

Diffusion Models Meet Remote Sensing: Principles, Methods, and Perspectives

Yidan Liu, Jun Yue, Shaobo Xia, Pedram Ghamisi, *Senior Member, IEEE*, Weiyang Xie, *Senior Member, IEEE*, and Leyuan Fang, *Senior Member, IEEE*

Abstract—As a newly emerging advance in deep generative models, diffusion models have achieved state-of-the-art results in many fields, including computer vision, natural language processing, and molecule design. The remote sensing community has also noticed the powerful ability of diffusion models and quickly applied them to a variety of tasks for image processing. Given the rapid increase in research on diffusion models in the field of remote sensing, it is necessary to conduct a comprehensive review of existing diffusion model-based remote sensing papers, to help researchers recognize the potential of diffusion models and provide some directions for further exploration. Specifically, this paper first introduces the theoretical background of diffusion models, and then systematically reviews the applications of diffusion models in remote sensing, including image generation, enhancement, and interpretation. Finally, the limitations of existing remote sensing diffusion models and worthy research directions for further exploration are discussed and summarized.

Index Terms—Diffusion Models, Remote Sensing, Generative Models, Deep Learning.

I. INTRODUCTION

REMOTE sensing (RS), as an advanced earth observation technology, has been widely used in civilian and military fields such as environmental monitoring, urban planning, disaster response, and camouflage detection [1]–[5]. Following the boom of artificial intelligence, employing deep learning models to interpret RS images has become a large-scale solution for these applications [6]. Early intelligent RS interpretation methods primarily relied on supervised deep neural networks, which were trained with massive data and high-quality annotations. However, the scarcity of annotations and the high acquisition costs of RS images have hindered further advancements in these methods.

This work was supported in part by the National Natural Science Foundation of China under Grant U22B2014 and Grant 62101072 and in part by the Science and Technology Plan Project Fund of Hunan Province under Grant 2022RSC3064. (Yidan Liu and Jun Yue contributed equally to this work.) (Corresponding author: Weiyang Xie; Leyuan Fang.)

Yidan Liu and Leyuan Fang are with the College of Electrical and Information Engineering, Hunan University, Changsha 410082, China (e-mail: liuyidan_bu@163.com; fangleyuan@gmail.com).

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

Shaobo Xia is with the Department of Geomatics Engineering, Changsha University of Science and Technology, Changsha 410114, China (e-mail: shaobo.xia@csust.edu.cn).

Pedram Ghamisi is with the Helmholtz-Zentrum Dresden-Rossendorf (HZDR), Helmholtz Institute Freiberg for Resource Technology, 09599 Freiberg, Germany, and also with Lancaster University, LA1 4YR Lancaster, U.K. (e-mail: p.ghamisi@gmail.com).

Weiyang Xie is with the State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China (e-mail: wxyxie@xidian.edu.cn).

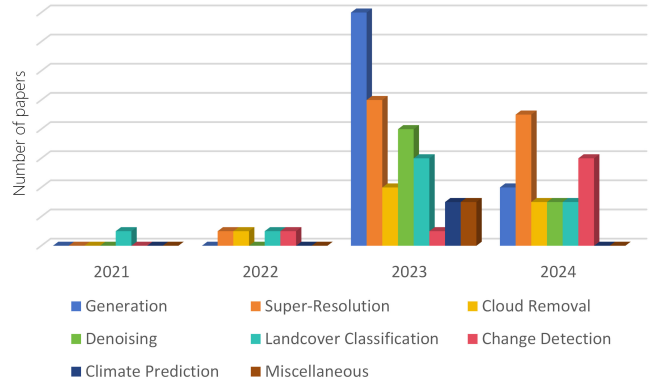


Fig. 1. Development of diffusion models in RS. Statistical data as of the first quarter of 2024.

The advent of deep generative models effectively solves the problems in the above supervised interpretation methods, bringing new opportunities for the intelligent processing of RS images. Specifically, deep generative models are capable of learning the data distribution from limited RS images to generate new data samples. At the same time, they can generate annotation information directly from low-quality or unlabeled RS images by learning the mapping relationships between images, reducing the need for high-quality manual annotations. Furthermore, they have also demonstrated excellent capabilities in learning representations for image details and complex scenes. Over the past decade, numerous works on deep generative models, such as Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and Normalizing Flows (NFs), have sprung up to tackle the challenges associated with RS images [7]–[11]. Despite the flourishing development of these generative models, each comes with its limitations. For example, VAE [12] requires a trade-off between reconstruction loss (similarity between the output and the input) and latent loss (proximity of the hidden nodes to the normal distribution), so that the generated images are often blurry. The structure of NF-based models [13] needs to comply with the calculation of probability density, resulting in limited scalability and flexibility. While GAN [14] has many variants and generates high-quality RS images, its training process is unstable and easy to collapse (i.e., the generated sample pattern is singular and cannot cover diverse patterns).

In this context, diffusion models [15], as a newly emerging type of deep generative models, have brought about a rev-

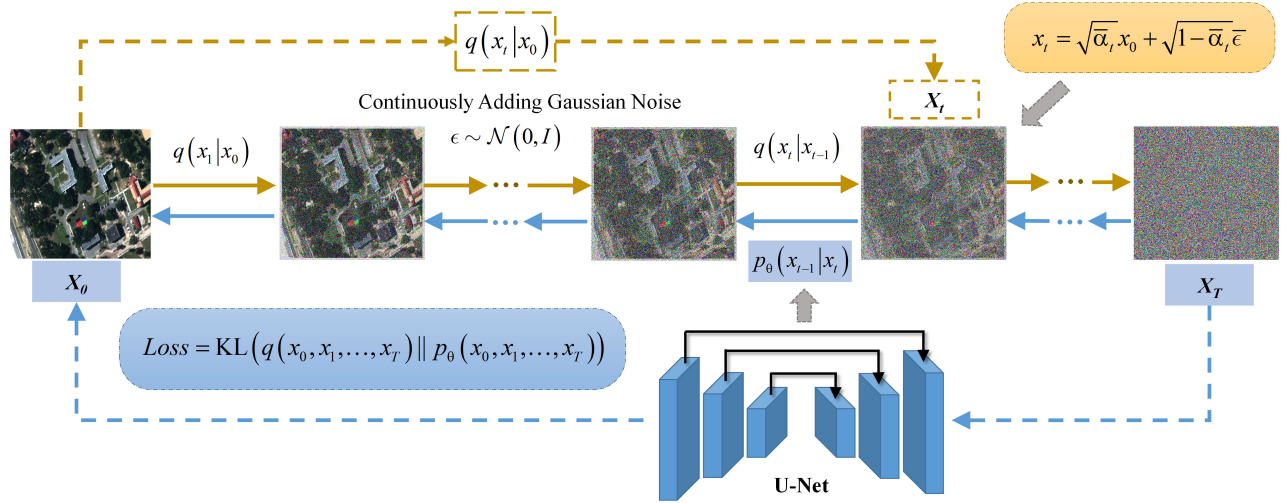


Fig. 2. The training procedure of denoising diffusion probabilistic model (DDPM), where yellow lines represent the forward diffusion process, and blue lines represent the backward diffusion process.

olutionary advancement in artificial intelligence. By modeling the inverse process of transforming regular images into random noise, diffusion models have demonstrated unprecedented performance in 2D and 3D image generation [16]–[20], image editing [21]–[24], image translation [25]–[28], and other computer vision tasks [29]–[32]. Moreover, they have achieved state-of-the-art results in many other fields, including natural language processing [33]–[35], audio synthesis [36]–[38], and molecular design [39]–[41], challenging the long-standing dominance of GANs.

Given these remarkable achievements, the RS community has also quickly applied diffusion models to a variety of tasks for image processing. Since 2021, the application of diffusion models in RS has shown a rapid development trend of expanding scope and increasing quantity (see Fig. 1). In fact, diffusion models have significant advantages over other deep generative models in processing and analyzing RS images.

- Firstly, due to the atmospheric interference and limitations of imaging equipment, RS images often contain noise. The inherent denoising ability of diffusion models can rightly eliminate these negative effects.
- Secondly, RS images are highly diverse due to differences in collection time, equipment, and environment. The architecture of diffusion models is flexible, allowing the introduction of conditional constraints to cope with various changes.
- Thirdly, RS images always contain diverse and complex scenes. The precise mathematical derivation and progressive learning process of diffusion models offer advantageous in learning such complex data distribution.
- Additionally, diffusion models can provide more stable training than GANs, which is suitable for training large-scale RS datasets.

In summary, diffusion models possess great development potential in the field of RS. Therefore, it is necessary to review and summarize existing diffusion model-based RS

papers to help researchers gain a comprehensive understanding of the current research status, and identify the gaps in the application of diffusion models in RS, thereby promoting further development in this field.

The remainder of this paper is organized as follows. **Section II** introduces the theoretical background of diffusion models. **Section III** reviews the application of diffusion models across various RS image processing tasks, and demonstrates the superiority of diffusion models through a series of visual experimental results and quantitative metrics. **Section IV** discusses the limitations of the existing RS diffusion models and reveals possible research directions in the future. Finally, conclusions are drawn in **Section V**.

II. THEORETICAL BACKGROUND OF DIFFUSION MODELS

Diffusion models, also known as diffusion probabilistic models, constitute a family of deep generative models. In essence, a generative model aims to transform a random distribution (i.e., noise) into a ‘probability distribution’ that matches the distribution of the observed dataset, thereby producing desired outcomes through sampling from this ‘probability distribution’.

Obtaining the target distribution directly is evidently challenging. However, disrupting a regular distribution into random noise is comparatively straightforward and can be accomplished by continually adding Gaussian noise, as depicted in Fig. 2. The concept of diffusion models is inspired by this process and involves learning the reverse denoising process, which dates back to 2015 [42] and gained popularity following the publication of the denoising diffusion probabilistic model (DDPM) in 2020 [15].

A. Denoising Diffusion Probabilistic Model (DDPM)

As shown in Fig. 2, the training procedure involves two phases: the forward diffusion process and the backward diffusion process.

Forward Diffusion Process: Given the original image x_0 , this process generates the noise-contaminated images x_1, x_2, \dots, x_T through T iterations of noise addition. where the image x_t obtained at each step is only related to x_{t-1} . Thus, this process can be represented by a Markov chain:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t I) \quad (1)$$

$$q(x_{1:T}|x_0) = \prod_{t=1}^T q(x_t|x_{t-1}) = \prod_{t=1}^T \mathcal{N}(x_t; \sqrt{1 - \beta_{t-1}}x_0, \beta_t I) \quad (2)$$

where $q(x_t|x_{t-1})$ is the transition probability of the Markov chain, which represents the distribution of Gaussian noise added at each step. β_t is a hyperparameter for the variance of the Gaussian distribution, linearly increasing with t . I denotes the identity matrix with the same dimensions as the input image x_0 .

An important property of the forward process is that it allows to directly obtain any noised image x_t from the original image x_0 and β_t , which is achieved through the reparameterization technique [12]. Specifically, with the notation of $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$, Eq. (1) could be expanded as

$$\begin{aligned} x_t &= \sqrt{\alpha_t}x_{t-1} + \sqrt{1 - \alpha_t}\epsilon_{t-1} \\ &= \sqrt{\alpha_t} \left(\sqrt{\alpha_{t-1}}x_{t-2} + \sqrt{1 - \alpha_{t-1}}\epsilon_{t-2} \right) + \sqrt{1 - \alpha_t}\epsilon_{t-1} \\ &= \sqrt{\alpha_t\alpha_{t-1}}x_{t-2} + \sqrt{1 - \alpha_{t-1}\alpha_t}\bar{\epsilon}_{t-2} \\ &= \dots \\ &= \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\bar{\epsilon} \end{aligned} \quad (3)$$

where $\epsilon_t, \epsilon_{t-1} \sim \mathcal{N}(0, I)$ and $\bar{\epsilon}_{t-2}$ is their merged result. According to the additivity of the independent Gaussian distribution, i.e., $\mathcal{N}(0, \sigma_1^2 I) + \mathcal{N}(0, \sigma_2^2 I) \sim \mathcal{N}(0, (\sigma_1^2 + \sigma_2^2)I)$, the third line of Eq. (3) conforms to Gaussian distribution, which means that the final derivation result also conforms to Gaussian distribution. Therefore, any noised image x_t satisfies:

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

In this way, when $T \rightarrow \infty$, x_T can converge to the standard normal distribution $\mathcal{N}(0, I)$, consistent with the original design intention.

Backward Diffusion Process: This process aims to obtain the reversed transition probability $q(x_{t-1}|x_t)$, thereby gradually restoring the image \hat{x}_0 from the noise. However, $q(x_{t-1}|x_t)$ is difficult to solve explicitly, so a neural network is employed to learn this distribution.

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \quad (5)$$

where θ represents the parameters of the neural network to be optimized, and the network is typically based on a U-Net architecture [43]. Accordingly, the backward diffusion process can be expressed as

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^T p_\theta(x_{t-1}|x_t) \quad (6)$$

The training goal of the network is to match the backward diffusion process $p_\theta(x_0, x_1, \dots, x_T)$ with the forward diffusion process $p_\theta(x_0, x_1, \dots, x_T)$, which can be achieved by minimizing the Kullback-Leibler (KL) divergence:

$$\begin{aligned} \mathcal{L}(\theta) &= \text{KL}(q(x_0, x_1, \dots, x_T) || p_\theta(x_0, x_1, \dots, x_T)) \\ &= -\mathbb{E}_{q(x_{0:T})} [\log p_\theta(x_0, x_1, \dots, x_T)] + \text{const} \\ &= -\mathbb{E}_{q(x_{0:T})} [-\log p(x_T) - \sum_{t=1}^T \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})}] + \text{const} \end{aligned} \quad (7)$$

Here, *const* denotes a constant independent of θ , and the first term of Eq. (7) represents the variational lower-bound of the negative log-likelihood, similar to VAE.

Notably, when the prior x_0 is introduced in $q(x_{t-1}|x_t)$, it can be converted by Bayes Rule

$$\begin{aligned} q(x_{t-1}|x_t, x_0) &= \frac{q(x_t, x_0, x_{t-1})}{q(x_t, x_0)} \\ &= \frac{q(x_0)q(x_{t-1}|x_0)q(x_t|x_{t-1}, x_0)}{q(x_0)q(x_t|x_0)} \\ &= q(x_t|x_{t-1}, x_0) \frac{q(x_{t-1}|x_0)}{q(x_t|x_0)} \end{aligned} \quad (8)$$

where $q(x_t|x_{t-1}, x_0)$ is defined in Eq. (1), $q(x_{t-1}|x_0)$ and $q(x_t|x_0)$ can be obtained by Eq. (4). After simplification, Eq. (8) can be rewritten as

$$q(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}; \tilde{\mu}_t(x_t), \tilde{\beta}_t I) \quad (9)$$

$$\tilde{\mu}_t(x_t) = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1 - \alpha_t}{\sqrt{1 - \alpha_t}} \bar{\epsilon} \right) \quad (10)$$

$$\hat{\beta}_t = \frac{1 - \alpha_{t-1}}{1 - \alpha_t} \beta_t \quad (11)$$

α_t and β_t are both constants in the above equation, only $\bar{\epsilon}$ in Eq. (10) can be parameterized by the neural network as

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right) \quad (12)$$

In other words, the constructed neural network is to learn the noise $\epsilon_\theta(x_t, t)$, which is reasonable since the process from x_t to x_{t-1} is essentially a denoising process.

