DDFM: Denoising Diffusion Model for Multi-Modality Image Fusion

Zixiang Zhao[1,2] Haowen Bai[1] Yuanzhi Zhu[2] Jiangshe Zhang[1*] Shuang Xu[3]
Yulun Zhang[2] Kai Zhang[2] Deyu Meng[1,5] Radu Timofte[2,4] Luc Van Gool[2]

[1]Xi'an Jiaotong University [2]Computer Vision Lab, ETH Zürich
[3]Northwestern Polytechnical University [4]University of Würzburg
[5]Macau University of Science and Technology

zixiangzhao@stu.xjtu.edu.cn, jszhang@mail.xjtu.edu.cn

Abstract

Multi-modality image fusion aims to combine different modalities to produce fused images that retain the complementary features of each modality, such as functional highlights and texture details. To leverage strong generative priors and address challenges such as unstable training and lack of interpretability for GAN-based generative methods, we propose a novel fusion algorithm based on the denoising diffusion probabilistic model (DDPM). The fusion task is formulated as a conditional generation problem under the DDPM sampling framework, which is further divided into an unconditional generation subproblem and a maximum likelihood subproblem. The latter is modeled in a hierarchical Bayesian manner with latent variables and inferred by the expectation-maximization (EM) algorithm. By integrating the inference solution into the diffusion sampling iteration, our method can generate high-quality fused images with natural image generative priors and cross-modality information from source images. Note that all we required is an unconditional pre-trained generative model, and no fine-tuning is needed. Our extensive experiments indicate that our approach yields promising fusion results in infrared-visible image fusion and medical image fusion. The code is available at https://github.com/Zhaozixiang1228/MMIF-DDFM.

1. Introduction

Image fusion integrates essential information from multiple source images to create high-quality fused images [37, 70, 27, 42], encompassing various source image types like digital [20, 67, 74], multi-modal [58, 72], and remote sensing [62, 76]. This technology provides a clearer representation of objects and scenes, and has diverse applications such as saliency detection [43, 40, 41], object detection [12, 2, 10, 55], and semantic segmentation [28, 11, 56]. Among the different subcategories of image fusion, *Infrared-*

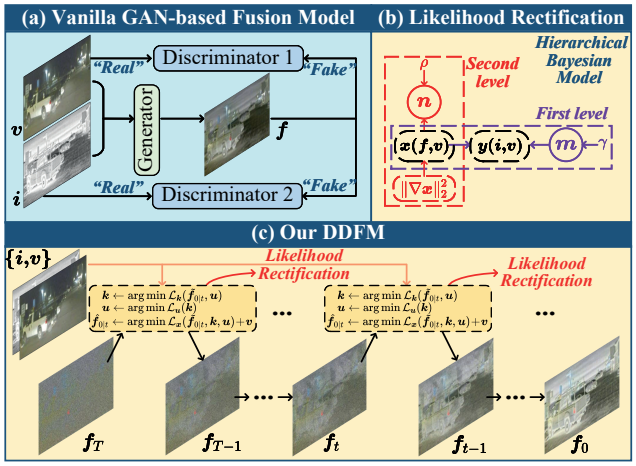

Figure 1: (a) Existing GAN-based fusion method workflow. (b) Graph of the hierarchical Bayesian model in likelihood rectification, linking the MMIF loss and our statistical inference model. (c) Our DDFM workflow: the unconditional diffusion sampling (UDS) module generates $\boldsymbol{f}_t$, while the likelihood rectification module, based on (b), rectifies UDS output with source image information.

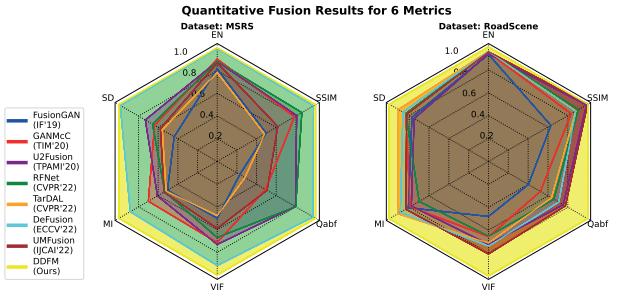

Figure 2: Visualization of results on MSRS [51] and RoadScene [59] in Tab. 1. Hexagons formed by lines of different colors represent the values of different methods across six metrics. Our DDFM (marked in yellow) outperforms all other methods.

*Corresponding author.

Visible image Fusion (IVF) and *Medical Image Fusion* (MIF) are particularly challenging in *Multi-Modality Image Fusion* (MMIF) since they focus on modeling cross-modality features and preserving critical information from all sensors and modalities. Specifically, in IVF, fused images aim to retain both thermal radiation from infrared images and detailed texture information from visible images, thereby avoiding the limitations of visible images being sensitive to illumination conditions and infrared images being noisy and low-resolution. While MIF can assist in diagnosis and treatment by fusing multiple medical imaging modalities for precise detection of abnormality locations [16, 9].

There have been numerous methods devised recently to address the challenges posed by MMIF [26, 65, 29], and generative models [7, 38] have been extensively utilized to model the distribution of fused images and achieve satisfactory fusion effects. Among them, models based on Generative Adversarial Networks (GANs) [34, 35, 33, 26] are dominant. The workflow of GAN-based models, illustrated in Fig. 1a, involves a generator that creates images containing information from source images, and a discriminator that determines whether the generated images are in a similar manifold to the source images. Although GAN-based methods have the ability to generate high-quality fused images, they suffer from unstable training, lack of interpretability and mode collapse, which seriously affect the quality of the generated samples. Moreover, as a black-box model, it is difficult to comprehend the internal mechanisms and behaviors of GANs, making it challenging to achieve controllable generation.

Recently, *Denoising Diffusion Probabilistic Models* (DDPM) [13] has garnered attention in the machine learning community, which generates high-quality images by modeling the diffusion process of restoring a noise-corrupted image towards a clean image. Based on the Langevin diffusion process, DDPM utilizes a series of reverse diffusion steps to generate promising synthetic samples [46]. Compared to GAN, DDPM does not require the discriminator network, thus mitigating common issues such as unstable training and mode collapse in GAN. Moreover, its generation process is interpretable, as it is based on denoising diffusion to generate images, enabling a better understanding of the image generation process [57].

Therefore, we propose a **D**enoising **D**iffusion image **F**usion **M**odel (**DDFM**), as shown in Fig. 1c. We formulate the conditional generation task as a DDPM-based posterior sampling model, which can be further decomposed into an unconditional generation diffusion problem and a maximum likelihood estimation problem. The former satisfies natural image prior while the latter is inferred to restrict the similarity with source images via likelihood rectification. Compared to discriminative approaches, modeling the natural image prior with DDPM enables better generation of details that are difficult to control by manually designed loss functions, resulting in visually perceptible images. As a generative method, DDFM achieves stable and controllable generation of fused images without discriminator, by applying likelihood rectification to the DDPM output.

Our contributions are organized in three aspects:

- We introduce a DDPM-based posterior sampling model for MMIF, consisting of an unconditional generation module and a conditional likelihood rectification module. The sampling of fused images is achieved solely by a pre-trained DDPM without fine-tuning.
- In likelihood rectification, since obtaining the likelihood explicitly is not feasible, we formulate the optimization loss as a probability inference problem involving latent variables, which can be solved by the EM algorithm. Then the solution is integrated into the DDPM loop to complete conditional image generation.
- Extensive evaluation of IVF and MIF tasks shows that DDFM consistently delivers favorable fusion results, effectively preserving both the structure and detail information from the source images, while also satisfying visual fidelity requirements.

2. Background

2.1. Score-based diffusion models

Score SDE formulation. Diffusion models aim to generate samples by reversing a predefined forward process that converts a clean sample $\boldsymbol{x}_0$ to almost Gaussian signal $\boldsymbol{x}_T$ by gradually adding noise. This forward process can be described by an Itô Stochastic Differential Equation (SDE) [49]:

$$\mathrm{d}\boldsymbol{x} = -\frac{\beta(t)}{2}\boldsymbol{x}_t\mathrm{d}t + \sqrt{\beta(t)}\mathrm{d}\boldsymbol{w}, \tag{1}$$

where $\mathrm{d}\boldsymbol{w}$ is standard Wiener process and $\beta(t)$ is predefined noise schedule that favors the variance-preserving SDE [49].

This forward process can be reversed in time and still in the form of SDE [1]:

$$\mathrm{d}\boldsymbol{x} = \left[-\tfrac{\beta(t)}{2}\boldsymbol{x}_t - \beta(t)\nabla_{\boldsymbol{x}_t}\log p_t(\boldsymbol{x_t})\right]\mathrm{d}t + \sqrt{\beta(t)}\mathrm{d}\overline{\boldsymbol{w}}, \tag{2}$$

where $\mathrm{d}\overline{\boldsymbol{w}}$ corresponds to the standard Wiener process running backward and the only unknown part $\nabla_{\boldsymbol{x}_t}\log p_t(\boldsymbol{x_t})$ can be modeled as the so-called *score function* $\boldsymbol{s}_\theta(\boldsymbol{x}_t, t)$ with denoising score matching methods, and this score function can be trained with the following objective [15, 48]:

$$\mathbb{E}_t\mathbb{E}_{\boldsymbol{x}_0}\mathbb{E}_{\boldsymbol{x}_t|\boldsymbol{x}_0}\left[\|\boldsymbol{s}_\theta(\boldsymbol{x}_t, t) - \nabla_{\boldsymbol{x}_t}\log p_{0t}(\boldsymbol{x}_t|\boldsymbol{x}_0)\|_2^2\right], \tag{3}$$

where t is uniformly sampled over $[0, T]$ and the data pair $(\boldsymbol{x}_0, \boldsymbol{x}_t) \sim p_0(\boldsymbol{x})p_{0t}(\boldsymbol{x}_t|\boldsymbol{x}_0)$.

