RS-Mamba for Large Remote Sensing Image Dense Prediction

Sijie Zhao, *Graduate Student Member, IEEE*, Hao Chen, Xueliang Zhang, *Senior Member, IEEE*,
Pengfeng Xiao, *Senior Member, IEEE*, Lei Bai, and Wanli Ouyang

***Abstract*— Context modeling is critical for remote sensing image dense prediction tasks. Nowadays, the growing size of very-high-resolution (VHR) remote sensing images poses challenges in effectively modeling context. While transformer-based models possess global modeling capabilities, they encounter computational challenges when applied to large VHR images due to their quadratic complexity. The conventional practice of cropping large images into smaller patches results in a notable loss of contextual information. To address these issues, we propose the remote sensing Mamba (RSM) for dense prediction tasks in large VHR remote sensing images. RSM is specifically designed to capture the global context of remote sensing images with linear complexity, facilitating the effective processing of large VHR images. Considering that the land covers in remote sensing images are distributed in arbitrary spatial directions due to characteristics of remote sensing over-head imaging, the RSM incorporates an omnidirectional selective scan module (OSSM) to globally model the context of images in multiple directions, capturing large spatial features from various directions. We designed simple yet effective models based on RSM, achieving state-of-the-art performance on dense prediction tasks in VHR remote sensing images without fancy training strategies. Extensive experiments on semantic segmentation (SS) and change detection (CD) tasks across various land covers demonstrate the effectiveness of the proposed RSM. Leveraging the linear complexity and global modeling capabilities, RSM achieves better efficiency and accuracy than transformer-based models on large remote sensing images. Interestingly, we also demonstrated that our model generally performs better with a larger image size on dense prediction tasks.**

***Index Terms*— Change detection (CD), deep learning, dense prediction, large remote sensing images, semantic segmentation (SS), state space model (SSM), very high resolution (VHR).**

Manuscript received 10 April 2024; revised 17 June 2024; accepted 5 July 2024. Date of publication 9 July 2024; date of current version 31 July 2024. This work was supported in part by the Shanghai Artificial Intelligence Laboratory, in part by the National Natural Science Foundation of China under Grant 42071297, in part by the AI and AI for Science Project of Nanjing University under Grant 02091480605203, in part by the Fundamental Research Funds for the Central Universities under Grant 020914380119, and in part by the Youth Innovation Team of China Meteorological Administration under Grant CMA2024QN02. *(Corresponding authors: Hao Chen; Xueliang Zhang.)*

Sijie Zhao, Xueliang Zhang, and Pengfeng Xiao are with Jiangsu Provincial Key Laboratory of Geographic Information Science and Technology, Key Laboratory for Land Satellite Remote Sensing Applications of Ministry of Natural Resources, School of Geography and Ocean Science, Nanjing University, Nanjing 210023, China (e-mail: zsj@smail.nju.edu.cn; zxl@nju.edu.cn; xiaopf@nju.edu.cn).

Hao Chen, Lei Bai, and Wanli Ouyang are with Shanghai Artificial Intelligence Laboratory, Shanghai 200000, China (e-mail: justchenhao@buaa.edu.cn; bailei@pjlab.org.cn; ouyangwanli@pjlab.org.cn).

Our code is available at https://github.com/walking-shadow/Official_Remote_Sensing_Mamba.

Digital Object Identifier 10.1109/TGRS.2024.3425540

I. Introduction

The advent of increasingly high spatial resolution in remote sensing images has marked a transformative period in the field, facilitating a deeper understanding and more nuanced analysis across a multitude of applications. These high-resolution images serve as pivotal resources in various domains, including urban planning [1], agricultural management [2], environmental monitoring [3], image super-resolution [4], [5], and disaster response [6].

Due to the unprecedented availability of very-high-resolution (VHR) images, the field of remote sensing has undergone rapid expansion in recent years. VHR remote sensing images are characterized by rich contextual information, which is crucial for dense prediction tasks such as semantic segmentation (SS) and change detection (CD). In these images, due to the very high spatial resolution, there is a wealth of spatial features within individual objects and among multiple objects, which often span large spatial scales. Additionally, since remote sensing images are captured from a downward-looking camera, the camera can acquire images from any direction, indicating that the spatial features of these images can exist in any direction. Therefore, the ability to globally model the context of VHR remote sensing images and extract large spatial features from multiple directions is essential for dense prediction tasks in VHR remote sensing.

In recent years, deep learning models based on transformers [7] have been widely applied to VHR remote sensing dense prediction tasks [8], [9], [10], [11]. The transformer architecture, famous for its ability to capture global contextual information and model spatial dependencies effectively through self-attention mechanisms, has achieved impressive results in this domain. However, due to the quadratic complexity of transformers, training and inference with these models on large VHR remote sensing images necessitate dividing these images into smaller patches, as shown in Fig. 1. This preprocessing step inevitably results in each patch containing only a portion of an object, offering limited contextual information. Consequently, the loss of internal spatial features within individual objects and the spatial dependencies among multiple objects can adversely affect the performance of VHR remote sensing tasks. It is important to note that this preprocessing step differs from the patch embedding used in transformer-based models. The former involves segmenting a large remote sensing image into multiple smaller images, which are then individually fed into the model for training. Since there is no

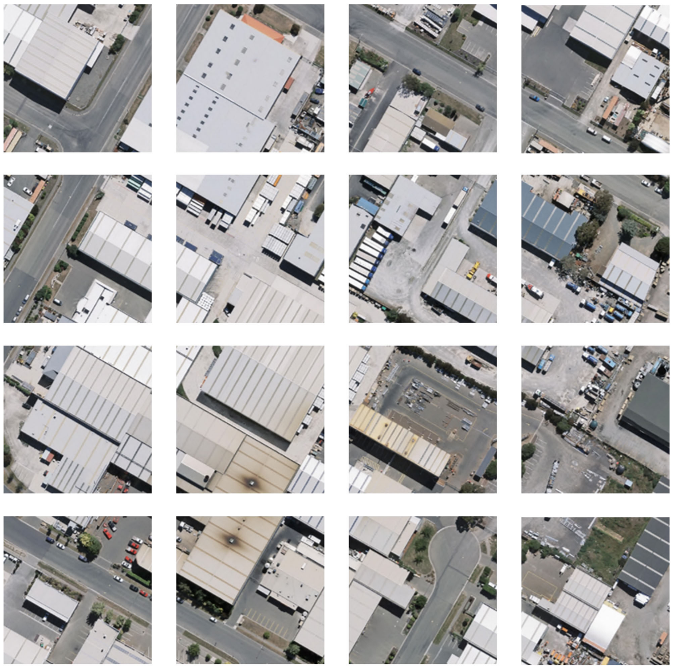


Fig. 1. Illustration of the large image preprocessing strategy of transformer-based models. Dividing a large VHR remote sensing image into small patches results in the loss of many spatial features. Each patch contains very limited contextual information compared to the original large image.

information linkage between these smaller images, the model can only capture limited contextual information from each one. In contrast, the latter method involves feeding the entire remote sensing image into the model, which then segments it into several patches. These patches are interrelated through self-attention mechanisms, enabling the model to acquire global contextual information from the remote sensing image. This limitation underscores the need for innovative solutions that can efficiently process whole images or large segments to preserve and leverage the rich contextual information inherent in large VHR remote sensing images.

The recent work Mamba [12] integrated time-varying parameters into the state space model (SSM) and proposed a hardware-aware algorithm that facilitates highly efficient training and inference processes. The SSM, which draws inspiration from the classical Kalman filter model [13], excels at capturing long-range dependencies and benefits from hardware efficient implementation. Research on Mamba has illustrated its potential as a promising alternative to transformers in language modeling due to its robust contextual modeling capacity and linear complexity [12]. However, Mamba is designed to process data along a specific direction, where preceding data cannot establish connections with subsequent data. This directional processing limitation renders it less applicable for image data, which lack a specific orientation and where spatial relationships are crucial across all dimensions.

Recent works such as Vim [14] and VMamba [15] have harnessed SSM to achieve linear complexity and a global effective receptive field, tackling tasks such as image classification and segmentation on natural images. To address the challenge of image data nondirectionality, Vim proposed SS1D module which employs SSM for selective scanning in both the forward and backward directions along the horizontal axis of an image. VMamba extends this approach with SS2D module which conducts selective scanning with SSM in both the horizontal and vertical directions, ensuring that every segment of the image can establish connections with other parts from both the horizontal and vertical directions. The visualization of the effective receptive field in VMamba shows that the effective receptive field is enhanced across both the horizontal and vertical directions [15]. This indicates that the selective scanning direction of the SSM can significantly impact the effective receptive field in specific directions.

However, Vim and VMamba are not ideally suited for VHR remote sensing images. Objects in natural images are captured from an eye-level view and adhere to the law of gravitation, which means that their main spatial features are distributed in the horizontal and vertical directions. For example, a natural image of a cat sitting on the ground cannot be rotated arbitrarily, as the cat's posture needs to conform to the law of gravity. In contrast, remote sensing images can be rotated freely as they are captured from a top-down satellite perspective, which means that their main spatial features can be distributed in any direction. Given that objects within VHR remote sensing images often span large spatial scales, the spatial features of individual objects and the dependencies among multiple objects can vary in direction. Therefore, VHR remote sensing images contain large spatial features in multiple directions. Due to the impact of the SSM' selective scanning direction on the effective receptive field in specific orientations, Vim's horizontal scanning and VMamba's horizontal and vertical scanning, while effective for natural images with primary features along these axes, cannot adequately address the diverse directional large spatial features inherent in VHR remote sensing images.

