Boosting Vision State Space Model with Fractal Scanning

Haoke Xiao*, Lv Tang*✉, Peng-Tao Jiang, Hao Zhang, Jinwei Chen, Bo Li✉

vivo Mobile Communication Co., Ltd, Shanghai, China
{xiaohaoke, lvtang, pt.jiang, haozhang, jinwei.chen, libra}@vivo.com

Abstract

Recently, foundational models have significantly advanced in different tasks, accompanied by Transformer as the general backbone. However, Transformer's quadratic complexity poses challenges for handling longer sequences and higher resolution images, which may limit foundational models further development. To alleviate this issue, various efficient State Space Models (SSMs) like Mamba have emerged, initially matching Transformer performance and gradually surpassing it. To improve the performance of SSMs in computer vision tasks, one crucial viewpoint is effective serialization of images. Existing vision Mambas, which rely on a linear scanning mechanism, often struggle to capture complex spatial relationships in 2D images. This results in feature loss during serialization and negatively impacts model performance. To overcome this limitation, we propose the use of fractal scanning curves for image serialization to enhance the Mambas' ability to accurately model complex spatial dependencies. Additionally, unlike existing vision Mambas, which are designed with various curve scanning directions that increase the complexity, contradicting the original intent of Mamba to enhance model performance. We novelty introduce the Fractal Fusion Pathway (FFP) for our FractalMamba, which can enhance its performance efficiently. Extensive experiments underscore the superiority of our proposed FractalMamba.

Introduction

Foundational models have rapidly evolved within the domains of natural language processing (NLP) and computer vision (CV), giving rise to a plethora of works (Devlin et al. 2019; Radford et al. 2021; Li et al. 2022a; Chowdhery et al. 2023; OpenAI 2023; Touvron et al. 2023; Oquab et al. 2023; Kirillov et al. 2023; Li et al. 2023; Zheng et al. 2024), such as LLaVA (Liu et al. 2023), InternVL (Chen et al. 2024b) and SAM2 (Ravi et al. 2024). The common used backbone of modern foundational models is Transformer (Vaswani et al. 2017), a type of sequential model. Although the self-attention mechanism in Transformer ensures effective modeling of complex data across varied contexts, its quadratic complexity poses significant efficiency challenges, thus presenting challenges to the further advancement of foundational models when facing longer language lengths or larger image resolutions.

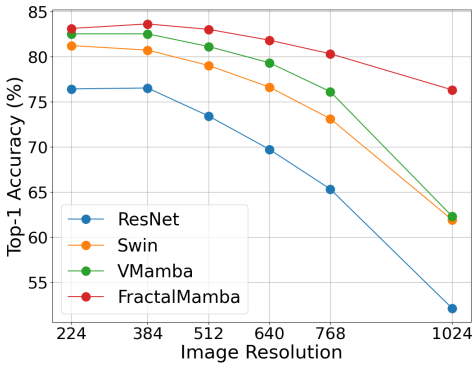

Figure 1: Classification results of different models varies across different scales on ImageNet-1K (Deng et al. 2009).

Recently, Mamba (Gu and Dao 2023; Dao and Gu 2024), based on the State Space Model (SSM), has achieved performance comparable to Transformers in NLP, while maintaining linear time complexity. Subsequently, researchers have made further improvements to Mamba (Huang et al. 2024; Yang et al. 2024; Zhu et al. 2024a; Liu et al. 2024), enabling it to surpass the Vision Transformer (ViTs) in different CV tasks, such as image classification. These developments have inspired further exploration into SSMs to ascertain their potential as viable alternatives to ViTs.

For fully leveraging the modeling capabilities of SSMs in CV tasks, one of the primary challenges is the effective serialization of image patches (Liu et al. 2024; Huang et al. 2024; Zhu et al. 2024b,b). Effective serialization can accurately represent and retain the intricate spatial relationships and structural properties inherent in images as far as possible, which are essential for SSMs to capture the representation of images (Huang et al. 2024; Zhu et al. 2024b). Consequently, numerous studies have designed various scanning mechanisms to optimize the serialization. For example, ViM (Zhu et al. 2024a) employs tion a bi-directional scanning curve that processes image patches horizontally (row by row) to capture spatial dependencies. Building upon this, VMamba (Liu et al. 2024) adds serialization along vertical columns (column by column), enriching the analysis of spatial relationships. LocalMamba (Huang et al. 2024) introduces a window mechanism that divides the image into windows, serializing

* Equal contribution.
✉ Corresponding author.

image patches within each window first, and then similarly processing between windows, which helps SSMs in capturing comprehensive image features. PlainMamba (Yang et al. 2024) introduces a continuous scanning strategy designed to address this issue by simply adjusting the propagation direction at points of discontinuity.

Although the above SSM-based methods have achieved notable performance, they generally utilize linear scanning curves, such as Zigzag, which may have inherent limitations and fail to fully capture the complex spatial relationships and structural properties within images. In other words, when serializing image patches, these linear scanning curves find it difficult to maintain the structural association of these image patches in 2D images, which may lead to the loss of structural information. Additionally, the row-by-row or column-by-column property causes linear scanning to fall into the repetitive pattern, which would introduce biases in the subsequent modeling by SSMs, as the network might over-fit to these repetitive patterns rather than capturing the true underlying dynamics of the image. For example, the reliance on the repetitive scanning patterns may limit the model's ability to adapt to dynamic changes in image resolution, further constraining its applicability across various resolutions.