Sampling with diffusion models. Specifically, an unconditional diffusion generation process starts with a random

noise vector $\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and updates according to the discretization of Eq. (2). Alternatively, we can understand the sampling process in the DDIM fashion [46], where the score function can also be considered to be a denoiser and predict the denoised $\tilde{\boldsymbol{x}}_{0|t}$ from any state $\boldsymbol{x}_t$ at iteration t:

$$\tilde{\boldsymbol{x}}_{0|t} = \frac{1}{\sqrt{\bar{\alpha}_t}}(\boldsymbol{x}_t + (1-\bar{\alpha}_t)\boldsymbol{s}_\theta(\boldsymbol{x}_t, t)), \tag{4}$$

and $\tilde{\boldsymbol{x}}_{0|t}$ denotes the estimation of $\boldsymbol{x}_0$ given $\boldsymbol{x}_t$. We use the same notation $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ following Ho *et al*. [13]. With this predicted $\tilde{\boldsymbol{x}}_{0|t}$ and the current state $\boldsymbol{x}_t$, $\boldsymbol{x}_{t-1}$ is updated from

$$\boldsymbol{x}_{t-1} = \frac{\sqrt{\alpha_t}\,(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\tilde{\boldsymbol{x}}_{0|t} + \tilde{\sigma}_t \boldsymbol{z}, \tag{5}$$

where $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\tilde{\sigma}_t^2$ is the variance which is usually set to 0. This sampled $\boldsymbol{x}_{t-1}$ is then fed into the next sampling iteration until the final image $\boldsymbol{x}_0$ is generated. Further details about this sampling process can be found in the supplementary material or the original paper [46].

Diffusion models applications. Recently, diffusion models have been improved to generate images with better quality than previous generative models like GANs [5, 39]. Moreover, diffusion models can be treated as a powerful generative prior and be applied to numerous conditional generation tasks. One representative work with diffusion models is stable diffusion which can generate images according to given text prompts [44]. Diffusion models are also applied to many low-level vision tasks. For instance, DDRM [19] performs diffusion sampling in the spectral space of degradation operator $\mathcal{A}$ to reconstruct the missing information in the observation $\boldsymbol{y}$. DDNM [64] shares a similar idea with DDRM by refining the null-space of the operator $\mathcal{A}$ iteratively for image restoration tasks. DPS [3] endorses Laplacian approximation to calculate the gradient of log-likelihood for posterior sampling and it is capable of many noisy non-linear inverse problems. In ΠGDM [47], the authors employ few approximations to make the log-likelihood term tractable and hence make it able to solve inverse problems with even non-differentiable measurements.

2.2. Multi-modal image fusion

The deep learning-based multi-modality image fusion algorithms achieve effective feature extraction and information fusion through the powerful fitting ability of neural networks. Fusion algorithms are primarily divided into two branches: generative methods and discriminative methods. For generative methods [34, 31, 35], particularly the GAN family, adversarial training [7, 36, 38] is employed to generate fusion images following the same distribution as the source images. For discriminative methods, auto encoder-based models [72, 23, 21, 28, 53, 22, 65] use encoders and decoders to extract features and fuse them on a high-dimensional manifold. Algorithm unfolding models [4, 6, 73, 63, 75, 24] combine traditional optimization methods and neural networks, balancing efficiency and interpretability. Unified models [59, 66, 58, 68, 18] avoid the problem of lacking training data and ground truth for specific tasks. Recently, CDDFuse [69] addresses cross-modality feature modeling and extracts modality-specific/shared features through a dual-branch Transformer-CNN architecture and correlation-driven loss, achieving promising fusion results in multiple fusion tasks. On the other hand, fusion methods have been combined with pattern recognition tasks such as semantic segmentation [50] and object detection [26] to explore the interactions with downstream tasks. Specifically, TarDAL [26] demonstrates an obvious advantage in dealing with challenge scenarios with high efficiency. Self-supervised learning [25] is employed to train fusion networks without paired images. Moreover, the pre-processing registration module [60, 14, 54, 61] can enhance the robustness for unregistered input images. Benefiting from the multi-modality data, MSIS [17] achieves realizable and outstanding stitching results.

2.3. Comparison with existing approaches

The methods most relevant to our model are optimization-based methods and GAN-based generative methods. Conventional optimization-based methods are often limited by manually designed loss functions, which may not be flexible enough to capture all relevant aspects and are sensitive to changes in the data distribution. While incorporating natural image priors can provide extra knowledge that cannot be modeled by the generation loss function alone. Then, in contrast to GAN-based generative methods, where unstable training and pattern collapse may occur, our DDFM achieves more stable and controllable fusion by rectifying the generation process towards source images and performing likelihood-based refinement in each iteration.

3. Method

In this section, we first present a novel approach for obtaining a fusion image by leveraging DDPM posterior sampling. Then, starting from the well-established loss function for image fusion, we derive a likelihood rectification approach for the unconditional DDPM sampling. Finally, we propose the DDFM algorithm, which embeds the solution of the hierarchical Bayesian inference into the diffusion sampling. In addition, the rationality of the proposed algorithm will be demonstrated. For brevity, we omit the derivations of some equations and refer interested readers to the *supplementary material*. It is worth noting that we use IVF as a case to illustrate our DDFM, and MIF can be carried out analogously to IVF.

3.1. Fusing images via diffusion posterior sampling

We first give the notation of the model formulation. Infrared, visible and fused images are denoted as $\boldsymbol{i} \in \mathbb{R}^{HW}$, $\boldsymbol{v} \in \mathbb{R}^{3HW}$ and $\boldsymbol{f} \in \mathbb{R}^{3HW}$, respectively.

We expect that the distribution of $\boldsymbol{f}$ given $\boldsymbol{i}$ and $\boldsymbol{v}$, *i.e.*, $p\left(\boldsymbol{f}|\boldsymbol{i},\boldsymbol{v}\right)$, can be modeled, thus $\boldsymbol{f}$ can be obtained by sampling from the posterior distribution. Inspired by Eq. (2), we can express the reverse SDE of diffusion process as:

$$d\boldsymbol{f} = \left[-\frac{\beta(t)}{2}\boldsymbol{f} - \beta(t)\nabla_{\boldsymbol{f}_t}\log p_t\left(\boldsymbol{f}_t|\boldsymbol{i},\boldsymbol{v}\right)\right]dt + \sqrt{\beta(t)}d\overline{\boldsymbol{w}}, \quad (6)$$

and the score function, *i.e.*, $\nabla_{\boldsymbol{f}_t}\log p_t(\boldsymbol{f}_t|\boldsymbol{i},\boldsymbol{v})$, can be calculated by:

$$\begin{aligned}\nabla_{\boldsymbol{f}_t}\log p_t(\boldsymbol{f}_t|\boldsymbol{i},\boldsymbol{v}) &= \nabla_{\boldsymbol{f}_t}\log p_t(\boldsymbol{f}_t) + \nabla_{\boldsymbol{f}_t}\log p_t(\boldsymbol{i},\boldsymbol{v}|\boldsymbol{f}_t) \\ &\approx \nabla_{\boldsymbol{f}_t}\log p_t(\boldsymbol{f}_t) + \nabla_{\boldsymbol{f}_t}\log p_t(\boldsymbol{i},\boldsymbol{v}|\tilde{\boldsymbol{f}}_{0|t})\end{aligned} \quad (7)$$

where $\tilde{\boldsymbol{f}}_{0|t}$ is the estimation of $\boldsymbol{f}_0$ given $\boldsymbol{f}_t$ from the unconditional DDPM. The equality comes from Bayes' theorem, and the approximate equation is proved in [3].

In Eq. (7), the first term represents the score function of unconditional diffusion sampling, which can be readily derived by the pre-trained DDPM. In the next section, we explicate the methodology for obtaining $\nabla_{\boldsymbol{f}_t}\log p_t(\boldsymbol{i},\boldsymbol{v}|\tilde{\boldsymbol{f}}_{0|t})$.

3.2. Likelihood rectification for image fusion

Unlike the traditional image degradation inverse problem $\boldsymbol{y} = \mathcal{A}(\boldsymbol{x}) + \boldsymbol{n}$ where $\boldsymbol{x}$ is the ground truth image, y is measurement and $\mathcal{A}(\cdot)$ is known, we can explicitly obtain its posterior distribution. However, it is not possible to explicitly express $p_t(\boldsymbol{i},\boldsymbol{v}|\boldsymbol{f}_t)$ or $p_t(\boldsymbol{i},\boldsymbol{v}|\tilde{\boldsymbol{f}}_{0|t})$ in image fusion. To address this, we start from the loss function and establish the relationship between the optimization loss function $\ell(\boldsymbol{i},\boldsymbol{v},\tilde{\boldsymbol{f}}_{0|t})$ and the likelihood $p_t(\boldsymbol{i},\boldsymbol{v}|\tilde{\boldsymbol{f}}_{0|t})$ of a probabilistic model. For brevity, $\tilde{\boldsymbol{f}}_{0|t}$ is abbreviated as $\boldsymbol{f}$ in Secs. 3.2.1 and 3.2.2.

