Towards Evaluating the Robustness of Visual State Space Models

Hashmat Shadab Malik[1] Fahad Shamshad[1] Muzammal Naseer[2] Karthik Nandakumar[3]
Fahad Shahbaz Khan[1,4]
Salman Khan[1,5]

[1]Mohamed Bin Zayed University of AI, UAE
[2]Center of Secure Cyber-Physical Security Systems, Khalifa University, UAE
[3]Michigan State University
[4]Linköping University, Sweden
[5]Australian National University, Australia

{hashmat.malik, fahad.shamshad, fahad.khan, salman.khan}@mbzuai.ac.ae
muhammadmuzammal.naseer@ku.ac.ae
nandakum@msu.edu

Abstract

Vision State Space Models (VSSMs), a novel architecture that combines the strengths of recurrent neural networks and latent variable models, have demonstrated remarkable performance in visual perception tasks by efficiently capturing long-range dependencies and modeling complex visual dynamics. However, their robustness under natural and adversarial perturbations remains a critical concern. In this work, we present a comprehensive evaluation of VSSMs' robustness under various perturbation scenarios, including occlusions, image structure, common corruptions, and adversarial attacks, and compare their performance to well-established architectures such as transformers and Convolutional Neural Networks. Furthermore, we investigate the resilience of VSSMs to object-background compositional changes on sophisticated benchmarks designed to test model performance in complex visual scenes. We also assess their robustness on object detection and segmentation tasks using corrupted datasets that mimic real-world scenarios. To gain a deeper understanding of VSSMs' adversarial robustness, we conduct a frequency-based analysis of adversarial attacks, evaluating their performance against low-frequency and high-frequency perturbations. Our findings highlight the strengths and limitations of VSSMs in handling complex visual corruptions, offering valuable insights for future research. Our code and models are available on GitHub.

1. Introduction

Deep learning models such as Convolutional Neural Networks (CNNs) [24] and Vision Transformers [9] have achieved remarkable success across various visual perception tasks, including image classification, object detection, and semantic segmentation. However, their robustness across different distribution shifts of the data remains a significant concern for their deployment in security-critical applications. Several works [2, 18, 39, 50] have extensively evaluated the robustness of CNNs and Transformers against common corruptions, domain shifts, information drop, and adversarial attacks, highlighting that a model's design impacts its ability to handle adversarial and natural corruptions, with robustness varying across different architectures. This observation motivates us to investigate the robustness of the recently proposed Vision State-Space Models (VSSMs) [14, 29, 51], a novel architecture designed to efficiently capture long-range dependencies in visual data.

CNNs are particularly adept at extracting hierarchical image features due to their shared weights across features, which help in capturing local-level information. In contrast, transformer-based models employ an attention mechanism that captures global information, effectively increasing the model's receptive field [9]. This allows transformers to excel at modeling long-range dependencies. However, a significant drawback of transformers is their quadratic computational scaling with input size, which makes them computationally expensive for downstream tasks [32]. Recently, state space sequence models (SSMs) have been adapted from the natural language domain to vision tasks. Un-

like transformers, vision-based SSMs offer the capability to handle long-range dependencies while maintaining a linear computational cost, providing a more efficient alternative for vision applications [4, 14, 15, 26, 29, 51].

VSSMs, such as the VMamba [29] and the hybrid Mamba-Transformer variant MambaVision [16], have gained attention in the vision domain due to their impressive performance. These models offer a unique approach to managing spatial dependencies, which is critical for handling dynamic visual environments. Their ability to selectively adjust interactions between states promises enhanced adaptability, a trait that could be pivotal in improving resilience against perturbations. Given their potential in safety-critical applications such as autonomous vehicles, robotics, and healthcare, it is crucial to thoroughly assess the robustness of these models.

In this paper, we present a comprehensive analysis of the performance of VSSMs, Vision Transformers, and CNNs in handling various nuisances for classification, detection, and segmentation tasks, aiming to provide valuable insights into their robustness and suitability for real-world applications. Our evaluation is divided into three main parts, each addressing a crucial aspect of model robustness.

❶ **Occlusions and Information Loss:** We rigorously assess the robustness of pure and hybrid VSSMs against information loss along scanning directions and severe occlusions affecting foreground objects, non-salient background regions, and random patch drops at multiple levels. This analysis is crucial for understanding how well VSSMs can handle partial information drop and maintain performance despite occlusions. Additionally, we explore the sensitivity of VSSMs to the overall image structure and global composition through patch shuffling experiments, providing insights into their ability to capture global context.

Findings: *Our experiments reveal that ConvNext [31] and VSSM models are superior in handling of sequential information loss along the scanning direction compared to ViT and Swin models. In scenarios involving random, salient, and non-salient patch drops, VSSMs exhibit the highest overall robustness, although Swin models perform better under extreme information loss. Additionally, VSSM models show greater resilience to spatial structure disturbances caused by patch shuffling compared to Swin models.*

❷ **Common Corruptions:** We evaluate the robustness of VSSM-based classification models against common corruptions that mimic real-world scenarios. This includes both ***global corruptions*** such as noise, blur, weather, and digital-based corruptions at multiple intensity levels, and ***fine-grained corruptions*** like object attribute editing and background manipulations. Furthermore, we extend the evaluation to VSSM-based detection and segmentation models to demonstrate their robustness in dense prediction tasks. By testing the models under these diverse and challenging conditions, we aim to provide a comprehensive understanding of their resilience in real-world applications.

Findings: *For global corruptions, VSSM models experience the least average performance drop compared to Swin and ConvNext models. When subjected to fine-grained corruptions, the VSSM family outperforms all transformer-based variants and maintains performance that is either better than or comparable to the advanced ConvNext models. In dense prediction tasks such as detection and segmentation, VSSM-based models generally demonstrate greater resilience and outperform other models.*

❸ **Adversarial Attacks:** We analyze the robustness of VSSMs against adversarial attacks in both white-box and black-box settings. In addition to the standard adversarial evaluation, we conduct a frequency analysis to demonstrate the resilience of VSSM models against low-frequency and high-frequency adversarial perturbations. This analysis provides insights into VSSMs ability to withstand adversarial perturbations at different frequency levels.

