Enabling True Global Perception in State Space Models for Visual Tasks

Jie Hui , Zhenxiang Zhang , Wenyu Mi , Jianji Wang

西安交通大學 XI'AN JIAOTONG UNIVERSITY

IAIR Est. 1986 Institute of Artificial Intelligence and Robotics

ICLR 2026

Enabling True Global Perception in State Space Models for Visual Tasks

Jie Hui[1], *Zhenxiang Zhang*[1], *Wenyu Mi*[1], *Jianji Wang*[1,*]

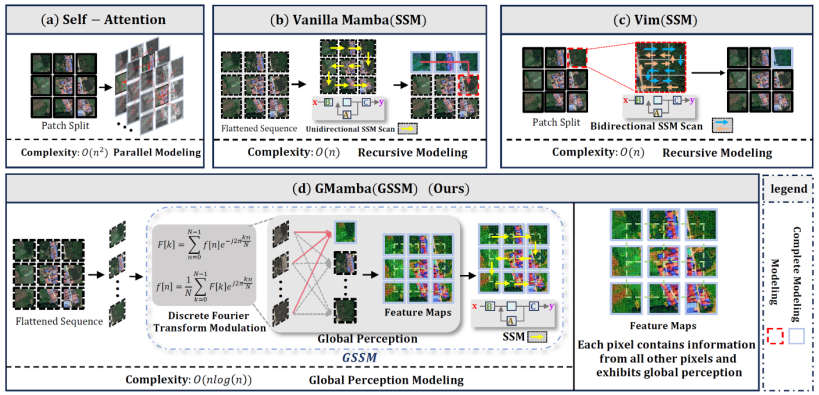
[1] State Key Laboratory of Human-Machine Hybrid Augmented Intelligence, Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University

Institute of Artificial Intelligence and Robotics

Motivation

Global modeling is crucial for capturing long-range dependencies in visual recognition and semantic understanding.

- Existing global modeling approaches (Transformer-based and Mamba-based) are mainly evaluated through empirical evidence (e.g., ablations or visualizations), while a rigorous definition of true global perception is still missing.
- Transformer-based methods enable explicit global interactions, but their quadratic computational complexity limits efficiency in high-resolution vision tasks.
- Mamba/SSM-based methods achieve linear complexity, yet rely on sequential state propagation and causal assumptions, leading to localized perception and structural mismatch with non-causal image data. Although sophisticated scanning strategies can partially alleviate this issue, they do not fundamentally overcome the intrinsic limitations of SSM-based modeling and often introduce additional complexity.



Main ideas

- First formal definition of global image modeling, providing a rigorous mathematical foundation for analyzing and guaranteeing true global perception in visual tasks.
- Theoretical framework for frequency-modulated SSM, connecting frequency domain analysis with state-space modeling to guide global feature propagation.
- Design of a plug-and-play GMamba (GSSM) module that injects frequency-guided global semantics into SSM, adaptable to diverse CNN backbones and multiple vision tasks while maintaining computational efficiency.

Mathematical Definition of Global Image Modeling

Definition: For an image modeling function $f : \mathbb{R}^{H\times W\times C} \to \mathbb{R}^{H\times W\times C}$ which is differentiable almost everywhere in $\mathcal{D}_f$, and a global influence function $\mathcal{I} : \{1,\dots,H\}\times\{1,\dots,W\}\times\{1,\dots,C\} \to \mathbb{R}^+$ whose derivative exists at all points in $\mathcal{D}_f$, if the Frobenius norm of the derivative of f with respect to any input pixel is greater than $\mathcal{I}$, then f exhibits global gradient dependency.

$$\left\|\frac{\partial f(\mathbf{X})}{\partial X_{i,j,c}}\right\|_F \geq \mathcal{I}(i,j,c) > 0, \quad \forall (i,j,c) \in \mathcal{D}_f; \quad \inf_{(i,j,c)} \mathcal{I}(i,j,c) \geq \tau > 0, \tag{1}$$

where $X_{i,j,c}$ represents the pixel value at position (i,j) and channel c in the input image $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$, $\|\cdot\|_F$ denotes the Frobenius norm of the gradient tensor which measures the degree of influence of input pixels at different positions on the output image, $\mathcal{D}_f$ denotes the set of points where both f and $\mathcal{I}$ are differentiable, inf represents the infimum, and τ ensures that each input pixel exerts a stable and non-negligible influence on the reconstructed output.