To address the aforementioned challenges, we introduce SSM to VHR remote sensing dense prediction tasks for the first time, aiming to achieve global modeling capability and linear complexity. We propose remote sensing Mamba (RSM) to process VHR remote sensing images, leveraging the strengths of SSM to extract large and multidirectional spatial features in VHR remote sensing images with rich contextual information.

The RSM can globally model the context of large VHR remote sensing images without self-attention operations. Furthermore, rather than dividing large VHR remote sensing images into small patches, the RSM can handle whole images without losing contextual information due to its linear complexity. Thus, the RSM is suitable for efficiently handling VHR remote sensing images.

Furthermore, we propose the omnidirectional selective scan module (OSSM) to extract large spatial features from multiple directions in VHR remote sensing images. The OSSM employs SSM for selective scanning in the forward and backward directions across the horizontal, vertical, diagonal, and anti-diagonal axes. This approach enhances the global modeling capability of the contextual information in multiple directions, allowing for the extraction of comprehensive global spatial features.

In summary, our contributions are as follows.

1) We introduce the SSM for the first time to perform dense prediction tasks in VHR remote sensing. The proposed RSM employs SSM to process VHR remote sensing images with linear complexity, enabling direct handling of large remote sensing images without the necessity of segmenting them into small patches, which preserves the rich contextual information inherent in remote sensing images.
2) We design an OSSM to extract large spatial features in multiple directions. Considering that the land covers can be distributed in any spatial direction as remote sensing images are captured from a top-down perspective, RSM utilizes OSSM to selectively scan remote sensing images in multiple directions, thereby modeling the global context of remote sensing images.
3) We demonstrate the efficiency and superiority of the RSM in VHR remote sensing tasks. Experiments on the SS datasets (WHU and Massachusetts Road) and the CD datasets (LEVIR-CD and WHU-CD) show that the RSM achieves state-of-the-art performance on both SS and CD tasks.

II. Related Works

A. Dense Prediction of VHR Remote Sensing

Dense prediction tasks in VHR remote sensing mainly include SS and CD tasks, where remote sensing images have very high spatial resolution and abundant contextual information. Large remote sensing images span large spatial ranges, thereby offering rich contextual information, which is crucial for dense prediction tasks. Experiments in FCCDN [16] and SwinB-CNN [17] show that models utilizing larger remote sensing images can achieve superior performance in dense prediction tasks. This underscores the importance of leveraging rich contextual information in large remote sensing images for dense prediction tasks.

There are three main types of deep learning models for VHR remote sensing dense prediction tasks: convolutional neural network (CNN)-based models, transformer-based models, and CNN-transformer hybrid-based models.

CNN-based models excel in image processing due to their ability to efficiently capture local spatial features through their hierarchical structure. Papadomanolaki et al. [18] proposed an urban CD framework that combines U-Net [19] for feature extraction and LSTMs [20] for temporal modeling. Yue et al. [21] proposed an adaptive network to increase the classification rate at the pixel level based on a deep semantic model infrastructure. Zhao et al. [22] introduced a novel CD framework named EDED, which operates by exchanging features between two encoder branches. This approach enables the separate identification of changed objects in bitemporal images to produce CD results. CNN-based models often incorporate attention modules to focus on important areas within images, thereby enhancing feature extraction capabilities. Yang et al. [23] proposed a multipath attention-fused block module to fuse multipath features, and a refinement attention-fused block module to fuse high-level abstract features and low-level spatial features. Han et al. [24] proposed a change guide module for CD, which can effectively capture the long-distance dependency among pixels and effectively overcomes the problem of the insufficient receptive field of traditional CNNs.

However, the inherently large spatial scales of objects in VHR remote sensing images pose a challenge for CNN models. Due to their limited ability to capture global receptive fields, CNNs struggle to extract comprehensive global spatial features and dependencies within these images. Conversely, transformers excel at global context modeling across entire images by using self-attention mechanisms, thereby overcoming the limitations of CNNs in capturing global spatial relationships. This attribute has led to the widespread application of transformer-based models in VHR remote sensing tasks, demonstrating their ability to perform SS and CD with an enhanced understanding of the global context. Bandara and Patel [9] proposed a transformer-based Siamese network architecture for CD, which unified a hierarchically structured transformer encoder with a multilayer perception decoder in a Siamese network architecture to efficiently render multiscale long-range details. Zhang et al. [11] proposed a purely transformer-based architecture for CD tasks, constructing a model based on the swin transformer [25] architecture.

The global modeling capabilities of transformer-based models have led to their widespread application in dense prediction tasks for remote sensing. However, their quadratic complexity and substantial computational demands pose challenges for model training and inference. In addition to the global features of remote sensing images, local features are equally important for dense prediction tasks. Consequently, CNN-transformer hybrid-based models have been introduced and extensively applied to dense prediction tasks in remote sensing. On the one hand, hybrid-based models can reduce the number of parameters and computations compared to transformer-based models, which facilitates more efficient training and inference. Chen et al. [8] integrated transformers into CD tasks, utilizing a ResNet [26] as the encoder and self-attention modules as decoders. On the other hand, these models are capable of leveraging both the local features extracted by CNNs and the global features extracted by transformers, thus paying attention to both local and global aspects of images for dense prediction tasks. Li et al. [27] proposed an encoding–decoding hybrid transformer model for CD, which has the advantages of both transformers and UNet [19] in learning global context and low-level details.

Despite their strengths, transformer-based models and hybrid-based models still encounter challenges due to the quadratic complexity of their self-attention mechanisms, particularly when processing large VHR remote sensing images. This necessitates transformer-based models and hybrid-based models to divide large images into smaller segments, which can result in a significant loss of contextual information. While transformers can model global contexts, the reduced contextual information limits their effectiveness in VHR remote sensing dense prediction tasks. In response, we propose the SSM-based RSM for VHR remote sensing dense prediction tasks, which has linear complexity and global modeling capability.

The RSM is adept at handling large images with rich contextual information, thereby providing a more effective solution for processing VHR remote sensing images.

B. State Space Models

SSMs have gained significant traction in the field of deep learning in recent years, marking a remarkable evolution in the way long-range dependencies and sequential data are handled [28], [29], [30]. Initially inspired by their traditional application in control systems, SSMs were innovatively adapted to deep learning, leveraging the strengths of continuous state spaces to model complex temporal dynamics. The integration of SSM into deep learning was catalyzed by the introduction of the highest polynomial powered operator (HiPPO) initialization method [31], which significantly improved the models' ability to capture long-range dependencies.

The LSSL model demonstrated the potential of SSM in addressing long-range dependency challenges, setting a foundation for subsequent research in the field [29]. However, LSSL faces critical hurdles related to computational and memory efficiency, limiting its practical application. Addressing these limitations, the S4 model introduced by Gu et al. [28] emerged as a pivotal advancement, proposing a normalized parameterization strategy that reduced computational overhead, thereby making SSM more feasible for practical applications.

Following the breakthrough of the S4 model, the landscape of SSM research expanded rapidly, with several variants being developed to enhance the model's structure and efficiency. Notable among these are models incorporating complex-diagonal structures to improve temporal modeling capabilities [32], [33], as well as those supporting multiple-input multiple-output configurations to increase model flexibility [30]. Additionally, innovations such as the decomposition of operations into diagonal plus low-rank structures [34] and the introduction of selection mechanisms [12] have further refined the adaptability and performance of SSMs.

However, the aforementioned models are only capable of processing unidirectional sequence data and cannot handle image data that lack a specific direction. Recent works [14], [15], [35] have achieved global modeling of the context of images using SSM by conducting selective scanning both forward and backward in certain directions. Vim [14] proposed SS1D module to perform selective scanning in the horizontal direction, enabling each part of the image to perceive global information. VMamba [15] extends this with SS2D module by conducting selective scanning both horizontally and vertically, enhancing the model's global effective receptive field in both dimensions.

Nevertheless, since the primary spatial features of natural images are distributed in both the horizontal and vertical directions, and VHR remote sensing images exhibit large spatial features in multiple directions, although Vim and VMamba achieved good performance on natural images, they are not suitable for VHR remote sensing images. The proposed omnidirectional selective scan module conducts selective scanning in multiple directions, and is capable of capturing the large spatial features of VHR remote sensing images in various directions.

III. Methodology

A. Preliminaries: SSMs

In the realm of deep learning, SSMs have gained prominence for their ability to encapsulate dynamic systems that map an input sequence $x(t) \in \mathbb{R}^L$ to an output $y(t) \in \mathbb{R}$. SSMs are grounded in the principles of control theory and are defined by a set of linear ordinary differential equations (ODEs)

$$h'(t) = Ah(t) + Bx(t) \tag{1}$$

$$y(t) = Ch(t) + Dx(t) \tag{2}$$

where $A \in \mathbb{C}^{N\times N}$, $B \in \mathbb{R}^{N\times L}$, $C \in \mathbb{R}^N$, and $D \in \mathbb{R}^L$ are the system matrices and $h(t)$ denotes the hidden state vector at time t.

The model's state-transition matrix A governs the evolution of the state vector $h(t)$, while the input matrix B, output matrix C, and feedthrough matrix D articulate the relationships between the input $x(t)$, state $h(t)$, and output $y(t)$, respectively. In discrete-time settings, which are typical in deep learning applications, these continuous equations must be discretized for computational tractability and alignment with data sampling rates.