Findings: *For adversarial attacks, smaller VSSM models exhibit higher robustness against white-box attacks than Swin Transformers, though this does not scale to larger VSSMs. VSSMs maintain over 90% robustness against low-frequency perturbations, even at high perturbation strengths, but degrade quickly under high-frequency attacks. Across standard attacks, VSSMs outperform ConvNext, ViT, and Swin models under smaller perturbation budgets. In adversarial fine-tuning, VSSMs excel in both clean and robust accuracy on high-resolution images, but ViT models outperform them on low-resolution datasets like CIFAR.*

Our findings reveal that VSSM models have both strengths and limitations in handling various nuisances and adversarial attacks. They also indicate that a variety of metrics is essential to fully evaluate the diverse capabilities of different architectures. While VSSMs often demonstrate superior robustness, ConvNext and ViT architectures occasionally outperform them. These insights can inform model selection for specific applications, considering robustness requirements in real-world scenarios.

2. Related Work

Robustness of Deep Learning Models: Robustness refers to a conventionally trained model's ability to maintain satisfactory performance under natural and adversarial distribution shifts [10, 12]. In practice, deep learning-based models often encounter various types of corruptions, such as noise, blur, compression artifacts, and adversarial pertur-

bations, which can significantly degrade their performance. To ensure the reliability and robustness of these models, it is essential to systematically evaluate their performance under such challenging conditions. Recent studies have investigated the robustness of deep learning-based models across a wide range of areas, including image classification [18, 20], semantic segmentation [22], object detection [37], video classification [46], point cloud processing [21], and transformer-based architectures [25, 40, 41]. However, there is a lack of similar investigations for vision state space models (VSSMs), despite their growing popularity and potential applications [6, 28, 45, 47]. In this work, we aim to bridge this gap by examining how the performance of VSSMs is affected by adversarial and common corruptions.Considering the increasing adoption of VSSMs, our findings can provide valuable insights for researchers and practitioners working on developing robust and reliable vision systems.

State Space Models: State space models (SSMs) [36, 43] have emerged as a promising method for modeling sequential data in deep learning. These models map a 1-dimensional sequence $x(t) \in \mathbb{R}^L$ to $y(t) \in \mathbb{R}^L$ via an implicit latent state $h(t) \in \mathbb{R}^N$ as:

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

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 governing the dynamics and output mapping. To enhance computational efficiency, the continuous SSM is discretized using a zero-order hold assumption, leading to a discretized form:

$$h_t = \overline{\boldsymbol{A}}h_{t-1} + \overline{\boldsymbol{B}}x_t, \qquad y_t = \boldsymbol{C}h_t, \tag{2}$$

where $\overline{\boldsymbol{A}}$ and $\overline{\boldsymbol{B}}$ are the discrete counterparts of $\boldsymbol{A}$ and $\boldsymbol{B}$, obtained via a specified sampling timescale $\boldsymbol{\Delta} \in \mathbb{R}> 0$. The iterative process in Eq. 2 can be further expedited through parallel computation using a global convolution operation:

$$\boldsymbol{y} = \boldsymbol{x} \circledast \overline{\boldsymbol{K}}, \quad \text{with} \quad \overline{\boldsymbol{K}} = (\boldsymbol{C}\overline{\boldsymbol{B}}, \boldsymbol{C}\overline{\boldsymbol{A}\boldsymbol{B}}, ..., \boldsymbol{C}\overline{\boldsymbol{A}}^{L-1}\overline{\boldsymbol{B}}), \tag{3}$$

where $\overline{\boldsymbol{K}} \in \mathbb{R}^L$ is the kernel used and $\circledast$ denotes convolution operator. Recent advancements in SSMs, like Mamba models [14], introduce dynamic, input-dependent parameterization for managing sequential state interactions. This inspired vision models such as VMamba [29] and MambaVision [16], which combine Mamba with ViT-like hierarchies. While these models have been studied for tasks like detection and segmentation, their robustness against natural and adversarial corruptions remains underexplored. We aim to evaluate their resilience to these perturbations, which is crucial for understanding their potential in real-world applications and identifying areas for improvement.

3. Robustness of Vision State Space Models

We have broadly categorized the experiments into natural and adversarial corruption categories to evaluate the robustness of CNNs, transformers, and VSSMs across tiny, small, and base-model families. For natural corruptions, we conduct experiments on classification, detection, and segmentation tasks. For the classification task, we employ the recent ConvNext model [31] as a representative of the CNN family, while selecting the ViT [9] and Swin [30] family models for transformer architectures. For VSSMs, we report results on pure VMamba *v2* pretrained models [29] and hybrid MambaVision models [16]. For detection and segmentation, we report results using ImageNet-pretrained backbones of the specified models. These models are fine-tuned with the MMDetection [5] and MMSegmentation [7] frameworks. For detection, we utilize Mask-RCNN [17], and for segmentation, we use UperNet [44] as the network architecture. To evaluate the robustness of VSSMs against adversarial attacks, we consider imagenet trained classification models. Furthermore, we use imagenet pretrained models for adversarial fine-tuning on two downstream datasets; CIFAR-10 [23] and Imagenette [1]. Evaluations are done on 224×224 images for classification (except 32×32 for CIFAR-10), 800×1216 for detection, and 512×512 for segmentation.

3.1. Robustness against Natural Corruptions

We categorize natural corruptions into **information drop** and **ImageNet-based corruption benchmarks**. Information drop experiments assess models' robustness against various patch perturbations, such as occlusions, random patch drops, and patch shuffling, assessing their ability to handle partial information loss and local distortions. All information drop experiments are conducted on 5000 images from the ImageNet validation set, following [38]. ImageNet-based corruptions mimic real-world issues such as noise, blur, weather, and digital corruptions at various intensities. We also evaluate **VSSM-based detectors and segmentation models** under these natural corruptions.

3.1.1. Robustness against Information Drop

Information Drop along the Scanning Axis: VSSM models scan image patches sequentially along four paths (top-bottom, bottom-top, left-right, right-left) to capture spatial information. To study the effectiveness of this 2D-Selective Scan operation, we investigate the models' response to information drop along these scanning directions. We consider two settings: (1) linearly increasing information drop along the scanning direction, with maximum drop in last-scanned patches, and (2) linearly increasing drop from start to the center, then linearly decreasing until the end.