Constraint: For function f, due to the absence of inherent sequential regularity in image features, f should not impose strict order-dependent constraints on the input.

Frequency-Domain Modulated SSM Framework

Discrete-time SSM formulation: Continuous-time SSMs are discretized using a step size Δ and the zero-order hold method:

$$x_t = \bar{A}(\Delta)x_{t-1} + \bar{B}(\Delta)u_t, \quad y_t = Cx_t,$$

enabling efficient sequence modeling in neural networks.

SSM as dynamic convolution: The output can be equivalently expressed as

$$y_t = \sum_{k=0}^{t} K_k u_{t-k}, \quad K_k = C\bar{A}^k\bar{B},$$

showing that SSM performs step-by-step filtering on input sequences.

Frequency-domain equivalence: The discrete-time transfer function of SSM is

$$H(\omega) = C\left(e^{j\omega}I - \bar{A}\right)^{-1}\bar{B},$$

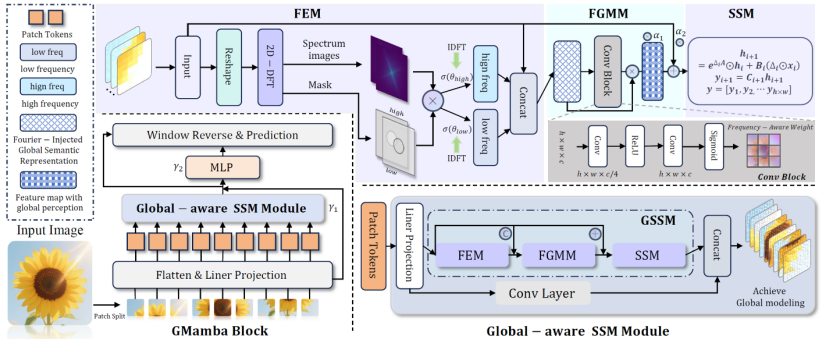
which is mathematically equivalent to the discrete-time Fourier transform of the convolution kernel K_k, linking time-domain convolution with frequency-domain representation.

$$\hat{K}(\omega) = \sum_{k=0}^{\infty} K_k e^{-j\omega k} = C\left(I - \bar{A}e^{-j\omega}\right)^{-1}\bar{B} = e^{j\omega}C\left(e^{j\omega}I - \bar{A}\right)^{-1}\bar{B} = e^{j\omega}H(\omega).$$

Theoretical Foundation of Frequency-Domain Modulated SSM.

- **Information Preservation:** The encoded hidden state information x_t in SSMs can be transformed into a convolutional form K_k, where frequency-transformed information of K_k directly maps with original information, ensuring frequency domain operations preserve the representational capacity of the original model.
- **Bidirectional Convertibility:** The dynamic convolution kernel K_k of SSM and its frequency domain transfer function $H(\omega)$ form a Fourier transform pair, enabling lossless bidirectional conversion between time-domain and frequency-domain representations.
- **Linear Information Propagation:** Both the discrete convolution $y_t = \sum_{k=0}^{t} K_k u_{t-k}$ and the frequency-domain multiplication $\hat{Y}(\omega) = H(\omega)\cdot\hat{U}(\omega)$ are linear relations, ensuring consistent and stable linear information propagation through the network in either domain.

GMamba (GSSM) Block



Experiments

Experiments are conducted on five datasets across three backbones and vision tasks to validate GMamba's effectiveness. For semantic segmentation, four remote sensing datasets are used: Vaihingen, Potsdam, LoveDA, and UAVid. For instance segmentation and object detection, the MS-COCO dataset is employed. Results demonstrate GMamba's strong generalization and consistent improvements across different tasks and architectures.