The discretization of SSM involves transforming the continuous ODE into a discrete-time representation. Employing a zero-order hold on the input signal, the discrete-time SSM can be represented as

$$h_k = \Phi h_{k-1} + \Gamma x_k \tag{3}$$

$$y_k = Ch_k + Dx_k \tag{4}$$

where h_k is the hidden state at discrete time step k, y_k is the output, $\Phi = e^{A\Delta T}$ is the state transition matrix for time step ΔT, and Γ is derived as $\Gamma = (e^{A\Delta T} - I)A^{-1}B$, assuming that the input remains constant over each interval ΔT.

The Mamba [12] methodology distinguishes itself within the SSM framework by adopting a selective scan mechanism. This mechanism enhances the standard SSM structure by permitting dynamic adjustments to system matrices B and D, based on the current and historical context of the input sequence. Consequently, Mamba's SSM can model complex temporal dynamics more effectively, as these matrices adapt in response to the evolving features of the input data.

B. Overall Architecture

VHR remote sensing images are primarily utilized in SS and CD tasks. Consequently, we have developed two specialized frameworks: RSM for SS (RSM-SS) for the SS task and RSM for CD (RSM-CD) for the CD task, as illustrated in Fig. 2. To demonstrate the effectiveness of SSM in processing VHR remote sensing images, RSM-SS and RSM-CD employ the simplest architectures for SS and CD, respectively. Our objective is to show that RSM can achieve state-of-the-art

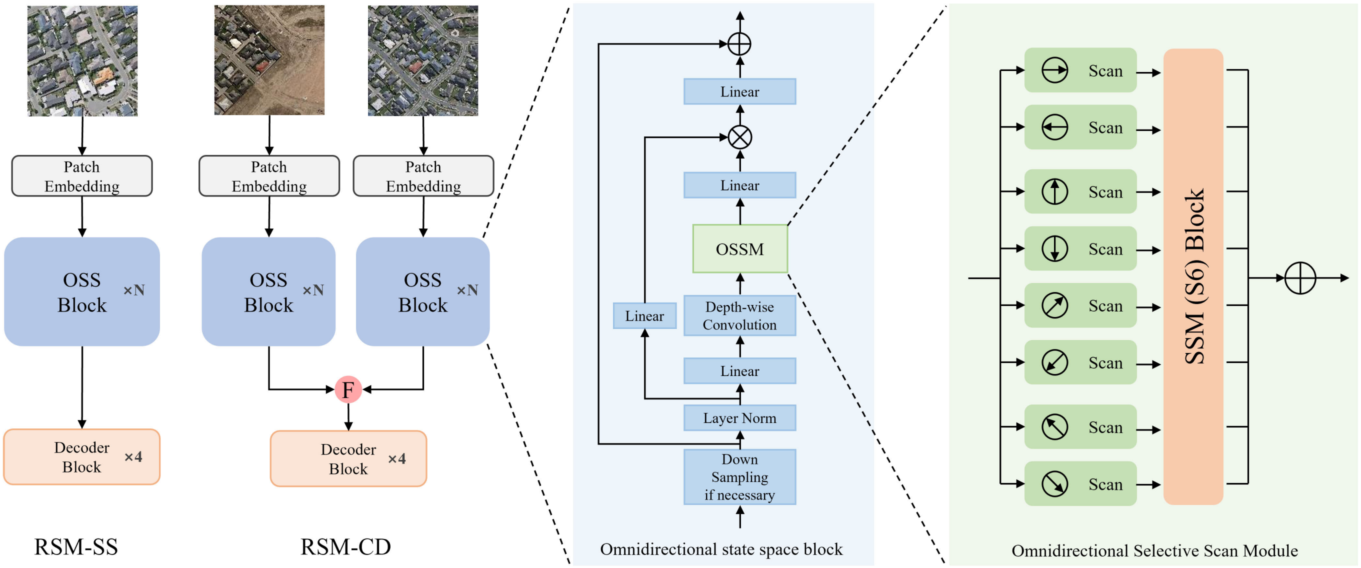


Fig. 2. Illustration of the overall structure of RSM-SS and RSM-CD. RSM-SS and RSM-CD can globally model the context of images in multiple directions with linear complexity using the omnidirectional selective scan.

performance with the simplest architectures, thus demonstrating the substantial potential of SSM-based models for remote sensing dense prediction tasks.

The RSM-SS architecture utilizes the U-Net [19] encoder–decoder framework, where input VHR remote sensing images are first transformed into a sequence of image patches through patch embedding. These patches are then fed into the encoder to extract features, which are subsequently upsampled by the decoder to produce SS results. The encoder consists of five stages, each comprising several omnidirectional state space (OSS) blocks. Stage 1 extracts features from the input VHR remote sensing images, while stages 2–5 progressively downsample the encoder features and double the number of feature channels at each stage. The decoder is composed of four decoder blocks, where features are upsampled and then concatenated with the encoder features along the channel dimension through skip connections followed by convolution. This process fuses the semantic information of decoder features with the spatial information of encoder features, facilitating SS from both global and local perspectives.

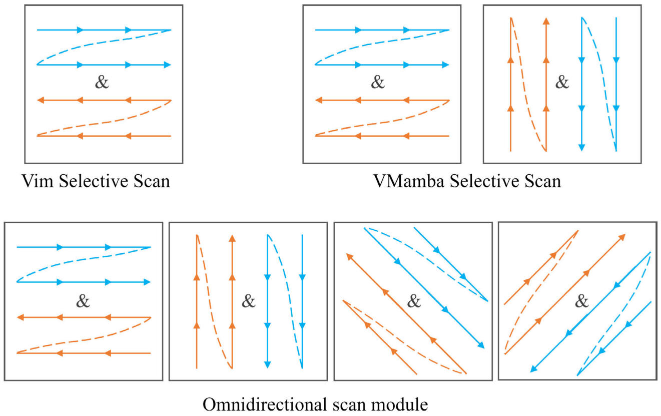
The RSM-CD employs an FC-Siam-Conc [36] Siamese network architecture. Bitemporal VHR remote sensing images are first converted into bitemporal sequences of image patches using patch embedding, which are then fed into bitemporal encoders with shared weights to extract features. These bitemporal encoder features are simply fused and upsampled in a single decoder to obtain the final CD results. Similar to RSM-SS, the shared-weight encoders in RSM-CD consist of five stages with several OSS blocks each, and the decoder comprises four decoder blocks. After feature extraction by the shared-weight encoders, bitemporal features of the same size are concatenated along the channel dimension and convolved. This fusion captures the information of both temporal phases of VHR remote sensing images, enabling the effective segmentation of changed objects. The fused features are upsampled in the decoder and concatenated with fusion features of the same size through skip connections and convolution, thus preserving rich semantic and spatial information.

C. OSS Block

The OSS block is a novel feature extraction unit designed for SS and CD tasks in VHR remote sensing images, as shown in Fig. 2. Central to the OSS block is the oriented scanning module, which serves as the core for global contextual modeling across multiple orientations within an image. The OSSM selectively scans the input image in various directions, capturing the intricate spatial relationships and providing a comprehensive understanding of the context.

The architecture begins with a layer normalization that standardizes the input data, enhancing the model training stability. Following this, a linear transformation adjusts the dimensionality of the data, preparing it for the depth-wise convolution process. This convolution operates on each input channel separately, reducing the parameter count and focusing on extracting spatial features. After convolution, the features pass through the OSSM. The OSSM performs selective scanning on the features in the forward and backward directions along the horizontal, vertical, diagonal, and anti-diagonal directions, which are then added together, as shown in Fig. 2. The output from the OSSM then undergoes a linear transformation and a gating operation, adjusting deep features with outputs of a linear transformation of the normalized features. Finally, the features are passed to a final linear layer, which are then added with input features through a residual connection.

The OSS block is crafted with a keen focus on balancing computational efficiency and the ability to extract rich spatial features from VHR remote sensing images. Therefore, as the OSS block is efficient and lightweight, we can stack more blocks with a similar total model depth budget when building the model.



Vim Selective Scan

VMamba Selective Scan

Omnidirectional scan module

Fig. 3. Illustration of the selective scan directions of Vim, VMamba, and OSSM.

D. Omnidirectional Selective Scan Module

Vim and VMamba have demonstrated commendable performance in natural images, where primary spatial features are distributed along the horizontal and vertical directions. Vim used SS1D module to conduct selective scanning in the forward and backward directions along the horizontal axis, and VMamba designed SS2D module which introduces selective scanning across both the horizontal and vertical directions, allowing every part of the image to globally attend in both forward and backward directions along a specific direction, as illustrated in Fig. 3.

However, Vim and VMamba are not suitable for VHR remote sensing images. These images contain large-scale spatial features in multiple directions, such as the edges of objects and their arrangements. Modeling images globally only in the horizontal and vertical directions makes it challenging for the model to extract spatial features in other directions. Therefore, we propose the OSSM, which selectively scans images in horizontal, vertical, diagonal, and anti-diagonal directions. This approach enables global modeling of VHR remote sensing images in multiple directions, effectively extracting large-scale spatial features from various orientations, as illustrated in Fig. 3.

Specifically, the structure of the OSSM is illustrated in Fig. 4. OSSM begins with the input image patches undergoing omnidirectional scanning in the horizontal, vertical, diagonal, anti-diagonal directions, and their respective reverse directions, resulting in eight sequences of image patches. These sequences are then stacked along a new dimension and fed into the S6 block. The S6 block's selective scanning mechanism independently processes each image patch sequence, performing global modeling in specific directions [12]. Finally, all the image patch sequences are added after unstacking, merging global modeling information from multiple directions. This method enables the extraction of large spatial features from various orientations within VHR remote sensing images.

