State Space Models: A Naturally Robust Alternative to Transformers in Computer Vision

Anonymous Authors[1]

Abstract

Visual State Space Models (VSSMs) have recently emerged as a promising architecture, exhibiting remarkable performance in various computer vision tasks. However, its robustness has not yet been thoroughly studied. In this paper, we delve into the robustness of this architecture through comprehensive investigations from multiple perspectives. Firstly, we assess its adversarial robustness using whole-image and patch-specific attacks, finding it superior to Transformers in whole-image attacks but vulnerable to patch-specific attacks. Secondly, we evaluate VSSMs' robustness across diverse scenarios, including natural adversarial examples, out-of-distribution data, and common corruptions. VSSMs generalize well to OOD and corrupted data but struggle with natural adversarial examples. We also analyze their gradients in white-box attacks, revealing unique vulnerabilities and defenses. Lastly, we examine their sensitivity to image structure variations, identifying weaknesses tied to disturbance distribution and spatial information. Through these comprehensive studies, we contribute to a deeper understanding of VSSMs's robustness, providing valuable insights for refining and advancing the capabilities of deep neural networks in computer vision applications.

1. Introduction

Deep neural networks represent a cornerstone of contemporary research (Han et al., 2022; Wang et al., 2018), but their robustness in the face of adversarial attacks (Goodfellow et al., 2014; Madry et al., 2017; Fu et al., 2022) and other perturbations (Hendrycks et al., 2021b;a; Hendrycks & Dietterich, 2019) remains a critical concern. Researchers are increasingly focused on developing models that not only excel in specific tasks but also demonstrate resilience against adversarial attacks while maintaining strong generalization capabilities across diverse scenarios. Recently, a novel and promising addition to the landscape of neural network architectures for visual representation learning has emerged, known as the Visual State Space Model (Liu et al., 2024; Zhu et al., 2024; Huang et al., 2024). This architecture has gained significant attention for its outstanding performance across various computer vision tasks, demonstrating the potential to replace Transformers in a wide range of applications. Despite the considerable successes in various applications, an aspect that has not yet been thoroughly studied is the robustness of VSSMs.

This paper addresses the existing gap in the understanding of VSSMs' robustness by undertaking a comprehensive investigation. To comprehensively evaluate the robustness of VSSMs, our analysis takes a multi-faceted approach, thoroughly exploring the robustness of VSSMs from various perspectives.

Firstly, we analyze the robustness of VSSMs against adversarial attacks. We employ two types of adversarial attacks. The first type targets the entire image, which includes Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2014), and Projected Gradient Descent (PGD) (Madry et al., 2017). The second type focuses on attacking only several patches in an image, which includes Patch-Fool (Fu et al., 2022). This analysis reveals:

1. VSSMs have better adversarial robustness than Transformer architectures (Steiner et al., 2021; Touvron et al., 2021; Liu et al., 2021).
2. The scalability of VSSMs is relatively weak against adversarial attacks.
3. The robustness of VSSMs significantly decreases after removing input-dependent parameter gradients.

Secondly, we assess the general robustness of VSSMs, evaluating its performance against natural adversarial examples in ImageNet-A (Hendrycks et al., 2021b), out-of-distribution data in ImageNet-R (Hendrycks et al., 2021a), and common corruptions in ImageNet-C (Hendrycks & Dietterich, 2019). Understanding the model's behavior in these diverse scenarios is crucial for establishing its reliability in real-world

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

applications. This analysis reveals:

4. VSSMs exhibit superior generalizability when faced with out-of-distribution data and common corruptions.
5. The scalability of VSSMs is relatively weak against natural adversarial examples and common corruptions.

Furthermore, we perform experiments to inspect the gradients of VSSMs and the back-propagation process within VSSMs when subjected to white-box attacks. This exploration aims to investigate how the novel components in VSSMs behave under adversarial attacks. These novel components have not been previously observed in architectures designed for vision tasks. This analysis reveals:

6. Mamba block demonstrates strong defense capabilities against white-box attacks.
7. The vulnerabilities of input-dependent parameters are more easily exploited by patch attacks.
8. The parameter Δ exhibited stronger robustness in hierarchical architectures.
9. The robustness variations of parameter B in hierarchical and non-hierarchical architectures are opposite to the Δ.
10. The parameter Δ and B becomes the main vulnerability in patch-wise attack.
11. The parameter C demonstrates defense capabilities across all VSSMs against White-box attacks.

Finally, we turn our attention to exploring the sensitivity of VSSMs to the structures of images. This investigation encompasses a comparative analysis involving the random removal of image patches and pixels. Additionally, we introduce permutations in the order of image patches to assess their impact. We also employ the Patch-Fool to systematically attack specific positions within the image patches. This multifaceted examination aims to provide a nuanced understanding of how VSSMs respond to variations in image structure, offering valuable insights into its robustness and potential vulnerabilities.

This analysis reveals:

12. VSSMs are more reliant on the integrity of the input patch sequence.
13. VSSMs are more robust to pixel-level perturbations compared to Transformer models.
14. VSSMs are highly sensitive to the spatial information of images, especially the non-hierarchical VSSMs.
15. The closer the perturbation is to the center of the image, the more vulnerable VMamba will be.

Our research offers a comprehensive understanding of the robustness of VSSMs from various perspectives. Through empirical analysis, we delve into the factors that may influence VSSMs's robustness, shedding light on critical aspects that warrant attention for further refinement. By systematically examining its performance under different conditions and stressors, our study contributes valuable insights that can inform future developments in enhancing the VSSMs architecture (as discussed in Section 8). The findings provide a roadmap for researchers to iteratively refine and optimize VSSMs, ultimately advancing their robustness to achieve superior performance compared to their current state.

2. Preliminaries

2.1. Vision Transformers

The **Transformer** (Vaswani et al., 2017) is a model architecture that relies solely on attention mechanisms, initially designed for Natural Language Processing (NLP) tasks. Following the successes of Transformers in NLP tasks, Vision Transformer (**ViT**) (Dosovitskiy et al., 2020; Su et al., 2022) explores the application of a standard Transformer directly to images with minimal modifications. The approach involves dividing an image into patches, treating them akin to tokens (words) in an NLP application, and presenting the sequence of linear embeddings of these patches as input to the Transformer.

The **Swin** Transformer (Liu et al., 2021) introduces a novel hierarchical Transformer architecture with a distinctive emphasis on Shifted windows for representation computation. The proposed shifted windowing scheme enhances efficiency by confining self-attention computation to non-overlapping local windows, concurrently facilitating cross-window connections. This hierarchical design offers flexibility in modeling at multiple scales and maintains linear computational complexity concerning image size. The Swin Transformer's notable attributes, including its efficient windowing strategy, render it versatile and applicable across various vision tasks.

2.2. State Space Models

A novel category of sequence models for deep learning is emerging, which is known as Structured State Space Sequence (S4) models. Using an implicit latent state $h(t) \in \mathbb{R}^N$, S4 models can map a 1-dimensional function or sequence $x(t) \in \mathbb{R}^L \mapsto y(t) \in \mathbb{R}^L$:

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

where $\boldsymbol{A} \in \mathbb{R}^{N\times N}$, $\boldsymbol{B} \in \mathbb{R}^{N\times 1}$, and $\boldsymbol{C} \in \mathbb{R}^{N\times 1}$ are continuous parameters.

