Dif-Fusion: Toward High Color Fidelity in Infrared and Visible Image Fusion With Diffusion Models

Jun Yue, Leyuan Fang, *Senior Member, IEEE*, Shaobo Xia, Yue Deng, *Senior Member, IEEE*, and Jiayi Ma, *Senior Member, IEEE*

***Abstract*— Color plays an important role in human visual perception, reflecting the spectrum of objects. However, the existing infrared and visible image fusion methods rarely explore how to handle multi-spectral/channel data directly and achieve high color fidelity. This paper addresses the above issue by proposing a novel method with diffusion models, termed as Dif-Fusion, to generate the distribution of the multi-channel input data, which increases the ability of multi-source information aggregation and the fidelity of colors. In specific, instead of converting multi-channel images into single-channel data in existing fusion methods, we create the multi-channel data distribution with a denoising network in a latent space with forward and reverse diffusion process. Then, we use the the denoising network to extract the multi-channel diffusion features with both visible and infrared information. Finally, we feed the multi-channel diffusion features to the multi-channel fusion module to directly generate the three-channel fused image. To retain the texture and intensity information, we propose multi-channel gradient loss and intensity loss. Along with the current evaluation metrics for measuring texture and intensity fidelity, we introduce Delta E as a new evaluation metric to quantify color fidelity. Extensive experiments indicate that our method is more effective than other state-of-the-art image fusion methods, especially in color fidelity. The source code is available at https://github.com/GeoVectorMatrix/Dif-Fusion.**

***Index Terms*— Image fusion, color fidelity, multimodal information, diffusion models, latent representation, deep generative model.**

I. Introduction

DUE to the theoretical and technical limitations of the optical imaging hardware equipment, the image acquired by a single sensor or a single shooting setting can only obtain part of the image information [1], [2]. Therefore, the fusion of images from different sensors or different shooting settings helps to enrich the image information. Among various image fusion tasks, infrared and visible image fusion is one of the most widely used [3], [4]. The infrared sensor can capture the thermal radiation from the object, but it is vulnerable to noise and difficult to capture the texture information. On the contrary, visible images usually contain rich structure and texture information, but are vulnerable to illumination and occlusion. The complementary between them makes it possible to generate fusion images containing both thermal objects and texture details. In this context, infrared and visible image fusion servers as a technique that creates a fused image by combining information from infrared and visible image pairs. This method makes use of both spectra's advantages to improve the final image's overall quality and interpretability. A common and widely usage of fused images is to make visual interpretation faster and more accurate, such as the analysis of multi-modal remote sensing images. The two-modal images are often captured simultaneously in surveillance applications, thus fused images that provide essential complementary information are also important. In recent years, the image fusion has also attracted attention in other fields, such as semantic segmentation [5], human re-identification [6], object detection and tracking [7], and many others.

In order to achieve effective fusion of infrared and visible images, many image fusion technologies have been proposed in the past decades [3], including traditional methods [8] and deep learning-based methods [9]. Traditional infrared and visible image fusion algorithms can be generally divided into the following categories, including sparse representation-based methods [10], [11], multi-scale transformation-based methods [12], [13], subspace-based methods [14], saliency detection-based methods [15], and hybrid methods [16]. Although the above algorithms can meet the needs of specific scenes in most cases, there are still some problems: 1) The existing traditional methods usually use the same method to express the image features, and rarely consider the distinctive characteristics of infrared and visible images; 2) The activity level measurement and fusion rules need to be set manually, which cannot meet the needs of complex scenarios [17].

In recent years, with the rapid development of deep learning technology [18], [19], researchers have explored fusion algorithms based on deep neural networks. Generally, the current mainstream deep fusion methods can be divided into three categories: methods based on autoencoder (AE) [20], methods

Manuscript received 20 January 2023; revised 18 August 2023; accepted 28 September 2023. Date of publication 16 October 2023; date of current version 24 October 2023. This work was supported in part by the National Natural Science Foundation of China under Grant U22B2014, Grant 62101072, and Grant 42201481; in part by the Science and Technology Plan Project Fund of Hunan Province under Grant 2022RSC3064; and in part by the Hunan Provincial Natural Science Foundation of China under Grant 2021JJ40570 and Grant 2023JJ40024. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Vittoria Bruni. *(Jun Yue and Leyuan Fang contributed equally to this work.) (Corresponding author: Shaobo Xia.)*

Jun Yue is with the School of Automation, Central South University, Changsha 410083, China (e-mail: jyue@pku.edu.cn).

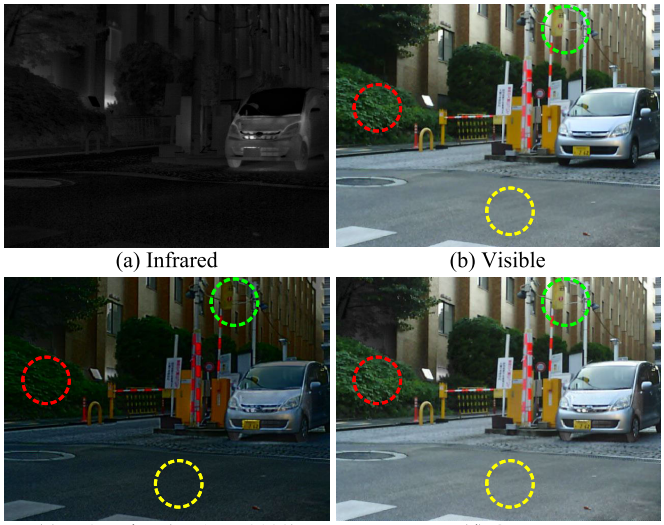
Leyuan Fang is with the College of Electrical and Information Engineering, Hunan University, Changsha 410082, China, and also with the Peng Cheng Laboratory, Shenzhen 518000, China (e-mail: fangleyuan@gmail.com).

Shaobo Xia is with the Department of Geomatics Engineering, Changsha University of Science and Technology, Changsha 410114, China (e-mail: shaoboxia2020@gmail.com).

Yue Deng is with the School of Astronautics, Beihang University, Beijing 100083, China (e-mail: yuedeng.thu@gmail.com).

Jiayi Ma is with the School of Electronic Information, Wuhan University, Wuhan 430072, China (e-mail: jyma2010@gmail.com).

Digital Object Identifier 10.1109/TIP.2023.3322046



(a) Infrared (b) Visible

(c) U2Fusion (TPAMI 2022) (d) Ours

Fig. 1. Illustration of color fidelity. From (a) to (d): infrared image, visible image, fused images of U2Fusion [1] and our proposed Dif-Fusion. The dotted circles in green, yellow, and red show the color differences between the visible and the fused images of the wall, the pavement, and the vegetation, respectively. Compared with existing method, Dif-Fusion achieves higher color fidelity.

based on convolutional neural network (CNN) [4], [21] and methods based on generative adversarial network (GAN) [9], [22]. As an image generation task, the existing infrared and visible image fusion methods lack in-depth exploration of the generation model. The existing methods based on generation model are mainly based on GAN, including FusionGAN [9] and GANMcC [22]. However, the distribution of infrared and visible images cannot be built due to the additional constraints that these methods impose on the generator.

Although the existing fusion methods based on deep learning can achieve relatively satisfactory performance, there are still some issues that need to be taken into consideration. First, the existing methods mainly focus on preserving the thermal targets in the infrared image and the background texture structure in the visible image, and less on how to preserve the color information in the visible image [4]. However, color reflects the spectrum of objects which is of vital importance in digital images. The human visual system is highly sensitive to color (spectrum) and there are many studies on the significance of color in understanding visual scenes [23], [24]. Extensive theoretical and empirical studies on color clearly indicates that color has a significant influence on people's cognition, affect, and behavior [25], [26], [27]. As seen in Fig. 1 (c), the current method (U2Fusion [1]) does not effectively utilize multi-spectral information and performs poorly in maintaining the color information in visible images, which will affect human perception negatively. In addition to gradient fidelity and intensity fidelity, we believe that it is also necessary to maintain the color of visible image in the fusion task, which helps to retain information that is critical to human perception [28].

On the other hand, how to extract multi-channel complementary information within the input data is not well studied. The existing methods usually convert the visible images stored in three channels (i.e., RGB channels) from RGB space to YCbCr space, and use the Y channel for fusion [1], [29]. After the single-channel fused image is generated, it needs to be converted to a three-channel image through post-processing [30], [31]. Since not all channels are presented in the input data, it is hard to construct the multi-channel distribution and extract multi-channel complementary information, resulting in color distortion.

To address the above challenges, a novel infrared and visible image fusion method based on diffusion models, namely, Dif-Fusion, is proposed. First, we directly feed multi-channel data composed of three-channel visible image and one-channel infrared image, and construct multi-channel distribution in the latent space through diffusion process. The diffusion process is a Markov process, which is divided into a forward process and a reverse process [32]. In the forward process, Gaussian noise is incrementally added to the multi-channel input data, and in the reverse process, the noise added in the forward process is eliminated with multiple timesteps. The multi-channel distribution is constructed by training the denoising network in the reverse process to estimate the noise added in the forward process. Second, we extract the multi-channel diffusion features from the denoising network, which includes both infrared and visible features. Third, the multi-channel diffusion features are fed into the multi-channel fusion module to directly generate three-channel fused images. Furthermore, we propose multi-channel gradient loss $\mathcal{L}_{MCG}$ and multi-channel intensity loss $\mathcal{L}_{MCI}$ to preserve the texture and gradient information of three-channel fused images.

The existing methods mostly concentrate on fusing texture/gradient in the visible images and intensity in the infrared images, without paying attention to the preservation of color information and the extraction of multi-channel complementary information. The proposed method establishes the distribution of multi-channel input data based on diffusion models and extracts multi-channel complementary information to achieve high color fidelity. As shown in Fig. 1 (d), our fused image has high color fidelity and is more suitable for human visual perception. In terms of fusion result evaluation, in addition to existing indicators used to quantify intensity and gradient fidelity, we introduce an indicator to quantify color fidelity. With the proposed Dif-Fusion, the infrared and visible images can be simply fed into the model without color space transformation. To sum up, the main contributions of this work are threefold.

- We propose an infrared and visible image fusion framework based on diffusion models that can generate chromatic fused image directly and achieve color, gradient and intensity fidelity simultaneously.
- We formulate the construction of multi-channel distribution as a diffusion process, which is the first study to apply the diffusion models to infrared and visible image fusion to the best of our knowledge.
- In order to measure the color fidelity of fused images, a new evaluation metric is introduced to quantify the color fidelity. Extensive experiments show the proposed method outperforms the existing state-of-the-art methods.

The rest of this paper is organized as follows. In Section II, we briefly introduce the related work of image fusion and diffusion models. In Section III, the proposed method is described in detail. In Section IV, the experimental settings and results are shown and discussed. In Section V, the conclusions of this article are summarized.

II. Related Work

In this section, we introduce background materials and related work that are highly relevant to the method proposed in this paper, including traditional infrared and visible image fusion methods, deep learning based fusion methods, and diffusion models.

A. Infrared and Visible Image Fusion

In the past decades, researchers have proposed many infrared and visible image fusion techniques, including traditional methods and deep learning-based methods [2], [17]. Traditional infrared and visible image fusion algorithms can be generally divided into five categories, i.e., sparse representation, multi-scale transformation, subspace representation, saliency detection, and hybrid methods [3].