We split the image into $n \times n$ non-overlapping patches and perform the information drop experiments, with the

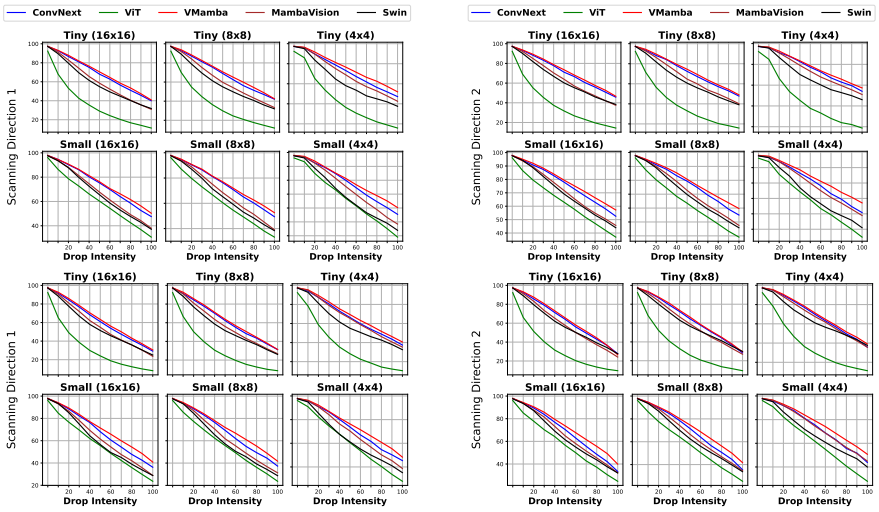

Figure 1. Information drop of Tiny and Small family of models along the scanning direction: the image is split into a sequence of fixed-size non-overlapping patches of size 16x16, 8x8, and 4x4. The first row shows the results of linearly increasing the number of pixels dropped from each patch to the maximum threshold (*Drop Intensity*) along different scanning directions. The bottom row presents results of linearly increasing the number of pixels dropped from each patch to the maximum threshold (*Drop Intensity*) until the center of the scanning direction. More detailed analysis is provided in Section A.1 of the Appendix.

maximum drop in a patch (*Drop Intensity*) varying from 10% to 100%. In Fig. 1, we report results on information drop along top-to-bottom *(Direction 1)* and left-to-right *(Direction 2)* scanning directions. For both settings (1) and (2) we observe that VMamba and ConvNext models show high robustness to sequential information drop across various thresholds. Overall, *'T'* and *'S'* versions of VMmamba model demonstrate superior performance compared to their counterparts across different patch sizes. Pure transformer-based ViT models show poor performance in this experimental setup. We also observe that as we reduce the patch size for splitting the image, leading to a gradual loss of information in the scanning direction, the performance of all the models improves. This implies that handling more abrupt information loss in fewer and larger patches is challenging for these models. In Section A.1 of Appendix, we expand the analysis across base models and all the scanning directions with varying number of patch sizes. ***Overall, we observe that both VMamba and ConvNext are more adept at handling sequential drop of information along different scanning directions, compared to hybrid MambaVision, ViT and Swin Models.***

Random Patch Drop: We assess the robustness of VSSMs in occluded scenarios by randomly dropping patches from images. We split the image into $n \times n$ patches and randomly select the patches whose values will be set to zero. As shown in Tab.1 (*top*), when image is split into 16×16 patches, VSSMs consistently outperform MambaVision, ResNet, ConvNeXt, and ViT models in maintaining accuracy with increasing numbers of dropped patches. However, under conditions of extreme spatial information loss, Swin models demonstrate superior performance, whereas the hybrid architecture-based MambaVision performs worst. ***This trend persists when the image is split into*** 8×8 ***patches, as illustrated in Tab.1 (bottom), highlighting the robustness of VSSMs and the exceptional resilience of Swin models in extreme patch drop scenarios.*** Results on different patch sizes showing similar trend are reported in Appendix A.2.

Salient and Non-Salient Patch Drop: We evaluate the robustness of VSSMs against salient (foreground) and non-salient (background) patch drop. Using a self-supervised ViT model, DINO [3], we effectively segment salient objects by exploiting the spatial positions of information flowing into the final feature vector within the last attention block. This allows us to control the amount of salient information captured within the selected patches by thresholding. Similarly, we also select the least salient regions of the image and drop the patches containing the lowest foreground information [39]. Similar to [39], the patch size is fixed to 16×16 for this experiment. Tab. 2 *(top)* shows that ***VSSM models, including VMamba and MambaVision, demonstrate notable robustness when foreground content is removed, outperforming both convolutional (ResNet and ConvNeXt) and the ViT transformer family. Their performance is on par with the Swin family until a 50%*** ***salient patch drop, beyond which Swin transformers exhibit better robustness***, maintaining higher accuracy compared to VSSMs. The trend for non-salient patch drops is similar and is shown in the Appendix A.3.

Patch Shuffling: VSSMs process images as a sequence of patches, and the order of these patches represents the overall image structure and global composition. To evaluate the robustness of VSSMs to patch permutations, we define a shuffling operation on the input image patches, which destroys the image structure by changing the order of the patch sequence. Based on the dimensions of the patch size, we split the image into either $4, 8, 16, 32, 64$ and 256 patches, which is then followed by random shuffling of the patches to evaluate the performance of models on spatial rearrangement of information. In Tab. 2 *(bottom)*, we observe that VMamba family overall performs better than other models when the spatial structure of the input image is disturbed. ***VMamba models generally demonstrate greater resilience to spatial structure disturbances than Swin models.***

Table 1. Top-1 classification accuracy of various architectures under random patch drop occlusions.