Model Variant	Params (M)	FLOPs (G)	mIoU (%)	mF1 (%)	OA (%)
UNet-ResNet34 (Baseline)	25.33	30.32	81.65	89.24	91.86
+ Swin (×7)	35.81	43.45	83.24 (↑1.59)	90.63 (↑1.39)	93.08 (↑1.22)
+ SwinV2 (×7)	35.86	43.46	83.10 (↑1.45)	90.54 (↑1.30)	93.04 (↑1.18)
+ ViM (×7)	28.20	35.81	83.02 (↑1.37)	90.51 (↑1.27)	92.93 (↑1.07)
+ VMamba (×7)	32.45	38.49	83.24 (↑1.59)	90.62 (↑1.38)	93.04 (↑1.18)
+ TinyViM (×7)	31.19	36.51	83.17 (↑1.52)	90.59 (↑1.35)	92.99 (↑1.13)
+ Mamba Version (×7)	34.87	39.78	83.20 (↑1.55)	90.62 (↑1.38)	93.00 (↑1.14)
+ Spatial Mamba (×7)	30.92	35.72	82.90 (↑1.25)	90.40 (↑1.16)	92.85 (↑0.99)
+ FreqMamba (×7)	31.98	36.21	83.00 (↑1.35)	90.50 (↑1.26)	92.90 (↑1.04)
+ Group Mamba (×7)	29.72	35.15	82.99 (↑1.34)	90.41 (↑1.17)	92.86 (↑1.00)
+ GMamba (×7) (Ours)	30.96	36.30	**84.74** (↑1.09)	**91.56** (↑1.32)	**93.72** (↑1.86)
UNet-SwinT (Baseline)	36.48	44.46	82.44	89.75	92.18
+ Swin (×7)	60.01	73.27	83.70 (↑1.26)	90.92 (↑1.17)	93.12 (↑0.94)
+ SwinV2 (×7)	60.07	73.28	83.73 (↑1.29)	90.94 (↑1.19)	93.19 (↑1.01)
+ ViM (×7)	42.83	57.99	83.62 (↑1.18)	90.88 (↑1.13)	93.05 (↑0.87)
+ VMamba (×7)	52.06	61.99	84.07 (↑1.63)	91.14 (↑1.39)	93.32 (↑1.14)
+ TinyViM (×7)	49.38	58.01	84.04 (↑1.60)	91.13 (↑1.38)	93.28 (↑1.10)
+ Mamba Version (×7)	52.47	62.95	83.68 (↑1.24)	90.90 (↑1.15)	93.10 (↑0.92)
+ Spatial Mamba (×7)	48.93	56.88	83.40 (↑0.96)	90.70 (↑0.95)	92.92 (↑0.74)
+ FreqMamba (×7)	49.95	57.98	83.55 (↑1.11)	90.82 (↑1.07)	93.00 (↑0.82)
+ Group Mamba (×7)	46.31	55.15	83.51 (↑1.07)	90.80 (↑1.05)	92.97 (↑0.79)
+ GMamba (×7) (Ours)	49.13	57.81	**84.83** (↑2.39)	**91.61** (↑1.86)	**93.65** (↑1.47)
UNet-ConvNeXt(S) (Baseline)	58.42	68.88	83.11	90.19	92.30
+ Swin (×7)	81.95	97.69	84.82 (↑1.71)	91.59 (↑1.40)	93.57 (↑1.27)
+ SwinV2 (×6)	81.77	93.61	84.36 (↑1.25)	91.31 (↑1.12)	93.32 (↑1.02)
+ ViM (×7)	64.76	78.09	84.24 (↑1.13)	91.22 (↑1.03)	93.37 (↑1.07)
+ VMamba (×7)	73.99	86.41	84.56 (↑1.45)	91.45 (↑1.26)	93.56 (↑1.26)
+ TinyViM (×7)	71.31	82.43	84.38 (↑1.27)	91.33 (↑1.14)	93.41 (↑1.11)
+ Mamba Version (×7)	74.87	88.47	84.80 (↑1.69)	91.55 (↑1.36)	93.55 (↑1.25)
+ Spatial Mamba (×7)	70.93	84.92	84.50 (↑1.39)	91.32 (↑1.13)	93.35 (↑1.05)
+ FreqMamba (×7)	71.95	85.98	84.60 (↑1.49)	91.42 (↑1.23)	93.42 (↑1.12)
+ Group Mamba (×7)	68.20	79.60	84.56 (↑1.45)	91.44 (↑1.25)	93.38 (↑1.08)
+ GMamba (×7) (Ours)	71.06	85.66	**86.00** (↑2.89)	**92.31** (↑2.12)	**93.99** (↑1.69)

