IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 8 December 2023, accepted 2 January 2024, date of publication 8 January 2024, date of current version 18 January 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3350641

METHODS

Leveraging Diffusion Modeling for Remote Sensing Change Detection in Built-Up Urban Areas

RAN WAN[1], JIAXIN ZHANG[1,2], YIYING HUANG[1], YUNQIN LI[1,2], BOYA HU[3], AND BOWEN WANG[4], (Member, IEEE)

[1]Architecture and Design College, Nanchang University, Nanchang 330047, China
[2]Division of Sustainable Energy and Environmental Engineering, Graduate School of Engineering, Osaka University, Osaka 565-0871, Japan
[3]College of Architecture and Urban Planning, Tongji University, Shanghai 200070, China
[4]Department of Science and Technology, Graduate School of Information, Osaka University, Osaka 565-0871, Japan

Corresponding author: Jiaxin Zhang (jiaxin.arch@ncu.edu.cn)

This work was supported in part by the Jiangxi Province Compact Residential Lot Detailed Planning and Architectural Design Research Project 2011ZBBH80014.

ABSTRACT In the evolving domain of built-up area surveillance, remote sensing technology emerges as an essential instrument for Change Detection (CD). The introduction of deep learning has notably augmented the precision and efficiency of CD. This study focuses on the integration of deep learning methodologies, specifically the diffusion model, into remote sensing CD tasks for built-up urban areas. The goal is to explore the potential of a pre-trained Text-to-Image Stable Diffusion model for CD tasks and propose a new model called the Difference Guided Diffusion Model (DGDM). DGDM incorporates multiple pre-training techniques for image feature extraction and introduces the Difference Attention Module (DAM) and an Image-to-Text (ITT) adapter to improve the correlation between image features and text semantics. Additionally, DGDM utilizes attention generated from pre-trained Denoise UNet to enhance CD predictions. The effectiveness of the proposed method is evaluated through comparative assessments on four datasets, demonstrating its superiority over previous deep learning methods and its ability to produce more precise and detailed CD results. This innovative approach offers a promising direction for future research in urban remote sensing, emphasizing the potential of diffusion models in enhancing urban CD precision and automation. Our implementation code is available at https://github.com/morty20200301/cd-diffusion.

INDEX TERMS Remote sensing data, change detection, diffusion model, built-up areas.

I. INTRODUCTION

The escalating phenomenon of urban sprawl worldwide underscores the pressing need for efficacious Change Detection (CD) mechanisms to meticulously monitor, analyze, and adeptly manage urban transformations. The essence of CD lies in discerning variations in terrestrial objects over temporal scales, typically employing a pair or more of images captured from identical geographical coordinates, a task significantly facilitated by remote sensing technology [1], [2], [3], [4]. This technology has proven indispensable in a plethora of realms, notably in disaster monitoring and urban planning, by virtue of its capacity for real-time surveillance and analysis [5].

In the specific context of built-up urban territories, CD unveils a unique array of challenges and prospects, thereby warranting a focused scrutiny within the domain of remote sensing technologies [6], [7], [8], [9]. The inherent density, complexity, and dynamism of urban landscapes demand the inception of advanced, resilient, and automated CD methodologies [10]. These methodologies are quintessential for aptly mirroring the subtle alterations transpiring over time, especially against the canvas of swift urbanization, climate change ramifications, and burgeoning

The associate editor coordinating the review of this manuscript and approving it for publication was Prakasam Periasamy.

infrastructural advancements [11]. The rigorous surveillance of urban locales is instrumental in unearthing invaluable insights into a multitude of urban dynamics including, but not limited to, shifts in land-utilization, residential expansions, and the intrusion upon verdant expanses [12]. In light of the progressive strides in remote sensing technology, there has been a palpable surge of interest among researchers toward honing CD methodologies [13], [14]. The ongoing endeavor to refine these methods not only holds promise for elevating the accuracy and efficiency of urban change detection but also for contributing to more prudent and sustainable urban planning and disaster management strategies.

In the past, change detection was primarily addressed through traditional methods and handcrafted features [15], [16]. However, traditional methods have limitations, and it can be challenging to choose the most suitable approach in practice. The limitations of traditional methods are becoming more apparent due to factors like diverse remote sensing data sources, improved spatial resolution, and richer image details. The poor expressiveness of features extracted by traditional methods can significantly reduce change detection accuracy and make the process sensitive to factors such as seasonal variation, illumination conditions, satellite sensors, and solar altitude angle. Traditional methods often require manual extraction of features, which can be time-consuming and tedious. Additionally, they heavily rely on domain-specific knowledge, which hampers the automation capability of change detection technology. In summary, traditional methods that depend on expert knowledge tend to be suboptimal, and the features they use are not strong in representing images.

With the emergence of Deep Learning (DL) [17], [18], change detection has been extensively explored [19], [20]. DL-based methods offer a promising solution to overcome the limitations of traditional methods, by automatically learning and extracting features from remote sensing images, making the process more efficient and less reliant on human expertise. Furthermore, DL's nonlinear characterization and remarkable feature extraction capabilities grant it a profound understanding of complex scenes, resulting in performance that consistently surpasses that of traditional methods. Capitalizing on these advantages, DL-based methods have witnessed an exponential rise in their adoption for tackling remote sensing challenges [21], [22]. From tasks such as image classification, object detection, and scene upstanding, to image segmentation, DL techniques have been harnessed to enhance the quality and efficiency of remote sensing applications [3], [4]. Notably, the highly discriminative features offered by DL methods have proven invaluable in addressing CD problems. Researchers have undertaken numerous studies to acquire richer insights into changes through the design of well-crafted model structures [23], [24], [25], [26].

Recently, the diffusion model [27] has emerged as a powerful tool in generative AI. One of its standout features is the ability to model intricate data distributions without the need for adversarial training. This sidesteps challenges commonly associated with Generative Adversarial Networks (GANs) [28], such as mode collapse. Furthermore, the Denoise Unet [29] demonstrates proficiency in extracting meaningful image features during its reverse operation. Notably, the ddpm-CD [26] employs the diffusion model for CD tasks and surpasses preceding methodologies by a significant margin. However, it relies solely on image features, overlooking the rich potential of high-level text semantics [30].

In this paper, we introduce the Difference Guided Diffusion Model (DGDM) for remote sensing CD tasks. Our objective is to harness the capabilities of a pre-trained text-to-image diffusion model [29]. Drawing inspiration from recent research [30], [31], we employ a frozen VQ-GAN [32] encoder to extract features from the subsequent scene and a frozen MAE [33] to obtain features from both the preceding and subsequent scenes. Building on this foundation, we developed a Difference Attention Module (DAM) to pinpoint the feature variations between scenes. To bridge the disparity between image features and high-level text semantics from pre-trained models, we introduced an Image-to-Text (ITT) adapter. Using a pre-trained diffusion model, the cross-attention map integrates with multi-scale feature maps sourced from the Denoise UNet. This combined output is then concatenated through a streamlined decoder for CD prediction. Our evaluations reveal that the DGDM outperforms the previous State-of-the-Art (SOTA) on four renowned CD datasets, showcasing significant improvements.

The contributions of this paper are summarized as follows: (1) We introduce DGDM, a pioneering approach that leverages pre-trained knowledge from a text-to-image diffusion model for CD tasks. (2) We developed the ITT adapter, specifically designed to bridge the gap between image features and learned text semantics. (3) We demonstrate the effectiveness of employing multiple pre-training technologies to enhance learning in new tasks. (4) Across four datasets, we assess and establish the superiority of our methods.

