality images*, (*i.e.,* desired fusion results), *which hinders our method from constructing a complete Markov process.* Moreover, the distinct properties of different modalities in these datasets may adversely *affect model optimization*. For example, grayscale IR and CT images may lead the model to overlook color information.

All images are randomly cropped into 128×128 patches, and the batch size is set to 8. Training is conducted for 3,500 epochs using the AdamW optimizer with an initial learning rate of 5×10^{-5} and a weight decay of 1×10^{-4}. In the dual masking scheme, we adopt an 8×8 grid to implement the masks and

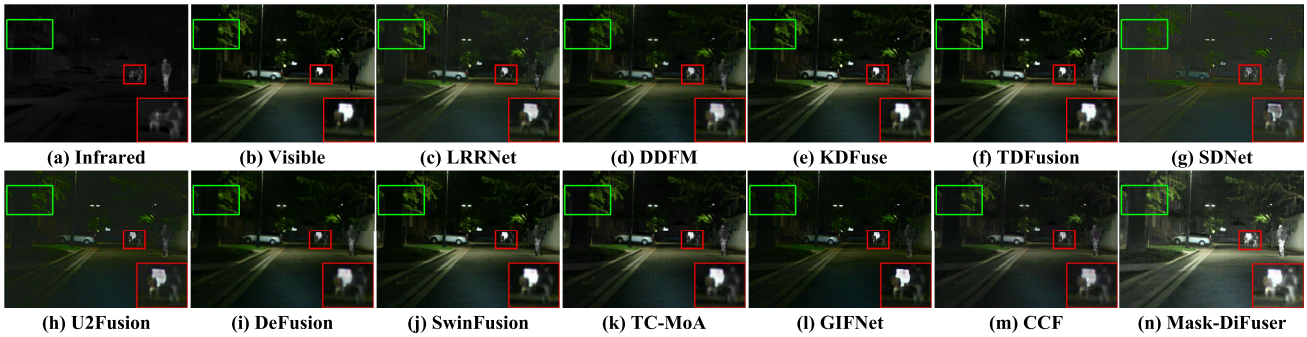


Fig. 4. Visual comparison of infrared-visible image fusion results for night scenes on the MSRS dataset.

set $\mathcal{M}_1 : \mathcal{M}_2 : \mathcal{M}_3 : \mathcal{M}_4$ to $4 : 1 : 4 : 1$. The hyper-parameters for balancing different loss terms are empirically set as $\lambda_1 = 0.05$, $\lambda_2 = 0.05$, $\lambda_3 = 0.1$, and $\lambda_4 = 1.0$. During inference, we employ the DDIM sampling strategy with a step of 5 to generate fusion results. All experiments are conducted on the NVIDIA RTX 4090 GPUs and 2.50 GHz Intel(R) Xeon(R) Platinum 8180 CPU with PyTorch.

2) Experiment Configurations: We demonstrate the effectiveness and generalizability of our Mask-DiFuser across six representative fusion tasks, *i.e.*, infrared and visible image fusion (IVIF), near-infrared and visible image fusion (NVIF), multi-polarization image fusion (MPF), medical image fusion (MIF), multi-exposure image fusion (MEF), and multi-focus image fusion (MFF). Comprehensive comparisons are conducted on eight typical datasets, *i.e.*, 361 image pairs in the MSRS dataset [56] and 100 image pairs in the RoadScene dataset [11] for IVIF, 100 image pairs in the Nirscene dataset [57] for NVIF, 40 image pairs in the Polarization dataset [58] for MPF, 100 image pairs in the Harvard dataset [59] for MIF, 360 image pairs in the SICE dataset [17] and 100 image pairs in the MEFB dataset [60] for MEF, and 20 image pairs in the Lytro dataset [61] for MFF.

For each task, we compare our Mask-DiFuser with four task-specific and seven unified methods. The task-specific methods are LRRNet [19], DDFM [41], KDFuse [52], and TDFusion [53] for IVIF and NVIF, LRRNet, DDFM, PAPIF [62], and CPIFuse [63] for MPF, EMFusion [7], DDFM [41], ALMFNet [64], and MMIF-INet [65] for MIF, SAMT-MEF [24], HSDS-MEF [66], CRMEF [67], and EAT [68] for MEF, and ZMFF [69], DB-MFIF [25], Fusion2Void [70] and MDLSR_RFM [71] for MFF. The unified alternatives include SDNet [34], U2Fusion [11], DeFusion [12], SwinFusion [13], TC-MoA [20], GIFNet [54] and CCF [55]. To ensure a fair comparison, all compared methods are evaluated using their publicly available pretrained models without any additional retraining. Six metrics, *i.e.*, EN, AG, SF, SD, SCD [72], and PI [73], are employed to quantify fusion performance. EN, AG, SF, and SD serve as fundamental statistical metrics that comprehensively reflect the texture richness, edge sharpness, spatial activity level, and contrast distribution properties of fused images, respectively. Meanwhile, SCD measures the correlation between fused and source images. Furthermore, PI assesses image quality from a perceptual perspective.

B. Infrared and Visible Image Fusion

Qualitative analysis reveals that the fusion results generated by our Mask-DiFuser demonstrate the most prominent pedestrians, the highest overall contrast, and the optimal exposure levels. Specifically, as shown in Fig. 4, LRRNet, DDFM, U2Fusion, DeFusion, SwinFusion, TC-MoA, GIFNet, and CCF are ineffective in preserving the prominence of pedestrians for the night scene on the MSRS dataset, while our Mask-DiFuser enhances the infrared information. More attractively, our method enhances details in low light, such as lawns, roads, and buildings, making the overall scene brighter and more recognizable. In daytime scenes, as can be noticed in Fig. 5, the fusion results obtained by Mask-DiFuser are also outstanding, especially the prominence of the characters as well as the texture details of the doors. Additionally, on the RoadScene dataset, we can find that the overexposure issue in visible images poses a challenge for U2Fusion, DeFusion, SwinFusion, TC-MoA, and CCF, as these methods fail to effectively mitigate this problem, shown in Fig. 6. In contrast, Mask-DiFuser successfully corrects the overexposure artifacts in the source images. Table I presents the quantitative results of Mask-DiFuser and other algorithms. It is evident that our method consistently demonstrates top-tier performance across all evaluation metrics. The superior EN and AG values, along with the competitive SF score, indicate that the fused images generated by our approach effectively retain the most detailed texture information. The highest SD value strongly demonstrates the optimal contrast, while the best PI value represents the superior visual quality. Overall, the experimental results confirm that Mask-DiFuser is capable of learning high-quality prior information, such as proper exposure, sharpness, and high contrast, which further strengthens the advantage of Mask-DiFuser in handling complex fusion tasks.

C. Near-Infrared and Visible Image Fusion

Qualitative results for the near-infrared and visible image fusion task are presented in Fig. 7. It is evident that the fusion results of other methods are more susceptible to haze degradation, resulting in blurred cloud and mountain structures, reduced color vividness, and diminished overall contrast. This primarily stems from their inability to fully exploit the rich texture details inherent in near-infrared (NIR) images. In contrast, Mask-DiFuser generates more natural colors in the sky and preserves cloud

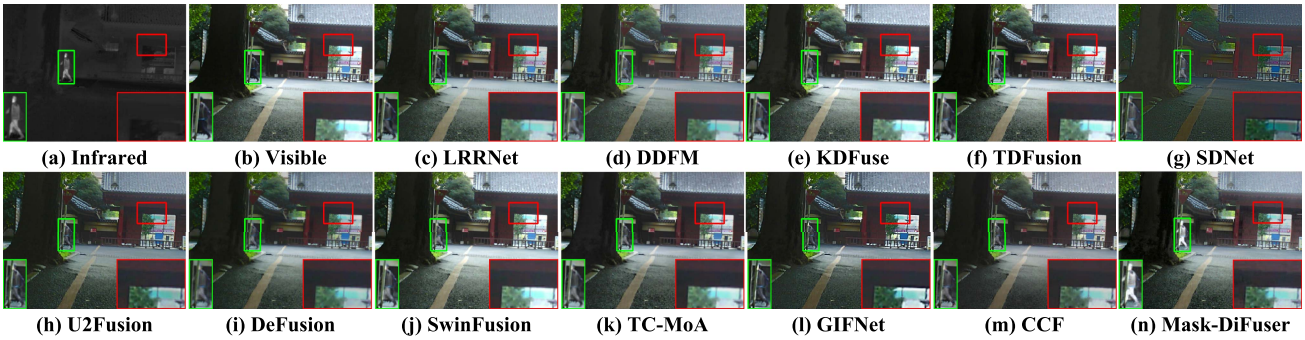


Fig. 5. Visual comparison of infrared-visible image fusion results for daytime scenes on the MSRS dataset.

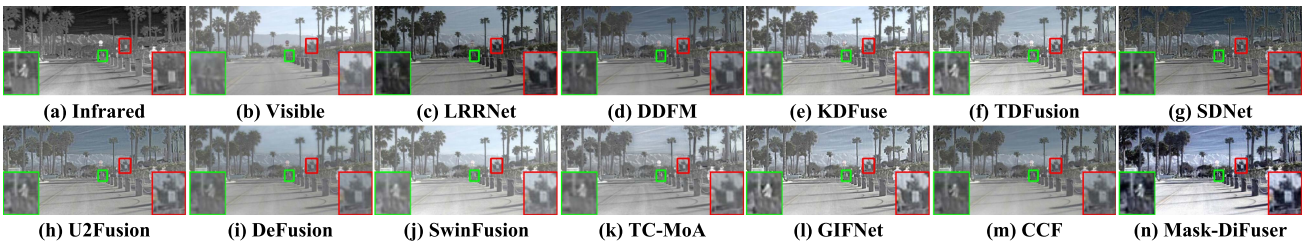


Fig. 6. Visual comparison of infrared-visible image fusion results on the RoadScene dataset.