According to [15], the optimization goal of the network can be further simplified with the help of Eq. (12) and (3) as the following form:

$$\mathcal{L}_{\text{simple}}(\theta) = \mathbb{E}_{x_0, \epsilon, t} \left[\left\| \epsilon - \epsilon_\theta \left(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, t \right) \right\|^2 \right] \quad (13)$$

which intuitively shows that the core of the diffusion model is to minimize the distance between the predicted noise ϵ_θ and the actual noise ϵ .

Sampling Process: In the inference process, which is also known as the sampling process, a new image y_0 can be generated from either Gaussian noise or a noisy image y_t by iteratively sampling y_{t-1} until $t = 1$ according to the following expanded Eq. (9)

$$y_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(y_t - \frac{1 - \alpha_t}{\sqrt{1 - \alpha_t}} \epsilon_\theta(y_t, t) \right) + \hat{\beta}_t z \quad (14)$$

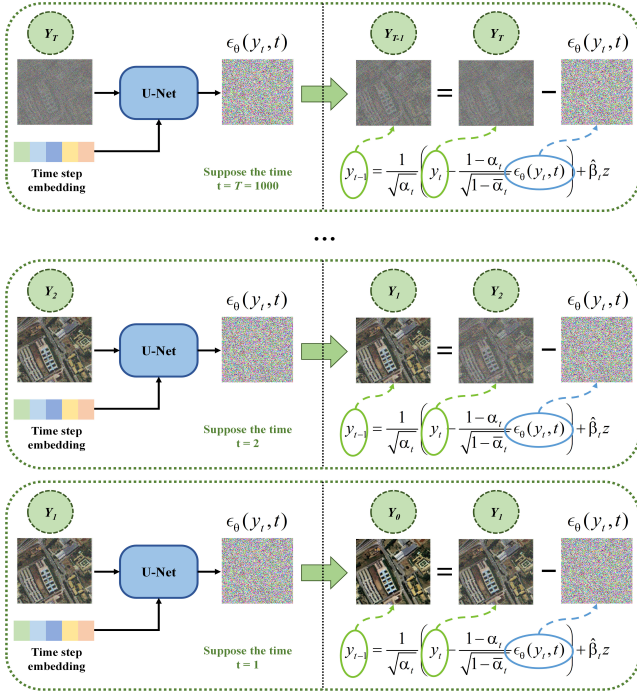


Fig. 3. The sampling process of DDPM. Supposing that sampling begins at $T=1000$, the noise distribution $\epsilon_\theta(y_t, t)$ is obtained from the well-trained diffusion model. Then, the noised image Y_t is used to subtract the noise $\epsilon_\theta(y_t, t)$, resulting in a denoised image Y_{t-1} . This denoised image Y_{t-1} is then input into the diffusion model to obtain the noise image for the next timestep. This process is repeated until $t = 1$, at which point the denoised image is quite clear.

where $z \sim \mathcal{N}(0, I)$, and $\hat{\beta}_t$ is usually approximated as β_t in practical [15]. Such a sampling process is illustrated in Fig. 3.

B. Conditional Diffusion Model

Similar to the development of GAN, the diffusion model was first proposed with unconditional generation, and then followed with the conditional generation [16]. Unconditional generation is often used to explore the upper limits of model capabilities, while conditional generation is more conducive to applications since it allows for the output to be controlled based on human wishes.

The first work to introduce the condition in a diffusion model is [44], which guides the generation of the diffusion model by adding a classifier to the well-trained diffusion model, so it is also called the *Guided Diffusion Model*. Although this method is less expensive to train, it also increases the inference cost by utilizing classification results to guide the sampling process of the diffusion model. More importantly, it has poor control over details and fails to produce satisfactory required images. As a result, the Google team [45] decided to adopt a straightforward idea to control the generated results by retraining the DDPM with conditions, and named it as *Classifier-Free Guidance*.

Formally, given conditional information c , the distribution

of DDPM that need to be learned is changed as

$$p_\theta(x_{t-1}|x_t, c) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, c, t), \tilde{\beta}_t I) \quad (15)$$

$$\mu_\theta(x_t, c, t) = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1 - \alpha_t}{\sqrt{1 - \alpha_t}} \epsilon_\theta(x_t, c, t) \right) \quad (16)$$

Correspondingly, the optimization objective (13) and sampling process (14) are modified into following forms:

$$\mathcal{L}_{con}(\theta) = \mathbb{E}_{x_0, \epsilon, c, t} \left[\left\| \epsilon - \epsilon_\theta(\sqrt{\alpha_t}x_0 + \sqrt{1 - \alpha_t}\epsilon, c, t) \right\|^2 \right] \quad (17)$$

$$y_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(y_t - \frac{1 - \alpha_t}{\sqrt{1 - \alpha_t}} \epsilon_\theta(y_t, c, t) \right) + \hat{\beta}_t z \quad (18)$$

Compared to the *Guided Diffusion Model*, this method is more widely used and is the basis of many attractive models (such as DALL-E2 [46], Imagen [47], Stable Diffusion [48] etc.), as well as the theoretical foundation of the conditional diffusion model in RS mentioned below.

III. APPLICATIONS OF DIFFUSION MODELS IN REMOTE SENSING

In this section, we will review and summarize existing related work, all of which involve the use of diffusion models in addressing RS image-related problems. To better organize our review, we categorize these papers according to their applications in RS and provide subdivisions for some common applications, as illustrated in Fig. 4. It is important to note that some applications may overlap with each other, but our categorization attempts to align with the core problems addressed by each paper.

A. RS Image Generation

As one of the most impressive deep generative models, diffusion models are expected to synthesize realistic RS images from existing images or given textual descriptions to support the development of various RS applications, as an alternative to autoregressive models [49]. According to the data sources, these image-generation methods can be mainly divided into two categories: text-to-image generation and image-to-image generation.

1) *Text-to-Image*: Over the past two years, numerous text-to-image diffusion models have come out in computer vision [16], [47], [48], especially the Stable Diffusion (SD) [48] model that has been widely adopted since its release [50]–[52]. However, its success mainly depends on training with billions of text-image pairs from the internet [53], which makes it difficult to extend to the field of RS since such vast and diverse RS datasets are not readily available, as mentioned in [49]. To address this problem, a straightforward idea is to produce trainable RS text-image pairs. Ou *et al.* [54] realized this idea with the help of pre-trained large models. Specifically, they first caption the existing RS images through a vision-language pre-training model to obtain initial textual prompts. Then, refining these prompts with human feedback and GPT-4 to improve semantic accuracy and suitability, successfully enabling the SD model to synthesize the required RS images. Instead of directly supplementing text prompts, Khanna *et al.*

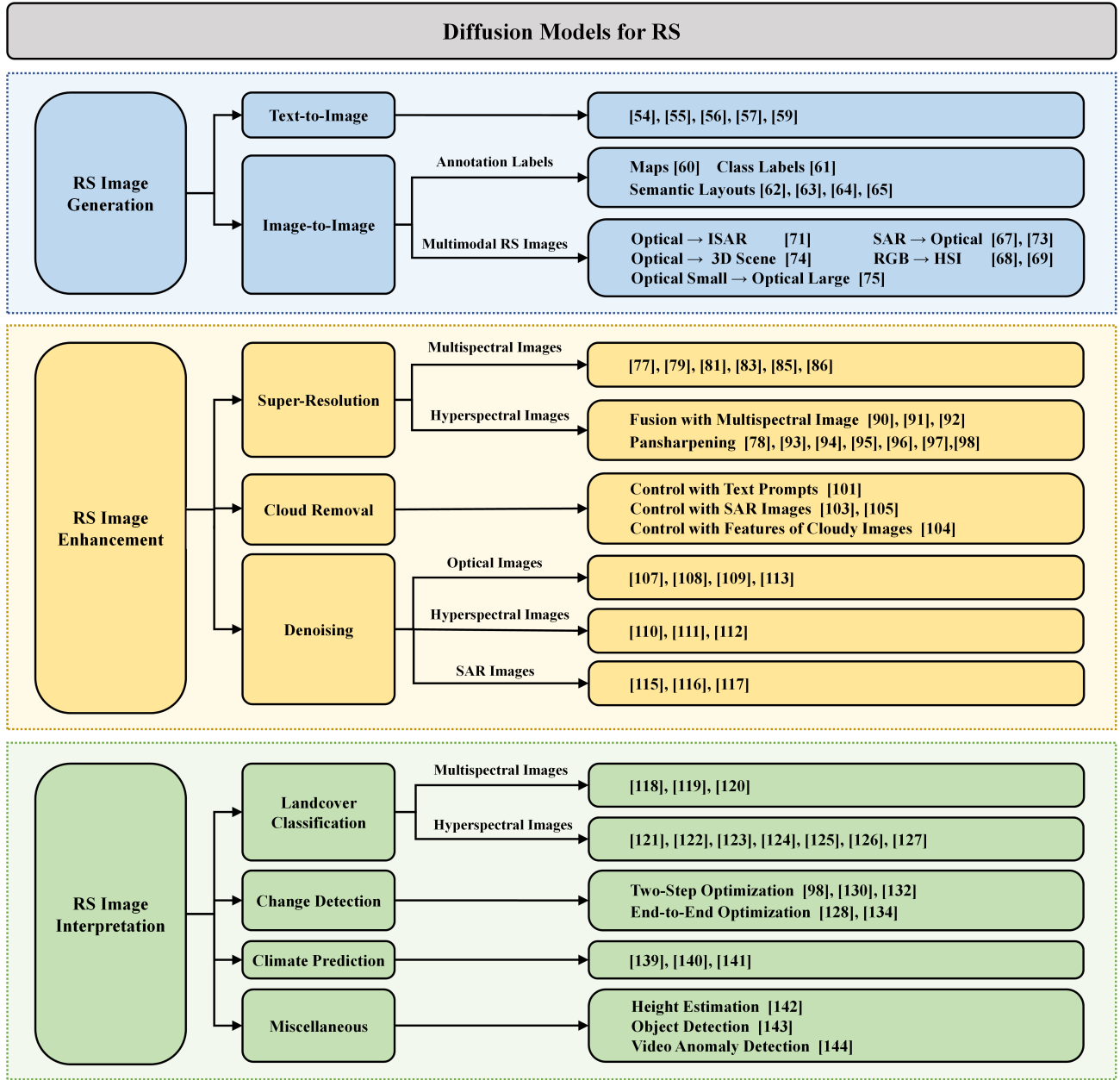


Fig. 4. The proposed taxonomy of diffusion model applications in RS.

[55] used various numerical information related to satellite images, such as geolocation and sampling time as new prompts of the SD model, which effectively enriches the SD model's input and enhances its ability to generate high-quality satellite images. Furthermore, Tang *et al.* [56] refined the generation process with SD model by incorporating RS image-related features as control conditions. They treated textual descriptions and numerical information as global control information, and used the depth map, segmentation mask, object boundaries, and other result images obtained through a series of pre-trained networks as local control information. By flexibly selecting control conditions, this approach achieves effective integration of multiple control information, expanding the RS image generation space.

Despite the significant progress made by the improved SD

model on optical RS images, researchers recently encountered new challenges when adapting it to other modal RS images [57]. For example, directly using SAR images to fine-tune the SD model will degrade the model's representational ability, resulting in the failure to generate satisfactory SAR images. This is because there are significant differences in capture perspectives and data modalities between SAR and natural images. In view of this, Tian *et al.* [57] proposed to fine-tune the SD model with optical RS images before using SAR images, so as to transition the model from regular view to the bird's-eye view. Meanwhile, they suggested training the SD model's Low-Rank Adaptation network [58], rather than the whole model, thus ensuring that semantic knowledge learned from natural images can be successfully transferred to the learning process of SAR images.

Apart from fine-tuning the SD model, researchers in the RS community have attempted to design a new architecture for RS text-to-image generation with diffusion model [59]. The proposed pipeline consists of two cascaded diffusion models, where the first one is designed to generate low-resolution satellite images from text prompts, and the second one is to increase the resolution of the generated images based on the text descriptions. The benefits of this two-stage generation approach are twofold. On the one hand, separating the generation of low and high-resolution images is advantageous for capturing scene information from global to local perspectives. On the other hand, the low-resolution image generation stage can alleviate the computational burden of generating high-resolution satellite images directly from text descriptions, which is more feasible for practical deployment.

2) *Image-to-Image*: Compared to the text-to-image generation, image-to-image generation is more popular in the field of RS. In this task, diffusion models are guided by existing images to generate new ones. The guiding images can take various forms. Some researchers prefer using masks, such as maps [60], class labels [61], and semantic layouts [62]–[65], as the guiding images for the diffusion models. Although these images only contain some specific information, they still present excellent performance in the RS image generation. For example, [60] successfully produces realistic satellite images by training the ControlNet model [66] with maps, even historical maps. [61] effectively addresses the issue of insufficient samples and unbalanced classes captured from actual battlefield environments by using class labels as conditional constraints. [63] achieves the generation of RS image-annotation pairs by implementing a two-phase training process on the SD model, addressing the expensive high quality annotation problem. One of the most noteworthy works is [65], which not only generates high-quality RS images with the guides of semantic masks, but also addresses the inherent problem of diffusion models requiring long training time for the model convergence. Specifically, this work introduces a lightweight diffusion model obtained through a customized distillation process, which ensures the quality of image generation via a multi-frequency extraction module and achieves rapid convergence by adjusting the image size at different stages of the diffusion process.