II. RELATED WORKS

A. REMOTE SENSING CHANGE DETECTION

There are a wide variety of methods available for detecting changes in multi-temporal images, such as ImageRatio [15], DPCA [36], MAD [37], CVA [38], and IRMAD [39]. Recent advancements in remote sensing CD have been significantly influenced by deep learning, owing to its superior learning capabilities [4], [40], [41]. Deep learning models, especially Convolutional Neural Networks (CNNs), have demonstrated exceptional performance in extracting intricate features from high-dimensional data, which is especially beneficial for analyzing remote sensing images (e.g., FC-EF [19], DS-IFN [42], and DASNet [43]). These neural network-based models are adept at automatically learning hierarchical representations without the necessity of manual

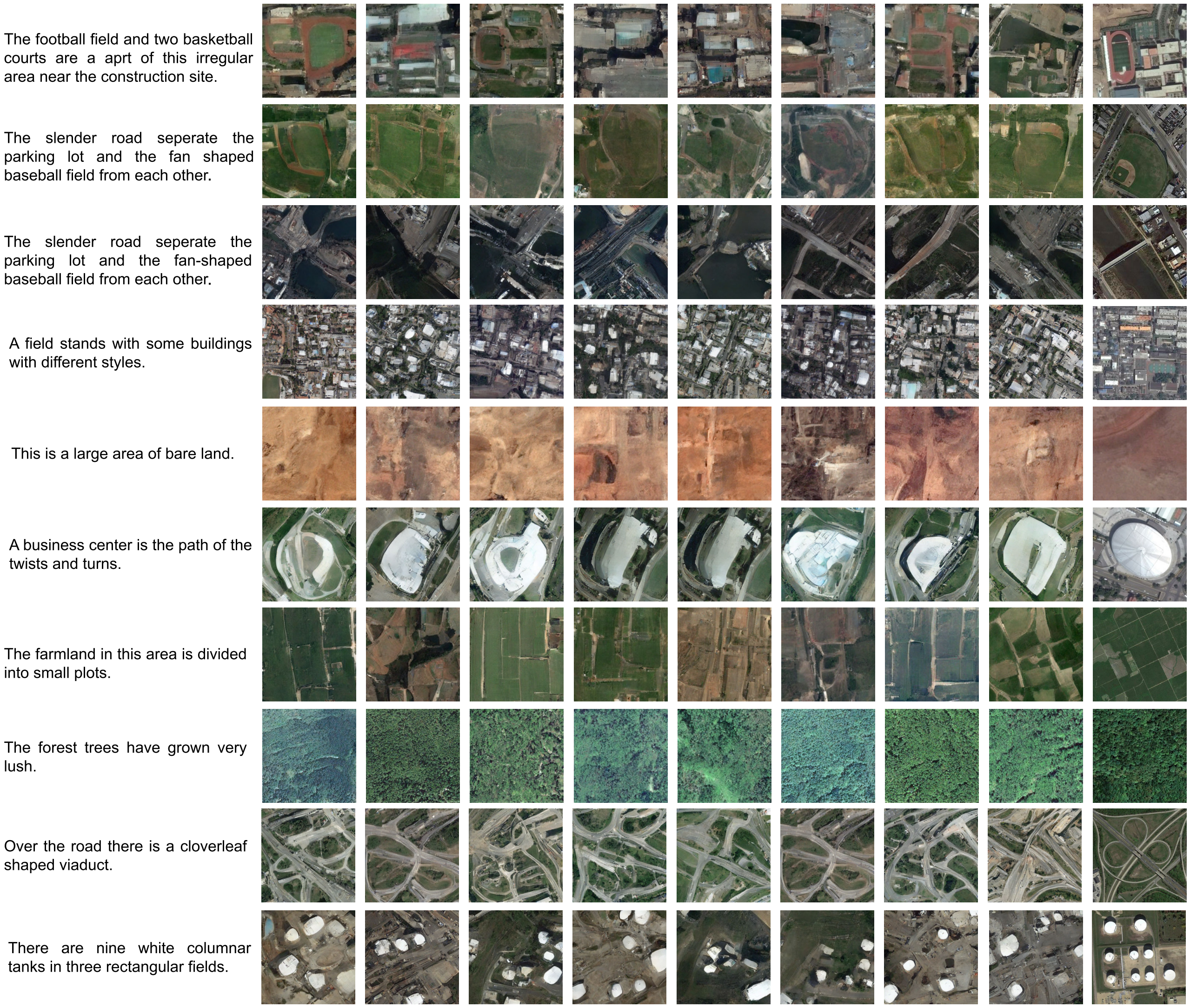

The football field and two basketball courts are a aprt of this irregular area near the construction site.
The slender road seperate the parking lot and the fan shaped baseball field from each other.
The slender road seperate the parking lot and the fan-shaped baseball field from each other.
A field stands with some buildings with different styles.
This is a large area of bare land.
A business center is the path of the twists and turns.
The farmland in this area is divided into small plots.
The forest trees have grown very lush.
Over the road there is a cloverleaf shaped viaduct.
There are nine white columnar tanks in three rectangular fields.

FIGURE 1. **Samples for text-to-image generation. All the samples are generated through a Stable Diffusion [29] model finetuned on RSICD dataset [34]. Input texts are shown on the left, and columns 1-8 are generated images (the last column is ground-truth).**

feature extraction, making them particularly advantageous for large-scale and high-resolution datasets [21], [22]. Attention modules [44], [45], [46] and transformer networks [47] tend to be a new aspect of CD tasks, which focus more on the global information during CD calculation. Moreover, the emergence of weakly supervised and contrastive techniques in the context of deep learning has further opened new horizons in remote sensing CD. These methodologies allow models to enable a training pipeline by pseudo label [35] or contrastive learning, thereby reducing the need for large labeled datasets in the remote sensing domain. Recently, diffusion probabilistic models have emerged as potent tools, addressing the challenges present in contemporary remote sensing pre-training techniques [48]. ddpm-CD [26] leverages the diffusion model to compute feature differences during the denoising process, yielding superior DC results compared to earlier studies. However, it still lacks the ability to fully utilize the potential of a pre-trained diffusion model.

B. DIFFUSION MODEL

Generative models have been a cornerstone of machine learning, particularly in image synthesis. Prominent models such as Variational Autoencoders (VAEs) [49] and Generative Adversarial Networks (GANs) [28] have paved the way for generating realistic images, with Diffusion models [27], in particular, showing remarkable success in this domain.

Diffusion processes have been explored to model the data generation procedure as a stochastic process, where images or data are gradually transformed from a noise initial state. One of the earliest works leveraging this idea was the denoising diffusion probabilistic models [27]. This model established a framework where data is transformed into noise through a reverse diffusion process, and the generation is the forward process from noise to data. Stable Diffusion [29] is designed to make the training and generation process of diffusion models more stable, consistent, and efficient. This can be