To address these shortcomings, inspired by fractal theories (Gotsman and Lindenbaum 1996), we propose the use of fractal scanning curves as a more advanced approach for image patch serialization and design the FractalMamba. Unlike linear curves, fractal curves can help maintain spatial structural relationships throughout the serialization process as complete as possible. As shown in Fig. 2, the image patches in the red area, although connected in the 2D image, are no longer adjacent after being serialized through a linear curve, which to some extent destroys the spatial structure of the original image. However, for fractal curves, they can maintain these spatial structures as much as possible. Additionally, the intricate and complex path planning of fractal curves effectively avoids the repetitive scanning cycles typical of other methods, fully leveraging SSMs to dynamically capture and analyze complex patterns in images. Furthermore, due to their self-similarity, fractal curves consistently preserve structural capturing capabilities across various resolutions, ensuring the model's adaptability to different scales. As shown in Fig. 1, at low resolutions, the performance of FractalMamba in classification is still close to that of some typical methods, but as the resolution gradually increases, the performance advantages of FractalMamba become apparent.

However, due to the complexity of the objects and structures represented in different images, serialization using fractal curves, although capable of maintaining some structural integrity, still results in some loss of information. Current methods(Huang et al. 2024; Yang et al. 2024; Zhu et al. 2024a; Liu et al. 2024) address this by designing scanning curves that traverse the image from different directions to better simulate these structures. Nonetheless, using curves from multiple directions implies different serialization methods, which undeniably introduces additional computational complexity. To solve this problem, we propose the Fractal Fusion Pathway (FFP). This method aims to efficiently enhance the ability of fractal curves to model complex structures. The

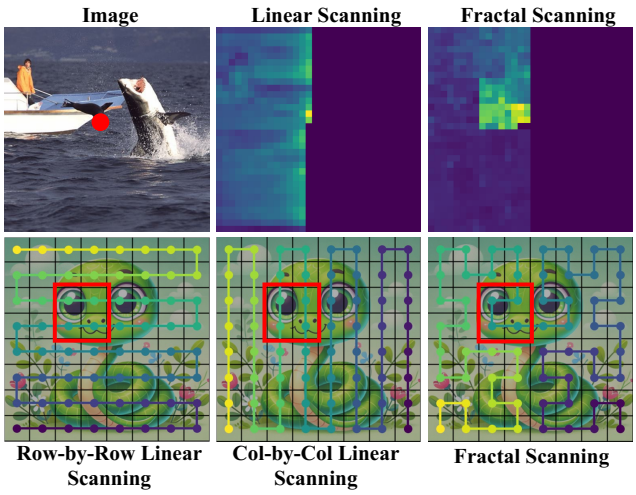

Figure 2: Comparison between Linear Scanning and Fractal Scanning. From the attention map and the curve path, it can be seen that the fractal curve can better capture the spatial relationship between an image patch and its adjacent patches, thus better preserving the spatial structure.

contributions of this paper are listed as follows:

- We identify and address the inherent limitations of linear scanning curves traditionally used in Vision Mambas. By analyzing their inefficiencies, we advocate for the adoption of fractal scanning curves, which are better suited to preserving the spatial correlation of images.
- We analyze the limitations of fractal curves in Vision Mambas and introduce the FFP. This aims to efficiently enhance the fractal curves' ability to model complex structures while effectively reducing complexity.
- We evaluate the efficacy of FractalMamba across various-resolutions vision tasks: image classification task, semantic segmentation task, remote sensing detection task and common object detection task. The experimental results unequivocally demonstrate enhanced performance and broad applicability of our proposed FractalMamba.

Related Work

Vision Backbone Architecture

In the domain of vision backbone architectures, significant advancements have been achieved through the development and refinement of several key frameworks. The existing widely used backbones contain CNN-based and ViT-based, each offering unique advantages and suited for different tasks within the field of CV. CNNs (Krizhevsky, Sutskever, and Hinton 2012; Simonyan and Zisserman 2015; He et al. 2016; Howard et al. 2017; Radosavovic et al. 2020) have been foundational in the progress of vision models, predominantly due to their ability to capture spatial hierarchies in images. Pioneering models like AlexNet (Krizhevsky, Sutskever, and Hinton 2012), VGG (Simonyan and Zisserman 2015), and ResNet (He et al. 2016) have set the benchmarks in various vision tasks and continue to be pivotal in many applications. ViTs (Vaswani et al. 2017; Dosovitskiy et al. 2021; Liu et al. 2021; Touvron, Cord, and Jégou 2022) represent a paradigm

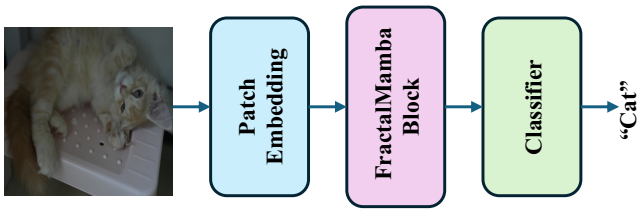

Figure 3: The architecture of our FractalMamba Backbone.

shift in vision processing by applying the principles of self-attention across patches of an image, treating them akin to tokens in natural language processing. This architecture not only capitalizes on the Transformer's ability to handle long-range dependencies but also sets a robust foundation for the explosion of modern foundational models (Devlin et al. 2019; Radford et al. 2021; Li et al. 2022a; Chowdhery et al. 2023; OpenAI 2023; Touvron et al. 2023; Oquab et al. 2023; Liu et al. 2023; Li et al. 2023). However, despite their remarkable performance, ViTs face significant efficiency challenges due to their quadratic complexity, which poses obstacles to further advancements in foundational models, particularly when dealing with longer sequences or larger image resolutions.

More recently, SSMs have been proposed as an alternative backbone architecture for vision tasks (Huang et al. 2024; Yang et al. 2024; Zhu et al. 2024a; Liu et al. 2024; Pei, Huang, and Xu 2024), offering linear time complexity. SSMs employ a novel approach by modeling data sequences through state transitions, which is particularly advantageous for capturing dynamic changes in images. This emerging field seeks to provide more efficient solutions compared to traditional CNNs and Transformers. A critical issue in designing SSM-based visual backbones is how to effectively serialize image patches, transitioning from 2D to 1D while preserving the structural information in the image. This ensures that the SSM can accurately capture the current feature representation of the image. In this paper, we highlight the problems with current serialization methods and introduce a fractal serialization mechanism to enhance the performance of SSM-based backbones.

