RS-Mamba for Large Remote Sensing Image Dense Prediction

Sijie Zhao, Hao Chen*, Xueliang Zhang*, Pengfeng Xiao, Lei Bai, and Wanli Ouyang

***Abstract*—The spatial resolution of remote sensing images is becoming increasingly higher, posing challenges in handling large very-high-resolution (VHR) remote sensing images for dense prediction tasks. Models based on convolutional neural networks are limited in their ability to model global features of remote sensing images due to local convolution operations. Transformer-based models, despite their global modeling capabilities, face computational challenges with large VHR images due to their quadratic complexity. The common practice of cropping large images into smaller patches leads to a significant loss of contextual information. To address these issues, we propose the Remote Sensing Mamba (RSM) for dense prediction tasks in VHR remote sensing. RSM is designed to model global features of remote sensing images with linear complexity, enabling it to process large VHR images effectively. It employs an omnidirectional selective scan module to globally model the images in multiple directions, capturing large spatial features from various directions. Experiments on semantic segmentation and change detection tasks across various objects demonstrate the effectiveness of RSM. With simple model architecture and training approach, RSM achieves state-of-the-art performance on the dense prediction tasks of VHR remote sensing. The code for this work will be available at *https://github.com/walking-shadow/Official_Remote_Sensing_Mamba*.**

***Index Terms*—Large remote sensing images, Dense prediction, Very high resolution, State space model, Deep learning.**

I. Introduction

THE advent of increasingly high spatial resolution in remote sensing image 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 a pivotal resource in various domains, including urban planning [1], agricultural management [2], environmental monitoring [3], and disaster response [4].

In recent years, the field of remote sensing has witnessed a rapid expansion, due to the unprecedented availability of very-high-resolution (VHR) image. VHR remote sensing images are characterized by spatial features of large spatial scales across multiple directions, which are crucial for dense prediction tasks such as semantic segmentation and change detection. 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 horizontal direction, indicating that the spatial features of these images can exist in any direction. Therefore, the ability to globally model 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 have been widely applied to VHR remote sensing dense prediction tasks [5, 6, 7, 8]. The transformer architecture, famous for its ability to capture global spatial information and model spatial dependencies effectively through self-attention mechanisms, has demonstrated 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 Figure 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. This limitation underscores the need for innovative solutions that can efficiently process whole images or larger segments to preserve and leverage comprehensive spatial relationships inherent in VHR remote sensing image.

The recent work, Mamba [9], integrated time-varying parameters into State Space Models (SSM), and proposed a hardware-aware algorithm, facilitating highly efficient training and inference processes. SSM, which draw inspiration from the classical Kalman filter model [10], excel at capturing long-range dependencies and benefit from parallel training capabilities. Research on Mamba illustrates its potential as a promising alternative to transformers in language modeling due to its robust contextual modeling capacity and linear complexity [9]. However, Mamba is design 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 lacks a specific orientation and where spatial relationships are crucial across all dimensions.

Recent works such as Vim [11] and VMamba [12] have harnessed SSM to achieve linear complexity and a global effective receptive field, tackling tasks like image classification

Sijie Zhao, Xueliang Zhang, and Pengfeng Xiao are with the 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 is with the Shanghai Artificial Intelligence Laboratory, Shanghai 200000, China (e-mail: chenhao1@pjlab.org.cn, bailei@pjlab.org.cn, ouyangwanli@pjlab.org.cn).

Corresponding Author: Hao Chen and Xueliang Zhang.

This work was done during his internship at Shanghai Artificial Intelligence Laboratory.

Fig. 1. Illustration of the image preprocess strategy of transformer-based model when using large images. Dividing large VHR remote sensing image into small patches would lose lots of spatial features. Each patch contains very limited contextual information compared to original large image.

and segmentation on natural images. To address the challenge of image data's non-directionality, Vim employs SSM for selective scanning in both forward and backward directions along the horizontal axis of an image. VMamba extends this approach by conducting selective scanning with SSM in both horizontal and vertical directions, ensuring that every segment of the image can establish connections with other parts. The visualization of the effective receptive field in VMamba demonstrates its global effective receptive field and enhanced effective receptive field across both horizontal and vertical directions [12]. This indicates that the selective scanning direction of SSM can significantly impact the effective receptive field in specific orientations.

However, Vim and VMamba are not ideally suited for VHR remote sensing image. Natural images adhere to certain physical laws and can not rotate freely, their main spatial features distributed along horizontal and vertical directions. In contrast, remote sensing images can be rotated freely as they are captured from a top-down satellite perspective, which makes 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 image contain large spatial features in multiple directions. Due to the significant impact of 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 a global receptive field and linear complexity. We propose Remote Sensing Mamba (RSM) to process VHR remote sensing images, leveraging the strengths of SSM to extract large and multi-directional spatial features in VHR remote sensing images with rich contextual information.

RSM have a global receptive field capable of effectively modeling the context of VHR remote sensing images without self-attention operations. Furthermore, due to its linear complexitym, rather than dividing large VHR remote sensing images into small patches, RSM can handle the whole images and process them without losing spatial feature of individual objects or spatial dependency among multiple objects. Thus, RSM is well-equipped to handle VHR remote sensing images efficiently.

Furthermore, we propose the Omnidirectional Selective Scan Module (OSSM) to extract spatial features in VHR remote sensing images that span large spatial scales and multiple directions. OSSM employs SSM for selective scanning in forward and backward directions across horizontal, vertical, diagonal, and anti-diagonal axes. This approach enhances the global effective receptive field of the remote sensing images in multiple directions, allowing for the extraction of comprehensive global spatial features.