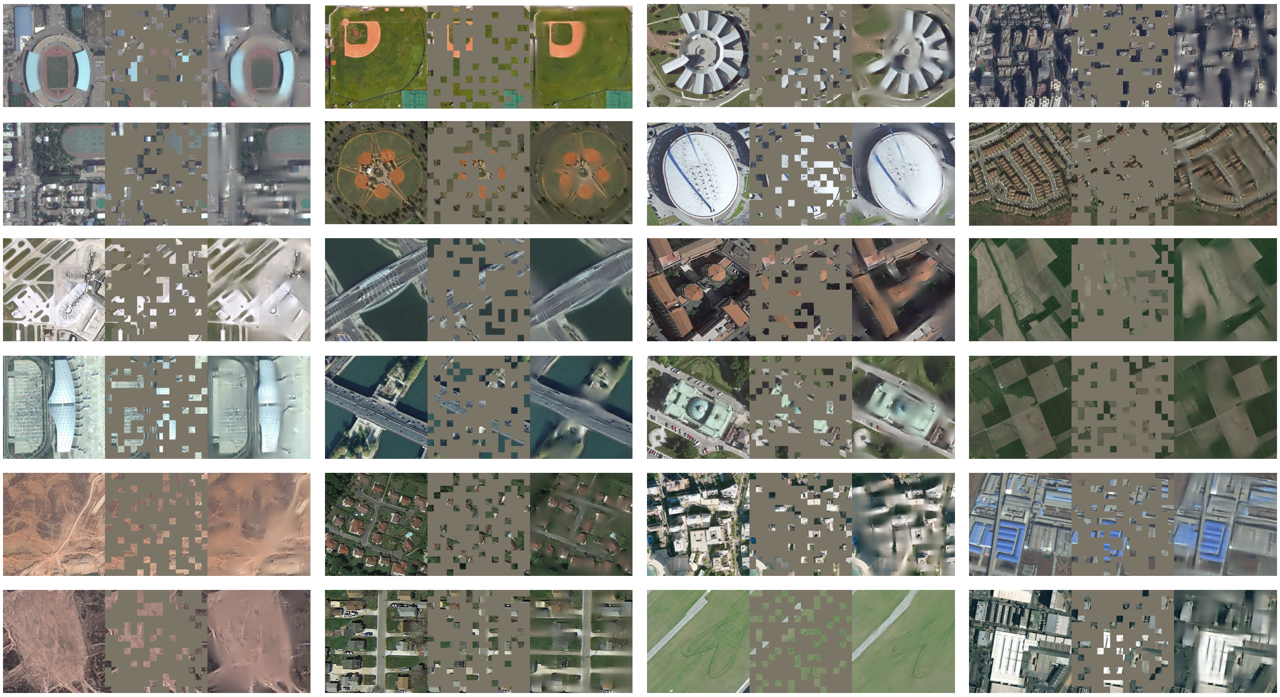


FIGURE 2. Reconstruction samples from MAE Pre-training. We employ the RSICD dataset [34] and the BA Dataset [35] for pre-training, using a mask rate of 75%. Within each sample, the first column displays the original input image, the second showcases the image with 75% of its pixels masked, and the final column presents the image as reconstructed by MAE.

in terms of convergence during training, the fidelity of the generated samples, or even the computational efficiency of the process. Text-to-image generation is inherently more complex than image synthesis from noise. This is because the model has to align the diffusion process with the semantic content of the text. Stability in this context would imply that small changes in the text description or slight variations in the diffusion process wouldn't lead to drastically different or nonsensical images. As shown in Figure 1, we demonstrate some generated samples from stable diffusion finetuned on RSICD dataset [34]. They depict scenes across city centers, bridges, land, and more, boasting high quality and aligning accurately with the provided texts.

A recent work VPD [30] further extends the denoise process of text-to-image by extracting the visual insights garnered by advanced diffusion models for tasks related to visual perception, e.g., segmentation [50] and depth estimation [51]. Rather than following the incremental diffusion process, they suggest using an autoencoder like UNet as the primary model. This allows for the direct processing of clean images without introducing noise. They then incorporate a singular denoising step using specific text prompts to derive the semantic information. The cross-attention between text and image feature [52] tends to be a strong guidance for downstream tasks. Inspired by VPD, we take advantage of Denoise UNet to instruct the detection of remote sensing change through feature differences between input images.

C. MASKED AUTOENCODERS

The Masked Visual Modeling method [53] offers a novel approach to learning representations from images that have been purposefully disrupted by masking. Essentially, this technique functions akin to denoise autoencoders, perturbing the input signals and subsequently reconstructing the original, undistorted signals to acquire effective representations. This paradigm has given rise to various derivatives, such as methods for reconstructing masked pixels [54] or for restoring lost color channels [55]. The remarkable success of Masked Language Modeling in Natural Language Processing (NLP), exemplified by BERT [56], in the context of self-supervised pre-training, coupled with the growing popularity of Vision Transformers (ViT) [57], has ignited a significant surge in research exploring the use of transformer-based architectures for Masked Visual Modeling within the domain of computer vision.

Some recent works naturally extend the foundational principles laid out by BERT [56] and propose learning representations from images by predicting discrete tokens. Meanwhile, MaskFeat [58] showcases the remarkable performance, particularly when employing the Histogram of Oriented Gradients (HoG) as the prediction target. iGPT [59], with its pioneering works in predicting pixel sequences, has charted the course in this direction. Additionally, some other research efforts [33] focus on using pixels themselves as prediction targets, a more straightforward approach. MAE [33] and its variants [60], by learning to reconstruct missing patches

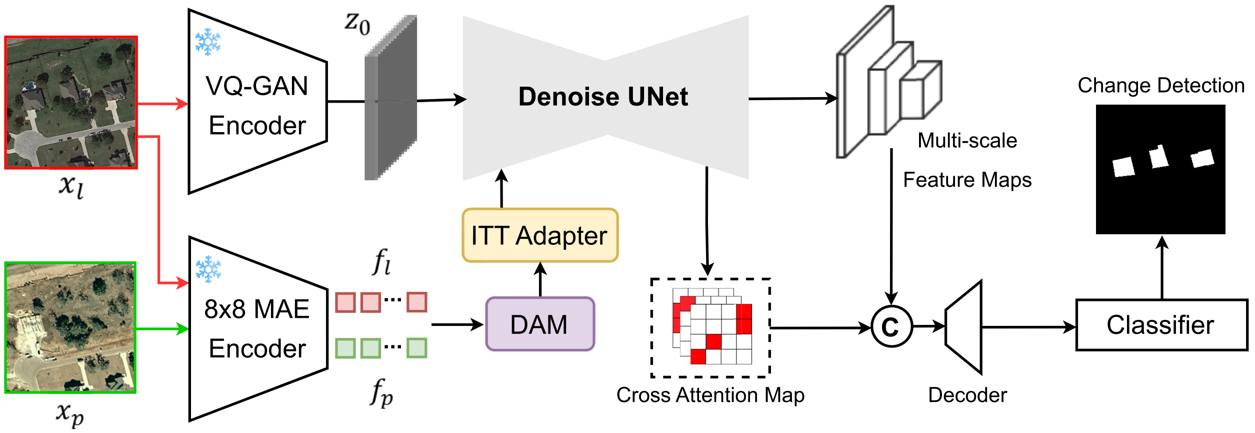


FIGURE 3. Overview of our proposed DGDM for remote sensing change detection: DGDM employs a frozen VQ-GAN encoder to extract features from the subsequent scene. It utilizes a frozen MAE to capture features from both preceding and later scenes. Following this, the DAM module computes the difference and aligns it with the Denoise UNet via the ITT adapter. We concatenate the cross-attention map with multi-scale feature maps sourced from the Denoise UNet, then forward them to a decoder, culminating in a classifier for change detection.

from randomly masked input image patches, have excelled in generating high-quality visual representations. Notably, MAE's encoder is optimized to handle visible patches, even under a high mask ratio, significantly expediting training and translating into improved transfer performance. We adopt MAE (patch 16 × 16) as the encoder to encode the input image and then use the extracted features to compute the difference prompt for a Denoise UNet. In Figure 2, we demonstrate some reconstruction samples from MAE pre-trained on RSICD dataset [34] and BA Dataset [35]. We observe that even when 75% of the pixels are masked in the original input image, MAE demonstrates the capability to reconstruct them using visible patches. Therefore, MAE is adeptly suited to serve as the encoder for feature difference calculation.

III. METHOD

A. OVERVIEW

In Figure 3, we present the overall workflow of our Difference Guided Diffusion Model (DGDM) designed for detecting changes in remote sensing imagery. The DGDM framework employs a frozen VQ-GAN encoder to distill features specifically from later scenes. Simultaneously, a static Masked Autoencoder (MAE) extracts features from both earlier and later scenes. Following this, our Difference Attention Module (DAM) assesses the disparities between these scenes and aligns them using the Image-to-Text (ITT) adapter before the input to a Denoise UNet architecture. The denoising process will generate cross-attention maps with multi-scale feature maps. They are channeled through a decoder, culminating in a specialized classifier for change detection. We will provide a detailed discussion of these components in this section.