In practice, the continuous parameters in Eq. 1 need to be first discretized. This can be achieved using a zero-order hold (ZOH):

$$\overline{\boldsymbol{A}} = \exp(\Delta\boldsymbol{A}) \quad \overline{\boldsymbol{B}} = (\boldsymbol{\Delta A})^{-1}(\exp(\Delta\boldsymbol{A}) - \boldsymbol{I}) \cdot \Delta\boldsymbol{B}, \tag{2}$$

where $\overline{\boldsymbol{A}}$, $\overline{\boldsymbol{B}}$ are the discrete counterparts of the continuous parameters $\boldsymbol{A}$ and $\boldsymbol{B}$, and $\boldsymbol{\Delta} \in \mathbb{R}> 0$ is a specified sampling timescale for the discretization. The discretization leads to a discretized form of the model as follows:

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

A remaining issue is that the iterative process in Eq. 3 is not computationally efficient. To enhance efficiency, it can be sped up through parallel computation. With a global convolution operation (denoted by $\circledast$), we obtain:

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{x} \circledast \overline{\boldsymbol{K}} \\ \text{with} \quad \overline{\boldsymbol{K}} &= (\boldsymbol{C}\overline{\boldsymbol{B}}, \boldsymbol{C}\overline{\boldsymbol{A}\boldsymbol{B}}, ..., \boldsymbol{C}\overline{\boldsymbol{A}}^{L-1}\overline{\boldsymbol{B}}), \end{aligned} \tag{4}$$

where $\overline{\boldsymbol{K}} \in \mathbb{R}^L$ is a kernel used in the S4 model. This method uses convolution to generate outputs across the sequence at the same time, improving computational efficiency and scalability.

2.3. Selective State Space Models

Traditional State Space Models (S4) are known for their linear time complexity but face limitations in capturing sequence context due to fixed parameterization. The Mamba models (Gu & Dao, 2023), overcome these limitations by implementing a dynamic and selective approach for managing interactions between sequential states. Unlike standard SSMs that rely on constant transition parameters ($\overline{\boldsymbol{A}}, \overline{\boldsymbol{B}}$), Mamba utilizes parameters that depend on the input, enabling more complex, sequence-aware parameterization. This approach involves directly deriving parameters $\boldsymbol{B}, \boldsymbol{C}$, and $\boldsymbol{\Delta}$ from the input sequence $\boldsymbol{x}$, which allows for a richer representation of sequence context.

By adopting selective SSMs, Mamba models not only maintain linear scalability with respect to sequence length but also demonstrate strong performance in language modeling tasks. This innovation has paved the way for their application in vision tasks as well, inspiring the development of new models that integrate Mamba. For example, Vim (Zhu et al., 2024) combines Mamba with a ViT-like structure by including bi-directional Mamba blocks instead of the usual Transformer blocks. Similarly, VMamba (Liu et al., 2024) presents an innovative 2D selective scanning method for processing images in both horizontal and vertical directions and builds a hierarchical model that is reminiscent of the Swin Transformer (Liu et al., 2021).

3. Adversarial Robustness

We employ two types of adversarial attacks. The first type targets the entire image. The second type focuses on attacking only several patches in an image.

Table 1: Evaluation of SOTA methods on ImageNet-1K. The top-1 accuracy is used to assess performance on clean ImageNet-1K and under adversarial attacks (FGSM and PGD). All models utilize input dimensions of 224×224.

Categories	Models	Clean	FGSM		PGD	
			1/255	4/255	1/255	4/255
Transformer	ViT-S/16 + AugReg (Steiner et al., 2021)	74.7	17.9	6.0	5.4	0.2
	ViT-B/16 + AugReg (Steiner et al., 2021)	76.8	28.1	10.9	12.1	0.7
	DeiT-Ti (Touvron et al., 2021)	72.2	22.3	11.6	6.2	0.3
	DeiT-S (Touvron et al., 2021)	79.8	40.6	29.0	16.4	2.2
	DeiT-B (Touvron et al., 2021)	81.8	46.3	36.3	22.1	7.3
	Swin-T (Liu et al., 2021)	81.2	33.7	24.2	8.0	1.1
	Swin-S (Liu et al., 2021)	83.2	45.7	37.8	18.7	7.0
	Swin-B (Liu et al., 2021)	83.5	49.2	43.3	22.8	9.1
VSSM	VMamba-T (Liu et al., 2024)	82.6	43.1	33.5	29.8	3.0
	VMamba-S	83.6	47.7	39.4	23.2	7.7
	VMamba-B	83.9	49.1	40.8	23.1	7.5
	Vim-Tiny (Zhu et al., 2024)	76.1	38.8	28.8	19.0	5.2
	Vim-Small	80.5	49.6	42.5	28.6	13.2
	Vim-Base	81.9	50.9	38.6	32.0	10.6
VSSM w/o gradient of Δ, B and C	VMamba-T	82.6	38.8	22.7	8.6	0.3
	VMamba-S	83.6	44.6	32.1	13.5	1.1
	VMamba-B	83.9	41.0	24.0	11.4	0.3
	Vim-Tiny	76.1	31.9	10.1	8.8	0.3
	Vim-Small	80.5	38.4	15.0	13.9	0.7
	Vim-Base	81.9	48.5	16.6	22.7	1.8

3.1. Attack Entire Image

To attack the entire image, we utilize Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2014) and Projected Gradient Descent (PGD) (Madry et al., 2017) with a perturbation magnitude of $\varepsilon = 1/255$ and $\varepsilon = 4/255$. FGSM operates in a single step, whereas PGD represents a multi-step variant of FGSM. Specifically, we iterate PGD for 5 steps. Results in Table 1 reveals: **1. VSSMs have better adversarial robustness than Transformer architectures:** In various attack scenarios, the VSSM model generally outperforms Transformer-based models, demonstrating superior adversarial robustness. This difference is particularly noticeable in smaller models. For example, with a noise magnitude of 1/255, the Vim-tiny model shows a robustness accuracy that is 16.5% and 12.8%, higher than DeiT-Ti under FGSM and PGD attacks, respectively. Similarly, Vmamba-T exhibits a robustness accuracy that is 9.4%, 21.8%, and 13.0% higher than Swin-T under the same attacks.
2. The scalability of VSSMs is relatively weak against adversarial attacks: However, this robustness advantage diminishes as the model size increases. For example, with a noise intensity of 1/255, the Vim-S model only outperforms the DeiT-S model by 9.0%, 12.2%, and 10.3%, which is significantly lower than the advantage seen in the tiny version.
3. The robustness of VSSMs significantly decreases after removing input-dependent parameter gradients: The adversarial robustness of VSSMs heavily relies on input-dependent parameters (Δ, B, and C). When the gradients of these parameters are excluded during the generation of adversarial examples, the robustness of VSSMs drops sharply. For instance, under FGSM with $\varepsilon = 4/255$, the robustness accuracy of VMamba-T decreases from 33.5% to 22.7%, while Vim-Small drops from 42.5% to 15.0%. This shows that the gradients of input-dependent parameters are crucial in enhancing the robustness of VSSMs against adversarial attacks. Detailed discussion will be provided in 5.

Table 2: Robust Accuracy under Patch-fool attack, "P1" to "P4" represent the robust accuracy when a certain number of patches are under attack.

Model	Clean	P1	P2	P3	P4
ViT-S/16 + AugReg	75.9	0.1	0.0	0.0	0.0
ViT-B/16 + AugReg	78.3	12.2	0.9	0.0	0.0
DeiT-T	72.6	4.9	0.0	0.0	0.0
DeiT-S	81.4	6.6	0.2	0.0	0.0
DeiT-B	81.9	25.4	1.7	0.0	0.0
Swin-T	81.5	40.3	16.6	5.7	2.2
Swin-S	84.0	52.6	22.9	10.7	6.0
Swin-B	84.5	43.7	15.7	5.1	2.2
VMamba-T	83.5	30.9	4.3	0.4	0.1
VMamba-S	84.3	40.6	8.3	1.7	0.2
VMamba-B	83.7	34.3	4.1	0.7	0.2
Vim-tiny	74.8	44.1	25.6	15.5	11.5
Vim-tiny†	76.7	47.0	26.6	16.0	11.6
Vim-small	84.7	63.5	45.1	29.6	21.0
Vim-small†	86.0	55.2	28.2	16.5	9.6

Table 3: Top-1 accuracy measures performance on ImageNet-A and -R, while the mean Corruption Error (mCE) is used for ImageNet-C.