ResNet-50	ConvNext-T	ConvNext-S	ConvNext-B	ViT-T	ViT-S	ViT-B	VMamba-T	VMamba-S	VMamba-B	MambaVision-T	MambaVision-S	MambaVision-B	Swin-T	Swin-S	Swin-B
Patch Size 16 × 16 *(Percentage of patch drop increasing from top to bottom (10% to 90%))*															
96.70	97.24	97.78	97.84	92.30	96.08	97.54	97.38	97.94	97.96	97.36	97.82	97.60	97.24	97.60	97.60
75.27	96.49	97.19	97.37	90.83	95.39	96.85	96.49	96.61	97.25	96.24	96.80	97.09	96.76	97.38	97.32
39.93	94.63	95.27	96.48	88.09	94.29	96.27	95.16	92.97	96.25	92.91	94.67	95.74	96.11	96.84	96.79
17.91	89.99	91.12	95.29	85.26	92.35	95.08	93.45	89.74	95.21	87.14	91.19	93.25	94.88	95.88	96.17
6.73	81.43	84.63	93.03	80.08	90.15	92.78	90.52	84.82	93.46	77.02	84.56	88.05	93.38	94.35	95.03
2.43	70.07	74.44	88.76	72.49	85.10	89.21	86.52	78.41	90.89	61.79	76.07	78.92	91.05	92.21	93.25
1.05	57.59	60.15	82.35	61.34	76.56	82.08	80.52	68.40	87.03	42.16	62.44	63.89	87.96	87.84	90.43
0.56	44.67	44.71	71.25	45.63	62.63	70.25	70.39	52.72	79.96	20.73	42.16	43.33	80.65	79.71	84.70
0.45	31.29	28.82	57.71	25.86	41.73	50.08	56.23	34.51	67.56	5.62	21.07	22.21	70.37	66.60	74.38
0.43	16.73	14.98	33.67	7.85	15.86	19.68	34.83	16.82	41.55	1.95	7.08	11.08	47.16	47.85	53.54
Patch Size 8 × 8 *(Percentage of patch drop increasing from top to bottom (10% to 90%))*															
96.70	97.24	97.78	97.84	92.32	96.08	97.54	97.38	97.94	97.96	97.36	97.82	97.60	97.24	97.60	97.60
44.91	86.69	91.39	95.63	70.18	88.90	94.23	87.86	85.93	90.20	89.57	91.25	92.32	96.44	96.77	96.76
12.44	68.38	81.13	90.83	42.19	79.72	88.37	79.91	78.23	84.43	73.34	79.03	84.32	95.04	94.87	95.84
3.79	55.12	68.35	84.08	16.91	65.39	76.63	70.40	70.95	78.47	51.09	60.21	71.70	92.87	92.08	94.49
1.35	39.05	54.58	73.51	4.59	46.17	60.12	57.34	59.09	70.04	29.64	39.81	55.01	90.13	88.21	92.40
0.56	23.89	37.94	58.05	1.25	25.91	39.97	43.09	44.08	58.29	13.73	23.86	34.67	85.40	81.87	88.36
0.33	13.33	21.97	40.37	0.45	11.03	21.56	28.25	27.85	43.00	4.90	11.86	16.70	78.76	71.63	81.85
0.21	5.95	9.85	21.51	0.21	3.78	8.85	14.69	12.45	25.75	1.51	4.59	5.90	68.40	53.81	70.03
0.24	2.08	2.31	7.85	0.14	1.02	2.49	5.11	2.45	10.07	0.50	1.06	1.16	52.41	29.58	49.35
0.25	0.46	0.56	1.33	0.16	0.39	0.59	0.75	0.43	1.65	0.22	0.31	0.21	28.46	8.32	23.67

Table 2. Top-1 classification accuracy reported under salient patch drop occlusion and patch shuffling.

ResNet-50	ConvNext-T	ConvNext-S	ConvNext-B	ViT-T	ViT-S	ViT-B	VMamba-T	VMamba-S	VMamba-B	MambaVision-T	MambaVision-S	MambaVision-B	Swin-T	Swin-S	Swin-B
Salient Patch Drop *(Percentage of patch drop from top to bottom(10% to 100%))*															
92.70	96.88	97.38	97.50	90.46	95.36	94.90	97.04	97.66	97.40	96.88	97.24	97.20	96.94	97.44	97.26
84.86	95.98	96.98	96.86	88.62	94.56	93.92	96.58	96.98	97.14	95.78	96.30	96.62	96.14	97.02	96.90
73.64	94.60	95.86	96.52	85.40	92.70	92.52	95.34	95.82	96.74	93.56	94.82	95.34	95.02	96.22	96.24
60.16	92.52	94.04	94.92	80.94	89.82	89.72	93.88	93.14	95.04	90.08	91.88	93.62	93.28	94.92	94.76
44.78	88.22	90.06	92.60	75.12	85.26	85.34	91.04	89.64	92.74	85.00	87.90	89.84	90.24	92.74	92.98
29.16	81.40	84.02	88.64	65.42	78.06	78.36	86.34	83.66	89.30	76.24	80.32	83.98	86.76	88.88	89.78
16.12	72.32	74.64	82.36	51.60	67.60	68.34	79.24	74.26	83.44	63.44	70.16	74.80	80.32	82.08	83.54
7.38	56.96	58.28	69.44	35.10	50.18	50.80	66.42	59.26	72.88	45.52	52.86	59.08	68.84	71.18	73.30
2.14	33.80	34.96	46.52	14.96	24.26	24.72	43.90	35.38	50.94	22.26	29.24	35.12	46.42	49.58	51.80
0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10
Patch Shuffling *(From top to bottom, the image is split into* 4, 8, 16, 32, 64, *and* 256 *patches)*															
90.79	95.62	96.57	96.69	85.71	92.88	95.18	95.75	96.73	96.88	95.71	96.33	96.75	95.41	96.05	96.31
82.01	93.61	94.81	95.43	78.67	88.81	92.55	94.66	95.75	96.12	93.94	95.15	95.69	93.99	94.55	95.25
67.09	89.99	91.66	93.05	67.95	81.22	87.88	91.65	93.26	94.27	91.12	92.40	93.48	90.69	91.93	92.77
33.34	80.29	82.06	85.31	42.65	63.97	75.68	85.34	87.10	90.08	80.51	84.34	86.98	85.01	86.33	88.18
10.93	63.33	64.75	69.44	19.05	45.21	55.52	73.72	73.68	80.79	63.59	68.87	73.87	76.60	75.46	80.10
1.07	7.57	5.45	8.02	1.00	3.14	5.00	15.45	13.83	22.50	3.50	5.09	4.62	25.79	12.83	21.08

3.1.2. Robustness against ImageNet Corruptions

To evaluate the robustness of VSSMs in real-world scenarios, we performed experiments on several corrupted versions of the ImageNet dataset, which can be grouped into two categories based on the type of changes they introduce to the images. The **first category** includes datasets that make overall global compositional changes, such as ***ImageNet-C*** [18], which evaluates robustness against 19 common distortions across five categories (noise, blur, weather, and digital-based corruptions) with varying severity levels. Additionally, datasets like ***ImageNetV2*** [42], ***ImageNet-A*** [20], ***ImageNet-R*** [19], and ***ImageNet-S*** [11] also fall into this category, as they introduce domain shifts that affect the global composition of the images. For details of these datasets, see Appendix B.1. The **second category** consists of datasets that provide more control in editing the images and focus on fine-grained details. ***ImageNet-E*** [27] assesses classifiers' robustness to changes in background, object size, position, and direction, while ***ImageNet-B*** [35] introduces diverse object-to-background changes using text-to-image, image-to-text, and image-to-segment models, preserving original object semantics while varying backgrounds to include both natural and adversarial changes. These datasets use diffusion models to generate various object-to-background compositional changes to test the resilience of the models.