In summary, our contributions are as follows:

1) We introduce Remote Sensing Mamba for VHR remote sensing tasks. RSM first introduced SSM to process VHR remote sensing image, which is capable of handling large VHR remote sensing images with rich contextual information.
2) We design an Omnidirectional Selective Scan Module to extract spatial features that span large spatial scales and multiple directions within VHR remote sensing images. Through selective scanning across multiple directions using SSM, OSSM enhances the global effective receptive field in multiple directions, thus extracting global spatial features in multiple directions.
3) We demonstrate the efficiency and superiority of RSM in VHR remote sensing tasks. Experiments on the semantic segmentation dataset (WHU, Massachusetts Road) and the change detection dataset (LEVIR-CD, WHU-CD) show that RSM achieves state-of-the-art performance on both semantic segmentation and change detection tasks.

II. Related Works

A. Very high resolution resolution remote sensing

Deep learning models for VHR remote sensing tasks can be categorized into two types: Convolutional Neural Networks (CNNs) based models and transformer based models.

CNNs excel in image processing due to their ability to efficiently capture local spatial features through their hierarchical structure. Papadomanolaki et al. [13] proposed an urban change detection framework , which combines U-Net [14] for feature extraction and LSTMs [15] for temporal modeling.

Zhao et al. [16] introduced a novel change detection 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 the change detection results. Gu et al. [17] focused on exploiting the multi-scale feature differences between bitemporal images to concentrate on the detailed information of the changing areas.

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, showcasing their ability to perform semantic segmentation and change detection with enhanced global context understanding. Chen et al. [5] integrated transformers into change detection tasks, utilizing the self-attention mechanism to model the global context of remote sensing images. Zhang et al. [8] proposed a purely transformer-based architecture for change detection tasks, constructing a model based on the Swin Transformer architecture.

Despite their strengths, transformer models encounter challenges with the quadratic complexity of their self-attention mechanisms, particularly when processing VHR images. This necessitates dividing these images into smaller segments, which can result in significant loss of spatial features and dependencies. While transformers can model global contexts, the reduced contextual information limits their effectiveness in VHR remote sensing tasks. In response, we propose the RSM for VHR remote sensing tasks, which operates with linear complexity and achieves a global receptive field. RSM is adept at handling images with rich contextual information, thereby providing a more effective solution for processing VHR remote sensing images.

B. State Space Models

State Space Models (SSM) have gained significant traction in the field of deep learning over recent years, marking a remarkable evolution in the way long-range dependencies and sequential data are handled [18, 19, 20]. Initially inspired by their traditional application in control systems, SSM 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 [21], 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 [19]. However, LSSL faced critical hurdles related to computational and memory efficiency, limiting its practical application. Addressing these limitations, the S4 model introduced by Gu et al. emerged as a pivotal advancement, proposing a normalized parameterization strategy that reduced computational overhead, thereby making SSM more feasible for practical applications [18].

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 [22, 23], as well as those supporting multiple-input multiple-output configurations to increase model flexibility [20]. Additionally, innovations such as the decomposition of operations into diagonal plus low-rank structures [24] and the introduction of selection mechanisms [9] have further refined the adaptability and performance of SSM.

However, the aforementioned models are only capable of processing unidirectional sequence data, and cannot handle image data that lacks a specific direction. Recent works, Vim [11] and VMamba [12], have achieved global modeling on images using SSM by conducting selective scanning both forwards and backwards in certain directions. Vim [11] performs selective scanning in the horizontal direction, enabling each part of the image to perceive global information. VMamba [12] extends this 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 horizontal and vertical directions, and VHR remote sensing images exhibit large spatial features in multiple directions, although Vim and VMamba achieved commendable performance on natural images, they are not suitable for VHR remote sensing images. The Omnidirectional Selective Scan Module we proposed conducts selective scanning in multiple directions, capable of capturing the large spatial features of VHR remote sensing images in various directions.

III. Methodology

A. Preliminaries: State Space Models

In the realm of deep learning, State Space Models (SSM) have gained prominence for their ability to encapsulate dynamic systems that map an input sequence $x(t) \in \mathbb{R}^L$ to an output $y(t) \in \mathbb{R}$. SSM 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 [9] 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 the system matrices B and D, based on the current and historical context of the input sequence. Consequently, Mamba's SSM are able to model complex temporal dynamics more effectively, as these matrices adapt in response to the input data's evolving features.

B. Overall Architecture

VHR remote sensing images are primarily utilized in semantic segmentation and change detection tasks. Consequently, we have developed two specialized frameworks: Remote Sensing Mamba for Semantic Segmentation (RSM-SS) for the semantic segmentation task and Remote Sensing Mamba for Change Detection (RSM-CD) for the change detection task, as illustrated in Figure 2. To demonstrate the effectiveness of SSM in processing VHR remote sensing images, RSM-SS and RSM-CD employ the simplest architectures for semantic segmentation and change detection, respectively.

The RSM-SS architecture utilizes the U-Net Encoder-Decoder framework, where input ultra-high resolution 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 semantic segmentation results. The Encoder consists of five stages, each comprising several OSS blocks. Stage 1 extracts features from the input ultra-high resolution remote sensing images, while stages 2-5 progressively downsample the encoder features and double the 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 semantic segmentation from both global and local perspectives.

