Siamese Meets Diffusion Network: SMDNet for Enhanced Change Detection in High-Resolution RS Imagery

Jia Jia, Geunho Lee, Zhibo Wang, Zhi Lyu, and Yuchu He

***Abstract*—In recent years, the application of deep learning to change detection (CD) has significantly progressed in remote sensing images. CD tasks have mostly used architectures, such as CNN and Transformer to locate image changes. However, these architectures have shortcomings in representing boundary details and are prone to false alarms and missed detections under complex lighting and weather conditions. For that, we propose a new network, Siamese meets diffusion network (SMDNet), a CD model that combines discriminative and generative architecture. By leveraging the power of the Siam-U2Net feature differential encoder (SU-FDE) and denoising diffusion implicit model (DDIM), it not only improves the accuracy of object edge detection but also enhances the data through iterative denoising and thinning reconstruction detail detection accuracy. Improves the model's robustness under environmental changes. First, we propose an SU-FDE module that uses shared weight features to capture differences between time series images, refine edge detection, and combine it with the attention mechanism to identify vital coarse features, thereby improving model sensitivity and accuracy. Finally, the progressive sampling of DDIM is used to integrate further these key features, and the adaptability of the model in different environments is enhanced with the help of the denoising ability of the diffusion model and the accurate capture of the probability distribution of image data. The performance evaluation of SMDNet on LEVIR-CD, DSIFN-CD, and CDD datasets yields validated F1 scores of 89.17%, 88.48%, and 88.23%, respectively. This substantiates the advanced capabilities of our model in accurately identifying variations and intricate details.**

***Index Terms*—Change detection (CD), deep learning, diffusion model (DM), remote sensing (RS), Siamese network.**

I. Introduction

REMOTE sensing (RS) [1], [2], [3], [4], [5] technology has seen rapid advancements, fueled by Earth Observation System projects launched by nations and international coalitions, utilizing satellites like Landsat, Gaofen, SPOT, RADARSAT, and Sentinel. These advancements have led to the generation of high-resolution RS images, which provide detailed ground information for various applications. Progress in multimodal high-resolution data analysis [6] and spectral RS image processing [7] has further expanded our capabilities to understand and apply RS imagery for detailed monitoring of the Earth's surface. RS image analysis [8], [9], particularly for CD, allows us to compare alterations in objects or phenomena over different time intervals within a consistent location. It is crucial for tracking temporal changes in specific locations, proving essential in environmental monitoring, disaster assessment, and urban planning. This process generates extensive data that deepens our semantic understanding across disciplines and significantly improves our grasp of the Earth's dynamic landscape and ongoing evolution.

Although CD technology in RS images [10], [11] has made rapid progress, many problems still need to be solved in practical applications, especially when dealing with high-resolution imagery. Issues, such as object occlusion, spectral confusion from elements with similar spectral characteristics (e.g., trees, buildings) [12], and environmental factors like climate and lighting variations persist. These challenges impede the accurate identification of distractors and the extraction of precise edge details in high-resolution RS images [13], complicating the reliable extraction of segmentation maps before and after changes.

Traditionally, CD has evolved from manual visual analysis to using algebraic techniques to calculate image pixel differences [14], [15] to employing data reduction strategies, such as principal component analysis (PCA) [16]. With the rise in image resolution, notably through CNN [17] and Transformer [18] in deep learning, they have become the mainstream methods to improve the accuracy and efficiency of RS image CD. However, there is still room for improvement in feature processing and detailed description.

The success of U-Net [19] in medical image segmentation has inspired its application in RS image CD. The efficient multiresolution network (EMRN) [20] addresses inconsistencies in multiresolution images. UNet++ [21] enhances information retention with its encoder–decoder architecture. Siamese NestedUNet (SNUNet)-CD [22] merges a Siamese network with UNet++ to blend deep and shallow features effectively, while STANet [23] advances spatiotemporal analysis through attention mechanisms. Despite these advancements, CNNs

Manuscript received 15 January 2024; revised 20 March 2024; accepted 31 March 2024. Date of publication 3 April 2024; date of current version 17 April 2024. *(Corresponding author: Geunho Lee.)*

Jia Jia is with the Department of Culture Technology (Artificial Intelligence Direction), Jeonju University, Jeonju 55069, South Korea (e-mail: kimmit628@jj.ac.kr).

Geunho Lee is with the Department of Artificial Intelligence, Jeonju University, Jeonju 55069, South Korea (e-mail: ghlee@jj.ac.kr).

Zhibo Wang and Zhi Lyu are with the Department of Graduate School of Artificial Intelligence, Jeonju University, Jeonju-si 55069, South Korea (e-mail: wang970128@jj.ac.kr; lyuzhi@jj.ac.kr).

Yuchu He is with the Department of Information Science and Technology College, Zhengzhou Normal University, Zhengzhou 450044, China (e-mail: heyuchu@foxmail.com).

Digital Object Identifier 10.1109/JSTARS.2024.3384545

continue to face challenges in processing multitemporal images within dynamic and complex settings.

The Transformer architecture was originally used in natural language processing and is now widely used in the CD field of RS. Models like the bitemporal image transformer (BIT) [23] leverage CNN and Transformer synergy for enhanced change recognition. PBSL [24] and Changeformer [25] refine feature focus and information capture for CD, while SwinSUNet [26] and CTS-Unet [27] integrate Siamese structures for improved global information capture capabilities with low-resolution images and temporal relationships. Despite its proficiency in global information processing, the Transformer architecture faces computational and large-scale data challenges. We discovered the advantages of Siamese networks in CNN and Transformer to effectively analyze image pairs and adapt to the complex spatiotemporal relationships in remote sensing image CD through a shared weight mechanism.

Recent attention has turned toward diffusion models (DMs) [28], [29], [30], [31], [32], advanced generative models that produce intricate and lifelike images by systematically reversing added noise. Their main advantage is the ability to create extremely realistic images due to their ability to process complex patterns and their powerful learning capabilities to simulate and reconstruct the subtle structure of the data. Unlike traditional models, DMs excel in generating images with unparalleled clarity and diversity, marking them as formidable tools in fields, such as medicine [33], art, and media production. Notable applications include virtual dress-up [34], pose enhancement [35], and video creation [36], [37]. Moreover, the authors in [38] highlighted DMs' efficiency in discriminative tasks–image segmentation [39], [40], [41], [42], classification [43], and anomaly detection [44], [45] underscoring their widespread applicability. This breadth of use signals considerable potential for future explorations in DM applications.

Despite the nascent application of DMs in RS image processing, notable efforts have begun to explore their utility in extracting essential semantic features [46]. Inspired by the success of DDPM in multiple fields, we propose the Siamese meets diffusion network (SMDNet) model. This model combines Siam-U2Net feature differential encoder (SU-FDE) and the denoising diffusion implicit model (DDIM) to innovatively fuse discriminative learning and generation processes into the change detection (CD) task to improve edge detection accuracy in CD and robustness under different environmental conditions. This hybrid architecture leverages the advantages of both models in feature extraction and image generation to achieve more refined recognition and reconstruction of changed areas. Among them, the feature differential encoder SU-FDE is used to improve the accuracy of edge description and is combined with Siamese network contrastive learning to analyze the similarity and difference of image pairs. In addition, multiscale information fusion using nested U2Net improves edge detail description, and identifies and enhances key coarse features through the spatial attention (SA) [47] module. Subsequently, the key feature maps are introduced into the encoder of the diffusion model to enhance the robustness to illumination and climate changes. DMs stepwise sampling integrates key features effectively and iteratively generates more accurate CD maps. The main contributions of our work are summarized as follows.

1) We propose SMDNet, a new model that combines Siamese encoder and DDIM architecture to fuse discriminative learning and generative processes in CD tasks. This model not only optimizes the quality of CD maps but also efficiently identifies changing areas.
2) SMDNet's SU-FDE module adopts a deeply nested U-shaped structure and multiscale feature extraction technology. It uses the characteristics of shared weights to enhance the model's ability to identify spatial correlation and difference and improve edge detail detection capabilities.
3) SMDNet uses denoising U-Net (DU) to learn pixel distribution under different lighting and weather conditions to improve model robustness and achieve good F1 scores of 89.17%, 88.48%, and 88.23% on three public datasets.

The rest of this article is as follows: Section II discusses related work in DL and DMs applied to CD. Section III details the methodology we propose. Section IV covers the experimental setup and analysis of results. Finally, Section V concludes this article.

II. Related Work

A. Deep Learning-Based CD Methods

With the swift advancement of deep learning technology in recent years, its strong potential in CD in RS images has received extensive research attention. In particular, the ability of CNN in pixel-level CD is outstanding in learning feature representations and identifying changed areas. In the realm of CD, the approaches can be categorized into two types: 1) single-stream and 2) dual-stream network methods.

Single-Stream Network: A singular deep learning network structure is employed to learn the evolving features between dual-temporal images. In this approach, a fully convolutional neural network (FCNN) [48] primarily handles demanding prediction tasks. To achieve this, two dual-temporal images before and after changes are concatenated into a single input image, and the CD feature map is extracted using a convolutional network. Recognizing that the FCNN's sampling process may result in information loss and weak global perception ability, the method incorporates U-Net [19]. This involves merging dual-temporal images into a single image, fed into the network with modifications. A recurrent neural network is introduced in the skip connection to enhance responsiveness to temporal changes in perception ability. PBSL [24] introduces a multimodal alignment approach to highlight relevant features and suppress irrelevant information. To capture global and edge detail information, the UNet++ network [21], [49] employs dense skip connections in the encoder for CD tasks, enhancing the segmentation network. The method adopts a deep supervision strategy to improve detection in high-resolution images. Nevertheless, it encounters challenges in understanding temporal data relationships, limiting its CD capability.