3.2.1 Formulation of the likelihood model

We first give a commonly-used loss function [22, 65, 30, 69] for the image fusion task:

$$\min_{\boldsymbol{f}} \|\boldsymbol{f} - \boldsymbol{i}\|_1 + \phi\|\boldsymbol{f} - \boldsymbol{v}\|_1. \quad (8)$$

Then simple variable substitution $\boldsymbol{x} = \boldsymbol{f} - \boldsymbol{v}$ and $\boldsymbol{y} = \boldsymbol{i} - \boldsymbol{v}$ are implemented, and we get

$$\min_{\boldsymbol{x}} \|\boldsymbol{y} - \boldsymbol{x}\|_1 + \phi\|\boldsymbol{x}\|_1. \quad (9)$$

Since $\boldsymbol{y}$ is known and $\boldsymbol{x}$ is unknown, this ℓ_1-norm optimization equation corresponds to the regression model: $\boldsymbol{y} = \boldsymbol{k}\boldsymbol{x} + \boldsymbol{\epsilon}$ with $\boldsymbol{k}$ fixed to $\mathbf{1}$. According to the relationship between regularization term and noise prior distribution, $\boldsymbol{\epsilon}$ should be a Laplacian noise and $\boldsymbol{x}$ is governed by the Laplacian distribution. Thus, in Bayesian fashion, we have:

$$\begin{aligned} p(\boldsymbol{x}) &= \mathcal{LAP}\left(\boldsymbol{x}; 0, \rho\right) = \prod_{i,j}\frac{1}{2\rho}\exp\left(-\frac{|x_{ij}|}{\rho}\right), \\ p(\boldsymbol{y}|\boldsymbol{x}) &= \mathcal{LAP}\left(\boldsymbol{y}; \boldsymbol{x}, \gamma\right) = \prod_{i,j}\frac{1}{2\gamma}\exp\left(-\frac{|y_{ij} - x_{ij}|}{\gamma}\right), \end{aligned} \quad (10)$$

where $\mathcal{LAP}(\cdot)$ is the Laplacian distribution. ρ and γ are scale parameters of $p(\boldsymbol{x})$ and $p(\boldsymbol{y}|\boldsymbol{x})$, respectively.

In order to prevent ℓ_1-norm optimization in Eq. (9) and inspired by [30, 71], we give the Proposition 1:

Proposition 1. *For a random variable (RV) ξ which obeys a Laplace distribution, it can be regarded as the coupling of a normally distributed RV and an exponentially distributed RV, which in formula:*

$$\mathcal{LAP}(\xi; \mu, \sqrt{b/2}) = \int_0^\infty \mathcal{N}(\xi; \mu, a)\mathcal{EXP}(a; b)da. \quad (11)$$

Remark 1. *In Proposition 1, we transform ℓ_1-norm optimization into an ℓ_2-norm optimization with latent variables, avoiding potential non-differentiable points in ℓ_1-norm.*

Therefore, $p(\boldsymbol{x})$ and $p(\boldsymbol{y}|\boldsymbol{x})$ in Eq. (10) can be rewritten as the following hierarchical Bayesian framework:

$$\begin{cases} y_{ij}|x_{ij}, m_{ij} \sim \mathcal{N}\left(y_{ij}; x_{ij}, m_{ij}\right) \\ m_{ij} \sim \mathcal{EXP}\left(m_{ij}; \gamma\right) \\ x_{ij}|n_{ij} \sim \mathcal{N}\left(x_{ij}; 0, n_{ij}\right) \\ n_{ij} \sim \mathcal{EXP}\left(n_{ij}; \rho\right) \end{cases} \quad (12)$$

where $i = 1, \ldots, H$ and $j = 1, \ldots, W$. Through the above probabilistic analysis, the optimization problem in Eq. (9) can be transformed into a maximum likelihood inference problem.

In addition, following [30, 50], the total variation penalty item $r(\boldsymbol{x}) = \|\nabla\boldsymbol{x}\|_2^2$ can be also added to make the fusion image $\boldsymbol{f}$ better preserve the texture information from $\boldsymbol{v}$, where ∇ denotes the gradient operator. Ultimately, the log-likelihood function of the probabilistic inference issue is:

$$\begin{aligned} \ell(\boldsymbol{x}) &= \log p(\boldsymbol{x}, \boldsymbol{y}) - r(\boldsymbol{x}) \\ &= -\sum_{i,j}\left[\frac{\left(x_{ij} - y_{ij}\right)^2}{2m_{ij}} + \frac{x_{ij}^2}{2n_{ij}}\right] - \frac{\psi}{2}\|\nabla\boldsymbol{x}\|_2^2, \end{aligned} \quad (13)$$

and probabilistic graph of this hierarchical Bayesian model is in Fig. 1b. Notably, in this way, we transform the optimization problem Eq. (8) into a maximum likelihood problem of a probability model Eq. (13). Additionally, unlike traditional optimization methods that require manually specified tuning coefficients ϕ in Eq. (8), ϕ in our model can be adaptively updated by inferring the latent variables, enabling the model to better fit different data distributions. The validity of this design has also been verified in ablation experiments in Sec. 4.3. We will then explore how to infer it in the next section.

3.2.2 Inference the likelihood model via EM algorithm

In order to solve the maximum log-likelihood problem in Eq. (13), which can be regarded as an optimization problem with latent variables, we use the *Expectation Maximization* (EM) algorithm to obtain the optimal $\boldsymbol{x}$. In *E-step*, it calculates the expectation of log-likelihood function with respect to $p\left(\boldsymbol{a}, \boldsymbol{b} | \boldsymbol{x}^{(t)}, \boldsymbol{y}\right)$, *i.e.*, the so-called $\mathcal{Q}$-function:

$$\mathcal{Q}\left(\boldsymbol{x} | \boldsymbol{x}^{(t)}\right)=\mathbb{E}_{\boldsymbol{a}, \boldsymbol{b} | \boldsymbol{x}^{(t)}, \boldsymbol{y}}[\ell(\boldsymbol{x})]. \tag{14}$$

Then in *M-step*, the optimal $\boldsymbol{x}$ is obtained by

$$\boldsymbol{x}^{(t+1)}=\arg \max _{\boldsymbol{x}} \mathcal{Q}\left(\boldsymbol{x} | \boldsymbol{x}^{(t)}\right). \tag{15}$$

Next, we will show the implementation detail in each step.

E-step. Proposition 2 gives the calculation results for the conditional expectation of latent variables, and then gets the derivation of $\mathcal{Q}$-function.

Proposition 2. *The conditional expectation of the latent variable $1/m_{ij}$ and $1/n_{ij}$ in Eq. (13) are:*

$$\begin{aligned} \mathbb{E}_{m_{ij} | x_{ij}^{(t)}, y_{ij}}\left[\frac{1}{m_{ij}}\right] &=\sqrt{\frac{2(y_{ij}-x_{ij}^{(t)})^2}{\gamma}}, \\ \mathbb{E}_{n_{ij} | x_{ij}^{(t)}}\left[\frac{1}{n_{ij}}\right] &=\sqrt{\frac{2[x_{ij}^{(t)}]^2}{\rho}}. \end{aligned} \tag{16}$$

Proof. For convenience, we set $\tilde{m}_{ij} \equiv 1/m_{ij}$ and $\tilde{n}_{ij} \equiv 1/n_{ij}$. From Eq. (12) we know that $m_{ij} \sim \mathcal{EXP}\left(m_{ij} ; \gamma\right)=\Gamma(m_{ij} ; 1, \gamma)$. Thus, $\tilde{m}_{ij} \sim \mathcal{IG}\left(1, \gamma\right)$, where $\Gamma(\cdot, \cdot)$ and $\mathcal{IG}(\cdot, \cdot)$ are the gamma distribution and inverse gamma distribution, respectively.