TABLE I
QUANTITATIVE RESULTS OF MASK-DIFUSER VS. SOTA METHODS ON THE MSRS AND ROADSCENE DATASETS. THE BEST AND SECOND-BEST RESULTS ARE HIGHLIGHTED WITH RED \COLORBOX[HTML]{FFC7CE}{RED} AND PURPLE \COLORBOX[HTML]{D9D9FF}{PURPLE}, RESPECTIVELY.

Methods	MSRS Infrared-Visible Fusion Dataset						RoadScene Infrared-Visible Fusion Dataset					
	EN ↑	AG ↑	SF ↑	SD ↑	SCD ↑	PI ↓	EN ↑	AG ↑	SF ↑	SD ↑	SCD ↑	PI ↓
LRRNet [19]	6.216	2.665	7.337	31.761	0.791	4.290	7.110	4.452	9.842	41.628	1.613	3.024
DDFM [41]	6.190	2.425	6.398	29.230	1.444	4.354	7.176	3.867	8.691	40.768	1.717	3.766
KDFuse [52]	6.673	3.688	9.588	42.041	1.623	3.824	7.304	5.943	12.53	49.012	1.411	3.228
TDFusion [53]	6.759	3.722	9.758	42.950	**1.864**	3.673	7.451	5.951	13.003	54.597	**1.872**	2.480
SDNet [34]	5.270	2.697	7.324	17.322	0.986	4.008	7.303	5.782	12.290	44.231	1.481	2.987
U2Fusion [11]	5.250	2.535	6.990	22.679	1.144	4.204	6.906	4.619	9.582	33.260	1.349	3.100
DeFusion [12]	6.369	2.621	7.002	34.887	1.291	4.358	6.926	3.305	7.079	35.242	1.337	3.767
SwinFusion [13]	6.641	3.562	9.619	42.980	1.686	3.948	6.993	4.407	10.005	44.491	1.630	3.058
TC-MoA [20]	6.514	3.162	8.438	36.105	1.555	4.689	7.166	4.613	9.966	40.332	1.476	3.669
GIFNet [54]	5.958	3.355	10.739	32.878	1.408	3.925	7.335	6.169	**14.741**	47.555	1.740	3.098
CCF [55]	6.213	2.822	7.29	30.339	1.439	3.679	7.286	4.283	8.733	45.148	1.801	2.905
Mask-DiFuser	**6.978**	**4.603**	**10.823**	**47.347**	1.860	**2.940**	**7.686**	**6.956**	14.568	**68.613**	**1.872**	**2.299**

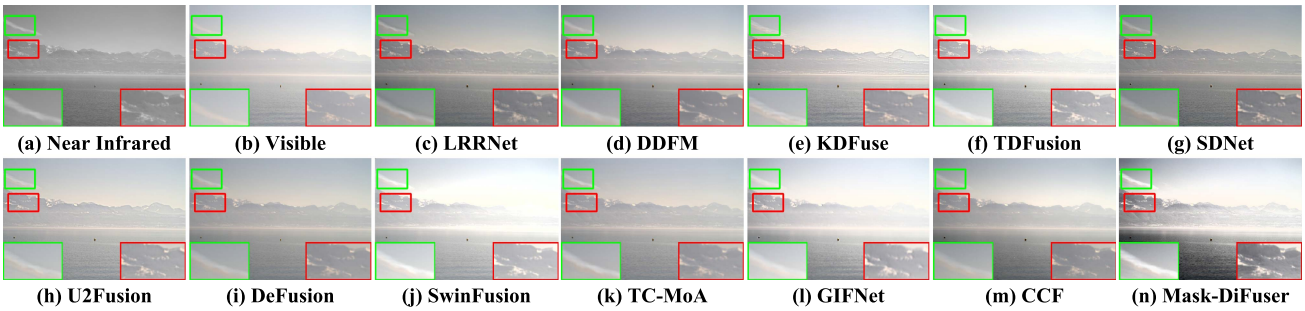


Fig. 7. Visual comparison of near-infrared and visible image fusion results on the Nirscene dataset.

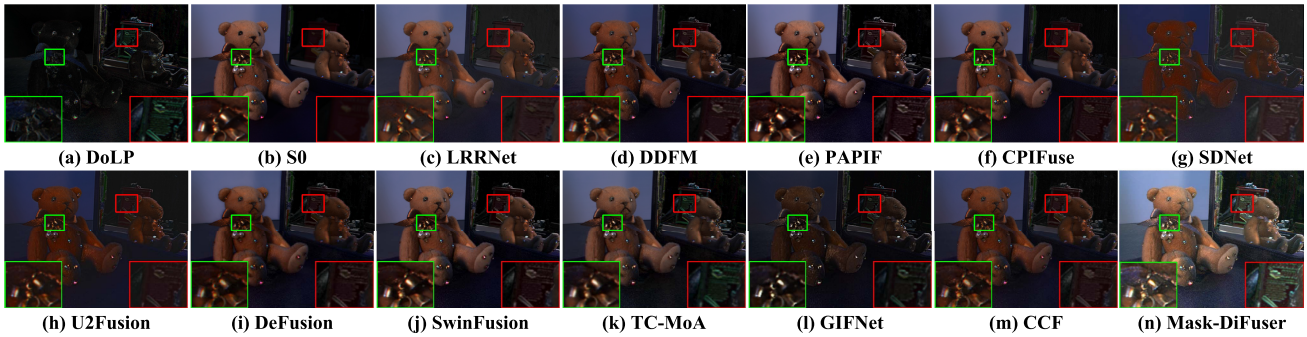


Fig. 8. Visual comparison of multi-polarization fusion results on the Polarization dataset.

TABLE II
QUANTITATIVE RESULTS OF MASK-DIFUSER VS. SOTA METHODS ON THE NIRSCENE DATASET

Methods	EN ↑	AG ↑	SF ↑	SD ↑	SCD ↑	PI ↓
LRRNet [19]	7.050	4.251	9.704	44.517	0.920	3.511
DDFM [41]	7.158	4.041	9.630	45.478	1.272	3.447
KDFuse [52]	7.214	5.414	11.880	46.242	1.148	3.083
TDFusion [53]	7.323	6.108	14.020	55.089	1.641	2.854
SDNet [34]	6.950	5.862	13.227	37.338	0.567	2.953
U2Fusion [11]	7.073	5.353	11.892	42.178	0.880	3.123
DeFusion [12]	6.875	3.426	7.894	36.915	0.647	3.751
SwinFusion [13]	7.021	5.444	12.514	48.176	1.382	2.930
TC-MoA [20]	6.996	3.856	8.851	39.839	0.748	3.529
GIFNet [54]	7.229	6.868	**17.210**	54.393	1.443	3.037
CCF [55]	7.286	4.396	10.383	51.423	1.600	3.265
Mask-DiFuser	**7.553**	**7.625**	16.715	**71.296**	**1.708**	**2.660**

TABLE III
QUANTITATIVE RESULTS OF MASK-DIFUSER VS. SOTA METHODS ON THE POLARIZATION DATASET

Methods	EN ↑	AG ↑	SF ↑	SD ↑	SCD ↑	PI ↓
LRRNet [19]	6.470	3.122	7.829	32.242	0.844	3.508
DDFM [41]	6.518	3.699	9.172	29.230	**1.875**	3.048
PAPIF [62]	6.798	4.216	11.254	39.564	1.674	2.958
CPIFuse [63]	6.422	3.201	7.535	30.001	1.548	3.924
SDNet [34]	5.350	3.425	8.764	16.917	1.220	3.555
U2Fusion [11]	5.759	3.343	8.759	21.254	1.501	3.530
DeFusion [12]	6.351	2.696	6.464	27.870	1.483	3.758
SwinFusion [13]	6.597	3.443	8.838	33.625	1.742	3.371
TC-MoA [20]	6.580	3.212	8.085	32.245	1.640	3.483
GIFNet [54]	5.913	3.591	10.059	21.392	1.539	3.419
CCF [55]	6.332	3.637	8.972	25.867	1.713	2.944
Mask-DiFuser	**7.126**	**5.552**	**13.444**	**45.118**	1.854	**2.734**

details more clearly, demonstrating its effective integration of complementary information from both near-infrared and visible images. Moreover, as reported in Table II, our Mask-DiFuser achieves the best performance in EN, AG, SD, SCD, and PI metrics, and ranks second in SF, demonstrating its superiority in information preservation, sharpness, visual fidelity, and edge retention.

D. Multi-Polarization Image Fusion

In the multi-polarization image fusion task, as shown in Fig. 8, the results synthesized by SDNet, U2Fusion, and GIFNet appear overly dark, with most of the details in the red box being lost. Although other methods manage to preserve the texture of the scarf around the neck of the bear, they fail to take advantage of the distinctive polarization cues to enhance the structural details in the photo frame region. In contrast, our method integrates the complementary information from the Degree of Linear Polarization (DoLP) and total intensity (S0) images, enabling better preservation of fine textures and enhanced visibility of details under low-light conditions, while maintaining a more natural overall brightness.

The quantitative results in Table III further demonstrate the significant advantage of the proposed method in the multi-polarization image fusion task. The highest EN score indicates that our results contain the most abundant information. The top performance in AG, SF, and SD reflects the richness of texture details and contrast, which is highly consistent with the qualitative observations. In addition, the best PI score suggests that our fusion results better align with human visual perception. Although our method slightly lags behind DDFM in SCD, it still indicates a higher correlation between our fused results and the source images.