B. PRELIMINARIES: DIFFUSION MODELS

Firstly, we offer a concise introduction to diffusion models. These models emulate data distribution by learning the reverse stage of a diffusion process. Denote z_t as the random variable at the t-th time step; the diffusion process can be characterized as:

$$z_t \sim \mathcal{N}(\sqrt{a_t} z_{t-1}, (1 - a_t)I), \tag{1}$$

where a_t are pre-defined coefficients for noise schedule. We can further get a simple close form of $p(z_t|z_0)$ by:

$$z_t = \sqrt{\bar{a}_t} z_0 + \sqrt{1 - \bar{a}_t}\epsilon, \tag{2}$$

$$\bar{a}_t = \prod_{i=1}^{t} a_i, \epsilon \sim \mathcal{N}(0, I). \tag{3}$$

This function allows model to sample an arbitrary z_t during training. By the re-parameterization [27], the objective function of diffusion models can be converted to:

$$L_{DM} = \mathbb{E}_{z_0,\epsilon,t}[||\epsilon - \epsilon_\theta(z_t(z_0, \epsilon)), t; \mathcal{C}||_2^2], \tag{4}$$

where ϵ_θ is the Denoise UNet [61] which is trained to predict ϵ based on the conditioning inputs $\mathcal{C}$ and z_t is calculated in Equation 2.

Equation 4 ensures consistent training of diffusion models, even when faced with sophisticated conditioning inputs. The approach of Stable Diffusion [29] has been pivotal in text-to-image generation, showcasing impressive outcomes in image synthesis guided by natural language. In their method, a VQGAN is initially trained to bridge the pixel space and the latent space. Following that, a diffusion model is trained in this latent space, aligning with the objective outlined in Equation 4. VPD [30] delves deeper, exploring ways to harness the extensive knowledge embedded in the pre-trained text-to-image diffusion model for subsequent visual tasks. In our work, we leverage the feature disparity between prior and subsequent scenes as the conditioning input, enhancing CD predictions using the pre-trained text-to-image diffusion model.

C. DIFFERENCE GUIDED DIFFUSION MODEL

Given an image $x_p \in \mathbb{R}^{\times H \times W \times 3}$ of the preceding scene and an image $x_l \in \mathbb{R}^{H \times W \times 3}$ of the later scene, our target is to predict the changing area $y \in \mathbb{R}^{H \times W \times 1}$ between them where y is a binary mask.

1) VQ-GAN AND MAE ENCODER

We denote a frozen VQ-GAN's encoder as E_{vq_gan}. Note that we obtain the VQ-GAN pre-trained by [62]. Thus, using x_l as the input, the output of the encoder is treated as z_0:

$$z_0 = E_{vq\_gan}(x_l), \in \mathbb{R}^{h \times w \times c}. \tag{5}$$

In order to compute the feature difference between x_p and x_l, we utilize an 8×8 MAE [33] encoder pre-trained on RSICD dataset [34] and the BA Dataset [35]. Denoting E_{mae} as the encoder, the feature extraction can be defined as:

$$f_p = E_{mae}(x_p), \in \mathbb{R}^{s \times c}, \tag{6}$$

$$f_l = E_{mae}(x_l), \in \mathbb{R}^{s \times c}, \tag{7}$$

where s is the flattened spatial dimension ($s = 64$ for 8×8 MAE) and c is the same feature dimension as Equation 5.

2) DIFFERENCE ATTENTION MODULE

Our DAM is an adaptation of the cross-attention mechanism. To preserve spatial information, we incorporate position embeddings p into each vector in f_{x_l}, leading to $f'_{x_l} = f_{x_l} + p$. f'_{x_l} serves as both the Key and the Value, while the feature f_{x_p} from the preceding scene acts as the Query. The linear transformations for Query, Key, and Value are symbolized by $Q(\cdot)$, $K(\cdot)$, and $V(\cdot)$, respectively. As shown in Figure 4, the formulation of DAM is as follows:

$$A_d = Q(f_p)K(f'_l), \tag{8}$$

$$A'_d = softmax(\phi(A_d)), \tag{9}$$

$$f_d = A'_d V(f'_l), \tag{10}$$

$$f'_d = norm(f_d + f'_l), \in \mathbb{R}^{s \times c}, \tag{11}$$

where ϕ represents a Feed Forward Network (FFN) and norm denotes layer normalization. After DAM we get feature tokens that describe the difference between preceding and later scenes.

3) IMAGE-TO-TEXT ADAPTER

Our DAM quantifies the difference between two input images and represents it by f'_d. However, as we introduced in Section III-B, the conditioning inputs are original text data and there is a gap to our f'_d. We thus need a module to covert image features and align them into the semantic space of texts. Image-to-Text (ITT) Adapter is inspired by recent visual prompt works [63], [64]. We set ITT as a simple structure with only two layers of full connection and this processing can be formulated as:

$$\bar{f}_d = ITT(f'_d), \in \mathbb{R}^{s \times c}. \tag{12}$$

We find that ITT alignment between image feature and pre-trained text semantic is essential for the training of continues Denoise UNet. This can be found in Section IV. We use $\bar{f}_d$ as the conditioning input $\mathcal{C}$ defined in Equation 4.

4) DENOISE UNET

A pre-trained diffusion model encapsulates sufficient information for sampling from the data distribution, as model ϵ_θ can be interpreted as the gradient derived from data density [30], [69]. We believe that a text-to-image model already has enough high-level knowledge. After ITT, the aligned difference features can instruct the denoising process. Through Denoise UNet ϵ_θ, we want to extract multi-scale feature maps $\mathcal{F}$ and corresponding attention $\mathcal{A}$ between $\bar{f}_d$ and z_0. The process of Denoise UNet can be formulated as:

$$\mathcal{F}, \mathcal{A} = \epsilon_\theta(z_0, \bar{f}_d). \tag{13}$$

Observe that we set $t = 0$, ensuring that the latent feature map remains no noise. The multi-scale feature $\mathcal{F}$ is readily derived from the final layer of each output block across varying resolutions. This comprises four feature maps, with the spatial size of the i-th feature map F_i defined as $h_i = w_i = 2^{i+2}$, $i = 1, 2, 3, 4$.

It has been observed in [30] and [52] that the attention maps $\mathcal{A}$ generated from cross attention module is essential for downstream task. Cross-attention is operated through each of the 4 resolutions of the Denoise UNet. Following [30], we take the mean of all cross-attention maps corresponding to a particular resolution to yield an average map A_i for the i-th resolution. Given that cross-attention maps are derived by using the conditioning inputs $\mathcal{C}$ as both key and value, the shape of the averaged attention map is $A_i \in \mathbb{R}^{h_i \times w_i \times s}$.

We then resize (to block $i = 4$) and concatenate $\mathcal{F}$ and $\mathcal{A}$ into an entirety M as:

$$M = \oint_i^4 re([F_i, A_i]), \in \mathbb{R}^{h_4 \times w_4 \times \bar{c}}, \tag{14}$$

where $\oint$ is the loop calculation, $[\cdot]$ is the concatenation and re is the resize operation. M merge both the image features of z_0 derived from Denoise UNet and high-level semantic describing difference. We find this mixture of information can greatly contribute to the downstream CD tasks.

5) CHANGE DETECTION

The final target of our DGDM is to realize the CD and we employed a simple decoder (3 layers of convolution) followed by a one-layer fully connection as a classifier. The process can be formulated as follows:

$$y = cls(Decoder(M)), \tag{15}$$

where $cls(\cdot)$ is the classifier and y is the final change detection.

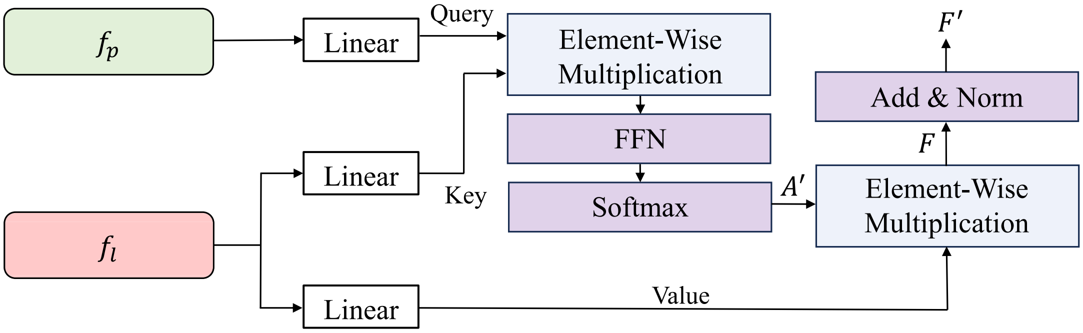


FIGURE 4. The structure of our Difference Attention Module (DAM). It takes features from the preceding scene as query and the later scene as Key and Value.

