RESPONES-TO-REVIEWER-RPRR

Anonymous authors
Paper under double-blind review

Dear Reviewer rPrr, we sincerely appreciate your valuable feedback on our submission. Below, we provide our responses to the concerns you raised. We have incorporated these revisions into the updated version of the manuscript, and we believe they will help improve the overall quality of the submission.

WEAKNESSES:

Comment 1

Insufficient analysis of scaling to ultra-high-resolution images (e.g., 4K+) and inference speed (FPS).

We appreciate your attention to the scalability of our module under ultra-high-resolution settings, which we also consider crucial. Due to the scarcity of 4K datasets, we were unable to conduct experiments at that resolution. Nevertheless, we performed alternative experiments and included them in **Appendix K.6** (Table 1), which presents an efficiency analysis of GSSM.

Using a UNet-ConvNeXt(S) backbone on the Vaihingen dataset with an input resolution of 1024×1024 on a single RTX 4090 GPU, the results demonstrate that GSSM achieves an optimal trade-off between accuracy and efficiency. Compared to high-cost self-attention mechanisms (latency 68.9 ms, memory 12.78 GB) and the slightly less accurate VSSM (84.01% mIoU), GSSM attains the highest performance (86.00% mIoU) while maintaining low latency (58.1 ms) and moderate memory usage (10.13 GB).

Furthermore, we compared inference latency and peak GPU memory of different global modeling methods across various resolutions, as shown in Fig. 1. As the resolution increases, both latency and memory consumption of GSSM grow approximately linearly, further demonstrating its superior scalability in high-resolution scenarios. We hope these experiments adequately address your concerns.

Table 1: Ablation of GMamba with Different Global Modeling Modules

Global Module	Latency (ms / image)	GPU Memory (GB)	mIoU (%)
Self-Attention (Transformer)	68.9	12.78	84.78
VSSM (Mamba)	51.6	8.71	84.01
GSSM (GMamba, Ours)	58.1	10.13	86.00

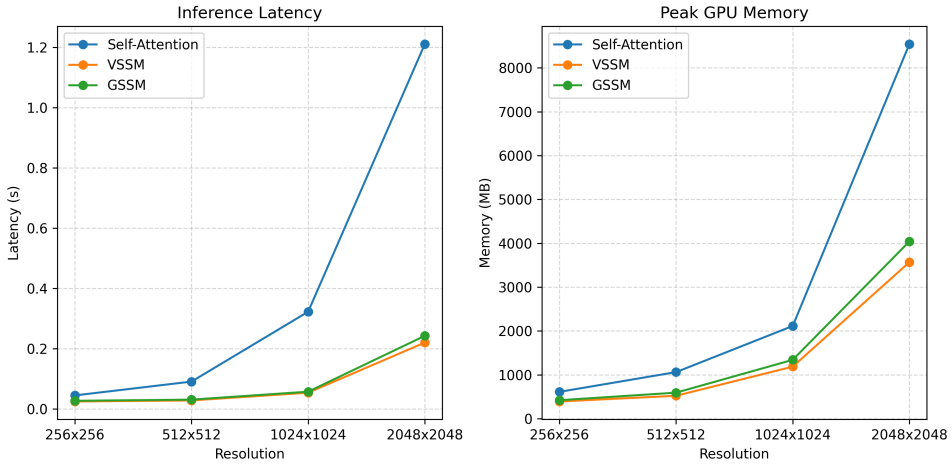

Figure 1: Performance comparison of three global modeling modules under different input resolutions.

Comment 2

Lack of explicit comparison with recent frequency-domain SSM variants (e.g., FreqMamba [1]).
[1] Freqmamba: Viewing mamba from a frequency perspective for image deraining

We sincerely apologize for the lack of comparisons with recent frequency-domain SSM variants. In the revised manuscript, we have added comparative analysis with FreqMamba[1], and, based on a thorough literature survey, we also include experimental comparisons with Mamba Version[2], Spatial Mamba[3], and Group Mamba[4].

As shown in the updated Tables 1 (semantic segmentation) and 3 (object detection) in the revised manuscript, GMamba consistently outperforms FreqMamba. For instance, on the Vaihingen dataset using ResNet34 as the backbone, GMamba achieves an mIoU of 84.74%, surpassing FreqMamba's 83.00%. FreqMamba employs a "triple-interaction structure," consisting of parallel spatial Mamba, frequency-domain Mamba, and Fourier global branches. This frequency integration approach has been further discussed in the Related Work section. Conceptually, this parallel structure corresponds to a "post-fusion" strategy: each branch processes sequences independently and without interaction, unable to leverage global frequency information during intermediate stages, and only performing feature fusion at the final step.