Then we can get the posterior of $\tilde{m}_{ij}$ by Bayes' theorem:

$$\begin{aligned} &\log p\left(\tilde{m}_{ij} | y_{ij}, x_{ij}\right)=\log p\left(y_{ij} | x_{ij}, m_{ij}\right)+\log p\left(\tilde{m}_{ij}\right) \\ =&-\frac{3}{2} \log \tilde{m}_{ij}-\frac{\tilde{m}_{ij}\left(y_{ij}-x_{ij}\right)^2}{2}-\frac{1}{\gamma \tilde{m}_{ij}}+\text{ constant}. \end{aligned} \tag{17}$$

Subsequently, we have

$$p\left(\tilde{m}_{ij} | y_{ij}, x_{ij}\right)=\mathcal{IN}\left(\tilde{m}_{ij} ; \sqrt{2\left(y_{ij}-x_{ij}\right)^2 / \gamma}, 2 / \gamma\right), \tag{18}$$

where $\mathcal{IN}(\cdot, \cdot)$ is the inverse Gaussian distribution. For the posterior of $\tilde{n}_{ij}$, it can be obtain similar to Eq. (17):

$$\begin{aligned} &\log p\left(\tilde{n}_{ij} | x_{ij}\right)=\log p\left(x_{ij} | n_{ij}\right)+\log p\left(\tilde{n}_{ij}\right) \\ =&-\frac{3}{2} \log \tilde{n}_{ij}-\frac{\tilde{n}_{ij} x_{ij}^2}{2}-\frac{1}{\rho \tilde{n}_{ij}}+\text{constant}, \end{aligned} \tag{19}$$

and therefore

$$p\left(\tilde{n}_{ij} | x_{ij}\right)=\mathcal{IN}\left(\tilde{n}_{ij} ; \sqrt{2 x_{ij}^2 / \rho}, 2 / \rho\right). \tag{20}$$

Finally, the conditional expectation of $1/m_{ij}$ and $1/n_{ij}$ are the mean parameters of the corresponding inverse Gaussian distribution Eqs. (18) and (20), respectively. ■

Remark 2. *The conditional expectation computed by Proposition 2 will be used to derive the $\mathcal{Q}$-function below.*

Afterwards, the $\mathcal{Q}$-function Eq. (14) is derived as:

$$\begin{aligned} \mathcal{Q} &=-\sum_{i, j}\left[\frac{\bar{m}_{ij}}{2}\left(x_{ij}-y_{ij}\right)^2+\frac{\bar{n}_{ij}}{2} x_{ij}^2\right]-\frac{\psi}{2}\|\nabla \boldsymbol{x}\|_2^2 \\ &\propto-\|\boldsymbol{m} \odot(\boldsymbol{x}-\boldsymbol{y})\|_2^2-\|\boldsymbol{n} \odot \boldsymbol{x}\|_2^2-\psi\|\nabla \boldsymbol{x}\|_2^2, \end{aligned} \tag{21}$$

where $\bar{m}_{ij}$ and $\bar{n}_{ij}$ represent $\mathbb{E}_{m_{ij} | x_{ij}^{(t)}, y_{ij}}[1/m_{ij}]$ and $\mathbb{E}_{n_{ij} | x_{ij}^{(t)}}[1/n_{ij}]$ in Eq. (16), respectively. $\odot$ is the element-wise multiplication. $\boldsymbol{m}$ and $\boldsymbol{n}$ are matrices with each element being $\sqrt{\bar{m}_{ij}}$ and $\sqrt{\bar{n}_{ij}}$, respectively.

M-step. Here, we need to minimize the negative Q-function with respect to $\boldsymbol{x}$. The half-quadratic splitting algorithm is employed to deal with this problem, i.e.,

$$\begin{aligned} &\min _{\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{k}}||\boldsymbol{m} \odot(\boldsymbol{x}-\boldsymbol{y})||_2^2+||\boldsymbol{n} \odot \boldsymbol{x}||_2^2+\psi||\boldsymbol{u}||_2^2, \\ &\text{s.t.} \quad \boldsymbol{u}=\nabla \boldsymbol{k}, \boldsymbol{k}=\boldsymbol{x}. \end{aligned} \tag{22}$$

It can be further cast into the following unconstraint optimization problem,

$$\begin{aligned} \min _{\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{k}} &||\boldsymbol{m} \odot(\boldsymbol{x}-\boldsymbol{y})||_2^2+||\boldsymbol{n} \odot \boldsymbol{x}||_2^2+\psi||\boldsymbol{u}||_2^2 \\ &+\frac{\eta}{2}\left(||\boldsymbol{u}-\nabla \boldsymbol{k}||_2^2+||\boldsymbol{k}-\boldsymbol{x}||_2^2\right). \end{aligned} \tag{23}$$

The unknown variables $\boldsymbol{k}, \boldsymbol{u}, \boldsymbol{x}$ can be solved iteratively in the coordinate descent fashion.

Update $\boldsymbol{k}$: It is a deconvolution issue,

$$\min _{\boldsymbol{k}} \mathcal{L}_{\boldsymbol{k}}=||\boldsymbol{k}-\boldsymbol{x}||_2^2+||\boldsymbol{u}-\nabla \boldsymbol{k}||_2^2. \tag{24}$$

It can be efficiently solved by the fast Fourier transform (fft) and inverse fft (ifft) operators, and the solution of $\boldsymbol{k}$ is

$$\boldsymbol{k}=\text{ifft}\left\{\frac{\text{fft}(\boldsymbol{x})+\overline{\text{fft}(\nabla)} \odot \text{fft}(\boldsymbol{u})}{1+\overline{\text{fft}(\nabla)} \odot \text{fft}(\nabla)}\right\}, \tag{25}$$

where $\bar{\cdot}$ is the complex conjugation.

Update $\boldsymbol{u}$: It is an ℓ_2-norm penalized regression issue,

$$\min _{\boldsymbol{u}} \mathcal{L}_{\boldsymbol{u}}=\psi||\boldsymbol{u}||_2^2+\frac{\eta}{2}||\boldsymbol{u}-\nabla \boldsymbol{k}||_2^2. \tag{26}$$

The solution of $\boldsymbol{u}$ is

$$\boldsymbol{u}=\frac{\eta}{2 \psi+\eta} \nabla \boldsymbol{k}. \tag{27}$$

Update $\boldsymbol{x}$: It is a least squares issue,

$$\min _{\boldsymbol{x}} \mathcal{L}_{\boldsymbol{x}}=||\boldsymbol{m} \odot(\boldsymbol{x}-\boldsymbol{y})||_2^2+||\boldsymbol{n} \odot \boldsymbol{x}||_2^2+\frac{\eta}{2}||\boldsymbol{k}-\boldsymbol{x}||_2^2. \tag{28}$$

The solution of $\boldsymbol{x}$ is

$$\boldsymbol{x}=(2 \boldsymbol{m}^2 \odot \boldsymbol{y}+\eta \boldsymbol{k}) \oslash(2 \boldsymbol{m}^2+2 \boldsymbol{n}^2+\eta), \tag{29}$$

where $\oslash$ denotes the element-wise division, and final estimation of $\boldsymbol{f}$ is

$$\hat{\boldsymbol{f}} = \boldsymbol{x} + \boldsymbol{v}. \tag{30}$$

Additionally, hyper-parameter γ and ρ in Eq. (10) can be also updated after the sampling from $\boldsymbol{x}$ (Eq. (29)) by

$$\gamma = \frac{1}{hw}\sum_{i,j}\mathbb{E}[m_{ij}],\ \rho = \frac{1}{hw}\sum_{i,j}\mathbb{E}[n_{ij}]. \tag{31}$$

3.3. DDFM

Overview. In Sec. 3.2, we present a methodology for obtaining a hierarchical Bayesian model from existing loss function and perform the model inference via the EM algorithm. In this section, we present our DDFM, where the inference solution and diffusion sampling are integrated within the same iterative framework for generating $\boldsymbol{f}_0$ given $\boldsymbol{i}$ and $\boldsymbol{v}$. The algorithm is illustrated in Algorithm 1 and Fig. 3.

There are two modules in DDFM, the *unconditional diffusion sampling* (UDS) module and the *likelihood rectification*, or say, EM module. The UDS module is utilized to provide natural image priors, which improve the visual plausibility of the fused image. The EM module, on the other hand, is responsible for rectifying the output of UDS module via likelihood to preserve more information from the source images.

Unconditional diffusion sampling module. In Sec. 2.1, we briefly introduce diffusion sampling. In Algorithm 1, UDS (in *grey*) is partitioned into two components, where the first part estimates $\tilde{\boldsymbol{f}}_{0|t}$ using $\boldsymbol{f}_t$, and the second part estimates $\boldsymbol{f}_{t-1}$ using both $\boldsymbol{f}_t$ and $\hat{\boldsymbol{f}}_{0|t}$. From the perspective of score-based DDPM in Eq. (7), a pre-trained DDPM can directly output the current $\nabla_{\boldsymbol{f}_t}\log p_t(\boldsymbol{f}_t)$, while $\nabla_{\boldsymbol{f}_t}\log p_t(\boldsymbol{i},\boldsymbol{v}|\tilde{\boldsymbol{f}}_{0|t})$ can be obtain by the EM module.

EM module. The role of the EM module is to update $\tilde{\boldsymbol{f}}_{0|t} \Rightarrow \hat{\boldsymbol{f}}_{0|t}$. In Algorithm 1 and Fig. 3, the EM algorithm (in *blue* and *yellow*) is inserted in UDS (in *grey*). The preliminary estimate $\tilde{\boldsymbol{f}}_{0|t}$ produced by DDPM sampling (*line 5*) is utilized as the initial input for the EM algorithm to obtain $\hat{\boldsymbol{f}}_{0|t}$ (*line 6-13*), which is an estimation of the fused image subjected to likelihood rectification. In other words, EM module rectify $\tilde{\boldsymbol{f}}_{0|t}$ to $\hat{\boldsymbol{f}}_{0|t}$ to meet the likelihood.