However, the masks used in these methods are essentially annotation labels that require expert knowledge and manual labeling, making them costly to acquire. Given the difficulty in acquiring these masks, some works have explored to use multi-modal RS images as the guide images of diffusion models [67]–[69]. Different modal RS images have their own strengths and weaknesses. For example, optical RS images are highly visualized and can intuitively reflect surface information, but they are limited by weather conditions and capture time, and are easily obscured by clouds. Conversely, SAR images can be captured in all weather conditions and penetrate clouds and fog, but their imaging process is complex and usually require experts to interpret. In view of these, Bai *et al.* [67] adopted SAR images as the guide images for a diffusion model to generate optical RS images, which gains higher clarity and better structural consistency than those generated

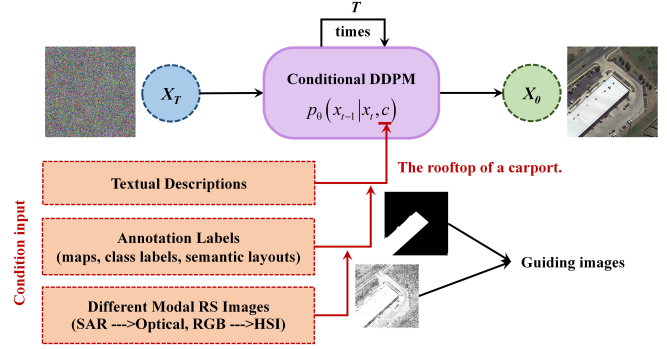


Fig. 5. Overview of different diffusion model-based methods for RS image generation. Note that also combinations of different condition inputs are possible.

by GAN models. Similarly, hyperspectral images (HSIs) can provide richer spectral-spatial information than multispectral images (e.g., RGB images), even though both belong to optical RS images. However, the acquisition cost of HSIs is much higher than that of multispectral images. Therefore, researchers would like to generate HSIs with the help of easy-obtainable multispectral RS images [68], [69]. Unfortunately, using diffusion model to generate HSIs requires matching the input noise dimensionality with the spectral bands of the HSI, resulting in an excessively large noise sampling space that hampers the model's convergence. To address this issue, Zhang *et al.* [68] proposed a spectral folding technology to convert the input HSI into a pseudo-color image before training the diffusion model. Liu *et al.* [69] used conditional vector quantized generative adversarial network (VQGAN) [70] to obtain a latent code space of HSIs, and performed the training and sampling processes of diffusion model within this space.

Overall, the above image-to-image generation methods are based on the conditional diffusion model [45], using the guide image as the condition input to the conditional DDPM, which essentially generates target RS images from noise rather than images directly. To realize the true image-to-image translation, Wang *et al.* [71] employed a straightforward idea: input Inverse Synthetic Aperture Radar (ISAR) images [72] to the diffusion model during the training phase to force the model to learn the distribution of ISAR images, and then input optical RS images in the testing phase to generate new ISAR images. Seo *et al.* [73] proposed to generate optical RS images from SAR images by sampling noise from the target images rather than using Gaussian noise, efficiently ensuring the consistency of the distribution between the generated image and the target image without using the conditional diffusion model. More recently, Li *et al.* [74] the diffusion model for generating 3D urban scenes from satellite images. They utilized a 3D diffusion model with sparse convolutions to generate texture and colors for the foreground point cloud, comprising buildings and roads, and employed a 2D diffusion model to synthesize the background sky, which enabled the direct generation of 3D scenes solely from satellite imagery, demonstrating a novel application of diffusion models in the RS community. Apart from generating images from one

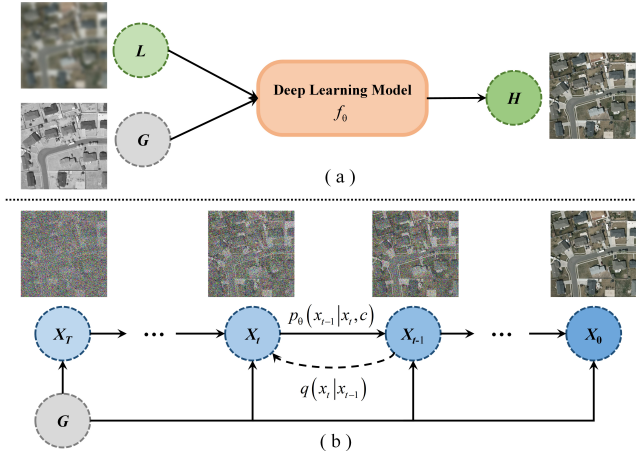


Fig. 6. Comparison of previous deep learning-based methods and diffusion model-based methods for the RSSR task. (a) The workflow of previous deep learning-based methods, where **L**, **G**, **H** indicate the LRRS image, guided image and HRRS image. (b) The workflow of diffusion model-based methods, where X_t represents the diffused HR image at t and **G** represents the condition input (such as the LR image and its features) for the diffusion model. The figure originally shown in [78]

modality to another, researchers have proposed that large satellite images can be generated from patch-level images within the same modality [75]. Specifically, they first adopted self-supervised learning to extract feature embeddings of the input patches, and then used these embeddings as condition input to guide the diffusion model's learning. Notably, they retained the spatial arrangement of each patch in the original image, making it possible to assemble the generated patches into a large and coherent image.

B. RS Image Enhancement

1) **Super-Resolution:** Remote sensing super-resolution (RSSR) aims to reconstruct high-resolution (HR) RS images with more details from low-resolution (LR) RS images [76], which are degraded by imaging equipment, weather condition, or downsampling. Compared to natural images, RS images suffer from much more details loss, making it more challenging to reconstruct the HR images [77]. Therefore, using a more powerful generative model, such as the diffusion model, to complete the RSSR task has received widespread attention in the RS community. Fig. 6 compares the workflow of previous deep learning-based and diffusion model-based RSSR methods. Unlike the previous deep learning-based methods, which attempt to find a suitable mapping function to fuse the detail information in just one step, diffusion model-based methods integrate the information from LR image and guided image into each step of the diffusion process, which facilitates the better a better fusion of different image information.

a) **Multispectral images:** Given that RS images always contain small and dense targets, Liu *et al.* [77] proposed a diffusion model with a detail supplement mechanism for the RSSR task, which requires two-step training to realize. Specifically, the first training aims to improve the model's capability for reconstructing small objects through randomly masking HR images, and the second training is to complete

the super-resolution task by utilizing a conditional diffusion model with LR images as condition input. Although superior performance is achieved, the dual training process in this method is complex and time-consuming. To simplify the training process, Han *et al.* [79] leveraged Transformer [80] and CNN to extract global features and local features from LR images, respectively, and used the fused feature images to guide the diffusion model generate HR images. In this way, the function of two-step training in [77] is successfully realized in one training. Similarly, Xiao *et al.* [81] extracted rich prior knowledge from the original LR images by using stacked residual channel attention blocks [82] to guide the optimization of the diffusion model. Furthermore, An *et al.* [83] departed from the commonly used U-Net architecture in diffusion models, implemented an encoder-decoder architecture through parameter-free approaches, and adopted denoising diffusion implicit models (DDIM) [84] to accelerate sampling process, which significantly improves the efficiency of generating HR images, and is more suitable for diverse RS scenarios.

The above RSSR methods are designed on the assumption that LR images are generated from a fixed degradation model, such as downsampling. However, the blurring in real remote sensing images are complex and varied, which can be modeled as many different degradation models. In view of this, Xu *et al.* [85] proposed to solve this problem with two diffusion models, where the first one is trained as a degradation kernel predictor, so that the predicted degradation kernel and LR image can be used together as conditions in the second diffusion model to generate the HR images. Feng *et al.* [86] achieved the learning of degradation kernel and the reconstruction of HR images within a single diffusion model through the use of the kernel gradient descent module and kernel proximal mapping module.

b) **HSIs:** Although HSIs possess high spectral resolution, their spatial resolution is relatively low [87], which may limit the performance of various applications based on HSIs. To obtain high spatial resolution HSIs, there are two categories methods: pansharpening [88] and multispectral and hyperspectral image fusion [89]–[92]. As the name suggests, the former enhances the HSI by injecting the detail information from panchromatic (PAN) images, while the latter leverages multispectral images to help the HSI learn spatial details. For example, Shi *et al.* [91] used the concatenated image of the multispectral and HSI as the condition input for the diffusion model, enabling the model to capture useful information from both image modalities to generate HSIs with high spatial resolution.

Compared to using multispectral images as the detail guides, researchers are increasingly dedicated to achieve HSI super-resolution with PAN images [78], [93]–[98]. Instead of crudely concatenating the PAN and HSI images directly, Meng *et al.* [93] proposed a Modal Intercalibration Module to enhance and extract features from both images, where the enhanced features is used as condition input to the diffusion model. Cao *et al.* [78] believed that the unique information in different image modalities should not be blended for processing. Thus, they designed two conditional modulation modules to extract coarse-grained style information and fine-grained frequency information respectively as the condition inputs. Still for fully

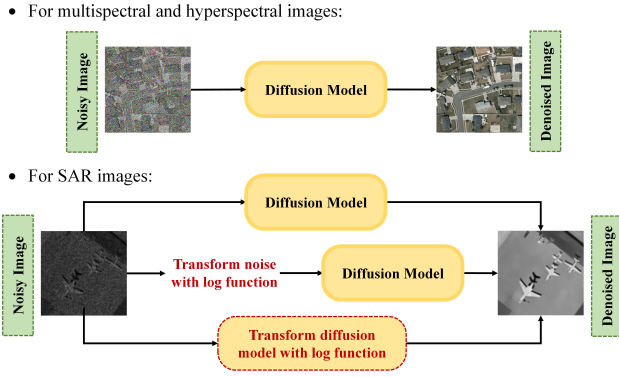


Fig. 7. Overview of diffusion model-based methods for RS image denoising. Since multispectral and hyperspectral images typically exhibit additive noise, diffusion models can be directly employed for denoising. In contrast, SAR images are corrupted by the multiplicative noise. Consequently, there are three approaches to address this problem: directly utilizing the diffusion model, transforming the multiplicative noise into additive noise, as well as transforming the diffusion model to be fit for removing multiplicative noise.

utilizing the unique information of different modalities, Li *et al.* [95] proposed a dual conditional diffusion models-based pansharpening network, which takes HSI and PAN as independent condition inputs to learn the spectral features and spatial texture respectively. Notably, the training and sampling of the proposed diffusion model are performed in a low-dimensional latent space constructed by an auto-encoder network, which not only reduces the computational cost but also maintains the model's generalization ability. [96] also performs the sampling process in a low-dimensional space. Based on the assumption that the HRHSI can be decomposed into the product of two low-rank tensors, this method first computes one of the low-rank tensors with the LRHSI. Then, this tensor is taken as the condition along with the LRHSI and PAN, input into a pre-trained RS diffusion model [99] for another low-rank tensor. This method, unlike the above-mentioned methods, is a completely unsupervised deep learning method that does not require the involvement of HRHSI in any process, providing feasibility for its application in practice.

2) **Cloud Removal:** In many cases, optical RS images will be partially obscured by clouds since the view of imaging equipment is limited. Cloud removal, in essence, is to reconstruct the areas corrupted by clouds, which means to generate content that is consistent with the surrounding environment to fill the missing areas of the image [100]. Therefore, it is suitable to use diffusion model for this task, since it can better control the generated content.

The control condition for cloudy image reconstruction comes in various forms. For example, Czerkawski *et al.* [101] adopted text prompts and edge information [102] as the guiding conditions, together with the input cloudy image, cloud mask, and diffused cloud-free image to control the generation process of SD [48]. Jing *et al.* [103] input both SAR image and cloudy optical RS image into the diffusion model for feature extraction, effectively enhancing the cloud removal results with the help of cloud-unaffected SAR image. Zou *et al.* [104] first extracted features from the cloudy image and noise level, and then inputs the extracted spatial and temporal

features into the diffusion model as control conditions. Instead of using more controlled conditions, they trained the model in a supervised manner with cloud-free images to further improve the quality of the reconstructed images. Moreover, Zhao *et al.* [105] integrated different images from multi-modalities and multi-time into a sequence input, and utilized a two-branch diffusion model to extract scene content from the optical RS and SAR images, respectively. Unlike the previous works, this method does not require the participation of cloud-free images and can handle image sequences of any length, offering greater flexibility and practical value.

3) **Denoising:** Due to the inherent constraints of imaging technology and environmental conditions, RS images are always accompanied by various noise [106]. In essence, the learning process of the diffusion model is equal to the denoising process [15], which makes it possible to achieve superior denoising performance in the context of RS. Different modalities of RS images confront different noise challenges. For example, optical RS images are more susceptible to blurring caused by atmospheric scattering and absorption [107], [108], as well as the noise from changing lighting conditions [109]. HSIs need to consider the uneven distribution of noise over the spectral dimensions [110] and the correlation of noise between different bands [111], [112]. Fortunately, these noises can be effectively removed through diffusion models. Huang *et al.* [107] proposed to crop the noisy RS image into small regions and rearrange them in a cyclic shift manner before fed into the diffusion model, so as to achieve finer local denoising as well as artifact elimination. He *et al.* [111] proposed a truncated diffusion model that starts denoising from the intermediate step of the diffusion process, instead of a pure noise, to avoid the destruction of the inherent effective information in the HSI. Moreover, Yu *et al.* [113] simulated the harsh imaging conditions of RS satellites by adding various attack disturbances to input images, enhancing the diffusion model's ability to counteract the system noise.

SAR images, as another modality of RS images, are usually contaminated by a multiplicative noise, speckle [114]. Unlike additive noise, the degradation caused by speckle varies across different areas in one image. Therefore, speckle noise significantly affects the disparity and interdependence between pixels, causing severe damage to the image. To eliminate this particular noise, Perera *et al.* [115] proposed to use the speckled SAR image, along with the Gaussian noised SAR image that conforms to the standard diffusion model for denoising training. However, this method of introducing the synthetic noise images may not be able to accurately and reasonably simulate the actual SAR images, causing suboptimal denoising performance. Thus, Xiao *et al.* [116] proposed to transform the multiplicative noise in SAR into additive noise by log function, enabling the transformed SAR images to match the standard diffusion model for independent training. Further advancing this methodology, Guha *et al.* [117] integrated the log operation into the derivation process of the diffusion model and obtained a one-step denoising network that can directly address the multiplicative noise.

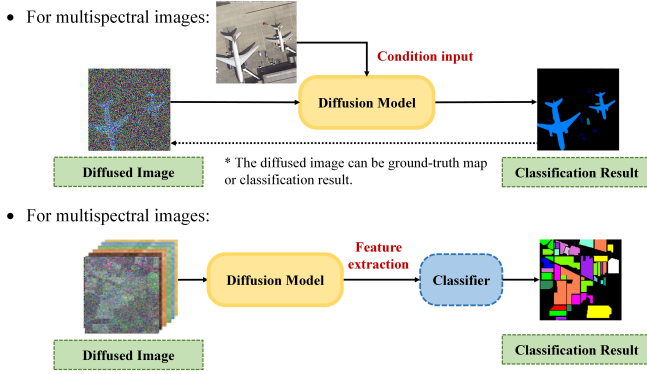


Fig. 8. Overview of diffusion model-based methods for landcover classification. For multispectral images, the ground-truth map is typically used as the diffused image, with the original RS image serving as the condition. The dashed line indicates that the classification results predicted by the network can also be used as the diffused image for the next input. As for HSIs, the prevailing approaches involve using the diffusion model as a feature extractor, after which the extracted features are fed into a classifier for classification.