Models	A	R	C (↓)
ViT-S/16 + AugReg	9.0	31.9	53.4
ViT-B/16 + AugReg	11.7	36.9	47.8
DeiT-T	7.6	32.7	53.6
DeiT-S	19.5	41.9	41.2
DeiT-B	27.8	44.6	36.7
Swin-T	21.2	41.2	45.9
Swin-S	32.6	44.8	41.0
Swin-B	36.0	46.4	40.2
VMamba-T	26.6	45.3	39.6
VMamba-S	32.8	49.3	36.1
VMamba-B	36.8	49.5	36.1
Vim-tiny	9.5	38.8	46.9
Vim-tiny†	17.2	39.7	44.0
Vim-small	19.7	44.7	38.9
Vim-small†	28.3	44.3	37.5

Table 4: Robust accuracy with the gradients of Δ, B, or C. **All** indicates all four parameters have gradients. **None** indicates none of the four parameters have gradients.

Attack Methods	Model	Have/Has Gradients				
		None	All	Δ	B	C
PGD	VMamba-T	11.0	**29.8**	12.5	10.0	14.2
	VMamba-S	16.4	**23.2**	17.1	15.0	21.9
	VMamba-B	12.7	**23.1**	16.3	12.8	17.9
	Vim-tiny	10.9	**19.0**	9.1	12.4	11.5
	Vim-tiny†	13.5	**21.1**	10.9	16.3	13.9
	Vim-small	14.3	**28.6**	15.3	16.6	15.9
	Vim-small†	17.7	**30.9**	18.2	21.5	19.3
P1	VMamba-T	**60.5**	30.9	40.9	59.9	60.8
	VMamba-S	**70.5**	40.6	46.2	65.8	70.6
	VMamba-B	**71.2**	34.3	41.5	66.6	71.8
	Vim-tiny	**65.1**	44.1	48.0	59.9	65.2
	Vim-tiny†	**65.9**	47.0	48.4	61.0	66.0
	Vim-small	**76.1**	63.5	66.0	69.6	76.7
	Vim-small†	**75.6**	55.2	63.9	71.9	75.8

3.2. Patch-wise Attack

To perform the patch-wise attack, we utilize Patch-Fool (Fu et al., 2022). The results under patch-wise attacks are reported in Table 2. In Patch-Fool experiments, the weight coefficient α is fixed at 0.002. The initial step size η is set to 0.2 and undergoes a 0.95 decay every 10 iterations, with a total of 250 iterations. The targeted patch for the attack is chosen randomly from all the 196 patches. Our experiments include different numbers of targeted patches, ranging from 1 to 4. Adversarial noise learning is optimized using Adam. We randomly select 2,500 images from the ImageNet validation set to evaluate patch-wise robustness, following (Fu et al., 2022). The "Clean" column shows accuracy on unperturbed images, while "P1" to "P4" indicates robust accuracy under increasing patch attacks.

Compared to DeiT and Swin models, the Vim model maintains higher accuracy across the P1 to P4 range, indicating that the Vim model exhibits stronger robustness against patch-wise white-box adversarial attacks. However, this robustness advantage is not sustained in the hierarchical architecture of VSSMs, specifically in VMamba. For instance, in the case of P1, the robust accuracy of the T, S, and B versions of VMamba is lower than that of the Swin transformer by 9.4%, 12.0%, and 9.5%, respectively. In response to this phenomenon, we will discuss the robustness differences between hierarchical and non-hierarchical architectures of VSSMs in Section **The role of parameters Δ, B, and C in model robustness**. In addition, in Section **Sensitivity to Information Loss**, we provide a more detailed discussion on the impact of the number of perturbed patches or pixels on the performance of VSSMs.

4. General Robustness

The ImageNet-A dataset (Hendrycks et al., 2021b) comprises natural adversarial examples that challenge models by placing ImageNet objects in unconventional contexts or orientations. This assesses the model's adaptability to unexpected scenarios. In contrast, the ImageNet-R dataset (Hendrycks et al., 2021a) introduces out-of-distribution data, presenting abstract or rendered versions of objects to test the model's ability to generalize beyond its trained data distribution. Lastly, the ImageNet-C dataset (Hendrycks & Dietterich, 2019) introduces common corruptions, incorporating 19 distortions across 5 categories, such as motion blur, Gaussian noise, fog, and JPEG compression. This dataset emulates real-world distortions, providing insights into a model's resilience to diverse environmental challenges. The results are reported in Table 3.

4. VSSMs exhibits superior generalizability when faced with out-of-distribution data (ImageNet-R) and common corruptions data (ImageNet-C): Notably, the VMamba and Vim models consistently demonstrate superior performance compared to the Swin and DeiT models, respectively, on both ImageNet-R and ImageNet-C. For instance, on the out-of-distribution data (ImageNet-R) and common corruption data (ImageNet-C), VMamba-T surpasses Swin-T by 4.1% and 6.3%, VMamba-S outperforms Swin-S by 4.5% and 4.9%, VMamba-B exceeds Swin-B by 3.1% and 4.1% respectively. Vim-Tiny surpasses DiT-T by 6.1% and 6.7%, and Vim-small surpasses DiT-S by 2.8% and 2.3%.

One potential explanation is that VSSMs exhibit a superior capability for managing long-range dependencies compared to transformers (Yu & Wang, 2024). This allows them to more effectively integrate features from various regions of an image, enhancing the overall feature representation. In other words, even if part of the image information is missing, the model may still be able to understand the image by using global information. Additionally, In the context of out-of-distribution data, this enhanced feature representation aids the model in recognizing and comprehending previously unseen patterns or variations, leading to improved classifi-

cation performance.

5. The scalability of VSSMs is relatively weak against natural adversarial examples (ImageNet-A): The smallest variants, VMamba-T and Vim-tiny, showcase substantial superiority, exhibiting significant improvements of 5.4% and 1.9% over Swin-T and DeiT-T on the ImageNet-A dataset, respectively. However, this advantage was not sustained in their larger variants. Specifically, VMamba showed only marginal improvements of 0.1% and 0.8% compared to the S and B versions of Swin, respectively. Similarly, Vim exhibited a minimal enhancement of 0.2% relative to the S version of DeiT. indicating that its performance may not uniformly improve with model size across different ImageNet variants. This phenomenon indicates a weakness in the scalability of VSSMs on natural adversarial examples datasets, suggesting that their performance may not uniformly improve with increases in model size.