3.4. Why does one-step EM work?

The main difference between our DDFM and conventional EM algorithm lies in that the traditional method requires numerous iterations to obtain the optimal $\boldsymbol{x}$, *i.e.*, the operation from line 6-13 in Algorithm 1 needs to be looped many times. However, our DDFM only requires one step of the EM algorithm iteration, which is embedded into the DDPM framework to accomplish sampling. In the following, we give Proposition 3 to demonstrate its rationality.

Algorithm 1 DDFM

Input:
Infrared image $\boldsymbol{i}$, Visible image $\boldsymbol{v}$, T, $\{\tilde{\sigma}_t\}_{t=1}^{T}$
Output:
Fused image $\boldsymbol{f}_0$.

1: $\boldsymbol{f}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T-1$ **to** 0 **do**
3: *% DDPM Part 1: Obtain* $\tilde{\boldsymbol{f}}_{0|t}$
4: $\hat{\boldsymbol{s}} \leftarrow \boldsymbol{s}_\theta(\boldsymbol{f}_t, t)$
5: $\tilde{\boldsymbol{f}}_{0|t} \leftarrow \frac{1}{\sqrt{\bar{\alpha}_t}}(\boldsymbol{f}_t + (1-\bar{\alpha}_t)\hat{\boldsymbol{s}})$
6: *% E-step: Update latent variables*
7: $\tilde{\boldsymbol{x}}_0 = \tilde{\boldsymbol{f}}_{0|t} - \boldsymbol{v}, \boldsymbol{y} = \boldsymbol{i} - \boldsymbol{v}$
8: Evaluate expectations by Eq. (16).
9: Update hyper-parameters γ, ρ by Eq. (31).
10: *% M-step: Obtain* $\hat{\boldsymbol{f}}_{0|t}$ *via Likelihood Rectification*
11: $\boldsymbol{k} \leftarrow \arg\min_{\boldsymbol{k}} \mathcal{L}_{\boldsymbol{k}}(\tilde{\boldsymbol{x}}_0, \boldsymbol{u})$ (Eq. (25))
12: $\boldsymbol{u} \leftarrow \arg\min_{\boldsymbol{u}} \mathcal{L}_{\boldsymbol{u}}(\nabla \boldsymbol{k})$ (Eq. (27))
13: $\hat{\boldsymbol{f}}_{0|t} \leftarrow \arg\min_{\boldsymbol{x}} \mathcal{L}_{\boldsymbol{x}}(\boldsymbol{k}, \boldsymbol{u}, \tilde{\boldsymbol{x}}_0) + \boldsymbol{v}$ (Eq. (29)&(30))
14: *% DDPM Part 2: Estimate* $\boldsymbol{f}_{t-1}$
15: $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
16: $\boldsymbol{f}_{t-1} \leftarrow \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{f}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\hat{\boldsymbol{f}}_{0|t} + \tilde{\sigma}_t \boldsymbol{z}$
17: **end for**

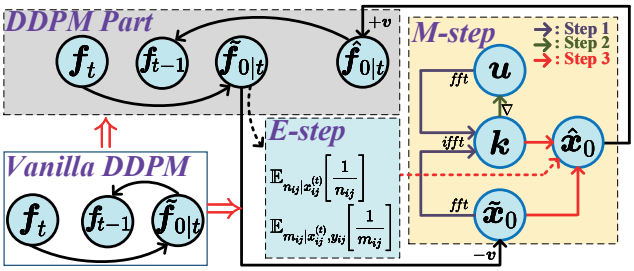

Figure 3: Computational graph of our DDFM in one iteration. Different from the vanilla DDPM, likelihood rectification is completed via the EM algorithm, *i.e.*, the update from $\tilde{\boldsymbol{f}}_{0|t} \Rightarrow \hat{\boldsymbol{f}}_{0|t}$.

Proposition 3. *One-step unconditional diffusion sampling combined with one-step EM iteration is equivalent to one-step conditional diffusion sampling.*

Proof. The estimation of $\hat{\boldsymbol{f}}_{0|t}$ in conditional diffusion sampling, refer to Eq. (4), could be expressed as:

$$\hat{f}_{0|t}(\boldsymbol{f}_t, \boldsymbol{i}, \boldsymbol{v}) = \frac{1}{\sqrt{\bar{\alpha}_t}}[\boldsymbol{f}_t + (1-\bar{\alpha}_t)\boldsymbol{s}_\theta(\boldsymbol{f}_t, \boldsymbol{i}, \boldsymbol{v})] \tag{32a}$$

$$= \frac{1}{\sqrt{\bar{\alpha}_t}}\{\boldsymbol{f}_t + (1-\bar{\alpha}_t)[\boldsymbol{s}_\theta(\boldsymbol{f}_t) + \nabla_{\boldsymbol{f}_t}\log p_t(\boldsymbol{i},\boldsymbol{v}|\boldsymbol{f}_t)]\} \tag{32b}$$

$$\approx \tilde{f}_{0|t}(\boldsymbol{f}_t) + \frac{1-\bar{\alpha}_t}{\sqrt{\bar{\alpha}_t}}\nabla_{\boldsymbol{f}_t}\log p_t\left(\boldsymbol{i}, \boldsymbol{v}|\tilde{\boldsymbol{f}}_{0|t}\right) \tag{32c}$$

$$= \tilde{f}_{0|t}(\boldsymbol{f}_t) - \zeta_t \nabla_{\tilde{\boldsymbol{x}}_0}\mathcal{L}_{\boldsymbol{x}}(\boldsymbol{i}, \boldsymbol{v}, \tilde{\boldsymbol{x}}_0). \tag{32d}$$

Eqs. (32a) to (32c) are respectively based on the definition of Score-based DDPM, Bayes' theorem, and proof in [3]. For

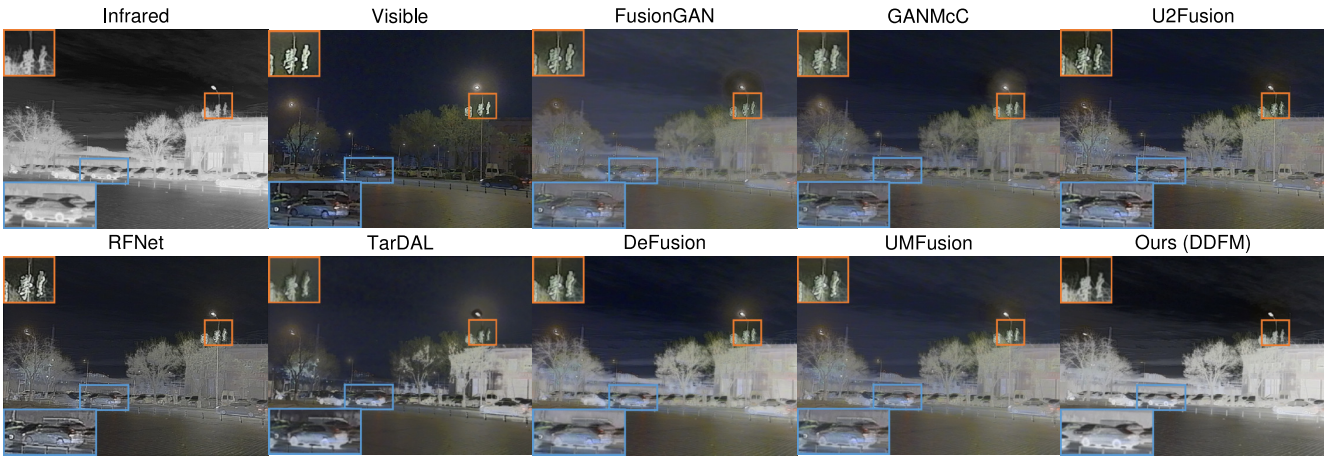

Figure 4: Visual comparison of "01462" from M^{3}FD IVF dataset.

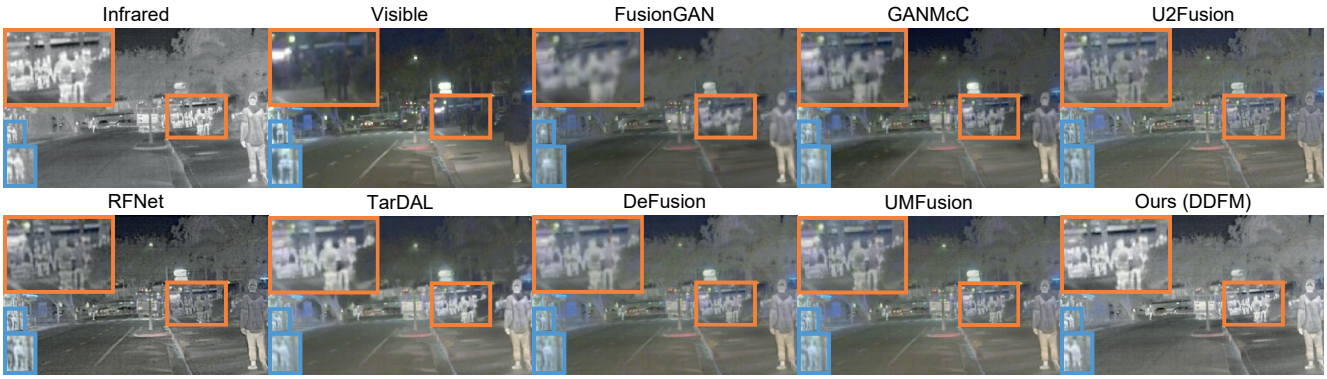

Figure 5: Visual comparison of "FLIR_08248" from RoadScene IVF dataset.