C. RS Image Interpretation

1) **Landcover Classification:** As one of the most common applications in the RS community, landcover classification aims to assign each pixel to a specific class, such as buildings or grass, to obtain useful landcover information. However, the RS images always contain diverse and complex scenes, increasing the difficulty of accurate classification. Given that the diffusion model can learn and simulate complex data distributions better than other deep learning models, researchers are trying to use it for the RS landcover classification.

The first application of the diffusion model to this task is presented in [118], based on the conditional diffusion model [44]. It takes the manually annotated ground-truth map and the original RS image as the diffused image and condition respectively, and decouples the commonly used U-Net architecture by adding two separate encoders for extracting features from both diffused and guiding images. Besides, it averages the results of multiple sampling for the final classification result to improve the stability and overall accuracy. This method was later tested by Ayala *et al.* [119] on wider RS datasets, validating the effectiveness and development potential of the diffusion model in the RS landcover classification. Instead of using the ground-truth map as the diffused image, Kolbeinson *et al.* [120] diffused the classification prediction result of the network from the previous step and input it along with the conditioning RS image into the diffusion model for next classification prediction. Notably, the parameters of their diffusion model are optimized not only through the MSE of predicted noise but also by minimizing the difference between the prediction result and ground-truth map at each step.

Compared to multispectral images, HSIs exhibit a more complex data distribution, posing a greater challenge for the application of conventional deep learning models in landcover classification. Fortunately, the diffusion model can better capture the spectral-spatial joint features of HSIs, facilitating the improvement of classification accuracy [121]. Based on this fact, Zhou *et al.* [122] constructed a timestep-wise feature bank by utilizing the temporal information of the diffusion model,

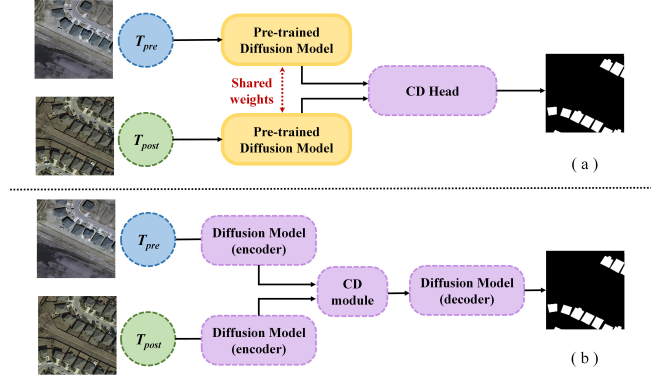


Fig. 9. Comparison of (a) two-step and (b) end-to-end diffusion model-based CD methods. The two-step methods separate the training process of diffusion model and CD head, where the diffusion models (in yellow) are pre-trained by millions of RS images and only the CD head is optimized by the CD dataset. In contrast, the end-to-end method enables the joint optimization of diffusion models and CD module. The figure originally shown in [128]

and proposed a dynamic fusion module to integrate spectral-spatial features with temporal features, making it possible to obtain sufficient image information before the classification. Li *et al.* [123] designed a dual-branch diffusion model for feature extraction from HSI and LiDAR images separately, achieving information complementarity between different modal RS images and enhancing the distinguishability among pixels, which also demonstrated superior performance in a multi-client RS task [124]. Qu *et al.* [125] further set the encoder of the dual-branch diffusion model to operate in parameter-sharing mode to ensure the extraction of shared features from multimodal RS images. In addition, Chen *et al.* [126] utilized the diffusion model to assist deep subspace construction, achieving excellent HSI classification performance in an unsupervised manner. Different from the aforementioned methods, Ma *et al.* [127] used the diffusion model to directly classify the HSI pixels into background and anomaly targets, rather than as a feature extractor. Specifically, they adopted a diffusion model to learn the background distribution based on the fact that the distribution of the background obeys a mixed Gaussian. Thus, the background is removed as noise during the inference process, effectively retaining the anomalous pixels of interest.

2) **Change Detection:** RS change detection (CD) aims to identify the differences between two images taken at different times of the same area [129], thereby providing support for environmental monitoring, and natural disaster assessment. Considering the exceptional performance of diffusion models in handling image details, such as textures and edges, which are crucial for distinguishing changes, some researchers have explored diffusion model-based CD methods.

Bandara *et al.* [99] employed a pre-trained diffusion model to extract multi-scale features of different temporal images for the CD module training, where the pre-training was accomplished by millions of free and unlabeled RS images to capture the key semantic. Tian *et al.* [130] integrated a diffusion model into the contrastive learning framework [131] to capture fine-grained information in the RS images, successfully extracting features with clearer boundaries and richer texture details for the CD task. Additionally, Zhang *et al.* [132] adopted the

Transformer [133] as the backbone for the diffusion model to extract spectral-spatial features from HSIs captured at different times. In essence, all of the above methods use the diffusion model as a feature extractor trained separately from the CD task, which results in the generated features not being fully suitable for CD and overlooks the potential benefits of gradual learning and controllability provided by the diffusion model. To address these issues, Wen *et al.* [128] and Jia *et al.* [134] proposed an end-to-end diffusion model for CD, where the diffused ground-truth map is input into the network along with the pre- and post-change images, and the difference between two temporal images is used as the condition to guide the direction of detection.

3) **Climate Prediction:** Climate prediction is another application for the diffusion model in the field of RS, which is a complex systematic project that requires the integration of multiple variables, including cloud amount, cyclone distribution, and water vapor. Therefore, its solution often consists of multiple diffusion models, enabling fine-grained, stepwise processing of these high-dimensional data. For example, Nath *et al.* [135] cascaded three independently trained diffusion models to generate future satellite imagery, increase the resolution, and predict precipitation. Hatanaka *et al.* [136] utilized two cascaded score-based diffusion models to generate high-resolution cloud cover images from coarse-resolution atmospheric variables. In addition, Leinonen *et al.* [137] addressed the issue of high computational costs by adopting the latent diffusion model concept proposed in [48], running the diffusion process within a latent variable space mapped by a 3D VAE network. These methods all demonstrate that diffusion models can better capture the complex spatio-temporal relationships and become a powerful tool in climate prediction.

4) **Miscellaneous Tasks:** Apart from the classic RS interpretation tasks reviewed above, there are some other tasks that may not fall into the above categories [138]–[140].

One such task is height estimation, which aims at providing pixel-wise height information of surface features (e.g., buildings, trees, terrain, etc.) to generate 3D models of surface scenes. Diffusion model has been reported to be a promising solution for surface estimation [138]. Unlike traditional methods that require multi-view geospatial imagery or LiDAR point clouds, it produces accurate height estimates only with single-view optical RS images.

The other task is object detection, which uses bounding boxes to locate the instances of a certain class (such as plane, vehicle, or ship) in RS images. One of the most challenging issues in RS object detection is the lack of sufficient training data [141]. This scarcity is due to the long-distance photography of RS images, which often results in the object of interest being small and sparsely distributed across different regions in the images. Thus, augmenting the object of interest with the diffusion model has become an effective solution [139]. Specifically, it works by training the SD model with object patches cropped from the available object detection training set, where the patches are 10 pixels larger than the corresponding ground-truth boxes to ensure that sufficient context is captured for seamlessly merging

TABLE I
QUANTITATIVE COMPARISON OF DIFFERENT CLOUD REMOVAL
METHODS ON THE SEN2-MTC-OLD DATASET

Methods	Metrics			
	PSNR \uparrow	SSIM \uparrow	FID \downarrow	LPIPS \downarrow
STNet [142]	26.321	0.834	146.057	0.438
DSen2-CR [143]	26.967	0.855	123.382	0.330
PMAA [144]	27.377	0.861	120.393	0.367
UnCRtainTS [145]	26.417	0.837	130.875	0.400
DDPM-CR [103]	27.060	0.854	110.919	0.320
DiffCR [104]	29.112	0.886	89.845	0.258

*The results are retrieved from [104].

with the background. Compared with other data augmentation methods, this method can extract the precise coordinates of the synthesized objects, and effectively mitigates the long-tailed distribution of target categories.

Additionally, diffusion model also shows superior performance over CNNs in the task of anomaly detection in satellite videos (such as wildfire detection) [140]. The past frames serve as the condition input for the diffusion model, enabling it to learn the data distribution of normal frames and generate high-quality data that closely resemble real images. Consequently, when an anomalous frame is input, the model outputs a significantly higher anomaly score. This means that the diffusion model can detect small wildfires promptly, preventing widespread fire outbreaks, which is as opposed to CNN-based methods that usually require the fire to reach a certain visual extent to be effective.

D. Experimental Evaluation

To effectively illustrate the superiority of diffusion models in processing RS images, we take the experimental results of cloud removal, landcover classification, and change detection as examples in this section to evaluate the performance of the diffusion model and other existing techniques through visual results and quantitative indicators.

Fig. 10 displays the visual results of six different cloud removal methods on multi-temporal optical satellite (Sentinel-2) images [156], where the first three cloudy images are taken at different times from the same location. As shown in Fig. 10, diffusion model-based methods DDPM-CR [103] and DiffCR [104] have successfully removed clouds without leaving excessive artifacts, restoring the RS image with detailed information. This observation is also confirmed by the quantitative indicators in Table I. Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) [157], Learned Perceptual Image Patch Similarity (LPIPS) [158] and Frechet Inception Distance (FID) [159] are used in Table I to evaluate the quality of cloud-free images generated by the comparison methods. As shown in Table I, DiffCR achieves the best performance on all four indicators, and DDPM-CR, another diffusion model-based method, also ranked second on FID and LPIPS indicators, which demonstrates that using diffusion models for cloud removal is highly competitive.

As illustrated in Fig. 11, the diffusion model-based method SpectralDiff [121] has significantly better classification results

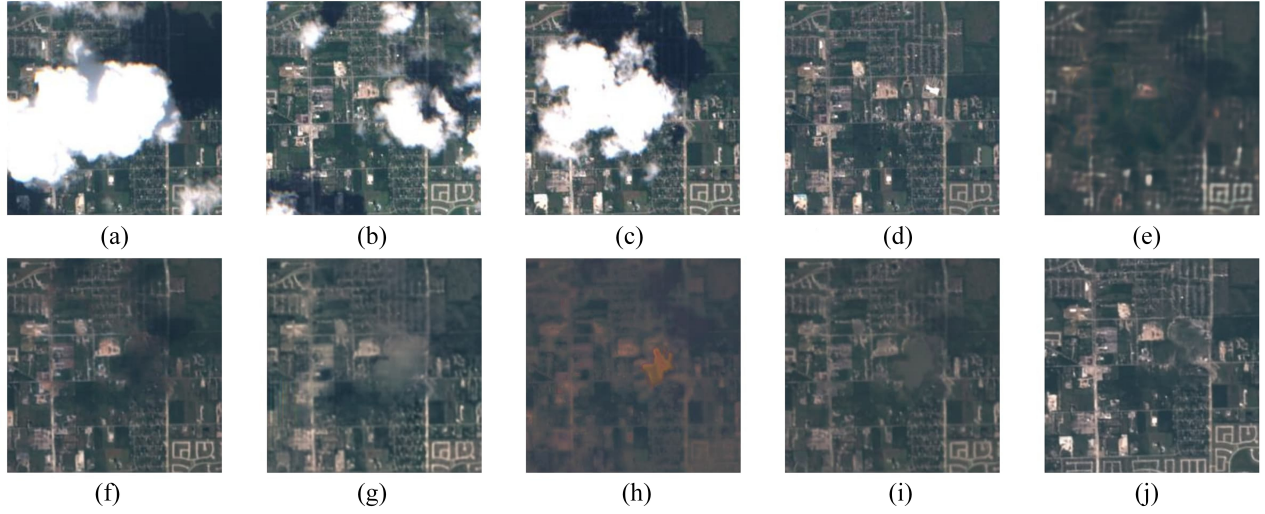


Fig. 10. Comparison of different cloud removal methods on the Sen2-MTC-Old dataset: (a) Cloudy Image T1. (b) Cloudy Image T2. (c) Cloudy Image T3. (d) Ground-Truth. (e) STNet [142]. (f) DSen2-CR [143]. (g) PMAA [144]. (h) UnCRtainTS [145]. (i) DDPM-CR [103]. (j) DiffCR [104]. The visual results are retrieved from [104].

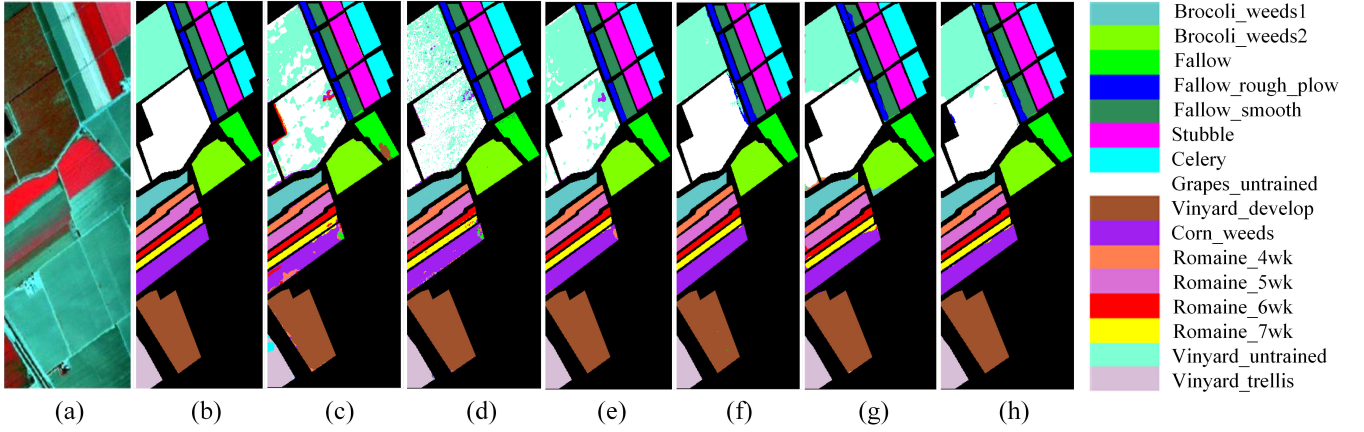


Fig. 11. Comparison of different HSI classification methods on the Salinas dataset: (a) Pseudo-Color Image. (b) Ground-Truth. (c) SF [146]. (d) miniGCN [147]. (e) SSFTT [148]. (f) DMVL [149]. (g) SSGRN [150]. (h) SpectralDiff [121]. The visual results are retrieved from [121].