5. The role of Input-dependent parameters in model robustness.

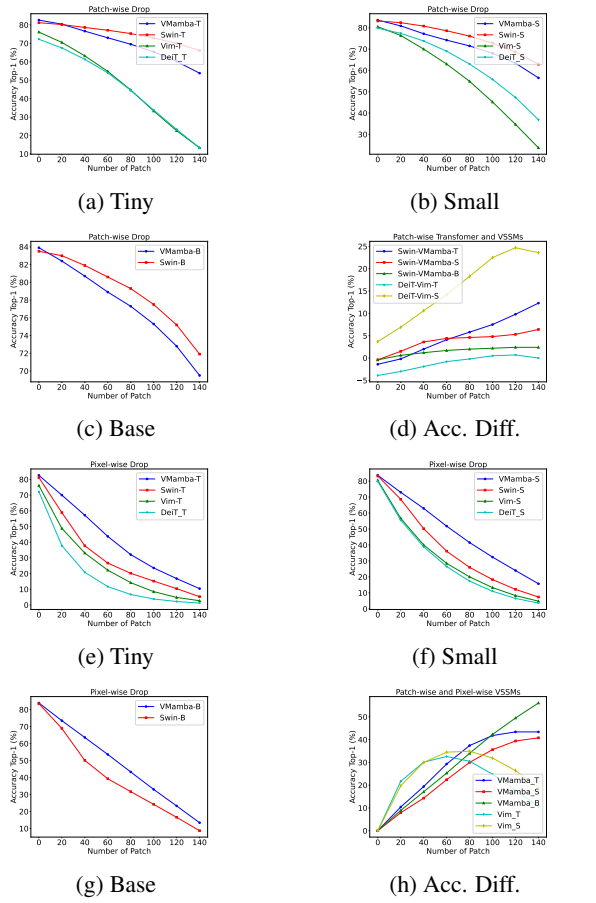
To analyze the robustness of the VSSMs under white-box adversarial attacks, we examine the effect of individual parameters on the robustness of models. This will be achieved by measuring the robust accuracy of models using PGD or Patch-Fool after individually activating the gradients of parameters B, C, and Δ. In Table 4, the **All** column represents the robust accuracy of the model under the original white-box setting, which estimates the full parameter gradients. The **None** column represents the robust accuracy under a white-box attack without the participation of gradients for all parameters Δ, B, and C. The appended data columns quantify the robust accuracy upon the gradient activation of each respective parameter.

6. Mamba block demonstrates strong defense capabilities against PGD attacks but shows vulnerability against Patch-Fool: Table 4 indicates that disabling the gradient of the mamba block leads to a substantial decrease in the robust accuracy of VMamba and Vim under PGD attacks. Specifically, the robust accuracy of VMamba's T, S, and B variants decreased by 4.3%, 6.8%, and 10.4%, respectively, while Vim's Tiny and Small variants decreased by 8.1% and 14.3%, respectively. These significant reductions in robust accuracy highlight the critical role of the mamba block in maintaining the robustness of VSSM against PGD attacks.

7. The vulnerabilities of input-dependent parameters are more easily exploited by patch attacks. However, for Patch-Fool attacks, without the gradient of the mamba block, the robust accuracy of VMamba and Vim consistently shows an upward trend. Specifically, VMamba's T, S, and B variants increase by 29.6%, 29.9%, and 36.9%, respectively, while Vim's Tiny and Small variants increase by 21% and 12%. This indicates that the Mamba block exhibits vulnerability under patch attacks. This difference raises an intriguing question: why are input-dependent parameters particularly critical in defending against patch attacks? The answer lies in their localized nature. Input-dependent parameters are derived directly from individual patches and lack a comprehensive understanding of the entire input context. This patch-centric transformation makes them inherently more susceptible to attacks that exploit isolated regions of the input, such as Patch-Fool.

8. The parameter Δ exhibited stronger robustness in hierarchical architectures under PGD attacks: In the scenario of PGD attack, by activating the gradient of the parameter Δ, the robust accuracy of hierarchical VSSMs, specifically the VMamba's T, S, and B variants, increased by 1.5%, 0.7%, and 3.6%, respectively. However, for the non-hierarchical VSSMs, the robust accuracy of the Vim-tiny decreased by 1.8%, while the Vim-small increased by 1%. This result suggests that the robustness conferred by the parameter Δ exhibits a more pronounced advantage within hierarchical architectures. In VSSMs, the Δ parameter transforms continuous-time system parameters into their discrete counterparts, enabling the model to capture the dynamics of input data across different time scales, thereby filtering out irrelevant information and noise from the data stream. Hierarchical architectures typically involve processing information at different levels, where each layer can be regarded as operating on a specific spatial or temporal scale. This design allows the model to capture the dynamics of input data across multiple levels, providing a more favorable environment for the parameter Δ to function effectively.

9. The robustness variations of parameter B in hierarchical and non-hierarchical architectures are opposite to the Δ parameter under PGD attacks: However, it is noteworthy that the robustness variations of parameter B in hierarchical and non-hierarchical architectures are opposite to the Δ parameter. By activating gradients of the parameter B, the VMamba's T, S, and B variants experienced respective decreases of 1.0% and 1.4%, and with a marginal increase of 0.1%. Conversely, Vim displayed consistent improvements in robustness across the Tiny and Small versions, with increases of 1.5% and 2.3%, respectively. This phenomenon may reflect their complementary roles in the model robustness. Parameter B primarily manages the mapping of those filtered input data by parameter Δ into the state space $h(t)$ in VSSMs. In hierarchical architectures, the parameter delta has demonstrated a strong ability to filter input data, which is often an advantage because it can help the model ignore irrelevant information and focus on important features. However, it could also make parameter B overfit to those stable, undisturbed data distributions. Thereby reducing the robustness of the parameter B in the hierarchical architecture.



(a) Tiny (b) Small

(c) Base (d) Acc. Diff.

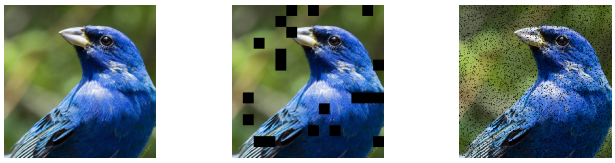
(e) Tiny (f) Small

(g) Base (h) Acc. Diff.

Figure 1: Robust accuracy under varying information loss.

10. The parameter Δ and B becomes the main vulnerability in patch-wise attack: Under the attack from Patch-Fool, activating the gradient of the parameters Δ or B leads to a consistent decrease in the robustness of all VSSMs. Specifically, the VMamba variants T, S, and B see reductions of 19.7%, 24.3% and 29.7% for parameters Δ, and 0.6%, 4.7%, and 4.6% for parameters B. For the tiny and small versions of Vim, the decreases are 17.1% and 10.1% for parameter Δ, and 5.1% and 6.6% for parameter B. This phenomenon indicates that compared to attacks on the entire image(PGD), VSSMs seem to be more sensitive to patch attacks. We will discuss this in detail in the Section on Sensitivity to Information Loss.

11. The parameter C demonstrates defense capabilities across all VSSMs against White-box attack: In VSSMs, the primary role of parameter C is to selectively transform the hidden state $h(t)$ into the final output $y(t)$. By activating gradients of the parameter C, both VMamba and Vim consistently exhibited increases in robust accuracy. Specifically, VMamba's T, S, and B variants improved by 3.2%, 5.5%, and 5.2%, respectively, while Vim's Tiny and Small versions show increases of 0.6% and 1.6%, respectively.



(a) Origin Image (b) Patch-wise (c) Pixel-wise

Figure 2: Patch-wise and pixel-wise drop.

6. Sensitivity to Information Loss

The Transformer and VSSMs both require converting images into sequences of patches of a specified length L to predict labels Y. The Transformer models utilize self-attention to facilitate parallel interactions between patches. In contrast, the VMamba model introduces an innovative 2D selective scanning method to process images in both horizontal and vertical directions.

To analyze the impact of the patch interaction mechanisms adopted by the two models on model robustness, we design two different experiments from the perspectives of both dense and sparse perturbations. *Dense perturbation* means destroying all information within a single patch, while *sparse perturbation* means distributing an equal amount of perturbation across multiple patches. The main difference between these two methods lies in the granularity of the information omission. Patch-wise drop affects larger, contiguous areas of the image, while the pixel-wise drop is employed to evenly distribute perturbations across each patch, which means the perturbation is more sparse and covers a wider range.

The experimental results are shown in the Fig.1. The horizontal axis (number of patches) quantifies the amount of information equivalent to how many patches were lost, for example, dropping 10 patches of size 16×16 is equivalent to 256×10 pixels, and the vertical axis represents robust accuracy. In this section, the drop (or loss) of patches or pixels means setting the corresponding values to zero.