The RSM-CD employs a FC-Siam-Conc [25] 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 change detection 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 ultra-high resolution 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. Omnidirectional state space block

The Omnidirectional state space (OSS) block is a novel feature extraction unit designed for semantic segmentation and change detection tasks in VHR remote sensing image, as shown in Figure 2. Central to the OSS block is the Oriented Scanning Module (OSSM), 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 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 parameter count and focusing on extracting spatial features. Subsequent to the convolution, the features pass through the OSSM. 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. The output from the OSSM then undergoes a linear transformation and a gating operation, adjusting deep features with outputs of linear transformation of the normalized features. Lastly, the features pass to a final linear layer and add input features through residual connection.

The OSS block is crafted with a keen focus on balancing computational efficiency and the capability 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 budget of total model depth when building the model.

D. Omnidirectional selective scan module

Vim and VMamba have demonstrated commendable performance in the natural images, where primary spatial features distributed along horizontal and vertical directions. Vim conducts selective scanning in the forward and backward directions along the horizontal axis, and VMamba introduces

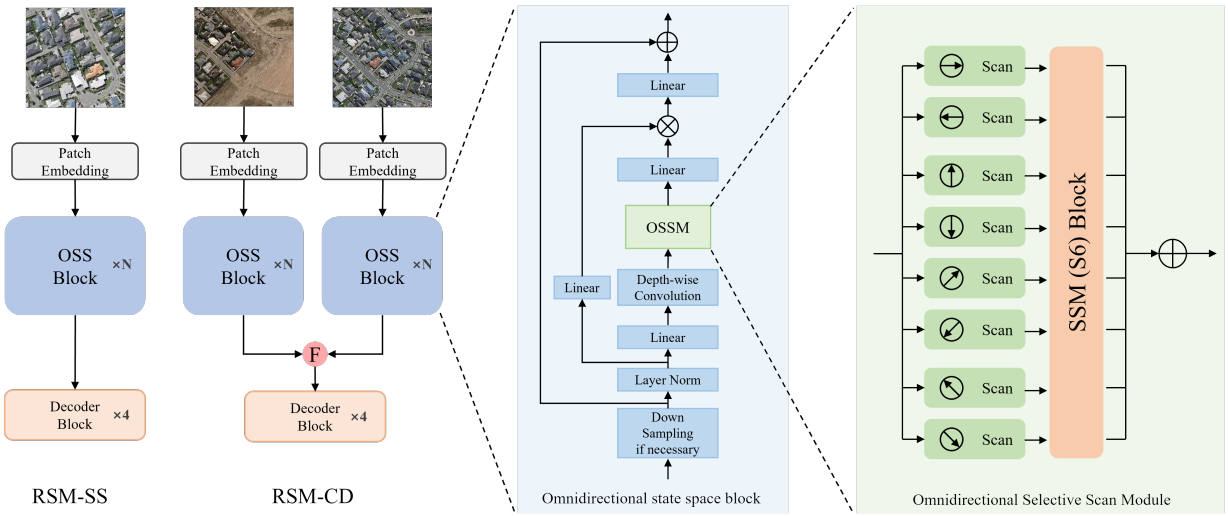


Fig. 2. Illustration of the Overall structure of RSM-SS and RSM-CD. RSM-SS and RSM-CD can globally model the images in multiple directions with linear complexity using omnidirectional selective scan.

selective scanning across both horizontal and vertical directions, allowing every part of the image to globally attend in both forward and backward directions along specific direction, as illustrated in Figure 3.

However, Vim and VMamba fall short when applied to VHR remote sensing image, where spatial features exist in arbitrary directions. VHR remote sensing image contains large spatial features in multiple directions, such as the edges of objects and their arrangements. Relying solely on global modeling along horizontal and vertical axes impedes the model's ability to extract spatial features in other directions. To address this, we propose the Omnidirectional Selective Scan Module (OSSM), which performs selective scanning in the forward and backward directions along the horizontal, vertical, diagonal, and anti-diagonal directions, as depicted in Figure 3. This approach enables global modeling of VHR remote sensing images across multiple directions, facilitating the extraction of large spatial features from various angles.

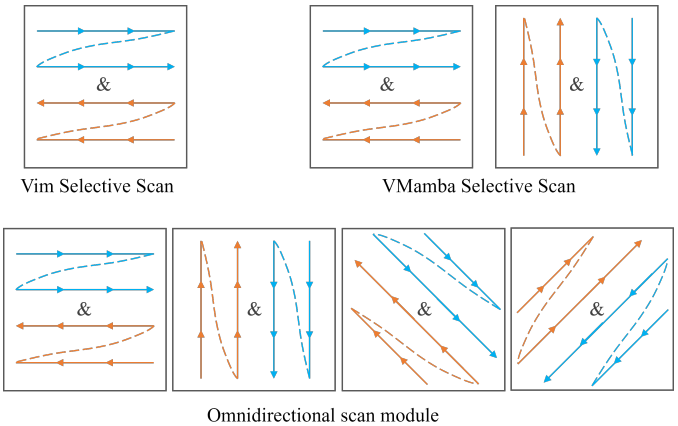


Fig. 3. Illustration of the selective scan directions of Vim, VMamba and OSSM.

Specifically, the structure of OSSM is illustrated in Figure 5. OSSM begins with the input image patches undergoing omnidirectional scanning in horizontal, vertical, diagonal, anti-diagonal, 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 [9]. 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: semantic segmentation and change detection. For the semantic segmentation task, we evaluated the effectiveness of the RSM-SS model on the WHU dataset and the Massachusetts Road dataset. For the change detection task, we evaluated the effectiveness of the RSM-SS model on the WHU-CD dataset and the LEVIR-CD dataset.