E. Medical Image Fusion

The visual comparison of medical image fusion is presented in Fig. 9. From the results, we can observe that DDFM, U2Fusion, DeFusion, SwinFusion, TC-MoA, and CCF tend to weaken the fundamental information from source images, specifically failing to highlight the soft tissue details in MRI images. Conversely, our Mask-DiFuser can highlight texture information, such as physiological structure while maintaining functional distribution, achieving sharper edge details. While task-specific methods such as EMFusion, ALMFNet, and MMIF-INet, strike a balance between the detailed textures in MRI images and the functional information in PET images, their fusion results tend to be somewhat dull, whereas our method minimizes artifacts. For the quantitative results in Table IV, Mask-DiFuser performs best on EN, SD, SCD, and PI metrics, which shows that our results have more information, superior contrast, and better visualization. Although GIFNet achieves the best performance on the AG

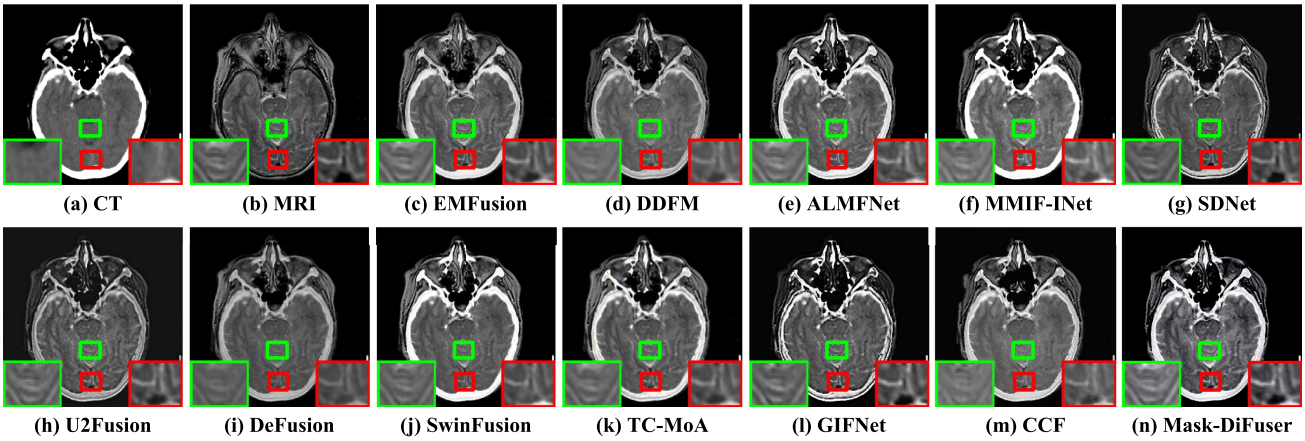


Fig. 9. Visual comparison of medical image fusion results on the Harvard dataset.

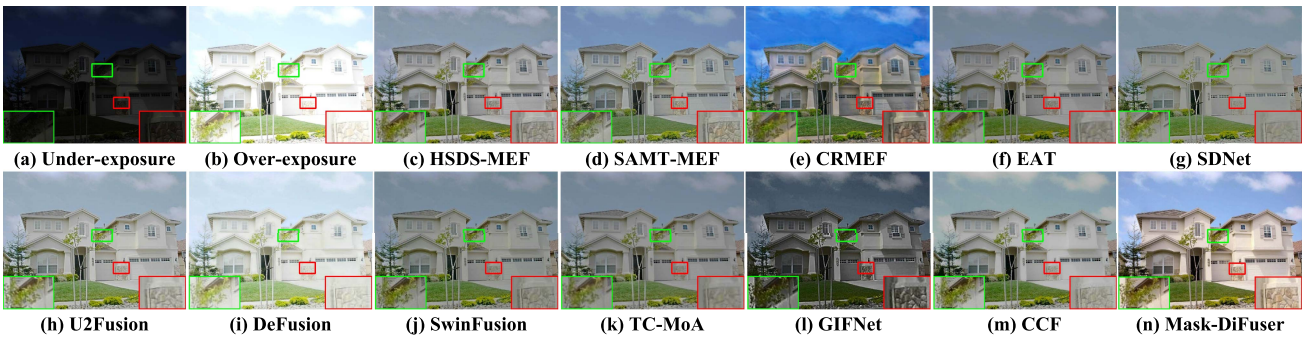


Fig. 10. Visual comparison of multi-exposure image fusion results on the SICE dataset.

TABLE IV
QUANTITATIVE RESULTS OF MASK-DIFUSER VS. SOTA METHODS ON THE HARVARD DATASET

Methods	EN ↑	AG ↑	SF ↑	SD ↑	SCD ↑	PI ↓
EMFusion [7]	4.584	5.830	18.720	78.333	1.245	3.819
DDFM [41]	4.196	4.910	16.027	61.654	1.020	3.319
ALMFNet [64]	4.387	7.955	26.578	85.860	1.519	3.436
MMIF-INet [65]	4.356	7.729	27.485	91.188	1.645	3.623
SDNet [34]	4.913	7.558	22.781	62.226	0.966	3.432
U2Fusion [11]	4.693	6.018	17.954	53.179	0.716	3.800
DeFusion [12]	4.469	5.314	17.697	69.290	1.019	3.413
SwinFusion [13]	4.217	7.297	25.104	88.860	1.472	3.325
TC-MoA [20]	4.721	7.055	22.280	79.797	1.320	3.617
GIFNet [54]	4.770	**9.712**	**30.973**	71.592	1.082	3.921
CCF [55]	4.596	6.284	21.208	72.989	1.268	3.130
Mask-DiFuser	**5.372**	7.301	23.782	**98.969**	**1.765**	**2.928**

and SF metrics, the qualitative result indicates that their visual performance is inferior to our method. Additionally, our method achieves competitive performance on the AG and SF metrics, surpassing specialized medical image fusion algorithms, such as EMFusion, which highlights the superior capability of our approach in handling cross-modal image fusion by effectively preserving structural details and enhancing overall perceptual quality.

F. Multi-Exposure Image Fusion

The qualitative comparison of multi-exposure image fusion is presented in Figs. 10 and 11. As shown in Fig. 10, it can be observed that EAT, SDNet, U2Fusion, SwinFusion, TC-MoA, and CCF show varying levels of texture degradation, especially in foliage and building surfaces, which negatively affect visual quality. Additionally, DeFusion suffers from severe overexposure, revealing its inability to balance different exposure levels effectively. CRMEF exhibits noticeably distorted tones and prominent artifacts, particularly in the sky and adjacent building areas. Moreover, GIFNet produces overly dark results, causing substantial loss of fine structures in regions such as the sky, buildings, and lawns. In contrast, HSDS-MEF, SAMT-MEF and our Mask-DiFuser produce well-balanced overall exposure and retain considerable detail fidelity. In particular, Mask-DiFuser delivers more balanced exposure, superior detail preservation, and overall superior performance. It is worth noting that HSDS-MEF and SAMT-MEF fail to retain critical fine details in the second scenario, such as the cake and food inside the glass, as presented in Fig. 11. Moreover, EAT, SDNet, U2Fusion, SwinFusion, TC-MoA, and CCF still exhibit texture blurring. Meanwhile, DeFusion continues to suffer from pronounced overexposure, and the fusion result of GIFNet remains excessively dark, losing much of the original color information. On the contrary, Mask-DiFuser not only preserves intricate structures

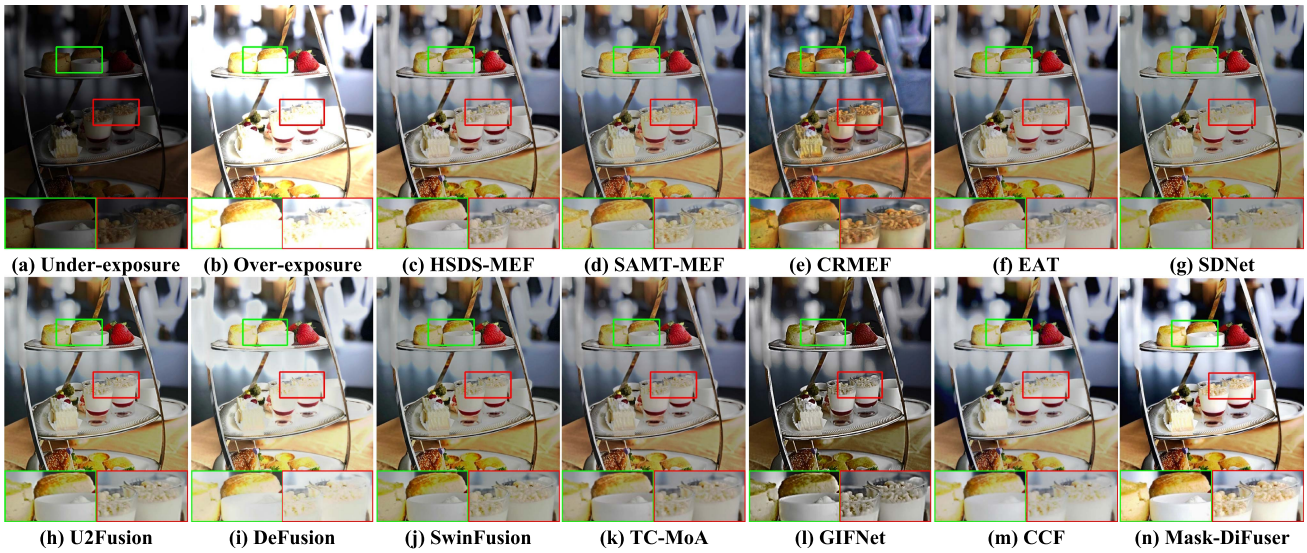


Fig. 11. Visual comparison of multi-exposure image fusion results on the MEFB dataset.