IV. Experimental Settings and Results

To validate the efficiency and superiority of the RSM in VHR remote sensing tasks, we conducted experiments across two distinct tasks: SS and CD. For the SS task, we evaluated the effectiveness of the RSM-SS model on the WHU [37] dataset and the Massachusetts Road [38] dataset. For the CD task, we evaluated the effectiveness of the RSM-CD model on the WHU-CD [37] dataset and the LEVIR-CD [39] dataset.

TABLE I
Brief Introduction of the Experimental Datasets

Name	Task	Resolution (m)	Images	Image size
WHU [37]	Seg	0.3	8189	512×512
M-Road [38]	Seg	1	1171	1500 × 1500
WHU-CD [37]	CD	0.075	1	32207×15354
LEVIR-CD [39]	CD	0.5	637	1024×1024

A. Datasets

We offer a brief description of the experimental SS and CD datasets in Table I.

1) SS Datasets: The WHU [37] building dataset is composed of two distinct subsets: one featuring satellite images and the other showcasing aerial photographs. Our investigation employs the aerial images subset, which includes a total of 8189 images. The images are divided into 4736 images designated for training, 1036 for validation, and 2416 for testing purposes, each with a spatial resolution of 0.3 m. This subset collectively captures approximately 22 000 buildings across an expanse of more than 450 km^2. We conducted our experiments using the data partitioning mode and image sizes (512 × 512) from the original WHU dataset.

The Massachusetts [38] Road dataset incorporates 1171 aerial photographs from Massachusetts, with each image measuring 1500 × 1500 pixels and encompassing 2.25 km^2. This dataset is organized into 1108 training images, 14 validation images, and 49 test images. It encompasses a diverse array of environments, including urban, suburban, and rural areas, spanning more than 2600 km^2, with the test segment covering more than 110 km^2. We follow the data partitioning mode of origin Massachusetts Road dataset. For analytical purposes, we segment the images into 1024 × 1024-pixel patches with a 548-pixel overlap on both the horizontal and vertical axes.

2) CD Datasets: The WHU-CD [37] dataset includes bitemporal VHR aerial images from 2012 and 2016, revealing significant alterations in building structures. We segment the dataset into 1024 × 1024 pixel patches that do not overlap and distribute these patches into training, validation, and test sets at a 7:1:2 ratio.

The LEVIR-CD [39] dataset is an extensive CD dataset comprising VHR (0.5 m/pixel) Google Earth images that document a range of building transformations over a period of 5–14 years. This dataset is particularly focused on changes related to buildings, such as construction and demolition. The bitemporal images are meticulously labeled with binary masks by specialists, indicating changes (1) and no changes (0), featuring a total of 31 333 instances of building changes. We segment the dataset into nonoverlapping 256 × 256 pixel patches. We follow the data partitioning mode of origin LEVIR-CD dataset.

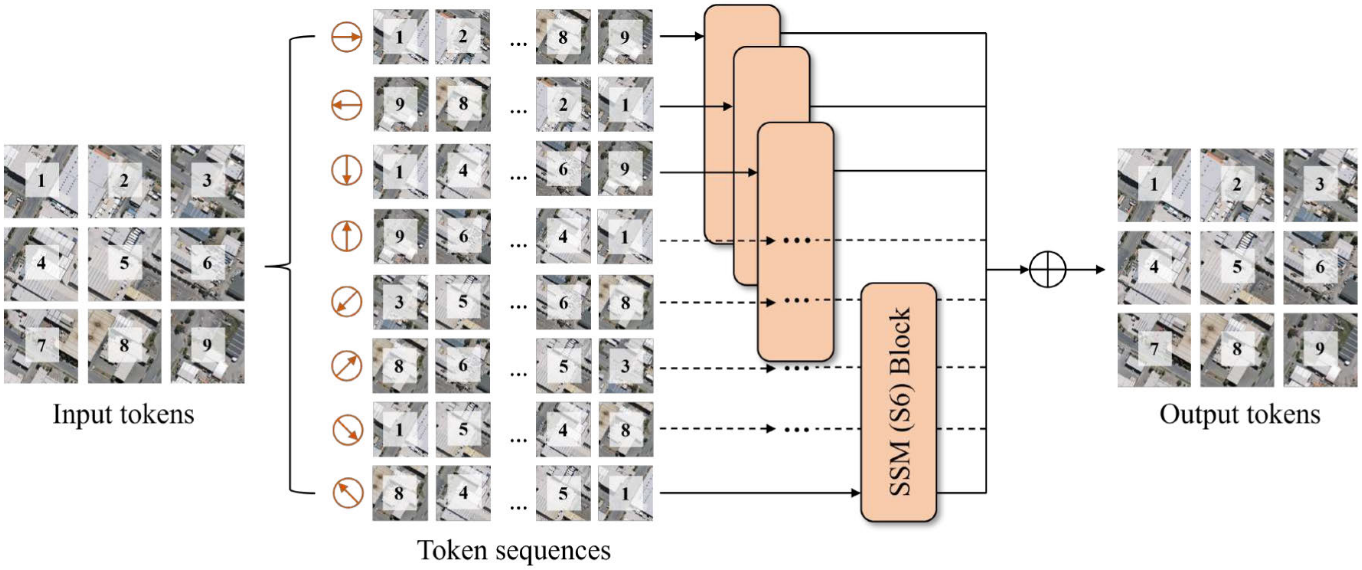


Fig. 4. Illustration of the structure of the OSSM.

B. Benchmark Methods

To evaluate the effectiveness of the proposed RSM, we conducted comparative experiments with various benchmark methods on the SS and CD tasks. The benchmark methods tested on the same dataset are based on the same splitting of the dataset and use the same data.

On the SS task, the compared CNN-based models include FCN [40], SegNet [41], U-Net [19], PSPNet [42], HRNet [43], MA-FCN [44], Deeplabv3+ [45], ResUNet [46], MAP-Net [47], D-LinkNet [48], and SIINet [49], the compared transformer-based models include Segformer [50] and Road-Former [10], and the CNN-transformer hybrid models include BDTNet [51], TransUNet [52], and CMTFNet [53].

On the CD task, the compared CNN-based models include FC-EF [36], FC-Siam-Diff [36], FC-Siam-Conc [36], STANet [39], DTCDSCN [54], SNUNet [55], CDNet [56], DDCNN [57], DASNet [58], DSIFN [59], and HANet [60], the compared transformer-based models include Change-Former [9], and the CNN-transformer hybrid models include BIT [8], MTCNet [61], MSCANet [62], and AMTNet-50 [63].

C. Implementation Details

1) Data Augmentation: To demonstrate the effectiveness of the proposed methods, we only employed straightforward data augmentation techniques, avoiding the use of any elaborate tricks. For the SS task, the data augmentation methods used for the RSM-SS model included flipping ($p = 0.5$) and transposing ($p = 0.5$). For the CD task, the data augmentation methods used for the RSM-CD model included flipping ($p = 0.5$), transposing ($p = 0.5$), and swapping of bitemporal images ($p = 0.5$).

2) Training and Inference: We utilized PyTorch [64] to construct and deploy RSM-SS and RSM-CD on a single RTX A100 GPU (80G). Given the variable image sizes across datasets, we adjusted the batch sizes accordingly: 16 for the WHU dataset, 4 for both the Massachusetts Road and WHU-CD datasets, and 64 for the LEVIR-CD dataset. Our loss function integrates binary cross-entropy loss with Dice coefficient loss to optimize performance. We employed the AdamW [65] optimizer with an initial learning rate of 0.001 and a weight decay of 0.001. The learning rate adjustment strategy is to reduce the learning rate by a factor of 0.1 if there is no improvement in the $F1$-score on the validation set over a span of 10 epochs. The models were trained over 150 epochs to ensure ample training and convergence. Checkpoints capturing the highest $F1$-scores on the validation sets were preserved for subsequent testing. To maintain consistency with other CD methodologies, we initialized our models using the default settings provided by PyTorch for all datasets.

3) Evaluation Metrics: To evaluate the performance of the proposed models, we employ four key evaluation metrics: precision (P), recall (R), $F1$-score, and intersection over union (IoU). Precision quantifies the rate of false positives within the results, whereas recall measures the rate of false negatives. Achieving high scores in both precision and recall simultaneously poses a significant challenge due to their inversely proportional relationship. The $F1$-score, which represents the harmonic mean of precision and recall, serves as a balance between the two by simultaneously considering both metrics. Additionally, the IoU metric measures the proportion of overlap between the predicted and actual changed pixels relative to the total area of union, providing a spatial accuracy assessment of the model's predictions.

D. Ablation Study

To verify the effectiveness of OSSM, comparative experiments were conducted on the Massachusetts Road dataset for the SS task and the WHU-CD dataset for the CD task. We compared the performance of three variations: SS1D [14], which employs selective scanning in the horizontal direction and its reverse direction; SS2D [15], which includes selective scanning in both the horizontal and vertical directions and their

TABLE II

Ablation Study of OSSM on the Massachusetts Road Dataset and WHU-CD Dataset. The Best Values Are Highlighted in Bold in Each Dataset

Dataset	Task	Module	P (%)	R (%)	F1 (%)	IoU (%)
M-Road	Seg	SS1D	85.17	74.37	79.40	65.84
M-Road	Seg	SS2D	85.57	74.78	79.81	66.41
M-Road	Seg	OSSM	**86.52**	**75.24**	**80.49**	**67.35**
WHU-CD	CD	SS1D	91.86	89.33	90.58	82.78
WHU-CD	CD	SS2D	92.25	89.66	90.94	83.38
WHU-CD	CD	OSSM	**93.37**	**90.42**	**91.87**	**84.96**

reverses; and OSSM, which extends selective scanning to eight directions—horizontal, vertical, diagonal, and anti-diagonal, along with their reverse directions. This comparison aimed to demonstrate the superiority of employing eight-directional selective scanning in VHR remote sensing images.