A. Datasets

We offer a brief description of the experimental semantic segmentation and change detection datasets in Table I.

TABLE I
Brief Introduction of The Experimental Datasets.

Name	Task	Resolution (m)	Images	Image size
WHU [26]	Seg	0.3	8189	512×512
M-Road [27]	Seg	1	1171	1500 × 1500
WHU-CD [26]	CD	0.075	1	32207×15354
LEVIR-CD [28]	CD	0.5	637	1024×1024

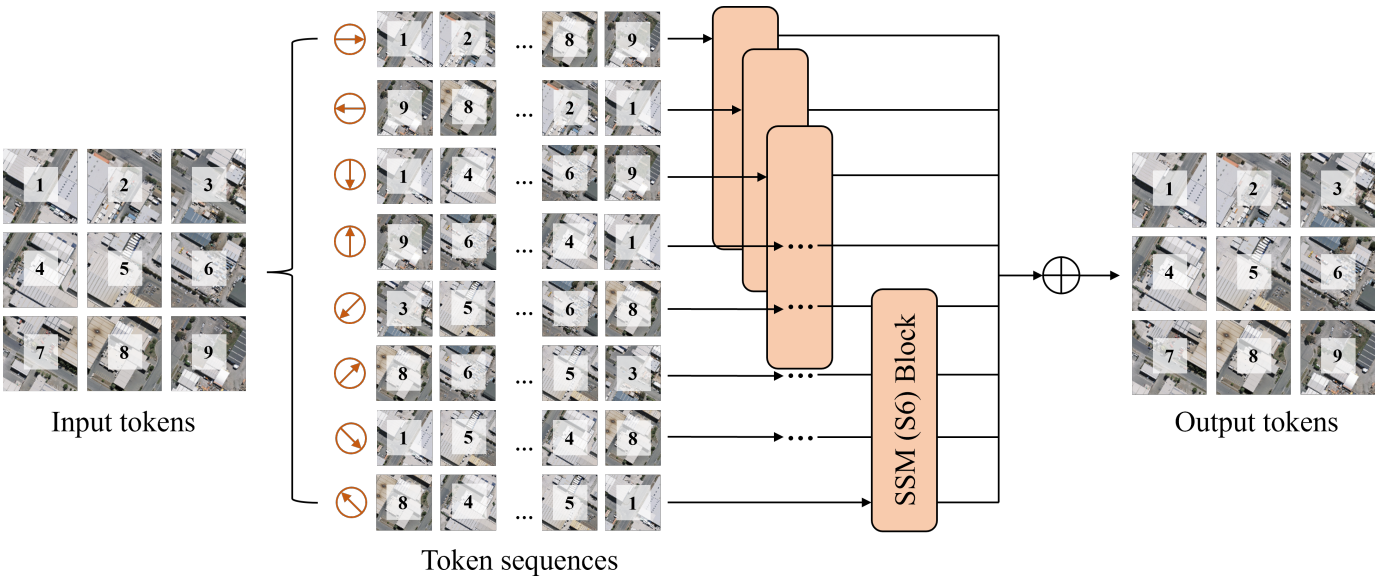


Fig. 4. Illustration of the structure of Omnidirectional selective scan module. OSSM

1) Semantic Segmentation Datasets: The WHU [26] Building Dataset is comprised 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 8,189 images. They are divided into 4,736 images designated for training, 1,036 for validation, and 2,416 for testing purposes, each with a spatial resolution of 0.3 meters. This subset collectively captures approximately 22,000 buildings across an expanse of more than 450 square kilometers.

The Massachusetts [27] Roads Dataset incorporates 1,171 aerial photographs from Massachusetts, with each image measuring 1500×1500 pixels and encompassing 2.25 square kilometers. This dataset is organized into 1,108 training images, 14 validation images, and 49 test images. It encompasses a diverse array of environments, including urban, suburban, and rural areas, spanning over 2,600 square kilometers, with the test segment alone covering in excess of 110 square kilometers. 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) Change Detection Datasets: The WHU-CD [26] 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 into training, validation, and test sets in a 7:1:2 ratio.

The LEVIR-CD [28] dataset is an extensive change detection dataset comprising VHR (0.5 m/pixel) Google Earth images that document a range of building transformations over a period of 5 to 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 1024× 1024 pixel patches.

B. Benchmark Methods

To evaluate the effectiveness of the proposed Remote Sensing Mamba, we conducted comparative experiments with various benchmark methods on the semantic segmentation and change detection 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 semantic segmentation task, the compared CNN-based models include FCN [29], SegNet [30], U-Net [14], PSPNet [31], HRNet [32], MA-FCN [33], Deeplabv3+ [34], ResUNet [35], MAP-Net [36], D-LinkNet [37] and SI-INet [38], and the compared transformer-based models include Segformer [39], RoadFormer [7] and BDTNet [40].

On the change detection task, the compared CNN-based models include FC-EF [25], FC-Siam-Diff [25], FC-Siam-Conc [25], STANet [41], DTCDSCN [42], SNUNet [43], CDNet [44], DDCNN [45], DASNet [46] and DSIFN [47], and the compared transformer-based models include BIT [5], ChangeFormer [6], MTCNet [48] and MSCANet [49].