Dual-Stream Network: The approach involves processing images from two temporal nodes through two parallel neural networks, establishing a coupled architecture to discern features

with significant material relationships, thus improving CD map accuracy. Notably, FMCD [50] integrates feature encoding, CD, and domain adaptation tasks within a cohesive two-stream framework. Similarly, WS-Net [51] utilizes wavelet transform to analyze spatial and frequency domain differences, while MLA-Net [52] introduces a mask-guided attention mechanism to enhance detection accuracy. In contrast to a single network structure, the dual-stream network is divided into an asymmetric dual-stream network (pseudo-Siamese) and a Siamese network, each handling inputs from different time points. The asymmetric network learns distinct features with independent weights, whereas the Siamese network compares input similarities with shared weights adapted to different time series data analysis tasks. In addition, recurrent convolutional neural networks [53] integrate CNNs with recurrent neural networks to form a pseudo-Siamese network, and the dual-task constrained deep Siamese convolutional network [54] utilizes fully convolutional networks (FCNs) to extract multilevel features. Addressing the lack of sufficient supervision for change feature learning, an intensely supervised image fusion network (IFN) [55] is proposed using a pretrained VGG16 as an encoder through the attention module of the dual-stream architecture. In addition, TCRL [56] introduces a contrastive learning method, and the densely connected Siamese network (SNUNet) [22] merges Siamese network principles with UNet++, further enriching the landscape of dual-stream network methodologies.

B. Diffusion Model

DDPM as a generative model, is focused on reconstructing data by inversely simulating the diffusion process of data from its true distribution to the noise distribution. During training, the model gradually masters the process of recovering from the noisy state to the original data. Compared with traditional deep learning frameworks, such as single-stream or dual-stream networks, DDPM demonstrates multiple advantages: they can capture pixel changes under different time and meteorological conditions and enhance key features of the image when reconstructing the real data, which improves image contrast and clarity [57]. DDPM has shown its superiority in synthesizing and recovering high-quality images [58]. In remote sensing image analysis, diffusion models have proven effective, especially in enhancing image representation and detail supplementation [59], [60], [61]. Furthermore, the DM also demonstrates its utility in cloud removal [62], [63], [64] and image segmentation [65] tasks. In the field of CD in RS images, these models effectively distinguish real changes from pseudo-changes due to noise through an iterative denoising process, thus improving the detection accuracy of details and edges. Although there are not many cases of utilizing DMs in research literature on RS image CD, recent innovative research has started investigating the application of DDPM in processing RS images, especially in feature extraction and construction of pretrainer encoders for large-scale RS data [66]. Inspired by the successful application of DDPM in other fields, this study proposes an innovative method for RS image CD: the Siamese U2Net denoising diffusion implicit model (SMDNet).

III. Method

This section will introduce the proposed SMDNet network. Initially, the overall architecture of the network is presented. The proposed feature differential encoder module (SU-FDE) is described in detail. Next, is an explanation of the added attention mechanism. Finally, DDIM is briefly explained.

A. Framework Overview

Fig. 1 mainly comprises an SU-FDE and a denoising module (denosing U-net). SU-FDE is a bitemporal U2-Net feature differential encoder. Different from traditional CD methods and typical deep learning methods, DMs learn the process of denoising and learn from the noise to produce clear detection results. Since high-resolution remote sensing images have richer texture and geometric semantic information than standard optical remote sensing images, there is a higher demand for more advanced feature extraction capabilities.

First, the dual-temporal image pairs (T1, T2) are input into SU-FDE. This module utilizes the U-shaped structure and the shared weight characteristics of the Siamese network to effectively extract and aggregate the multiscale features of the image pairs, enhancing the spatial correlation and difference of the model pairs' recognition ability. Next, SA is used to enhance key spatial features, while L_1 distance is used to calculate the difference between feature maps at different scales to obtain the difference map ($\widehat{f}_i$). Further improve the detection accuracy of edge details. The noise label map GT_t is obtained by adding t step noise to the binary classification (changed/unchanged) label GT_0. Bitemporal T1 and T2 and the added noise GT_t are combined along the channel dimension before being input to the encoder of denoise-UNet (DU) to extract multiscale features (I_S). The model can learn from images before and after changes and enhance its denoising ability by processing the noisy label map GT_t, allowing the encoder to obtain a more comprehensive feature perspective at multiple scales. Since $\widehat{f}_i$ and I_S have the same number of features, we combine their corresponding scale features to obtain fused features. Finally, the obtained fusion features are input into the decoder of DU to obtain the prediction result $\widehat{\mathrm{GT}_0} \in \mathbb{R}^{C*W*H}$

$$\widehat{\mathrm{GT}_0} = \mathrm{DU}(\mathrm{cat}(T1, T2, \mathrm{GT}_t), t, \widehat{f}_i). \tag{1}$$

Here $\mathrm{cat}(\cdot)$ is the concatenation operator.

The CD task is to detect changed and unchanged areas between pixels in an image. This is a discrete binary classification task, and in CD tasks, changed areas usually look much smaller than unchanged areas. Therefore, simultaneously using Dice Loss and BCE Loss can improve the model's sensitivity to boundary pixels and overall pixel classification accuracy. The total loss here is

$$L_{\text{total}} = L_{\text{dice}}(\widehat{\mathrm{GT}_0}, \mathrm{GT}_0) + L_{\text{bce}}(\widehat{\mathrm{GT}_0}, \mathrm{GT}_0). \tag{2}$$

Here, L_{dice} is dice loss, L_{bce} is BCE loss, $\widehat{\mathrm{GT}_0}$ is the predicted label value, and GT_0 is the ground true label value.

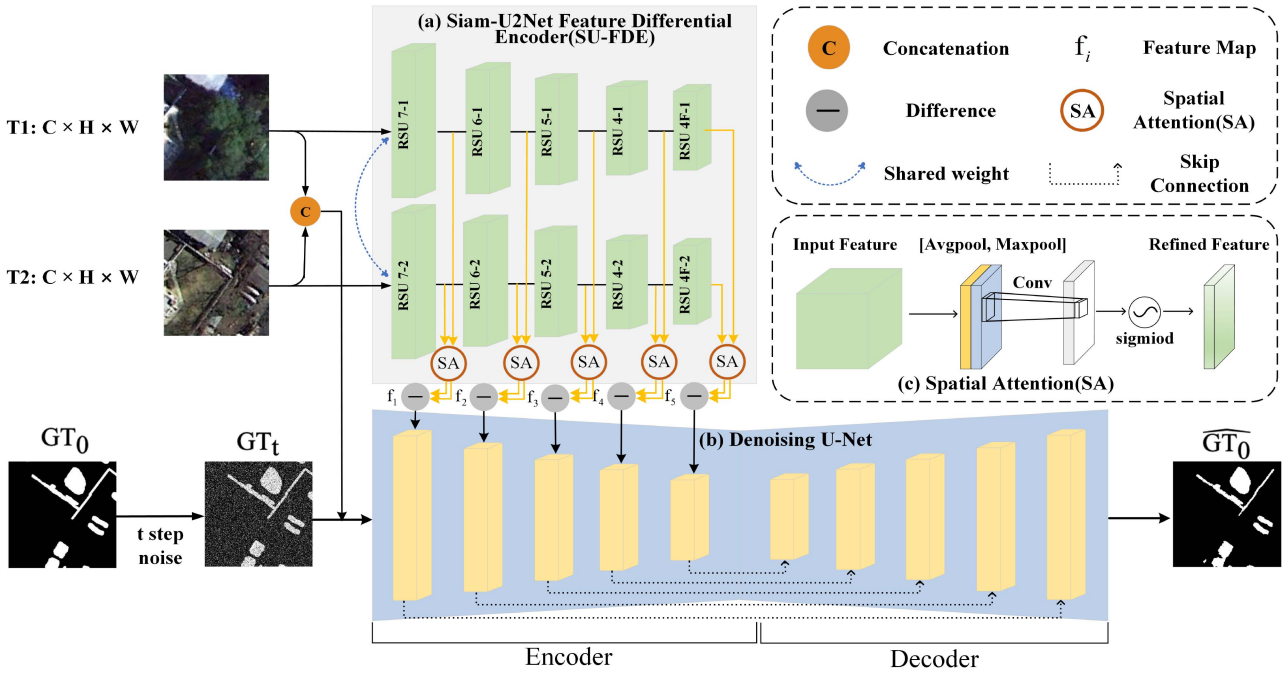

Fig. 1. Structural layout of the SMDNet network design. T1, T2, and GT_0 are the prechange, postchange, and labeled images in the CDD dataset. (a) Proposed SU-FDE. (b) DU for denoising. (c) SA Mechanism. In the model, SU-FDE extracts and processes spatiotemporal features from dual-temporal image pairs uses SA to emphasize important regions, and calculates differential feature maps. After DU fuses the noise label map and the difference map, it gradually samples and iterates the prediction results.