Table II presents the comparative results of SS1D, SS2D, and OSSM, indicating that OSSM outperforms both SS1D and SS2D in SS and CD tasks. Specifically, for the SS task on the Massachusetts Road dataset, the presence of roads extending in multiple directions with significant spatial scales necessitates selective scanning across multiple directions, thereby extracting large road features oriented in various directions. Similarly, in the CD task on the WHU-CD dataset, the spatial features of buildings, such as edge characteristics and arrangement directions, require selective scanning in multiple directions to capture large architectural features from various orientations. Compared to SS1D and SS2D, the omnidirectional selective scan of OSSM enables the extraction of large object features from multiple directions, making it more suitable for VHR remote sensing images.

E. Overall Comparison

1) SS Task: This section presents the results of comparing RSM-SS with other models on the SS task on two datasets (the Massachusetts Road and WHU datasets).

The accuracy comparison results on the Massachusetts Road dataset are presented in Table III. The RSM-SS surpasses all the compared models and achieves the highest IoU (0.6735) and $F1$-score (0.8049) on the Massachusetts Road dataset. The Massachusetts Road dataset is characterized by roads with extensive spatial scales, where contextual information plays a critical role in road SS. Due to its linear complexity, RSM-SS is adept at processing VHR remote sensing images with rich contextual information. This capability allows it to accurately segment roads by effectively leveraging the vast amount of contextual information in remote sensing images.

The accuracy comparison results on the WHU dataset are summarized in Table IV. The accuracy comparison results show that the RSM-SS achieves the highest IoU (0.9081) and $F1$-score (0.9518) on this dataset. In the WHU dataset, buildings are arranged in a variety of orientations and cover large spatial scales. Extracting large spatial features of buildings in multiple directions is crucial for accurate building detection. RSM-SS performs selective scanning across multiple directions on VHR remote sensing images, which enables the extraction of substantial spatial features of buildings from various angles, thus achieving precise segmentation of buildings in VHR remote sensing images.

TABLE III

Accuracy Comparison on the Massachusetts Road Dataset. The Best Values Are Highlighted in Bold

Methods	P (%)	R (%)	F1 (%)	IoU (%)
SegNet [41]	76.09	78.23	77.15	62.79
U-Net [19]	77.53	77.82	77.67	63.50
ResUNet [46]	78.77	77.45	78.10	64.07
D-LinkNet [48]	78.34	77.91	78.12	64.10
HRNetv2 [43]	79.01	78.20	78.60	64.75
Deeplabv3+ [45]	75.14	72.56	73.83	58.51
SIINet [49]	85.36	74.13	79.35	65.77
RoadFormer [10]	80.54	**78.90**	79.71	66.27
BDTNet [51]	82.99	76.37	79.54	66.03
RSM-SS	**86.52**	75.24	**80.49**	**67.35**

TABLE IV

Accuracy Comparison on the WHU Dataset. The Best Values Are Highlighted in Bold

Methods	P (%)	R (%)	F1 (%)	IoU (%)
FCN [40]	92.29	92.84	92.56	86.16
SegNet [41]	93.42	91.71	92.56	86.15
U-Net [19]	94.50	90.88	92.65	86.31
PSPNet [42]	93.19	94.21	93.70	88.14
HRNet [43]	91.69	92.85	92.27	85.64
MA-FCN [44]	94.75	94.92	94.83	90.18
Deeplabv3+ [45]	94.31	94.53	94.42	89.43
ResUNet [46]	94.49	94.71	94.60	89.75
MAP-Net [47]	93.99	94.82	94.40	89.40
Segformer [50]	94.72	94.42	94.57	89.70
TransUNet [52]	94.05	93.07	93.56	87.89
CMTFNet [53]	90.12	**95.21**	92.59	86.21
RSM-SS	**95.25**	95.12	**95.18**	**90.81**

We show some inference results of the test set of the Massachusetts Road and WHU datasets in Fig. 5. It shows that the RSM-SS can accurately segment all the roads in the Massachusetts Road dataset and buildings in the WHU dataset. On the Massachusetts Road dataset, despite roads extending in various directions and covering extensive spatial scales, RSM-SS manages to accurately segment roads by leveraging the rich contextual information available in large VHR remote sensing images. Similarly, on the WHU dataset, despite the dense arrangement of buildings and the presence of spatial features in multiple directions, the RSM-SS is capable of extracting building features across various directions.

2) CD Task: This section presents the results of comparing RSM-CD with other CD models on the CD task with two datasets (WHU-CD and LEVIR-CD datasets).

The accuracy comparison results on the WHU-CD dataset are shown in Table V. RSM-CD achieves the highest IoU (0.8496) and $F1$-score (0.9187) on this dataset, outperforming all other CD models. Given the very high spatial resolution of the WHU-CD dataset remote sensing images (0.075 m/pixel), a remote sensing image with a normal size (256 × 256 pixels) may only contain a few buildings or parts of buildings, thereby losing a significant amount of contextual information. Due to its linear complexity, RSM-CD is capable of processing

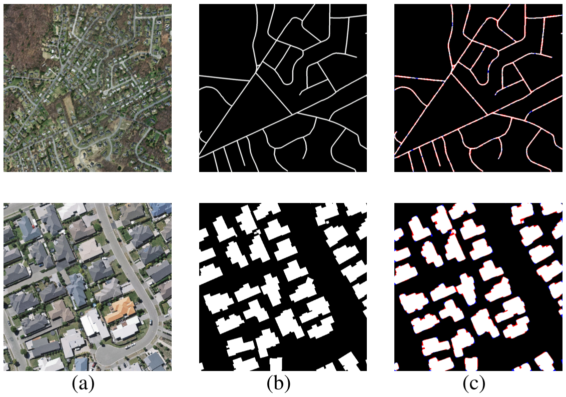


Fig. 5. Sample inference results of RSM-SS on the SS task. The results on the Massachusetts Road and WHU datasets are shown in the first and second rows, respectively. Red areas denote false positives and blue areas denote false negatives. (a) Input image. (b) Ground truth image. (c) RSM-SS result.

TABLE V

ACCURACY COMPARISON ON THE WHU-CD DATASET. THE BEST VALUES ARE HIGHLIGHTED IN BOLD

Methods	P (%)	R (%)	F1 (%)	IoU (%)
FC-EF [36]	78.86	78.64	78.75	64.95
FC-Siam-Diff [36]	84.73	87.31	86.00	75.44
FC-Siam-Conc [36]	78.86	78.64	78.75	64.95
DTCDSCN [54]	63.92	82.30	71.95	56.19
DSIFN [59]	91.44	89.75	90.59	82.79
STANet [39]	79.37	85.50	82.32	69.95
SNUNet [55]	85.60	81.49	83.49	71.67
DASNet [58]	88.23	84.62	86.39	76.04
HANet [60]	88.30	88.01	88.16	78.82
CDNet [56]	91.75	86.89	89.25	80.59
DDCNN [57]	**93.71**	89.12	91.36	84.09
BIT [8]	86.64	81.48	83.98	72.39
MTCNet [61]	75.10	**91.90**	82.65	70.43
MSCANet [62]	91.10	89.86	90.47	82.60
RSM-CD	93.37	90.42	**91.87**	**84.96**

TABLE VI

ACCURACY COMPARISON ON THE LEVIR-CD DATASET. THE BEST VALUES ARE HIGHLIGHTED IN BOLD

Methods	P (%)	R (%)	F1 (%)	IoU (%)
FC-EF [36]	86.91	80.17	83.40	71.53
FC-Siam-Diff [36]	89.53	83.31	86.31	75.91
FC-Siam-Conc [36]	91.99	76.77	83.69	71.96
DTCDSCN [54]	88.53	86.83	87.67	78.05
DSIFN [59]	**94.02**	82.93	88.13	78.77
STANet [39]	83.81	**91.00**	87.26	77.39
SNUNet [55]	89.18	87.17	88.16	78.83
HANet [60]	91.21	89.36	90.28	82.27
CDNet [56]	91.60	86.50	88.98	80.14
DDCNN [57]	91.85	88.69	90.24	82.22
BIT [8]	89.24	89.37	89.30	80.68
ChangeFormer [9]	92.05	88.80	90.40	82.47
MTCNet [61]	90.87	89.62	90.24	82.22
MSCANet [62]	91.30	88.56	89.91	81.66
AMTNet-50 [63]	91.82	89.71	90.76	83.08
RSM-CD	92.52	89.73	**91.10**	**83.66**

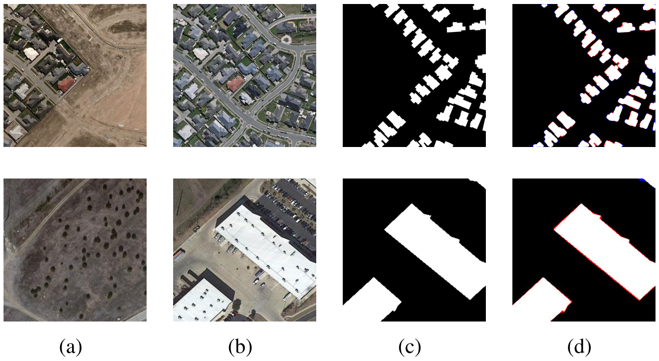


Fig. 6. Sample inference results of RSM-CD on the CD task. The results on the WHU-CD and LEVIR-CD datasets are shown in the first and second rows, respectively. Red areas denote false positives and blue areas denote false negatives. (a) T1 image. (b) T2 image. (c) Ground-truth image. (d) RSM-CD result.

large VHR remote sensing images. This allows RSM-CD to utilize the rich contextual information present in large images to accurately identify changed buildings.