C. Implementation Details

1) Data Augmentation: To demonstrate the effectiveness of the proposed methods, we only employed the straightforward data augmentation techniques, avoiding the use of any elaborate tricks. For the semantic segmentation task, the data augmentation methods used for the RSM-SS model included flipping (p=0.5) and transposing (p=0.5). For the change detection task, the data augmentation methods used for the RSM-SS 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 [50] to construct and deploy RSM-SS and RSM-CD on a single RTX A100 GPU. Given the variable image sizes across datasets, we

adjusted the batch sizes accordingly: 16 for the WHU dataset, 4 for both the Massachusetts Roads 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 [51] optimizer with an initial learning rate of 0.001 and a weight decay of 0.001. The learning rate adjustment strategy is reducing 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 achieve convergence. Checkpoints capturing the highest F1-scores on the validation sets were preserved for subsequent testing. To maintain consistency with other change detection 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 targets the false negatives. Achieving high scores in both precision and recall simultaneously poses a significant challenge due to their inversely proportional relationship. The F1-score, representing 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 the OSSM, comparative experiments were conducted on the Massachusetts Roads dataset for the semantic segmentation task and the WHU-CD dataset for the change detection task. We compared the performance of three variations: SS1D [11], which employs selective scanning in the horizontal direction and its reverse; SS2D [12], which includes selective scanning in both horizontal and vertical directions and their 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 tasks of semantic segmentation and change detection. Specifically, for the semantic segmentation task on the Massachusetts Roads dataset, the presence of roads extending in multiple directions with significant spatial scales necessitates selective scanning across multiple directions, therefore extracting large road features oriented in various directions. Similarly, in the change detection 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 enable the extraction of large object features from multiple directions, making it more suited for VHR remote sensing images.

TABLE II
ABLATION STUDY OF OSSM ON 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**

E. Experimental Results

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

The accuracy comparison results on the Massachusetts Road dataset are presented in Table III. It shows that RSM-SS surpasses all the compared models and achieves the highest IoU (0.6852) and F1-score (0.8132) 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 semantic segmentation. 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 contextual information in the images.

TABLE III
ACCURACY COMPARISON ON THE MASSACHUSETTS ROAD DATASET. THE BEST VALUES ARE HIGHLIGHTED IN BOLD.

Methods	P (%)	R (%)	F1 (%)	IoU (%)
SegNet [30]	76.09	78.23	77.15	62.79
U-Net [14]	77.53	77.82	77.67	63.50
ResUNet [35]	78.77	77.45	78.10	64.07
D-LinkNet [37]	78.34	77.91	78.12	64.10
HRNetv2 [32]	79.01	78.2	78.60	64.75
Deeplabv3+ [34]	75.14	72.56	73.83	58.51
SIINet [38]	85.36	74.13	79.35	65.77
RoadFormer [7]	80.54	**78.9**	79.71	66.27
BDTNet [40]	82.99	76.37	79.54	66.03
RSM-SS	**86.52**	75.24	**80.49**	**67.35**

The accuracy comparison results on the WHU dataset are summarized in Table IV. The accuracy comparison results show that RSM-SS achieves the highest IoU (0.9081) and F1-score (0.9518) on this dataset. In the WHU dataset, the 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 the VHR remote sensing image.

TABLE IV
ACCURACY COMPARISON ON THE WHU DATASET. THE BEST VALUES ARE HIGHLIGHTED IN BOLD.

Methods	P (%)	R (%)	F1 (%)	IoU (%)
FCN [29]	92.29	92.84	92.56	86.16
SegNet [30]	93.42	91.71	92.56	86.15
U-Net [14]	94.50	90.88	92.65	86.31
PSPNet [31]	93.19	94.21	93.70	88.14
HRNet [32]	91.69	92.85	92.27	85.64
MA-FCN [33]	94.75	94.92	94.83	90.18
Deeplabv3+ [34]	94.31	94.53	94.42	89.43
ResUNet [35]	94.49	94.71	94.60	89.75
MAP-Net [36]	93.99	94.82	94.40	89.40
Segformer [39]	94.72	94.42	94.57	89.70
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 Figure 5. It shows that 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, RSM-SS is capable of extracting building features across various directions.

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

The accuracy comparison results on the WHU-CD dataset are shown in Table V. It shows that RSM-CD achieves the highest IoU (0.8496) and F1-score (0.9187) on this dataset, outperforming all other change detection models. Given the very high spatial resolution of the WHU-CD dataset remote sensing images (0.075m/pixel), a normal size image (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 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. It shows that RSM-CD outperforms all 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 bi-temporal remote sensing images underscores the importance of extracting large features in multiple directions. The omnidirectional selective scan module of RSM-CD can extract large spatial features of buildings from various directions, thereby accurately identifying changed buildings.

We show some inference results of the test set of the WHU-CD and LEVIR-CD datasets in Figure 6. It shows that RSM-CD can accurately detect all the changed buildings in WHU-CD and LEVIR-CD datasets. On the WHU-CD dataset, the high spatial resolution of the 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 bi-temporal images, it accurately identifies changed buildings. On the LEVIR-CD dataset, where buildings exhibit edge features in multiple directions, RSM-CD's ability to perform selective scanning in multiple directions allows it to extract building features from various orientations, accurately identifying changed buildings.