The main idea of sparse representation theory is that an image signal can be represented as a linear combination of the least possible atoms or transformation primitives in an overly complete dictionary [18]. Over-completeness indicates that the number of atoms in the dictionary is greater than the dimension of the signal [10], [11]. In image fusion, sparse representation usually learns a complete dictionary from a group of training images, which captures the inherent data-driven image representation. The over-complete dictionary contains abundant base atoms, allowing for more meaningful and stable source image representation [12]. Multi-scale transformation can decompose the original image into subimages of different scales [3]. The multi-scale transformation is similar to the human visual process, which can make the fused image have a good visual effect [12], [33], [34], [35].

The method based on subspace representation aims to project high dimensional features into low dimensional subspace [3]. Projection into low dimensional subspace can help capture the inherent structure of the original input image [36]. In addition, data processing in a low dimensional subspace can save time and memory compared to that in a high-dimensional space. Common subspace representation-based methods include principal component analysis (PCA) [37], [38], independent component analysis (ICA) [39], [40] and non-negative matrix factorization (NMF) [14], [41].

The saliency detection model simulates human behavior and captures the most prominent regions/objects from images or scenes [15]. It has many important applications in computer vision and pattern recognition tasks [42]. In recent years, infrared and visible image fusion methods based on saliency detection can be mainly divided into two categories, namely, weight calculation [15], [16], [43] and salient target extraction [44], [45], [46]. The above fusion methods all have advantages and disadvantages. Researchers explore hybrid methods to make full use of the advantages of various methods and improve image fusion performance. Common hybrid methods include hybrid multi-scale transformation and sparse representation [12], [47], hybrid multi-scale transformation and saliency detection [15], [48], etc.

Due to the excellent feature learning ability and nonlinear fitting ability of neural networks, researchers have explored data-driven infrared and visible image fusion methods based on deep learning [2], [49], [50]. These methods mainly include AE-based methods [20], [51], [52], CNN-based methods [4], [29], [53], and GAN-based methods [9], [22], [54].

Researchers have proposed many fusion methods based on AE [3]. Most of them use the encoder structure to extract features from the source image, and use decoder structure to complete image reconstruction. DenseFuse [20] is a typical AE-based method. It integrates convolutional layers, fusion layers, and dense blocks within its encoding network, establishing interconnections between the output of each layer. Subsequently, a decoder is employed to reconstruct the fused image. In order to improve the feature extraction ability of the encoder, researchers proposed a method named NestFuse [51]. From a multi-scale perspective, this method can preserve a large amount of information from the input data based on nest connections. SEDRFuse [52] is a symmetric encoder-decoder with residual networks. In the fusion phase, the trained extractor is used to extract intermediate and compensated features, and then two attention maps derived from the intermediate features are multiplied by the intermediate features for fusion [52].

For CNN-based fusion methods, a typical method is PMGI [53]. This approach represents a fast image fusion network centered on the preservation of gradient and intensity. Additionally, it incorporates a pathwise transfer block to facilitate information exchange across various pathways, enabling the pre-fusion of gradient and intensity data to augment the fusion process. In order to adaptively determine the proportion of gradient information preserved and retain more complete texture structure, SDNet is proposed [29]. For gradient fidelity, this method determines the optimization objective of gradient distribution according to the texture richness, and guides the fusion image to contain more texture details through an adaptive decision block. To provide spatial guidance for the integration of multi-source information, STDFusionNet employs a salient target mask to help with the fusion task [4]. To combine fusion tasks with a high-level visual task, a fusion method assisted by a high-level semantic task is proposed [55]. In addition, there are a few methods to investigate how lighting condition affects image fusion [21], [56]. In recent years, the multimodal image fusion framework combining CNN and Transformer has also attracted considerable attention [50], [57]. Their core idea is to leverage Transformer's self-attention mechanism and the ability in long-range relationship learning to explore the complementary of features between different modalities, demonstrating significant potential. However, it should be pointed out that current research in this area requires feature extraction based on CNN and still focuses on single-channel image fusion frameworks, without directly modeling the correlations between input channels.

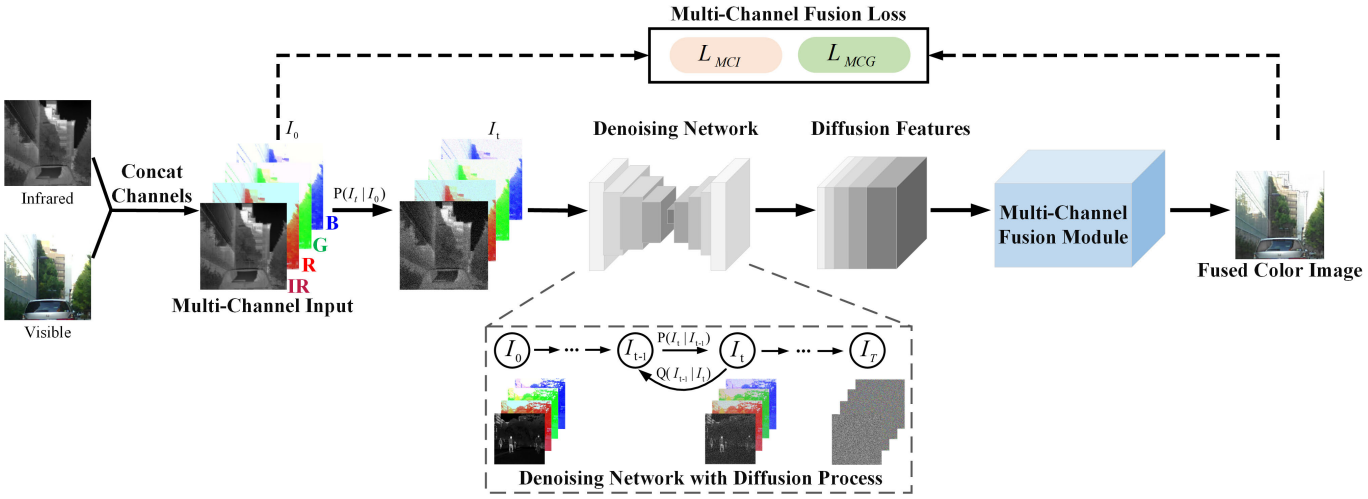


Fig. 2. The overall framework of Dif-Fusion. $\boldsymbol{I_0}$ and $\boldsymbol{I_t}$ denote the multi-channel input and the multi-channel data in the forward diffusion process with t timesteps. $P(\cdot|\cdot)$ and $Q(\cdot|\cdot)$ stand for the forward diffusion possess and reverse diffusion process. $\mathcal{L}_{MCI}$ and $\mathcal{L}_{MCG}$ represent multi-channel intensity loss and multi-channel gradient loss.

Since GAN has the ability to estimate the probability distribution in an unsupervised manner, some image fusion techniques based on GAN are proposed [2], [3]. Among them, FusionGAN [9] creates an adversarial game between the generator and the discriminator. The generator aims to generate the fused image, while the discriminator attempts to force the fused image to contain more details of the visible image [9]. In order to address the problem that the discriminator is only used to distinguish visible images, a dual-discriminator conditional generative adversarial network (DDcGAN) is proposed, which uses two discriminators to identify the structural differences between the fused image and the source images [54]. To help the generator focus on the foreground object information of the infrared image and the background details of the visible image, researchers exploit multi-scale attention mechanisms to fuse infrared and visible images for both generator and discriminator (AttentionFGAN) [58]. To balance the information between infrared and visible images, a fusion method named generative adversarial network with multiclassification constraints (GANMcC) is proposed [22]. However, the fusion methods mentioned above based on the generation model, whose generators add gradient fidelity and intensity fidelity constraints during training, cannot realize the distribution construction of infrared and visible images in the latent space. At the same time, the existing methods usually convert the three-channel visible image into a single channel image, making it challenging to fully utilize the multi-spectral information and achieve high color fidelity.

B. Diffusion Models

Diffusion models have become a powerful family of deep generation models [59], [60], with record breaking performance in many areas [61], including image generation [32], [62], [63], [64], image inpainting [65], image super-resolution [66], [67], [68], and image-to-image translation [69], [70]. In addition, the feature representations learned from the diffusion models are also found to be very useful in discriminative tasks, including image classification [71], image segmentation [72], [73] and object detection [74]. The diffusion model is a deep generative model with two processes, namely the forward process and the reverse process [75]. In the forward process, the input data is gradually disturbed in several time-steps by adding Gaussian noise. In the reverse process, the task of the model is to recover the original input data through multiple reverse time-steps by reducing the difference between the added noise and the predicted noise [72].

Diffusion models are widely used to generate samples because of the high quality and variety of samples generated by the models [75]. With its continuous development in various fields, diffusion models break the long-term dominant position of the generation adversarial network in the image generation field [59]. The fusion of infrared and visible images can also be regarded as an image generation task. This paper explores an effective way to achieve state-of-the-art fusion results with diffusion models.

III. Method

In this section, we describe the Diffusion-based image fusion framework for multimodal data in detail. The main idea of the proposed method is illustrated in Fig. 2. The visible image and infrared image pairs are concatenated along the channel dimension to make a multi-channel input for the diffusion models. In the forward process, Gaussian noise is gradually added to the multi-channel data until the data is close to pure noise (e.g., $P(\boldsymbol{I_t}|\boldsymbol{I_{t-1}})$), as shown in Fig. 2. Then, the reverse process tries to predict and remove the added noise with the help of a denoising network (e.g., $Q(\boldsymbol{I_{t-1}}|\boldsymbol{I_t})$). After that, diffusion features can be extracted from the diffusion models and fed to the proposed multi-channel fusion module, as shown in Fig. 2. The chromatic fusion images will be produced by the proposed framework directly under the guidance of the proposed multi-channel losses. It should be noted that the dashed box at the bottom of Fig. 2 illustrates the forward

diffusion process and the reverse diffusion process, without including the training process of the denoising model.

In the following subsections, we first go through how diffusion models learn the multi-channel distribution and generate new image pairs. Next, a multi-source information aggregation method based on diffusion models is presented in detail. Finally, we introduce multi-channel intensity loss and multi-channel gradient loss to guide the fusion network's training process.

A. Joint Diffusion With Infrared and Visible Images

Given a pair of registered infrared image $\boldsymbol{I_{ir}} \in \mathbb{R}^{H\times W\times 1}$ and visible image $\boldsymbol{I_{vis}} \in \mathbb{R}^{H\times W\times 3}$, where H and W represent the height and width, respectively. In order to learn the joint latent structure of multi-channel data, the 1-channel infrared image and 3-channel visible image are concatenated to form a 4-channel image, which is represented by a $\boldsymbol{I} \in \mathbb{R}^{H\times W\times 4}$. We adopt the diffusion process proposed in Denoising Diffusion Probabilistic Model (DDPM) [32] to construct the distribution of multi-channel data. The forward diffusion process of the multi-channel image is to gradually add noises with T timesteps. In the reverse process, the noise is gradually eliminated through T timesteps. The goal of training the diffusion models with the forward and reverse process is to learn the joint latent structure of the infrared and visible images by modeling the diffusion of the 4-channel images in the latent space [76], [77].