TABLE V
QUANTITATIVE RESULTS OF MASK-DIFUSER VS. SOTA METHODS ON THE SICE AND MEFB DATASETS

Methods	SICE Multi-exposure Image Fusion Dataset						MEFB Multi-exposure Image Fusion Dataset					
	EN ↑	AG ↑	SF ↑	SD ↑	SCD ↑	PI ↓	EN ↑	AG ↑	SF ↑	SD ↑	SCD ↑	PI ↓
HSDS-MEF [66]	7.146	7.444	16.156	47.636	0.785	2.351	7.165	5.948	14.389	55.514	0.788	**2.821**
SAMT-MEF [24]	7.011	6.912	14.885	43.348	0.653	2.468	7.076	5.129	12.562	50.910	0.634	3.001
CRMEF [67]	7.217	7.925	16.286	42.903	0.336	2.351	7.326	6.152	14.020	49.287	0.300	3.142
EAT [68]	7.051	5.850	12.567	45.796	0.749	2.839	7.154	4.631	11.736	54.319	0.824	3.377
SDNet [34]	6.853	7.082	16.387	43.793	0.480	2.482	6.961	5.316	13.853	50.498	0.527	3.125
U2Fusion [11]	6.930	6.436	14.670	45.602	0.816	2.635	6.900	4.887	12.736	52.021	0.810	3.393
DeFusion [12]	6.892	6.234	14.132	48.145	0.073	2.601	6.815	4.585	11.768	52.222	-0.117	3.455
SwinFusion [13]	6.964	7.404	16.656	42.095	0.506	2.334	7.068	5.622	14.084	49.116	0.524	2.939
TC-MoA [20]	6.931	5.728	12.685	40.573	0.459	2.520	7.077	4.576	11.346	48.364	0.481	3.473
GIFNet [54]	6.848	7.887	**20.041**	54.057	1.065	2.849	7.065	6.148	**18.169**	74.001	1.483	3.378
CCF [55]	7.149	6.776	15.518	52.396	1.164	2.471	7.178	4.776	12.485	61.137	1.206	3.195
Mask-DiFuser	**7.502**	**8.112**	18.071	**67.786**	**1.646**	**2.274**	**7.424**	**6.179**	15.307	**75.106**	**1.575**	3.045

comprehensively but also renders vivid, well-saturated colors. This indicates its ability to learn high-quality image priors, such as accurate exposure and vibrant color via self-supervised learning. Quantitative results in Table V further confirm these observations. The highest EN, AG, SD, and SCD scores indicate that our method achieves richer structural details and higher contrast. Although the SF and PI metrics on the MEFB dataset are not optimal, the proposed method remains highly competitive compared to fusion approaches.

G. Multi-Focus Image Fusion

As illustrated in Fig. 12, DeFusion, SwinFusion, TC-MoA, and CCF suffer from a loss of sharpness in the mane of the horse, the metallic buckle of the bridle, and the background trees, which compromises fine texture preservation and results in less visually coherent fused images. In addition, although ZMFF, DB-MFIF, Fusion2Void, MDLSR_RFM, SDNet, and U2Fusion perform relatively well in preserving detail sharpness, their outputs generally exhibit a slightly darker overall tone, as observed in the bridle, the eye of the horse, and the sky. Notably, GIFNet exhibits the best sharpness in the mane of the horse (*i.e.*, red box), even surpassing the sharpness in the corresponding focused regions of the source images. This is mainly attributed to its training strategy, which introduces an *auxiliary task branch trained on the multi-focus image fusion task* to provide additional pixel-level supervision. However, due to over-sharpening, GIFNet introduces slight halo artifacts along the edges of the hat and around the white regions on the horse head, highlighted with blue elliptical boxes. In contrast, Mask-DiFuser excels at preserving subtle textures, such as the woven patterns of the bridle and the details of the tree. Additionally, it ensures smooth and natural transitions between different focus regions, such as the boundary between the horse head and the background, without noticeable discontinuities. This produces a more coherent and visually realistic result, as highlighted in Fig. 12. Interestingly, Mask-DiFuser effectively enhances contrast and shadow details in the head of the horse and achieves natural

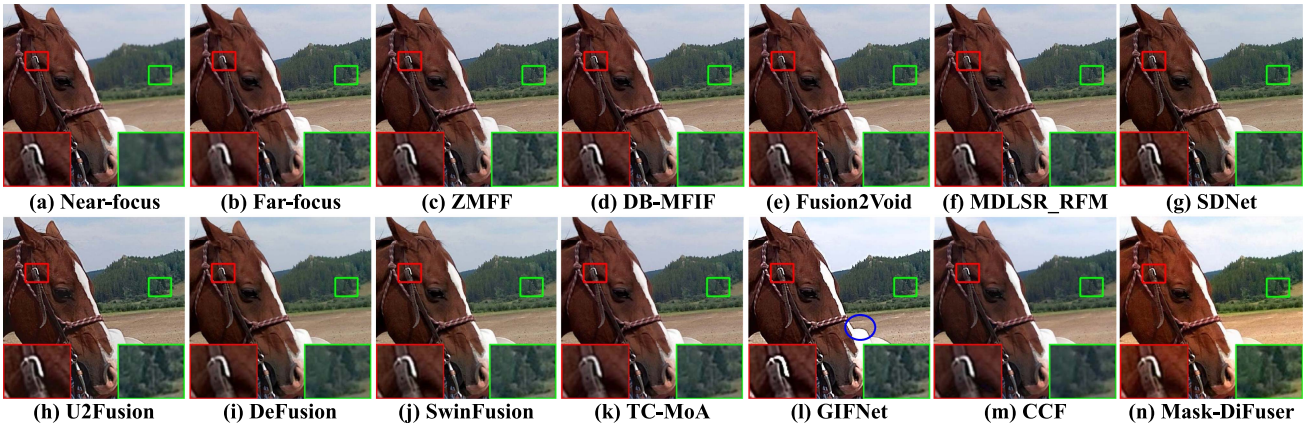


Fig. 12. Visual comparison of multi-focus image fusion results on the Lytro dataset.

TABLE VI
QUANTITATIVE RESULTS OF MASK-DIFUSER VS. SOTA METHODS ON THE LYTRO DATASET

Methods	EN ↑	AG ↑	SF ↑	SD ↑	SCD ↑	PI ↓
ZMFF [69]	7.550	6.715	16.040	56.853	0.393	3.115
DB-MFIF [25]	7.562	6.856	16.287	57.707	0.572	2.547
Fusion2Void [70]	7.556	6.779	16.156	57.509	0.534	**2.546**
MDLSR_RFM [71]	7.555	6.805	16.215	57.588	0.548	2.568
SDNet [34]	7.565	6.101	14.463	58.320	0.882	2.595
U2Fusion [11]	7.478	6.163	14.256	57.502	0.593	2.994
DeFusion [12]	7.490	4.276	9.924	54.474	0.280	3.647
SwinFusion [13]	7.547	5.931	14.050	56.666	0.556	2.742
TC-MoA [20]	7.537	5.697	13.586	56.664	0.510	3.350
GIFNet [54]	7.602	**7.927**	**20.626**	68.663	1.238	3.261
CCF [55]	7.607	4.373	10.636	61.988	1.335	4.133
Mask-DiFuser	**7.723**	6.301	14.496	**73.411**	**1.530**	3.048

color tones, resulting in a visually coherent all-in-focus image free of noticeable artifacts. This advantage comes from the fact that our diffusion model successfully captures the high-quality image generation priors, and thus represents imaging scenarios in a more natural manner, which other methods fail to achieve.

Furthermore, the quantitative results in Table VI confirm the effectiveness of our method, which achieves outstanding performance in the EN, SD, and SCD metrics, indicating that our fused images retain the richest information and highest contrast. The task-specific method Fusion2Void attains the best performance in the PI metric. Notably, GIFNet significantly surpasses others in the AG and SF metrics but performs poorly in PI, suggesting that its fusion outputs may not fully align with human visual preferences. This discrepancy is possibly caused by excessive enhancement that introduces noticeable artifacts, thereby degrading the perceived image quality. Moreover, due to its unique methodological design, GIFNet demonstrates strong performance only in the specific task of *multi-focus image fusion*, with limited applicability and generalization across other fusion scenarios. In contrast, our method, without tailoring to any particular task, consistently delivers promising results across all fusion tasks. Overall, Mask-DiFuser achieves a well-balanced trade-off between detail preservation and natural visual appearance, highlighting its effectiveness in multi-focus image fusion.

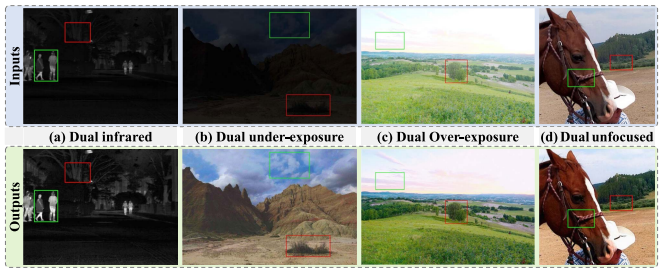


Fig. 13. Visual examples of generative priors applied to different image fusion and enhancement tasks.

H. Extended Experiments and Discussions

1) Generative Prior Visualization: As mentioned, the pre-trained masked diffusion model implicitly learns the priors (normal exposure, high contrast, high saturation, and clarity) of high-quality images. This enables our Mask-DiFuser to potentially present results in a more natural manner even if inputs suffer some degradations. To observe this property more intuitively, we dual-input an image with specific degradations into Mask-DiFuser. As shown in Fig. 13(a), when two infrared images are fed into Mask-DiFuser, it enhances the contrast of the input, making pedestrians and branches more distinct. Moreover, Fig. 13(b) and (c) demonstrate that when inputting two underexposed (or overexposed) images, Mask-DiFuser corrects the exposure levels, yielding satisfactory outputs with normal exposure. Finally, Fig. 13(d) illustrates that when two unfocused images are input, the model sharpens the input and presents it with appropriate saturation.