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

Methods	P (%)	R (%)	F1 (%)	IoU (%)
FC-EF [25]	78.86	78.64	78.75	64.95
FC-Siam-Diff [25]	84.73	87.31	86.00	75.44
FC-Siam-Conc [25]	78.86	78.64	78.75	64.95
DTCDSCN [42]	63.92	82.30	71.95	56.19
DSIFN [47]	91.44	89.75	90.59	82.79
STANet [41]	79.37	85.50	82.32	69.95
SNUNet [43]	85.60	81.49	83.49	71.67
DASNet [46]	88.23	84.62	86.39	76.04
CDNet [44]	91.75	86.89	89.25	80.59
DDCNN [45]	**93.71**	89.12	91.36	84.09
BIT [5]	86.64	81.48	83.98	72.39
MTCNet [48]	75.10	**91.90**	82.65	70.43
MSCANet [49]	91.10	89.86	90.47	82.60
RSM-CD	93.37	90.42	**91.87**	**84.96**

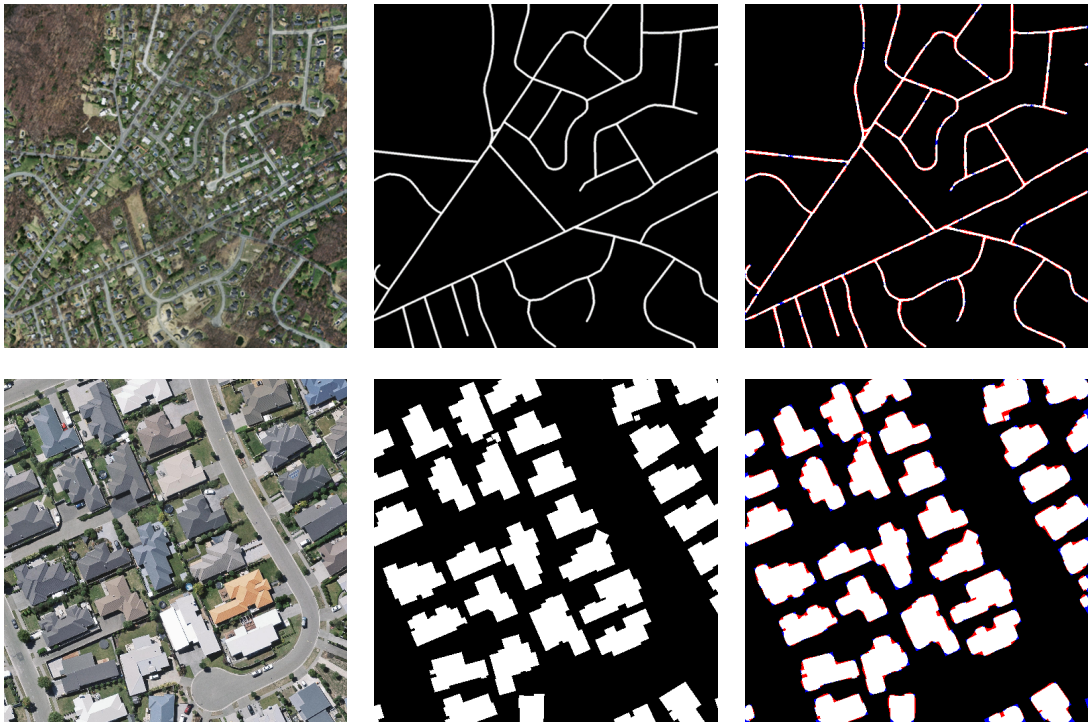
TABLE VI
ACCURACY COMPARISON ON THE LEVIR-CD DATASET. THE BEST VALUES ARE HIGHLIGHTED IN BOLD.

Methods	P (%)	R (%)	F1 (%)	IoU (%)
FC-EF [25]	86.91	80.17	83.40	71.53
FC-Siam-Diff [25]	89.53	83.31	86.31	75.91
FC-Siam-Conc [25]	91.99	76.77	83.69	71.96
DTCDSCN [42]	88.53	86.83	87.67	78.05
DSIFN [47]	**94.02**	82.93	88.13	78.77
STANet [41]	83.81	**91.00**	87.26	77.39
SNUNet [43]	89.18	87.17	88.16	78.83
CDNet [44]	91.6	86.5	88.98	80.14
DDCNN [45]	91.85	88.69	90.24	82.22
BIT [5]	89.24	89.37	89.30	80.68
ChangeFormer [6]	92.05	88.8	90.40	82.47
MTCNet [48]	90.87	89.62	90.24	82.22
MSCANet [49]	91.30	88.56	89.91	81.66
RSM-CD	92.52	89.73	**91.10**	**83.66**

V. DISCUSSION

A. 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 semantic segmentation and change detection 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.



(a) (b) (c)

Fig. 5. Sample inference results of RSM-SS on the semantic segmentation task. The results on 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.

(a) (b) (c) (d)

Fig. 6. Sample inference results of RSM-CD on the change detection task. The results on 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.

In our semantic segmentation experiments on the Massachusetts Roads 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, 128, 256, 512, and 1024 pixels. It's 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 ratio2/ratio1 times the spatial

range of the former. We used the F1 Score as a metric to evaluate the model's performance across different image sizes and downsampling ratios, with the results illustrated in Figure 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 contextual information and high-resolution spatial features are important for segmenting roads.

In the change detection 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 Roads 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 Figure 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 identify changed buildings accurately. Downsampling results in a significant loss of spatial features, substantially reducing the model's performance. 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 declines.