1) Forward Diffusion Process: The forward diffusion process inspired by non-equilibrium thermodynamics [62] can be viewed as a Markov chain that gradually adds Gaussian noise to the data with T timesteps [32]. At time-step t, the noisy multi-channel image $\boldsymbol{I_t}$ can be represented as follows:

$$P(\boldsymbol{I_t}|\boldsymbol{I_{t-1}}) = \mathcal{N}(\boldsymbol{I_t}; \sqrt{\alpha_t}\boldsymbol{I_{t-1}}, (1-\alpha_t)\boldsymbol{Z}) \tag{1}$$

where $\boldsymbol{Z}$ denotes the standard normal distribution. $\boldsymbol{I_t}$ and $\boldsymbol{I_{t-1}}$ represent the noisy 4-channel images generated by adding Gaussian noises for t and $t-1$ times, respectively. α_t is the variance schedule that controls the variance of the Gaussian noise added in timestep t. More specifically, for the first timestep, the noisy 4-channel image $\boldsymbol{I_1}$ can be formulated as:

$$\boldsymbol{I_1} = \sqrt{\alpha_1}\boldsymbol{I_0} + \sqrt{1-\alpha_1}\boldsymbol{\gamma} \tag{2}$$

where $\boldsymbol{I_0} = \boldsymbol{I}$, $\boldsymbol{\gamma} \in \mathbb{R}^{H\times W\times 4}$ is the Gaussian noise. Given the original input $\boldsymbol{I_0} \in \mathbb{R}^{H\times W\times 4}$, the expression of $\boldsymbol{I_t}$ can be deduced by Eq. (1) and Eq. (2):

$$P(\boldsymbol{I_t}|\boldsymbol{I_0}) = \mathcal{N}(\boldsymbol{I_t}; \sqrt{\bar{\alpha}_t}\boldsymbol{I_0}, (1-\bar{\alpha}_t)\boldsymbol{Z}) \tag{3}$$

where $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$. In the forward diffusion process, given the timestep t, variance schedule $\alpha_1, \ldots, \alpha_t$, and the sampled noise, the noisy multi-channel sample of timestep t can be directly calculated by Eq. (3).

2) Reverse Diffusion Process: In the reverse diffusion process, neural networks are used to perform a series of small denoising operations to obtain the original multi-channel image [72]. In each timestep of the reverse process, the denoising operation is performed on the noisy multi-channel image $\boldsymbol{I_t}$ to obtain the previous image $\boldsymbol{I_{t-1}}$ [78]. The probability distribution of $\boldsymbol{I_{t-1}}$ under the condition $\boldsymbol{I_t}$ can be formulated as [32]:

Fig. 3. Visible and infrared image pairs generated from the diffusion models.

$$Q(\boldsymbol{I_{t-1}}|\boldsymbol{I_t}) = \mathcal{N}(\boldsymbol{I_{t-1}}; \mu_\theta(\boldsymbol{I_t}, t), \sigma_t^2\boldsymbol{Z}) \tag{4}$$

where σ_t^2 is the variance of the conditional distribution $Q(\boldsymbol{I_{t-1}}|\boldsymbol{I_t})$, which can be formulated as:

$$\sigma_t^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t \tag{5}$$

where $\beta_t = 1-\alpha_t$. The mean $\mu_\theta(\boldsymbol{I_t}, t)$ of the conditional distribution $Q(\boldsymbol{I_{t-1}}|\boldsymbol{I_t})$ can be formulated as:

$$\mu_\theta(\boldsymbol{I_t}, t) = \frac{1}{\sqrt{\alpha_t}}(\boldsymbol{I_t} - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\boldsymbol{I_t}, t)) \tag{6}$$

where $\epsilon_\theta(\cdot,\cdot)$ is the denoising network. The inputs of $\epsilon_\theta(\cdot,\cdot)$ are the timestep t and the noisy multi-channel image $\boldsymbol{I_t}$.

3) Loss Function of Diffusion Process: First, we sample a pair of registered visible and infrared image pairs $(\boldsymbol{I_{ir}}, \boldsymbol{I_{vis}})$ in the training set to form the multi-channel image $\boldsymbol{I}$. Then we sample the noise $\boldsymbol{\gamma}$ from the standard normal distribution. Third, we sample the timestep $t \sim U(\{1, \ldots, T\})$ from the uniform distribution. After completing the above sampling, the loss function of the diffusion models can be formulated as:

$$\mathcal{L}_{diff} = \left\| \boldsymbol{\gamma} - \epsilon_\theta(\sqrt{\bar{\alpha}_t}\boldsymbol{I_0} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\gamma}, t) \right\|_2 \tag{7}$$

where $\|\cdot\|_2$ denotes the L2 norm.

4) Structure of the Denoising Network: In order to predict the noise added in the forward diffusion process, the structure of the denoising network $\epsilon_\theta(\cdot,\cdot)$ adopts the U-Net structure used in SR3 [66]. The SR3 backbone consists of a contracting path, an expansive path and a diffusion head. The contracting path and the expansive path are composed of 5 convolution layers. The diffusion head consists of a single convolution layer, which is used to generate the predicted noise.

Fig. 3 shows some paired visible and infrared images generated by our trained diffusion models. These image pairs can be seen to visually resemble the real visible and infrared images. The targets that are highlighted in the corresponding infrared images also appear plausible. These results demonstrate that the diffusion models is a powerful tool for constructing the distributions of multi-channel data.

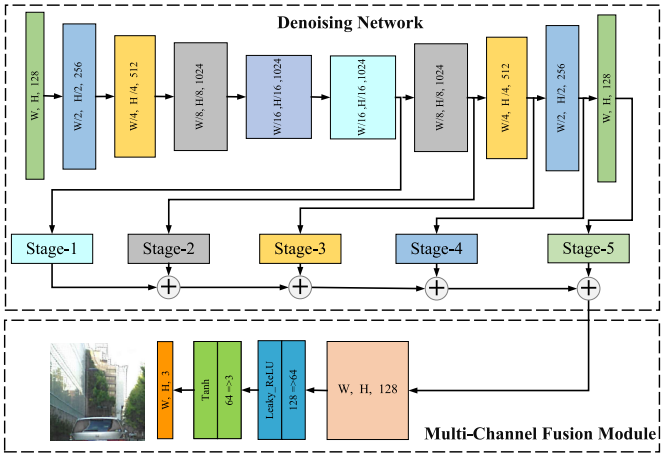


Fig. 4. The structure of the denoising network and the multi-channel fusion module.

B. Fusion With Multi-Channel Diffusion Features

After training the denoising network, we use the denoising network to extract the multi-channel features. In the image fusion training stage, we use two kinds of losses (i.e., multi-channel gradient loss and multi-channel intensity loss) for training. It is worth noting that by using the multi channel loss, the three-channel fused images can be generate directly without color space transformation.

1) Multi-Channel Diffusion Features: For the SR3 backbone, its expansive path contains five convolution layers, and the sizes of its output feature maps are $W/16$, $H/16$, $W/8$, $H/8$, $W/4$, $H/4$, $W/2$, $H/2$, W, H. We use a multi-channel fusion module to fuse multi-channel diffusion features from 5 stages of the denoising network [72], [78]. Here, we use the term "stage" to represent diffusion features from different decoder layers. Due to the varying output sizes of diffusion features, up-sampling and 1×1 convolutional layers are applied to combine them. For the five stage features of the five expansive layers, we add them up and feed them into the fusion head to generate the fused image $\boldsymbol{I_f} \in \mathbb{R}^{H \times W \times 3}$. Specifically, 3×3 convolutional layers are applied to map high-dimensional fused features to 3-channel outputs. Leaky ReLU and Tanh are adopted as the activation function. The structure of the denoising network and the multi channel fusion module is shown in Fig. 4.

2) Loss Function of Fusion Process: Since the visible image has abundant texture information, in order to retain sufficient texture information in the final fused image, we apply a gradient loss for gradient fidelity. However, the existing gradient loss is designed for single-channel fused images [55]. To directly generate three-channel fused image while keeping the gradient, we extends the existing gradient loss and proposes multi-channel gradient loss $\mathcal{L}_{MCG}$, which can be formulated as:

$$\mathcal{L}_{MCG} = \frac{1}{HW} \sum_{i=1}^{3} \left\| \nabla \boldsymbol{I}_f^i - \max(\nabla |\boldsymbol{I_{ir}}|, \nabla |\boldsymbol{I}_{vis}^i|) \right\|_1 \quad (8)$$

where ∇ represents the gradient operator. $\boldsymbol{I}_f^1$, $\boldsymbol{I}_f^2$ and $\boldsymbol{I}_f^3$ represent the there channels (i.e., red, green and blue) of the fused image $\boldsymbol{I}_f$. $\boldsymbol{I}_{vis}^1$, $\boldsymbol{I}_{vis}^2$ and $\boldsymbol{I}_{vis}^3$ denote the there channels of the input visible image $\boldsymbol{I_{vis}}$. $\|\cdot\|_1$ denotes the L1 norm. The thermal radiation is usually characterized by pixel intensity [79]. We apply intensity loss to make the fused image have a intensity distribution similar to the infrared image and the visible image. However, similar to gradient loss, the current intensity loss is designed for generating single channel fused images [55]. We extend the existing intensity loss into multi-channel intensity loss $\mathcal{L}_{MCI}$, which can be formulated as:

$$\mathcal{L}_{MCI} = \frac{1}{HW} \sum_{i=1}^{3} \left\| \boldsymbol{I}_f^i - \max(\boldsymbol{I_{ir}}, \boldsymbol{I}_{vis}^i) \right\|_1 \quad (9)$$

Existing fusion methods usually preserve color information through color space conversion. In order to solve this problem and make full use of diffusion features, this paper directly generates three-channel fused images with multi-channel gradient and intensity losses. The final loss $\mathcal{L}_f$ can be formulated as:

$$\mathcal{L}_f = \mathcal{L}_{MCG} + \mathcal{L}_{MCI} \quad (10)$$

IV. Experiments

In this section, we first describe the experiment details, including datasets, evaluation metrics, and the training process. Then, we conduct quantitative and qualitative analysis on three public datasets to evaluate the proposed framework. Also, we compare the performance of our method with six state-of-the-art models in order to demonstrate the benefits of Dif-Fusion. Finally, we reveal the effectiveness and advantages of using diffusion models in multi-channel information fusion based on the ablation study.

A. Experimental Settings

1) Datasets: We utilize the color and infrared image pairs from the MSRS [21], RoadSence [1], and M3FD datasets [30] to evaluate the proposed framework. We also compare our method with six state-of-the-art algorithms: FusionGAN [9], SDDGAN [31], GANMcC [22], SDNet [29], U2Fusion [1], and TarDAL [30]. SDNet and U2Fusion are fusion approaches based on CNN architectures, while FusionGAN, SDDGAN, GANMcC and TarDAL are based on generative models and their variants. For the methods that are being compared, fused images are generated and evaluated using publicly accessible codes and pre-trained models [55]. To produce color results for visual analysis and quantitative evaluation, those single-channel fused results from comparison methods will be converted to color images in post-processing.

2) Evaluation Metrics: Six statistical metrics are used in the quantitative evaluation, five of which are mutual information (MI) [80], visual information fidelity (VIF) [81], spatial frequency (SF) [82], Qabf [83], and standard deviation (SD). MI primarily assesses how well the information from the initial image pairs has been aggregated in the fused image. VIF evaluates the fidelity of the information present in the fused image. The spatial frequency-related information in the combined data is measured by SF. The edge information from

the source images is quantified using Qabf. SD primarily evaluates the contrast of composite images.

Specifically, we introduce the Delta E [84], a color difference calculation index built in CIELAB space that is believed to be more in line with the human perception system [85], to quantify the color distortion between the fused image and the original visible image. CIELAB is an abbreviation for the International Commission on Illumination (abbreviated CIE) $L^*a^*b^*$ color space, where L^* corresponds to perceptual lightness, and a^* and b^* represent the four unique colors of human vision: red, green, blue, and yellow. Delta E is a sort of color distance measurement. As the human eye is more sensitive to some colors than others due to perceptual non-uniformities, the Euclidean distance directly measured in the color space does not match human perception [84]. The Delta E is recommended as a solution to these problems, along with several corrections for neutral colors, lightness, chroma, hue, and hue rotation [85].