Robustness against Global Corruptions: The performance of various models on ImageNet-C is reported in terms of the mean corruption error (mCE) in Tab. 3. The mCE (*lower is better*) represents the average error of the

Table 3. Top-1 classification accuracy for the domain generalization setting across various architectures and datasets. Models trained on ImageNet are evaluated on datasets with domain shifts.

Model ↓	ImageNet	OOD Average(↑)	mCE(↓)
ConvNext-T	81.87	67.55	36.90(-44.97)
ViT-T	75.35	153.7	26.88(-48.47)
Swin-T	80.91	77.86	34.00(-46.91)
VMamba-T	82.28	65.05	37.32(-44.96)
MambaVision-T	82.10	66.58	37.66(-44.44)
ConvNext-S	82.82	59.52	39.81(-43.01)
ViT-S	81.40	79.28 v	36.74(-44.66)
Swin-S	82.90	63.04	37.81(-45.09)
VMamba-S	83.48	53.08	40.55(-42.93)
MambaVision-S	83.22	50.50	39.52(-43.70)
ConvNext-B	83.75	56.92	41.68(-42.07)
ViT-B	84.40	42.26	45.49(-38.91)
Swin-B	83.08	64.98	38.97(-44.11)
VMamba-B	83.76	53.76	41.52(-42.24)
MambaVision-B	83.96	49.61	41.98(-41.98)

model across various common corruptions at multiple intensity levels, normalized by the ResNet-50's standard accuracy. VMamba-T has the lowest mCE among all the *'T'* versions, followed by MambaVision-S, which performs significantly better than its *'S'* counterparts. However, among the *'B'* family, ViT-B achieves the lowest mCE score, , possibly due to ImageNet21k pretraining and ImageNet 1k finetuning. The table also shows top-1 accuracy on domain-shifted datasets (ImageNetV2, ImageNet-A, ImageNet-R, ImageNet-S). *'T'* and *'S'* versions of VMamba and MambaVision demonstrate the least average performance drop compared to Swin and ConvNext counterparts. For *'B'* models, ViT performs best, likely due to ImageNet21k pretraining. For more results, see Appendix B.1. Furthermore, in Appendix B.2, we report Expected Calibration Error(ECE) of models across the mentioned datasets to quantify the reliability of model's predicted confidence levels, accompanied by reliability diagrams for visualization.

Robustness against Fine-grained Corruptions: Tab. 4 (left) shows model robustness on ImageNet-E against variations in background complexity, object size, and positioning. Lower λ values indicate low background texture complexity, higher values indicate high complexity, with $\lambda = 20$ (adv) representing adversarially optimized high texture complexity. Objects are resized to (0.1, 0.08, 0.05) of original size and randomly placed. VSSM models (VMamba and MambaVision) demonstrate higher resilience to background changes compared to Swin and ViT transformer families, and perform comparably to ConvNeXt models. As background complexity increases or object size decreases, all models' performance declines, but the drop is less significant for VSSM and ConvNeXt models, showing their robustness to object size variations. Similar trends are observed for ImageNet-B (Tab. 4, right), where backgrounds are modified using a diffusion model with textual guidance (class/caption information) or color/texture prompts. ***The VSSM family, including VMamba and MambaVision, demonstrates superior performance compared to all Transformer-based variants and maintains performance better or comparable to the advanced ConvNext models. Pure VMamba and hybrid MambaVision exhibit comparable robustness against ImageNet corruption when no information is dropped.*** For further analysis across more models, see Appendix B.3.

3.1.3. Robustness on Object Detection

We evaluate VSSM, transformer, and CNN robustness for object detection using COCO-O [8], COCO-DC [35], and COCO-C datasets. COCO-O includes 6,782 images with 26,624 bounding boxes across six domains (sketch, cartoon, painting, weather, handmake, tattoo). COCO-DC contains 1,127 COCO 2017 validation images with diffusion model-induced background changes. COCO-C applies ImageNet-C style corruptions to the COCO-2017 evaluation set at various intensities. Performance is assessed using Average Precision (AP) and AP by object sizes (APs, APm, APl) on five models: ConvNext-T, Swin-T, Swin-S, VMamba-T, and VMamba-S.

Fig. 2 shows VMamba-S and VMamba-T consistently outperforming other architectures in most COCO-O scenarios, leading in original validation and out-of-distribution domains (cartoon, painting, sketch, weather). However, all models struggle with the tattoo domain. On average, VMamba-S and VMamba-T achieve AP scores of 42.2% and 41.1%, surpassing Swin and ConvNext models. Fig. 3 demonstrates VMamba models' superior robustness across various common corruptions in object detection, with even VMamba-T outperforming larger Swin-S in most scenarios. All models show performance drops with *'Glass Blur'* and *'Zoom Blur'* corruptions. For further details and results on COCO-DC, see Appendix B.4.

3.1.4. Robustness on Semantic Segmentation

We assess segmentation model robustness using 2,000 images from the ADE20K [49] validation set, corrupted with ImageNet-C [18] at various intensities. Performance is evaluated using mean Intersection over Union (mIoU).

Fig. 3 shows VMamba-T and VMamba-S consistently outperforming Swin counterparts across various ImageNet-C corruptions in segmentation, mirroring trends seen in object detection. The high performance of VMamba models on the original dataset also transfers effectively to the corrupted version of the dataset. Notably, VMamba-T surpasses the larger Swin-S in most corruption scenarios. Additional results are provided in Appendix B.4.

3.2. Robustness against Adversarial Attacks

In this section, we evaluate the robustness of VSSMs against spatial and frequency-based adversarial attacks. We

Table 4. Top-1 classification accuracy of various architectures on the ImageNet-E dataset [27] (*left*) and ImageNet-B dataset [35] (*right*).