Eq. (32d), although optimization Eq. (28) has a closed-form solution (Eq. (29)), it can also be solved by gradient descent:

$$\hat{\boldsymbol{x}}_0 = \tilde{\boldsymbol{x}}_0 + \nabla_{\tilde{\boldsymbol{x}}_0}\mathcal{L}_{\boldsymbol{x}}\left(\boldsymbol{k}, \boldsymbol{u}, \tilde{\boldsymbol{x}}_0\right) = \tilde{\boldsymbol{x}}_0 + \nabla_{\tilde{\boldsymbol{x}}_0}\mathcal{L}_{\boldsymbol{x}}\left(\boldsymbol{i}, \boldsymbol{v}, \tilde{\boldsymbol{x}}_0\right) \quad (33)$$

where the second equation holds true because as the input for updating $\hat{\boldsymbol{x}}_0$ (Eq. (29)), $\boldsymbol{k}$ and $\boldsymbol{u}$ are functions of $\boldsymbol{i}$ and $\boldsymbol{v}$. ζ_t in Eq. (32d) can be regraded as the update step size.

Hence, conditional sampling $\hat{f}_{0|t}(\boldsymbol{f}_t, \boldsymbol{i}, \boldsymbol{v})$ can be split as an unconditional diffusion sampling $\tilde{f}_{0|t}(\boldsymbol{f}_t)$ and one-step EM iteration $\nabla_{\tilde{\boldsymbol{x}}_0}\mathcal{L}_{\boldsymbol{x}}(\boldsymbol{i}, \boldsymbol{v}, \tilde{\boldsymbol{x}}_0)$, corresponding to UDS module (part 1) and EM module, respectively. ■

Remark 3. *Proposition 3 demonstrates the theoretical explanation for the rationality of inserting the EM module into the UDS module and explains why the EM module only involves one iteration of the Bayesian inference algorithm.*

4. Infrared and visible image fusion

In this section, we elaborate on numerous experiments for IVF task to demonstrate the superiority of our method. More related experiments are placed in *supplementary material*.

4.1. Setup

Datasets and pre-trained model. Following the protocol in [26, 25], IVF experiments are conducted on the four test datasets, *i.e.*, TNO [52], RoadScene [59], MSRS [51], and M^{3}FD [26]. Note that there is no training dataset due to that we do not need any fine-tuning for specific tasks but directly use the pre-trained DDPM model. We choose the pre-trained model proposed by [5], which is trained on ImageNet [45].

Metrics. We employ six metrics including entropy (EN), standard deviation (SD), mutual information (MI), visual information fidelity (VIF), $Q^{AB/F}$, and structural similarity index measure (SSIM) in the quantitative experiments to comprehensively evaluate the fused effect. The detail of metrics is in [32].

Implement details. We use a machine with one NVIDIA GeForce RTX 3090 GPU for fusion image generation. All input images are normalized to $[-1, 1]$. ψ and η in Eq. (23) are set to 0.5 and 0.1, respectively. Please refer to the *supplementary material* for selecting ψ and η via grid search.

Table 1: The quantitative results of IVF task, with the best and second-best values in **boldface** and underline, respectively.

	EN ↑	SD ↑	MI ↑	VIF ↑	Qabf ↑	SSIM ↑
Dataset: MSRS Fusion Dataset [51]						
FGAN [34]	5.60	17.81	1.29	0.40	0.13	0.47
GMcC [35]	6.20	25.95	1.79	0.57	0.28	0.74
U2F [58]	6.06	29.80	1.55	0.59	0.46	0.76
RFN [60]	6.07	26.82	1.36	0.54	0.46	0.81
TarD [26]	5.39	22.74	1.32	0.38	0.16	0.45
DeF [25]	6.85	40.20	2.25	0.74	0.56	0.92
UMF [54]	5.98	23.56	1.38	0.47	0.29	0.58
Ours	**6.88**	**40.75**	**2.35**	**0.81**	**0.58**	**0.94**
Dataset: M^3FD Fusion Dataset [26]						
FGAN [34]	6.51	28.14	2.07	0.44	0.30	0.75
GMcC [35]	6.68	32.23	2.01	0.58	0.36	0.93
U2F [58]	6.84	34.05	1.95	0.73	0.49	**0.98**
RFN [60]	6.67	31.04	1.71	0.67	0.44	0.91
TarD [26]	6.67	38.83	2.38	0.54	0.29	0.87
DeF [25]	6.79	36.39	2.32	0.65	0.44	0.94
UMF [54]	6.73	32.46	2.23	0.66	0.40	0.97
Ours	**6.86**	**38.95**	**2.52**	**0.80**	**0.49**	0.95
Dataset: RoadScene Fusion Dataset [59]						
FGAN [34]	7.12	40.13	1.90	0.36	0.26	0.61
GMcC [35]	7.26	43.44	1.86	0.49	0.34	0.81
U2F [58]	7.16	38.97	1.83	0.54	0.49	0.96
RFN [60]	7.30	43.37	1.64	0.49	0.43	0.88
TarD [26]	7.31	47.24	2.15	0.53	0.41	0.86
DeF [25]	7.31	44.91	2.09	0.55	0.46	0.86
UMF [54]	7.29	42.91	1.96	0.61	0.50	0.98
Ours	**7.41**	**52.61**	**2.35**	**0.75**	**0.65**	**0.98**
Dataset: TNO Fusion Dataset [52]						
FGAN [34]	6.74	34.41	1.78	0.42	0.25	0.66
GMcC [35]	6.86	35.51	1.64	0.53	0.28	0.83
U2F [58]	7.02	38.52	1.41	0.63	0.43	0.93
RFN [60]	6.93	34.95	1.21	0.55	0.37	0.87
TarD [26]	7.02	49.89	1.89	0.54	0.28	0.83
DeF [25]	7.06	40.70	2.04	0.64	0.43	0.92
UMF [54]	6.83	36.56	1.66	0.65	0.42	0.94
Ours	**7.06**	**51.42**	**2.21**	**0.81**	**0.49**	**0.95**

Table 2: Ablation experiment results. **Bold** indicates the best value.

	Configurations			EN	SD	MI	VIF	Qabf	SSIM
	DDPM	$r(\boldsymbol{x})$	ϕ						
I		√	\	7.19	41.82	2.11	0.60	0.42	0.92
II	√		\	7.33	44.12	2.29	0.69	0.52	0.93
III	√	√	0.1	7.25	43.16	2.26	0.66	0.49	0.90
IV	√	√	1	7.26	42.37	2.24	0.66	0.47	0.91
Ours	√	√	\	**7.41**	**52.61**	**2.35**	**0.75**	**0.65**	**0.98**

4.2. Comparison with SOTA methods

In this section, we compare our DDFM with the state-of-the-art methods, including the *GAN-based generative methods group*: FusionGAN [34], GANMcC [35], TarDAL [26], and UMFusion [54]; and the *discriminative methods group*: U2Fusion [58], RFNet [60], and DeFusion [25].

Qualitative comparison. We show the comparison of fusion results in Figs. 4 and 5. Our approach effectively combines thermal radiation information from infrared images with detailed texture information from visible images. As a result, objects located in dimly-lit environments are conspicuously accentuated, enabling easy distinguishing of foreground objects from the background. Moreover, previously indistinct background features due to low illumination now possess clearly defined edges and abundant contour information, enhancing our ability to comprehend the scene.

Quantitative comparison. Subsequently, six metrics previously mentioned are utilized to quantitatively compare the fusion outcomes, as presented in Tab. 1. Our method demonstrates remarkable performance across almost all metrics, affirming its suitability for different lighting and object categories. Notably, the outstanding values for MI, VIF and Qabf across all datasets signify its ability to generate images that adhere to human visual perception while preserving the integrity of the source image information.

4.3. Ablation studies

Numerous ablation experiments are conducted to confirm the soundness of our various modules. The above six metrics are utilized to assess the fusion performance for the experimental groups, and results on the Roadscene testset are displayed in Tab. 2.

Unconditional diffusion sampling module. We first verify the effectiveness of DDPM. In Exp. I, we eliminate the denoising diffusion generative framework, thus only the EM algorithm is employed to solve the optimization Eq. (8) and obtain the fusion image. In fairness, we keep the total iteration number consistent with DDFM.

EM module. Next, we verify the components in the EM module. In Exp. II, we removed the total variation penalty item $r(\boldsymbol{x})$ in Eq. (13). Then, we remove the Bayesian inference model. As mentioned earlier, ϕ in Eq. (8) can be automatically inferred in the hierarchical Bayesian model. Therefore, we manually set ϕ to 0.1 (Exp. III) and 1 (Exp. IV), and used the ADMM algorithm to infer the model.

In conclusion, the results presented in Tab. 2 demonstrate that none of the experimental groups is able to achieve fusion results comparable to our DDFM, further emphasizing the effectiveness and rationality of our approach.