State Space Model

Drawing on principles from control theory, linear state space equations have been integrated with deep learning to enhance sequential data modeling, as demonstrated in works like HiPPO (Gu et al. 2020) and LSSL (Gu et al. 2021). Recent advancements have enabled SSMs to increasingly compete with CNNs and Transformers in terms of performance. Notably, the Structured State Space Sequence Model (S4) (Gu, Goel, and Ré 2022) utilizes a linear state space for contextualization and has exhibited strong performance across various sequence modeling tasks, particularly with lengthy sequences. Subsequently, Mamba (Gu and Dao 2023) has achieved a significant breakthrough with its linear-time inference and efficient training process, incorporating a selection mechanism and hardware-aware algorithms. Building on the success of Mamba, subsequent studies have explored the potential of SSMs for CV tasks (Huang et al. 2024; Yang et al. 2024; Zhu et al. 2024a; Liu et al. 2024), achieving performance comparable to that of Transformers.

At the core of these endeavors is the development of various linear scanning mechanisms for serialization, as highlighted in recent surveys (Xu et al. 2024; Zhang et al. 2024). However, linear scanning curves have inherent limitations and often fail to fully capture the complex spatial relationships within images. Furthermore, these linear scans can lead to repetitive scanning cycles, creating serialized sequences that exhibit redundant patterns. This redundancy can introduce biases in the subsequent modeling by SSMs, as the model may overfit to these repetitive patterns rather than capturing the true underlying dynamics of the image. To overcome these deficiencies, in this paper, we propose the adoption of fractal scanning curves in vision Mamba, which offer a more advanced approach to image serialization. As shown in Fig. 3, following works (Liu et al. 2024; Huang et al. 2024), we use the S6 block in the FractalMamba block.

Method

Preliminaries

SSM. SSM provides a robust framework for modeling physical systems, particularly Linear Time-Invariant (LTI) systems. These models excel in representing such systems through a set of first-order differential equations, effectively capturing the dynamics of the system's state variables:

$$h'(t) = \mathbf{A}h(t) + \mathbf{B}x(t), y(t) = \mathbf{C}h(t), \tag{1}$$

where $h'(t)$ denotes time derivative of state vector $h(t)$, with matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}$ defining relationships between $h(t), x(t), y(t)$. This system uses $\mathbf{A} \in \mathbb{R}^{N\times N}$ as the evolution parameter and $\mathbf{B} \in \mathbb{R}^{N\times 1}, \mathbf{C} \in \mathbb{R}^{1\times N}$ as projection parameters.

Since SSMs operate on continuous sequences $x(t)$, they are unable to process discrete token inputs such as images and texts. Consequently, the adaptation to a discretized version of the SSM becomes essential. As described in Mamba (Gu and Dao 2023), the commonly used method for transformation $\mathbf{A}, \mathbf{B}$ from continuous to discrete form is zero-order hold (ZOH), which is defined as follows:

$$\begin{aligned} \overline{\mathbf{A}} &= \exp(\mathbf{\Delta A}), \\ \overline{\mathbf{B}} &= (\mathbf{\Delta A})^{-1}(\exp(\mathbf{\Delta A}) - \mathbf{I}) \cdot \mathbf{\Delta B}, \end{aligned} \tag{2}$$

where $\mathbf{\Delta}$ is the timescale parameter to transform the continuous parameters $\mathbf{A}, \mathbf{B}$ to discrete parameters $\overline{\mathbf{A}}, \overline{\mathbf{B}}$. After the operation, the discretized version of Eqn. 1 is rewritten as:

$$h_t = \overline{\mathbf{A}}h_{t-1} + \overline{\mathbf{B}}x_t, y_t = \mathbf{C}h_t. \tag{3}$$

Finally, the model computes output with global convolution:

$$\begin{aligned} \overline{\mathbf{K}} &= \left(\mathbf{C}\overline{\mathbf{B}}, \mathbf{C}\overline{\mathbf{A}\mathbf{B}}, \ldots, \mathbf{C}\overline{\mathbf{A}}^{L-1}\overline{\mathbf{B}}\right) \\ \mathbf{y} &= \mathbf{x} * \overline{\mathbf{K}}, \end{aligned} \tag{4}$$

where L is the length of the input $\mathbf{x}$, $\overline{\mathbf{K}} \in \mathbb{R}^L$ is a structured convolutional kernel and $*$ represents the convolution operation.

Selective SSMs. The inherent LTI characteristic of SSMs, characterized by the consistent application of matrices $\overline{\mathbf{A}}$,

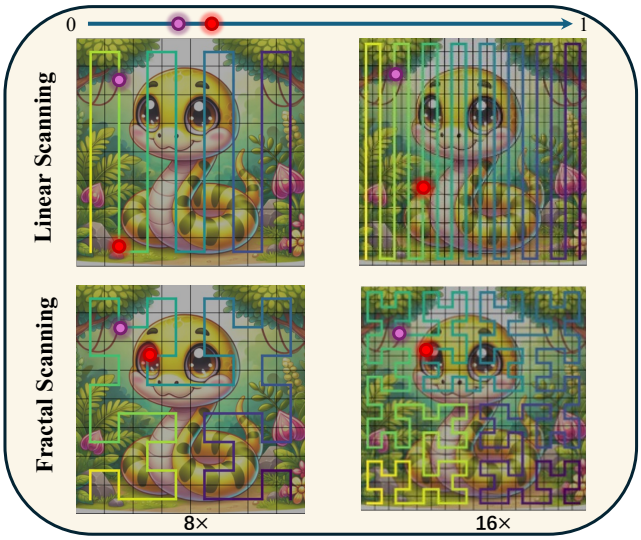

Figure 4: The distance between the two patches in linear scanning and fractal scaanning.