Dataset →	**ImageNet-E**									**ImageNet-B**				
Model ↓	$\lambda = -20$	$\lambda = 20$	$\lambda = 20$(adv)	Random-BG	0.1	0.08	0.05	Random Pos.	Original	Original	Caption	Class	Color	Texture
ResNet-50	88.74	86.76	73.02	84.05	89.19	86.60	77.34	73.30	94.55	98.60	94.00	96.60	88.20	85.70
ConvNext-T	90.95	90.03	76.88	88.09	93.01	90.87	83.09	80.19	96.09	98.20	93.20	95.10	88.80	87.40
ConvNext-S	91.96	90.76	78.52	88.99	93.61	91.66	85.34	82.19	96.07	98.80	94.00	96.70	90.70	89.60
ConvNext-B	92.30	91.52	80.44	90.00	93.91	93.01	86.65	83.75	96.41	99.20	93.60	96.40	90.60	91.40
ViT-T	80.81	77.07	46.78	69.07	81.06	76.55	64.13	57.86	91.08	95.20	85.50	90.40	67.30	64.50
ViT-S	86.77	83.46	63.19	80.58	87.98	84.05	74.29	69.94	94.74	97.70	89.20	94.30	84.20	80.60
ViT-B	90.07	87.48	71.28	84.88	91.01	88.64	79.99	76.42	95.66	98.00	90.40	93.80	86.20	84.80
VMamba-T	91.15	89.87	75.18	87.41	92.09	91.06	83.66	79.71	95.84	98.50	92.20	96.30	87.20	86.80
VMamba-S	92.03	90.79	76.15	88.81	93.22	92.25	85.57	81.89	96.37	99.20	94.10	97.40	90.90	89.50
VMamba-B	92.37	91.27	77.30	89.11	93.70	92.64	86.03	83.62	96.37	99.10	94.00	96.50	90.80	89.80
MambaVision-T	90.67	88.83	73.07	86.40	91.93	90.07	81.87	78.42	95.73	98.60	93.70	96.60	89.10	87.70
MambaVision-S	91.22	90.19	75.18	88.19	92.78	91.19	84.01	81.18	96.03	99.40	94.40	97.80	91.40	90.10
MambaVision-B	91.77	90.65	78.42	89.18	93.70	92.81	86.35	83.78	96.30	99.10	94.60	97.20	91.40	90.60
Swin-T	90.05	88.83	71.51	86.19	91.08	88.94	79.39	76.49	95.27	97.90	91.70	95.30	85.50	84.00
Swin-S	90.67	88.86	73.35	87.25	91.91	89.68	81.55	78.81	96.25	98.30	91.80	95.50	86.10	85.40
Swin-B	91.08	89.96	75.09	87.87	92.62	91.22	83.43	80.65	95.95	98.60	92.30	95.60	89.20	86.70

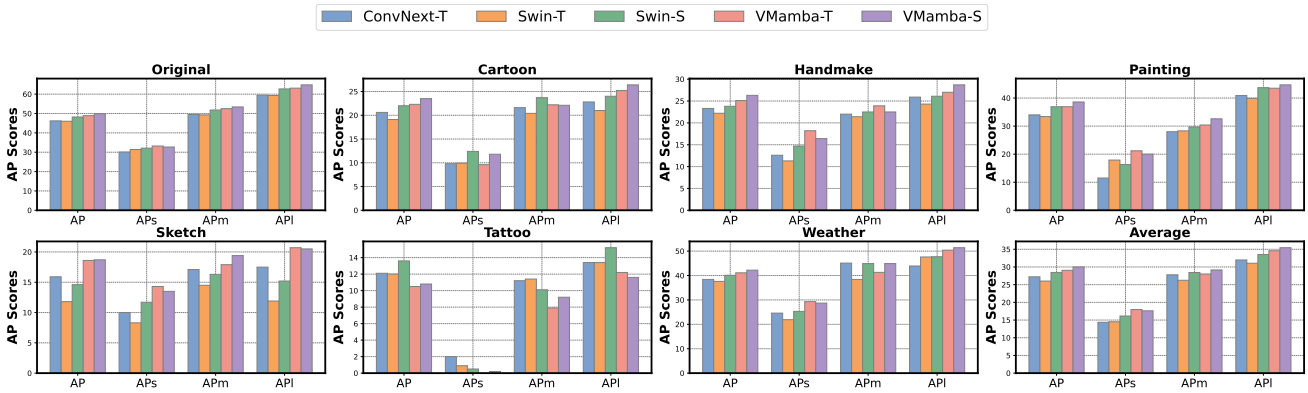

Figure 2. Average Precision (AP) scores for different architectures on the COCO-O dataset [8], detailing results for small (APs), medium (APm), and large objects (APl).

also compare the performance of adversarially fine-tuned VSSM models on CIFAR-10 [23] and Imagenette [1] downstream dataset with ViT and Swin models.

Adversarial Attacks in Spatial Domain: We conduct experiments in both white-box and black-box settings using the Fast Gradient Sign Method (FGSM) [13] and Projected Gradient Descent (PGD) [33] with an l_∞-norm and $\epsilon = 8/255$. FGSM is a single-step process, while PGD operates as a multistep method, iterating for 20 steps with a step size of $2/255$. Tab. 5 (*top*) displays the robust accuracy scores under white-box (diagonal entries) and black-box (off-diagonal entries) adversarial attacks for FGSM. ***'T'* and *'S'* versions of VMamba and MambaVision models exhibit higher white-box attack robustness compared to their Swin Transformer counterparts, but this pattern does not extend to the larger *'B'* models**. This indicates that VSSM's robustness advantage over Swin Transformers may not consistently scale with increased model size. For black-box settings, attacks transfer rate is high within the same architecture family than across different architectures. As expected, under stronger iteraive attack (PGD), all models' performance almost drops to zero in white-box settings. For the black-box transferability, we observe similar trends to FGSM attack (see Appendix C).

Adversarial Attacks in the Frequency Domain: We evaluate VMamba and MambaVision's robustness against frequency-specific PGD attacks and report the results in Appendix C. VMamba and MambaVision maintain above 90% robustness for low-frequency perturbations up to $\epsilon = 16$, comparable to ConvNext and Swin. For high-frequency attacks, all models' robustness decreases rapidly, with ViT models showing highest resilience. In standard full-frequency attacks, VSSM models display higher robustness than ConvNext, ViT, and Swin.

Adversarial Fine-tuning on Downstream Datasets: We adversarially fine-tune ImageNet pretrained VSSM, ViT, and Swin on downstream datasets using the TRADES [48] objective with varying robustness strength β and an l_∞