In contrast, our GSSM adopts a frequency pre-modulation mechanism: global frequency information is injected prior to the SSM state transition, enabling the global context to directly guide the state selection process and effectively overcoming the limitations of recursive models. This mechanism provides a more fundamental and direct solution than simple feature fusion, validating the effectiveness of our proposed theoretical framework for frequency-modulated SSMs.

Moreover, in **Appendix K.7** (Table 2), we added new experiments comparing different frequency integration strategies. We compared the "pre-modulation" strategy with the "dual-branch fusion" strategy used in existing frequency-based methods. The results show that our pre-modulation approach achieves an mIoU of 86.00%, significantly outperforming conventional methods (84.10% mIoU).

[1] Zhen Zou, Hu Yu, Jie Huang, and Feng Zhao. *Freqmamba: Viewing mamba from a frequency perspective for image deraining*. In Proceedings of the 32nd ACM International Conference on Multimedia, pp. 1905–1914, 2024.

[2] Ali Hatamizadeh and Jan Kautz. *Mambavision: A hybrid mamba-transformer vision backbone*. In Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 25261–25270, 2025.

[3] Chaodong Xiao, Minghan Li, Zhengqiang Zhang, Deyu Meng, and Lei Zhang. *Spatial-mamba: Effective visual state space models via structure-aware state fusion*. In The Thirteenth International Conference on Learning Representations, 2025.

[4] Abdelrahman Shaker, Syed Talal Wasim, Salman Khan, Juergen Gall, and Fahad Shahbaz Khan. *Groupmamba: Efficient group-based visual state space model*. In Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 14912–14922, 2025.

Table 2: Comparison of Frequency Enhancement Strategies in GMamba

Method	Frequency Strategy	Position	Mechanism	mIoU (%)
Baseline (GMamba)	None	None	None	83.11
GMamba w/ Dual-Branch Fusion	Dual-Branch Fusion	After SSM	Post-hoc feature fusion	84.10
GMamba w/ Pre-Modulation	Frequency Pre-Modulation	Before SSM	Frequency-guided state update	**86.00**

Comment 3

No analysis of failure cases.

We appreciate your attention to the failure case analysis. We have added a dedicated failure case analysis in **Appendix L** of the revised manuscript.

Appendix L Failure Case Analysis

We present some failure cases in Fig. 2. In the first three images, the scenes contain dense crowds. Our model fails to accurately identify individuals in the back rows, whose targets occupy only a few pixels and are partially occluded. In the last image, the scene shows food items on supermarket shelves. These items are small, closely packed, and exhibit similar colors and textures, and were also not accurately recognized. The GSSM module primarily enhances the recursive state selection process under global frequency guidance. For tiny objects, local information is easily diluted through multiple downsampling and state transitions. Consequently, GMamba achieves limited performance improvement in small-object scenarios. Future work could explore integrating more powerful local feature enhancement mechanisms to improve performance on small object recognition.

Figure 2: Visualization of failure cases.

Comment 4

Lack experiments on the Imagenet benchmark.

We thank you for raising this important point. We understand that ImageNet classification serves as a standard benchmark for validating general-purpose visual backbone networks. We did not constrain the scope to "dense prediction" in our title because the GMamba module is indeed a fundamental vision module. The extensive experiments we conducted on dense tasks, combined with fair baseline comparisons, sufficiently validate the module's generality and efficiency.

The core motivation of our work lies in addressing the lack of rigorous definition for global context modeling, which subsequently led to the design of the GMamba module to tackle the challenge of "efficiency and capability of global modeling at high resolution."

Classification is typically local: ImageNet classification is inherently object-centric. Convolutional Neural Networks (CNNs) and even Transformer models typically rely on local discriminative features (e.g., textures or specific object parts) rather than global structural understanding to solve classification tasks.

Dense prediction requires global context: In contrast, our chosen tasks—semantic segmentation and object detection—are context-centric. Distinguishing between "road" and "lane," or detecting small objects in cluttered backgrounds, necessitates understanding long-range dependencies and spatial relationships among pixels across the entire image.

We are not disregarding ImageNet; rather, we deliberately chose more challenging tasks that are more sensitive to spatial structural information as our core validation. Meanwhile, we designed GMamba as a plug-and-play module and conducted extensive experiments to demonstrate its efficiency, achieving strong performance on nearly every task, thereby validating its broad deployment capability. Since our central goal is to tackle the tension between efficiency and global modeling capability at high resolution, evaluations on high-resolution segmentation and detection tasks provide substantially stronger evidence than low-resolution (224×224) ImageNet classification.