The experiments on the Massachusetts Roads and WHU-CD datasets highlight the importance of large contextual information and high-resolution spatial features for dense prediction tasks in VHR remote sensing. The model's performance varies in response to the loss of contextual information and spatial features, with a more significant performance drop when downsampling road images compared to building images. However, a common finding is that the model performs better on higher spatial resolution images. Moreover, the higher the spatial resolution of the images, 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 RSM enables it to handle large VHR remote sensing images, thereby achieving excellent results in dense prediction tasks for VHR remote sensing.

B. Expectations and Limitations

In the realm of VHR remote sensing, contextual information within images is crucial for dense prediction tasks. However, current models based on Convolutional Neural Networks (CNNs) and transformers struggle with processing 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 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. Experimental results in semantic segmentation and change detection 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 RSM in dense prediction tasks for VHR remote sensing. We hope RSM can serve as a baseline in the field, promoting the development of methods based on SSM within VHR remote sensing.

Despite the notable performance of RSM in dense prediction tasks for VHR remote sensing, it also exhibits some limitations. On one hand, the models for semantic segmentation and change detection tasks are overly simplistic, not fully leveraging the potential of SSM. On the other hand, dense prediction tasks in VHR remote sensing require extensive training data, thus limiting RSM's applicability in tasks lacking labeled data.

VI. Conclusion

We proposed a Remote Sensing Mamba for dense prediction tasks in ultra-high resolution 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 ultra-high resolution remote sensing images with rich contextual information at linear complexity. Through selective scanning in multiple directions, RSM models global context information and extracts large spatial features across various directions, thereby efficiently accomplishing dense prediction tasks.

Experiments on semantic segmentation and change detection tasks demonstrate RSM's superior performance across different objects. Leveraging the State Space Model for processing large images and modeling global context information, RSM operates on ultra-high resolution remote sensing images without the need to segment these images into smaller patches,

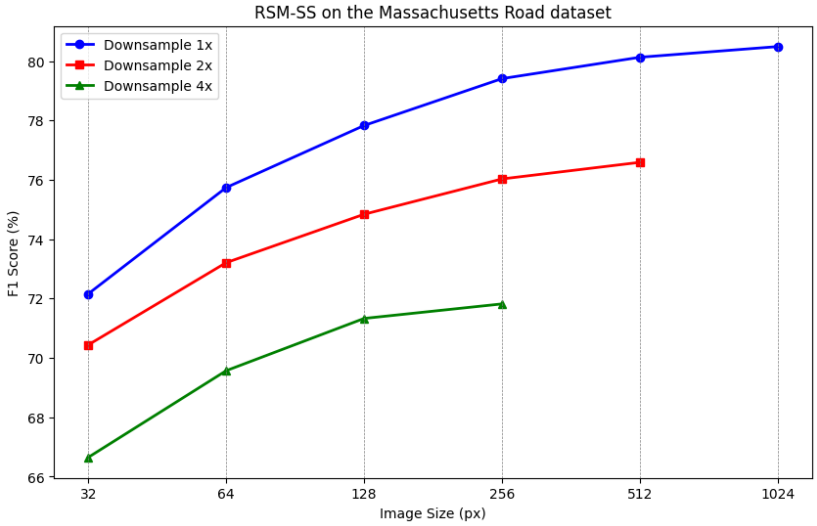


Fig. 7. Illustration of the performance of RSM-SS on the Massachusetts Roads dataset with different image sizes and downsampling ratios.

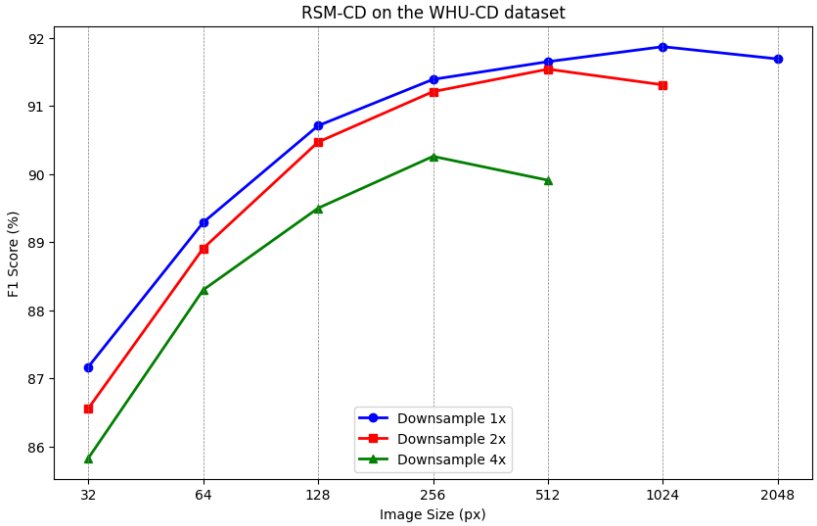


Fig. 8. Illustration of the performance of RSM-SS on the WHU-CD dataset with different image sizes and downsampling ratios.

which is achieved through its linear complexity. By modeling globally in various directions, 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 ultra-high resolution remote sensing, promoting further development of SSM-based approaches in this field.

ACKNOWLEDGMENT

This work was done during his internship at Shanghai Artificial Intelligence Laboratory. They would also like to thank the editor and the anonymous reviewers for their constructive comments.

REFERENCES

[1] T. Wellmann, A. Lausch, E. Andersson, S. Knapp, C. Cortinovis, J. Jache, S. Scheuer, P. Kremer, A. Mascarenhas, R. Kraemer *et al.*, "Remote sensing in urban planning: Contributions towards ecologically sound policies?" *Landscape and urban planning*, vol. 204, p. 103921, 2020.