The accuracy comparison results on the LEVIR-CD dataset are summarized in Table VI. RSM-CD outperforms all the other models, achieving the highest IoU (0.8366) and $F1$-score (0.9110) on this dataset. In the LEVIR-CD dataset, the presence of buildings with multiple orientations and arrangements in the bitemporal remote sensing images underscores the importance of extracting large features in multiple directions. The OSSM of RSM-CD can extract large spatial features of buildings from various directions, thereby accurately identifying changed buildings.

We show some inference results for the test sets of the WHU-CD and LEVIR-CD datasets in Fig. 6. It shows that RSM-CD can accurately detect all the changed buildings in the WHU-CD and LEVIR-CD datasets. On the WHU-CD dataset, the high spatial resolution of remote sensing images necessitates the use of large images to preserve ample contextual information. The linear complexity of RSM-CD enables it to process large VHR remote sensing images. By leveraging the rich contextual information available in the bitemporal images, RSM-CD can accurately identify changed buildings. On the LEVIR-CD dataset, where buildings exhibit edge features in multiple directions, the ability to perform selective scanning in multiple directions allows RSM-CD to extract building features from various orientations, accurately identifying changed buildings.

F. Impact of Image Size and Spatial Resolution

The extensive contextual information and high-resolution spatial features of VHR remote sensing images are crucial for dense prediction tasks. To investigate the impact of contextual information and spatial features on dense prediction tasks in VHR remote sensing, we experimented with SS and CD tasks using images of varying sizes and downsampling factors, as cropping images into small patches would lose contextual information and downsampling images would lose spatial features.

In our SS experiments on the Massachusetts Road dataset, where roads are the objects of interest, we first downsampled the remote sensing images by factors of 1 (no downsampling), 2, and 4. Then, we cropped the images to sizes of 32, 64,

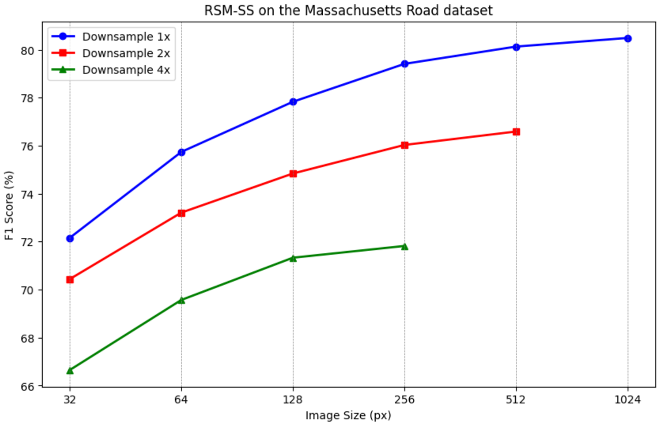


Fig. 7. Performance of RSM-SS on the Massachusetts Road dataset with different image sizes and downsampling ratios.

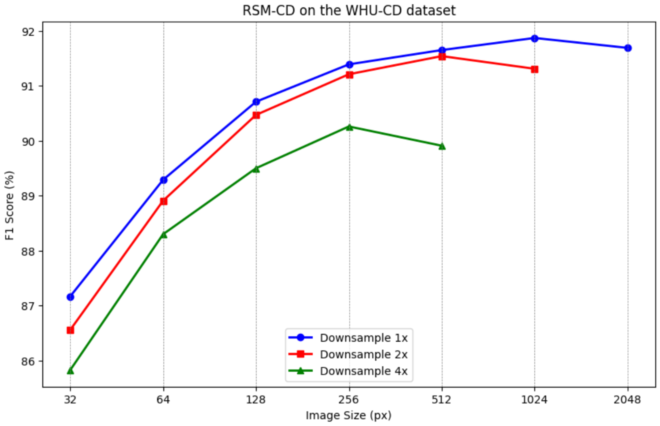


Fig. 8. Performance of RSM-CD on the WHU-CD dataset with different image sizes and downsampling ratios.

128, 256, 512, and 1024 pixels. It is important to note that images downsampled by a factor of 2 have a maximum size of 512 pixels, and those downsampled by a factor of 4 have a maximum size of 256 pixels. For two images of the same size but different downsampling ratios (ratio 1, ratio 2), the latter has a ratio 2/ratio 1 times the spatial range of the former, which means that an image with a size of 512 and a downsampling ratio of 1 has the same spatial range as an image with a size of 256 and a downsampling ratio of 2. We used the $F1$-score as a metric to evaluate the model's performance across different image sizes and downsampling ratios, and the results are illustrated in Fig. 7.

The results indicate that the model's performance improves with increasing image size, regardless of the downsampling ratio. For images of the same size, those with a higher downsampling ratio perform worse, despite having more contextual information. This could be attributed to the elongated nature of roads, which extend in various directions across the image. Downsampling the images results in a significant loss of road spatial features, making it difficult to segment roads. Cropping the images into smaller patches leads to a substantial loss of contextual information, which hampers the ability to determine the roads' extension directions and segment them effectively. Thus, both large amounts of contextual information and high-resolution spatial features are important for segmenting roads.

In the CD task, we conducted experiments on the WHU-CD dataset, where the changed objects are buildings. The strategies for image downsampling and cropping were the same as those applied to the Massachusetts Road dataset, with the $F1$-score serving as the evaluation metric. The performance of the model across different image sizes and downsampling ratios is illustrated in Fig. 8. The results show that the model's performance initially increases with the size of the image, reaching a peak before starting to decline. The image size peaks for downsampling ratios of 1, 2, and 4 correspond to 256, 512, and 1024, respectively, each offering the same level of spatial range. Furthermore, at any given size, the model performs worse on images with a higher downsampling ratio. This may be due to the presence of spatial features within individual buildings and between multiple buildings, necessitating a certain level of contextual information and high-resolution spatial features to accurately identify changed buildings. Downsampling results in a significant loss of spatial features, substantially reducing the model's performance. Due to the geospatial correlation in remote sensing images [66], buildings that are closer to each other exhibit stronger correlations, whereas buildings that are farther apart exhibit weaker correlations. Cropping images to a certain size makes larger patches containing more contextual information, which benefits the model in detecting changed buildings. However, when there is too much contextual information, including a lot of irrelevant details for identifying a specific changed building, the model's performance decreases.

The experiments on the Massachusetts Road and WHU-CD datasets highlight the importance of contextual information and high-resolution spatial features for dense prediction tasks in VHR remote sensing, which aligns with the experimental results from FCCDN [16] and SwinB-CNN [17]. The model's performance varies in response to the loss of contextual information and spatial features, with a more significant decrease in performance when downsampling road images than when downsampling building images. A common finding is that the model performs better on higher spatial resolution images. Moreover, the higher the spatial resolution of the images is, the larger the image size required to contain the same level of contextual information, thus larger images are needed for the model to perform optimally. Therefore, VHR remote sensing images and models capable of processing large images are crucial for dense prediction tasks in remote sensing. The linear complexity of the RSM enables it to handle large VHR remote sensing images, thereby achieving excellent results in dense prediction tasks for VHR remote sensing.

G. Handling Large Remote Sensing Images

Large remote sensing images contain rich contextual information, which is crucial for dense prediction tasks. To demonstrate the superiority of the RSM over CNN-based and transformer-based models in processing large remote sensing images, we conducted comparative experiments on the WHU-CD dataset for CD tasks. Given that the original

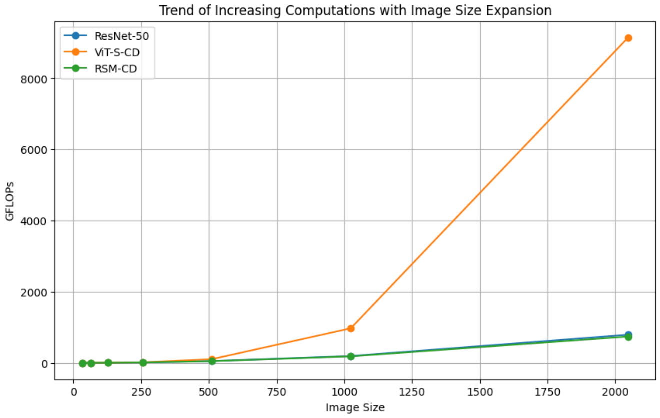


Fig. 9. Trend of increasing computational load with image size expansion in different models.

WHU-CD data are completely large remote sensing images, we can crop them into image patches of various sizes for our experiments. We compared RSM-CD with ResNet-50-CD and ViT-S-CD, which have a similar number of parameters. All these models have the same framework, with the only difference lying in the encoder. ResNet-50-CD replaces the encoder of RSM-CD with ResNet-50, and ViT-S-CD replaces it with ViT-S.

The comparison results are shown in Table VII and Fig. 9, where the batch size is set to 1 when calculating the FLOPs of the models. OOM stands for out of memory, indicating that the model training could not proceed because the required memory exceeded the machine's available memory. As the image size increased, the computational costs of ResNet-50-CD and RSM-CD exhibited a linear growth trend because they have linear complexity. In contrast, as ViT-S-CD has quadratic complexity, it exhibited a quadratic growth trend and significantly exceeded the computational costs of ResNet-50-CD and RSM-CD on large remote sensing images. This complexity led to the inability of the model to train on images of sizes 1024 and 2048 due to excessive memory requirements.