6.1. Sensitivity to the Dense Perturbation

The dense perturbation, i.e. patch-wise drops, randomly selects and drops from the 196 patches (Fig. 2b). In such a setup, the robust performance of the Swin model consistently surpasses that of VMamba, while the DeiT model demonstrates markedly superior robustness compared to Vim across all model variants as demonstrated in the first row of Fig. 1. This phenomenon indicates that **12.VSSMs are more reliant on the integrity of the input patch sequence.** There are two possible reasons for this phenomenon: Firstly, patch-wise dropout can disrupt the structure of the image to some extent, thereby affecting the model's ability to comprehend contextual information. This factor impacts both the Transformer models and VSSMs models. This reason will be further inspected in Section

Sensitivity to the Relative Position of Patches. Secondly, it may disrupt the continuity of the scanning trajectory of the VSSMs. Specifically, delving into Eq. 3, if the input x_t at time step t is an all-zero matrix, the state update can be simplified to $h_t = \overline{\boldsymbol{A}} h_{t-1}$, indicating that the new state is entirely reliant on the previous state. It is noteworthy that all elements of the matrix $\overline{\boldsymbol{A}}$ are strictly within the range of 0 to 1. As a result, with a sufficient number of dropped patches, the state will gradually approach zero. This means that the VSSMs may progressively simplify or forget previously stored information, potentially harming the model's long-term memory capability.

6.2. Sensitivity to Sparse Perturbation

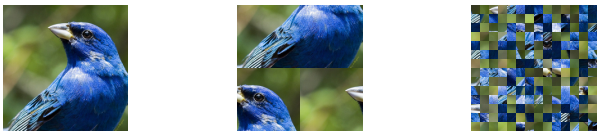
In the sparse perturbation, we introduced pixel-wise drop which employs a similar randomization strategy to a patch-wise drop but is applied to the individual pixels across the 224x224 pixel space of the image (Fig. 2c). In the second row of Fig. 1, we observed that **13. VSSMs are more robust to pixel-level perturbations compared to Transformer models**, as evidenced by the superior performance of both the VMamba and Vim models relative to their corresponding Transformer counterparts. This phenomenon may be attributed to VSSMs' superior long-range dependency capabilities compared to Transformer models. This strength likely allows the Mamba model to maintain a stable comprehension and processing of the overall input sequence, even when faced with pixel-level perturbations.

7. Sensitivity to the Relative Position and Absolute Position of Patches

When capturing contextual information within a sequence of image patches, Transformers use positional embeddings to encode both the position and spatial information of each patch. In contrast, VSSMs achieve this contextual understanding through the order of its scanning trajectory. This fundamental difference in how contextual relationships are understood may result in varying sensitivities to perturbations in image structure. In this section, we will conduct a detailed analysis of VSSM's robustness under image structure perturbations, considering both the relative and absolute positions of patches.

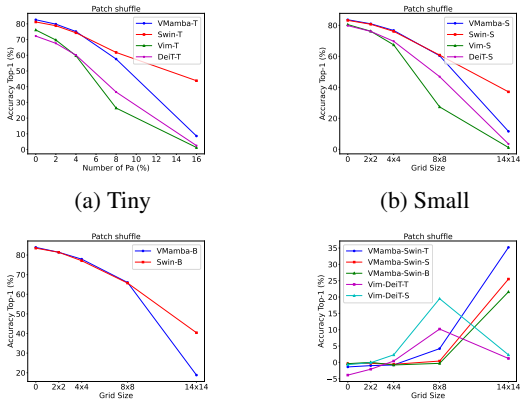
7.1. Relative Position

As shown in Fig.3, when the image is divided into a 2x2 grid, the main subject of the picture can still be easily identified. However, when the grid number is increased to 14x14, it becomes quite challenging to recognize the original subject of the image after the shuffle operation. Therefore, we will employ the number of grids as the horizontal axis in the following experiments to represent the extent of disorders.



(a) Origin Image (b) 2×2 Grid (c) 14×14 Grid

Figure 3: Example images and their different extend of disorder examples.



(a) Tiny (b) Small

(c) Base (d) Acc. Diff.

Figure 4: Figures (a), (b), and (c) represent the robust accuracy related to degrees of shuffle for the Tiny, Small, and Base versions of VSSMs and Transformer-based models, respectively. Figure (d) illustrates the difference in robust accuracy between the models used for comparison.

Fig.4 shows the correlation between robust accuracy and disorder. As the number of grids increases, both VSSMs and Transformer-based models exhibit a decline in accuracy. However, the decline is more pronounced in VMamba and Vim models. For example, after the grid size reaches 8x8, the robust accuracy of the three VMamba variants (T, S, B) is significantly lower than that of the Swin model. Likewise, the two Vim variants exhibit a marked decrease in robust accuracy compared to the DeiT model once the grid size exceeds 4x4. This indicates that **14. VSSMs are highly sensitive to the spatial information of images, especially the non-hierarchical VSSM.**

7.2. Absolute Position

The 3D surface plots Fig. 5 illustrate the performance of the VMamba and Swin models under a white-box adversarial attack targeting the absolute positions of 196 image patches. For the VMamba model, the plot on the left indicates a vulnerability trend where patches near the center are more prone to adversarial attacks, as evidenced by the dips in accuracy within the central region. This central susceptibility is further contextualized by the operational dynamics of the VMamba model, where the image's central region is generally the culmination point within its scanning trajectory. This indicates that **15. The closer the perturbation is to the center of the image, the more vulnerable**

(a) VMamba-T (b) Swin-T

Figure 5: Performance of the VMamba-T and Swin-T models under Patch-fool attacks targeting each of 196 image patches, For the efficiency of this experiment, we randomly selected 500 images from the ImageNet-1K validation set.

VMamba will be. Contrastingly, the Swin model, as shown on the right, exhibits a more irregular and turbulent response to the adversarial attacks. The sharp fluctuations in accuracy across different patch positions suggest that the Swin model's performance is unevenly affected by the perturbations, with certain areas being more resilient than others. Comparing the two, VMamba's uniformity in performance degradation across patch positions, with a notable central vulnerability, contrasts with Swin's more erratic response, indicating different internal processing and utilization of spatial information within the models.

8. Insights

8.1. Robustness of Input-dependent Parameters

Through a detailed analysis of the effects of parameters B, C, and Δ on the model's robustness, it was observed that parameters B and Δ demonstrate a complementary relationship. Notably, during patch-wise attacks, the defense capability associated with the Δ parameter significantly deteriorates when compared to its performance in full-image white-box attacks. These observations indicate that B and Δ may possess distinct strengths and vulnerabilities in responding to different types of adversarial attacks. To address this challenge and optimize the model's overall robustness, the following strategies can be considered:

- The first strategy involves Parameter-independent adversarial training. For example, adversarial examples could be generated specifically targeting attacks that impact B and Δ, respectively. These parameters can then be optimized separately to enhance their ability to resist specific types of attacks.
- The second strategy revolves around joint optimization. Consider the combined effects of B and Δ during adversarial training to find an optimal robustness balance between the two. This can be achieved through a multi-objective optimization strategy, where both robustness metrics are taken into account to seek a holistic optimization solution. This approach may involve adjusting the loss function to simultaneously reflect the robustness requirements of both parameters.

8.2. Reduce Sensitivity to Information Loss

For the two different types of information loss, Patch-wise drop and Pixel-wise drop, VSSMs show different vulnerabilities. Specifically, in the case of Patch-wise drop, the loss of information not only disrupts the overall structure of the image but also breaks the continuity of the model's scanning process, severely affecting the model's long-range dependency capabilities.

- To address this issue, we could develop adaptive scanning strategies that allow the model to dynamically adjust its scanning path in response to detected patch drops. By identifying areas of information loss in the image, the model can reroute its scanning trajectory to prioritize intact areas and infer missing information based on the context and spatial relationships of the remaining patches.