B. Siam-U2Net Feature Differential Encoder

In the current RS image CD research domain, dealing with complex geographical environments and diverse ground object types (Such as roads, buildings, plants, lakes, etc.) is a crucial challenge. To overcome this challenge, researchers usually employ various advanced techniques, such as feature pyramid fusion, inception modules, skip dense connections, and residual connections, to improve the performance of image detail capture and edge detection. In this context, we introduced the Siamese network and built a lightweight feature extractor named SU-FDE. This feature extractor aims to enhance the understanding of similarities and differences in images to improve the ability of multiscale information fusion and edge detection. We achieved satisfactory results by pretraining the complete Siam-U2Net architecture and extracting the posttraining weights. These weights are then used in the model's feature extractor and frozen.

We pretrained the complete Siam-U2Net architecture to shorten training time using the same dataset as the CD task. The complete Siam-U2Net architecture, including the decoder, is used in the pretraining stage. The decoder has a similar structure to the encoder, and each stage combines the upsampled feature map of the previous stage with the differential feature map of the corresponding encoder stage as input. The decoder contains five layers and is processed by a 3×3 convolutional layer to generate five side-output binary images. After upsampling and fusion, these binary images are fully trained through a 1×1 convolutional layer and a Sigmoid function. It is worth noting that the decoder is only used in the pretraining stage to assist the model in learning image features. In the subsequent training of the CD task, we extract and freeze the encoder weights and do not use the decoder.

The SU-FDE approach primarily consists of two main components.

U-Shaped Structure Feature Extractor: We design a feature encoder using a U-shaped structure containing ten residual Residual U Block (RSU). RSU combines receptive fields of various sizes and can capture contextual information at multiple scales. These RSU blocks form a coherent U-shaped feature extraction framework, which helps extract feature information of complex geographical environments and diverse ground object types. These RSU blocks can be divided into two main structures: 1) the first four layers and 2) the fifth layer. The first four layers are four pairs of RSUs of different depths, namely, RSU-7, RSU-6, RSU-5, and RSU-4. The numbers here (such as "7", "6", "5", "4") represent the depth (D) of the RSU block (as shown in Fig. 2). This depth can be adjusted based on the resolution of the input feature map. In high-resolution remote sensing images, deeper RSU blocks can be selected to capture more detailed information. In addition, the fifth layer adopts another RSU-4F (right in structure, where "F" represents the dilated convolution version. These blocks are used to downsample to $16\times$ deeper and to keep the resolution of the feature maps not reduced. We use dilated convolutions instead of traditional downsampling and upsampling. This helps prevent the loss of useful contextual information and ensures that the input and output resolutions of the RSU-4F remain consistent.

Multilevel Feature Difference and Fusion: In the second part, we leverage an SA module to augment the sensitivity of the feature map in each layer of the Siamese network toward spatial

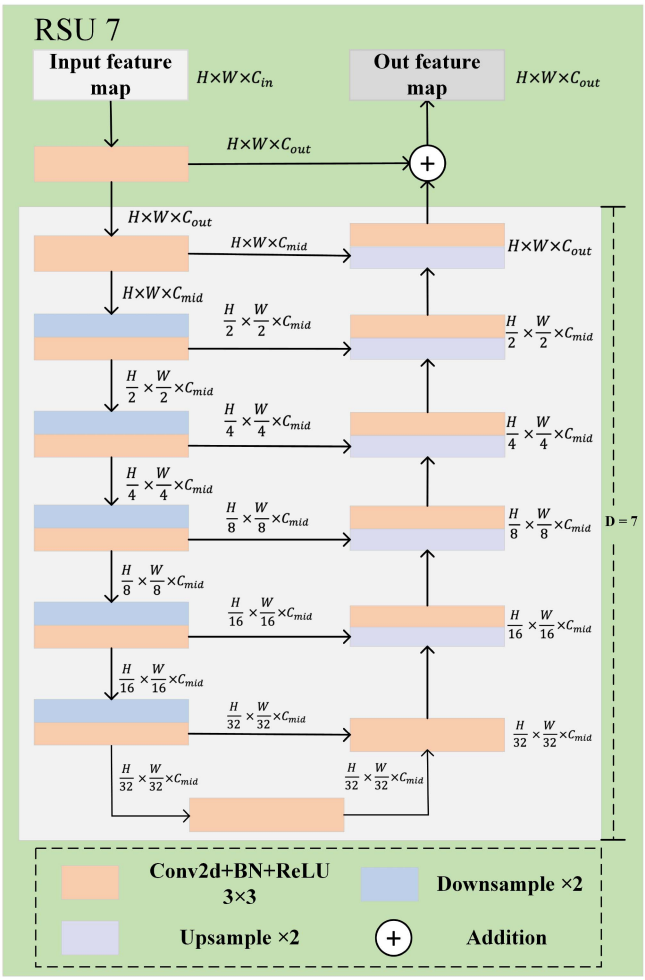

Fig. 2. Residual U Block (RSU). It adopts the encoder and decoder designed by U-Net, which is symmetrical and has D-layer depth. In the figure example, the depth is seven layers for the RSU seven block.

changes in the context of the CD task. The feature map is spatially enhanced through the SA module, which focuses on critical areas for CD in RS images. The SA module combines the information gathered through maximum and average pooling to generate a salient feature map. This salient feature map is then processed by the sigmoid activation function, resulting in an attention map of the same size as the original feature map. Attention maps highlight important spatial regions and suppress less relevant areas, improving the model's ability to recognize subtle changes. We measure the difference between processed feature maps through the L1 norm and effectively fuse the difference features into the encoding layer of Denoise-UNet, thereby enhancing CD performance while maintaining computational efficiency.

Combining the Siamese network and U2Net, the SU-FDE module is more potent for detecting RS image change. It improves the model's comprehension of image similarities and differences while enhancing the fusion of multiscale information and edge detection capability.

C. Denoising Diffusion Implicit Models

DDIM serves as another crucial component within the SMD-Net model. It plays a role in image denoising and enhancing the model's resilience to lighting and meteorological conditions changes. DDIM proposes an accelerated diffusion model based on DDPM. This generative model learns the noise distribution and removes noise in the image, improving CD accuracy. Here is a brief overview of DDPM and DDIM.

DDPM consists of a forward Markov denoising process, denoted as q, and a reverse denoising process, denoted as p ([28]). The forward process starts from x_0 with a label map $q(x_0)$ at T time steps, gradually introducing noise using a Gaussian distribution

$$q(x_1 : T|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}) \tag{3}$$

where

$$q(x_t|x_{t-1}) = N(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I). \tag{4}$$

Here, x_0 is the label map. x_t represents the state of the image at time step t, and β_t is a noise scale parameter.

In (3), as pointed out by [28], it is stated that x_t can be sampled for any time step t, eliminating the need to reuse q. Here, $\alpha_t := 1-\beta_t$, and $(\bar{\alpha}_t : \prod_{s=0}^{t} \alpha_t)$, can be written as

$$q(x_t|x_0) = N(x_t; \sqrt{\alpha_t}x_0, (1-\bar{\alpha}_t)I) \tag{5}$$

$$x_t = \sqrt{\alpha_t}x_0 + \sqrt{(1-\bar{\alpha}_t)}\epsilon, \epsilon \in N(0, I). \tag{6}$$

α_t is defined as $1-\beta_t$, where α_t is the complementary part of the noise scale parameter. DDIM modulates the label map and the impact of the noise introduced at each time step. $\bar{\alpha}_t$ is the cumulative product of α_t from 0 to t. That is the total amount of noise added at time step t. ϵ is a random noise added to the image.

In simpler terms, the forward process of DDPM gradually introduces noise to the image, transforming it from its initial state to a fully noisy state over T time steps.

The reverse process of DDPM involves the neural network's prediction of the noise added during the forward process, enabling the step-by-step reconstruction of the original image from its noisy state. This process portrays the neural network's ability to identify and subtract noise at each time step efficiently. The reverse process, utilizing the Bayes' theorem, discovers $p(x_{0:T})$ and decomposes it according to the chain rule formula

$$p(x_{0:T}) = p(x_T) \prod_{t=T}^{1} p(x_{t-1}|x_t) \tag{7}$$

where $(p(x_T) \in N(0, I))$ is a standard normal distribution

$$p(x_{t-1}|x_T) = N(x_{t-1}; \mu(x_t, t), \sum(x_t, t)). \tag{8}$$

$\mu(x_t, t)$ and $\sum(x_t, t)$ represent the mean and variance used in the reverse process, respectively.