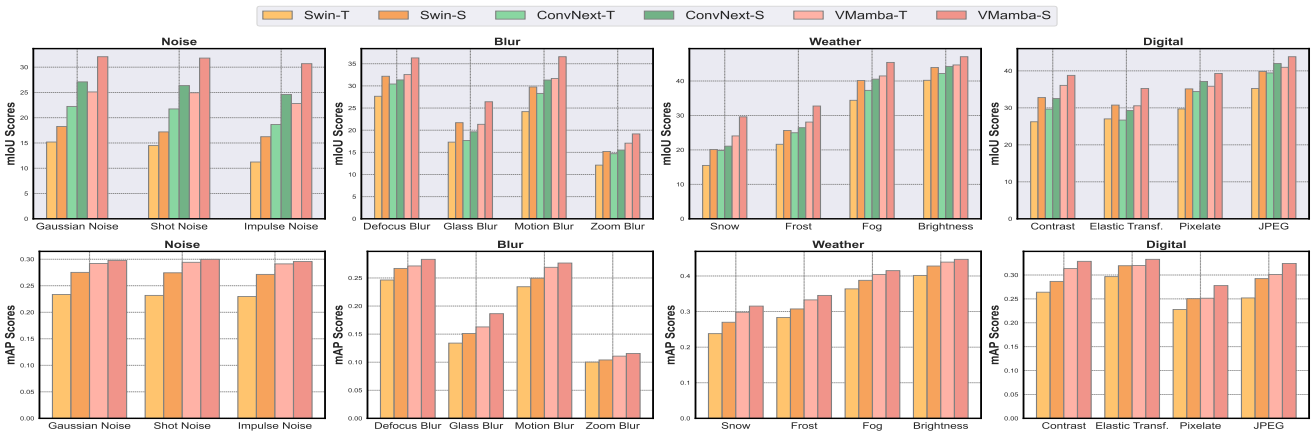

Figure 3. Performance comparison of different architectures on the AED20k-C and COCO-C datasets for segmentation (top) and detection (bottom) tasks. The top figure shows the Mean Intersection over Union (mIoU) score on the AED20k-C dataset, while the Mean Average Precision (mAP) score on the COCO-C dataset.

Table 5. Robust accuracy of models under white-box and black-box settings for FGSM attack. Adversarial examples crafted on surrogate models are used to evaluate robustness of target models.

Target →	VMamba-T	VMamba-S	VMamba-B	MambaVision-T	MambaVision-S	MambaVision-B	ResNet-50	ConvNext-T	ConvNext-S	ConvNext-B	ViT-T	ViT-S	ViT-B	Swin-T	Swin-S	Swin-B
Surrogate ↓	Fast Gradient Sign Method (FGSM) at $\epsilon = \frac{8}{255}$															
VMamba-T	42.90	66.34	65.10	74.44	73.00	73.16	80.20	72.66	73.80	72.84	79.46	83.64	86.94	72.22	76.38	74.60
VMamba-S	62.24	48.42	63.00	71.84	71.92	71.34	79.70	71.40	71.04	70.18	78.42	81.56	84.58	70.96	72.92	71.14
VMamba-B	65.52	66.96	51.24	75.22	73.82	73.08	81.22	73.54	73.20	72.32	79.12	83.62	86.24	73.24	76.30	73.88
MambaVision-T	77.74	79.04	79.78	46.18	73.30	75.64	82.06	79.50	80.88	81.82	78.30	84.20	87.86	77.78	81.24	80.80
MambaVision-S	77.86	79.56	79.24	74.96	53.42	73.84	83.60	80.40	81.14	81.26	79.90	85.64	88.52	78.50	82.40	81.88
MambaVision-B	75.90	77.84	77.00	75.46	71.96	52.68	83.46	80.24	80.52	79.92	79.82	85.18	88.12	78.46	81.60	80.30
ResNet-50	81.38	83.24	83.84	81.06	82.54	84.88	30.46	80.30	82.20	83.38	75.94	84.74	89.12	80.64	85.00	85.42
ConvNext-T	69.00	71.46	71.18	73.50	73.92	74.40	77.96	36.36	61.96	63.76	76.92	82.74	85.58	67.06	71.88	70.78
ConvNext-S	69.54	70.62	70.48	74.68	74.38	75.24	79.48	63.48	49.10	63.62	78.48	82.94	85.02	69.24	71.54	69.74
ConvNext-B	70.54	71.78	69.78	76.78	75.02	74.82	81.72	67.34	66.24	51.32	80.26	83.98	86.22	69.86	73.44	70.84
ViT-T	85.92	88.30	88.88	82.16	85.06	87.24	82.68	85.08	86.56	88.06	2.28	50.04	69.90	75.66	79.72	82.32
ViT-S	81.74	83.76	84.46	78.94	80.78	82.68	81.70	82.34	82.78	84.38	45.40	11.02	54.82	72.50	75.98	77.88
ViT-B	82.46	84.02	84.48	80.40	81.82	82.46	81.90	82.04	82.24	83.62	58.50	53.84	24.24	75.82	77.62	78.54
Swin-T	72.40	76.06	75.58	76.80	77.16	78.06	82.82	71.46	72.12	71.96	76.48	80.66	85.46	28.86	56.22	55.12
Swin-S	78.24	79.10	79.30	80.12	81.90	81.40	85.46	77.96	77.10	77.02	79.38	83.04	86.58	61.94	48.00	63.90
Swin-B	78.88	79.40	79.04	81.48	82.16	81.04	86.60	78.16	78.46	77.32	81.48	84.74	87.64	66.82	68.28	54.76

perturbation budget $\epsilon = \frac{8}{255}$. In Fig. 4, we plot the clean accuracy and robust accuracy under PGD-100 attack at $\epsilon = \frac{8}{255}$. On Imagennette [1], VSSM-T shows strong performance in both clean and robust accuracy across different β levels, followed closely by Swin-T, while ViT-T exhibits the lowest robustness. However, on the low-resolution CIFAR-10 dataset with significantly lower number of patches, ViT models perform better than Mamba-based VSSM models.

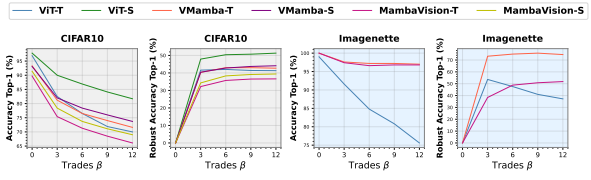

Figure 4. Clean and Robust accuracy of models evaluated on CIFAR-10 *(left)* and Imagenette *(right)*.

4. Conclusion

In conclusion, we present a comprehensive evaluation of the robustness of Vision State-Space Models (VSSMs) under diverse natural and adversarial image manipulations, highlighting both their strengths and weaknesses compared to transformers and CNNs. Through rigorous experiments, we demonstrated the capabilities and limitations of VSSM-based classifiers in handling occlusions, common corruptions, and adversarial attacks, as well as their resilience to object-background compositional changes in complex visual scenes. Additionally, we show that VSSM-based models are generally more robust to real-world corruptions in the dense prediction tasks, including detection and segmentation. As an early work in this area, Our findings provide insights into the robustness of Visual State Space Models across diverse settings, laying the groundwork for future research to improve the reliability of current visual perception systems that depend on these models.