$\overline{\mathbf{B}}$, $\overline{\mathbf{C}}$, and Δ across various inputs, limits their ability to adaptively filter and interpret contextual nuances within diverse input sequences. To overcome this limitation, we use Selective SSM as the core operator in FractalMamba. In Selective SSMs, the matrices $\overline{\mathbf{B}}$, $\overline{\mathbf{C}}$, and Δ are rendered as dynamic, input-responsive elements, effectively transitioning the SSM into a time-variant model. This modification enables the model to adapt more effectively to different input contexts, significantly enhancing its capacity to capture pertinent temporal features and relationships, and thereby improving the accuracy and efficiency of the input representation.

Fractal Scanning Curve

Limitations of Linear Curves. From Eqn. 1, it becomes evident that selecting the most appropriate inputs for each time step in the SSM is crucial for effectively capturing and modeling the feature representation of the current image. To achieve this, the serialization process from 2D images to 1D sequences must meticulously capture the inherent structural information within the image. This requires preserving the structural coherence among image patches, ensuring that the serialized form maintains the essential spatial relationships present in the original image.

Existing linear scanning methods, such as the Z-order or Zigzag curve, exhibit significant limitations in preserving spatial relationships. While these methods maintain adjacency information between neighboring patches within the same row, they fail to effectively capture inter-row relationships. This oversight leads to the loss of crucial structural links essential for understanding broader spatial relationships within the image. Consequently, such scanning techniques compromise the structural integrity of the original image, undermining the model's capacity to accurately reflect the image's true characteristics. This lack of structural fidelity results in a model that is insensitive to variations in image scale, meaning that correlations evident at lower resolutions may not be preserved at higher resolutions. As a result, patterns and features learned at a lower resolution often do not translate effectively to higher resolutions, limiting the model's applicability across different scales. As shown in Fig. 4 (Linear), linear scanning methods traversal alter the relative distances between two image patches when the image is scaled. This change can affect the consistency and accuracy of feature extraction across different resolutions.

To overcome the above challenge, we introduce fractal scanning. This scanning is designed to more adeptly follow the complex structure within an image, thereby preserving its integral structural characteristics across different viewing scales. In this paper, we choose the Hilbert curve, which is a typical fractal curve. As shown in Fig. 4 (Fractal), the Hilbert curve can maintain consistent relative distances between two image patches regardless of the image scale.

Fractal Curves. The Hilbert curve, a fractal defined through a recursive process, is particularly effective in image analysis because it maintains spatial and structural consistency at varying scales. Its property of self-similarity is vital for the analysis of high-resolution images, enabling it to capture features coherently across different scales. Starting from a single point, the Hilbert curve uses the midpoints of directional vectors $\vec{x}$ and $\vec{y}$ to dictate its path within the two-dimensional space. Through successive recursive subdivisions and traversals, the curve systematically accesses every part of the image, ensuring local continuity. This method effectively preserves the spatial relationships within the image, which is essential for delivering precise analysis and representation.

Fractal Fusion Pathway

Limitation of Fractal Curves. Although we introduce fractal curves in SSMs and design the FractalMamba, which better capture the spatial structural relationships of images compared to linear curves, it's important to note that using a single curve still has limitations in capturing structural relationships. For example, as illustrated in Fig. 5, adjacent image patches like (3,3) and (4,4) on the diagonal also struggle to maintain their spatial relationships during the serialization process with fractal curves. A common approach is to introduce scanning curves from multiple different directions like works (Liu et al. 2024; Zhu et al. 2024a). However, this undoubtedly increases computational complexity. Enhancing performance by increasing complexity contradicts the original intent behind the development of mamba. To address this flaw, we design the FFP, which can more efficiently enhance the performance of FractalMamba. FFP enhances the efficiency of FractalMamba by effectively organizing and integrating information from various image patches on the fractal curve, allowing it to preserve structural information without significantly increasing computational complexity.

Firstly, we derive the state transition equations for SSMs, transforming the original unidirectional transition process into a bidirectional one, thus image token can gather state information coming from multiple direction simultaneously. Specifically, the original SSMs equation can be expressed as:

$$y = \mathbf{C} \sum_{i=1}^{m} x_i \overline{\mathbf{B}}_i \prod_{j=i+1}^{m} \overline{\mathbf{A}}_j, \tag{5}$$

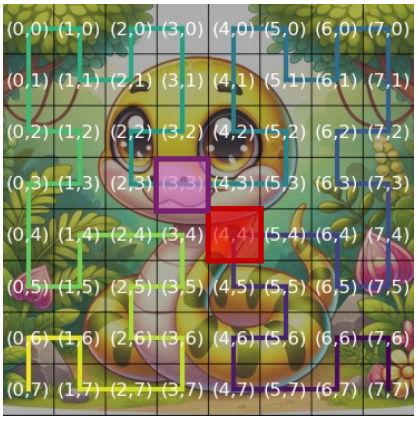

Figure 5: Limitation of Hilbert Curves.

where m is the index of the patch for which the output is being calculated. Then, our modified SSMs equation can be written as:

$$y = \mathbf{C}\sum_{i=1}^{m} x_i \overline{\mathbf{B}}_i \prod_{j=i+1}^{m} \overline{\mathbf{A}}_j + \underbrace{\mathbf{C}\sum_{i=m+1}^{n} x_i \overline{\mathbf{B}}_i \prod_{j=m}^{i-1} \overline{\mathbf{A}}_j}_{Supplement}, \quad (6)$$

where n is the total number of image patches. The supplement term enables the SSMs to perform state transitions from the opposite direction simultaneously. Observing Equation 6, we see that y depends not only on the information preceding m but also on the information following m. This dependency prevents Equation 6 from being efficiently computed using a recursive approach. To address this issue, we employ a memory mechanism to pre-store various states, enabling bidirectional transitions to be carried out recursively. Furthermore, to enable end-to-end training of the modified SSMs, we also derived the derivatives of the output with respect to the state parameters.