TABLE II
QUANTITATIVE COMPARISON OF DIFFERENT HSI CLASSIFICATION
METHODS ON THE SALINAS DATASET

Methods	Metrics		
	OA (%)	AA (%)	κ (%)
SF [146]	88.248	93.262	86.973
miniGCN [147]	88.181	94.297	86.823
SSFTT [148]	95.789	98.272	95.322
DMVL [149]	97.005	95.853	96.668
SSGRN [150]	96.539	96.354	96.144
SpectralDiff [121]	98.971	99.465	98.854

*The results are retrieved from [121].

TABLE III
QUANTITATIVE COMPARISON OF DIFFERENT CHANGE DETECTION
METHODS ON THE LEVIR DATASET

Methods	Metrics		
	F1(%)	IoU(%)	OA(%)
FC-SD [151]	86.31	75.92	98.67
STANet [152]	87.26	77.40	98.66
SNUNet [153]	88.16	78.83	98.82
BIT [154]	89.31	80.68	98.92
ChangeFormer [155]	90.40	82.48	99.04
DDPM-CD [99]	90.91	83.35	99.09

*The results are retrieved from [99].

than SF [146], miniGCN [147], and SSFTT [148] on the hyperspectral dataset Salinas. Although DMVL [149] and SSGRN [150] show comparable performance to SpectralDiff across most classes, they are not as accurate as SpectralDiff in assigning pixels at the boundaries of different classes. In

addition to the visualized results, Table II lists the quantitative results of these methods on overall accuracy (OA), average accuracy (AA), and Kappa coefficient, where SpectralDiff ranks first.

As for the change detection task, we selected five compar-

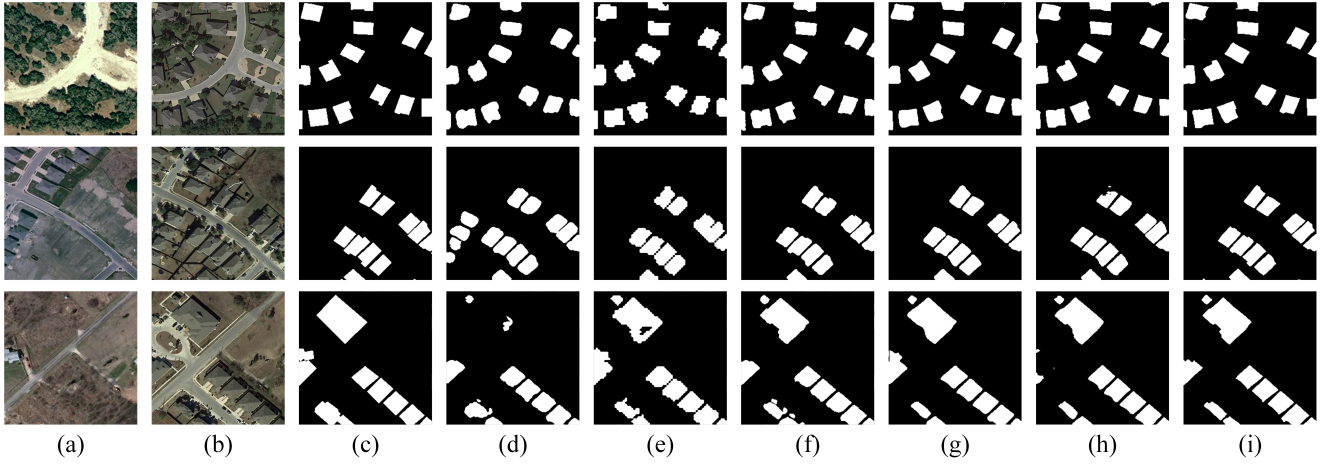


Fig. 12. Comparison of different change detection methods on the LEVIR dataset: (a) Pre-change Image. (b) Post-change Image. (c) Ground-Truth. (d) FC-SD [151]. (e) STANet [152]. (f) SNUNet [153]. (g) BIT [154]. (h) ChangeFormer [155]. (i) DDPM-CD [99].

ised methods based on different deep learning models along with a diffusion model-based method, DDPM-CD [99], for experimentation on the LEVIR dataset [160]. The corresponding visual results and the values of evaluation indicators, i.e., F1 score (F1), overall accuracy (OA), and intersection over union (IoU), are presented in Fig. 12 and Table III, respectively. It is evident from both qualitative analysis and quantitative comparisons that the change detection method based on the diffusion model is significantly superior to the others.

IV. DISCUSSIONS AND FUTURE DIRECTIONS FOR RS DIFFUSION MODELS

As discussed in the previous sections, diffusion models are rapidly evolving in the RS community, presenting great potential from generating RS images to enhancing the image quality, and further to recognition and detection. In fact, the research on diffusion models in RS is still at an early stage, with many tasks to be explored and further improvements to be achieved. In the following sections, we will discuss the possible future research directions from two aspects: extended applications and model deployment.

A. Extended Applications

1) **For Specific RS Tasks:** As shown in Fig. 13, diffusion models are most frequently used in the RS image generation task, followed by the task of reconstructing high-resolution RS images (i.e., the super-resolution task), which accounts for 21% of all reviewed papers. Comparatively, the use of diffusion models in advanced and complex interpretation tasks is quite less. Especially for object detection, an important and common RS task, there is only one paper related to the diffusion model [139], which implies a huge gap in the research of diffusion models for this task. Recently, Chen *et al.* [161] proposed a new object detection network named DiffusionDet, which finds the correct positions of objects by gradually denoising from random noise boxes to object boxes. This method ingeniously combines the object detection task with diffusion model, and achieves favorable performance on

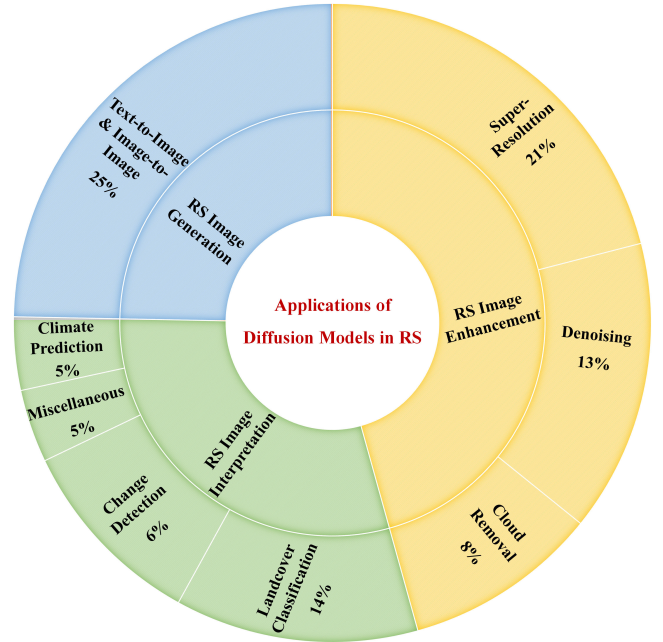


Fig. 13. Frequency of diffusion models in different RS applications.

the COCO dataset [162]. Although the performance of this method on small and sparse RS targets is still unclear, its emergence has paved an enlightening new path for researchers in the RS field to further explore in the object detection task. Nevertheless, Fig. 1 presents that the research on diffusion models for RS image generation has decreased since 2024, while research on their application in RS image interpretation tasks, especially in the change detection task, has significantly increased. This suggests that researchers have realized the limitations of diffusion models in the application to RS. Therefore, developing more effective RS diffusion models for image interpretation tasks is an important research direction for the future.

While diffusion models are also commonly applied in land-cover classification and change detection, there exists a sig-

TABLE IV
CATEGORIZATION OF DIFFUSION MODELS IN RS BASED ON IMAGE MODALITIES

Modality	Application	Correlation paper
Multispectral	Generation	[55] [59] [54] [65] [61] [62] [75] [60] [67] [73] [64] [56] [63]
	Cloud Removal	[100] [101] [103] [104] [105]
	Super-Resolution	[77] [79] [81] [83] [85] [86]
Hyperspectral	Super-Resolution	[78] [90] [91] [92] [93] [94] [95] [96] [97] [98]
	Classification	[121] [122] [123] [124] [125] [126] [127]
	Denosing	[111] [112] [110]
SAR	Denosing	[115] [116] [117]
	Generation	[57] [71] [167]

nificant limitation. In these methods, diffusion models mainly serve as feature extractors for the input image [99], [121]–[124], [130], [132], requiring an additional classifier/detector to execute the specific task. Such separation of feature extraction and detection is not straightforward and usually needs two-step training, which is likely to fall into a suboptimal solution due to the extracted features not being fully suitable for the specific task. Therefore, future research could focus on developing RS task-specific diffusion models, by incorporating task-related prior in the model design or designing end-to-end models to improve the performance of diffusion models on specific RS tasks.

It is worth noting that the vast majority of existing RS diffusion models are based on the U-Net architecture, with only a few works incorporating the Transformer architecture or its attention mechanism [79], [121], [132]. Among the most innovative is [132], which employs a completely Transformer-based diffusion model, U-ViT [133], for change detection. Nevertheless, the U-ViT consists of long skip residual connections, aligned with those of U-Net. In contrast, DiT [163], another diffusion model entirely based on the Transformer architecture, set the residual connections within each block, allowing the attention layer to perform global convolution and information extraction at a finer-grained scale, achieving state-of-the-art results in both image and video generation [163]–[166]. Thus, it can be seen that the diffusion model based on transformer architecture hold great potential for development, warranting further exploration by researchers in the field of RS.

2) **For Multi-Modal RS Images:** Unlike natural images, RS images are captured by different types of sensors, encompassing multiple modalities. Table IV lists the top three most common applications across different modal RS images. Obviously, most of the diffusion model-based methods are developed for multispectral images. For HSI images, the applications of diffusion models are focused on super-resolution, especially pansharpening [78], [93]–[96], and classification tasks. As for SAR images, diffusion models only appear in the process of generation and denoising. However, the spectral signatures of HSI and the robustness of SAR images play crucial roles in the recognition and detection tasks [168]–

[172]. Therefore, exploring how to effectively apply diffusion models to multi-modal RS images is a necessary future research direction. In this way, the unique information of different modalities can be leveraged to improve the accuracy of RS image analysis.

In addition, LiDAR data, as an important type of RS data that can provide surface height information and ground structure details, is seldom used in the existing RS diffusion models. Only Li *et al.* [123] introduced LiDAR images as auxiliary information when using diffusion models for hyperspectral classification. In fact, LiDAR images can not only be used as supplement to other modalities [173], [174], but also generate continuous 3D terrain models for topographic and geomorphological analysis [175], vegetation detection [176], [177], and urban planning [178], [179]. Recently, diffusion models have demonstrated satisfactory performance in 3D point cloud generation [19], [180], [181]. Such technological advancements may be borrowed to the RS LiDAR data, thereby filling the gap of LiDAR data in various RS tasks.

3) **For Realistic RS Images:** Although many diffusion model-based RS image generation methods have been developed, they often overlook some special features of RS images, resulting in noticeable gaps between the synthesized image and real images. For example, diffusion model-based HSI generation methods usually need to compress the spectral dimension [68], [69], which neglects the details of spectral curves and hinders the generation of accurate spectra. SAR images are complex in nature, comprising both amplitude and phase terms [182]. However, the generation methods developed so far have mainly focused on the image amplitude, ignoring the phase information. Therefore, future diffusion model-based generation methods should take these aspects into account to obtain more realistic RS images.

Another aspect that is often overlooked is the size of real RS images, which tend to be quite large (e.g., Gaofen-2 images are $29,200 \times 27,620$ pixels) [183], [184]. Existing diffusion model-based generation methods are primarily designed for patch-level images with 256×256 pixels, which means that a large-scale RS image can only be obtained by stitching multiple patch images [75]. From the visual perspective, this approach is suboptimal since it is difficult to ensure the continuity of scenes and easy to leave traces at the joints. Thus, how to obtain realistic and reasonable large-scale RS images with diffusion models is a worthy research direction for further exploration. Notably, the increase in image size will inevitably bring computational and storage burdens. Accordingly, how to generate large-scale RS images under limited resources also needs careful consideration.

4) **For General RS Model:** Nowadays, more and more researchers are devoting themselves to developing general intelligent models, which can provide more accessible and high-performing solutions to help both industry professionals and interested non-professionals [46], [185]–[188]. For example, ChatGPT [187] has greatly simplified the process of collecting and summarizing information, while DALL-E [188] has facilitated the rapid transformation of artistic ideas into practical examples. In the field of RS, some researchers are also attempting to develop universal multi-modal large

models [189]–[192]. However, these studies are still in the embryonic stage and have a lot of room for development. Fortunately, diffusion models have shown superior performance on various RS applications, as presented before. Therefore, a promising research direction is to construct a general RS intelligent model based on the diffusion model, which can span different modalities of RS images and accomplish multiple earth observation tasks.

B. Model Deployment

It is widely acknowledged that diffusion models require a substantial number of iterations to generate high-quality samples, which is a noticeable drawback of this technology. Moreover, these models often encompass a substantial number of parameters, necessitating deployment on devices with powerful neural computing units [193], [194]. However, the resources on satellites are extremely limited and lack the computational power for such requirements. What's more serious is that the deep-space environment is severe, often subject to extreme climate and illumination changes, which puts higher demands on the stability and reliability of the model. Therefore, how to deploy diffusion models on resource-limited satellites within harsh environments for real-time processing is a very meaningful research direction.

1) **Accelerate Processing:** Recently, some accelerated sampling approaches for diffusion models have been proposed [84], [195]–[198], which successfully reduce the necessary sampling steps from several hundred to dozens, or even a few or a single step [199]–[201]. However, Tuel *et al.* [167] found that these accelerated sampling methods, such as DDIM [84] or DPM-solver [197], did not perform well on SAR images. This observation suggests that the significant differences between RS images and natural images make these acceleration methods, originally designed for natural images, inapplicable to RS imagery. More recently, Kodaira *et al.* [202] proposed StreamDiffusion, which achieves the speed of generating images up to 91.07fps on a 4090GPU through pipelined batch processing. Unfortunately, they did not evaluate the proposed model on RS images and overlooked that some of the acceleration techniques are not suitable for devices with limited computing resources. It can be seen that developing acceleration methods for RS diffusion models is still an open problem.

Another common optimization strategy is to design lightweight diffusion models, which can be achieved by distillation techniques [65], [79], [203], [204] or changing the network structure [81], [83], [205], [206]. Nevertheless, the more prevalent solution in RS is to place the diffusion model in a lower-dimensional latent space through data compression [69], [95], [96], [137] thereby shrinking the sampling space and reducing computational overhead. However, the performance of such methods is limited by the quality of the constructed latent space [207]. If the latent space is incapable of extracting useful semantic information from RS imagery, or critical information is lost during the dimensionality reduction process, the final generated RS images will be adversely affected. Therefore, it is necessary to explore better lightweight structure design methods than compressing the data space.