Marginalize through $p(x_{0:T})$ to get the marginal probability $p(x_0)$ of x_0

$$p(x_0) = \int p(x_{0:T}) dx_{1:T}. \tag{9}$$

By applying the Jensen's inequality, we derive an evidence lower bound for the logarithm of the likelihood function. We then train the backward process to align its distribution with the forward process distribution: $-L(x_0) <= \log(p(x_0))$

$$L(x_0) = \mathrm{Eq}[L_T(x_0) + \sum_{t>1} D_{KL}(q(x_{t-1}|x_t, x_0)||p(x_{t-1}|x_t)) - \log p(x_0|x_1)] \tag{10}$$

where

$$L_T(x_0) = D_{KL}(q(x_T|x_0)||p(x_T)). \tag{11}$$

There exist multiple approaches to parameterize $\mu(x_t, t)$ (8) the priors, and we can predict the formula with a neural network$\mu(x_t, t)$, which can also predict noiseϵ. Here, we directly expect x_0 instead of noiseϵ and calculate the $\mu(x_t, t)$. In [28], the training objective of optimizing network parameters is simplified, and the optimization objective is proposed; in addition, a reweighted loss function is introduced as

$$E_{t,x_0,\epsilon} \parallel \epsilon - \epsilon_{x_t,t} \parallel^2 . \tag{12}$$

After the diffusion model is trained, x_T is sampled from $N(0, I)$, and x_t is denoised and iterated to obtain a new x_0 at each time step *t*

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \left(\frac{x_t - \sqrt{(1-\bar{\alpha}_t)}\epsilon_{(x_t,t)}}{\sqrt{\bar{\alpha}_t}} \right) + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2}\epsilon_{(x_t,t)} + \sigma_t \epsilon. \tag{13}$$

In [28], DDIM uses variance σ^2 as a hyperparameter that can be manually adjusted, and different effects can be obtained by adjusting σ^2

$$\sigma_t^2 = \frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}. \tag{14}$$

In DDPM

$$\sigma_t = \sqrt{(1-\bar{\alpha}_{t-1})/(1-\bar{\alpha}_t)}\sqrt{(1-\bar{\alpha}_t)/\bar{\alpha}_{t-1}}. \tag{15}$$

In DDIM, setting $\sigma_t = 0$ becomes deterministic sampling, and the generation process is deterministic. When x_{t-1} and x_0 are given, the forward process becomes deterministic, so the resulting model is an implicit probability model. Samples are generated from latent variables using a fixed process (from x_T to x_0). When the sampling length is much smaller than T, and the computational efficiency is substantially enhanced as a result of the iterative nature of the sampling process.

As one of the key components of the SMDNet model, leveraging DDIM accelerates the noise removal process in remote sensing images, consequently enhancing the overall performance of CD. DDIM gradually reduces noise through multiple iterative diffusion processes, thereby removing noise and improving the image. This capability contributes to the model's adeptness in detecting subtle changes in RS images, thereby enhancing the accuracy and reliability of the detection results.

TABLE I
DATASET STATICS

Dataset	Year	Resolution	Image Size	Pairs
LEVIR-CD [24]	2020	0.5m	1024×1024	10 192
DSIFN-CD [68]	2020	2m	512×512	15 760
CDD [69]	2018	0.03 ∼ 1m	4725×2200	16 000

IV. EXPERIMENTAL SETUP AND RESULT ANALYSIS

A. Datasets and Comparisons

In our experiments, we used three popular public datasets to evaluate the performance of our model across diverse lighting conditions, seasonal changes, different resolutions, and changing terrains and scenes. They are the LEVIR-CD dataset, the DSIFN-CD dataset, and the CDD dataset, as summarized in Table I. For the consistency of the test, the dataset was preprocessed, cropped into image pairs of 256×256 size, and randomly divided into three parts: train/test/val with 0.75/0.2/0.05, respectively.

LEVIR-CD Dataset [23] collected 637 pairs of image patches with very high resolution from Google Earth (GE). Each patch had a pixel resolution of 0.5 m and a size of 1024×1024 pixels. These image pairs come from more than 20 regions in the United States, with time blocks ranging from 5 to 14 years, with significant land changes and building growth, migration, etc. Due to its extensive period and diverse range of building types, this dataset presents a significant challenge for large-scale remote sensing building CD.

DSIFN-CD Dataset [55] is six large bitemporal high-resolution images collected from Google Earth. It covers six cities in China (i.e., Beijing, Chengdu, Shenzhen, Chongqing, Wuhan, and Xi'an) that contain significant differences among the land objects collected. The five large images in the dataset were segmented into 394 subimage pairs, each measuring 512×512 in size. Following data augmentation, a total of 3940 bitemporal image pairs were generated. Among them, the image pairs of Xi'an are cropped to 48 as the test set. This dataset is challenging due to its abundance of land objects and images that vary seasonally.

Google Earth Change detection dataset(CDD) [67] has remote sensing imagery of seasonal changes in an area. The dataset comprises 11 pairs of images, including 7 image pairs with dimensions of 4725×2700 pixels and 4 pairs of 1900×1000 pixels. The final output is a set of 16 000 cropped images, each measuring 256×256 pixels. it is divided into 10 000 image pairs for the training set and 3000 for testing and validation. With a high spatial resolution ranging from 3–100 cm, it offers detailed insights into changes in everyday structural objects like cars, buildings, roads, and seasonal variations in natural objects, from individual trees to extensive forest areas.

To validate the reliability of our method, we conducted comparisons with nine state-of-the-art CD methods from recent years on the same three public datasets, using image block sizes

consistent with those in the referenced papers. These methods include fully convolutional early fusion (FC-EF), fully convolutional siamese-difference (FC-Siam-diff), fully convolutional siamese-concatenation (FC-Siam-Conc), spatial-temporal attention neural network (STANet), a deeply supervised IFNet, SNUNet, BIT, Changeformer, and mask-guided local–global attentive network (MLA-Net). Each method demonstrates effectiveness in various aspects of CD through different network architectures. FC-EF, FC-Siam-diff, and FC-Siam-Conc are three pioneering CD methods based on deep learning, all proposed in the same research paper. STANet achieves better spatiotemporal relationship modeling by introducing multiscale subregion division and obtaining long-range spatiotemporal information in the self-attention module. IFNet employs a dual-stream architecture to extract deep features and then fuses these deep features with differential features via an attention module, reconstructing the CD map in the process. SNUNet, which merges the Siamese network with NestedUnet, enhances CD performance by integrating deep and shallow features. This integration is achieved through efficient information transmission and the application of attention modules, leading to a more effective feature synthesis for CD tasks. A simple CNN backbone (ResNet18) is combined with an end-to-end transformer to enhance the CD recognition of changing areas of interest. In addition, another Siamese network combines the Changeformer with a hierarchical Transformer encoder and MLP decoder to achieve efficient, multiscale CD, effectively capturing the detailed, long-range information required for precise CD tasks. MLA-Net accurately captures local and global contextual information by fusing a memory-efficient local-global attention module and introducing change masks.

B. Evaluation Metrics

In our evaluation framework, the primary metric for assessing performance is the F1-score derived from the test's precision and recall values. The formula to calculate the F1-score(F1) is

$$F1-\text{score}=2\times\frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}. \tag{16}$$

Furthermore, we also provide additional metrics for a comprehensive assessment, and These metrics take into account true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) in their calculations. TP represents correctly identified changes, FP indicates incorrectly marked changes, FN denotes missed actual changes, and TN refers to correctly identified nonchanges. Specifically included are Precision (Pre.) and Recall (Rec.) for the change category, the intersection over union (IoU) for the same, and the overall accuracy (OA). The formulas for these evaluation metrics are as follows:

$$\text{Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{17}$$

$$\text{Recall}=\frac{\text{TP}}{\text{TP+FN}} \tag{18}$$

$$\text{IoU}=\frac{\text{TP}}{\text{TP+FN+FP}} \tag{19}$$

$$\text{OA}=\frac{\text{TP+TN}}{\text{TP+TN+FN+FP}}. \tag{20}$$

C. Experimental Details and Analysis

Our network SMDNet is implemented using the PyTorch framework and executed on a single NVIDIA A100 GPU. We configured the batch size to 8 during training to ensure memory usage. The AdamW optimizer is used with the learning rate set to 1e-4, and the corresponding weight attenuation factor is also set to 1e-4 to prevent overfitting. The warm-up period is set to 1% of the total period, and the learning rate is updated according to the cosine annealing schedule. The model is regularly verified every 10 epochs. During training and testing, DDIM selects ten key points from 1000 time steps to guide the denoising process.

1) LEVIR-CD Dataset Analysis: LEVIR-CD is a dataset of high-resolution images with an extended period, rich in building types and changing scenes. In experiments conducted on the LEVIR-CD dataset, as shown in Table II, the SMDNet model achieved 92.71% and 89.17% in Pre. and F1, indicating that the model performed well in detecting subtle significant architectural changes, such as the emergence of new buildings or the demolition of old buildings. However, the model performed slightly worse regarding recall (85.89%) and IoU (82.71%). This may be because there are fewer samples of the changed category in the dataset, causing the model to prefer predicting the "no change" category. It improves accuracy but reduces recognition of actual changes. Here, IFNet leads in Pre. with 94.02%, but its performance in Rec. and IoU (82.93% and 78.77%, respectively) is slightly insufficient, which reflects the performance tradeoff between different models. MLA-Net performs outstandingly regarding Rec. (91.38%), indicating that the model may need to be selected and optimized according to specific application scenarios in different CD tasks.