TABLE 1. CD results on four datasets compared to previous SOTA methods. All results are shown as percentages and the best results are mark with bold and the second best are marked with underline.

	WHU-CD [65]			LEVIR-CD [66]			DSIFN-CD [42]			CDD [67]		
Method	F1	IoU	OA	F1	IoU	OA	F1	IoU	OA	F1	IoU	OA
FC-SC [19]	83.67	71.96	98.49	79.85	66.46	98.50	59.71	42.56	87.57	75.11	60.14	94.95
FC-EF [19]	83.40	71.53	98.39	76.73	62.24	98.31	72.17	56.45	83.37	66.93	50.30	93.28
FC-SD [19]	86.31	75.92	98.67	78.95	65.23	98.84	70.55	54.49	84.13	70.61	54.57	94.33
DT-SCN [68]	87.67	78.05	98.77	91.43	84.21	99.35	70.58	54.43	82.87	92.09	85.34	98.16
STANet [66]	87.26	77.40	98.66	82.32	69.95	98.52	64.56	47.66	88.49	84.12	72.22	96.13
SNUNet [24]	88.16	78.83	98.82	83.50	71.67	98.71	66.18	49.45	87.34	83.89	72.11	96.23
ChangeFormer [20]	90.40	82.48	99.04	88.57	79.49	99.12	94.67	88.71	93.23	94.63	89.80	98.74
BIT [23]	89.31	80.68	98.92	90.53	83.39	99.34	87.61	77.96	92.30	88.90	80.01	97.47
ddpm-CD [26]	90.91	83.35	99.09	92.65	86.31	99.42	**96.65**	**91.28**	**97.09**	95.62	91.62	98.98
DGDM *(ours)*	**91.25**	**83.91**	**99.18**	**92.80**	**86.55**	**99.43**	96.48	91.05	96.88	**95.70**	**91.78**	**99.09**

IV. EXPERIMENTS

A. DATASET AND TRAINING DETAILS

1) PRE-TRAING

We adopt a VQ-GAN model from [62] and finetune the Stable Diffusion using RSICD dataset [34] which collectively comprise 10,921 remote sensing images, each accompanied by five descriptive sentences. Some samples are shown in Figure 1. We also pre-train MAE with both the RSICD dataset [34] and BADataset [35]. This is realized by first pre-training a 16 × 16 MAE with 40 epochs. Then, we resized the learned position embedding into 8 × 8 and finetuned a 8 × 8 with a further 40 epochs. The learning rate starts from 0.0001 and undergoes a linear decay to zero across 40 epochs in both settings. We take this operation to ensure that even an 8 × 8 MAE can extract informative features. The reconstruction results can be found in Figure 2.

2) TRAINING OF DGDM

We adopt four popular datasets to evaluate CD tasks. They are WHU-CD [65], LEVIR-CD [66], DSIFN-CD [42], and CDD [67]. These four data describe the changes in surface architecture during urban construction. Following previous works, we adopt 256 × 256 as the input size for models. We retained the parameters of both the VQ-GAN's encoder and the MAE's encoder in a fixed state. The Denoise UNet was assigned a learning rate of 0.00005, while other modules received a rate of 0.0001. These rates undergo a linear decay to zero across 100 epochs. We fine-tuned DGDM using the validation set of the datasets and presented the outcomes on the test set. We employed the cross-entropy loss in conjunction with the AdamW [70] optimizer. All experiments are implemented using a Nvidia A100 GPU.

3) COMPARISON WITH STATE-OF-THE-ART (SOTA)

We compare our DGDM with several SOTA methods including DT-SCN [68] a daul-task constrained siamese network, STANet [66] spatial-temporal attention network, ChangeFormer [20] a transformer-based method, fully-convolutional early-fusion (FC-EF) [19], siamese-difference (FC-SD) [19], siamese-concatenation (FC-SC) [19], ddpm-CD [26] a diffusion model-based method, and BIT [23] bi-temporal image transformer.

4) EVALUATION METRICS

To evaluate the CD performance, we present the F1 and Intersection over Union (IoU) scores pertaining to the change class

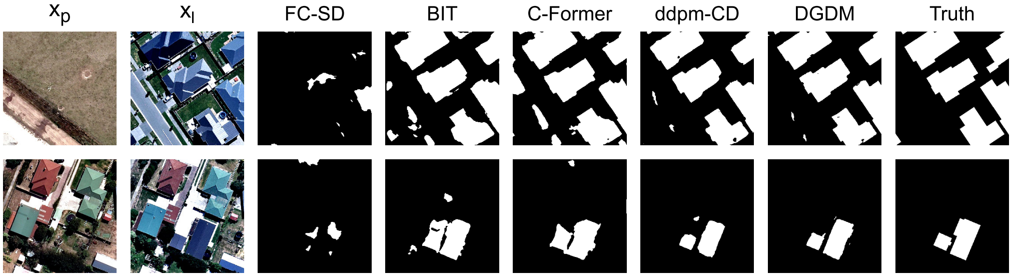


FIGURE 5. Comparision of our DGDM to SOTA methods FC-SD [19], BIT [23], ChangeFormer (C-Former) [20], and ddpm-CD [26] in WHU-CD dataset [65].

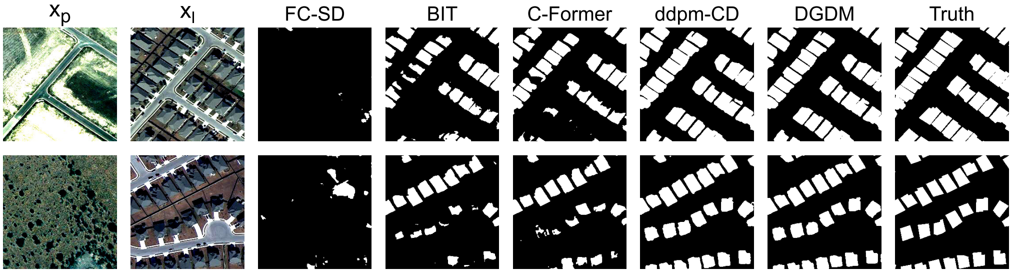


FIGURE 6. Comparision of our DGDM to SOTA methods FC-SD [19], BIT [23], ChangeFormer (C-Former) [20], and ddpm-CD [26] in LEVIR-CD dataset [66].

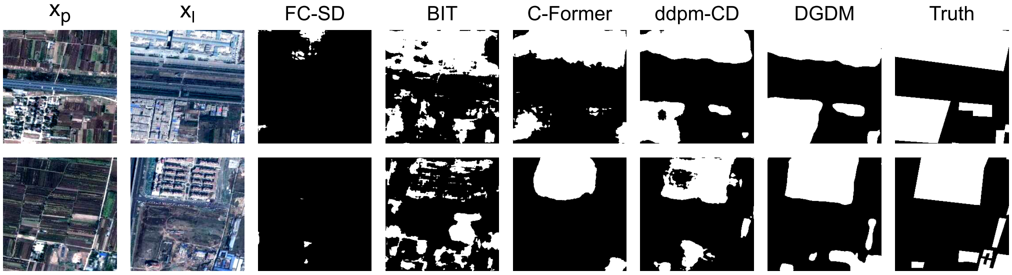


FIGURE 7. Comparision of our DGDM to SOTA methods FC-SD [19], BIT [23], ChangeFormer (C-Former) [20], and ddpm-CD [26] in DSIFN-CD dataset [42].