On the other hand, in the scenario of Pixel-wise drop, the impact on the overall structure of the image is relatively minor and does not completely disrupt the continuity of the scanning process. However, each scanning step generates minor errors, leading to a significant deviation in the final output from what is expected.

- A potential solution to mitigate this issue involves robust feature extraction techniques that tolerate minor errors. Specifically, during the training process, we can introduce a small perturbation to intermediate latent state $h(t)$ to reduce the model's dependency on the results of the previous scan, thereby enhancing the model's robustness.

9. Conclusion

In conclusion, our thorough examination of the VSSMs emphasizes their considerable promise in computer vision tasks. While excelling in performance across various domains, our focus on robustness reveals nuanced aspects. VSSMs demonstrate superior robustness to adversarial attacks compared to Transformer architectures yet expose scalability vulnerabilities. General robustness assessments showcase remarkable out-of-distribution generalizability but unveil weaknesses against natural adversarial examples and common corruptions. Exploring VSSMs' gradients and back-propagation during white-box attacks exposes unique vulnerabilities and defensive capabilities within its novel components. Furthermore, sensitivity analysis elucidates vulnerabilities associated with the distribution of disturbance area and spatial information, particularly accentuated near the image center. This comprehensive analysis offers several insights that can enhance the robustness of VSSMs.

References

Croce, F. and Hein, M. Sparse and imperceivable adversarial attacks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4724–4732, 2019.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Dong, X., Chen, D., Bao, J., Qin, C., Yuan, L., Zhang, W., Yu, N., and Chen, D. Greedyfool: Distortion-aware sparse adversarial attack. *Advances in Neural Information Processing Systems*, 33:11226–11236, 2020.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Fu, Y., Zhang, S., Wu, S., Wan, C., and Lin, Y. Patch-fool: Are vision transformers always robust against adversarial perturbations? *arXiv preprint arXiv:2203.08392*, 2022.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

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

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

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

Han, D., Wang, Z., Xia, Z., Han, Y., Pu, Y., Ge, C., Song, J., Song, S., Zheng, B., and Huang, G. Demystify mamba in vision: A linear attention perspective. *arXiv preprint arXiv:2405.16605*, 2024.

Han, K., Wang, Y., Xu, C., Guo, J., Xu, C., Wu, E., and Tian, Q. Ghostnets on heterogeneous devices via cheap operations. *International Journal of Computer Vision*, 130(4):1050–1069, 2022.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8349, 2021a.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples. *CVPR*, 2021b.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Huang, T., Pei, X., You, S., Wang, F., Qian, C., and Xu, C. Localmamba: Visual state space model with windowed selective scan. *arXiv preprint arXiv:2403.09338*, 2024.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

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

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Ma, J., Li, F., and Wang, B. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*, 2024.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Modas, A., Moosavi-Dezfooli, S.-M., and Frossard, P. Sparsefool: a few pixels make a big difference. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9087–9096, 2019.

Nguyen, E., Goel, K., Gu, A., Downs, G., Shah, P., Dao, T., Baccus, S., and Ré, C. S4nd: Modeling images and videos as multidimensional signals with state spaces. *Advances in neural information processing systems*, 35:2846–2861, 2022.

Patro, B. N. and Agneeswaran, V. S. Simba: Simplified mamba-based architecture for vision and multivariate time series. *arXiv preprint arXiv:2403.15360*, 2024.

Pei, X., Huang, T., and Xu, C. Efficientvmamba: Atrous selective scan for light weight visual mamba. *arXiv preprint arXiv:2403.09977*, 2024.

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

Steiner, A., Kolesnikov, A., Zhai, X., Wightman, R., Uszkoreit, J., and Beyer, L. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021.

Su, X., You, S., Xie, J., Zheng, M., Wang, F., Qian, C., Zhang, C., Wang, X., and Xu, C. Vitas: Vision transformer architecture search. In *European Conference on Computer Vision*, pp. 139–157. Springer Nature Switzerland Cham, 2022.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jegou, H. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, volume 139, pp. 10347–10357, July 2021.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wang, Y., Xu, C., Xu, C., Xu, C., and Tao, D. Learning versatile filters for efficient convolutional neural networks. *Advances in Neural Information Processing Systems*, 31, 2018.

Yang, C., Chen, Z., Espinosa, M., Ericsson, L., Wang, Z., Liu, J., and Crowley, E. J. Plainmamba: Improving non-hierarchical mamba in visual recognition. *arXiv preprint arXiv:2403.17695*, 2024.

Yu, W. and Wang, X. Mambaout: Do we really need mamba for vision? *arXiv preprint arXiv:2405.07992*, 2024.

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

Table 5: Transferability of adversarial samples between the SSM and transformer models. The first column "A vs. B" indicates the two models used for assessment. The second and third columns respectively show the accuracy of the A and B models on clean images. The fourth column, "Left to Right", represents the robust accuracy after transferring the adversarial samples generated using model A's gradients to model B, and vice versa. The last column, "Left to Right(w/o)", represents the robust accuracy of adversarial examples generated using the gradient of VSSMs (excluding the gradient of the mamba block) after being transferred to transformer models. All adversarial samples are generated by using PGD. Vim† is the version adapted for Long Sequence Fine-tuning. Specifically, they maintained the original patch size but adjusted the patch extraction stride to 8.

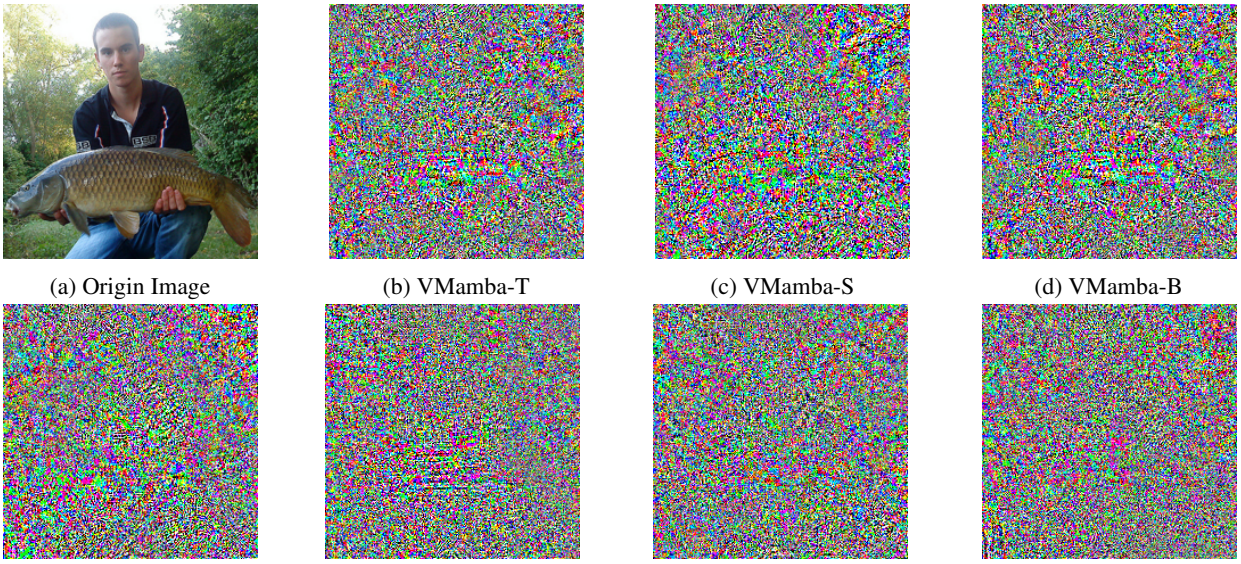
A vs. B	**SSM**	**Transformer**	**Left → Right**		**Right → Left**		**Left → Right(w/o)**	
			1/255	**4/255**	**1/255**	**4/255**	**1/255**	**4/255**
VMamba-T vs. DeiT-Ti	82.6	72.2	69.3	65.0	80.5	76.7	68.8	63.3
VMamba-T vs. Swin-T	82.6	81.2	71.9	61.0	77.9	72.0	69.8	56.5
Vim-Ti vs. DeiT-Ti	76.1	72.2	61.6	46.1	67.2	50.6	59.8	40.2
Vim-Ti vs. Swin-T	76.1	81.2	77.6	72.6	74.2	71.1	77.1	70.2
Vim-Ti† vs. DeiT-Ti	78.3	72.2	65.2	54.4	73.9	66.5	63.4	48.0
Vim-Ti† vs. Swin-T	78.3	81.2	76.4	69.9	76.2	73.2	75.4	65.5
VMamba-S vs. DeiT-S	83.6	79.8	76.7	72.1	80.6	75.4	75.9	69.4
VMamba-S vs. ViT-S	83.6	74.7	72.1	67.8	81.6	77.6	71.4	65.6
VMamba-S vs. Swin-S	83.6	83.2	73.6	62.8	79.1	73.7	71.3	57.3
Vim-S vs. DeiT-S	80.5	79.8	72.3	60.2	71.6	58.5	70.2	53.0
Vim-S vs. ViT-S	80.5	74.7	69.1	58.9	74.3	62.9	67.3	53.4
Vim-S vs. Swin-S	80.5	83.2	79.5	74.6	78.4	75.5	78.6	71.2
Vim-S† vs. DeiT-S	81.6	79.8	74.7	66.5	76.7	69.5	72.7	59.6
Vim-S† vs. ViT-S	81.6	74.7	70.8	63.7	78.0	71.8	68.9	58.4
Vim-S† vs. Swin-S	81.6	83.2	79.0	73.5	79.5	76.5	77.5	68.6
VMamba-B vs. DeiT-B	83.9	81.8	78.4	73.8	80.0	73.9	77.5	70.9
VMamba-B vs. ViT-B	83.9	76.8	74.7	71.0	81.5	77.1	74.1	69.0
VMamba-B vs. Swin-B	83.9	83.5	73.7	63.1	78.7	73.2	70.9	56.3