2) **Improve Stability and Reliability:** As for the deployment challenges posed by the harsh environment, Yu *et al.* [113] proposed a diffusion model-based adversarial defense approach to protect deep neural networks from a variety of unknown adversarial attacks, effectively improving the robustness of the model. Similar research on employing diffusion models to tackle the complex environments in RS should be further explored, especially to design specific diffusion models for different environmental disturbances or different sensors. Such efforts would contribute to the long-term operational stability of the deployed models and hold great practical value.

In addition, with the rapid development of Global Observation System (GOS), the distributed learning of intelligent models over multiple satellites has gradually become one of mainstream directions in RS [208]–[211]. More recently, Li *et al.* [212] reported that the speed of generating images with diffusion models can be effectively improved under this parallel learning mode, further confirming prospects of distributed diffusion models in RS. However, when deploying diffusion models in such multi-client distributed scenarios, ensuring the security of RS data is an essential issue [213]. Consequently, developing robust and trustworthy diffusion models has become an urgent necessity in the RS community.

V. CONCLUSION

In summary, the emergence of diffusion models has created a new era of intelligent RS image processing. Compared to other deep generative models, diffusion models are robust to the inherent noise in RS images, better adapt to their variability and complexity, and offer a more stable training process. Hence, applying diffusion models to various RS tasks has become an inevitable trend. For this reason, this paper first introduces the theoretical background of diffusion models to help understand how the diffusion model works for RS tasks. Then, it reviews and summarizes studies on the use of diffusion models in processing RS images, including image generation, super-resolution, cloud removal, denoising, and a series of interpretation tasks such as landcover classification, change detection, and climate prediction. Moreover, the paper takes cloud removal, landcover classification, and change detection as examples to demonstrate the superiority of diffusion models in various RS image processing tasks through visual results and quantitative indicators. Finally, the paper discusses the limitations of the existing diffusion models in RS and highlights that further exploration could be carried out on the extended applications and model deployment. We hope this paper can provide a valuable reference for researchers in related fields to stimulate more innovative studies to break the performance bottleneck of existing methods or to promote the development of diffusion models for more RS applications.

REFERENCES

- [1] S. P. Mertikas, P. Partsinevelos, C. Mavrocordatos, and N. A. Maximenko, "Environmental applications of remote sensing," in *Pollution assessment for sustainable practices in applied sciences and engineering*. Elsevier, 2021, pp. 107–163.

- [2] M. Jhawar, N. Tyagi, and V. Dasgupta, "Urban planning using remote sensing," *International Journal of Innovative Research in Science, Engineering and Technology*, vol. 1, no. 1, pp. 42–57, 2013.
- [3] A. Khan, S. Gupta, and S. K. Gupta, "Multi-hazard disaster studies: Monitoring, detection, recovery, and management, based on emerging technologies and optimal techniques," *International journal of disaster risk reduction*, vol. 47, p. 101642, 2020.
- [4] Z. Yang, X. Yu, S. Dedman, M. Rosso, J. Zhu, J. Yang, Y. Xia, Y. Tian, G. Zhang, and J. Wang, "Uav remote sensing applications in marine monitoring: Knowledge visualization and review," *Science of The Total Environment*, vol. 838, p. 155939, 2022.
- [5] M. Shimoni, R. Haelterman, and C. Perneel, "Hypersectral imaging for military and security applications: Combining myriad processing and sensing techniques," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 101–117, 2019.
- [6] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, 2017.
- [7] L. Zhang and Y. Liu, "Remote sensing image generation based on attention mechanism and vae-msgan for roi extraction," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2021.
- [8] S. Valero, F. Agulló, and J. Inglada, "Unsupervised learning of low dimensional satellite image representations via variational autoencoders," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, 2021, pp. 2987–2990.
- [9] P. Jian, K. Chen, and W. Cheng, "Gan-based one-class classification for remote-sensing image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2021.
- [10] S. Jozdani, D. Chen, D. Pouliot, and B. A. Johnson, "A review and meta-analysis of generative adversarial networks and their applications in remote sensing," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 108, p. 102734, 2022.
- [11] H. Wu, N. Ni, S. Wang, and L. Zhang, "Blind super-resolution for remote sensing images via conditional stochastic normalizing flows," *arXiv preprint arXiv:2210.07751*, 2022.
- [12] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.
- [13] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *Proc. Int. Conf. Mach. Learn. (ICML)*. PMLR, 2015, pp. 1530–1538.
- [14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 27, 2014.
- [15] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 33, pp. 6840–6851, 2020.
- [16] C. Zhang, C. Zhang, M. Zhang, and I. S. Kweon, "Text-to-image diffusion model in generative ai: A survey," *arXiv preprint arXiv:2303.07909*, 2023.
- [17] H. Cao, C. Tan, Z. Gao, Y. Xu, G. Chen, P.-A. Heng, and S. Z. Li, "A survey on generative diffusion models," *IEEE Transactions on Knowledge and Data Engineering*, 2024.
- [18] Z. Wang, C. Lu, Y. Wang, F. Bao, C. Li, H. Su, and J. Zhu, "Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation," *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 36, 2024.
- [19] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "Dreamfusion: Text-to-3d using 2d diffusion," *arXiv preprint arXiv:2209.14988*, 2022.
- [20] T. Yi, J. Fang, J. Wang, G. Wu, L. Xie, X. Zhang, W. Liu, Q. Tian, and X. Wang, "Gaussiandreamer: Fast generation from text to 3d gaussians by bridging 2d and 3d diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024.
- [21] O. Avrahami, D. Lischinski, and O. Fried, "Blended diffusion for text-driven editing of natural images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 18 208–18 218.
- [22] B. Kawar, S. Zada, O. Lang, O. Tov, H. Chang, T. Dekel, I. Mosseri, and M. Irani, "Imagic: Text-based real image editing with diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 6007–6017.
- [23] B. Yang, S. Gu, B. Zhang, T. Zhang, X. Chen, X. Sun, D. Chen, and F. Wen, "Paint by example: Exemplar-based image editing with diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 18 381–18 391.
- [24] Y. Huang, J. Huang, Y. Liu, M. Yan, J. Lv, J. Liu, W. Xiong, H. Zhang, S. Chen, and L. Cao, "Diffusion model-based image editing: A survey," *arXiv preprint arXiv:2402.17525*, 2024.
- [25] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi, "Palette: Image-to-image diffusion models," in *ACM SIGGRAPH 2022 conference proceedings*, 2022, pp. 1–10.
- [26] J. Seo, G. Lee, S. Cho, J. Lee, and S. Kim, "Midms: Matching interleaved diffusion models for exemplar-based image translation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 2, 2023, pp. 2191–2199.
- [27] B. Li, K. Xue, B. Liu, and Y.-K. Lai, "Bbdm: Image-to-image translation with brownian bridge diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 1952–1961.
- [28] N. Tumanyan, M. Geyer, S. Bagon, and T. Dekel, "Plug-and-play diffusion features for text-driven image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 1921–1930.
- [29] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang, "Diffusion models: A comprehensive survey of methods and applications," *ACM Computing Surveys*, vol. 56, no. 4, pp. 1–39, 2023.
- [30] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2023.
- [31] X. Li, Y. Ren, X. Jin, C. Lan, X. Wang, W. Zeng, X. Wang, and Z. Chen, "Diffusion models for image restoration and enhancement—a comprehensive survey," *arXiv preprint arXiv:2308.09388*, 2023.
- [32] B. B. Moser, A. S. Shanbhag, F. Raue, S. Frolov, S. Palacio, and A. Dengel, "Diffusion models, image super-resolution and everything: A survey," *arXiv preprint arXiv:2401.00736*, 2024.
- [33] X. Li, J. Thickstun, I. Gulrajani, P. S. Liang, and T. B. Hashimoto, "Diffusion-lm improves controllable text generation," *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 35, pp. 4328–4343, 2022.
- [34] H. Zou, Z. M. Kim, and D. Kang, "Diffusion models in nlp: A survey," *arXiv preprint arXiv:2305.14671*, 2023.
- [35] J. Lovelace, V. Kishore, C. Wan, E. Shekhtman, and K. Q. Weinberger, "Latent diffusion for language generation," *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 36, 2024.
- [36] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "Dif-fwave: A versatile diffusion model for audio synthesis," *arXiv preprint arXiv:2009.09761*, 2020.
- [37] Y. Leng, Z. Chen, J. Guo, H. Liu, J. Chen, X. Tan, D. Mandic, L. He, X. Li, T. Qin *et al.*, "Binauralgrad: A two-stage conditional diffusion probabilistic model for binaural audio synthesis," *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 35, pp. 23 689–23 700, 2022.
- [38] R. Huang, J. Huang, D. Yang, Y. Ren, L. Liu, M. Li, Z. Ye, J. Liu, X. Yin, and Z. Zhao, "Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models," in *Proc. Int. Conf. Mach. Learn. (ICML)*. PMLR, 2023, pp. 13 916–13 932.
- [39] B. Jing, G. Corso, J. Chang, R. Barzilay, and T. Jaakkola, "Torsional diffusion for molecular conformer generation," *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 35, pp. 24 240–24 253, 2022.
- [40] E. Hoogeboom, V. G. Satorras, C. Vignac, and M. Welling, "Equivariant diffusion for molecule generation in 3d," in *Proc. Int. Conf. Mach. Learn. (ICML)*. PMLR, 2022, pp. 8867–8887.
- [41] Z. Guo, J. Liu, Y. Wang, M. Chen, D. Wang, D. Xu, and J. Cheng, "Diffusion models in bioinformatics and computational biology," *Nat. Rev. Bioeng.*, vol. 2, no. 2, pp. 136–154, 2024.
- [42] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proc. Int. Conf. Mach. Learn. (ICML)*. PMLR, 2015, pp. 2256–2265.
- [43] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Medical Image Comput. Comput.-Assisted Intervent. (MICCAI)*, 2015, pp. 234–241.
- [44] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 34, pp. 8780–8794, 2021.
- [45] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022.
- [46] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.
- [47] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 35, pp. 36 479–36 494, 2022.