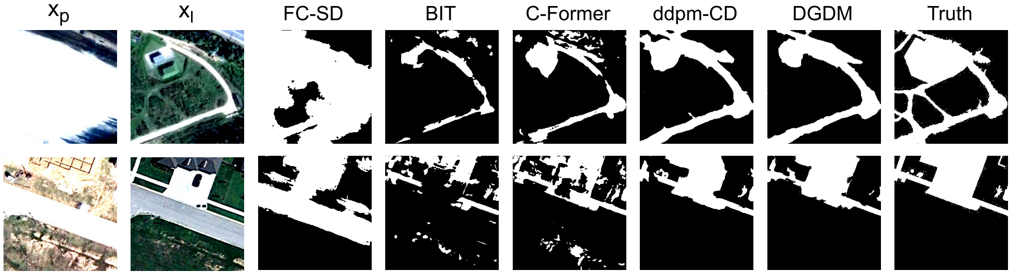


FIGURE 8. Comparision of our DGDM to SOTA methods FC-SD [19], BIT [23], ChangeFormer (C-Former) [20], and ddpm-CD [26] in CDD dataset [67].

as our primary quantitative metrics. Additionally, we provide the overall accuracy (OA) to assess the comprehensive quality of predictions, in alignment with [26].

B. RESULTS

1) QUANTITATIVE RESULTS

In Table 1, we juxtapose our DFDM with prior SOTA methods across four distinct datasets. Results are represented in percentages, with top performances highlighted in bold and second-best underlined. Notably, ChangeFormerm, BIT, and ddpm-CD represent the most recent advancements in CD tasks. Upon analysis, it's evident that DFDM consistently outperforms across three metrics on WHU-CD, LEVIR-CD, and CDD datasets. To illustrate, in the WHU-CD dataset, there is a marked improvement of 0.34%, 0.56%, and 0.09% in F1, IoU, and OA metrics respectively when compared to

TABLE 2. Ablation studies assess the impact of the DAM&ITT, MAE, and Attention modules on CD results across four datasets. All outcomes are presented as percentages, with the best results highlighted in bold.

				WHU-CD [65]			LEVIR-CD [66]			DSIFN-CD [42]			CDD [67]		
DAM	ITT	MAE	Attention	F1	IoU	OA	F1	IoU	OA	F1	IoU	OA	F1	IoU	OA
	✓	✓	✓	81.57	70.33	98.19	79.02	66.58	98.70	69.93	52.77	81.68	76.33	62.68	95.68
✓		✓	✓	83.69	72.92	98.36	80.50	67.14	98.83	70.44	53.29	80.53	74.19	60.27	95.39
✓	✓		✓	88.15	80.34	98.90	89.45	82.97	99.28	87.07	76.99	92.18	87.50	79.81	97.33
		✓	✓	80.22	70.18	98.02	77.56	66.02	98.43	68.56	52.50	80.27	73.48	59.82	95.26
✓	✓	✓		85.22	75.40	98.55	79.02	65.71	98.90	71.34	58.49	84.13	75.48	58.90	95.81
✓	✓	✓	✓	**91.25**	**83.91**	**99.11**	**92.80**	**86.55**	**99.44**	**96.48**	**91.05**	**96.88**	**95.70**	**91.78**	**98.99**

one of the leading methods, ddpm-CD. The incorporation of the diffusion model's semantic information enables DFDM to more effectively discern differences between scenes, a core competency of the CD task. However, in the DSIFN dataset, our performance is marginally less optimal. This may caused by the small data quantity of DSIFN, which made the learning of DFDM difficult. Despite this, the results achieved by DFDM remain competitive, given the minimal disparity.

2) QUALITATIVE RESULTS

We provide visual representations of CD predictions across four datasets, presenting two sample images for each from Figures 5 to 8. A comprehensive analysis will follow to validate the superiority of our DGDM.

In Figure 5, we present two samples from the WHU-CD dataset, which focuses on detecting changes in building structures. Our DGDM demonstrates superior accuracy in predicting the architectural contours, even capturing minor modifications in buildings, as illustrated by the second sample. The LEVIR-CD dataset shares a similar objective of identifying building alterations but boasts a more extensive data set with intricate scenes. As depicted in Figure 6, the samples spotlight changes in densely constructed areas. While methods like FC-SD, BIT, and C-Form struggle to detect comprehensive alterations, and ddpm-CD exhibits commendable results, our DGDM remains unmatched in its performance. The DSIFN-CD and CDD datasets delve into broader CD tasks. As shown in Figure 7, samples in DSIFN-CD highlight the evident superiority of DGDM. And CDD, with its massive collection of 15,998 data pairs, poses challenges for all methods. As evident from the first challenging sample in Figure 8, discerning pathways in green zones proves elusive, leading models to misidentify them as unchanged areas. Despite these challenges, DGDM exhibits enhanced precision, especially in predicting main roads.

C. ABLATION

In this section, we evaluate the significance of certain modules within our DGDM. As detailed in Table 2, our ablation study focuses on DAM, ITT, MAE, and Attention (generated cross-attention masks that concatenated to the input for the encoder). VQ-GAN is necessary for the Stable Diffusion pipeline introduced in VPD [30]. Thus, it is not included to this ablation study. A checkmark in the table signifies the inclusion of a module during training, while the last row with all check marks represents the complete DGDM configuration. When MAE is excluded, the VQ-GAN encoder is directly used for feature extraction. In the absence of the Attention module, multi-scale features are directly fed into the decoder, bypassing the concatenation with attention maps.

Our findings highlight the essential roles of DAM and ITT in training DGDM. As we hypothesized in Section III-C3, a significant gap exists between image features and pre-trained high-level semantics. The difference of images from two periods also need to be calculated (disscused in Section III-C2). Without these modules, we observed a marked decrease in performance across all datasets. Furthermore, our results indicate that merely adding either DAM or ITT independently does not lead to a noticeable improvement in performance. This suggests that each module performs distinct and critical functions in the overall process. The utility of Attention for downstream CD detection tasks is also evident. We posit that the attention maps between the difference feature represented as $\bar{f}_d$ and z_0 accentuate the areas of change, thereby contributing to the decoder. Regarding the use of MAE, the pre-trained features in MAE seem to facilitate a superior feature differentiation. This could be attributed to the ViT structure, which aligns closely with the text token encoder in the pre-trained Stable Diffusion.

D. DISCUSSION

The rapid pace of urbanization worldwide accentuates the importance of efficient CD in built-up urban regions. Such areas are characterized by their intricate landscapes, comprising a mix of ever-evolving architectural styles, dense infrastructural elements, and overlapping urban features. Historically, these complexities posed significant hurdles for traditional CD methodologies. Deep learning, with its superior feature extraction capabilities, has been a game-changer, especially in discerning subtle transformations within dense urban fabrics. Our DGDM, which adeptly combines both image features and high-level text semantics, emerges as an invaluable tool in this domain. It can pinpoint even the slightest changes in urban structures, an attribute that's paramount for applications like urban planning, disaster management, infrastructure assessments, and developmental monitoring.

It's crucial to acknowledge that there's room for refined enhancements, as illustrated by the outcomes from the DSIFN dataset. To optimize learning and improve the alignment between image features and semantics, DGDM requires a more comprehensive dataset. In the future, we plan to involve more data for training, which is the tend for large vision model pattern. Our ablation study has provided deeper insights into the functions of individual modules within DGDM. The pivotal roles of DAM&ITT, the significance of the Attention module in underscoring differences, and the synergy between multiple pre-training technologies like MAE, underscore the depth and sophistication of our model.

It is certain that urban change is vast, and DGDM's potential could paint far beyond the current experiment results. As cities evolve, the need for CD methodologies that are not just precise but also swift, scalable, and socially sensitive becomes paramount. DGDM is designed to fusion the image feature and text knowledge. It can further extent to incoperate information other than only image (e.g., year, landsacpe description) to improve its ability. With its innovative architecture, sets the stage to address these needs, offering a solution that is as agile as it is accurate. The potential applications are myriad: from guiding urban development with an environmentally conscious approach to enabling disaster resilience by rapid assessment of calamities. The model could serve as the cornerstone for smart city initiatives, where real-time data fusion from various sensors informs sustainable urban growth, and where community engagement in validating CD results fosters a participatory approach to urban planning.