A. Black-box Attacks and Transferability of Noises

In this section, we comprehensively assessed the transferability of adversarial samples between the VSSMs and transformer models. In Table 5, the first column "A vs. B" indicates the two models used for assessment. The second and third columns respectively show the accuracy of the A and B models on clean images. The fourth column, "Left to Right", represents the robust accuracy after transferring the adversarial samples generated using model A's gradients to model B, and vice versa. The last column, "Left to Right(w/o)", represents the robust accuracy of adversarial examples generated using the gradient of VSSMs (excluding the gradient of the mamba block) after being transferred to transformer models.

Table 5 reveals that adversarial samples generated by the Swin model lead to the lowest robust accuracy when transferred to all sizes of the VMamba model. Similarly, the Vim model demonstrates the lowest robust accuracy when adversarial samples are transferred from the DeiT model across all sizes, with the ViT model following closely behind. This indicates that VMamba shares similar image feature extraction characteristics with Swin, whereas Vim demonstrates a closer alignment with DeiT. This similarity is likely attributable to the design of their respective model architectures. This assumption is further illustrated in Fig.6 and Fig. 7, which show example perturbations generated for VMamba, Vim, and Transformer models using the PGD method. Therefore, to reasonably assess the robustness differences, we prefer to compare VMamba with Swin, and Vim with ViT and DeiT in the main paper.

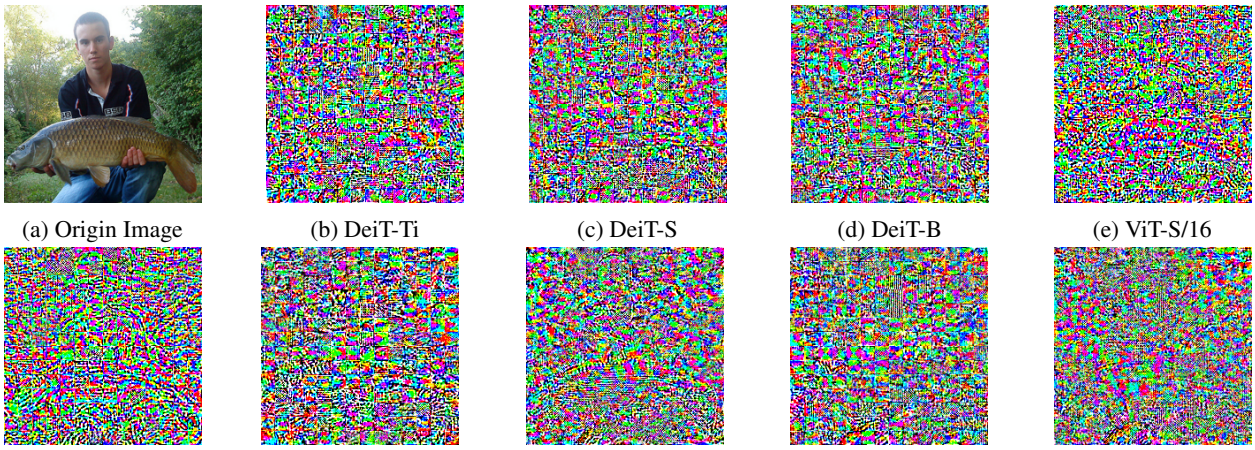
It is also noteworthy that adversarial samples generated by VSSMs without using the Mamba block exhibit stronger transferability. This further validates the discussion in the section **The role of parameters Δ, A, B, and C in model robustness**, where it is mentioned that the Mamba block demonstrates strong defense capabilities against PGD attacks.



(a) Origin Image (b) VMamba-T (c) VMamba-S (d) VMamba-B

(e) VMamba-B* (f) Swin-T (g) Swin-S (h) Swin-B

Figure 6: The example perturbations generated with PGD for VMamba and Swin model.



(a) Origin Image (b) DeiT-Ti (c) DeiT-S (d) DeiT-B (e) ViT-S/16

(f) ViT-B/16 (g) Vim-tiny (h) Vim-tiny+ (i) Vim-small (j) Vim-small+

Figure 7: The example perturbations generated with PGD for Vim, DeiT, and ViT model.

B. Related Works

B.1. State Space Models

State Space Models (SSMs) have emerged in deep learning, demonstrating their effectiveness in inefficient long sequence modeling. This success has garnered significant attention from both the Natural Language Processing and Computer Vision communities. The Linear State-Space Layer (LSSL) (Gu et al., 2021b) combines recurrent, convolutional, and continuous-time models to address their individual shortcomings. The LSSL model demonstrates state-of-the-art performance in time-series tasks, surpassing previous approaches on sequential image classification, healthcare regression, and speech tasks. The Structured State Space sequence model (S4) (Gu et al., 2021a) focuses on efficient modeling of long sequences by optimizing the fundamental SSM. This approach demonstrates strong empirical results across diverse benchmarks, achieving competitive accuracy on sequential CIFAR-10 and outperforming prior methods on the Long Range Arena benchmark, including the challenging Path-X task. The S5 model (Smith et al., 2022) extends the structured state space paradigm with the S5 layer, which leverages multi-input, multi-output SSMs for efficient parallel processing. The S5 layer achieves state-of-the-art results on long-range sequence modeling tasks, showcasing its prowess in tasks like the Long Range Arena benchmark's Path-X. Mamba (Gu & Dao, 2023) is a sequence model that leverages structured state spaces (SSMs) to achieve linear-time sequence modeling. Notably, Mamba surpasses the computational efficiency of Transformers with a $5\times$ higher throughput, excelling across modalities such as language, audio, and genomics.