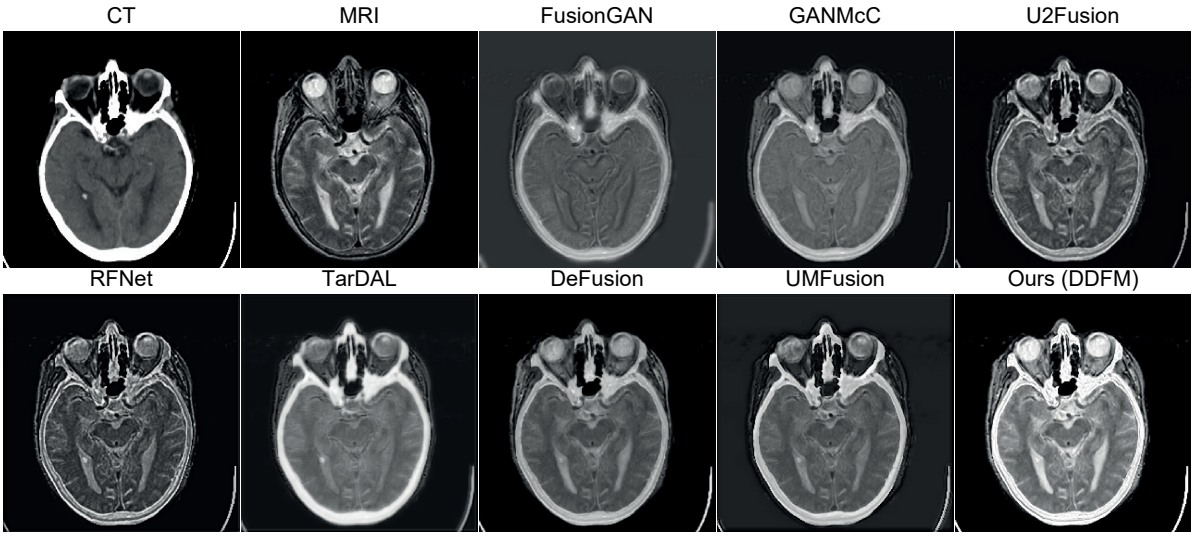

Figure 6: Visual comparison for MIF task.

Table 3: The quantitative results of the MIF task, with the best and second-best values in **boldface** and underline, respectively.

Dataset: Harvard Medical Fusion Dataset [8]						
	EN ↑	SD ↑	MI ↑	VIF ↑	Qabf ↑	SSIM ↑
FGAN [34]	4.05	29.20	1.53	0.39	0.18	0.23
GMcC [35]	4.18	42.49	1.74	0.50	0.42	0.35
U2F [58]	4.14	48.89	1.80	0.50	0.55	1.14
RFN [60]	**4.75**	40.81	1.62	0.43	0.56	0.40
TarD [26]	4.61	60.64	1.44	0.33	0.21	0.25
DeF [25]	4.21	61.65	1.85	0.62	0.59	1.40
UMF [54]	4.61	27.28	1.62	0.40	0.27	0.30
Ours	4.64	**63.11**	**1.99**	**0.76**	**0.60**	**1.41**

5. Medical image fusion

In this section, MIF experiments are carried out to verify the effectiveness of our method.

Setup. We choose 50 pairs of medical images from the Harvard Medical Image Dataset [8] for the MIF experiments, including image pairs of MRI-CT, MRI-PET and MRI-SPECT. The generation strategy and evaluation metrics for the MIF task are identical to those used for IVF.

Comparison with SOTA methods. Qualitative and quantitative results are shown in Fig. 6 and Tab. 3. It is evident that DDFM retains intricate textures while emphasizing structural information, leading to remarkable performance across both visual and almost all numerical metrics.

6. Conclusion

We propose DDFM, a novel generative image fusion algorithm based on the denoising diffusion probabilistic model (DDPM). The generation problem is split into an unconditional DDPM to leverage image generative priors and a maximum likelihood sub-problem to preserve cross-modality information from source images. We model the latter using a hierarchical Bayesian approach and its solution based on EM algorithm can be integrated into unconditional DDPM to accomplish conditional image fusion. Experiments on infrared-visible and medical image fusion demonstrate that DDFM achieves promising fusion results.

Acknowledgement

This work has been supported by the National Key Research and Development Program of China under grant 2018AAA0102201, the National Natural Science Foundation of China under Grant 61976174 and 12201497, the Macao Science and Technology Development Fund under Grant 061/2020/A2, Shaanxi Fundamental Science Research Project for Mathematics and Physics under Grant 22JSQ033, the Fundamental Research Funds for the Central Universities under Grant D5000220060, and partly supported by the Alexander von Humboldt Foundation.

References

[1] Brian DO Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982. 2

[2] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *CoRR*, abs/2004.10934, 2020. 1

[3] Hyungjin Chung, Jeongsol Kim, Michael T. McCann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *ICLR*, 2023. 3, 4, 6

[4] Xin Deng and Pier Luigi Dragotti. Deep convolutional neural network for multi-modal image restoration and fusion. *IEEE*

Trans. Pattern Anal. Mach. Intell., 43(10):3333–3348, 2021. 3

[5] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 3, 7

[6] Fangyuan Gao, Xin Deng, Mai Xu, Jingyi Xu, and Pier Luigi Dragotti. Multi-modal convolutional dictionary learning. *IEEE Trans. Image Process.*, 31:1325–1339, 2022. 3

[7] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, pages 2672–2680, 2014. 2, 3

[8] Harvard Medical website. http://www.med.harvard.edu/AANLIB/home.html. 9

[9] Chunming He, Kai Li, Guoxia Xu, Jiangpeng Yan, Longxiang Tang, Yulun Zhang, Xiu Li, and Yaowei Wang. Hqg-net: Unpaired medical image enhancement with high-quality guidance. *arXiv preprint arXiv:2307.07829*, 2023. 2

[10] Chunming He, Kai Li, Yachao Zhang, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. Camouflaged object detection with feature decomposition and edge reconstruction. In *CVPR*, 2023. 1

[11] Chunming He, Kai Li, Yachao Zhang, Guoxia Xu, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. Weakly-supervised concealed object segmentation with sam-based pseudo labeling and multi-scale feature grouping. *arXiv preprint arXiv:2305.11003*, 2023. 1

[12] Chunming He, Kai Li, Yachao Zhang, Yulun Zhang, Zhenhua Guo, Xiu Li, Martin Danelljan, and Fisher Yu. Strategic preys make acute predators: Enhancing camouflaged object detectors by generating camouflaged objects. *arXiv preprint arXiv:2308.03166*, 2023. 1

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2, 3

[14] Zhanbo Huang, Jinyuan Liu, Xin Fan, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Reconet: Recurrent correction network for fast and efficient multi-modality image fusion. In *ECCV*, 2022. 3

[15] Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005. 2

[16] Alex Pappachen James and Belur V. Dasarathy. Medical image fusion: A survey of the state of the art. *Inf. Fusion*, 19:4–19, 2014. 2

[17] Zhiying Jiang, Zengxi Zhang, Xin Fan, and Risheng Liu. Towards all weather and unobstructed multi-spectral image stitching: Algorithm and benchmark. In *ACM MM*, pages 3783–3791, 2022. 3

[18] Hyungjoo Jung, Youngjung Kim, Hyunsung Jang, Namkoo Ha, and Kwanghoon Sohn. Unsupervised deep image fusion with structure tensor representations. *IEEE Trans. Image Process.*, 29:3845–3858, 2020. 3

[19] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *arXiv preprint arXiv:2201.11793*, 2022. 3

[20] Hui Li, Kede Ma, Hongwei Yong, and Lei Zhang. Fast multi-scale structural patch decomposition for multi-exposure image fusion. *IEEE Trans. Image Process.*, 29:5805–5816, 2020. 1

[21] Hui Li, Xiao-Jun Wu, and Tariq S. Durrani. Nestfuse: An infrared and visible image fusion architecture based on nest connection and spatial/channel attention models. *IEEE Trans. Instrum. Meas.*, 69(12):9645–9656, 2020. 3

[22] Hui Li, Xiao-Jun Wu, and Josef Kittler. Rfn-nest: An end-to-end residual fusion network for infrared and visible images. *Inf. Fusion*, 73:72–86, 2021. 3, 4

[23] Hui Li and Xiao-Jun Wu. Densefuse: A fusion approach to infrared and visible images. *IEEE Trans. Image Process.*, 28(5):2614–2623, 2018. 3

[24] Hui Li, Tianyang Xu, Xiao-Jun Wu, Jiwen Lu, and Josef Kittler. Lrrnet: A novel representation learning guided fusion network for infrared and visible images. *IEEE transactions on pattern analysis and machine intelligence*, 2023. 3

[25] Pengwei Liang, Junjun Jiang, Xianming Liu, and Jiayi Ma. Fusion from decomposition: A self-supervised decomposition approach for image fusion. In *ECCV*, 2022. 3, 7, 8, 9

[26] 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 *CVPR*, pages 5792–5801. IEEE, 2022. 2, 3, 7, 8, 9

[27] Risheng Liu, Jinyuan Liu, Zhiying Jiang, Xin Fan, and Zhongxuan Luo. A bilevel integrated model with data-driven layer ensemble for multi-modality image fusion. *IEEE Trans. Image Process.*, 30:1261–1274, 2020. 1