References

[1] Imagenette. https://github.com/fastai/imagenette/. Accessed: 2024-05-10. 3, 7, 8

[2] Yutong Bai, Jieru Mei, Alan L Yuille, and Cihang Xie. Are transformers more robust than cnns? *Advances in neural information processing systems*, 34:26831–26843, 2021. 1

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 4

[4] Guo Chen, Yifei Huang, Jilan Xu, Baoqi Pei, Zhe Chen, Zhiqi Li, Jiahao Wang, Kunchang Li, Tong Lu, and Limin Wang. Video mamba suite: State space model as a versatile alternative for video understanding. *arXiv preprint arXiv:2403.09626*, 2024. 2

[5] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 3

[6] Keyan Chen, Bowen Chen, Chenyang Liu, Wenyuan Li, Zhengxia Zou, and Zhenwei Shi. Rsmamba: Remote sensing image classification with state space model. *arXiv preprint arXiv:2403.19654*, 2024. 3

[7] MMSegmentation Contributors. Mmsegmentation: Openmmlab semantic segmentation toolbox and benchmark, 2020. 3

[8] Edoardo Debenedetti, Vikash Sehwag, and Prateek Mittal. A light recipe to train robust vision transformers. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 225–253. IEEE, 2023. 6, 7

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 3

[10] Nathan Drenkow, Numair Sani, Ilya Shpitser, and Mathias Unberath. A systematic review of robustness in deep learning for computer vision: Mind the gap? *arXiv preprint arXiv:2112.00639*, 2021. 2

[11] Shanghua Gao, Zhong-Yu Li, Ming-Hsuan Yang, Ming-Ming Cheng, Junwei Han, and Philip Torr. Large-scale unsupervised semantic segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 2022. 5, 11

[12] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018. 2

[13] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 7

[14] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 1, 2, 3

[15] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021. 2

[16] Ali Hatamizadeh and Jan Kautz. Mambavision: A hybrid mamba-transformer vision backbone. *arXiv preprint arXiv:2407.08083*, 2024. 2, 3

[17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 3

[18] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 1, 3, 5, 6

[19] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, 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*, pages 8340–8349, 2021. 5, 11

[20] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15262–15271, 2021. 3, 5, 11

[21] Qiufan Ji, Lin Wang, Cong Shi, Shengshan Hu, Yingying Chen, and Lichao Sun. Benchmarking and analyzing robust point cloud recognition: Bag of tricks for defending adversarial examples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4295–4304, 2023. 3

[22] Christoph Kamann and Carsten Rother. Benchmarking the robustness of semantic segmentation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8828–8838, 2020. 3

[23] Alex Krizhevsky. Learning multiple layers of features from tiny images. *University of Toronto*, 2012. 3, 7

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 1

[25] Amandeep Kumar, Muzammal Naseer, Sanath Narayan, Rao Muhammad Anwer, Salman Khan, and Hisham Cholakkal. Multi-modal generation via cross-modal in-context learning. 2024. 3

[26] Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding. *arXiv preprint arXiv:2403.06977*, 2024. 2

[27] Xiaodan Li, Yuefeng Chen, Yao Zhu, Shuhui Wang, Rong Zhang, and Hui Xue. Imagenet-e: Benchmarking neural network robustness via attribute editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20371–20381, 2023. 5, 7, 27

[28] Xiao Liu, Chenxu Zhang, and Lei Zhang. Vision mamba: A comprehensive survey and taxonomy. *arXiv preprint arXiv:2405.04404*, 2024. 3

[29] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. Vmamba:

Visual state space model. *arXiv preprint arXiv:2401.10166*, 2024. 1, 2, 3, 11

[30] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 3

[31] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 2, 3

[32] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. *Advances in neural information processing systems*, 29, 2016. 1

[33] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 7

[34] Shishira R Maiya, Max Ehrlich, Vatsal Agarwal, Ser-Nam Lim, Tom Goldstein, and Abhinav Shrivastava. A frequency perspective of adversarial robustness. *arXiv preprint arXiv:2111.00861*, 2021. 12

[35] Hashmat Shadab Malik, Muhammad Huzaifa, Muzammal Naseer, Salman Khan, and Fahad Shahbaz Khan. Object-compose: Evaluating resilience of vision-based models on object-to-background compositional changes. *arXiv preprint arXiv:2403.04701*, 2024. 5, 6, 7, 27

[36] Harsh Mehta, Ankit Gupta, Ashok Cutkosky, and Behnam Neyshabur. Long range language modeling via gated state spaces. *arXiv preprint arXiv:2206.13947*, 2022. 3

[37] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*, 2019. 3

[38] Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Fahad Shahbaz Khan, and Fatih Porikli. On improving adversarial transferability of vision transformers. *arXiv preprint arXiv:2106.04169*, 2021. 3

[39] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. *Advances in Neural Information Processing Systems*, 34: 23296–23308, 2021. 1, 4

[40] Sayak Paul and Pin-Yu Chen. Vision transformers are robust learners. In *Proceedings of the AAAI conference on Artificial Intelligence*, pages 2071–2081, 2022. 3

[41] Francesco Pinto, Philip HS Torr, and Puneet K. Dokania. An impartial take to the cnn vs transformer robustness contest. In *European Conference on Computer Vision*, pages 466–480. Springer, 2022. 3

[42] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019. 5, 11

[43] Jue Wang, Wentao Zhu, Pichao Wang, Xiang Yu, Linda Liu, Mohamed Omar, and Raffay Hamid. Selective structured state-spaces for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6387–6397, 2023. 3

[44] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018. 3

[45] Rui Xu, Shu Yang, Yihui Wang, Bo Du, and Hao Chen. A survey on vision mamba: Models, applications and challenges. *arXiv preprint arXiv:2404.18861*, 2024. 3

[46] Chenyu Yi, Siyuan Yang, Haoliang Li, Yap-peng Tan, and Alex Kot. Benchmarking the robustness of spatial-temporal models against corruptions. *arXiv preprint arXiv:2110.06513*, 2021. 3

[47] Yubiao Yue and Zhenzhang Li. Medmamba: Vision mamba for medical image classification. *arXiv preprint arXiv:2403.03849*, 2024. 3

[48] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019. 7

[49] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 6

[50] Daquan Zhou, Zhiding Yu, Enze Xie, Chaowei Xiao, Animashree Anandkumar, Jiashi Feng, and Jose M Alvarez. Understanding the robustness in vision transformers. In *International Conference on Machine Learning*, pages 27378–27394. PMLR, 2022. 1

[51] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024. 1, 2