We first improve the capture of structural information by modifying the transition equations, and then further enhance the FractalMamba's capabilities through a shifting mechanism. As shown in Fig. 5, the two patches in the image, (3,3) and (4,4), are neighboring in terms of spatial location, but the distance between the two of them is not neighboring in the Hilbert scan. These issues can lead to a partial loss of local proximity information, which is crucial for accurate modeling by SSMs. To address this challenge, we apply a shifting operation, moving the curves up or down by one pixel. This minor modification significantly improves the local adjacency and continuity of pixel serialization. By aligning the curve more closely with the inherent spatial relationships within the image, this shifting helps to mitigate gaps and overlaps at critical junctions and complex parts of the curve, where adjacency might originally be lost.

Experiments

We conduct a series of experiments to evaluate and compare FractalMamba with different established benchmark models, including architectures based on CNNs, ViTs and SSMs. Our assessment covers a range of visual tasks with various resolution, such as image classification, object detection, remote sensing binary change object detection, and semantic segmentation. In these tasks, remote sensing binary change object detection involve high-resolution images, with resolutions of 1024×1024. Evaluating our model in high-resolution images further demonstrates its effectiveness. Following these evaluations, we provide a comprehensive analysis of FractalMamba's characteristics, with a particular emphasis on its standout feature: the remarkable capability to efficiently handle increasingly large input resolutions. All experiments are conducted using 8 NVIDIA H800 GPUs.

Models	Image size	#Param.	FLOPs	Top-1 Acc.
RegNetY-8G	224^2	39M	8.0G	81.7
EffNet-B5	456^2	30M	9.9G	83.6
ViT-B/16	384^2	86M	55.4G	77.9
DeiT-B	224^2	86M	17.5G	81.8
ConvNeXt-T	224^2	29M	4.5G	82.1
HiViT-S	224^2	38M	9.1G	83.5
Swin-T	224^2	28M	4.6G	81.3
ViM-S	224^2	26M	-	80.5
VMamba-T	224^2	31M	4.9G	82.5
LocalMamba-T	224^2	26M	5.7G	82.7
PlainMamba-L3	224^2	50M	14.4G	82.3
FractalMamba-T	224^2	31M	4.8G	83.0

Table 1: Performance comparison of our FractalMamba and other methods on the ImageNet-1K dataset.

Image Classification

Experiment Setting. We evaluate the performance of FractalMamba using the ImageNet-1K dataset (Deng et al. 2009), following the evaluation protocol described in the work (Liu et al. 2022a). The FractalMamba-T model is trained from scratch over 300 epochs, with an initial 20-epoch warm-up period, using a batch size of 1024. The training processes utilizes the AdamW optimizer (Loshchilov and Hutter 2017), with betas set to (0.9,0.999), momentum of 0.9, a cosine decay learning rate scheduler, an initial learning rate of 1×10^{-3}, and a weight decay of 0.05. Additional strategies such as label smoothing (0.1) and exponential moving average (EMA) were incorporated into the training regimen.

Model Performance. The comparison results of FractalMamba against benchmark backbone models on the ImageNet-1K dataset are summarized in Table. 1. Notably, with comparable FLOPs, FractalMamba-T achieves a performance of 83.0, outperforming RegNetY-8G (Radosavovic et al. 2020) by 1.3%, DeiT-B (Touvron et al. 2021) by 1.2%, and Swin-T by 1.7%. Moreover, FractalMamba can consistently outperform other SSM-based models, containing ViM-S (Zhu et al. 2024a), VMamba-T (Liu et al. 2024), LocalMamba-T (Huang et al. 2024) and PlainMamba-L3 (Yang et al. 2024). This also demonstrates that during serialization, fractal curves can better preserve the spatial structural relationships in images, lose fewer features, and thus enhance the model's performance.

Mask R-CNN 1 × schedule								
Backbone	AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}	#param.	FLOPs
Swin-T (Liu et al. 2022a)	42.7	65.2	46.8	39.3	62.2	42.2	48M	267G
ConvNeXt-T (Liu et al. 2022b)	44.2	66.6	48.3	40.1	63.3	42.8	48M	262G
ViT-Adapter-S (Dosovitskiy et al. 2021)	44.7	65.8	48.3	39.9	62.5	42.8	48M	403G
VMamba-T (Liu et al. 2024)	46.5	68.5	50.7	42.1	65.5	45.3	42M	286G
LocalMamba-T (Huang et al. 2024)	46.7	68.7	50.8	42.2	65.7	45.5	45M	291G
PlainMamba-L2 (Yang et al. 2024)	46.0	66.9	50.1	40.6	63.8	43.6	53M	542G
FractalMamba-T	46.8	68.7	50.8	42.4	65.9	45.8	41M	266G
Mask R-CNN 3 × MS schedule								
Swin-T (Liu et al. 2022a)	46.0	68.1	50.3	41.6	65.1	44.9	48M	267G
ConvNeXt-T (Liu et al. 2022b)	46.2	67.9	50.8	41.7	65.0	44.9	48M	262G
ViT-Adapter-S (Dosovitskiy et al. 2021)	48.2	69.7	52.5	42.8	66.4	45.9	48M	403G
VMamba-T (Liu et al. 2024)	48.5	69.9	52.9	43.2	66.8	46.3	48M	260G
LocalMamba-T (Huang et al. 2024)	48.7	70.1	53.0	43.4	67.0	46.4	45M	291G
FractalMamba-T	48.5	70.0	53.2	43.3	67.1	46.2	41M	266G

Table 2: The results of object detection and instance segmentation on the COCO dataset. FLOPs are calculated for an input size of 1280×800. The metrics AP^b and AP^m represent box AP and mask AP, respectively. The notation '1 ×' indicates that models were fine-tuned for 12 epochs, while '3 × MS' denotes the utilization of multi-scale training across 36 epochs.