Semantic Segmentation and Object Detection Results

Model Variant	AP (%)	AP_{50} (%)	AP_{75} (%)	AP_S (%)	AP_M (%)	AP_L (%)	Params (M)	FLOPs (G)
Faster R-CNN								
Baseline (ResNet50)	37.2	57.8	40.4	21.5	40.6	48.0	43.80(23.5)	207.07(76.50)
+ Swin (×3)	38.0	59.1	41.0	**23.0**	41.5	48.4	76.70(56.4)	210.57(80.00)
+ SwinV2 (×3)	37.7	58.7	40.9	21.5	41.3	48.8	76.90(56.6)	210.37(79.80)
+ TinyViM (×3)	37.6	58.6	40.6	21.6	41.3	48.4	61.40(41.1)	208.87(78.30)
+ VMamba (×3)	37.6	58.8	40.8	21.5	41.4	48.8	65.00(44.7)	209.67(79.10)
+ Mamba Version (×3)	37.7	58.9	41.0	21.6	41.4	48.5	65.40(45.1)	209.97(79.40)
+ Spatial Mamba (×3)	37.5	58.6	40.8	21.4	41.2	48.3	64.30(44.0)	209.27(78.70)
+ FreqMamba (×3)	37.6	58.7	40.9	21.5	41.3	48.4	64.90(44.6)	209.57(79.00)
+ Group Mamba (×3)	37.5	58.5	40.5	21.3	41.0	48.5	64.80(44.5)	209.73(78.80)
+ GMamba (×3) (Ours)	**38.5**	**59.6**	**42.2**	22.1	**42.0**	**49.9**	61.40(41.1)	210.22(79.65)
Baseline (Swin (T))	41.6	64.0	45.2	25.7	44.8	55.4	45.15(27.5)	213.00(86.10)
+ GMamba (×3) (Ours)	**42.9**	**65.6**	**46.9**	**27.1**	**46.4**	**56.4**	55.25(37.6)	216.15(89.25)
Mask R-CNN								
Baseline (ResNet50)	37.5	58.1	41.3	21.2	40.9	48.3	46.45(23.5)	260.14(76.50)
+ Swin (×3)	38.4	59.2	41.7	21.6	42.2	49.3	79.35(56.4)	263.64(80.00)
+ SwinV2 (×3)	38.3	59.1	41.7	22.2	41.9	49.6	79.55(56.6)	263.44(79.80)
+ TinyViM (×3)	38.2	58.8	41.6	21.5	41.6	49.9	64.05(41.1)	261.94(78.30)
+ VMamba (×3)	38.4	59.1	41.9	22.3	41.8	49.8	67.65(44.7)	262.74(79.10)
+ Mamba Version (×3)	38.2	59.8	42.2	22.3	42.0	49.2	68.05(45.1)	263.04(79.40)
+ Spatial Mamba (×3)	38.0	59.4	42.0	22.0	41.8	49.0	66.95(44.0)	262.34(78.70)
+ FreqMamba (×3)	38.1	59.6	42.1	22.1	41.9	49.1	67.55(44.6)	262.64(79.00)
+ Group Mamba (×3)	38.3	59.0	41.8	22.6	41.4	50.2	67.45(44.5)	262.44(78.80)
+ GMamba (×3) (Ours)	**39.1**	**60.1**	**42.8**	**23.3**	**42.5**	**50.3**	64.05(41.1)	263.29(79.65)
Baseline (SwinT))	42.2	64.7	46.0	26.7	45.5	55.6	47.79(27.5)	266.00(86.10)
+ GMamba (×3) (Ours)	**43.7**	**66.4**	**48.0**	**27.5**	**47.3**	**57.0**	57.88(37.6)	269.15(89.25)

Receptive field