It should be noted that whereas other measures require the original images, SF and SD metrics can be calculated directly on the fused images. A lower Delta E value suggests smaller color distortion and better fusion quality, but the other five metrics work in reverse, with a higher value indicating a better fusion result.

3) Training Details: The proposed model is trained on the MSRS dataset, which includes 1083 training pairs of visible and infrared images. There are 361 pairs of test images in the MSRS dataset. To train the diffusion models, we adopt the training settings in [78] and [66]. Specifically, we randomly crop 160 × 160 patches from visible and infrared images in the training process. Inspired by the recent work [72], [78], we extract diffusion features generated at three time-steps (e.g., 5, 50, 100) to form multi-channel diffusion features. When training the fusion module, the Adam optimizer is utilized to minimize loss, and the learning rate is set to 0.0001. We set the batch size to 24 and the model is trained for 300 epochs. In the test, the visible and infrared images are fed to the networks with the original size. The outputs of our model are color images, which are directly used in qualitative and quantitative analyses. The proposed model is implemented based on PyTorch. All the experiments involved were carried out on a workstation containing the NVIDIA RTX3090 GPU and 3.80 GHz Intel (R) Core (TM) i7-10700K CPU.

B. Fusion Performance Analysis

1) Qualitative Results: The MSRS dataset consists of both daytime and nighttime scenarios. To demonstrate the advantages of our method in complementary information fusion, texture preservation, and color fidelity, we select two image pairs from each of the two scenarios to show the results from different models. Infrared images highlight objects in daytime scenes that have high thermal radiation information, whereas visible images contain rich texture and color information. We hope that the fused color images will be able to emphasize the significant targets in the infrared images while preserving the original visible image's fine-grained texture and color information.

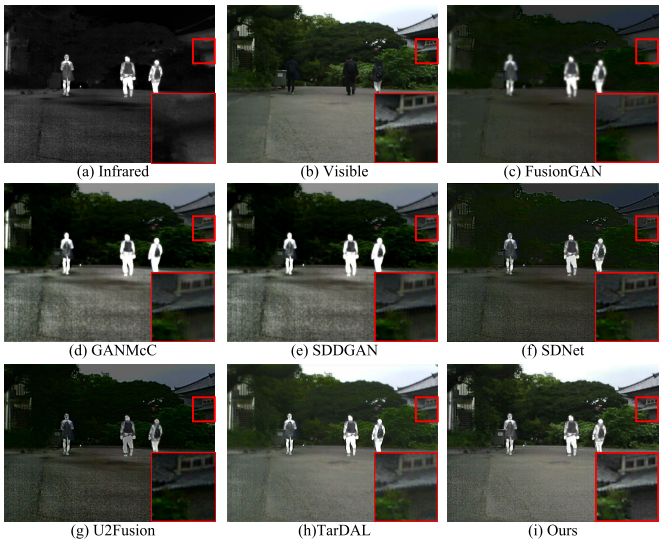


Fig. 5. Qualitative comparison of Dif-Fusion with six state-of-the-art methods on the 00634D image pair from the MSRS dataset.

The infrared image in Fig. 5 highlights three pedestrians, which are preserved in the fused images generated by all methods. However, only the results of our method and TarDAL closely resemble the original visible image. The composited images obtained by other methods (such as FusionGAN, GANMcC, etc.) are visually darker and have large color distortions, such as green trees turn into black in the fused images produced by SDDGAN, U2Fuison, SDNet. The red box in Fig. 5 enlarges the details of the windows under the eaves to demonstrate the benefits of our method in detail maintenance. Only the results from FusionGAN, SDDGAN, TarDAL, and our method represent the fact that the area beneath the eaves in the infrared image has a marginally higher brightness than the surroundings. But only our approach can clearly maintain the window's distinctive contours and arrangements. Additionally, our method makes it easy to discern between the foreground (greenery) and the background (walls beneath windows) in the fused image.

Another pair of daytime images is shown in Fig. 6. The infrared image primarily shows two bicyclists and some distant pedestrians as highlighted targets. This feature is apparent in the composite images produced by all approaches. Like Fig. 5, the results from TarDAL and our approach visually resemble the original visible image more closely. The regions in the red and green rectangles have been enlarged. The window structure is only conspicuously visible in the infrared image and not in the visible image. The green region, on the other hand, has a white sign that is only visible in the visible image and not in the infrared image. Some methods (e.g., FusionGAN, GANMcC) struggle to display these features clearly. U2Fusion and SDNet can display the structural information in the red box, while SDDGAN and TarDAL can emphasize the information in the green box. However, only our approach can simultaneously maintain the crucial characteristics in both rectangles. The analysis above reflects the advantages of our method in terms of color preservation,

Fig. 6. Qualitative comparison of Dif-Fusion with six state-of-the-art methods on the 00537D image pair from the MSRS dataset.

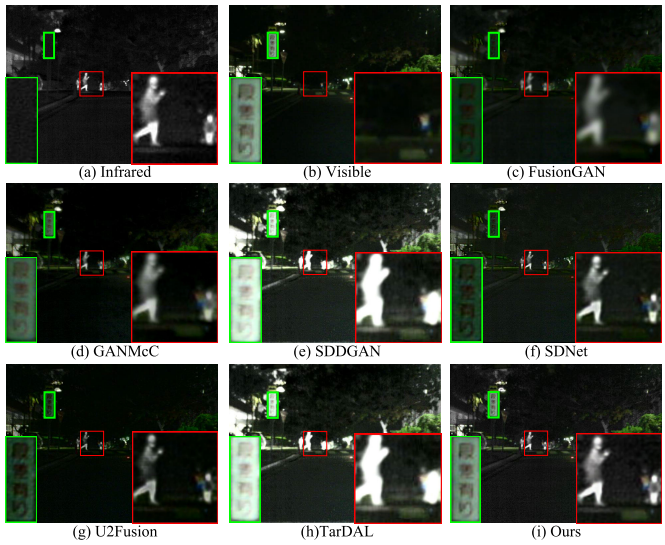


Fig. 7. Qualitative comparison of Dif-Fusion with six state-of-the-art methods on the 00878N image pair from the MSRS dataset.

learning complementary information, characterizing details, etc.

In nighttime scenarios, the infrared image still highlights the target with active thermal radiation, whereas the visible image only has rich color and texture when there is adequate illumination. Therefore, we anticipate that the fused three-channel image will be able to preserve the valuable information in the visible image as well as the infrared image's highlighted targets. We enlarged the red and green rectangular areas in Fig. 7 to highlight the benefits of our approach. In the red box, there are two pedestrians in the infrared image, one of whom is crossing the road. The green box has a signboard with text that is highlighted in the visible image but is entirely black in the infrared image. The red box also contains the zebra crossing from the visible image in the lower middle region. Nearly all methods emphasize the two pedestrians in the red box to varying degrees. However, two things should be mentioned. First, the body covered by clothes in the original infrared image is not as bright as the other areas. This difference is neglected by SDDGAN and TarDAL, i.e., the whole body is equally bright, resulting in the loss of structural information. Second, many methods (e.g., FusionGAN, GANMcC, U2Fusion, SDNet) overlook the zebra crossing information from the visible image. These two issues are both avoided by our method. In addition, compared with other methods, the proposed method better preserves the information from the visible image in the green box, including brightness, color, and clarity.

In the second pair of nighttime images in Fig. 8, we show the fusion results in a complex lighting scene. In the infrared image, the highlighted object is a pedestrian. The area with medium brightness is located inside the windows, and the region with weak brightness is the irregular wall surfaces on the right. Areas with various colors and rich textures in the visible image, such as a white car and road surfaces, are mostly found on the left side of the image. Also, window regions in the visible image are bright. We expect the fused image to contain critical information with different levels of brightness in the infrared image. In addition, we expect to maintain the authenticity of color and texture in the visible image. To illustrate the advantages of the proposed method, we zoom in on the details of the weak signal in the red box. Going through these fused images, we observe that it is challenging to distinguish surface characteristics in the enlarged images from FusionGAN, GANMcC, U2Fusion, and TarDAL. Although the results of SDDGAN and SDNet contain this structure, they are kind of blurred or contaminated by noise. Only the fused image generated by our method is close to the original infrared image in both clarity and brightness. The images produced by FusionGAN, SDNet, GANMcC, and U2Fusion all exhibit color distortions, e.g., the white vehicle inside the green box looks green. In short, the proposed method can still preserve color fidelity in visible images and weak information from infrared images against a nighttime background by extracting complementary information from multi-channel data.

2) Quantitative Results: We quantitatively compare the proposed method with six state-of-the-art methods. Fig. 9 shows the quantitative results of six statistical metrics on the MSRS dataset. For each metric, we generated cumulative distribution plots, illustrating the metric's cumulative probability across the entire test dataset. The horizontal axis of these plots represents cumulative probability, while the vertical axis represents metric values. Given a metric value on the y-axis, the cumulative probability on the x-axis indicates the proportion of samples in the test dataset that are less than or equal to that value. In these charts, better performance is denoted by higher values along the y-axis from the beginning point on the x-axis for all metrics except Delta E. Lower values on the y-axis indicates better performance for Delta E.

We can see that our method exhibits notable benefits in five metrics (i.e., MI, VIF, Qabf, SD, and Delta E). The highest MI indicates that our method successfully transfers the most information from multi-channel source images to fused images. The best VIF shows that the fused images generated

Fig. 8. Qualitative comparison of Dif-Fusion with six state-of-the-art methods on the 01061N image pair from the MSRS dataset.

by Dif-Fusion are more in line with the human visual system. Our Dif-Fusion exhibits the best Qabf, thus more edge information is maintained. In addition, the proposed method achieves the best SD, which means that our fused images have the largest contrast. Moreover, because multi-channel complementary information is been exploited with diffusion models, our method is significantly higher than the compared method in terms of color fidelity indicator (Delta E). In the SF metric, the proposed method is just marginally inferior to SDDGAN and TarDAL.

C. Generalization Experiment

1) Qualitative Results: On the RoadScene and M3FD datasets, we test the model trained on the MSRS dataset to assess the generalization performance. The comparison methods are also tested on these two datasets. We chose one example from each dataset for an in-depth qualitative study in order to highlight the advantages of our method.

Fig. 10 is from the RoadScene dataset. The visible image mainly consists of roads, trees, vehicles, and the sky, while the infrared image highlights the lower part of the car and part of the road surface. From the perspective of visual perception, the fused image produced by our method is the one that most resembles the original visible image. Although the bright areas in the infrared image have been preserved to some extent in the images fused using various methods, the colors of the sky and trees in the images fused by FusionGAN, GANMCC, SDDGAN, and SDNet have altered significantly. U2Fusion and TarDAL produce images with less color distortion than previous methods, but their output is blurry and lacks significant structure information (e.g., tree crown). The fused image of Dif-Fusion, in comparison, effectively preserves the salient information in the infrared image while maintaining the color and texture of prominent regions (such as the sky and trees) in the visible image. In the red rectangle, the rear of a van is enlarged. The outlines of the carriage and its wheel are muddled and cluttered in the fused image created by FusionGAN, GANMcC, SDDGAN, SDNet, U2Fusion, and TarDAL. Only our results preserve the region's color and structure details from the visible image. The ability of the proposed method to extract complementary information as well as its advantages in texture and color preservation are demonstrated by this phenomenon.