- [48] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 10684–10695.
- [49] Y. Xu, W. Yu, P. Ghamisi, M. Kopp, and S. Hochreiter, "T2img-mhn: Remote sensing image generation from text using modern hopfield networks," *IEEE Trans. Image Process.*, vol. 32, pp. 5737–5750, 2023.
- [50] J. Karras, A. Holynski, T.-C. Wang, and I. Kemelmacher-Shlizerman, "Dreampose: Fashion video synthesis with stable diffusion," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 22680–22690.
- [51] M. N. Everaert, M. Bocchio, S. Arpa, S. Süsstrunk, and R. Achanta, "Diffusion in style," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 2251–2261.
- [52] S. Shen, Z. Zhu, L. Fan, H. Zhang, and X. Wu, "Diffclip: Leveraging stable diffusion for language grounded 3d classification," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, 2024, pp. 3596–3605.
- [53] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman *et al.*, "Laion-5b: An open large-scale dataset for training next generation image-text models," *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 35, pp. 25278–25294, 2022.
- [54] R. Ou, H. Yan, M. Wu, and C. Zhang, "A method of efficient synthesizing post-disaster remote sensing image with diffusion model and llm," in *Proc. IEEE Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, 2023, pp. 1549–1555.
- [55] S. Khanna, P. Liu, L. Zhou, C. Meng, R. Rombach, M. Burke, D. B. Lobell, and S. Ermon, "Diffusionsat: A generative foundation model for satellite imagery," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2024. [Online]. Available: <https://openreview.net/forum?id=15webNFDgQ>
- [56] D. Tang, X. Cao, X. Hou, Z. Jiang, and D. Meng, "Crs-diff: Controllable generative remote sensing foundation model," *arXiv preprint arXiv:2403.11614*, 2024.
- [57] Z. Tian, Z. Chen, and Q. Sun, "Non-visible light data synthesis and application: A case study for synthetic aperture radar imagery," *arXiv preprint arXiv:2311.17486*, 2023.
- [58] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.
- [59] A. Sebaq and M. ElHelw, "Rsdif: Remote sensing image generation from text using diffusion model," *arXiv preprint arXiv:2309.02455*, 2023.
- [60] M. Espinosa and E. J. Crowley, "Generate your own scotland: Satellite image generation conditioned on maps," in *NeurIPS 2023 Workshop on Diffusion Models*, 2023.
- [61] Z. Wu, J. Qian, M. Zhang, Y. Cao, T. Wang, and L. Yang, "High-confidence sample augmentation based on label-guided denoising diffusion probabilistic model for active deception jamming recognition," *IEEE Geosci. Remote Sens. Lett.*, 2023.
- [62] O. Baghirli, H. Askarov, I. Ibrahimli, I. Bakhishov, and N. Nabiye, "Satdm: Synthesizing realistic satellite image with semantic layout conditioning using diffusion models," in *Proc. 74th Int. Astro. Cong. (IAC)*. Baku, Azerbaijan: IAF, October 2023.
- [63] C. Zhao, Y. Ogawa, S. Chen, Z. Yang, and Y. Sekimoto, "Label freedom: Stable diffusion for remote sensing image semantic segmentation data generation," in *Proc. IEEE Int. Conf. Big Data (BigData)*, 2023, pp. 1022–1030.
- [64] Z. Chen, D. Duggirala, D. Crandall, L. Jiang, and L. Liu, "Sepaint: Semantic map inpainting via multinomial diffusion," *arXiv preprint arXiv:2303.02737*, 2023.
- [65] Z. Yuan, C. Hao, R. Zhou, J. Chen, M. Yu, W. Zhang, H. Wang, and X. Sun, "Efficient and controllable remote sensing fake sample generation based on diffusion model," *IEEE Trans. Geosci. Remote Sens.*, 2023.
- [66] L. Zhang, A. Rao, and M. Agrawal, "Adding conditional control to text-to-image diffusion models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 3836–3847.
- [67] X. Bai, X. Pu, and F. Xu, "Conditional diffusion for sar to optical image translation," *IEEE Geosci. Remote Sens. Lett.*, 2023.
- [68] L. Zhang, X. Luo, S. Li, and X. Shi, "R2h-cdd: Hyperspectral imagery generation from rgb images based on conditional cascade diffusion probabilistic models," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, 2023.
- [69] L. Liu, B. Chen, H. Chen, Z. Zou, and Z. Shi, "Diverse hyperspectral remote sensing image synthesis with diffusion models," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023.
- [70] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 12873–12883.
- [71] X. Wang, H. Liao, Z. Yang, T. Han, and J. Xia, "Optical-sar image translation via denoising diffusion implicit model," in *Proc. IEEE Int. Conf. Image Process. Comput. Appl. (ICIPCA)*, 2023.
- [72] J. Wang, L. Du, Y. Li, G. Lyu, and B. Chen, "Attitude and size estimation of satellite targets based on isar image interpretation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2021.
- [73] M. Seo, Y. Oh, D. Kim, D. Kang, and Y. Choi, "Improved flood insights: Diffusion-based sar to eo image translation," *arXiv preprint arXiv:2307.07123*, 2023.
- [74] Z. Li, Z. Li, Z. Cui, M. Pollefeys, and M. R. Oswald, "Sat2scene: 3d urban scene generation from satellite images with diffusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024.
- [75] A. Graikos, S. Yellapragada, M.-Q. Le, S. Kapse, P. Prasanna, J. Saltz, and D. Samaras, "Learned representation-guided diffusion models for large-image generation," 2023.
- [76] P. Wang, B. Bayram, and E. Sertel, "A comprehensive review on deep learning based remote sensing image super-resolution methods," *Earth-Sci. Rev.*, p. 104110, 2022.
- [77] J. Liu, Z. Yuan, Z. Pan, Y. Fu, L. Liu, and B. Lu, "Diffusion model with detail complement for super-resolution of remote sensing," *Remote Sens.*, vol. 14, no. 19, p. 4834, 2022.
- [78] Z. Cao, S. Cao, L.-J. Deng, X. Wu, J. Hou, and G. Vivone, "Diffusion model with disentangled modulations for sharpening multispectral and hyperspectral images," *Inf. Fusion*, vol. 104, p. 102158, 2024.
- [79] L. Han, Y. Zhao, H. Lv, Y. Zhang, H. Liu, G. Bi, and Q. Han, "Enhancing remote sensing image super-resolution with efficient hybrid conditional diffusion model," *Remote Sens.*, vol. 15, p. 3452, 2023.
- [80] Z. Lu, J. Li, H. Liu, C. Huang, L. Zhang, and T. Zeng, "Transformer for single image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 457–466.
- [81] Y. Xiao, Q. Yuan, K. Jiang, J. He, X. Jin, and L. Zhang, "Ediffr: An efficient diffusion probabilistic model for remote sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, 2023.
- [82] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 286–301.
- [83] T. An, B. Xue, C. Huo, S. Xiang, and C. Pan, "Efficient remote sensing image super-resolution via lightweight diffusion models," *IEEE Geosci. Remote Sens. Lett.*, 2023.
- [84] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.
- [85] M. Xu, J. Ma, and Y. Zhu, "Dual-diffusion: Dual conditional denoising diffusion probabilistic models for blind super-resolution reconstruction in rsis," *IEEE Geosci. Remote Sens. Lett.*, 2023.
- [86] Y. Feng, Y. Yang, X. Fan, Z. Zhang, and J. Zhang, "A multiscale generalized shrinkage threshold network for image blind deblurring in remote sensing," *IEEE Trans. on Geosci. and Remote Sens.*, vol. 62, pp. 1–16, 2024.
- [87] J. Jia, J. Chen, X. Zheng, Y. Wang, S. Guo, H. Sun, C. Jiang, M. Karjalainen, K. Karila, Z. Duan *et al.*, "Tradeoffs in the spatial and spectral resolution of airborne hyperspectral imaging systems: A crop identification case study," *IEEE Trans. on Geosci. and Remote Sens.*, vol. 60, pp. 1–18, 2021.
- [88] A. Arienzo, G. Vivone, A. Garzelli, L. Alparone, and J. Chanussot, "Full-resolution quality assessment of pansharpening: Theoretical and hands-on approaches," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 3, pp. 168–201, 2022.
- [89] G. Vivone, "Multispectral and hyperspectral image fusion in remote sensing: A survey," *Inf. Fusion*, vol. 89, pp. 405–417, 2023.
- [90] J. Liu, Z. Wu, and L. Xiao, "A spectral diffusion prior for hyperspectral image super-resolution," *arXiv preprint arXiv:2311.08955*, 2023.
- [91] S. Shi, L. Zhang, and J. Chen, "Hyperspectral and multispectral image fusion using the conditional denoising diffusion probabilistic model," *arXiv preprint arXiv:2307.03423*, 2023.
- [92] J. Qu, J. He, W. Dong, and J. Zhao, "S2cyclediff: Spatial-spectral-bilateral cycle-diffusion framework for hyperspectral image super-resolution," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 5, 2024, pp. 4623–4631.
- [93] Q. Meng, W. Shi, S. Li, and L. Zhang, "Pandiff: A novel pansharpening method based on denoising diffusion probabilistic model," *IEEE Trans. on Geosci. and Remote Sens.*, 2023.
- [94] S. Du, H. Huang, F. He, H. Luo, Y. Yin, X. Li, L. Xie, R. Guo, and S. Tang, "Unsupervised stepwise extraction of offshore aquaculture

- ponds using super-resolution hyperspectral images,” *Int. J. Appl. Earth Observ. Geoinf.*, vol. 119, p. 103326, 2023.
- [95] S. Li, S. Li, and L. Zhang, “Hyperspectral and panchromatic images fusion based on the dual conditional diffusion models,” *IEEE Trans. on Geosci. and Remote Sens.*, 2023.
- [96] X. Rui, X. Cao, L. Pang, Z. Zhu, Z. Yue, and D. Meng, “Unsupervised hyperspectral pansharpening via low-rank diffusion model,” *Inf. Fusion*, p. 102325, 2024.
- [97] J. Wei, L. Gan, W. Tang, M. Li, and Y. Song, “Diffusion models for spatio-temporal-spectral fusion of homogeneous gaofen-1 satellite platforms,” *Int. J. Appl. Earth Observ. Geoinf.*, vol. 128, p. 103752, 2024.
- [98] Y. Xing, L. Qu, S. Zhang, X. Zhang, and Y. Zhang, “Crossdiff: Exploring self-supervised representation of pansharpening via cross-predictive diffusion model,” *arXiv preprint arXiv:2401.05153*, 2024.
- [99] W. G. C. Bandara, N. G. Nair, and V. M. Patel, “Ddpm-cd: Remote sensing change detection using denoising diffusion probabilistic models,” *arXiv preprint arXiv:2206.11892*, 2022.
- [100] F. Sanguigni, M. Czerkawski, L. Papa, I. Amerini, and B. L. Saux, “Diffusion models for earth observation use-cases: from cloud removal to urban change detection,” *arXiv preprint arXiv:2311.06222*, 2023.
- [101] M. Czerkawski and C. Tachatzis, “Exploring the capability of text-to-image diffusion models with structural edge guidance for multi-spectral satellite image inpainting,” *IEEE Geosci. Remote Sens. Lett.*, 2024.
- [102] S. Xie and Z. Tu, “Holistically-nested edge detection,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 1395–1403.
- [103] R. Jing, F. Duan, F. Lu, M. Zhang, and W. Zhao, “Denoising diffusion probabilistic feature-based network for cloud removal in sentinel-2 imagery,” *Remote Sens.*, vol. 15, no. 9, p. 2217, 2023.
- [104] X. Zou, K. Li, J. Xing, Y. Zhang, S. Wang, L. Jin, and P. Tao, “Diffcr: A fast conditional diffusion framework for cloud removal from optical satellite images,” *IEEE Trans. on Geosci. and Remote Sens.*, vol. 62, pp. 1–14, 2024.
- [105] X. Zhao and K. Jia, “Cloud removal in remote sensing using sequential-based diffusion models,” *Remote Sens.*, vol. 15, no. 11, p. 2861, 2023.
- [106] B. Rasti, Y. Chang, E. Dalsasso, L. Denis, and P. Ghamisi, “Image restoration for remote sensing: Overview and toolbox,” *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 2, pp. 201–230, 2021.
- [107] Y. Huang and S. Xiong, “Remote sensing image dehazing using adaptive region-based diffusion models,” *IEEE Geosci. Remote Sens. Lett.*, 2023.
- [108] Y. Huang, Z. Lin, S. Xiong, and T. Sun, “Diffusion models based null-space learning for remote sensing image dehazing,” *IEEE Geosci. Remote Sens. Lett.*, 2024.
- [109] Y. Zhu, L. Wang, J. Yuan, and Y. Guo, “Diffusion model based low-light image enhancement for space satellite,” *arXiv preprint arXiv:2306.14227*, 2023.
- [110] K. Deng, Z. Jiang, Q. Qian, Y. Qiu, and Y. Qian, “A noise-model-free hyperspectral image denoising method based on diffusion model,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, 2023, pp. 7308–7311.
- [111] J. He, Y. Li, Q. Yuan *et al.*, “Tdiffde: A truncated diffusion model for remote sensing hyperspectral image denoising,” *arXiv preprint arXiv:2311.13622*, 2023.
- [112] L. Pang, X. Rui, L. Cui, H. Wang, D. Meng, and X. Cao, “Hir-diff: Unsupervised hyperspectral image restoration via improved diffusion models,” *arXiv preprint arXiv:2402.15865*, 2024.
- [113] W. Yu, Y. Xu, and P. Ghamisi, “Universal adversarial defense in remote sensing based on pre-trained denoising diffusion models,” *arXiv preprint arXiv:2307.16865*, 2023.
- [114] G. Fracastoro, E. Magli, G. Poggi, G. Scarpa, D. Valsesia, and L. Verdoliva, “Deep learning methods for synthetic aperture radar image despeckling: An overview of trends and perspectives,” *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 2, pp. 29–51, 2021.
- [115] M. V. Perera, N. G. Nair, W. G. C. Bandara, and V. M. Patel, “Sar despeckling using a denoising diffusion probabilistic model,” *IEEE Geosci. Remote Sens. Lett.*, 2023.
- [116] S. Xiao, L. Huang, and S. Zhang, “Unsupervised sar despeckling based on diffusion model,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, 2023, pp. 810–813.
- [117] S. Guha and S. T. Acton, “Sddpm: Speckle denoising diffusion probabilistic models,” *arXiv preprint arXiv:2311.10868*, 2023.
- [118] T. Amit, T. Shaharabany, E. Nachmani, and L. Wolf, “Segdiff: Image segmentation with diffusion probabilistic models,” *arXiv preprint arXiv:2112.00390*, 2021.
- [119] C. Ayala, R. Sesma, C. Aranda, and M. Galar, “Diffusion models for remote sensing imagery semantic segmentation,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, 2023, pp. 5654–5657.
- [120] B. Kolbeinsson and K. Mikolajczyk, “Multi-class segmentation from aerial views using recursive noise diffusion,” in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, 2024, pp. 8439–8449.
- [121] N. Chen, J. Yue, L. Fang, and S. Xia, “Spectraldiff: A generative framework for hyperspectral image classification with diffusion models,” *IEEE Trans. on Geosci. and Remote Sens.*, 2023.
- [122] J. Zhou, J. Sheng, J. Fan, P. Ye, T. He, B. Wang, and T. Chen, “When hyperspectral image classification meets diffusion models: An unsupervised feature learning framework,” *arXiv preprint arXiv:2306.08964*, 2023.
- [123] D. Li, W. Xie, J. Zhang, and Y. Li, “Mdfi: Multi-domain diffusion-driven feature learning,” in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 8, 2024, pp. 8653–8660.
- [124] D. Li, W. Xie, Z. Wang, Y. Lu, Y. Li, and L. Fang, “Feddiff: Diffusion model driven federated learning for multi-modal and multi-clients,” *arXiv preprint arXiv:2401.02433*, 2023.
- [125] J. Qu, Y. Yang, W. Dong, and Y. Yang, “Lds2ae: Local diffusion shared-specific autoencoder for multimodal remote sensing image classification with arbitrary missing modalities,” in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 13, 2024, pp. 14 731–14 739.
- [126] J. Chen, S. Liu, Z. Zhang, and H. Wang, “Diffusion subspace clustering for hyperspectral images,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, 2023.
- [127] J. Ma, W. Xie, Y. Li, and L. Fang, “Bsdm: Background suppression diffusion model for hyperspectral anomaly detection,” *arXiv preprint arXiv:2307.09861*, 2023.
- [128] Y. Wen, X. Ma, X. Zhang, and M.-O. Pun, “Gcd-ddpm: A generative change detection model based on difference-feature guided ddpm,” *IEEE Trans. on Geosci. and Remote Sens.*, 2024.
- [129] A. Singh, “Review article digital change detection techniques using remotely-sensed data,” *Int. J. Remote Sens.*, vol. 10, no. 6, pp. 989–1003, 1989.
- [130] J. Tian, J. Lei, J. Zhang, W. Xie, and Y. Li, “Swimdiff: Scene-wide matching contrastive learning with diffusion constraint for remote sensing image,” *IEEE Trans. on Geosci. and Remote Sens.*, 2024.
- [131] X. Chen, H. Fan, R. Girshick, and K. He, “Improved baselines with momentum contrastive learning,” *arXiv preprint arXiv:2003.04297*, 2020.
- [132] X. Zhang, S. Tian, G. Wang, H. Zhou, and L. Jiao, “Diffucd: Unsupervised hyperspectral image change detection with semantic correlation diffusion model,” *arXiv preprint arXiv:2305.12410*, 2023.
- [133] F. Bao, S. Nie, K. Xue, Y. Cao, C. Li, H. Su, and J. Zhu, “All are worth words: A vit backbone for diffusion models,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 22 669–22 679.
- [134] J. Jia, G. Lee, Z. Wang, L. Zhi, and Y. He, “Siamese meets diffusion network: Smdnet for enhanced change detection in high-resolution rs imagery,” *arXiv preprint arXiv:2401.09325*, 2024.
- [135] P. Nath, P. Shukla, and C. Quilodrán-Casas, “Forecasting tropical cyclones with cascaded diffusion models,” *arXiv preprint arXiv:2310.01690*, 2023.
- [136] Y. Hatanaka, Y. Glaser, G. Galgon, G. Torri, and P. Sadowski, “Diffusion models for high-resolution solar forecasts,” *arXiv preprint arXiv:2302.00170*, 2023.
- [137] J. Leinonen, U. Hamann, D. Nerini, U. Germann, and G. Franch, “Latent diffusion models for generative precipitation nowcasting with accurate uncertainty quantification,” *arXiv preprint arXiv:2304.12891*, 2023.
- [138] I. Corley and P. Najafirad, “Single-view height estimation with conditional diffusion probabilistic models,” *arXiv preprint arXiv:2304.13214*, 2023.
- [139] Y. Jian, F. Yu, S. Singh, and D. Stamoulis, “Stable diffusion for aerial object detection,” in *NeurIPS 2023 Workshop on Synthetic Data Generation with Generative AI*, 2023.
- [140] A. Awasthi, S. Ly, J. Nizam, S. Zare, V. Mehta, S. Ahmed, K. Shah, R. Nemani, S. Prasad, and H. Van Nguyen, “Anomaly detection in satellite videos using diffusion models,” *arXiv preprint arXiv:2306.05376*, 2023.
- [141] S. Jozdani, D. Chen, D. Pouliot, and B. A. Johnson, “A review and meta-analysis of generative adversarial networks and their applications in remote sensing,” *Int. J. Appl. Earth Observ. Geoinf.*, vol. 108, p. 102734, 2022.
- [142] Y. Chen, Q. Weng, L. Tang, X. Zhang, M. Bilal, and Q. Li, “Thick clouds removing from multitemporal landsat images using spatiotem-