2) Object Detection: Effective information aggregation not only aids in visual perception but also advances machine vision. Thus, we validate the effectiveness of Mask-DiFuser by evaluating object detection performance on the MSRS dataset with the pre-trained YOLOv5 [74]. As shown in Fig. 14, the detector accurately identifies all objects in our fused images. Notably,

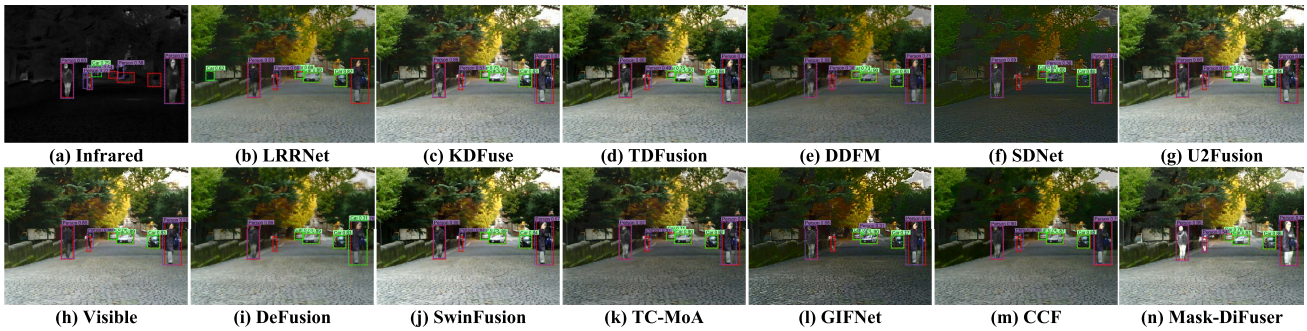


Fig. 14. Visual comparison of object detection on the MSRS dataset.

TABLE VII
QUANITATIVE RESULTS OF OBJECT DETECTION ON THE MSRS DATASET

Methods	AP@0.50			AP@0.65			AP@0.80			AP@0.95			mAP@[0.5:0.95]		
	Per.	Car	Avg.	Per.	Car	Avg.	Per.	Car	Avg.	Per.	Car	Avg.	Per.	Car	Avg.
LRRNet [19]	0.904	0.955	0.930	0.851	0.915	0.883	0.524	0.867	0.695	0.030	0.130	0.080	0.607	0.797	0.702
DDFM [41]	0.946	0.960	0.953	0.916	0.932	0.924	0.661	0.797	0.729	0.013	0.107	0.060	0.670	0.779	0.725
KDFuse [52]	0.936	0.934	0.935	0.907	0.917	0.912	0.635	0.834	0.735	0.010	0.127	0.068	0.674	0.777	0.725
TDFusion [53]	0.945	0.968	0.957	0.923	0.922	0.923	0.661	0.853	0.757	0.024	0.110	0.067	0.672	0.780	0.726
SDNet [34]	0.954	0.913	0.933	**0.932**	0.913	0.923	**0.706**	0.856	0.781	0.025	0.106	0.065	**0.696**	0.774	0.735
U2Fusion [11]	0.935	0.933	0.934	0.892	0.902	0.897	0.569	0.799	0.684	0.015	0.115	0.065	0.631	0.771	0.701
DeFusion [12]	0.936	0.940	0.938	0.930	0.920	0.925	0.640	0.827	0.734	0.020	**0.136**	0.078	0.673	0.780	0.726
SwinFusion	0.917	0.931	0.924	0.888	0.913	0.900	0.566	0.830	0.698	0.007	0.102	0.054	0.626	0.767	0.696
TC-MoA [20]	0.948	0.941	0.945	0.927	0.941	**0.934**	0.684	0.850	0.767	0.020	0.081	0.050	0.684	0.798	0.741
GIFNet [54]	0.928	0.965	0.946	0.892	0.937	0.915	0.612	0.862	0.737	0.017	0.120	0.069	0.646	0.794	0.720
CCF [55]	0.922	**0.971**	0.946	0.895	**0.943**	0.919	0.685	0.798	0.741	0.023	0.106	0.064	0.665	0.788	0.726
Mask-DiFuser	**0.957**	0.970	**0.963**	0.916	0.919	0.917	0.692	**0.896**	**0.794**	**0.037**	0.125	**0.081**	0.687	**0.800**	**0.744**

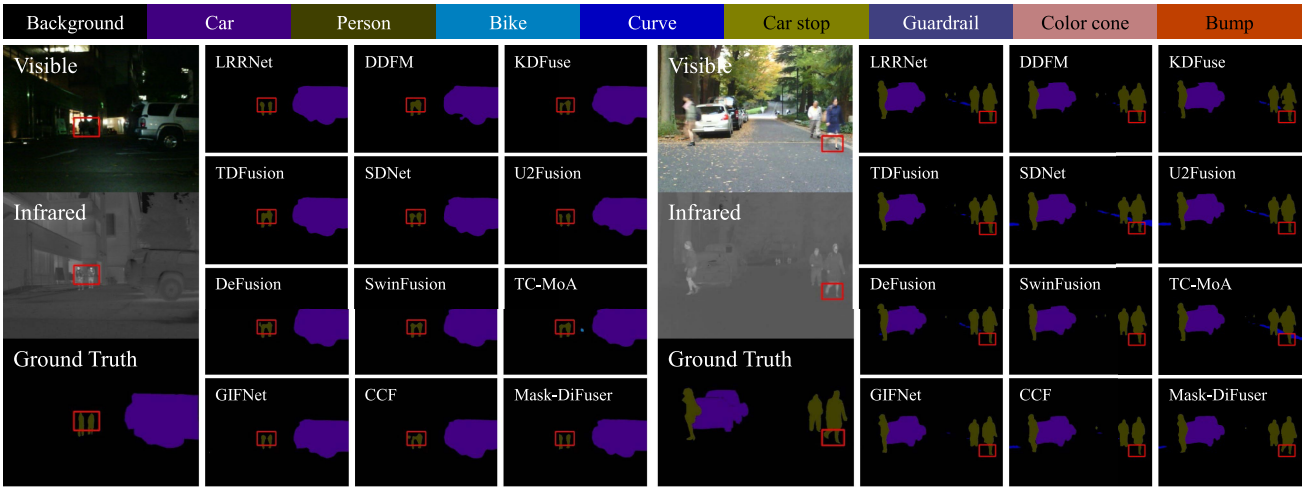
TABLE VIII
QUANTITATIVE RESULTS OF SEMANTIC SEGMENTATION ON THE MFNET DATASET. THE BEST AND SECOND-BEST MIOU SCORES ARE HIGHLIGHTED IN RED \COLORBOX[HTML]{FFC7CE}{RED} AND PURPLE \COLORBOX[HTML]{D9D9FF}{PURPLE}, RESPECTIVELY. KEY ABBREVIATIONS: BKG (BACKGROUND), C. STOP (CAR STOP), GUARD. (GUARDRAIL).

Methods	Bkg	Car	Person	Bike	Curve	C. Stop	Guard.	Cone	Bump	mIoU
LRRNet [19]	98.06	87.83	71.95	61.46	43.38	26.76	4.44	48.63	50.79	54.81
DDFM [41]	98.00	86.95	71.71	63.14	40.56	31.14	9.52	53.57	43.21	55.31
KDFuse [52]	98.17	87.79	72.97	64.05	46.03	22.39	5.17	50.12	50.37	55.23
TDFusion [53]	98.13	88.81	71.89	64.97	45.46	29.46	3.29	53.69	32.85	54.28
SDNet [34]	98.06	88.59	73.28	64.75	44.78	29.44	3.94	55.62	51.97	56.71
U2Fusion [11]	98.06	87.13	70.29	64.60	40.17	35.88	0.08	50.29	48.63	55.01
DeFusion [12]	98.10	87.69	72.81	65.61	45.86	23.93	4.18	51.94	42.49	54.73
SwinFusion [13]	98.11	87.75	72.80	64.73	46.28	29.59	6.68	50.76	46.40	55.90
TC-MoA [20]	98.16	88.56	72.51	65.38	43.76	31.27	6.34	51.11	53.26	56.71
GIFNet [54]	98.10	88.12	70.94	64.22	36.78	32.30	8.41	55.37	54.32	56.51
CCF [55]	97.95	88.17	72.12	63.26	42.08	31.89	0.28	46.55	44.10	54.05
Mask-DiFuser	98.21	88.25	74.30	64.98	45.42	34.24	6.95	53.93	52.89	**57.69**

our method enhances thermal targets with the generative priors from high-quality images, thus presenting higher confidence in pedestrian detection. Table VII further quantifies the efficacy of our method in downstream tasks, showing comparable average precision (AP) to other methods across various IoU thresholds.

3) Semantic Segmentation: We further validate the effectiveness of Mask-DiFuser by evaluating its performance on the semantic segmentation task using the MFNet dataset [75], with the retrained SegNeXt [76] employed as the segmentation backbone. Importantly, MFNet, as the parent dataset of MSRS, serves as a more comprehensive and representative benchmark for semantic segmentation, encompassing a wider variety of challenging scenes, such as low-contrast conditions.

The qualitative and quantitative results are presented in Fig. 15 and Table VIII, respectively. It can be observed that our method significantly improves segmentation quality compared to other



(a) Segmentation results for night scenes (b) Segmentation results for daytime scenes

Fig. 15. Visiual comparison of semantic segmentation on the MFNet dataset.