We chose an underground garage scenario from the M3FD dataset for qualitative analysis, as seen in Fig. 11. In this example, the pipeline structure and background wall are highlighted in the infrared image. First, the fused images produced by TarDAL and our approach are quite similar to the original visible image in terms of overall perception. Due to the improper combination of complementary information, FusionGAN, SDNet, and U2Fusion replace the brightness of the pillar in the visible image with the brightness of the infrared image, resulting in an excessively dark pillar on the right side of the image. GANMcC, SDDGAN, and TarDAL partially alleviate this issue. In their composite images, the pillar retains some texture and color information from the original visible image, but they are not as effective as the proposed method. Additionally, SDDGAN and TarDAL encounter the same problem as Fig. 7, i.e., the brightness is over-enhanced, which results in the loss of wall structural information. The reflecting corner guard, which is present in the original visible image, is indicated by the red rectangle and enlarged. A thorough examination of all the enlarged views reveals that only our fused image preserves the logo's color and structural details while preserving brightness. In short, the above analysis demonstrates that the proposed method has strong generalization abilities. It can mine complementary information from multimodal data in different scenes and has good capability in texture and color retention.

2) Quantitative Results: We follow the previous work [55] by choosing 25 pairs of images from two datasets other than the MSRS dataset to quantitatively evaluate the generalization performance of the proposed method. Tables I and II show the quantitative results of six statistical metrics on the M3FD and RoadScene datasets compared with six state-of-the-art methods. As shown in Table I, we can see that Dif-Fusion ranks first in six metrics on the M3FD dataset. The experimental results show that the fused image generated by our method has rich texture details, the highest contrast, and the best visual quality. Fig. 12 shows the cumulative distributions in M3FD test, and it also illustrates that the proposed method demonstrates superior performance in comparison to other methods, aligning with the evaluation findings on MSRS. Compared to other methods, our method shows noticeably higher values across MI, VIF, SF, and SD. In terms of Qabf, our approach ranks closely behind U2Fusion and SDNet, taking up the top three positions in order. Compared to alternative generative models like TarDAL, GANMcC, SDDGAN, and FusionGAN, our method also displays definite benefits in extracting edge information. Furthermore, the proposed method yields a low Delta E value, significantly distinguishing itself from all comparative methods. This highlights our model's ability to fuse multimodal information while maintaining high color fidelity.

According to Table II, Dif-Fusion outperforms the compared methods in terms of VIF, Qabf, and Delta E on the

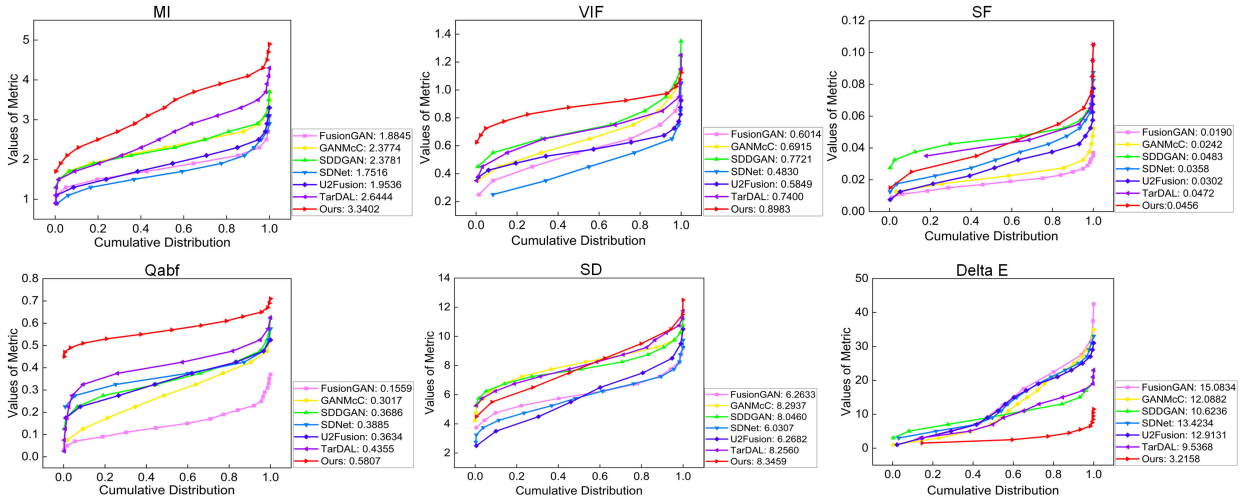


Fig. 9. Quantitative comparisons between Dif-Fusion and six state-of-the-art methods on MSRS dataset with six metrics, i.e., MI, VIF, SF, Qabf, SD, and Delta E.

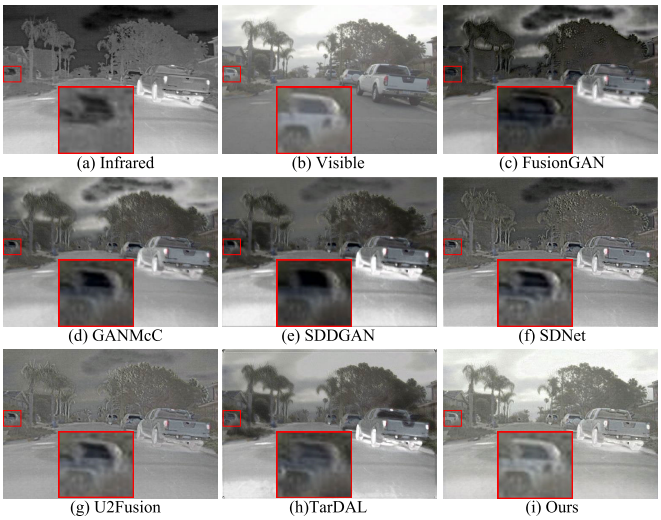


Fig. 10. Qualitative comparison of Dif-Fusion with six state-of-the-art methods on the FLIR00977 image pair from the RoadScene dataset.

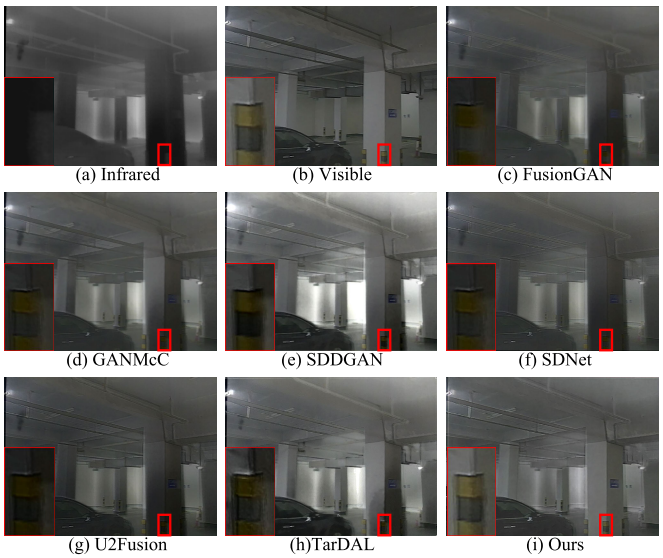


Fig. 11. Qualitative comparison of Dif-Fusion with six state-of-the-art methods on the 02757 image pair from the M3FD dataset.

TABLE I

FUSION QUALITY EVALUATION ON 25 IMAGE PAIRS FROM THE M3FD DATASET. BOLD INDICATES THE BEST RESULTS

Methods	MI	VIF	SF	Qabf	SD	Delta E
FusionGAN	2.4493	0.5530	0.0413	0.3539	9.4647	20.0319
GANMcC	2.3838	0.6893	0.0365	0.3376	9.8799	18.1298
SDDGAN	2.8018	0.8005	0.0477	0.3852	9.2755	19.3442
SDNet	2.6705	0.6870	0.0667	0.5609	9.6364	20.9597
U2Fusion	2.3374	0.7038	0.0522	0.5696	9.5848	16.5465
TarDAL	2.4727	0.7999	0.0604	0.4428	9.7364	13.5831
Ours	**2.9592**	**0.8543**	**0.0694**	**0.5771**	**10.0848**	**4.9644**

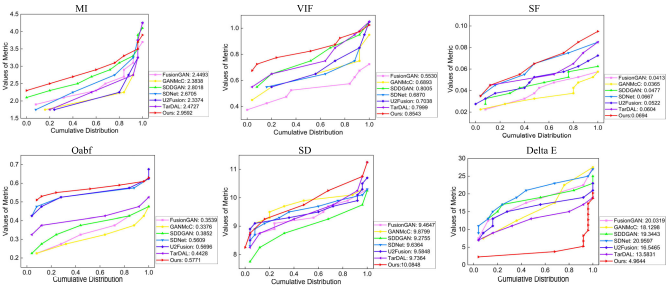


Fig. 12. Quantitative comparisons between Dif-Fusion and six state-of-the-art methods on 25 image pairs from M3FD dataset with six metrics, i.e., MI, VIF, SF, Qabf, SD, and Delta E.

RoadScene dataset. Moreover, Dif-Fusion ranks first in Delta E on both M3FD and RoadScene datasets, which implies that the proposed method can improve color fidelity while ensuring the amount of information. Fig. 13 shows the cumulative distributions in RoadScene test. The diagram shows that Dif-Fusion consistently ranks in the top three of all seven methods across all six metrics, even though the benefits of our results may not be as obvious when compared with the MSRS and M3FD datasets. In VIF, Qabf, and Delta E, the proposed model impressively claims to be the best. On MSRS and M3FD datasets, our method has a significant advantage over generative model-based methods. While TarDAL exhibits performance that is similar to ours on this dataset, it is

TABLE II

FUSION QUALITY EVALUATION ON 25 IMAGE PAIRS FROM THE ROAD-SCENE DATASET. BOLD INDICATES THE BEST RESULTS

Methods	MI	VIF	SF	Qabf	SD	Delta E
FusionGAN	2.8301	0.6137	0.0384	0.2985	10.0778	21.2553
GANMcC	2.8672	0.6957	0.0397	0.3740	10.1522	17.8119
SDDGAN	3.0513	0.7121	0.0425	0.3244	9.5164	21.8514
SDNet	3.3135	0.7799	**0.0606**	0.4724	9.8665	18.4423
U2Fusion	2.7190	0.6710	0.0479	0.4864	9.7104	11.6658
TarDAL	**3.3666**	0.8020	0.0570	0.4399	**10.2686**	12.8190
Ours	3.3073	**0.8054**	0.0516	**0.5181**	10.1065	**9.1714**

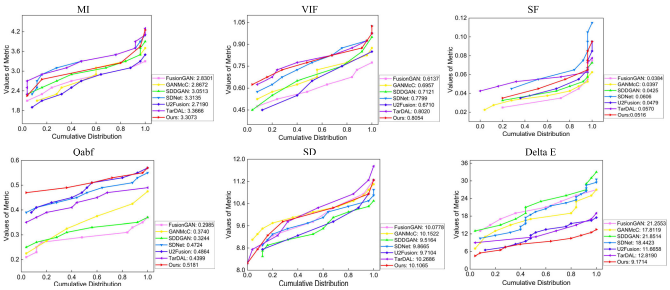


Fig. 13. Quantitative comparisons between Dif-Fusion and six state-of-the-art methods on 25 image pairs from RoadScene dataset with six metrics, i.e., MI, VIF, SF, Qabf, SD, and Delta E.