Type	Models	IoU	Pre	Rec	KC	F1	#param.	FLOPs
$\mathcal{T}$	ChangeFormerV6 (Bandara and Patel 2022)	57.1	67.6	78.5	71.4	72.7	41M	811G
	BIT-101 (Chen, Qi, and Shi 2022)	70.3	83.9	81.2	81.8	82.5	43M	381G
	TransUNetCD (Li et al. 2022b)	71.9	83.1	84.2	82.9	83.6	28M	245G
	SwinSUNet (Zhang et al. 2022)	74.8	85.3	85.9	85.0	85.6	39M	44G
	CTDFormer (Li et al. 2022b)	67.1	80.6	80.0	79.5	80.3	40M	304G
$\mathcal{M}$	ChangeMamba-T (Chen et al. 2024a)	78.6	88.8	87.3	87.5	88.0	17M	46G
	ChangeMamba-S (Chen et al. 2024a)	78.3	89.2	86.5	87.3	87.8	50M	115G
	ChangeMamba-B (Chen et al. 2024a)	79.2	89.2	87.6	87.9	88.4	85M	179G
	FractalMamba-T	80.0	89.3	88.4	88.4	89.9	35M	55G

Table 3: Accuracy assessment and efficiency comparison for different binary CD models on the LEVIR-CD+ dataset. The evaluation resolution of this dataset is 1024×1024. $\mathcal{T}$ means Transofrmer-based methods. $\mathcal{M}$ means vision Mamba methods.

Object Detection

Experiment Setting. We evaluate the performance of FractalMamba on object detection using the MSCOCO 2017 (Lin et al. 2014). We configure our training setup using the MMDetection library (Chen et al. 2019), adhering to the hyperparameters utilized in the Swin (Liu et al. 2021) model with the Mask-RCNN detector (He et al. 2017). Specifically, we use the AdamW optimizer (Loshchilov and Hutter 2017) and fine-tune the pre-trained classification models (originally trained on ImageNet-1K) over both 12 and 36 epochs. For FractalMamba-T, the drop path rate is set at 0.2%. The learning rate starts at 1×10^{-4} and is reduced by a factor of 10 at the 9th and 11th epochs. Multi-scale training and random flipping are implemented with a batch size of 16, aligning with established best practices for object detection evaluations.

Model Performance. The results on the MSCOCO 2017 dataset are summarized in Table. 2. FractalMamba-T consistently outperforms other models in both box and mask AP, regardless of the training schedule employed. Specifically, with a 12-epoch fine-tuning schedule, FractalMamba-T models achieve object detection mean Average Precision (mAP) of 46.8, which is superior to Swin-T by 4.1% mAP, ConvNeXt-T by 2.6% mAP and VMamba-T by 0.3%. In the same configuration, FractalMamba-T achieves instance segmentation AP^m of 42.4, surpassing Swin-T by 3.1% AP^m and ConvNeXt-T by 2.3% AP^m. These results highlight FractalMamba's capability to deliver robust performance in downstream tasks requiring dense prediction. Moreover, compared to some other vision Mamba methods, our FractalMamba-T achieves higher performance with a more optimized computational complexity, which also preliminarily highlights the potential of FractalMamba in downstream applications.

Remote Sensing Binary Change Detection

Experiment Setting. Following ChangeMamba (Chen et al. 2024a), we augment the pre-trained model with the same Change-Decoder (Chen et al. 2024a) in BCD (Remote Sensing Binary Change Detection) task. We conduct experiments on the building CD dataset LEVIR-CD+, which is advanced version of the LEVIR-CD. It comprises 985 pairs of very

Models	Crop Size	SS mIoU	MS mIoU	#param.	FLOPs
ResNet-50	512^2	42.1	42.8	67M	953G
DeiT-S+MLN	512^2	43.8	45.1	58M	1217G
Swin-T	512^2	44.4	45.8	60M	945G
ConvNeXt-T	512^2	46.0	46.7	60M	939G
VMamba-T	512^2	47.3	48.3	55M	964G
LocalVim-S	512^2	46.4	47.5	58M	297G
PlainMamba-L2	512^2	46.8	-	55M	285G
FractalMamba-T	512^2	48.0	48.9	53M	942G

Table 4: Results of semantic segmentation on ADE20K using UperNet (Xiao et al. 2018). FLOPs are calculated with input size of 512×2048. 'SS' and 'MS' denote single-scale and multi-scale testing, respectively.

high-resolution images at 0.5 meters/pixel, each with dimensions of 1024×1024 pixels. During training, we optimize the network using the AdamW optimizer with a learning rate of 0.0001 and a weight decay of 0.005. The batch size is set to 16. In evaluating the BCD performance of the model, we use the five commonly used metrics. They are intersection over union (IoU), precision rate (Pre), recall rate (Rec), Kappa coefficient (KC) and F1 score (F1). The higher the better for all five indicators.

Model Performance. Table.3 lists the performance of FractalMamba-T and comparison methods in BCD. It can be seen that our FractalMamba-T significantly outperforms both the Transformer-based architectures and Mamba-based approaches. Specifically, our FractalMamba-T, with less computational complexity, outperforms ChangeMamba-B across all five metrics, fully demonstrating the potential of the FractalMamba architecture for the remote sensing task. This also indicates that the use of fractal curves makes SSMs more suitable for high-resolution images.

Semantic Segmentation

Experiment Setting. Following Swin (Liu et al. 2021), we augment the pre-trained model with an UperHead (Xiao et al. 2018). We employ the AdamW optimizer (Loshchilov and Hutter 2017), setting the learning rate to 6×10^{-5}. The fine-tuning process extends over 160,000 iterations with a batch size of 16. The standard input resolution is 512×512, and we additionally provide experimental results using 640×640 inputs along with multi-scale (MS) and single-scale (SS) testing to evaluate performance enhancements at varied resolutions.