Fig. 3(a)–(c) show the images before and after the change and the ground truth, respectively. Fig. 3(a) has the initial surface state, such as lakes, farmland vegetation, and buildings. In Fig. 3(b), changes that may involve building expansions, land use conversions, or seasonal changes can be observed. The ground truth Fig. 3(c) marks the areas of change and provides a benchmark for evaluating the model's performance in different seasonal scenarios. It can be seen from the visualization effect in Fig. 3(j) that the SMDNet model can accurately identify the actual change areas in most cases, consistent with the real change annotations on the ground. The visual effects of other methods are also compared qualitatively. Judging from the visual structure, SMDNet has a more precise division of edge details. There is no adhesion between buildings, effectively suppressing false changes between plants and buildings. The model's accuracy is critical in detecting new construction or demolition of buildings and is essential for monitoring urban expansion or redevelopment activity. However, the imbalance of classes in the dataset may cause the model to be too conservative when dealing with broad unchanged areas, lacking sensitivity to subtle changes and hiding potential risks of missed detections. Therefore, future research should explore data augmentation or resampling techniques to balance class distribution or develop

TABLE II
QUANTITATIVE COMPARISON ON THREE DATASETS

	LEVIR-CD					DSIFN-CD					CDD				
	Pre.	Rec.	F1	IoU	OA	Pre.	Rec.	F1	IoU	OA	Pre.	Rec.	F1	IoU	OA
FC-EF [68]	86.91	80.17	83.40	71.53	98.39	72.61	52.73	61.09	43.98	88.59	80.81	64.39	71.67	55.85	85.85
FC-Siam-diff [68]	89.53	83.31	86.31	75.92	98.67	59.67	65.71	62.54	45.50	86.63	85.44	63.28	72.71	57.12	87.27
FC-Siam-Conc [68]	91.99	76.77	83.69	71.96	98.49	66.45	54.21	59.71	42.56	87.57	82.07	64.73	72.38	56.71	84.56
STANet [23]	83.81	91.00	87.26	77.40	98.66	67.71	61.68	64.56	47.66	88.49	89.37	65.02	75.27	60.35	82.58
IFNet [55]	**94.02**	82.93	88.13	78.77	98.87	67.86	53.94	60.10	42.96	87.83	-	-	-	-	-
SNUNet [22]	89.18	87.17	88.16	78.83	98.82	60.60	72.89	66.18	49.45	87.34	-	-	-	-	-
BIT [12]	89.24	89.37	89.31	80.68	98.92	68.36	70.18	69.26	52.97	89.41	**92.04**	72.03	80.82	67.81	96.59
ChangeFormer [25]	92.05	88.80	**90.40**	82.48	99.04	88.48	**84.94**	86.67	76.48	95.56	90.02	83.93	86.87	81.27	97.18
MLA-Net [52]	87.52	**91.38**	89.41	80.85	98.89	88.17	88.49	88.33	80.24	94.91	88.96	87.25	88.10	83.26	97.94
SMDNet(Ours)	92.71	85.89	89.17	**82.71**	**99.17**	**88.51**	88.46	**88.48**	**80.91**	**96.56**	89.13	**87.35**	**88.23**	**83.30**	**99.29**

Best-performing values are highlighted in bold. All scores are presented in percentage format (%).

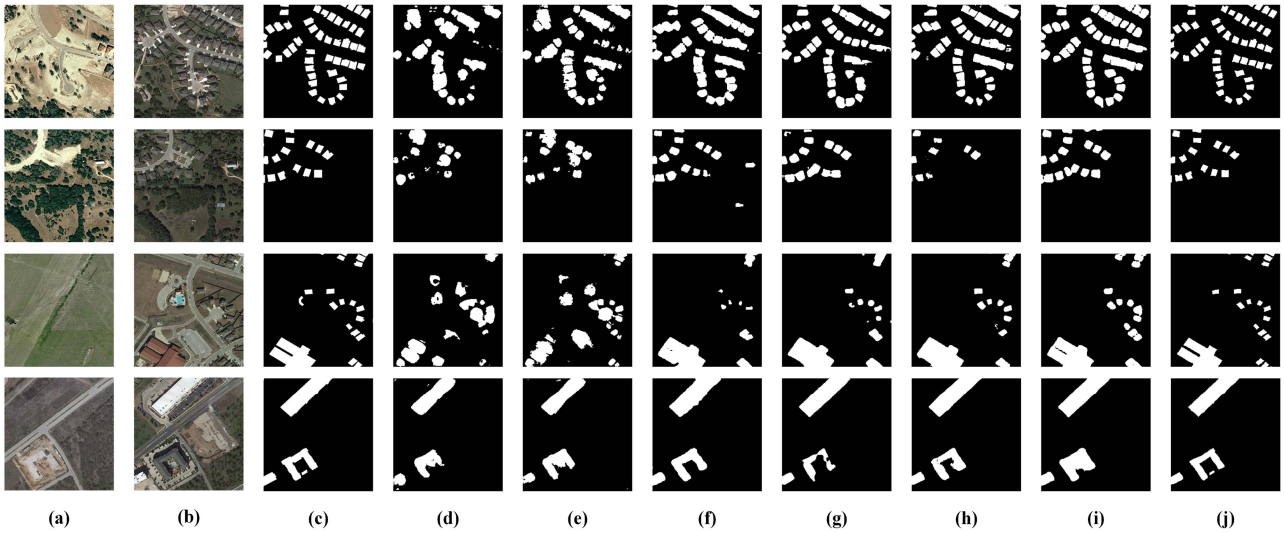

Fig. 3. Visualization of experimental results on the LEVIR-CD test set. (a) T1 image. (b) T2 image. (c) Ground truth. (d) FC-Siam-diff. (e) FC-Siam-Conc. (f) STANet. (g) BIT. (h) ChangeFormer. (i) MLA-Net. (j) Our SMDNet.

more sensitive CD algorithms to improve recall and IoU to ensure the model maintains high accuracy without sacrificing detection capabilities.

2) DSIFN-CD Dataset Analysis: The DSIFN-CD dataset has ground object scenes that change with seasons in different cities. As can be seen from Table II, SMDNet surpasses other comparison methods on multiple key performance indicators. Especially in terms of precision and recall, SMDNet achieves a higher balance, which shows that it can maintain a high CD rate while reducing false detections. SMDNet demonstrates its excellent overall performance in terms of F1 score, which is an essential indicator for evaluating the accuracy of CD methods. The IoU metric, combined with its high precision and recall, performs well in distinguishing changing areas from nonchanging areas. From the visual comparison of Fig. 4, we discovered the accuracy and boundary recognition capabilities of SMDNet in ground object CD. The model can more accurately identify the changing areas between images, and the predicted shape is consistent with the actual change boundaries of ground objects, showing the model's good sensitivity to the edges of ground objects. However, there are also certain shortcomings. For example, some misjudgment areas may appear in some complex backgrounds. These areas may be caused by the model's sensitivity to subtle changes in the image. In addition, it can be seen from Fig. 4 that the detailed part of the prediction may capture tiny changes in labels that are not annotated, which may be because the model has a certain sensitivity to these small-scale changes when extracting deep features. At the same time, it has specific suppression on some shadow changes.

3) CDD Dataset Analysis: The CDD dataset is mainly a seasonal variation dataset developed for RS. Observation of the graph shows that seasonal changes and different lighting conditions may cause spurious changes, thus challenging the model's ability to discern real changes. Prediction result (j) shows the sensitivity of our model SMDNet to various changes. It can be seen from the visual comparison diagram in Fig. 5 that due to the shadow of the building and the change of the lawn in the image centered on the first row, method (d)–(h) did not complete good detection. In contrast, our method performs well on changes in vegetation areas and building boundaries. Highly sensitive and resistant to false shadow changes. At the same time, the image pairs in the third and fourth rows also have

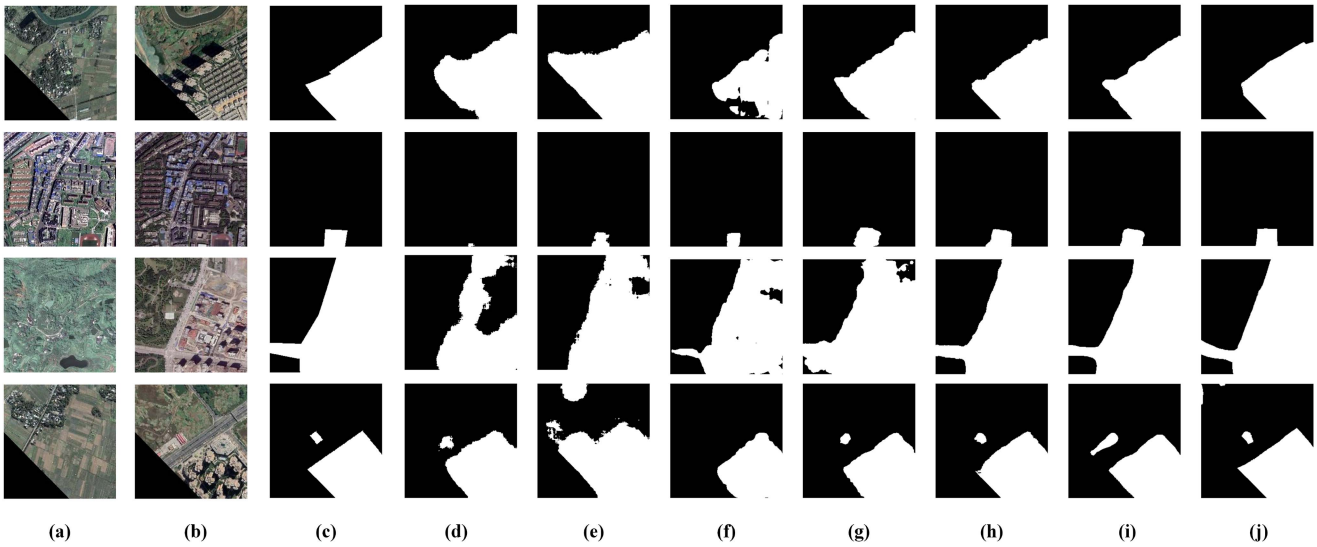

Fig. 4. Visualization of experimental results on the DSIFN-CD test set. (a) T1 image. (b) T2 image. (c) Ground truth. (d) FC-Siam-diff. (e) FC-Siam-Conc. (f) STANet. (g) BIT. (h) ChangeFormer. (i) MLA-Net. (j) Our SMDNet.