TABLE III

THE FUSION PERFORMANCE WITH AND WITHOUT THE DIFFUSION PROCESS ON THREE DATASETS

	MI	VIF	SF	Qabf	SD	Delta E
Dataset			MSRS			
w/o Dif.	2.9364	0.8181	0.0376	0.4228	**8.4097**	4.2378
w/ Dif.	**3.3402**	**0.8983**	**0.0456**	**0.5807**	8.3459	**3.2158**
Dataset			M3FD			
w/o Dif.	2.5870	0.7903	0.0588	0.4088	9.8860	9.6635
w/ Dif.	**2.9592**	**0.8543**	**0.0694**	**0.5771**	**10.0848**	**4.9644**
Dataset			RoadScene			
w/o Dif.	3.2656	0.7615	0.0431	0.3478	8.9175	10.1185
w/ Dif.	**3.3073**	**0.8054**	**0.0516**	**0.5181**	**10.1065**	**9.1714**

noteworthy that the Delta E of TarDAL greatly exceeds ours, indicating low color fidelity.

D. Ablation Study

The proposed framework adopts the diffusion models to extract multi-channel information, which improves color fidelity and visual quality with the help of multi-channel complementary information. In order to verify the effectiveness of the diffusion models, we ablate the diffusion process. More specifically, for the sake of fairness, we designed an autoencoder network following a UNet-style architecture and used this AE network to replace the diffusion module. We summarize the results of the ablation study in Table III. In the MSRS dataset, after removing the diffusion process, the performance of our method decreases on five metrics (i.e., MI, VIF, SF, Qabf, and Delta E). On the M3FD dataset and the RoadScene dataset, after removing the diffusion process, the performance of our method decreases on all six metrics. It is worth noting that in the M3FD dataset, the color fidelity decreases significantly after removing the diffusion process, which indicates that the distribution of multi-channel information and the extraction of multi-channel complementary information play a very important role in color preservation.

TABLE IV

THE FUSION PERFORMANCE WITH AND WITHOUT THEMULTI-CHANNEL LOSS ON MSRS DATASET

Losses	MI	VIF	SF	Qabf	SD	Delta E
w/o MCL	2.8284	0.8787	**0.0616**	0.4677	**8.5766**	25.7281
w/ MCL	**3.3402**	**0.8983**	0.0456	**0.5807**	8.3459	**3.2158**

In the second ablation study, we verify the effectiveness of our multi-channel gradient and intensity losses (MCL) by replacing them with single-channel gradient and intensity losses. On the MSRS dataset, multiple accuracy indicators exhibit a decrease, as shown in Table IV. Specifically, MI decreases from 3.3402 to 2.8284, VIF decreases from 0.8983 to 0.8787, and Qabf decreases from 0.5807 to 0.4677. Importantly, it is worth noting that a substantial drop in color preservation performance occurs when we switch from the multi-channel loss function to the single-channel loss function. Therefore, to directly input and generate multi-channel color images without experiencing color distortion caused by color conversion, it is essential to employ the proposed multi-channel gradient and intensity losses during the training process.

E. Discussions

1) Compared With Generative Model-Based Fusion Methods: The existing infrared and visible image fusion methods can be categorized into three major categories: CNN-based, generative model-based, and Transformer-based. The proposed Dif-Fusion falls under the second group. It is necessary to analyze and compare the proposed method with existing fusion algorithms based on generative models. As shown in Fig. 14, Dif-Fusion is compared with three recent generative model-based fusion algorithms (GANMcC, SDDGAN, TarDAL), where TarDAL can be recognized as the state-of-the-art in generative model-based methods. Overall, our method is capable of aggregating multi-modal information while preserving fine textures and original colors. Specifically, in the enlarged green box, the visible image contains rich facade structural details. SDDGAN and TarDAL struggle to incorporate these details into the fused image. GANMcC manages to extract a somewhat blurry facades but still falls significantly short compared to our results. Similarly, in the enlarged red box, our approach clearly displays the letters and numbers on the license plate, with minimal visual difference from the original visible image. The results from other methods are more contaminated by noise, and their outlines are less distinct. Notably, within the cyan dashed circle, we observe highlighted building regions in the near-infrared image. However, GANMcC completely loses this information in the fused image.

Together with the quantitative results and analysis in the previous sections, such as Fig. 9, we can conclude that

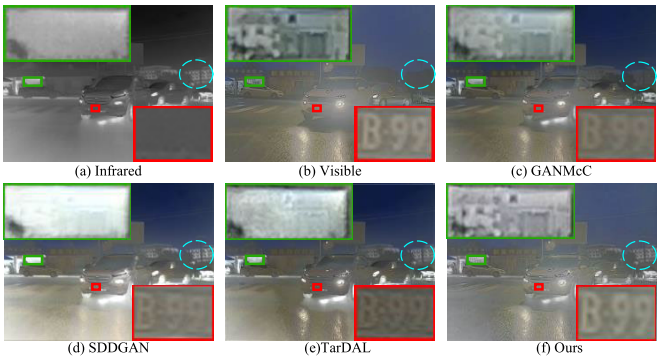


Fig. 14. Comparison with generative model-based fusion methods.

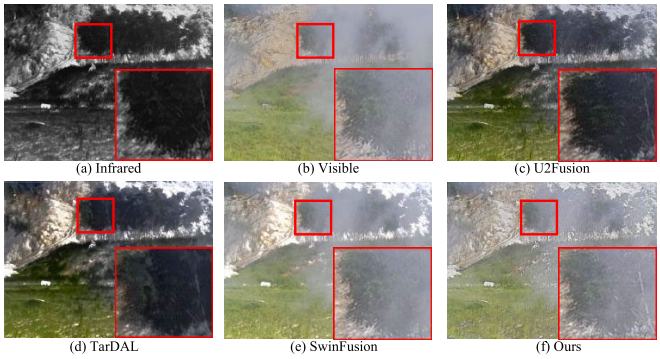


Fig. 15. Performance Comparison of U2Fusion, TarDAL, SwinFusion, and Our Method under Smoky Conditions.

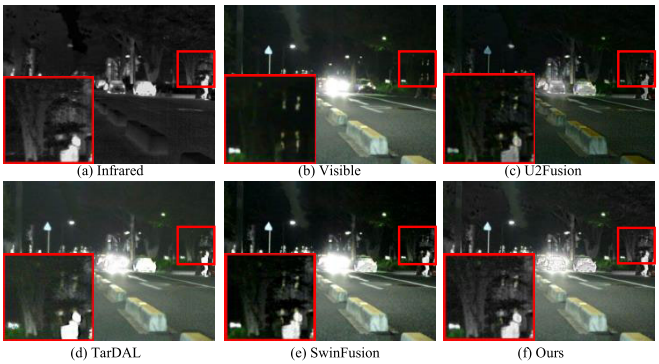


Fig. 16. Performance Comparison of U2Fusion, TarDAL, SwinFusion, and Our Method under Glare Conditions.

the Diffusion-based fusion model considerably improves the performance of the generative model-based frameworks. It is important to note that Diffusion is still under rapid development, we believe that the Diffusion-based fusion method will exhibit more compelling performance.

2) Compared With Different Types of Fusion Frameworks: The proposed model not only achieves new state-of-the-art (SOTA) performance among generative models, but also partially surpasses methods from other frameworks according to results reported in Fig. 9 and Table I. U2Fusion [1], a SOTA algorithm based on CNN, TarDAL [30], a SOTA algorithm based on generative models, and SwinFusion [50], a representative algorithm based on Transformers, are chosen to conduct additional analysis on challenging scenes (such as smoke and glare).

In the smoke scene, as depicted in Figure 15, both U2Fusion and TarDAL cannot keep color information for vegetation, while our method and SwinFusion maintain color fidelity well. However, the fusion process of SwinFusion is significantly affected by smoke interference, nearly hiding the tree canopy, and causing the loss of tree trunk details, as indicated in the enlarged red box. In addition, the TarDAL and U2Fusion results almost completely missing the smoke information. Achieving a balance between prominent objects and environmental information (such as weather conditions) in the fused images necessitates further discussion.

In the glare scene, as illustrated in Fig. 16, the CNN-based method effectively removes the influence of glare, revealing the complete outline of the car. However, this comes at the cost of overlooking the details of other regions of the image, such as the tree trunks and building facade within the enlarged red box, where U2Fusion's results are mostly dark. In comparison to the outcomes of TarDAL and SwinFusion, our approach better accentuates the car's contour and simultaneously maintains clear brightness and object structural information within the red box. Furthermore, in the results of U2Fusion and TarDAL, the sky among trees is completely invisible, with only SwinFusion and Dif-Fusion preserving the information of sky regions.

3) Different Training Datasets: In the previous experiments, the Dif-Fusion was trained on MSRS dataset and test on other datasets. This setting is widely used in assessing the generalization performance of the proposed method. To further analyze the impact of training dataset on performance across different datasets, we conducted retraining on the M3FD dataset based on the settings in [30]. The experimental results are summarized in Table V. The next-to-last row reports the metrics derived from the original model which was trained on the MSRS dataset, while the last row represents the results after retraining the Dif-Fusion model on M3FD dataset. Under this setting, Dif-Fusion still maintains an advantage in SF, Qabf, SD, and Delta E. Although its advantage is not as pronounced in the MI and VIF metrics, it still achieves first and second rankings. Comparing the metrics in the last two rows, we observe an improvement in performance for the retrained model, particularly in metrics like MI and Qabf. This result indicates that the fusion model based on Diffusion exhibits relatively stable performance across different scenes, and it also demonstrates a certain level of dependency on the training dataset.

In the RoadScene dataset, although our method ranks in the top three across all six metrics, with three of them being in the first position in Table II, the superiority of Dif-Fusion no longer matches its performance on the MSRS dataset. To further analyze the performance on RoadScene dataset, we randomly select half of the RoadScene dataset images to fine-tune the fusion model while leaving the rest for test. The experimental results are reported in Table VI. It demonstrates that our approach exhibits noticeable improvements in MI and VIF, and slight changes in Qabf, SF and SD while still retaining a low Delta E when using new samples to fine-tune the fusion module. However, compared to the performance on

TABLE V
FUSION QUALITY EVALUATION ON M3FD DATASET. BOLD, UNDERLINE, AND ITALIC RESPECTIVELY REPRESENT FIRST, SECOND, AND THIRD PLACE

Methods	MI	VIF	SF	Qabf	SD	Delta E
FusionGAN	3.0801	0.6048	0.0339	0.3191	9.5955	18.9728
GANMcC	2.9091	0.7764	0.0323	0.3648	9.8011	16.7376
SDDGAN	*3.3087*	*0.8600*	0.0388	0.3560	8.9258	20.4283
SDNet	3.4257	0.7518	0.0546	0.5148	9.3931	21.1140
U2Fusion	2.8999	0.7581	0.0431	*0.5267*	9.3647	16.9625
TarDAL	3.1877	**0.8752**	0.0554	0.4180	*9.8851*	*14.2018*
Ours	3.2085	0.8209	**0.0566**	0.5309	9.9281	**6.5598**
Ours w/ Re.	**3.5205**	0.8651	*0.0554*	**0.5758**	**9.9775**	6.6665

TABLE VI
FUSION QUALITY EVALUATION ON ROADSCENE DATASET WITH INCREASED SAMPLES

Methods	MI	VIF	SF	Qabf	SD	Delta E
w/o Fine.	3.0125	0.7538	**0.0580**	0.5173	**10.2184**	**8.9564**
w/ Fine.	**3.2019**	**0.7840**	0.0563	**0.5290**	10.1341	9.2264

TABLE VII
COMPARISONS OF DIFFERENT TIME-STEP COMBINATIONS

Time steps	(50,150,250) [72]	(50,100,400) [78]	(5, 50,100)
MI	3.0039	3.0352	**3.3402**
VIF	0.8199	0.8284	**0.8983**
SF	0.0434	0.0439	**0.0456**
Qabf	0.4798	0.4823	**0.5807**
SD	**8.4022**	8.3716	8.3459
Delta E	3.9025	3.6849	**3.2158**

the MSRS and M3FD datasets, there is still ample room for improvement. Through analysis the images of three datasets, we find that the RoadScene dataset is relatively small and exhibit significant differences in properties such as brightness and saturation when compared to the other two datasets. As the Diffusion-based fusion models primarily learn the distribution of different modalities in the training dataset, the evident differences between datasets might be the explanation for different performance. We think that the data diversity in training Diffusion-based fusion models worth further exploration in the future.