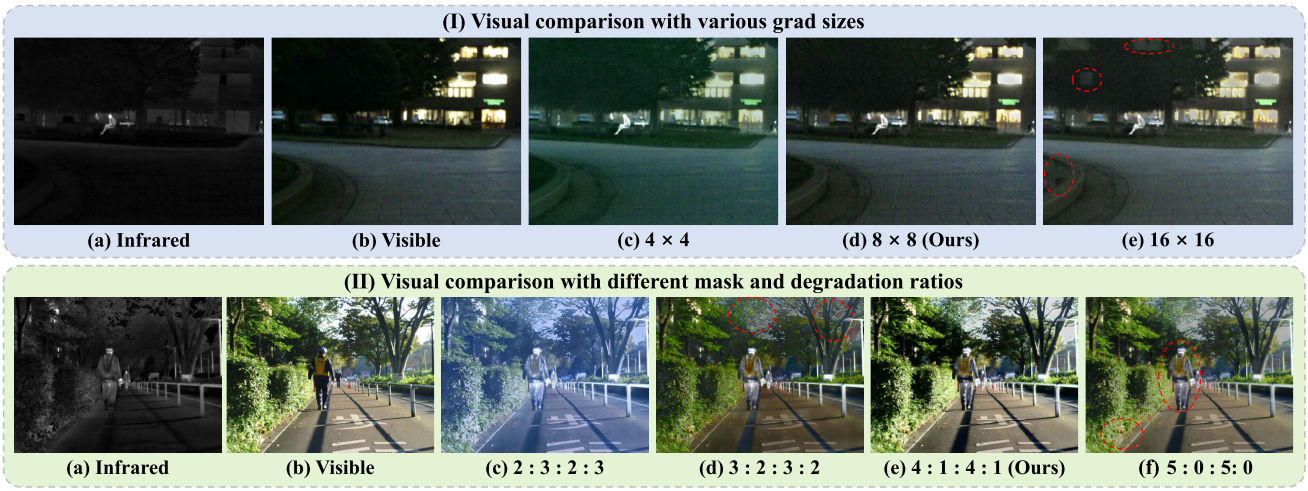


Fig. 16. Visual comparison with different grid sizes and mask ratios ($\mathcal{M}_1 : \mathcal{M}_2 : \mathcal{M}_3 : \mathcal{M}_4$) on the MSRS datasets.

TABLE IX
QUANTITATIVE RESULTS WITH VARIOUS GRID SIZES ON THE MSRS AND SICE DATASETS

Grid Sizes	IVIF (MSRS Dataset)						MEF (SICE Dataset)					
	EN ↑	AG ↑	SF ↑	SD ↑	SCD ↑	PI ↓	EN ↑	AG ↑	SF ↑	SD ↑	SCD ↑	PI ↓
4×4	6.94	4.38	9.55	41.55	1.8	**2.87**	7.44	7.19	15.88	60.92	1.52	**2.22**
8×8 **(Ours)**	6.98	4.60	**10.82**	**47.35**	**1.86**	2.94	7.50	8.11	**18.07**	**67.79**	**1.65**	2.27
16×16	**7.13**	**4.69**	10.576	46.79	1.76	3.05	**7.54**	**8.22**	18.06	66.21	1.53	2.28

fusion approaches, particularly in preserving object boundaries. Moreover, our method not only achieves the highest mean Intersection over Union (mIoU) but also attains the best IoU for the person category. The improvement is primarily driven by the ability of the proposed method to learn high-contrast priors from high-quality images, which in turn enhances the contrast of thermal targets and supplies more informative semantic features for subsequent tasks.

4) Effects of Grid Sizes: As shown in Fig. 16(I) and Table IX, we explore the impact of different grid sizes ($\mathbf{4 \times 4}$, $\mathbf{8 \times 8}$ and $\mathbf{16 \times 16}$) on fusion performance. Experimental results demonstrate that our default setting $\mathbf{8 \times 8}$ delivers a balanced performance across all metrics. Although increasing the grid size to $\mathbf{16 \times 16}$ can improve some quantitative metrics (e.g., EN from 6.98 to 7.13 and AG from 4.60 to 4.69), the fused images introduce noticeable block artifacts demonstrated in Fig. 16(I)(e),

TABLE X
QUANTITATIVE RESULTS WITH DIFFERENT MASK RATIONS ($\mathcal{M}_1 : \mathcal{M}_2 : \mathcal{M}_3 : \mathcal{M}_4$) ON THE MSRS AND SICE DATASETS

$\mathcal{M}_1 : \mathcal{M}_2 : \mathcal{M}_3 : \mathcal{M}_4$	IVIF (MSRS Dataset)						MEF (SICE Dataset)					
	EN ↑	AG ↑	SF ↑	SD ↑	SCD ↑	PI ↓	EN ↑	AG ↑	SF ↑	SD ↑	SCD ↑	PI ↓
2 : 3 : 2 : 3	6.43	3.69	8.92	33.82	1.49	3.17	7.10	6.59	14.78	45.22	0.77	2.23
3 : 2 : 3 : 2	6.23	3.33	8.41	31.90	1.48	3.23	6.96	5.97	13.58	45.83	0.94	2.26
4 : 1 : 4 : 1 **(Ours)**	**6.98**	**4.60**	**10.82**	**47.35**	**1.86**	**2.94**	**7.50**	**8.11**	**18.07**	**67.79**	**1.65**	2.27
5 : 0 : 5 : 0	6.48	3.47	8.31	36.59	1.76	3.48	7.36	6.72	15.12	57.05	1.45	**2.20**

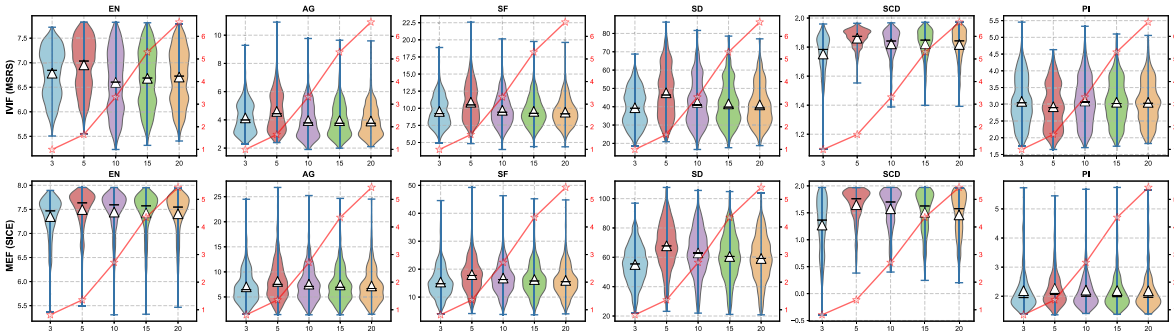


Fig. 17. Quantitative analysis of sampling steps. The red dash denotes inference time. Violin plots display the distribution of metrics across various sampling steps, with white triangles and black lines indicating mean and median values.

TABLE XI
COMPARISON OF COMPUTATIONAL COST BETWEEN OUR METHOD AND SOTA METHODS

	LRRNet	DDFM	KDFuse	TDFusion	SDNet	U2Fusion	DeFusion	SwinFusion	TC-MoA	GIFNet	CCF	Ours
Parm. (M)	0.05	552.66	2.492	0.059	0.067	0.66	7.87	0.97	340.35	0.613	552.66	171.26
Flops (G)	14.2	5220.5	765.434	18.211	41.3	405.2	71.6	307.5	3932.1	186.590	5220.5	4667.4
Times (s)	0.014	34.502	0.301	0.070	0.016	0.082	0.086	1.29	0.432	0.1989	210.883	1.655

leading to a decline in visuals. Conversely, reducing the grid size to $\mathbf{4} \times \mathbf{4}$ results in degradation of multiple metrics (e.g., SF from 10.82 to 9.55 and SCD from 1.86 to 1.80), and introduces color distortions as revealed in Fig. 16(I)(c), indicating that it struggles to capture sufficient details. This phenomenon further indicate that our chosen grid size ($\mathbf{8} \times \mathbf{8}$) obtains the best trade-off between quantitative performance and visual quality.

5) Effects of Mask Ratios: Fig. 16(II) and Table X illustrate the impact of various mask and degradation ratios (*i.e.,* $\mathcal{M}_1 : \mathcal{M}_2 : \mathcal{M}_3 : \mathcal{M}_4$) on fusion performance. From the qualitative and quantitative results, we can find that increasing the degradation ratio (e.g., $\mathbf{5} : \mathbf{0} : \mathbf{5} : \mathbf{0}$) forces the model to be unable to adapt to some degradation, such as a certain degree of blurring in the fusion results in Fig. 16(II)(f), and a decline in most metrics. Inversely, reducing the degradation ratio (e.g., $\mathbf{2} : \mathbf{3} : \mathbf{2} : \mathbf{3}$) weakens the generative ability of diffusion models, also resulting in degraded fusion performance, the most intuitive of which is the serious color distortion in Fig. 16(II)(c) and the dimming of the overall style of Fig. 16(II)(d). In conclusion, our default setting ($\mathbf{4} : \mathbf{1} : \mathbf{4} : \mathbf{1}$) achieves the best overall performance, balancing fusion quality and complementary information utilization.

6) Effects of Sampling Steps: As illustrated in Fig. 17, there is a nearly linear relationship between inference time and the number of sampling steps. It is worth noting that our model reaches the optimal values on most metrics when the sampling step is set to 5. Consequently, we employ the DDIM sampling strategy with a step of 5 to generate fusion results, which can balance the fusion quality and computational efficiency.

7) Comparison of Computational Cost: Table XI presents a detailed comparison of the computational costs between Mask-DiFuser and other algorithms, highlighting that our method incurs costs similar to those of Transformer approaches, indicating its substantial efficiency in managing complex fusion tasks. Notably, Mask-DiFuser demonstrates a significant runtime advantage over both DDFM and CCF, delivering faster processing without sacrificing performance. While our method may lag behind lightweight methods in terms of operational efficiency, its performance gains are remarkable, making it a compelling choice for cases where both high performance and reasonable computational cost are essential.