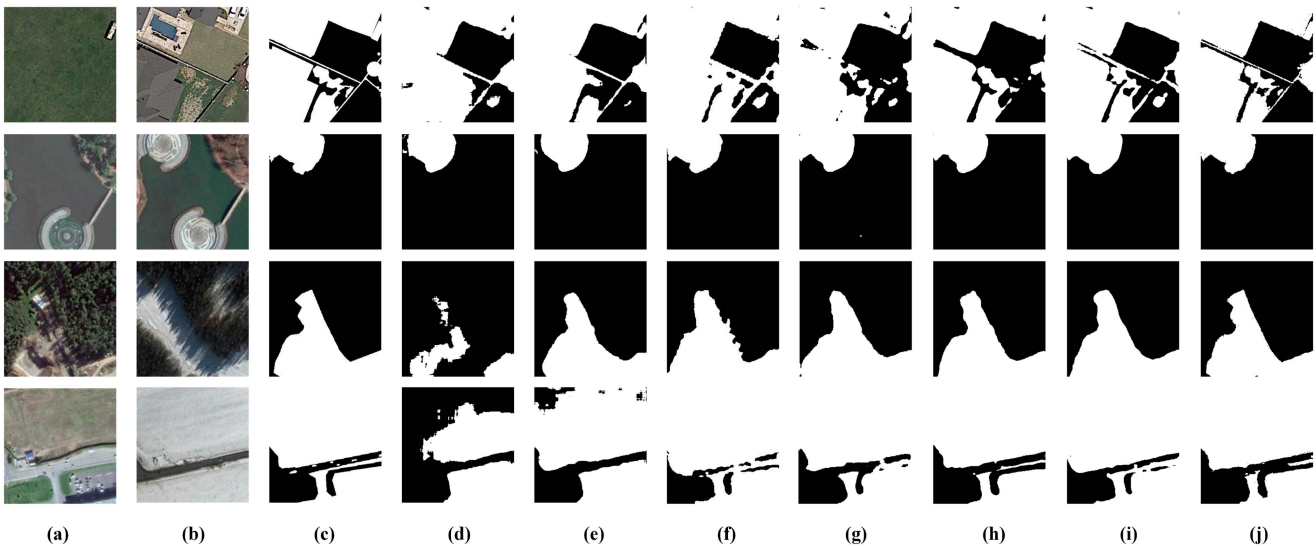

Fig. 5. Visualization of experimental results on the CDD test set. (a) T1 image. (b) T2 image. (c) Ground truth. (d) FC-Siam-diff. (e) FC-Siam-Conc. (f) STANet. (g) BIT. (h) ChangeFormer. (i) MLA-Net. (j) Our SMDNet.

corresponding performance. At the same time, in the second row (j) in Fig. 5, a detailed depiction of the edge of the water body is successfully identified. As can be seen from the visualization, the model SMDNet can predict the boundary changes of objects under different weather conditions, showing good sensitivity and accuracy.

The performance evaluation results of the model in Table II show its effectiveness on the CDD dataset. Although the Pre. (89.13%) lags slightly behind the BIT method, reflecting the model's efficiency in noise suppression and classification accuracy. At the same time, SDMNet also performs well regarding Rec., reaching 87.35%, ensuring that most real changes are successfully detected and avoiding missed detections. The F1 is 88.23%, balancing precision and recall, making it a reliable indicator of the model's overall performance. The IoU ratio is 83.3%, indicating that the model has high consistency between predicted and actual change areas. Furthermore, the OA reached 99.3%, demonstrating the model's high prediction accuracy on the entire dataset. However, OA may be overestimated due to the prevalence of unchanged regions in the dataset, highlighting the importance of comprehensive and accurate model evaluation using multiple performance metrics.

D. Ablation Experiments and Analysis

This subsection will help better understand the effectiveness of the proposed architectural and modular changes on the performance of SMDNet models in remote sensing image CD tasks. We designed experiments from three aspects:

1) The impact of different layer depths in SU-FDE and DU.

TABLE III
MODEL DEPTH RESULTS AT DIFFERENT LAYERS ON DSIFN-CD AND CDD DATASETS

Layer Depth	DSIFN-CD				CDD				Params Memory
	Precision	Recall	F1-score	OA	Precision	Recall	F1-score	OA	
4-layer	86.27	86.27	86.27	95.90	82.51	81.84	82.17	99.00	11.89MiB
5-layer	88.46	88.22	88.34	96.56	85.07	85.03	85.05	99.18	19.19MiB
6-layer	91.18	89.48	90.32	97.67	89.88	89.26	89.57	99.43	28.31MiB

All scores are presented in percentage format (%).

2) Mainly the impact of adding different attention mechanisms to the model.
3) Performance evaluation of added components in the model.
4) The impact of the choice of diffusion step size in the diffusion model on feature extraction in the dataset. The ablation experiments used four evaluation metrics: F1-score, Precision, Recall, and OA.

1) Effect of Layer Depth: The effect of different layers (four, five, and six) of the SU-FDE and DU on the performance of remote sensing image CD in the SMDNet model was first evaluated. The number of channels for the three types of depths were configured as follows: Four-layer network (64, 64, 64, 128), five-layer network (64, 64, 64, 128, 128), and six-layer network (64, 64, 64, 128, 128, 256). These experiments were conducted on the CDD and DSIFN-CD datasets to explore the relationship between network depth and model performance and parameter memory consumption.

The results of the two datasets, DSIFN-CD and CDD in Table III show that the F1 score and accuracy of the model increase as the number of layers increases, and parameter memory consumption also increases accordingly. Regarding the evaluation metrics, the six-layer SMDNet performs well on the DSIFN-CD and CDD datasets. The complexity of the scene is higher due to the composite nature of the dataset, which includes a variety of ground objects, such as buildings, roads, forests, and lakes. The six-layer configuration demonstrates that the deeper model structure can extract more complex and abstract features, resulting in good evaluation metrics. However, it is interesting to note that parameter memory consumption in the six-layer network is about 28.31MiB, which is about 2.4 times more than the four-layer network (approximately 11.89MiB). The five-layer network is not as good as the six-layer network regarding F1 score and accuracy, but it requires fewer parameter memory consumption, about 19.19 MiB. Balancing performance and parameter ratio efficiency, the five-layer network structure maintains a certain level of performance while effectively controlling the model size to ensure the practicability of practical applications.

2) Impact of Attention Mechanism on the Model: We conduct ablation experiments using different attention mechanisms after each SU-FDE layer and evaluate their impact on model performance on the CDD dataset. As shown in Table IV and Fig. 6, we examine nonlocal (NL) [69], axial attention (AX) [70], efficient channel attention (ECA) [71], and SA mechanisms. NL enhances the network's ability to integrate long-range features by capturing global dependencies; AX attention maintains computational efficiency when processing large-scale data; the ECA mechanism maintains a low parameter burden while enhancing channel correlation; and the SA mechanism maintains a low parameter burden in space. This enhances the recognition of local details and boundaries of the image. Experimental results show that adding any attention mechanism can improve the F1-score, Precision, Recall, and OA of SU-FDE. In particular, when the SA mechanism [SU-FDE+(SA)] is integrated, the model performance is most significantly improved, with F1-score increasing to 88.23%, Precision increasing to 89.13%, and Recall reaching 87.35%. At the same time, OA also improved to 99.29%. This result highlights the detailed sensitivity of the SA mechanism in capturing spatial dependencies in images to enhance feature extraction capabilities and help maintain the spatial consistency of detected changes, which helps interpret and understand changes. The SA mechanism can potentially improve model performance compared with other attention mechanisms.

TABLE IV
IMPACT OF USING DIFFERENT ATTENTION MECHANISMS IN SU-FDE ON SMDNET ON THE CDD DATASET

	F1-score	Precision	Recall	OA
SU-FDE(+NL)	86.54	86.31	86.78	99.24
SU-FDE(+AX)	86.72	86.30	87.14	99.27
SU-FDE(+ECA)	86.99	86.14	87.85	99.27
Ours	88.23	89.13	87.35	99.29

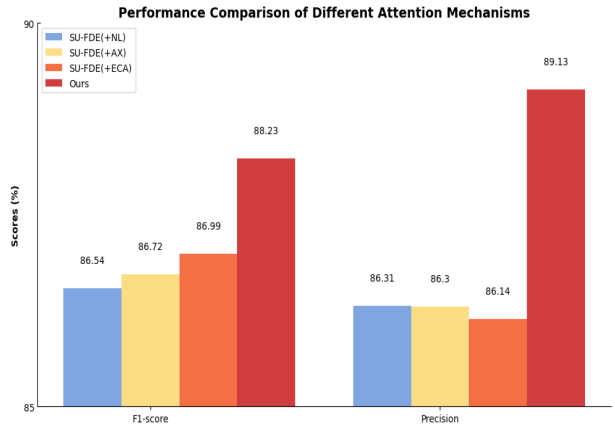

Fig. 6. Histogram comparison on CDD dataset using different attention mechanisms in SU-FDE of SMDNet.