[2] M. Weiss, F. Jacob, and G. Duveiller, "Remote sensing for agricultural applications: A meta-review," *Remote sensing of environment*, vol. 236, p. 111402, 2020.

[3] S. Asadzadeh, W. J. de Oliveira, and C. R. de Souza Filho, "Uav-based remote sensing for the petroleum industry and environmental monitoring: State-of-the-art and perspectives," *Journal of Petroleum Science and Engineering*, vol. 208, p. 109633, 2022.

[4] 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 Sensing of Environment*, vol. 265, p. 112636, 2021.

[5] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2021.

[6] W. G. C. Bandara and V. M. Patel, "A transformer-based siamese network for change detection," in *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2022, pp. 207–210.

[7] X. Jiang, Y. Li, T. Jiang, J. Xie, Y. Wu, Q. Cai, J. Jiang, J. Xu, and H. Zhang, "Roadformer: Pyramidal deformable vision transformers for road network extraction with remote sensing images," *International Journal of Applied Earth Observation and Geoinformation*, vol. 113, p. 102987, 2022.

[8] C. Zhang, L. Wang, S. Cheng, and Y. Li, "Swinsunet: Pure transformer network for remote sensing image change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

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

[10] R. E. Kalman, "A new approach to linear filtering and prediction problems," 1960.

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

[12] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, and Y. Liu, "Vmamba: Visual state space model," *arXiv preprint arXiv:2401.10166*, 2024.

[13] M. Papadomanolaki, S. Verma, M. Vakalopoulou, S. Gupta, and K. Karantzalos, "Detecting urban changes with recurrent neural networks from multitemporal Sentinel-2 data," in *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2019, pp. 214–217.

[14] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[15] 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," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[16] 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 Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023.

[17] F. Gu, P. Xiao, X. Zhang, Z. Li, and D. Muhtar, "Fdff-net: A full-scale difference feature fusion network for change detection in high-resolution remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2023.

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

[19] A. Gu, I. Johnson, K. Goel, K. Saab, T. Dao, A. Rudra, and C. Ré, "Combining recurrent, convolutional, and continuous-time models with linear state space layers," *Advances in neural information processing systems*, vol. 34, pp. 572–585, 2021.

[20] J. T. Smith, A. Warrington, and S. W. Linderman, "Simplified state space layers for sequence modeling," *arXiv preprint arXiv:2208.04933*, 2022.

[21] A. Gu, T. Dao, S. Ermon, A. Rudra, and C. Ré, "Hippo: Recurrent memory with optimal polynomial projections," *Advances in neural information processing systems*, vol. 33, pp. 1474–1487, 2020.

[22] A. Gu, K. Goel, A. Gupta, and C. Ré, "On the parameterization and initialization of diagonal state space models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 35 971–35 983, 2022.

[23] A. Gupta, A. Gu, and J. Berant, "Diagonal state spaces are as effective as structured state spaces," *Advances in Neural Information Processing Systems*, vol. 35, pp. 22 982–22 994, 2022.

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

[25] R. C. Daudt, B. Le Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 4063–4067.

[26] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery dataset," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 1, pp. 574–586, 2018.

[27] V. Mnih, *Machine learning for aerial image labeling*. University of Toronto (Canada), 2013.

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

[29] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[30] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[31] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2881–2890.

[32] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5693–5703.

[33] S. Wei, S. Ji, and M. Lu, "Toward automatic building footprint delineation from aerial images using cnn and regularization," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 3, pp. 2178–2189, 2019.

[34] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[35] Y. Xu, L. Wu, Z. Xie, and Z. Chen, "Building extraction in very high resolution remote sensing imagery using deep learning and guided filters," *Remote Sensing*, vol. 10, no. 1, p. 144, 2018.

[36] Q. Zhu, C. Liao, H. Hu, X. Mei, and H. Li, "Mapnet: Multiple attending path neural network for building footprint extraction from remote sensed imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 7, pp. 6169–6181, 2020.

[37] L. Zhou, C. Zhang, and M. Wu, "D-linknet: Linknet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 182–186.

[38] C. Tao, J. Qi, Y. Li, H. Wang, and H. Li, "Spatial information inference net: Road extraction using road-specific contextual information," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 158, pp. 155–166, 2019.

[39] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in neural information processing systems*, vol. 34, pp. 12 077–12 090, 2021.

[40] 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 Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

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

[42] 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 Geoscience and Remote Sensing Letters*, vol. 18, no. 5, pp. 811–815, 2020.

[43] S. Fang, K. Li, J. Shao, and Z. Li, "SNUNet-CD: A densely connected siamese network for change detection of vhr images," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2021.

[44] H. Chen, W. Li, and Z. Shi, "Adversarial instance augmentation for building change detection in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2021.

[45] X. Peng, R. Zhong, Z. Li, and Q. Li, "Optical remote sensing image change detection based on attention mechanism and image difference," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 9, pp. 7296–7307, 2020.

[46] 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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 1194–1206, 2020.

[47] 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 Journal of Photogrammetry and Remote Sensing*, vol. 166, pp. 183–200, 2020.

[48] W. Wang, X. Tan, P. Zhang, and X. Wang, "A cbam based multiscale transformer fusion approach for remote sensing image change detection," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 6817–6825, 2022.

[49] M. Liu, Z. Chai, H. Deng, and R. Liu, "A cnn-transformer network with multiscale context aggregation for fine-grained cropland change detection," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 4297–4306, 2022.

[50] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035.

[51] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.