4) Time-Steps in Diffusion: In Diffusion-based applications, diffusion features of several time-steps are often combined. Selecting the optimal combination of time-steps remains an open problem. Our research primarily focuses on incorporating the diffusion model into the task of image fusion, and we implement the idea using three time-steps based on empirical experiments. To analyze the impacts of different combinations of time-steps, we selected two time-steps from the existing work [72], [78], and the evaluation results are shown in Table VII.

The experiments shows that different time-steps have a relatively minor impact on metrics such as MI, VIF, SF, SD, and Delta E. It is noteworthy that there is significant variation in the Qabf values. Among the three combinations, the largest Qabf value is 0.5807, while the smallest is 0.4798. However, all these values remain higher than the largest value of 0.4355 achieved by TarDAL, the best of the six comparative methods on this metric. These findings suggest that the performance of the proposed Diffusion-based fusion model is not heavily influenced by the choice of time-steps. Carefully selected time-steps may potentially enhance accuracy metrics, although achieving this might necessitate more engineering tricks and could be influenced by factors such as datasets and Diffusion model variations, warranting further exploration.

TABLE VIII
COMPARISONS OF THE AVERAGED PER IMAGE INFERENCE TIME FOR DIFFERENT METHODS IN MSRS DATASET

Methods	Times (s)
FusionGAN	0.041158
GANMcC	0.069583
SDDGAN	0.016002
SDNet	0.0077
TarDAL	0.01276
U2Fusion	0.037195
Dif-Fusion DP	1.0230
Dif-Fusion MCFM	0.0001425

TABLE IX
TIME COST FOR DIFFUSION PROCESS AND MULTI-CHANNEL FUSION MODULE FOR DIFFERENT IMAGE SIZES

Image Size (W×H)	DP (s)	MCFM (s)
160 × 160	0.1279	0.0001374
320 × 320	0.3762	0.0001381
480 × 480	0.7920	0.0001387
720 × 720	1.7719	0.0001481

5) Inference Time Comparisons: As all methods were tested on GPUs, we report the average inference time for each method on the MSRS dataset on the NVIDIA RTX3090 GPU, and the results are presented in Table VIII. It should be noted that the time cost of our method consists of two components: the Diffusion Process (DP) and the Multi-Channel Fusion Module (MCFM). In Table IX, we provide the individual cost times for two components as the image size varies from 160 × 160 to 720 × 720. It can be observed that our fusion head is quite lightweight. However, the DP consumes significant time, and this time cost increases rapidly as the image size grows. We believe that the model size and the speed of these steps could potentially be reduced further with the advancement of the diffusion models [86], [87]. Efficient diffusion has become a crucial direction in diffusion researches.

F. Limitations

The newly proposed Dif-Fusion is based on Diffusion, which falls within the category of generative model-based image fusion frameworks, achieving state-of-the-art performance. However, there are still several directions that need to be further explored in the future.

First, compared to other generative model-based or CNN-based fusion models in image fusion area, the entire process of Dif-Fusion is computationally expensive, with over 99% of the time spent on the Diffusion process. In recent years, there has been increasing attention towards accelerating the diffusion process, indicating potential for significant improvements in

the efficiency of Diffusion-based algorithms. Second, in cross-dataset generalization tests, we observed that the performance of the Diffusion-based fusion model may be influenced by the data distribution of the training datasets. If there are substantial differences between datasets (e.g., saturation, illumination, intensity, etc.), the fusion performance may degrade. Therefore, there is a need for more exploration into the relationship between Diffusion-based fusion models and the diversity of multimodal data. Third, the Diffusion-based fusion model involves the selection and optimization of certain parameters, such as combinations of time-steps, which also requires further study.

The primary focus of this study is pioneering the integration of the Diffusion model into the infrared and visible image fusion, proposing methods such as the diffusion feature fusion module and multi-channel losses. The image fusion study will be inspired by Diffusion variants and their applications to diverse vision tasks, and new fusion approaches or strategies based on Diffusion might be proposed and show even greater potential in the future.

V. Conclusion

In this paper, an infrared and visible image fusion method based on diffusion models is proposed to achieve multi-channel complementary information extraction and effective maintenance of color and visual quality. On the one hand, we construct the distribution of multi-channel input data in the latent space with forward and reverse diffusion process. By training a denoising network in the reverse process to predict the Gaussian noise added in the forward process, the distribution of multi-channel data is built. On the other hand, we propose a method to generate three-channel images directly. In order to preserve the gradient and intensity of the three-channel image directly, we propose multi-channel gradient and intensity losses. Moreover, in terms of fused image evaluation, in addition to the existing texture and intensity fidelity metrics, we introduce Delta E to quantify color fidelity. Extensive experiments show that Dif-Fusion is superior to existing state-of-the-art methods.

Overall, we investigate a framework for extracting multi-channel complementary information based on diffusion models, and try to directly generate chromatic fusion images from multi-modal input. In the future, we might explore more multi-channel information learning models and end-to-end chromatic fusion image generation methods.

References

[1] 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.

[2] H. Zhang, H. Xu, X. Tian, J. Jiang, and J. Ma, "Image fusion meets deep learning: A survey and perspective," *Inf. Fusion*, vol. 76, pp. 323–336, Dec. 2021.

[3] J. Ma, Y. Ma, and C. Li, "Infrared and visible image fusion methods and applications: A survey," *Inf. Fusion*, vol. 45, pp. 153–178, Jan. 2019.

[4] J. Ma, L. Tang, M. Xu, H. Zhang, and G. Xiao, "STDFusionNet: An infrared and visible image fusion network based on salient target detection," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021.

[5] W. Zhou, J. Liu, J. Lei, L. Yu, and J.-N. Hwang, "GMNet: Graded-feature multilabel-learning network for RGB-thermal urban scene semantic segmentation," *IEEE Trans. Image Process.*, vol. 30, pp. 7790–7802, 2021.

[6] Y. Lu et al., "Cross-modality person re-identification with shared-specific feature transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13376–13386.

[7] C. Li, C. Zhu, Y. Huang, J. Tang, and L. Wang, "Cross-modal ranking with soft consistency and noisy labels for robust RGB-T tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 808–823.

[8] Z. Zhou, B. Wang, S. Li, and M. Dong, "Perceptual fusion of infrared and visible images through a hybrid multi-scale decomposition with Gaussian and bilateral filters," *Inf. Fusion*, vol. 30, pp. 15–26, Jul. 2016.

[9] 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, Aug. 2019.

[10] S. Li, H. Yin, and L. Fang, "Remote sensing image fusion via sparse representations over learned dictionaries," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 9, pp. 4779–4789, Sep. 2013.

[11] Y. Liu, X. Chen, R. K. Ward, and Z. Jane Wang, "Image fusion with convolutional sparse representation," *IEEE Signal Process. Lett.*, vol. 23, no. 12, pp. 1882–1886, Dec. 2016.

[12] Y. Liu, S. Liu, and Z. Wang, "A general framework for image fusion based on multi-scale transform and sparse representation," *Inf. Fusion*, vol. 24, pp. 147–164, Jul. 2015.

[13] S. Li, X. Kang, and J. Hu, "Image fusion with guided filtering," *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2864–2875, Jul. 2013.

[14] W. Kong, Y. Lei, and H. Zhao, "Adaptive fusion method of visible light and infrared images based on non-subsampled shearlet transform and fast non-negative matrix factorization," *Infr. Phys. Technol.*, vol. 67, pp. 161–172, Nov. 2014.

[15] D. P. Bavirisetti and R. Dhuli, "Two-scale image fusion of visible and infrared images using saliency detection," *Infr. Phys. Technol.*, vol. 76, pp. 52–64, May 2016.

[16] J. Ma, Z. Zhou, B. Wang, and H. Zong, "Infrared and visible image fusion based on visual saliency map and weighted least square optimization," *Infr. Phys. Technol.*, vol. 82, pp. 8–17, May 2017.

[17] S. Li, X. Kang, L. Fang, J. Hu, and H. Yin, "Pixel-level image fusion: A survey of the state of the art," *Inf. Fusion*, vol. 33, pp. 100–112, Jan. 2017.

[18] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Deep learning for hyperspectral image classification: An overview," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6690–6709, Sep. 2019.

[19] L. Wu, L. Fang, X. He, M. He, J. Ma, and Z. Zhong, "Querying labeled for unlabeled: Cross-image semantic consistency guided semi-supervised semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 7, pp. 8827–8844, Jul. 2022.

[20] 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.

[21] 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, Jul. 2022.

[22] J. Ma, H. Zhang, Z. Shao, P. Liang, and H. Xu, "GANMcC: A generative adversarial network with multiclassification constraints for infrared and visible image fusion," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–14, 2021.

[23] F. A. A. Kingdom, "Color brings relief to human vision," *Nature Neurosci.*, vol. 6, no. 6, pp. 641–644, Jun. 2003.

[24] M. S. Castelhano and J. M. Henderson, "The influence of color on the perception of scene gist," *J. Express Psychol., Hum. Percept. Perform.*, vol. 34, no. 3, p. 660, 2008.

[25] A. J. Elliot and M. A. Maier, "Color psychology: Effects of perceiving color on psychological functioning in humans," *Annu. Rev. Psychol.*, vol. 65, no. 1, pp. 95–120, Jan. 2014.

[26] R. Mehta and R. Zhu, "Blue or red? Exploring the effect of color on cognitive task performances," *Science*, vol. 323, no. 5918, pp. 1226–1229, Feb. 2009.

[27] R. A. Hill and R. A. Barton, "Red enhances human performance in contests," *Nature*, vol. 435, no. 7040, p. 293, May 2005.

[28] A. J. Elliot and M. A. Maier, "Color and psychological functioning," *Current Directions Psychol. Sci.*, vol. 16, no. 5, pp. 250–254, Oct. 2007.

[29] 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, Oct. 2021.

[30] 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/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5792–5801.

[31] H. Zhou, W. Wu, Y. Zhang, J. Ma, and H. Ling, "Semantic-supervised infrared and visible image fusion via a dual-discriminator generative adversarial network," *IEEE Trans. Multimedia*, vol. 25, pp. 635–648, 2023.

[32] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. (NIPS)*, vol. 33, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds. Vancouver, BC, Canada: Curran Associates, 2020, pp. 6840–6851.

[33] P. Zhu, X. Ma, and Z. Huang, "Fusion of infrared-visible images using improved multi-scale top-hat transform and suitable fusion rules," *Infr. Phys. Technol.*, vol. 81, pp. 282–295, Mar. 2017.