- poral neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2020.
- [143] A. Meraner, P. Ebel, X. X. Zhu, and M. Schmitt, "Cloud removal in sentinel-2 imagery using a deep residual neural network and sar-optical data fusion," *ISPRS-J. Photogramm. Remote Sens.*, vol. 166, pp. 333–346, 2020.
- [144] X. Zou, K. Li, J. Xing, P. Tao, and Y. Cui, "Pmaa: A progressive multi-scale attention autoencoder model for high-performance cloud removal from multi-temporal satellite imagery," in *Proc. Eur. Conf. Artif. Intell. (ECAI)*, 2023.
- [145] P. Ebel, V. S. F. Garnot, M. Schmitt, J. D. Wegner, and X. X. Zhu, "Uncrtaints: Uncertainty quantification for cloud removal in optical satellite time series," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 2086–2096.
- [146] D. Hong, Z. Han, J. Yao, L. Gao, B. Zhang, A. Plaza, and J. Chanussot, "Spectralformer: Rethinking hyperspectral image classification with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2021.
- [147] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5966–5978, 2020.
- [148] L. Sun, G. Zhao, Y. Zheng, and Z. Wu, "Spectral-spatial feature tokenization transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.
- [149] B. Liu, A. Yu, X. Yu, R. Wang, K. Gao, and W. Guo, "Deep multiview learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7758–7772, 2020.
- [150] D. Wang, B. Du, and L. Zhang, "Spectral-spatial global graph reasoning for hyperspectral image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, 2023.
- [151] R. C. Daudt, B. Le Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, 2018, pp. 4063–4067.
- [152] H. Chen and Z. Shi, "A spatial-temporal attention-based method and a new dataset for remote sensing image change detection," *Remote Sens.*, vol. 12, no. 10, p. 1662, 2020.
- [153] S. Fang, K. Li, J. Shao, and Z. Li, "Snunet-cd: A densely connected siamese network for change detection of vhr images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2021.
- [154] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2021.
- [155] W. G. C. Bandara and V. M. Patel, "A transformer-based siamese network for change detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, 2022, pp. 207–210.
- [156] V. Sarukkai, A. Jain, B. UzKent, and S. Ermon, "Cloud removal from satellite images using spatiotemporal generator networks," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, 2020, pp. 1796–1805.
- [157] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.
- [158] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 586–595.
- [159] D. Dowson and B. Landau, "The fréchet distance between multivariate normal distributions," *J. Multivariate Anal.*, vol. 12, no. 3, pp. 450–455, 1982.
- [160] H. Chen and Z. Shi, "A spatial-temporal attention-based method and a new dataset for remote sensing image change detection," *Remote Sens.*, vol. 12, no. 10, p. 1662, 2020.
- [161] S. Chen, P. Sun, Y. Song, and P. Luo, "Diffusiondet: Diffusion model for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 19830–19843.
- [162] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 740–755.
- [163] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 4195–4205.
- [164] S. Mo, E. Xie, R. Chu, L. Hong, M. Niessner, and Z. Li, "Dit-3d: Exploring plain diffusion transformers for 3d shape generation," *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 36, 2024.
- [165] J. Chen, J. Yu, C. Ge, L. Yao, E. Xie, Y. Wu, Z. Wang, J. Kwok, P. Luo, H. Lu *et al.*, "Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis," *arXiv preprint arXiv:2310.00426*, 2023.
- [166] N. Ma, M. Goldstein, M. S. Albergo, N. M. Boffi, E. Vanden-Eijnden, and S. Xie, "Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers," *arXiv preprint arXiv:2401.08740*, 2024.
- [167] A. Tuel, T. Kerdreux, C. Hulbert, and B. Rouet-Leduc, "Diffusion models for interferometric satellite aperture radar," *arXiv preprint arXiv:2308.16847*, 2023.
- [168] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Deep learning for hyperspectral image classification: An overview," *IEEE Trans. on Geosci. and Remote Sens.*, vol. 57, no. 9, pp. 6690–6709, 2019.
- [169] N. M. Nasrabadi, "Hyperspectral target detection: An overview of current and future challenges," *IEEE Signal Process. Mag.*, vol. 31, no. 1, pp. 34–44, 2013.
- [170] Sneha and A. Kaul, "Hyperspectral imaging and target detection algorithms: a review," *Multimedia Tools Appl.*, vol. 81, no. 30, pp. 44 141–44 206, 2022.
- [171] S. Chen, H. Wang, F. Xu, and Y.-Q. Jin, "Target classification using the deep convolutional networks for sar images," *IEEE Trans. on Geosci. and Remote Sens.*, vol. 54, no. 8, pp. 4806–4817, 2016.
- [172] Y. Zhang and Y. Hao, "A survey of sar image target detection based on convolutional neural networks," *Remote Sens.*, vol. 14, no. 24, p. 6240, 2022.
- [173] K. Ding, T. Lu, W. Fu, S. Li, and F. Ma, "Global-local transformer network for hsi and lidar data joint classification," *IEEE Trans. on Geosci. and Remote Sens.*, vol. 60, pp. 1–13, 2022.
- [174] S. Kahraman and R. Bacher, "A comprehensive review of hyperspectral data fusion with lidar and sar data," *Annu. Rev. Control.*, vol. 51, pp. 236–253, 2021.
- [175] A. Bhardwaj, L. Sam, A. Bhardwaj, and F. J. Martín-Torres, "Lidar remote sensing of the cryosphere: Present applications and future prospects," *Remote Sens. Environ.*, vol. 177, pp. 125–143, 2016.
- [176] D. Xu, H. Wang, W. Xu, Z. Luan, and X. Xu, "Lidar applications to estimate forest biomass at individual tree scale: Opportunities, challenges and future perspectives," *Forests*, vol. 12, no. 5, p. 550, 2021.
- [177] C. Webster, G. Mazzotti, R. Essery, and T. Jonas, "Enhancing airborne lidar data for improved forest structure representation in shortwave transmission models," *Remote Sens. Environ.*, vol. 249, p. 112017, 2020.
- [178] M. Mahdianpari, J. E. Granger, F. Mohammadimanesh, S. Warren, T. Puestow, B. Salehi, and B. Brisco, "Smart solutions for smart cities: Urban wetland mapping using very-high resolution satellite imagery and airborne lidar data in the city of st. john's, nl, canada," *J. Environ. Manage.*, vol. 280, p. 111676, 2021.
- [179] X. Wang and P. Li, "Extraction of urban building damage using spectral, height and corner information from vhr satellite images and airborne lidar data," *ISPRS-J. Photogramm. Remote Sens.*, vol. 159, pp. 322–336, 2020.
- [180] G. K. Nakayama, M. A. Uy, J. Huang, S.-M. Hu, K. Li, and L. Guibas, "Diffacto: Controllable part-based 3d point cloud generation with cross diffusion," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 14 257–14 267.
- [181] L. Melas-Kyriazi, C. Rupprecht, and A. Vedaldi, "Pc2: Projection-conditioned point cloud diffusion for single-image 3d reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 12 923–12 932.
- [182] L. Abady, E. Cannas, P. Bestagini, B. Tondi, S. Tubaro, M. Barni *et al.*, "An overview on the generation and detection of synthetic and manipulated satellite images," *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, 2022.
- [183] X. Xie, G. Cheng, Q. Li, S. Miao, K. Li, and J. Han, "Fewer is more: Efficient object detection in large aerial images," *Sci. China Inform. Sci.*, vol. 67, no. 1, pp. 1–19, 2024.
- [184] X. Gu, P. P. Angelov, C. Zhang, and P. M. Atkinson, "A semi-supervised deep rule-based approach for complex satellite sensor image analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2281–2292, 2020.
- [185] X. Zhu, J. Zhu, H. Li, X. Wu, H. Li, X. Wang, and J. Dai, "Uniperceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 16 804–16 815.

- [186] J. Zhu, X. Zhu, W. Wang, X. Wang, H. Li, X. Wang, and J. Dai, "Uniperceiver-moe: Learning sparse generalist models with conditional moes," *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 35, pp. 2664–2678, 2022.
- [187] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 33, pp. 1877–1901, 2020.
- [188] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *Proc. Int. Conf. Mach. Learn. (ICML)*. PMLR, 2021, pp. 8821–8831.
- [189] X. Sun, P. Wang, W. Lu, Z. Zhu, X. Lu, Q. He, J. Li, X. Rong, Z. Yang, H. Chang *et al.*, "Ringmo: A remote sensing foundation model with masked image modeling," *IEEE Trans. on Geosci. and Remote Sens.*, 2022.
- [190] Y. Hu, J. Yuan, C. Wen, X. Lu, and X. Li, "Rsgpt: A remote sensing vision language model and benchmark," *arXiv preprint arXiv:2307.15266*, 2023.
- [191] X. Guo, J. Lao, B. Dang, Y. Zhang, L. Yu, L. Ru, L. Zhong, Z. Huang, K. Wu, D. Hu *et al.*, "Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery," *arXiv preprint arXiv:2312.10115*, 2023.
- [192] W. Zhang, M. Cai, T. Zhang, Y. Zhuang, and X. Mao, "Earthgpt: A universal multi-modal large language model for multi-sensor image comprehension in remote sensing domain," *arXiv preprint arXiv:2401.16822*, 2024.
- [193] Y. Zhao, Y. Xu, Z. Xiao, and T. Hou, "Mobilediffusion: Subsecond text-to-image generation on mobile devices," *arXiv preprint arXiv:2311.16567*, 2023.
- [194] Y.-H. Chen, R. Sarokin, J. Lee, J. Tang, C.-L. Chang, A. Kulik, and M. Grundmann, "Speed is all you need: On-device acceleration of large diffusion models via gpu-aware optimizations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 4650–4654.
- [195] D. Watson, W. Chan, J. Ho, and M. Norouzi, "Learning fast samplers for diffusion models by differentiating through sample quality," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.
- [196] D. Watson, J. Ho, M. Norouzi, and W. Chan, "Learning to efficiently sample from diffusion probabilistic models," *arXiv preprint arXiv:2106.03802*, 2021.
- [197] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, "Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps," *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 35, pp. 5775–5787, 2022.
- [198] Q. Zhang and Y. Chen, "Fast sampling of diffusion models with exponential integrator," *arXiv preprint arXiv:2204.13902*, 2022.
- [199] X. Liu, X. Zhang, J. Ma, J. Peng *et al.*, "Instaflo: One step is enough for high-quality diffusion-based text-to-image generation," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2024.
- [200] S. Luo, Y. Tan, L. Huang, J. Li, and H. Zhao, "Latent consistency models: Synthesizing high-resolution images with few-step inference," *arXiv preprint arXiv:2310.04378*, 2023.
- [201] Y. Xu, Y. Zhao, Z. Xiao, and T. Hou, "Ufogen: You forward once large scale text-to-image generation via diffusion gans," *arXiv preprint arXiv:2311.09257*, 2023.
- [202] A. Kodaira, C. Xu, T. Hazama, T. Yoshimoto, K. Ohno, S. Mitsuohori, S. Sugano, H. Cho, Z. Liu, and K. Keutzer, "Streamdiffusion: A pipeline-level solution for real-time interactive generation," *arXiv preprint arXiv:2312.12491*, 2023.
- [203] C. Meng, R. Rombach, R. Gao, D. Kingma, S. Ermon, J. Ho, and T. Salimans, "On distillation of guided diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 14 297–14 306.
- [204] S. Lin, A. Wang, and X. Yang, "Sdxl-lightning: Progressive adversarial diffusion distillation," *arXiv preprint arXiv:2402.13929*, 2024.
- [205] S. Li, T. Hu, F. S. Khan, L. Li, S. Yang, Y. Wang, M.-M. Cheng, and J. Yang, "Faster diffusion: Rethinking the role of unet encoder in diffusion models," *arXiv preprint arXiv:2312.09608*, 2023.
- [206] S. Calvo-Ordóñez, J. Huang, L. Zhang, G. Yang, C.-B. Schonlieb, and A. I. Aviles-Rivero, "Beyond u: Making diffusion models faster & lighter," *arXiv preprint arXiv:2310.20092*, 2023.
- [207] Y. Fan, H. Liao, S. Huang, Y. Luo, H. Fu, and H. Qi, "A survey of emerging applications of diffusion probabilistic models in mri," *arXiv preprint arXiv:2311.11383*, 2023.
- [208] J. M. Haut, M. E. Paoletti, S. Moreno-Álvarez, J. Plaza, J.-A. Rico-Gallego, and A. Plaza, "Distributed deep learning for remote sensing data interpretation," *Proc. IEEE*, vol. 109, no. 8, pp. 1320–1349, 2021.
- [209] R. Sedona, G. Cavallaro, J. Jitsev, A. Strube, M. Riedel, and J. A. Benediktsson, "Remote sensing big data classification with high performance distributed deep learning," *Remote Sens.*, vol. 11, no. 24, p. 3056, 2019.
- [210] Z. M. Fadlullah and N. Kato, "On smart iot remote sensing over integrated terrestrial-aerial-space networks: An asynchronous federated learning approach," *IEEE Netw.*, vol. 35, no. 5, pp. 129–135, 2021.
- [211] D. Li, W. Xie, Y. Li, and L. Fang, "Fedfusion: Manifold driven federated learning for multi-satellite and multi-modality fusion," *IEEE Trans. Geosci. Remote Sens.*, 2023.
- [212] M. Li, T. Cai, J. Cao, Q. Zhang, H. Cai, J. Bai, Y. Jia, M.-Y. Liu, K. Li, and S. Han, "Distrifusion: Distributed parallel inference for high-resolution diffusion models," *arXiv preprint arXiv:2402.19481*, 2024.
- [213] Y. Xu, T. Bai, W. Yu, S. Chang, P. M. Atkinson, and P. Ghamisi, "Ai security for geoscience and remote sensing: Challenges and future trends," *IEEE Geosci. Remote Sens. Mag.*, vol. 11, no. 2, pp. 60–85, 2023.