TABLE V
PERFORMANCE COMPARISON OF DIFFERENT COMPONENTS ON THE CDD DATASET

Ablation Settings	SU-FDE	SA	DU	F1	Pre.	Rec.	OA
1	✓	✗	✗	74.57	67.72	82.96	98.68
2	✓	✓	✗	77.02	70.95	84.23	98.77
3	✗	✗	✓	80.97	81.39	80.41	98.77
4	✓	✗	✓	85.05	85.07	85.04	99.19
Ours	✓	✓	✓	88.23	89.13	87.35	99.29

3) Module Ablation Analysis: Table V presents the results of our ablation experiments on different model components to evaluate their impact on the final model performance. The first row contains only models of the SU-FDE encoding component designed to capture key features in the image. We equipped a simple decoder to form a complete network and test its performance. This component model achieved 74.57%, 67.72%, 82.96%, and 98.68% in F1-score, Precision, Recall, and OA, respectively. These results illustrate the potential of the SU-FDE encoder in extracting effective features. In the second row, we added the SA mechanism to the first complete Siam-U2Net network, highlighting that SA can consider key spatial information for enhancement on SU-FDE and has a 2.45% improvement on F1. In the third row, we adopt the DU in the diffusion model as our main architecture. Compared to using the SU-FDE encoder add SA, this model significantly improves F1 score and accuracy. At the same time, the recall decreases slightly, which may indicate that the DU component is better at suppressing noise and improving the model's discriminative ability. The model in the fourth row combines the SU-FDE encoding component and the DU, improving all performance metrics. SU-FDE improves model accuracy by introducing deep feature details into the diffusion model. Finally, our proposed SMDNet method integrates all components, where adding the SA mechanism improves detection efficiency by prioritizing spatial information when processing data. The last row shows that the comprehensive model performs optimally on all performance indicators. Compared with only DU, the F1 score increased by 7.26%, the precision and recall rates increased by 7.74% and 6.94%, respectively, and the overall accuracy was as high as 99.29%. This result shows that the interaction between SU-FDE, SA, and DU has no conflict and significantly improves the model's overall performance in remote sensing image CD.

4) Comparison of Efficiency: Table VI compares the number of parameters and computational complexity of SMDNet and several other models. The parameter amount of SMDNet is controlled at 5.031 M. At the same time, the parameter amount of our SU-FDE is only 0.617 M. Compared with FC-Siam-diff and FC-Siam-Conc, our model has fewer parameters to balance low parameter usage and sufficient learning ability. Regarding computational complexity, SMDNet has reached 266.8 G GFLOPs, higher than other models. However, this is mainly due to the denoising diffusion model we use, which improves the quality of model generation by repeatedly iteratively generating data.

TABLE VI
COMPARISON OF COMPUTATIONAL EFFICIENCY AND PARAMETER AMOUNTS OF DIFFERENT MODELS

Methods	GFLOPs(G)	Params(M)
FC-Siam-diff	4.285	1.350
FC-Siam-Conc	4.889	1.545
STANet	6.576	16.97
IFNet	41.18	50.71
SNUNet	27.44	12.03
BIT	10.90	12.40
ChangeFormer	202.8	41.03
MLA-Net	29.63	176.9
SMDNet(ours)	266.8	5.031

TABLE VII
CHOICE OF DIFFUSION STEP T CHANGES THE PERFORMANCE OF CD ON THE CDD DATASET

step t	F1-score	Precision	Recall	OA
500	84.38	84.82	83.94	99.06
750	85.20	85.31	85.10	99.10
1000	88.23	89.13	87.35	99.29

Although the computational overhead is relatively large, accuracy and processing details improvement have certain potential value. Compared with the 202.8 G computational complexity of ChangeFormer, the number of parameters of SMDNet is greatly reduced, successfully reducing the parameter burden of the model. We will further study how to reduce computational complexity while maintaining high accuracy in the future.

5) Effect of Diffusion Step Size: We conducted ablation experiments to understand the impact of the time step t in the denoising diffusion model on the feature extraction capability for CD datasets. During the training phase of the model, we considered multiple step settings to determine the appropriate time step length, including values of t set to 500, 750, and 1000. These experiments aimed to observe how different step length settings affect the model's performance on the CDD dataset. Table VII provides a detailed breakdown of the model's F1 scores, precision, recall rates, and overall accuracy at varying step lengths. Upon examination of the results, it was noted that the model achieved high values in all the aforementioned metrics at $t = 1000$. This observation in our model, which tracks CD over time, allows for a more detailed capture of changes in data details. The choice of step size t is crucial to the model's ability to process RS image data. The model can perform CD tasks more effectively under the setting of $t = 1000$.

V. CONCLUSION

In this study, we proposed the SMDNet model by introducing a Siamese network to propose a combination of SU-FDE and a DDIM. We successfully applied it to the CD task of high-resolution remote sensing images. Experimental results show that SMDNet's CD performance on LEVIR-CD, DSIFN-CD, and CDD datasets has been significantly improved compared

with existing technologies. Use the denoising capabilities of the diffusion model and the advantages of capturing data distribution. Add SU-FDE further to enhance the model's capture of edge details, thereby achieving more accurate CD in complex scenes. Despite the good performance of this model, there are still limitations.

1) In RS, due to the large amount of data and many resolutions from low to high, if you want to obtain more information, this will increase the time and cost of training, and it is difficult to implement in practice. Therefore, how to construct lightweight models to reduce training costs is an important direction.
2) We will continue to pay attention to the impact of the attention mechanism on the model. Explore an attention mechanism more suitable for RS images and conduct further research.
3) Considering the rapid advancements in RS hardware, a substantial volume of unlabeled data remains underutilized. Manually labeling data is time-consuming and often requires specific prior knowledge, yet it is still prone to inaccuracies, such as underlabeling or mislabeling. Therefore, it is necessary to explore semisupervised or self-supervised methods to leverage more data.

Future research will focus on further optimizing the model structure to reduce training costs, improve robustness under complex environmental conditions, and explore the potential of the SMDNet model in processing tasks in unlabeled remote sensing images.

REFERENCES

[1] M. Amani et al., "Evaluation of the landsat-based canadian wetland inventory map using multiple sources: Challenges of large-scale wetland classification using remote sensing," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 32–52, 2021.

[2] B. Tu, W. He, W. He, X. Ou, and A. Plaza, "Hyperspectral classification via global-local hierarchical weighting fusion network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 184–200, 2022.

[3] G. Yao et al., "An empirical study of the convolution neural networks based detection on object with ambiguous boundary in remote sensing imagery—A case of potential loess landslide," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 323–338, 2022.

[4] F. Safavi and M. Rahnemoonfar, "Comparative study of real-time semantic segmentation networks in aerial images during flooding events," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 4–20, 2023.

[5] S. M. M. Nejad, D. Abbasi-Moghadam, A. Sharifi, N. Farmonov, K. Amankulova, and M. Lászlź, "Multispectral crop yield prediction using 3D-convolutional neural networks and attention convolutional LSTM approaches," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 254–266, 2023.

[6] D. Hong et al., "Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks," *Remote Sens. Environ.*, vol. 299, 2023, Art. no. 113856.

[7] D. Hong et al., "SpectralGPT: Spectral remote sensing foundation model," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, doi: 10.1109/TPAMI.2024.3362475.

[8] J. Hu, Z. Huang, F. Shen, D. He, and Q. Xian, "A bag of tricks for fine-grained roof extraction," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2023, pp. 678–680.

[9] C. Qiao et al., "A novel multi-frequency coordinated module for SAR ship detection," in *Proc. IEEE 34th Int. Conf. Tools Artif. Intell.*, 2022, pp. 804–811.

[10] J. Hu, Z. Huang, F. Shen, D. He, and Q. Xian, "A rubust method for roof extraction and height estimation," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2023, pp. 770–771.

[11] W. Weng, W. Ling, F. Lin, J. Ren, and F. Shen, "A novel cross frequency-domain interaction learning for aerial oriented object detection," in *Proc. Chin. Conf. Pattern Recognit. Comput. Vis.*, 2023, pp. 292–305.

[12] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022, doi: 10.1109/TGRS.2021.3095166.

[13] Y. Liu, S. Nishiyama, and T. Yano, "Analysis of four change detection algorithms in Bi-temporal space with a case study," *Int. J. Remote Sens.*, vol. 25, no. 11, pp. 2121–2139, 2004.

[14] L. Bruzzone and D. F. Prieto, "Automatic analysis of the difference image for unsupervised change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 38, no. 3, pp. 1171–1182, May 2000.

[15] M. Hussain, D. Chen, A. Cheng, H. Wei, and D. Stanley, "Change detection from remotely sensed images: From pixel-based to object-based approaches," *ISPRS J. Photogrammetry Remote Sens.*, vol. 80, pp. 91–106, 2013.

[16] W. Shi, M. Zhang, R. Zhang, S. Chen, and Z. Zhan, "Change detection based on artificial intelligence: State-of-the-art and challenges," *Remote Sens.*, vol. 12, no. 10, 2020, Art. no. 1688.

[17] F. Shen, J. Zhu, X. Zhu, Y. Xie, and J. Huang, "Exploring spatial significance via hybrid pyramidal graph network for vehicle re-identification," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 8793–8804, Jul. 2022.