I. Ablation Studies

To validate the effectiveness of the specific designs in our proposed Mask-DiFuser, we conduct ablation studies on the IVIF (MSRS) and MEF (SICE) tasks, involving six configurations: (I) **w/o Content**: removing the content branch; (II) **w/o**

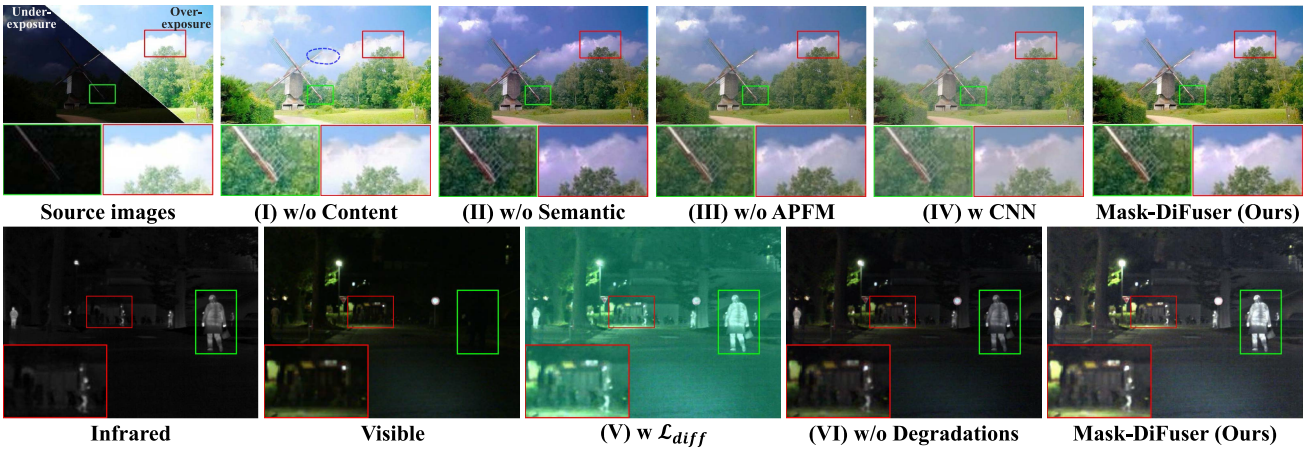


Fig. 18. Visual comparison of ablation studies on the MSRS and SICE datasets.

TABLE XII
QUANTITATIVE COMPARISON OF ABLATION STUDIES ON THE MSRS AND SICE DATASETS

	IVIF (MSRS Dataset)						MEF (SICE Dataset)					
	EN ↑	AG ↑	SF ↑	SD ↑	SCD ↑	PI ↓	EN ↑	AG ↑	SF ↑	SD ↑	SCD ↑	PI ↓
I	6.72	3.73	9.35	43.27	1.85	3.57	7.27	7.77	17.72	64.38	1.46	2.35
II	6.66	4.13	9.66	42.16	1.82	**2.91**	7.45	7.58	16.90	62.80	1.52	**2.17**
III	6.71	3.95	8.10	34.62	1.69	3.08	7.33	6.42	13.91	52.90	1.22	2.24
IV	6.75	3.59	8.43	36.99	1.71	3.26	7.11	5.67	12.85	47.93	0.95	2.37
V	6.71	3.35	7.30	34.33	1.62	3.67	7.17	5.05	11.44	49.96	1.12	2.49
VI	6.48	3.47	8.31	36.59	1.76	3.48	7.36	6.72	15.12	57.05	1.45	2.20
Ours	**6.98**	**4.60**	**10.82**	**47.35**	**1.86**	2.94	**7.50**	**8.11**	**18.07**	**67.79**	**1.65**	2.27

Semantic: removing the semantic branch; (III) **w/o APFM**: replacing APFM with a simple average weighted fusion strategy for coarse fusion; (IV) **w CNN**: using CNN as the backbone network instead of the diffusion model; (V) **w** $\mathcal{L}_{diff}$: training the model solely with the diffusion loss $\mathcal{L}_{diff}$, excluding other image-level consistency losses; and (VI) **w/o Degradations**: applying the masking scheme in (1) without introducing degradations. As shown in Fig. 18, removing any of these critical components leads to noticeable performance degradation. In particular, omitting the content branch introduces significant patch block effects in the fusion results, as highlighted in the blue box. When CNN is employed as the backbone, the model fails to capture high-quality priors from natural images, leading to less vivid fused images. Additionally, training the network solely with the diffusion loss leads to noticeable color distortions in the fusion results. If degradations are not introduced into the masking scheme, the generative capacity of the masked diffusion model becomes somewhat restricted. The quantitative results in Table XII further corroborate the importance of these key designs, despite a slight decline in the PI metric.

V. CONCLUSION

This work addresses a common challenge in existing image fusion techniques, *i.e.,* the lack of ground truth for fused images. To tackle this issue, we propose Mask-DiFuser, a novel unified image fusion framework based on masked image modeling and diffusion models. On the one hand, a tailored masking scheme is devised to ingeniously transform the GT-unavailable image fusion task into a self-supervised masked image restoration task. On the other hand, a masked diffusion model learning generative priors from high-quality images via self-supervised learning is developed to synthesize impressive fusion results closely aligned with human perception. Extensive experiments on the IVIF, MIF, MEF, and MFF tasks demonstrate the versatility and superiority of our method.

REFERENCES

[1] F. Bao et al., "Heat-assisted detection and ranging," *Nature*, vol. 619, no. 7971, pp. 743–748, 2023.
[2] D. K. Jain, X. Zhao, G. González-Almagro, C. Gan, and K. Kotecha, "Multimodal pedestrian detection using metaheuristics with deep convolutional neural network in crowded scenes," *Inf. Fusion*, vol. 95, pp. 401–414, 2023.
[3] J. Zhang, H. Liu, K. Yang, X. Hu, R. Liu, and R. Stiefelhagen, "CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 12, pp. 14679–14694, Dec. 2023.
[4] K. R. Prabhakar, V. Sai Srikar, and R. Venkatesh Babu, "Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4714–4722.
[5] H. Li and X.-J. Wu, "DenseFuse: A fusion approach to infrared and visible images," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2614–2623, May 2019.