In the era of swift urban transformations, the demand for precise, efficient, and scalable CD methodologies escalates. Our proposed method DGDM, with its innovative features and high accuracy in CD tasks, stands ready to meet this pressing urban challenge.

V. CONCLUSION

As cities continue to sprawl at an unprecedented pace, the ability to monitor and understand these changes is becoming ever more vital. Remote sensing technologies for urban CD are at the forefront of this challenge. Our latest research introduces a breakthrough in this vital field, blending the advanced capabilities of a learned diffusion model with a diverse suite of pre-training technologies. This innovative approach enables us to capture the nuanced dynamics of urban expansion with greater precision than ever before. Our DGDM is a testament to the power of this fusion, marrying the inherent strengths of diffusion models with the latest developments in text-to-image conversion. By doing so, the DGDM not only carves a path for future exploration in urban CD tasks but also sets a new benchmark for the discipline. We have achieved this by integrating VQ-GAN, MAE, and an intricate ITT adapter. The well-designed orchestration of these advanced technologies within the DGDM framework pushes it to the forefront, surpassing existing state-of-the-art methods. This assertion is backed by robust empirical data gathered from a comprehensive range of urban CD datasets. The practical applications of our research are manifold. From improving urban planning and management to aiding in disaster response, the implications of our work are profound. By increasing both the precision and automation of CD techniques, we can offer planners and policymakers tools that were previously unimaginable.

Looking forward, we envision our research serving as a springboard for further exploration. The potential to refine diffusion models for even more complex urban environments presents an avenue for future studies. It is our hope that the DGDM will encourage a new wave of innovation in remote sensing, leading to smarter, more sustainable cities. In sum, our research not only represents a improvment in urban remote sensing capabilities but also serves as a clarion call for the community to continue pushing the boundaries of what these technologies can achieve.

REFERENCES

[1] A. Singh, "Review article digital change detection techniques using remotely-sensed data," *Int. J. Remote Sens.*, vol. 10, no. 6, pp. 989–1003, Jun. 1989.

[2] Z. Zheng, Y. Zhong, J. Wang, A. Ma, and L. Zhang, "Building damage assessment for rapid disaster response with a deep object-based semantic change detection framework: From natural disasters to man-made disasters," *Remote Sens. Environ.*, vol. 265, Nov. 2021, Art. no. 112636.

[3] L. Moya, A. Muhari, B. Adriano, S. Koshimura, E. Mas, L. R. Marval-Perez, and N. Yokoya, "Detecting urban changes using phase correlation and ℓ_1-based sparse model for early disaster response: A case study of the 2018 Sulawesi Indonesia earthquake-tsunami," *Remote Sens. Environ.*, vol. 242, Jun. 2020, Art. no. 111743.

[4] H. Jiang, M. Peng, Y. Zhong, H. Xie, Z. Hao, J. Lin, X. Ma, and X. Hu, "A survey on deep learning-based change detection from high-resolution remote sensing images," *Remote Sens.*, vol. 14, no. 7, p. 1552, Mar. 2022.

[5] R. Liu, M. Kuffer, and C. Persello, "The temporal dynamics of slums employing a CNN-based change detection approach," *Remote Sens.*, vol. 11, no. 23, p. 2844, Nov. 2019.

[6] M. K. Firozjaei, A. Sedighi, M. Kiavarz, S. Qureshi, D. Haase, and S. K. Alavipanah, "Automated built-up extraction index: A new technique for mapping surface built-up areas using Landsat 8 OLI imagery," *Remote Sens.*, vol. 11, no. 17, p. 1966, Aug. 2019.

[7] L. Li, P. Liu, J. Wu, L. Wang, and G. He, "Spatiotemporal remote-sensing image fusion with patch-group compressed sensing," *IEEE Access*, vol. 8, pp. 209199–209211, 2020.

[8] H. Li and P. Tang, "Dps-MuSyQ: A distributed parallel processing system for multi-source data synergized quantitative remote sensing products producing," *IEEE Access*, vol. 8, pp. 79510–79520, 2020.

[9] N. Kim, S.-S. Han, and C.-S. Jeong, "ADOM: ADMM-based optimization model for stripe noise removal in remote sensing image," *IEEE Access*, vol. 11, pp. 106587–106606, 2023.

[10] D. Yu and C. Fang, "Urban remote sensing with spatial big data: A review and renewed perspective of urban studies in recent decades," *Remote Sens.*, vol. 15, no. 5, p. 1307, Feb. 2023.

[11] A. Asokan and J. Anitha, "Change detection techniques for remote sensing applications: A survey," *Earth Sci. Informat.*, vol. 12, no. 2, pp. 143–160, Jun. 2019.

[12] S. Sheng and H. Lian, "The spatial pattern evolution of rural settlements and multi-scenario simulations since the initiation of the reform and opening up policy in China," *Land*, vol. 12, no. 9, p. 1763, Sep. 2023.

[13] Z. Zhang, G. Vosselman, M. Gerke, D. Tuia, and M. Y. Yang, "Change detection between multimodal remote sensing data using Siamese CNN," 2018, *arXiv:1807.09562*.

[14] P. de Bem, O. de Carvalho Junior, R. F. Guimarães, and R. T. Gomes, "Change detection of deforestation in the Brazilian Amazon using Landsat data and convolutional neural networks," *Remote Sens.*, vol. 12, no. 6, p. 901, Mar. 2020.

[15] D. Lu, P. Mausel, E. Brondizio, and E. Moran, "Change detection techniques," *Int. J. Remote Sens.*, vol. 25, no. 12, pp. 2365–2401, Apr. 2004.

[16] R. Qin, J. Tian, and P. Reinartz, "3D change detection—Approaches and applications," *ISPRS J. Photogramm. Remote Sens.*, vol. 122, pp. 41–56, Dec. 2016.

[17] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[18] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[19] 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)*, Oct. 2018, pp. 4063–4067.

[20] 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)*, Jul. 2022, pp. 207–210.

[21] Y. Long, G.-S. Xia, S. Li, W. Yang, M. Y. Yang, X. X. Zhu, L. Zhang, and D. Li, "On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-AID," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4205–4230, 2021.

[22] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl, "Bigearthnet: A large-scale benchmark archive for remote sensing image understanding," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2019, pp. 5901–5904.

[23] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5920416.

[24] 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, 2022.

[25] B. Wang, J. Zhang, R. Zhang, Y. Li, L. Li, and Y. Nakashima, "Improving facade parsing with vision transformers and line integration," 2023, *arXiv:2309.15523*.

[26] W. G. C. Bandara, N. G. Nair, and V. M. Patel, "DDPM-CD: Remote sensing change detection using denoising diffusion probabilistic models," 2022, *arXiv:2206.11892*.

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

[28] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9.

[29] 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)*, Jun. 2022, pp. 10674–10685.

[30] W. Zhao, Y. Rao, Z. Liu, B. Liu, J. Zhou, and J. Lu, "Unleashing text-to-image diffusion models for visual perception," in *Proc. Int. Conf. Comput. Vis.*, 2023, pp. 5729–5739.

[31] J. Zhang, T. Fukuda, and N. Yabuki, "Automatic object removal with obstructed Façades completion using semantic segmentation and generative adversarial inpainting," *IEEE Access*, vol. 9, pp. 117486–117495, 2021.

[32] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12868–12878.

[33] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16000–16009.

[34] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2183–2195, Apr. 2018.

[35] Y. Cao, X. Huang, and Q. Weng, "A multi-scale weakly supervised learning method with adaptive online noise correction for high-resolution change detection of built-up areas," *Remote Sens. Environ.*, vol. 297, Nov. 2023, Art. no. 113779.