The S4ND (Nguyen et al., 2022) model extends the continuous-signal modeling prowess of state space models (SSMs) to multidimensional data like images and videos. S4ND excels in modeling large-scale visual data in 1D, 2D, and 3D as continuous multidimensional signals, showcasing superior performance on practical tasks. When integrated into existing state-of-the-art models by replacing Conv2D and self-attention layers, S4ND outperforms a Vision Transformer baseline on ImageNet-1k and matches ConvNeXt in 2D image modeling. For video tasks, S4ND improves activity classification on HMDB-51 compared to an inflated 3D ConvNeXt. VSSMs (Liu et al., 2024; Zhu et al., 2024; Ma et al., 2024; Han et al., 2024; Patro & Agneeswaran, 2024; Huang et al., 2024; Yang et al., 2024; Pei et al., 2024) introduces a Visual State Space Model inspired by state space models, and designed to achieve linear complexity while preserving global receptive fields. Specifically, VMamba addresses direction-sensitive issues with the Cross-Scan Module (CSM) and exhibits promising capabilities across various visual perception tasks, outperforming established benchmarks as image resolution increases. These works collectively advance the understanding and efficiency of visual representation learning models.

B.2. Adversarial Robustness

Szegedy *et al.* (Szegedy et al., 2013) uncovered a significant vulnerability in state-of-the-art neural networks and machine learning models. Their discovery highlighted the vulnerability of these models to adversarial examples. Adversarial examples are instances that lead to misclassifications when they are slightly altered. Building upon the work of Szegedy *et al.* (Szegedy et al., 2013), numerous novel methods have been developed to generate adversarial noises, enabling the effective alteration of inputs to models.

Fast Gradient Sign Method attack (Goodfellow et al., 2014) (FGSM) has proven that the linear behavior in high-dimensional spaces is adequate to induce adversarial examples, marking a crucial insight in the realm of adversarial machine learning. This perspective has facilitated the development of a rapid adversarial example generation method, thereby rendering adversarial training more practical. Let $\boldsymbol{\theta}$ denote the parameters of a model, $\boldsymbol{x}$ denote the input, y denote the targets, and $J(\boldsymbol{\theta}, \boldsymbol{x}, y)$ denote the loss function for training the neural network. The optimal max-norm constrained perturbation, denoted as $\boldsymbol{\eta}$ can be calculated by:

$$\boldsymbol{\eta} = \varepsilon \operatorname{sign}\left(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y)\right), \tag{5}$$

where ε is a step size. Then, the adversarial example $\boldsymbol{x}' = \boldsymbol{x} + \boldsymbol{\eta}$ is obtained by adding the perturbation $\boldsymbol{\eta}$ to the original input $\boldsymbol{x}$.

Projected Gradient Descent attack (Madry et al., 2017) (PGD) offers a distinctive perspective on the adversarial attack and defense problem by framing it as a saddle point problem

$$\min_{\boldsymbol{\theta}} \rho(\boldsymbol{\theta}), \quad \text{where } \rho(\boldsymbol{\theta}) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}\left[\max_{\boldsymbol{\delta}\in\mathcal{S}} J(\boldsymbol{\theta}, \boldsymbol{x}+\boldsymbol{\delta}, y)\right]. \tag{6}$$

This formulation allows PGD to interpret the FGSM attack as a straightforward one-step scheme aimed at maximizing the inner part of the saddle point formulation. A more powerful adversary is introduced through the multi-step variant, essentially aligning with the principles of Projected Gradient Descent applied to the negative loss function:

$$\boldsymbol{x}^{t+1} = \prod_{\boldsymbol{x}+\mathcal{S}} \left(\boldsymbol{x}^{t} + \varepsilon \operatorname{sign} \left(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y) \right) \right). \tag{7}$$

This approach broadens the understanding of adversarial attacks, providing a novel view that extends beyond the simple one-step scheme.

Patch-Fool attack (Fu et al., 2022) introduces a novel strategy by constraining perturbed pixels within one patch or several patches, unlike previous approaches that limit perturbation strength onto each pixel. This method can be viewed as a variant of sparse attacks (Modas et al., 2019; Croce & Hein, 2019; Dong et al., 2020). This approach produces adversarial examples with noisy patches, visually resembling and emulating natural corruptions within a small region of the original image. The objective of Patch-Fool can be formulated as:

$$\underset{1 \leq p \leq n, \boldsymbol{E} \in \mathbb{R}^{n \times d}}{\arg\max} \; J\left(\boldsymbol{x} + \mathbf{1}_p \odot \boldsymbol{E}, y \right) \tag{8}$$

where $\boldsymbol{E}$ is the adversarial perturbation, $\mathbf{1}_p \in \mathbb{R}^n$ is a one-hot vector, and $\odot$ represents the penetrating face product.

B.3. General Robustness

ImageNet-A (Hendrycks et al., 2021b) is a challenging dataset designed to expose vulnerabilities in machine learning model performance. Created through a simple adversarial filtration technique that minimizes spurious cues, ImageNet-A presents a formidable challenge for existing models, surpassing the difficulty level of the conventional ImageNet (Deng et al., 2009) test set. Notably, a DenseNet-121 (Huang et al., 2017) model achieves a mere 2% accuracy on ImageNet-A, reflecting a drastic 90% drop in performance. The dataset comprises real-world, unmodified examples that consistently challenge diverse models, unveiling shared weaknesses in computer vision algorithms.

ImageNet-R (Hendrycks et al., 2021a) is a novel test set comprising 30,000 images that offers a distinctive challenge for evaluating the robustness of machine learning models. This dataset includes diverse renditions of ImageNet object classes, such as paintings and embroidery, introducing natural variations in textures and local image statistics not present in conventional ImageNet (Deng et al., 2009) images. By incorporating these naturally occurring renditions, ImageNet-R allows for a meaningful assessment of model performance in the face of realistic visual variations. The dataset serves as a valuable benchmark to gauge the effectiveness of previously proposed methods aimed at enhancing out-of-distribution robustness. Researchers can leverage ImageNet-R to rigorously test and compare various strategies for improving model performance on real-world, visually diverse renditions, offering a more comprehensive evaluation of robustness in the realm of image classification.

ImageNet-C (Hendrycks & Dietterich, 2019) is a dataset designed to evaluate the robustness of machine learning models to various common visual corruptions. Comprising a collection of 75 widely encountered visual corruptions, this dataset applies these distortions to images from the ImageNet (Deng et al., 2009). The introduction of ImageNet-C aims to establish a standardized benchmark for assessing the robustness of models to image corruptions, addressing concerns related to shifting evaluation criteria and cherry-picking results. By systematically subjecting images to a diverse set of corruptions, the dataset provides a comprehensive framework for benchmarking the performance of current deep learning systems. The findings from the evaluation underscore the considerable room for improvement in the robustness of models when confronted with the challenges presented by ImageNet-C.

B.4. Vision Transformer

Vaswani *et al.* (Vaswani et al., 2017) first introduces the Transformer, a revolutionary architecture solely based on attention mechanisms, showcasing superior performance in machine translation tasks. Transitioning to computer vision, Vision Transformer (ViT) (Dosovitskiy et al., 2020) challenges the convention of coupling attention with convolutional networks, proposing a direct sequence-based alternative based solely on attention mechanisms. ViT excels in image classification while demanding fewer computational resources. Swin Transformer (Liu et al., 2021) further refines transformer architecture

for vision tasks, introducing a hierarchical design and a shifted window approach, yielding state-of-the-art results in image classification, object detection, and semantic segmentation. While ViT requires extensive pre-training, Steiner *et al.* (Steiner et al., 2021) proposed a novel approach to minimize training costs. They conducts an empirical study on Vision Transformers, highlighting their competitive performance with augmented regularization and increased compute, even when trained on smaller datasets. DeiT (Touvron et al., 2021) relies on knowledge distillation to reduce training costs. They demonstrates competitive results with a teacher-student strategy and introducing token-based distillation for effective knowledge transfer in attention-based models.