Model Performance. The results of semantic segmentation on the ADE20K dataset are summarized in Table. 4. In line with findings from previous experiments, FractalMamba exhibits superior accuracy. Specifically, FractalMamba-T achieves a mean Intersection over Union (mIoU) of 48.0 with a resolution of 512×512, and 48.9 mIoU with multiscale (MS) input. These results surpass those of all benchmarked methods, including ResNet (He et al. 2016), DeiT (Touvron et al. 2021), Swin (Liu et al. 2021), and ConvNeXt (Liu et al. 2022b), confirming FractalMamba's effectiveness.

Models	Curve type	Curve num	224^2	384^2	512^2	640^2	768^2	1024^2
Vim	Linear	2	76.1	70.4	67.4	51.4	30.6	16.1
Vim	Fractal	2	77.9	75.1	73.8	62.3	51.3	39.7
VMamba	Linear	4	82.5	82.5	81.1	79.3	76.1	62.3
VMamba	Fractal	2	82.7	82.8	82.0	80.6	78.1	72.3
VMamba	Fractal	4	82.9	83.3	82.5	81.6	79.7	74.9
Fractal Mamba (FFP w/o shifting)	Fractal	1	82.8	83.0	82.0	81.1	79.6	74.7
Fractal Mamba (FFP w/ shifting)	Fractal	2	83.0	83.9	83.0	81.8	80.3	76.3

Table 5: Ablation study of our proposed FractalMamba. w/o means without operation.

Ablation Studies

We conduct ablation studies of our FractalMamba across various resolutions on the ImageNet-1K dataset. Herein, FractalMamba differs from VMamba only in the scanning curves; all other modules are identical to those in VMamba.

The scalability of Fractal Curves. We note that fractal curves can help vision Mamba adapt more effectively to images of varying scales. As shown in Table. 5, simply applying the fractal curve to Vmamba- and Vim, resulting in performance improvements across different input resolutions while maintaining parameter numbers.

The effectiveness of FFP. In Table.5, in the FFP, when we use only the transforming operation without adding the shifting operation, FractalMamba requires only one curve. With this reduced computational complexity, it can surpass models that include two scanning curves and achieve competitive performance with models that contain four curves. Furthermore, reintroducing the shifting mechanism into the FFP led to performance improvements across all resolutions. This also demonstrates that our shifting operation can better assist fractal curves in capturing spatial structural relationships.

Conclusion

In this paper, we introduce a novel method for serializing image patches using fractal scanning curves to enhance the performance of SSMs in various CV tasks. Unlike traditional linear scanning curves, fractal curves exhibit superior by maintaining high spatial proximity and adapting seamlessly to different image resolutions. This approach not only reduces redundancy but also more accurately captures complex patterns within images. We validate our method across a range of computer vision tasks, including image classification, detection, and segmentation. The experimental results unequivocally demonstrate that fractal curve scanning significantly outperforms linear curve scanning in these applications. These findings underscore the practicality of fractal curves in vision tasks and pave the way for future research, such as exploring additional fractal scanning methods to further enhance model performance. We believe that with continued refinement, fractal curve usage in SSMs will become increasingly pivotal in future CV applications, particularly in processing high-resolution and large-scale image data.

References

Bandara, W. G. C.; and Patel, V. M. 2022. A Transformer-Based Siamese Network for Change Detection. In *IGARSS*, 207–210. IEEE.

Chen, H.; Qi, Z.; and Shi, Z. 2022. Remote Sensing Image Change Detection With Transformers. *IEEE Trans. Geosci. Remote. Sens.*, 60: 1–14.

Chen, H.; Song, J.; Han, C.; Xia, J.; and Yokoya, N. 2024a. ChangeMamba: Remote Sensing Change Detection with Spatio-Temporal State Space Model. arXiv 2024. *arXiv preprint arXiv:2404.03425*.

Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Xu, J.; et al. 2019. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*.

Chen, Z.; Wang, W.; Tian, H.; Ye, S.; Gao, Z.; Cui, E.; Tong, W.; Hu, K.; Luo, J.; Ma, Z.; et al. 2024b. How Far Are We to GPT-4V? Closing the Gap to Commercial Multimodal Models with Open-Source Suites. *arXiv preprint arXiv:2404.16821*.

Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H. W.; Sutton, C.; Gehrmann, S.; Schuh, P.; Shi, K.; Tsvyashchenko, S.; Maynez, J.; Rao, A.; Barnes, P.; Tay, Y.; Shazeer, N.; Prabhakaran, V.; Reif, E.; Du, N.; Hutchinson, B.; Pope, R.; Bradbury, J.; Austin, J.; Isard, M.; Gur-Ari, G.; Yin, P.; Duke, T.; Levskaya, A.; Ghemawat, S.; Dev, S.; Michalewski, H.; Garcia, X.; Misra, V.; Robinson, K.; Fedus, L.; Zhou, D.; Ippolito, D.; Luan, D.; Lim, H.; Zoph, B.; Spiridonov, A.; Sepassi, R.; Dohan, D.; Agrawal, S.; Omernick, M.; Dai, A. M.; Pillai, T. S.; Pellat, M.; Lewkowycz, A.; Moreira, E.; Child, R.; Polozov, O.; Lee, K.; Zhou, Z.; Wang, X.; Saeta, B.; Diaz, M.; Firat, O.; Catasta, M.; Wei, J.; Meier-Hellstern, K.; Eck, D.; Dean, J.; Petrov, S.; and Fiedel, N. 2023. PaLM: Scaling Language Modeling with Pathways. *J. Mach. Learn. Res.*, 24: 240:1–240:113.

Dao, T.; and Gu, A. 2024. Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality. *CoRR*, abs/2405.21060.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, 248–255. IEEE.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*, 4171–4186.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*. OpenReview.net.

Gotsman, C.; and Lindenbaum, M. 1996. On the metric properties of discrete space-filling curves. *IEEE Trans. Image Process.*, 5(5): 794–797.