In terms of model performance, RSM-CD achieved the best results on image size 1024 with a similar parameter count and lower computational cost, outperforming ResNet-50-CD and ViT-S-CD of all valid image sizes. Although ResNet-50-CD has linear complexity, it cannot globally model the context of remote sensing images and only has a local effective receptive field, which fails to utilize the rich contextual information of large remote sensing images effectively. While ViT-S-CD can globally model the context of remote sensing images, its quadratic complexity results in a rapid increase in computational cost and memory usage, restricting it to processing only small image patches. These small patches contain very limited contextual information, disadvantaging ViT-S-CD in CD tasks. RSM-CD has linear complexity and can model the context of remote sensing images globally across multiple directions, extracting large spatial features from various directions, thus achieving superior results in CD tasks.

TABLE VII
COMPARATIVE EXPERIMENTS ON THE COMPUTATIONAL COST AND MODEL PERFORMANCE, WHICH WERE CONDUCTED ON THE WHU-CD DATASET

Model	Param (M)	Image size	GFLOPs	F1 (%)
ResNet-50-CD	30.4	32	1.7	86.42
ViT-S-CD	26.9	32	1.3	86.74
RSM-CD	27.9	32	2.9	87.17
ResNet-50-CD	30.4	64	3.9	87.33
ViT-S-CD	26.9	64	3.6	88.53
RSM-CD	27.9	64	4.1	89.17
ResNet-50-CD	30.4	128	6.8	88.07
ViT-S-CD	26.9	128	7.6	89.74
RSM-CD	27.9	128	7.2	90.71
ResNet-50-CD	30.4	256	13.5	88.41
ViT-S-CD	26.9	256	16.8	90.33
RSM-CD	27.9	256	15.7	91.39
ResNet-50-CD	30.4	512	53.7	88.53
ViT-S-CD	26.9	512	107.9	90.52
RSM-CD	27.9	512	50.2	91.65
ResNet-50-CD	30.4	1024	194.7	88.37
ViT-S-CD	26.9	1024	973.4	OOM
RSM-CD	27.9	1024	185.2	91.87
ResNet-50-CD	30.4	2048	793.9	88.45
ViT-S-CD	26.9	2048	9148.7	OOM
RSM-CD	27.9	2048	742.8	91.49

V. DISCUSSION

In the realm of VHR remote sensing, contextual information within images is crucial for dense prediction tasks. However, current models based on CNNs and transformers struggle to process large VHR remote sensing images effectively. CNN-based models, limited by their local convolution operations, fail to model the global contextual information of VHR remote sensing images. Transformer-based models, due to their quadratic complexity, are incapable of handling large VHR images. These models typically resort to processing smaller image patches, which contain limited contextual information, thus hindering their performance on dense prediction tasks.

To address these issues, we propose the RSM for dense prediction tasks in VHR remote sensing. The RSM, characterized by its linear complexity and global modeling capabilities, can process large VHR remote sensing images and model their global contextual information. It is capable of extracting large spatial features in multiple directions, thus effectively facilitating dense prediction tasks. The experimental results of the SS and CD tasks demonstrate that, despite RSM-SS and RSM-CD employing simple model architectures without any sophisticated modules or training techniques, they achieve state-of-the-art performance in their respective tasks. This validates the potential of the RSM in dense prediction tasks for VHR remote sensing. We hope that the RSM can serve as a baseline in the field, promoting the development of SSM-based methods within VHR remote sensing.

Despite the notable performance of RSM in dense prediction tasks for ultrahigh-resolution remote sensing, it also exhibits some limitations. On one hand, the models for SS and CD tasks are overly simplistic, not fully leveraging the potential of SSMs. The experimental results of RSM-CD on the WHU-CD dataset indicate optimal performance when the image size is 1024. However, a reduction in model performance is observed as the image size increases to 2048. This decline could be attributed to the inherent simplicity of the RSM-CD, which may limit its capability to effectively process the global contextual information contained within large remote sensing images. On the other hand, dense prediction tasks in ultrahigh-resolution remote sensing require extensive training data, thus limiting RSM's applicability in tasks lacking labeled data.

In the future, we will develop more complex and effective SSM-based models to enhance their capacity for processing global contextual information in large remote sensing images, further exploring the potential of SSM-based models in remote sensing dense prediction tasks. At the same time, we will investigate the potential of SSM-based models in remote sensing self-supervised learning, enabling the use of vast quantities of unlabeled remote sensing images for self-supervised training of SSM-based models, and making these models applicable to tasks and scenarios where labeled data are scarce or unavailable.

VI. Conclusion

We have proposed the RSM for dense prediction tasks in VHR remote sensing imagery, addressing the limitations of CNN-based models in global context information modeling and the challenges of transformer-based models handling large remote sensing images. Our model can process large VHR remote sensing images with rich contextual information with linear complexity. Through selective scanning in multiple directions, the RSM models global context information and extracts large spatial features across various directions, thereby efficiently accomplishing dense prediction tasks.

Experiments on SS and CD tasks demonstrate the superior performance of the RSM across different objects. Leveraging the SSM for processing large images and modeling global context information, the RSM operates on VHR remote sensing images without the need to segment these images into smaller patches, which is achieved through its linear complexity. By modeling globally in various directions, the RSM captures large spatial features from multiple perspectives, leading to outstanding performance in dense prediction tasks. We envision RSM to serve as a baseline for dense prediction tasks in VHR remote sensing, promoting further development of SSM-based approaches in this field.

Acknowledgment

This work was done during the internship of Sijie Zhao at Shanghai Artificial Intelligence Laboratory, Shanghai, China. The authors would like to thank the editor and the anonymous reviewers for their constructive comments.

References

[1] T. Wellmann et al., "Remote sensing in urban planning: Contributions towards ecologically sound policies?" *Landscape Urban Planning*, vol. 204, Dec. 2020, Art. no. 103921.

[2] M. Weiss, F. Jacob, and G. Duveiller, "Remote sensing for agricultural applications: A meta-review," *Remote Sens. Environ.*, vol. 236, Jan. 2020, Art. no. 111402.

[3] S. Asadzadeh, W. J. D. Oliveira, and C. R. D. Souza Filho, "UAV-based remote sensing for the petroleum industry and environmental monitoring: State-of-the-art and perspectives," *J. Petroleum Sci. Eng.*, vol. 208, Jan. 2022, Art. no. 109633.

[4] S. Lei, Z. Shi, and W. Mo, "Transformer-based multistage enhancement for remote sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, 2021.

[5] S. Lei, Z. Shi, and Z. Zou, "Coupled adversarial training for remote sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3633–3643, May 2020.

[6] 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.

[7] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–7.

[8] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2021.

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

[10] X. Jiang et al., "RoadFormer: Pyramidal deformable vision transformers for road network extraction with remote sensing images," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 113, Sep. 2022, Art. no. 102987.

[11] 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.

[12] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," 2023, *arXiv:2312.00752*.

[13] R. E. Kalman, "A new approach to linear filtering and prediction problems," *J. Basic Eng.*, vol. 82, no. 1, pp. 35–45, 1960.

[14] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision mamba: Efficient visual representation learning with bidirectional state space model," 2024, *arXiv:2401.09417*.

[15] Y. Liu et al., "VMamba: Visual state space model," 2024, *arXiv:2401.10166*.

[16] P. Chen, B. Zhang, D. Hong, Z. Chen, X. Yang, and B. Li, "FCCDN: Feature constraint network for VHR image change detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 187, pp. 101–119, May 2022.

[17] C. Zhang, W. Jiang, Y. Zhang, W. Wang, Q. Zhao, and C. Wang, "Transformer and CNN hybrid deep neural network for semantic segmentation of very-high-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4408820.

[18] M. Papadomanolaki, S. Verma, M. Vakalopoulou, S. Gupta, and K. Karantzalos, "Detecting urban changes with recurrent neural networks from multitemporal Sentinel-2 data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2019, pp. 214–217.

[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. Intervent.*, vol. 9351. Cham, Switzerland: Springer, 2015, pp. 234–241.

[20] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–11.

[21] K. Yue, L. Yang, R. Li, W. Hu, F. Zhang, and W. Li, "TreeUNet: Adaptive tree convolutional neural networks for subdecimeter aerial image segmentation," *ISPRS J. Photogramm. Remote Sens.*, vol. 156, pp. 1–13, Oct. 2019.

[22] S. Zhao, X. Zhang, P. Xiao, and G. He, "Exchanging dual-encoder–decoder: A new strategy for change detection with semantic guidance and spatial localization," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4508016.

[23] X. Yang et al., "An attention-fused network for semantic segmentation of very-high-resolution remote sensing imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 177, pp. 238–262, Jul. 2021.

[24] C. Han, C. Wu, H. Guo, M. Hu, J. Li, and H. Chen, "Change guiding network: Incorporating change prior to guide change detection in remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 8395–8407, 2023.

[25] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[26] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. CVPR*, vol. 16, Sep. 2016, pp. 770–778.

[27] Q. Li, R. Zhong, X. Du, and Y. Du, “TransUNetCD: A hybrid transformer network for change detection in optical remote-sensing images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5622519.

[28] A. Gu, K. Goel, and C. Ré, “Efficiently modeling long sequences with structured state spaces,” 2021, *arXiv:2111.00396*.

[29] A. Gu et al., “Combining recurrent, convolutional, and continuous-time models with linear state space layers,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 572–585.

[30] J. T. H. Smith, A. Warrington, and S. W. Linderman, “Simplified state space layers for sequence modeling,” 2022, *arXiv:2208.04933*.

[31] A. Gu, T. Dao, S. Ermon, A. Rudra, and C. Ré, “Hippo: Recurrent memory with optimal polynomial projections,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 1474–1487.