[28] Risheng Liu, Zhu Liu, Jinyuan Liu, and Xin Fan. Searching a hierarchically aggregated fusion architecture for fast multi-modality image fusion. In *ACM MM*, pages 1600–1608. ACM, 2021. 1, 3

[29] Zhu Liu, Jinyuan Liu, Guanyao Wu, Long Ma, Xin Fan, and Risheng Liu. Bi-level dynamic learning for jointly multi-modality image fusion and beyond. *arXiv preprint arXiv:2305.06720*, 2023. 2

[30] Jiayi Ma, Chen Chen, Chang Li, and Jun Huang. Infrared and visible image fusion via gradient transfer and total variation minimization. *Inf. Fusion*, 31:100–109, 2016. 4

[31] Jiayi Ma, Pengwei Liang, Wei Yu, Chen Chen, Xiaojie Guo, Jia Wu, and Junjun Jiang. Infrared and visible image fusion via detail preserving adversarial learning. *Inf. Fusion*, 54:85–98, 2020. 3

[32] Jiayi Ma, Yong Ma, and Chang Li. Infrared and visible image fusion methods and applications: A survey. *Inf. Fusion*, 45:153–178, 2019. 7

[33] Jiayi Ma, Han Xu, Junjun Jiang, Xiaoguang Mei, and Xiao-Ping (Steven) Zhang. Ddcgan: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Trans. Image Process.*, 29:4980–4995, 2020. 2

[34] Jiayi Ma, Wei Yu, Pengwei Liang, Chang Li, and Junjun Jiang. Fusiongan: A generative adversarial network for infrared and visible image fusion. *Inf. Fusion*, 48:11–26, 2019. 2, 3, 8, 9

[35] Jiayi Ma, Hao Zhang, Zhenfeng Shao, Pengwei Liang, and Han Xu. Ganmcc: A generative adversarial network with multiclassification constraints for infrared and visible image fusion. *IEEE Trans. Instrum. Meas.*, 70:1–14, 2021. 2, 3, 8, 9

[36] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *ICCV*, pages 2794–2802, 2017. 3

[37] Bikash Meher, Sanjay Agrawal, Rutuparna Panda, and Ajith Abraham. A survey on region based image fusion methods. *Inf. Fusion*, 48:119–132, 2019. 1

[38] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2, 3

[39] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, pages 8162–8171, 2021. 3

[40] Haotong Qin, Yifu Ding, Xiangguo Zhang, Jiakai Wang, Xianglong Liu, and Jiwen Lu. Diverse sample generation: Pushing the limit of generative data-free quantization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1

[41] Haotong Qin, Mingyuan Zhang, Yifu Ding, Aoyu Li, Ziwei Liu, Fisher Yu, and Xianglong Liu. Bibench: Benchmarking and analyzing network binarization. In *ICML*, 2023. 1

[42] Haotong Qin, Xiangguo Zhang, Ruihao Gong, Yifu Ding, Yi Xu, and Xianglong Liu. Distribution-sensitive information retention for accurate binary neural network. *International Journal of Computer Vision*, 2022. 1

[43] Xuebin Qin, Zichen Vincent Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jägersand. Basnet: Boundary-aware salient object detection. In *CVPR*, pages 7479–7489. CVF / IEEE, 2019. 1

[44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 3

[45] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252, 2015. 7

[46] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 2, 3

[47] Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In *ICLR*, 2023. 3

[48] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[49] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021. 2

[50] 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. *Inf. Fusion*, 82:28–42, 2022. 3, 4

[51] Linfeng Tang, Jiteng Yuan, Hao Zhang, Xingyu Jiang, and Jiayi Ma. Piafusion: A progressive infrared and visible image fusion network based on illumination aware. *Inf. Fusion*, 83-84:79–92, 2022. 1, 7, 8

[52] Alexander Toet and Maarten A. Hogervorst. Progress in color night vision. *Optical Engineering*, 51(1):1 – 20, 2012. 7, 8

[53] Vibashan VS, Jeya Maria Jose Valanarasu, Poojan Oza, and Vishal M. Patel. Image fusion transformer. *CoRR*, abs/2107.09011, 2021. 3

[54] Di Wang, Jinyuan Liu, Xin Fan, and Risheng Liu. Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration. In *IJCAI*, pages 3508–3515. ijcai.org, 2022. 3, 8, 9

[55] Jiakai Wang, Aishan Liu, Zixin Yin, Shunchang Liu, Shiyu Tang, and Xianglong Liu. Dual attention suppression attack: Generate adversarial camouflage in physical world. In *CVPR*, pages 8565–8574, 2021. 1

[56] Jiakai Wang, Zixin Yin, Pengfei Hu, Aishan Liu, Renshuai Tao, Haotong Qin, Xianglong Liu, and Dacheng Tao. Defensive patches for robust recognition in the physical world. In *CVPR*, pages 2456–2465, 2022. 1

[57] Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion gans. In *ICLR*, 2022. 2

[58] Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2fusion: A unified unsupervised image fusion network. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(1):502–518, 2022. 1, 3, 8, 9

[59] Han Xu, Jiayi Ma, Zhuliang Le, Junjun Jiang, and Xiaojie Guo. Fusiondn: A unified densely connected network for image fusion. In *AAAI Conference on Artificial Intelligence, AAAI*, pages 12484–12491, 2020. 1, 3, 7, 8

[60] Han Xu, Jiayi Ma, Jiteng Yuan, Zhuliang Le, and Wei Liu. Rfnet: Unsupervised network for mutually reinforcing multi-modal image registration and fusion. In *CVPR*, pages 19647–19656. IEEE, 2022. 3, 8, 9

[61] Han Xu, Jiteng Yuan, and Jiayi Ma. Murf: Mutually reinforcing multi-modal image registration and fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 3

[62] Shuang Xu, Jiangshe Zhang, Zixiang Zhao, Kai Sun, Junmin Liu, and Chunxia Zhang. Deep gradient projection networks for pan-sharpening. In *CVPR*, pages 1366–1375. CVF / IEEE, 2021. 1

[63] Shuang Xu, Zixiang Zhao, Yicheng Wang, Chunxia Zhang, Junmin Liu, and Jiangshe Zhang. Deep convolutional sparse coding networks for image fusion. *CoRR*, abs/2005.08448, 2020. 3

[64] Wang Yinhuai, Yu Jiwen, and Zhang Jian. Zero shot image restoration using denoising diffusion null-space model. *arXiv:2212.00490*, 2022. 3

[65] Hao Zhang and Jiayi Ma. Sdnet: A versatile squeeze-and-decomposition network for real-time image fusion. *Int. J. Comput. Vis.*, 129(10):2761–2785, 2021. 2, 3, 4

[66] Hao Zhang, Han Xu, Yang Xiao, Xiaojie Guo, and Jiayi Ma. Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity. In *AAAI*, pages 12797–12804. AAAI Press, 2020. 3

[67] Xingchen Zhang. Deep learning-based multi-focus image fusion: A survey and a comparative study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1

[68] Yu Zhang, Yu Liu, Peng Sun, Han Yan, Xiaolin Zhao, and Li Zhang. IFCNN: A general image fusion framework based on

convolutional neural network. *Inf. Fusion*, 54:99–118, 2020. 3

[69] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Shuang Xu, Zudi Lin, Radu Timofte, and Luc Van Gool. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In *CVPR*, pages 5906–5916, June 2023. 3, 4

[70] 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. *arXiv preprint arXiv:2305.11443*, 2023. 1

[71] Zixiang Zhao, Shuang Xu, Chunxia Zhang, Junmin Liu, and Jiangshe Zhang. Bayesian fusion for infrared and visible images. *Signal Processing*, 177, 2020. 4

[72] Zixiang Zhao, Shuang Xu, Chunxia Zhang, Junmin Liu, Jiangshe Zhang, and Pengfei Li. DIDFuse: Deep image decomposition for infrared and visible image fusion. In *IJCAI*, pages 970–976, 2020. 1, 3

[73] Zixiang Zhao, Shuang Xu, Jiangshe Zhang, Chengyang Liang, Chunxia Zhang, and Junmin Liu. Efficient and model-based infrared and visible image fusion via algorithm unrolling. *IEEE Trans. Circuits Syst. Video Technol.*, 32(3):1186–1196, 2022. 3

[74] Zixiang Zhao, Jiangshe Zhang, Xiang Gu, Chengli Tan, Shuang Xu, Yulun Zhang, Radu Timofte, and Luc Van Gool. Spherical space feature decomposition for guided depth map super-resolution. In *ICCV*, 2023. 1

[75] Zixiang Zhao, Jiangshe Zhang, Shuang Xu, Zudi Lin, and Hanspeter Pfister. Discrete cosine transform network for guided depth map super-resolution. In *CVPR*, pages 5697–5707, June 2022. 3

[76] Zixiang Zhao, Jiangshe Zhang, Shuang Xu, Kai Sun, Lu Huang, Junmin Liu, and Chunxia Zhang. FGF-GAN: A lightweight generative adversarial network for pansharpening via fast guided filter. In *ICME*, pages 1–6. IEEE, 2021. 1