[34] H. Tang, G. Liu, L. Tang, D. P. Bavirisetti, and J. Wang, "MdedFusion: A multi-level detail enhancement decomposition method for infrared and visible image fusion," *Infr. Phys. Technol.*, vol. 127, Dec. 2022, Art. no. 104435.

[35] G. Li, Y. Lin, and X. Qu, "An infrared and visible image fusion method based on multi-scale transformation and norm optimization," *Inf. Fusion*, vol. 71, pp. 109–129, Jul. 2021.

[36] Z. Fu, X. Wang, J. Xu, N. Zhou, and Y. Zhao, "Infrared and visible images fusion based on RPCA and NSCT," *Infr. Phys. Technol.*, vol. 77, pp. 114–123, Jul. 2016.

[37] H. Li, L. Liu, W. Huang, and C. Yue, "An improved fusion algorithm for infrared and visible images based on multi-scale transform," *Infr. Phys. Technol.*, vol. 74, pp. 28–37, Jan. 2016.

[38] X. Zhang, X. Dai, X. Zhang, and G. Jin, "Joint principal component analysis and total variation for infrared and visible image fusion," *Infr. Phys. Technol.*, vol. 128, Jan. 2023, Art. no. 104523.

[39] N. Cvejic, D. Bull, and N. Canagarajah, "Region-based multimodal image fusion using ICA bases," *IEEE Sensors J.*, vol. 7, no. 5, pp. 743–751, May 2007.

[40] N. Mitianoudis and T. Stathaki, "Pixel-based and region-based image fusion schemes using ICA bases," *Inf. Fusion*, vol. 8, no. 2, pp. 131–142, Apr. 2007.

[41] J. Wang, J. Peng, X. Feng, G. He, and J. Fan, "Fusion method for infrared and visible images by using non-negative sparse representation," *Infr. Phys. Technol.*, vol. 67, pp. 477–489, Nov. 2014.

[42] S. Wang, J. Yue, J. Liu, Q. Tian, and M. Wang, "Large-scale few-shot learning via multi-modal knowledge discovery," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2020, pp. 718–734.

[43] G. Cui, H. Feng, Z. Xu, Q. Li, and Y. Chen, "Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition," *Opt. Commun.*, vol. 341, pp. 199–209, Apr. 2015.

[44] J. Han, E. J. Pauwels, and P. de Zeeuw, "Fast saliency-aware multi-modality image fusion," *Neurocomputing*, vol. 111, pp. 70–80, Jul. 2013.

[45] B. Zhang, X. Lu, H. Pei, and Y. Zhao, "A fusion algorithm for infrared and visible images based on saliency analysis and non-subsampled shearlet transform," *Infr. Phys. Technol.*, vol. 73, pp. 286–297, Nov. 2015.

[46] C. H. Liu, Y. Qi, and W. R. Ding, "Infrared and visible image fusion method based on saliency detection in sparse domain," *Infr. Phys. Technol.*, vol. 83, pp. 94–102, Jun. 2017.

[47] H. Yin, "Sparse representation with learned multiscale dictionary for image fusion," *Neurocomputing*, vol. 148, pp. 600–610, Jan. 2015.

[48] J. Zhao, Q. Zhou, Y. Chen, H. Feng, Z. Xu, and Q. Li, "Fusion of visible and infrared images using saliency analysis and detail preserving based image decomposition," *Infr. Phys. Technol.*, vol. 56, pp. 93–99, Jan. 2013.

[49] 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. Autom. Sinica*, vol. 9, no. 12, pp. 2121–2137, Dec. 2022.

[50] 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. Autom. Sinica*, vol. 9, no. 7, pp. 1200–1217, Jul. 2022.

[51] H. Li, X.-J. Wu, and T. Durrani, "NestFuse: An infrared and visible image fusion architecture based on nest connection and spatial/channel attention models," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 12, pp. 9645–9656, Dec. 2020.

[52] L. Jian, X. Yang, Z. Liu, G. Jeon, M. Gao, and D. Chisholm, "SEDR-Fuse: A symmetric encoder–decoder with residual block network for infrared and visible image fusion," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–15, 2021.

[53] 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.*, Apr. 2020, vol. 34, no. 7, pp. 12797–12804.

[54] J. Ma, H. Xu, J. Jiang, X. Mei, and X.-P. Zhang, "DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 4980–4995, 2020.

[55] 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, Jun. 2022.

[56] 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, Mar. 2023.

[57] Z. Wang, Y. Chen, W. Shao, H. Li, and L. Zhang, "SwinFuse: A residual Swin transformer fusion network for infrared and visible images," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–12, 2022.

[58] J. Li, H. Huo, C. Li, R. Wang, and Q. Feng, "AttentionFGAN: Infrared and visible image fusion using attention-based generative adversarial networks," *IEEE Trans. Multimedia*, vol. 23, pp. 1383–1396, 2021.

[59] L. Yang et al., "Diffusion models: A comprehensive survey of methods and applications," 2022, *arXiv:2209.00796*.

[60] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8162–8171.

[61] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021, pp. 1–22.

[62] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 2256–2265.

[63] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020, pp. 1–36.

[64] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," in *Proc. Adv. Neural Inf. Process. (NIPS)*, vol. 34, 2021, pp. 8780–8794.

[65] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool, "RePaint: Inpainting using denoising diffusion probabilistic models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11451–11461.

[66] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4713–4726, Apr. 2023.

[67] M. Daniels, T. Maunu, and P. Hand, "Score-based generative neural networks for large-scale optimal transport," in *Proc. Adv. Neural Inf. Process. (NIPS)*, vol. 34, 2021, pp. 12955–12965.

[68] H. Chung, B. Sim, and J. C. Ye, "Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12403–12412.

[69] C. Saharia et al., "Palette: Image-to-image diffusion models," in *Proc. ACM SIGGRAPH*, 2022, pp. 1–10.

[70] M. Zhao, F. Bao, C. Li, and J. Zhu, "EGSDE: Unpaired image-to-image translation via energy-guided stochastic differential equations," in *Proc. Adv. Neural Inf. Process. (NIPS)*, A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, Eds. 2022, pp. 3609–3623.

[71] R. S. Zimmermann, L. Schott, Y. Song, B. A. Dunn, and D. A. Klindt, "Score-based generative classifiers," 2021, *arXiv:2110.00473*.

[72] D. Baranchuk, A. Voynov, I. Rubachev, V. Khrulkov, and A. Babenko, "Label-efficient semantic segmentation with diffusion models," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021, pp. 1–15.

[73] T. Amit, T. Shaharbany, E. Nachmani, and L. Wolf, "SegDiff: Image segmentation with diffusion probabilistic models," 2021, *arXiv:2112.00390*.

[74] S. Chen, P. Sun, Y. Song, and P. Luo, "DiffusionDet: Diffusion model for object detection," 2022, *arXiv:2211.09788*.

[75] F.-A. Croitoru, V. Hondru, R. Tudor Ionescu, and M. Shah, "Diffusion models in vision: A survey," 2022, *arXiv:2209.04747*.

[76] Y. Song and S. Ermon, "Improved techniques for training score-based generative models," in *Proc. Adv. Neural Inf. Process. (NIPS)*, vol. 33, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds. Vancouver, BC, Canada: Curran Associates, 2020, pp. 12438–12448.

[77] S. Gu et al., "Vector quantized diffusion model for text-to-image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10686–10696.
[78] W. G. C. Bandara, N. G. Nair, and V. M. Patel, "DDPM-CD: Remote sensing change detection using denoising diffusion probabilistic models," 2022, *arXiv:2206.11892*.
[79] J. Ma, C. Chen, C. Li, and J. Huang, "Infrared and visible image fusion via gradient transfer and total variation minimization," *Inf. Fusion*, vol. 31, pp. 100–109, Sep. 2016.
[80] G. Qu, D. Zhang, and P. Yan, "Information measure for performance of image fusion," *Electron. Lett.*, vol. 38, no. 7, pp. 313–315, 2002.
[81] Y. Han, Y. Cai, Y. Cao, and X. Xu, "A new image fusion performance metric based on visual information fidelity," *Inf. Fusion*, vol. 14, no. 2, pp. 127–135, Apr. 2013.
[82] A. M. Eskicioglu and P. S. Fisher, "Image quality measures and their performance," *IEEE Trans. Commun.*, vol. 43, no. 12, pp. 2959–2965, 1995.
[83] C. S. Xydeas and V. Petrovic, "Objective image fusion performance measure," *Electron. Lett.*, vol. 36, no. 4, pp. 308–309, 2000.
[84] G. Sharma, W. Wu, and E. N. Dalal, "The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations," *Color Res. Appl.*, vol. 30, no. 1, pp. 21–30, Feb. 2005.
[85] W. G. Backhaus, R. Kliegl, and J. S. Werner, *Color Vision: Perspectives From Different Disciplines*. Berlin, Germany: Walter de Gruyter, 2011.
[86] J. Wu, D. Zhu, L. Fang, Y. Deng, and Z. Zhong, "Efficient layer compression without pruning," *IEEE Trans. Image Process.*, vol. 32, pp. 4689–4700, 2023.
[87] Q. Zhang and Y. Chen, "Fast sampling of diffusion models with exponential integrator," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2023, pp. 1–33.

Jun Yue received the B.Eng. degree in geodesy from Wuhan University, Wuhan, China, in 2013, and the Ph.D. degree in GIS from Peking University, Beijing, China, in 2018.

He is currently an Assistant Professor with the School of Automation, Central South University. His research interests include satellite image understanding, pattern recognition, and few-shot learning. He serves as a reviewer for IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, *ISPRS Journal of Photogrammetry and Remote Sensing*, IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING, *Information Fusion*, and *Information Sciences*.

Leyuan Fang (Senior Member, IEEE) received the Ph.D. degree from the College of Electrical and Information Engineering, Hunan University, Changsha, China, in 2015.

From August 2016 to September 2017, he was a Postdoctoral Researcher with the Department of Biomedical Engineering, Duke University, Durham, NC, USA. He is currently a Professor with the College of Electrical and Information Engineering, Hunan University. His research interests include sparse representation and multi-resolution analysis in remote sensing and medical image processing. He was a recipient of the one Second-Grade National Award from the Nature and Science Progress of China in 2019. He is an Associate Editor of IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and *Neurocomputing*.

Shaobo Xia received the bachelor's degree in geodesy and geomatics from the School of Geodesy and Geomatics, Wuhan University, Wuhan, China, in 2013, the master's degree in cartography and geographic information systems from the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing, China, in 2016, and the Ph.D. degree in geomatics from the University of Calgary, Calgary, AB, Canada, in 2020.

He is currently an Assistant Professor with the Department of Geomatics Engineering, Changsha University of Science and Technology, Changsha, China. His research interests include point cloud processing, image processing, and remote sensing.

Yue Deng (Senior Member, IEEE) received the Ph.D. degree (Hons.) in control science and engineering from the Department of Automation, Tsinghua University, Beijing, China, in 2013.

He is currently a Professor with the School of Astronautics, Beihang University, Beijing. His research interests include machine learning, signal processing, and computational biology.

Jiayi Ma (Senior Member, IEEE) received the B.S. degree in information and computing science and the Ph.D. 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 the School of Electronic Information, Wuhan University, Wuhan. He has authored or coauthored more than 300 refereed journals and conference papers, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, *IJCV*, CVPR, ICCV, and ECCV. His research interests include computer vision, machine learning, and robotics. He is an Area Editor of *Information Fusion* and an Associate Editor of *Neurocomputing*. He has been identified in the 2019–2022 Highly Cited Researcher lists from the Web of Science Group.