[32] A. Gu, K. Goel, A. Gupta, and C. Ré, “On the parameterization and initialization of diagonal state space models,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 35971–35983.

[33] A. Gupta, A. Gu, and J. Berant, “Diagonal state spaces are as effective as structured state spaces,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 22982–22994.

[34] R. Hasani, M. Lechner, T.-H. Wang, M. Chahine, A. Amini, and D. Rus, “Liquid structural state-space models,” 2022, *arXiv:2209.12951*.

[35] H. Chen, J. Song, C. Han, J. Xia, and N. Yokoya, “ChangeMamba: Remote sensing change detection with spatio-temporal state space model,” 2024, *arXiv:2404.03425*.

[36] 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.

[37] 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.

[38] V. Mnih, *Machine Learning for Aerial Image Labeling*. Toronto, ON, Canada: Univ. Toronto (Canada), 2013.

[39] H. Chen and Z. Shi, “A spatial–temporal attention-based method and a new dataset for remote sensing image change detection,” *Remote Sens.*, vol. 12, no. 10, p. 1662, 2020.

[40] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[41] V. Badrinarayanan, A. Kendall, and R. Cipolla, “SegNet: A deep convolutional encoder–decoder architecture for image segmentation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[42] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.

[43] K. Sun, B. Xiao, D. Liu, and J. Wang, “Deep high-resolution representation learning for human pose estimation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Sep. 2019, pp. 5693–5703.

[44] S. Wei, S. Ji, and M. Lu, “Toward automatic building footprint delineation from aerial images using CNN and regularization,” *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 2178–2189, Mar. 2019.

[45] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder–decoder with atrous separable convolution for semantic image segmentation,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.

[46] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, “ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data,” *ISPRS J. Photogramm. Remote Sens.*, vol. 162, pp. 94–114, Apr. 2020.

[47] Q. Zhu, C. Liao, H. Hu, X. Mei, and H. Li, “MAP-Net: Multiple attending path neural network for building footprint extraction from remote sensed imagery,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 6169–6181, Jul. 2020.

[48] L. Zhou, C. Zhang, and M. Wu, “D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 182–186.

[49] C. Tao, J. Qi, Y. Li, H. Wang, and H. Li, “Spatial information inference net: Road extraction using road-specific contextual information,” *ISPRS J. Photogramm. Remote Sens.*, vol. 158, pp. 155–166, Dec. 2019.

[50] E. Xie et al., “SegFormer: Simple and efficient design for semantic segmentation with transformers,” in *Proc. Adv. Neural Inf. Process. Sys. (NIPS)*, vol. 34, Dec. 2021, pp. 12077–12090.

[51] L. Luo, J.-X. Wang, S.-B. Chen, J. Tang, and B. Luo, “BDTNet: Road extraction by bi-direction transformer from remote sensing images,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[52] J. Chen et al., “TransUNet: Transformers make strong encoders for medical image segmentation,” 2021, *arXiv:2102.04306*.

[53] H. Wu, P. Huang, M. Zhang, W. Tang, and X. Yu, “CMTFNet: CNN and multiscale transformer fusion network for remote-sensing image semantic segmentation,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 2004612.

[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] S. Fang, K. Li, J. Shao, and Z. Li, “SNUNet-CD: A densely connected Siamese network for change detection of VHR images,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2021.

[56] H. Chen, W. Li, and Z. Shi, “Adversarial instance augmentation for building change detection in remote sensing images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5603216.

[57] X. Peng, R. Zhong, Z. Li, and Q. Li, “Optical remote sensing image change detection based on attention mechanism and image difference,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7296–7307, Sep. 2020.

[58] J. Chen et al., “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, 2020.

[59] C. Zhang et al., “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.

[60] C. Han, C. Wu, H. Guo, M. Hu, and H. Chen, “HANet: A hierarchical attention network for change detection with bitemporal very-high-resolution remote sensing images,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 3867–3878, 2023.

[61] W. Wang, X. Tan, P. Zhang, and X. Wang, “A CBAM based multiscale transformer fusion approach for remote sensing image change detection,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 6817–6825, 2022.

[62] M. Liu, Z. Chai, H. Deng, and R. Liu, “A CNN-transformer network with multiscale context aggregation for fine-grained cropland change detection,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 4297–4306, 2022.

[63] W. Liu, Y. Lin, W. Liu, Y. Yu, and J. Li, “An attention-based multiscale transformer network for remote sensing image change detection,” *ISPRS J. Photogramm. Remote Sens.*, vol. 202, pp. 599–609, Aug. 2023.

[64] A. Paszke et al., “PyTorch: An imperative style, high-performance deep learning library,” in *Proc. NIPS*, Dec. 2019, pp. 8024–8035.

[65] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, *arXiv:1412.6980*.

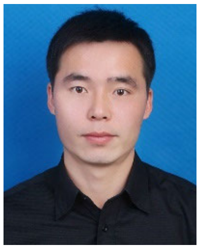
[66] T. Serrano, C. Chevrier, L. Multigner, S. Cordier, and B. Jegou, “International geographic correlation study of the prevalence of disorders of male reproductive health,” *Hum. Reproduction*, vol. 28, no. 7, pp. 1974–1986, Jul. 2013.

Sijie Zhao (Graduate Student Member, IEEE) received the B.S. degree in geographic information science from the School of Geography and Ocean Science, Nanjing University, Nanjing, China, in 2023, where he is currently pursuing the M.S. degree in cartography and geographic information system.

His research interests include change detection and application of generative models to Earth surface understanding and forecasting.

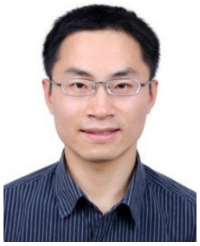
Hao Chen received the B.S. and Ph.D. degrees from the Image Processing Center, School of Astronautics, Beihang University, Beijing, China, in 2017 and 2023, respectively.

He is currently a Researcher at Shanghai Artificial Intelligence (AI) Laboratory, Shanghai, China. His research interests include geospatial machine learning, remote sensing, Earth monitoring, and prediction.



Xueliang Zhang (Senior Member, IEEE) received the B.S. degree in geographical information system and the Ph.D. degree in remote sensing of resources and environment from Nanjing University, Nanjing, China, in 2010 and 2015, respectively.

From 2014 to 2015, he was a Visiting Student with the Informatics Institute, University of Missouri, Columbia, MO, USA. From 2016 to 2018, he was an Associate Researcher with the Department of Geographic Information Science, Nanjing University, where he is currently an Associate Professor. His research interests include high-resolution remote sensing image analysis, semantic segmentation, and deep learning for remote sensing.



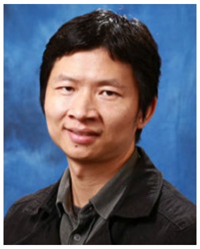
Pengfeng Xiao (Senior Member, IEEE) received the B.M. degree in land resource management from Hunan Normal University, Changsha, China, in 2002, and the Ph.D. degree in cartography and geographical information system from Nanjing University, Nanjing, China, in 2007.

From 2007 to 2009, he was a Lecturer at the School of Geography and Ocean Science, Nanjing University, where he was an Associate Professor, from 2010 to 2018, and he has been a Professor, since 2019. He was a Visiting Scholar with the Department of Geography, University of Giessen, Giessen, Hesse, Germany, from 2011 to 2012, and the Department of Environmental Science, Policy, and Management, University of California at Berkeley, Berkeley, CA, USA, from 2014 to 2015. He has authored four books and over 160 articles. His research interests include high- resolution remote sensing image analysis, remote sensing of snow cover, and land use and land cover change.

Lei Bai received the Ph.D. degree from the University of New South Wales, Sydney, NSW, Australia, in 2021.

He was a Post-Doctoral Researcher with the University of Sydney, Camperdown, NSW, Australia. He is currently a Research Scientist with Shanghai Artificial Intelligence (AI) Laboratory, Shanghai, China. He has authored or co-authored a set of peer-reviewed papers in top AI conferences and journals, such as Neural Information Processing Systems (NeurIPS), Conference on Computer Vision and Pattern Recognition (CVPR), International Joint Conference on Artificial Intelligence (IJCAI), Knowledge Discovery and Data Mining (KDD), International Conference on Computer Vision (ICCV), International Conference on Ubiquitous Computing (Ubicomp), IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (TPAMI), and IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS (TITS). His research interests include machine learning, spatial-temporal learning, and their applications (e.g., Earth System Science and Smart City).

Dr. Bai is or was a Program Committee Member or Reviewer for IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, NeurIPS, International Conference on Machine Learning (ICML), International Conference on Learning Representations (ICLR), CVPR, ICCV, Association for the Advancement of Artificial Intelligence (AAAI), IJCAI, KDD, European Conference on Computer Vision (ECCV), IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON MULTIMEDIA, and *ACM Transactions on Sensor Networks*. He was a recipient of the 2020 Google Ph.D. Fellowship, the 2020 UNSW Engineering Excellence Award, and the 2021 Dean's Award for Outstanding Ph.D. Theses.



Wanli Ouyang received the Ph.D. degree from the Department of Electronic Engineering, Chinese University of Hong Kong, Hong Kong, in 2010.

He was an Associate Professor with The University of Sydney, Camperdown, NSW, Australia. He is now a Professor with Shanghai Artificial Intelligence (AI) Laboratory, Shanghai, China. His research interests include pattern recognition, machine learning, and AI for Science.

Dr. Ouyang served as an Associate Editor for *International Journal of Computer Vision* (IJCV) and *Pattern Recognition* (PR), the Senior Area Chair for Conference on Computer Vision and Pattern Recognition (CVPR), and the Guest Editor for IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (TPAMI).