[18] F. Shen, Y. Xie, J. Zhu, X. Zhu, and H. Zeng, "GiT: Graph interactive transformer for vehicle re-identification," *IEEE Trans. Image Process.*, vol. 32, pp. 1039–1051, 2023.

[19] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2015, pp. 234–241.

[20] F. Shen et al., "An efficient multiresolution network for vehicle reidentification," *IEEE Internet Things J.*, vol. 9, no. 11, pp. 9049–9059, Jun. 2022.

[21] D. Peng, Y. Zhang, and H. Guan, "End-to-end change detection for high resolution satellite images using improved UNet," *Remote Sens.*, vol. 11, no. 11, 2019, Art. no. 1382.

[22] 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, 2022, Art. no. 8007805.

[23] 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, 2020, Art. no. 1662.

[24] F. Shen, X. Shu, X. Du, and J. Tang, "Pedestrian-specific bipartite-aware similarity learning for text-based person retrieval," in *Proc. 31th ACM Int. Conf. Multimedia*, 2023, pp. 8922–8931.

[25] W. G. C. Bandara and V. M. Patel, "A transformer-based Siamese network for change detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 207–210.

[26] C. Zhang, L. Wang, S. Cheng, and Y. Li, "SwinSUNet: Pure transformer network for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5224713.

[27] F. Heidary, M. Yazdi, P. Setoodeh, and M. Dehghani, "CTS-UNet : Urban change detection by convolutional siamese concatenate network with swin transformer," *Adv. Space Res.*, vol. 72, no. 10, pp. 4272–4281, Nov. 2023.

[28] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 6840–6851.

[29] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," 2022, *arXiv:2204.06125*.

[30] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2021, pp. 8780–8794.

[31] B. Kawar, M. Elad, S. Ermon, and J. Song, "Denoising diffusion restoration models," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2022, vol. 35, pp. 23593–23606.

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

[33] A. Kazerouni et al., "Diffusion models for medical image analysis: A comprehensive survey," 2023, *arXiv:2211.07804*.

[34] S. Zhang, M. Ni, L. Wang, W. Ding, S. Chen, and Y. Liu, "A two-stage personalized virtual try-on framework with shape control and texture guidance," 2023, *arXiv:2312.15480*.

[35] F. Shen, H. Ye, J. Zhang, C. Wang, X. Han, and W. Yang, "Advancing pose-guided image synthesis with progressive conditional diffusion models," 2024, *arXiv:2310.06313*.
[36] J. Ho et al., "Imagen video: High definition video generation with diffusion models," 2022, *arXiv:2210.02303*.
[37] J. Z. Wu et al., "Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 7623–7633.
[38] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 10850–10869, Sep. 2023.
[39] D. Baranchuk, I. Rubachev, A. Voynov, V. Khrulkov, and A. Babenko, "Label-efficient semantic segmentation with diffusion models," 2022, *arXiv:2112.03126*.
[40] T. Amit, T. Shaharbany, E. Nachmani, and L. Wolf, "Segdiff: Image segmentation with diffusion probabilistic models," 2022, *arXiv:2112.00390*.
[41] A. Graikos, N. Malkin, N. Jojic, and D. Samaras, "Diffusion models as plug-and-play priors," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2022, vol. 35, pp. 14715–14728.
[42] J. Wolleb, R. Sandkühler, F. Bieder, P. Valmaggia, and P. C. Cattin, "Diffusion models for implicit image segmentation ensembles," in *Proc. Int. Conf. Med. Imag. Deep Learn.*, 2022, pp. 1336–1348.
[43] S. Mukhopadhyay et al., "Diffusion models beat GANs on image classification," 2023, *arXiv:2307.08702*.
[44] W. H. Pinaya et al., "Fast unsupervised brain anomaly detection and segmentation with diffusion models," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2022, pp. 705–714.
[45] J. Wolleb, F. Bieder, R. Sandkühler, and P. C. Cattin, "Diffusion models for medical anomaly detection," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2022, pp. 35–45.
[46] W. G. C. Bandara, N. Gopalakrishnan Nair, and V. M. Patel, "DDPM-CD: Denoising diffusion probabilistic models as feature extractors for change detection," 2024, *arXiv:2206.11892*.
[47] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.
[48] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.
[49] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Proc. Deep Learn. Med. Image Anal. Multimodal Learn. Clin. Decis. Support: 4th Int. Workshop, 8th Int. Workshop, Held Conjunction Med. Image Comput. Comput.-Assist. Intervention* 2018, pp. 3–11.
[50] Y. Tang et al., "An object fine-grained change detection method based on frequency decoupling interaction for high-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5600213.
[51] T. Li, F. Xiong, W. Zheng, Z. Li, J. Zhou, and Y. Qian, "Wavelet Siamese network for change detection in remote sensing images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2023, pp. 5455–5458.
[52] F. Xiong, T. Li, J. Chen, J. Zhou, and Y. Qian, "Mask guided local-global attentive network for change detection in remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 3366–3378, 2024.
[53] L. Mou, L. Bruzzone, and X. X. Zhu, "Learning spectral-spatial-temporal features via a recurrent convolutional neural network for change detection in multispectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 924–935, Feb. 2019.
[54] Y. Liu, C. Pang, Z. Zhan, X. Zhang, and X. Yang, "Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 811–815, May 2021.
[55] C. Zhang et al., "A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 166, pp. 183–200, 2020.
[56] F. Shen, X. Du, L. Zhang, and J. Tang, "Triplet contrastive learning for unsupervised vehicle re-identification," 2023, *arXiv:2301.09498*.
[57] O. Özdenizci and R. Legenstein, "Restoring vision in adverse weather conditions with patch-based denoising diffusion models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 10346–10357, Aug. 2023.
[58] 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.*, 2022, pp. 10684–10695.
[59] 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, 2022, Art. no. 4834.
[60] Y. Xiao, Q. Yuan, K. Jiang, J. He, X. Jin, and L. Zhang, "EDiffSR: An efficient diffusion probabilistic model for remote sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5601514.
[61] L. Han et al., "Enhancing remote sensing image super-resolution with efficient hybrid conditional diffusion model," *Remote Sens.*, vol. 15, no. 13, 2023, Art. no. 3452.
[62] X. Zhao and K. Jia, "Cloud removal in remote sensing using sequential-based diffusion models," *Remote Sens.*, vol. 15, no. 11, 2023, Art. no. 2861.
[63] Y. Huang and S. Xiong, "Remote sensing image dehazing using adaptive region-based diffusion models," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 8001805.
[64] 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.*, 2023, pp. 5654–5657.
[65] X. Li et al., "Superpixel segmentation based on anisotropic diffusion model for object-oriented remote sensing image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, to be published, doi: 10.1109/JSTARS.2023.3324770.
[66] W. G. C. Bandara, N. G. Nair, and V. M. Patel, "DDPM-CD: Denoising diffusion probabilistic models as feature extractors for change detection," 2024, *arXiv:2206.11892*.
[67] M. Lebedev, Y. V. Vizilter, O. Vygolov, V. Knyaz, and A. Y. Rubis, "Change detection in remote sensing images using conditional adversarial networks," *Int. Arch. Photogrammetry Remote Sens. Spatial Inf. Sci.*, vol. 42, no. 2, pp. 565–571, 2018.
[68] R. C. Daudt, B. Le Saux, and A. Boulch, "Fully convolutional Siamese networks for change detection," in *Proc. IEEE 25th Int. Conf. Image Process.*, 2018, pp. 4063–4067.
[69] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.
[70] J. Ho, N. Kalchbrenner, D. Weissenborn, and T. Salimans, "Axial attention in multidimensional transformers," 2019, *arXiv:1912.12180*.
[71] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11534–11542.

Jia Jia received the B.S. degree in intelligent media and the M.S. degree in artificial intelligence, in 2019 and 2021, respectively, from Jeonju University, Jeonju, South Korea, where she is currently working toward the Ph.D. degree in artificial intelligence.

Her research interests include deep learning, remote sensing, and image segmentation.

Geunho Lee received the doctorate degree in circuits and systems from the Jeonbuk National University, Jeonju, South Korea, in 2000.

He has been a Professor with the Department of Artificial Intelligence, Jeonju University, since 2002. His research interests include linear algebra, optimization theory and deep learning, image generation and signal processing.

Zhibo Wang received the B.S. degree in information science and technology from Zhengzhou Normal University, Zhengzhou, China, in 2019, and the M.S. degree in artificial intelligence, in 2022, from Jeonju University, Jeonju, South Korea, where he is currently working toward the Ph.D. degree in artificial intelligence.

His research interests include deep learning and image restoration.

Zhi Lyu received the master's degree in agricultural informatics from Shanxi Agricultural University, Jinzhong, China, in 2017.

Since 2019, he has studied artificial intelligence with the Department of Culture and Technology, Jeonju University, Jeonju, South Korea. His research interests include deep learning and video captioning.

Yuchu He received the B.S degree in computer science and technology from Henan Normal University, Xinxiang, China, in 2009, the master's degree in computer science and technology from Sichuan University, Chengdu, China, in 2012, and the doctor's degree in control theory and control engineering from Chongqing University, Chongqing, China, in 2018.

He is currently an Associate Professor with Zhengzhou Normal University, Zhengzhou, China. His research interests include intelligent transportation systems and machine learning.