[6] J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang, "FusionGAN: A generative adversarial network for infrared and visible image fusion," *Inf. Fusion*, vol. 48, pp. 11–26, 2019.
[7] H. Xu and J. Ma, "Emfusion: An unsupervised enhanced medical image fusion network," *Inf. Fusion*, vol. 76, pp. 177–186, 2021.
[8] L. Tang, J. Yuan, and J. Ma, "Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network," *Inf. Fusion*, vol. 82, pp. 28–42, 2022.
[9] Y. Zhang, Y. Liu, P. Sun, H. Yan, X. Zhao, and L. Zhang, "IFCNN: A general image fusion framework based on convolutional neural network," *Inf. Fusion*, vol. 54, pp. 99–118, 2020.
[10] H. Zhang, H. Xu, Y. Xiao, X. Guo, and J. Ma, "Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12797–12804.
[11] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2Fusion: A unified unsupervised image fusion network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 502–518, Jan. 2022.
[12] P. Liang, J. Jiang, X. Liu, and J. Ma, "Fusion from decomposition: A self-supervised decomposition approach for image fusion," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 719–735.
[13] J. Ma, L. Tang, F. Fan, J. Huang, X. Mei, and Y. Ma, "Swinfusion: Cross-domain long-range learning for general image fusion via swin transformer," *IEEE/CAA J. Automatica Sinica*, vol. 9, no. 7, pp. 1200–1217, Jul. 2022.
[14] L. Tang, H. Zhang, H. Xu, and J. Ma, "Deep learning-based image fusion: A survey," *J. Image Graph.*, vol. 28, no. 1, pp. 3–36, 2023.
[15] J. Liu et al., "Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5802–5811.
[16] H. Zhang, L. Tang, X. Xiang, X. Zuo, and J. Ma, "Dispel darkness for better fusion: A controllable visual enhancer based on cross-modal conditional adversarial learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 26487–26496.
[17] J. Cai, S. Gu, and L. Zhang, "Learning a deep single image contrast enhancer from multi-exposure images," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 2049–2062, Apr. 2018.
[18] H. Zhang, Z. Le, Z. Shao, H. Xu, and J. Ma, "MFF-GAN: An unsupervised generative adversarial network with adaptive and gradient joint constraints for multi-focus image fusion," *Inf. Fusion*, vol. 66, pp. 40–53, 2021.
[19] H. Li, T. Xu, X.-J. Wu, J. Lu, and J. Kittler, "LRRNet: A novel representation learning guided fusion network for infrared and visible images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 11040–11052, Sep. 2023.
[20] P. Zhu, Y. Sun, B. Cao, and Q. Hu, "Task-customized mixture of adapters for general image fusion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 7099–7108.
[21] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16000–16009.
[22] Z. Xie et al., "SimMIM: A simple framework for masked image modeling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 9653–9663.
[23] J. Liu, R. Lin, G. Wu, R. Liu, Z. Luo, and X. Fan, "CoCoNet: Coupled contrastive learning network with multi-level feature ensemble for multi-modality image fusion," *Int. J. Comput. Vis.*, vol. 132, no. 5, pp. 1748–1775, 2024.
[24] Q. Huang, G. Wu, Z. Jiang, W. Fan, B. Xu, and J. Liu, "Leveraging a self-adaptive mean teacher model for semi-supervised multi-exposure image fusion," *Inf. Fusion*, vol. 112, 2024, Art. no. 102534.
[25] J. Zhang, Q. Liao, H. Ma, J.-H. Xue, W. Yang, and S. Liu, "Exploit the best of both end-to-end and map-based methods for multi-focus image fusion," *IEEE Trans. Multimedia*, vol. 26, pp. 6411–6423, 2024.
[26] D. Wang, J. Liu, X. Fan, and R. Liu, "Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration," in *Proc. Int. Joint Conf. Artif. Intell.*, 2022, pp. 3508–3515.
[27] H. Xu, J. Yuan, and J. Ma, "MURF: Mutually reinforcing multi-modal image registration and fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 12148–12166, Oct. 2023.
[28] L. Tang, Y. Deng, Y. Ma, J. Huang, and J. Ma, "SuperFusion: A versatile image registration and fusion network with semantic awareness," *IEEE/CAA J. Automatica Sinica*, vol. 9, no. 12, pp. 2121–2137, Dec. 2022.
[29] H. Li, J. Liu, Y. Zhang, and Y. Liu, "A deep learning framework for infrared and visible image fusion without strict registration," *Int. J. Comput. Vis.*, vol. 132, no. 5, pp. 1625–1644, 2024.
[30] L. Tang, X. Xiang, H. Zhang, M. Gong, and J. Ma, "DivFusion: Darkness-free infrared and visible image fusion," *Inf. Fusion*, vol. 91, pp. 477–493, 2023.
[31] X. Yi, H. Xu, H. Zhang, L. Tang, and J. Ma, "Text-IF: Leveraging semantic text guidance for degradation-aware and interactive image fusion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 27026–27035.
[32] L. Tang, Y. Deng, X. Yi, Q. Yan, Y. Yuan, and J. Ma, "DRMF: Degradation-robust multi-modal image fusion via composable diffusion prior," in *Proc. ACM Int. Conf. Multimedia*, 2024, pp. 8546–8555.
[33] W. Zhao, S. Xie, F. Zhao, Y. He, and H. Lu, "Metafusion: Infrared and visible image fusion via meta-feature embedding from object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 13955–13 965.
[34] H. Zhang and J. Ma, "SDNet: A versatile squeeze-and-decomposition network for real-time image fusion," *Int. J. Comput. Vis.*, vol. 129, no. 10, pp. 2761–2785, 2021.
[35] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 6840–6851.
[36] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10684–10695.
[37] C.-H. Lin et al., "Magic3D: High-resolution text-to-3D content creation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 300–309.
[38] G. Kim, T. Kwon, and J. C. Ye, "DiffusionCLIP: Text-guided diffusion models for robust image manipulation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 2426–2435.
[39] B. Xia et al., "Diffir: Efficient diffusion model for image restoration," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 13095–13105.
[40] X. Yi, H. Xu, H. Zhang, L. Tang, and J. Ma, "Diff-Retinex: Rethinking low-light image enhancement with a generative diffusion model," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 12302–12311.
[41] Z. Zhao et al., "DDFM: Denoising diffusion model for multi-modality image fusion," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 8082–8093.
[42] J. Yue, L. Fang, S. Xia, Y. Deng, and J. Ma, "Dif-Fusion: Towards high color fidelity in infrared and visible image fusion with diffusion models," *IEEE Trans. Image Process.*, vol. 32, pp. 5705–5720, 2023.
[43] H. Zhang, L. Cao, and J. Ma, "Text-DiFuse: An interactive multi-modal image fusion framework based on text-modulated diffusion model," in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, pp. 39552–39572.
[44] X. Yi, L. Tang, H. Zhang, H. Xu, and J. Ma, "Diff-IF: Multi-modality image fusion via diffusion model with fusion knowledge prior," *Inf. Fusion*, vol. 110, 2024, Art. no. 102450.
[45] J. Devlin, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
[46] T. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 1877–1901.
[47] O. Henaff, "Data-efficient image recognition with contrastive predictive coding," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 4182–4192.
[48] H. Bao, L. Dong, S. Piao, and F. Wei, "BEiT: BERT pre-training of image transformers," in *Proc. Int. Conf. Learn. Representations*, 2022.
[49] L. Qu, S. Liu, M. Wang, and Z. Song, "TransMEF: A transformer-based multi-exposure image fusion framework using self-supervised multi-task learning," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 2126–2134.
[50] Y. Yin, D. Xu, C. Tan, P. Liu, Y. Zhao, and Y. Wei, "CLE diffusion: Controllable light enhancement diffusion model," in *Proc. ACM Int. Conf. Multimedia*, 2023, pp. 8145–8156.
[51] E. Agustsson and R. Timofte, "NTIRE 2017 challenge on single image super-resolution: Dataset and study," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 126–135.
[52] C. Yang, X. Luo, Z. Zhang, Z. Chen, and X.-j. Wu, "KDFuse: A high-level vision task-driven infrared and visible image fusion method based on cross-domain knowledge distillation," *Inf. Fusion*, vol. 118, 2025, Art. no. 102944.
[53] H. Bai et al., "Task-driven image fusion with learnable fusion loss," in *Proc. Comput. Vis. Pattern Recognit. Conf.*, 2025, pp. 7457–7468.
[54] C. Cheng et al., "One model for all: Low-level task interaction is a key to task-agnostic image fusion," in *Proc. Comput. Vis. Pattern Recognit. Conf.*, 2025, pp. 28102–28112.
[55] B. Cao, X. Xu, P. Zhu, Q. Wang, and Q. Hu, "Conditional controllable image fusion," in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, pp. 120311–120335.
[56] L. Tang, J. Yuan, H. Zhang, X. Jiang, and J. Ma, "PIAFusion: A progressive infrared and visible image fusion network based on illumination aware," *Inf. Fusion*, vol. 83, pp. 79–92, 2022.

[57] M. Brown and S. Süsstrunk, "Multi-spectral sift for scene category recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 177–184.
[58] M. Morimatsu, Y. Monno, M. Tanaka, and M. Okutomi, "Monochrome and color polarization demosaicking using edge-aware residual interpolation," in *Proc. IEEE Int. Conf. Image Process.*, 2020, pp. 2571–2575.
[59] E. D. Vidoni, "The whole brain atlas: www.med.harvard.edu/aanlib," *J. Neurologic Phys. Ther.*, vol. 36, no. 2, 2012, Art. no. 108.
[60] X. Zhang, "Benchmarking and comparing multi-exposure image fusion algorithms," *Inf. Fusion*, vol. 74, pp. 111–131, 2021.
[61] M. Nejati, S. Samavi, and S. Shirani, "Multi-focus image fusion using dictionary-based sparse representation," *Inf. Fusion*, vol. 25, pp. 72–84, 2015.
[62] H. Xu, Y. Sun, X. Mei, X. Tian, and J. Ma, "Attention-guided polarization image fusion using salient information distribution," *IEEE Trans. Comput. Imag.*, vol. 8, pp. 1117–1130, 2022.
[63] Y. Luo, J. Zhang, and C. Li, "CPIFuse: Toward realistic color and enhanced textures in color polarization image fusion," *Inf. Fusion*, vol. 120, 2025, Art. no. 103111.
[64] P. Mu, G. Wu, J. Liu, Y. Zhang, X. Fan, and R. Liu, "Learning to search a lightweight generalized network for medical image fusion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 7, pp. 5921–5934, Jul. 2024.
[65] D. He, W. Li, G. Wang, Y. Huang, and S. Liu, "MMIF-INet: Multimodal medical image fusion by invertible network," *Inf. Fusion*, vol. 114, 2025, Art. no. 102666.
[66] G. Wu, H. Fu, J. Liu, L. Ma, X. Fan, and R. Liu, "Hybrid-supervised dual-search: Leveraging automatic learning for loss-free multi-exposure image fusion," in *Proc. AAAI Conf. Artif. Intell.*, 2024, pp. 5985–5993.
[67] Z. Liu, J. Liu, G. Wu, Z. Chen, X. Fan, and R. Liu, "Searching a compact architecture for robust multi-exposure image fusion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 7, pp. 6224–6237, Jul. 2024.
[68] W. Tang and F. He, "EAT: Multi-exposure image fusion with adversarial learning and focal transformer," *IEEE Trans. Multimedia*, vol. 27, pp. 3744–3754, 2025.
[69] X. Hu, J. Jiang, X. Liu, and J. Ma, "ZMFF: Zero-shot multi-focus image fusion," *Inf. Fusion*, vol. 92, pp. 127–138, 2023.
[70] H. Lin et al., "Fusion2Void: Unsupervised multi-focus image fusion based on image inpainting," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 4, pp. 3328–3341, Apr. 2025.
[71] J. Wang, H. Qu, Z. Zhang, and M. Xie, "New insights into multi-focus image fusion: A fusion method based on multi-dictionary linear sparse representation and region fusion model," *Inf. Fusion*, vol. 105, 2024, Art. no. 102230.
[72] V. Aslantas and E. Bendes, "A new image quality metric for image fusion: The sum of the correlations of differences," *AEU- Int. J. Electron. Commun.*, vol. 69, no. 12, pp. 1890–1896, 2015.
[73] Y. Blau, R. Mechrez, R. Timofte, T. Michaeli, and L. Zelnik-Manor, "The 2018 PIRM challenge on perceptual image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 334–355.
[74] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.
[75] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, and T. Harada, "MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2017, pp. 5108–5115.
[76] M.-H. Guo, C.-Z. Lu, Q. Hou, Z. Liu, M.-M. Cheng, and S.-M. Hu, "SegNeXt: Rethinking convolutional attention design for semantic segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 1140–1156.

Linfeng Tang received the BE degree from the School of Computer Science and Engineering, Central South University, Changsha, China, in 2020. He is currently working toward the PhD degree with Electronic Information School, Wuhan University. His research interests include computer vision, machine learning, and pattern recognition.

Chuyu Li received the BS degree from the School of Communication Engineering, Jilin University, Changchun, China, in 2024. He is currently working toward the MS degree with Multi-Spectral Vision Processing Lab, Wuhan University. His current research interests include computer vision and pattern recognition.

Jiayi Ma (Senior Member, IEEE) received the BS degree in information and computing science, and the PhD degree in control science and engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2008 and 2014, respectively. He is currently a professor with Electronic Information School, Wuhan University, Wuhan. He has coauthored more than 400 refereed journal and conference papers, including *Cell*, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, and IJCV. He is the recipient of Information Fusion Best Paper Award 2024 and Hsue-shen Tsien Paper Award 2023. He is the Area Editor of *Information Fusion*, Associate Editor for *IEEE/CAA Journal of Automatica Sinica*, *Neurocomputing*, *Geo-spatial Information Science*, and *Image and Vision Computing*, and Youth Editor of *The Innovation* and *Fundamental Research*.