Gu, A.; and Dao, T. 2023. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *CoRR*, abs/2312.00752.

Gu, A.; Dao, T.; Ermon, S.; Rudra, A.; and Ré, C. 2020. HiPPO: Recurrent Memory with Optimal Polynomial Projections. In *NeurIPS*.

Gu, A.; Goel, K.; and Ré, C. 2022. Efficiently Modeling Long Sequences with Structured State Spaces. In *ICLR*. OpenReview.net.

Gu, A.; Johnson, I.; Goel, K.; Saab, K.; Dao, T.; Rudra, A.; and Ré, C. 2021. Combining Recurrent, Convolutional, and Continuous-time Models with Linear State Space Layers. In *NeurIPS*, 572–585.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. B. 2017. Mask R-CNN. In *ICCV*, 2980–2988. IEEE Computer Society.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *CVPR*, 770–778. IEEE.

Howard, A. G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; and Adam, H. 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *CoRR*, abs/1704.04861.

Huang, T.; Pei, X.; You, S.; Wang, F.; Qian, C.; and Xu, C. 2024. LocalMamba: Visual State Space Model with Windowed Selective Scan. *CoRR*, abs/2403.09338.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollar, P.; and Girshick, R. 2023. Segment Anything. In *ICCV*, 4015–4026.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *NIPS*, 1106–1114.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. C. H. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *ICML*, volume 202, 19730–19742.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. C. H. 2022a. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, 12888–12900. PMLR.

Li, Q.; Zhong, R.; Du, X.; and Du, Y. 2022b. TransUNetCD: A Hybrid Transformer Network for Change Detection in Optical Remote-Sensing Images. *IEEE Trans. Geosci. Remote. Sens.*, 60: 1–19.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*, 740–755. Springer.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. *CoRR*, abs/2304.08485.

Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; and Liu, Y. 2024. VMamba: Visual State Space Model. *CoRR*, abs/2401.10166.

Liu, Z.; Hu, H.; Lin, Y.; Yao, Z.; Xie, Z.; Wei, Y.; Ning, J.; Cao, Y.; Zhang, Z.; Dong, L.; et al. 2022a. Swin transformer v2: Scaling up capacity and resolution. In *CVPR*, 12009–12019.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *ICCV*, 9992–10002. IEEE.

Liu, Z.; Mao, H.; Wu, C.; Feichtenhofer, C.; Darrell, T.; and Xie, S. 2022b. A ConvNet for the 2020s. In *CVPR*, 11966–11976. IEEE.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

OpenAI. 2023. GPT-4 Technical Report. *CoRR*, abs/2303.08774.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; Assran, M.; Ballas, N.; Galuba, W.; Howes, R.; Huang, P.; Li, S.; Misra, I.; Rabbat, M. G.; Sharma, V.; Synnaeve, G.; Xu, H.; Jégou, H.; Mairal, J.; Labatut, P.; Joulin, A.; and Bojanowski, P. 2023. DINOv2: Learning Robust Visual Features without Supervision. *CoRR*, abs/2304.07193.

Pei, X.; Huang, T.; and Xu, C. 2024. EfficientVMamba: Atrous Selective Scan for Light Weight Visual Mamba. *CoRR*, abs/2403.09977.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, volume 139, 8748–8763. PMLR.

Radosavovic, I.; Kosaraju, R. P.; Girshick, R. B.; He, K.; and Dollár, P. 2020. Designing Network Design Spaces. In *CVPR*, 10425–10433. IEEE.

Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; Mintun, E.; Pan, J.; Alwala, K. V.; Carion, N.; Wu, C.-Y.; Girshick, R.; Dollár, P.; and Feichtenhofer, C. 2024. SAM 2: Segment Anything in Images and Videos. *arXiv preprint arXiv:2408.00714*.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *ICML*, volume 139, 10347–10357. PMLR.

Touvron, H.; Cord, M.; and Jégou, H. 2022. DeiT III: Revenge of the ViT. In *ECCV*, volume 13684, 516–533. Springer.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. *CoRR*, abs/2302.13971.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *NeurIPS*, 5998–6008.

Xiao, T.; Liu, Y.; Zhou, B.; Jiang, Y.; and Sun, J. 2018. Unified perceptual parsing for scene understanding. In *ECCV*, 418–434.

Xu, R.; Yang, S.; Wang, Y.; Du, B.; and Chen, H. 2024. A survey on vision mamba: Models, applications and challenges. *arXiv preprint arXiv:2404.18861*.

Yang, C.; Chen, Z.; Espinosa, M.; Ericsson, L.; Wang, Z.; Liu, J.; and Crowley, E. J. 2024. PlainMamba: Improving Non-Hierarchical Mamba in Visual Recognition. *CoRR*, abs/2403.17695.

Zhang, C.; Wang, L.; Cheng, S.; and Li, Y. 2022. SwinSUNet: Pure Transformer Network for Remote Sensing Image Change Detection. *IEEE Trans. Geosci. Remote. Sens.*, 60: 1–13.

Zhang, H.; Zhu, Y.; Wang, D.; Zhang, L.; Chen, T.; and Ye, Z. 2024. A Survey on Visual Mamba. *arXiv preprint arXiv:2404.15956*.

Zheng, T.; Jiang, P.; Wan, B.; Zhang, H.; Chen, J.; Wang, J.; and Li, B. 2024. Beta-Tuned Timestep Diffusion Model. In *ECCV (3)*, volume 15061 of *Lecture Notes in Computer Science*, 114–130. Springer.

Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; and Wang, X. 2024a. Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model. *CoRR*, abs/2401.09417.

Zhu, Q.; Fang, Y.; Cai, Y.; Chen, C.; and Fan, L. 2024b. Rethinking Scanning Strategies with Vision Mamba in Semantic Segmentation of Remote Sensing Imagery: An Experimental Study. *CoRR*, abs/2405.08493.