[36] J. S. Deng, K. Wang, Y. H. Deng, and G. J. Qi, "PCA-based land-use change detection and analysis using multitemporal and multisensor satellite data," *Int. J. Remote Sens.*, vol. 29, no. 16, pp. 4823–4838, Aug. 2008.

[37] A. A. Nielsen, K. Conradsen, and J. J. Simpson, "Multivariate alteration detection (MAD) and MAF postprocessing in multispectral, bitemporal image data: New approaches to change detection studies," *Remote Sens. Environ.*, vol. 64, no. 1, pp. 1–19, Apr. 1998.

[38] S. Saha, F. Bovolo, and L. Bruzzone, "Unsupervised deep change vector analysis for multiple-change detection in VHR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 3677–3693, Jun. 2019.

[39] A. A. Nielsen, "The regularized iteratively reweighted MAD method for change detection in multi- and hyperspectral data," *IEEE Trans. Image Process.*, vol. 16, no. 2, pp. 463–478, Feb. 2007.

[40] Y. You, J. Cao, and W. Zhou, "A survey of change detection methods based on remote sensing images for multi-source and multi-objective scenarios," *Remote Sens.*, vol. 12, no. 15, p. 2460, Jul. 2020.

[41] B. Wang, L. Li, Y. Nakashima, R. Kawasaki, H. Nagahara, and Y. Yagi, "Noisy-LSTM: Improving temporal awareness for video semantic segmentation," *IEEE Access*, vol. 9, pp. 46810–46820, 2021.

[42] C. Zhang, P. Yue, D. Tapete, L. Jiang, B. Shangguan, L. Huang, and G. Liu, "A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 166, pp. 183–200, Aug. 2020.

[43] J. Chen, Z. Yuan, J. Peng, L. Chen, H. Huang, J. Zhu, Y. Liu, and H. Li, "DASNet: Dual attentive fully convolutional Siamese networks for change detection in high-resolution satellite images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1194–1206, 2021.

[44] D. Wang, X. Chen, M. Jiang, S. Du, B. Xu, and J. Wang, "ADS-Net: An attention-based deeply supervised network for remote sensing image change detection," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 101, Sep. 2021, Art. no. 102348.

[45] B. Wang, L. Li, M. Verma, Y. Nakashima, R. Kawasaki, and H. Nagahara, "MTUNet: Few-shot image classification with visual explanations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 2294–2298.

[46] B. Wang, L. Li, M. Verma, Y. Nakashima, R. Kawasaki, and H. Nagahara, "Match them up: Visually explainable few-shot image classification," *Appl. Intell.*, vol. 53, no. 9, pp. 10956–10977, 2023.

[47] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[48] O. Mañas, A. Lacoste, X. Giró-i-Nieto, D. Vazquez, and P. Rodríguez, "Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9414–9423.

[49] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.

[50] S. Hao, Y. Zhou, and Y. Guo, "A brief survey on semantic segmentation with deep learning," *Neurocomputing*, vol. 406, pp. 302–321, Sep. 2020.

[51] C. Godard, O. M. Aodha, M. Firman, and G. Brostow, "Digging into self-supervised monocular depth estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3828–3838.

[52] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross attention control," 2022, *arXiv:2208.01626*.

[53] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. 25th Int. Conf. Mach. Learn. (ICML)*, 2008, pp. 1096–1103.

[54] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2536–2544.

[55] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Proc. 14th Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands. Cham, Switzerland: Springer, Sep. 2016, pp. 649–666.

[56] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[57] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16 × 16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[58] C. Wei, H. Fan, S. Xie, C.-Y. Wu, A. Yuille, and C. Feichtenhofer, "Masked feature prediction for self-supervised visual pre-training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14668–14678.

[59] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, and I. Sutskever, "Generative pretraining from pixels," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1691–1703.

[60] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu, "SimMIM: A simple framework for masked image modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9653–9663.

[61] 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. Intervent.*, Munich, Germany. New York, NY, USA: Springer, 2015, pp. 234–241.

[62] Y. Xu, W. Yu, P. Ghamisi, M. Kopp, and S. Hochreiter, "Txt2Img-MHN: Remote sensing image generation from text using modern Hopfield networks," *IEEE Trans. Image Process.*, vol. 32, pp. 5737–5750, 2023.

[63] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *Int. J. Comput. Vis.*, vol. 130, no. 9, pp. 2337–2348, Sep. 2022.

[64] P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, and Y. Qiao, "CLIP-adapter: Better vision-language models with feature adapters," *Int. J. Comput. Vis.*, pp. 1–15, Sep. 2023.

[65] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, Jan. 2019.

[66] 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, May 2020.

[67] M. A. Lebedev, Y. V. Vizilter, O. V. Vygolov, V. A. Knyaz, and A. Y. Rubis, "Change detection in remote sensing images using conditional adversarial networks," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 42, pp. 565–571, May 2018.

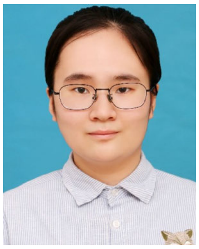
[68] Y. Liu, C. Pang, Z. Zhan, X. Zhang, and X. Yang, "Building change detection for remote sensing images using a dual-task constrained deep Siamese convolutional network model," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 811–815, May 2021.

[69] G. Batzolis, J. Stanczuk, C.-B. Schönlieb, and C. Etmann, "Conditional image generation with score-based diffusion models," 2021, *arXiv:2111.13606*.

[70] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.

RAN WAN is currently an Associate Professor with the Department of Industrial Design, School of Architecture and Design, Nanchang University. He is also the Deputy Dean of the National Industrial Design Institute, specializing in the field of ecological design. His work is dedicated to the fusion of design and technology and advocating for the innovative D+X (design and multiple disciplines integration) design model. His research interests include smart city, deep learning, and remote sensing.

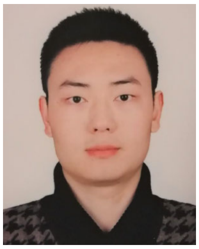
JIAXIN ZHANG received the bachelor's degree in architecture from Nanchang University, in 2016, the master's degree from Southeast University, in 2019, and the Ph.D. degree from the School of Energy and Environment, Osaka University, in 2022. Since November 2022, he has been a specially-appointed Researcher with the Laboratory of Environmental Design and Information Technology, Osaka University. His research interests include city perception, integrating machine learning in urban studies, and pioneering in automated measurements of architectural facades. His work reflects his architectural acumen and passion for innovative technology.



YIYING HUANG is currently pursuing the degree in architecture with the School of Architecture and Design, Nanchang University. Her academic pursuits are centered around the digitalization of architecture. She has been actively involved in research projects focusing on human settlements and has participated in rural revitalization work camps, taking on the role of the Team Leader.

YUNQIN LI received the bachelor's degree from Nanchang University, in 2016, the master's degree from Southeast University, in 2019, and the Ph.D. degree from Osaka University, in 2022. Currently, she is a Visiting Researcher with Osaka University. Her research interests include spatial auditing, measuring, perception, understanding, interaction, backed by new data, technologies, methods, street perception, and explainable machine learning. Despite her achievements, she remains humble in her continuous exploration of the intricate relationship between technology, space, and urban environments.

BOYA HU is currently pursuing the Graduate degree with the College of Architecture and Urban Planning, Tongji University. She is also a part of the Architecture and Urban Space Team, focusing her research on architectural criticism and intelligent design.



BOWEN WANG (Member, IEEE) received the B.S. degree in computer science from Anhui University, China, in 2016, and the M.S. degree in medical information and the Ph.D. degree in computer science from Osaka University, Japan, in 2020 and 2023, respectively. He is currently a specially-appointed Researcher with the Institute for Datability Science (IDS), Osaka University. His research interests include computer vision, explainable AI, city perception, and medical AI. He is a member of ACM and IPSJ. He has received the Best Paper Award in APAMI 2020.