Investigating the Adversarial Robustness of Visual Mamba - A Comparative Study

Mathias Duedahl
KTH Royal Institute of Technology
duedahl@kth.se

Abstract

This study investigates the adversarial robustness of Visual Mamba (ViM) models, focusing on a comparative analysis of different model architectures. We evaluate the performance of ResNet and Vision Transformer (ViT) models under adversarial conditions, employing fine-tuning techniques and adversarial data augmentation and blur preprocessing. Our findings indicate that while ResNet initially appears more robust, requiring substantial image perturbations for misclassification, and ViT more sensitive, the fine-tuning process likely compromised the inherent robustness of both models. This suggests the need for a more comprehensive study, with greater computational resources, to better understand and enhance adversarial robustness in ViM models.

1 Introduction

In recent years, the transformer architecture has revolutionized the field of natural language processing and has extended its influence to computer vision with the advent of the Vision Transformer (ViT), which currently represents the pinnacle of state-of-the-art vision models [6]. A newer development in the realm of state-space models has given rise to the Mamba architecture [2]. Known for its efficiency in both inference and training rivaling that of the transformer, the Mamba architecture's adaptation to vision tasks has led to the creation of the Vision Mamba model (ViM) [7]. Despite its potential, the ViM is still relatively unexplored, particularly in the context of its vulnerability to adversarial attacks.

Adversarial attacks are subtle modifications to a model's input designed to elicit incorrect outputs, posing significant risks in practical applications. Given the nascent stage of the ViM model, its adversarial robustness remains an open question. This study aims to fill this gap by conducting comparative experiments to assess the adversarial robustness of the ViM, ViT, and ResNet models. Additionally, we will explore their susceptibility to transfer attacks, providing insights into the security dimensions of these advanced visual models.

All projcet code is publicly available on Github at duedahl/AdversarialRobustness.

2 Background

While the adversarial robustness of Vision Mamba (ViM) models remains largely unexplored, research has demonstrated that Vision Transformer (ViT) models exhibit superior robustness compared to MLP-Mixer and CNNs [4]. In the context of adversarial attacks, robustness quantifies how much an input image must be altered to cause misclassification. To this end, the L^2-norm, its closely related Mean Squared Error (MSE), and the L^∞ norm are commonly utilized. These metrics are even sufficiently well understood to allow for provable robustness guarantees in the framework of L^p-norms [3].

KTH Royal Institute of Technology - DD2424 Project Report (2024).

These norms measure absolute changes in the image. To bridge the gap between these changes and human perception, which is central to computer vision tasks, we incorporate a metric that quantifies perceptual impact. We will use the following three metrics to assess adversarial robustness:

1. Mean Squared Error (MSE) - quantifies the average squared differences between the pixels of the original and perturbed images, providing insights into the intensity of distortions.
2. L^∞-norm - measures the maximum alteration in any single pixel, highlighting the worst-case scenario for perturbations.
3. Structural Similarity Index Measure (SSIM) - evaluates changes in luminance, contrast, and structure, elements crucial for mimicking human visual perception, with $SSIM = 1$ meaning perfect similarity, and $SSIM = 0$ meaning no similarity [5].

To perform the adversarial attacks, we make use of the well known Fast Gradient Sign Method (FGSM) [1] to move images toward a higher loss region of feature space with the goal of causing misclassification whilst still representing the same object.

3 Data

In this project, we employ three foundational models, each trained on the ImageNet-1K dataset, as outlined in Table 1. To evaluate the robustness of these models, we utilize ImageNette, a smaller subset of ImageNet-1K, ensuring that the data remains within the same domain.

Table 1: Overview of models.

Nickname	Name	Description	Parameters (M)
Vit	google/vit-base-patch16-224	Vision transformer	86.6
ResNet	microsoft/resnet-152	Residual network	60.3
Vim	hustvl/Vim-small-midclstok	Vision mamba	26

For subsequent domain adaptation, we select a subset from the PubFig dataset. We specifically extracted the 10 classes with the highest number of samples and partitioned them in a 40-40-10-10 split, creating two disjoint training sets. A comprehensive overview of this data configuration is presented in Table 2.

Table 2: Summary of Datasets

Subset Name	Number of Classes	Number of Samples	Subset of
ImageNet	1000	14M	
ImageNette	10	13K	ImageNet
ImageNette-Val	10	4K	ImageNette
PubFig	200	59K	
CelebTrainA	10	923	PubFig
CelebTrainB	10	923	PubFig
CelebVal	10	231	PubFig
CelebTest	10	230	PubFig

To establish a baseline for robustness, we applied the Fast Gradient Sign Method (FGSM) to generate the minimum perturbation adversarial examples for each image in ImageNette that was initially correctly classified by each model. The resulting adversarial datasets are detailed in Table 7 in appendix A.3.

4 Methods

4.1 Fine-tuning

To evaluate the models in a different domain from their original training, we adapt them to new classes in this new domain. We replace the last fully connected layer of each model with a linear

layer matching the number of classes, then train with all weights frozen, except for those in the last layer. All fine-tuning is performed with the hyperparameters shown in Table 5 in appendix A.1.

4.2 Adversarial Robustness

We assess the robustness of each model using least-perturbation FGSM attacks on correctly classified image samples from the dataset. We then calculate the difference between the adversarial and original images using the metrics detailed in Section 2.

4.3 Transfer Attack

To investigate susceptibility to transfer attacks, we fine-tune two copies, A and B, of each model on disjoint training data, then evaluate model A on adversarial examples generated from model B, simulating a black-box transfer attack scenario.

4.4 Adversarial Data Augmentation

To evaluate hardening potential, we re-fine-tune the models with adversarial examples added to the training pool as data augmentation. While adversarial training with regularization would be ideal, we used this simpler method due to time constraints.

4.5 Blur Preprocessing

To explore wrapper-type robustness augmentations, we apply a standard Gaussian blur as a preprocessing step before model inference, inspired by techniques such as denoised smoothing [3].

5 Experiments

5.1 Adversarial Robustness Baseline

To establish a baseline adversarial robustness for each of the base-models using the procedure outlined in section 4.2 with the ImageNette-Val dataset. See the distributions of resulting metrics in figure 1.

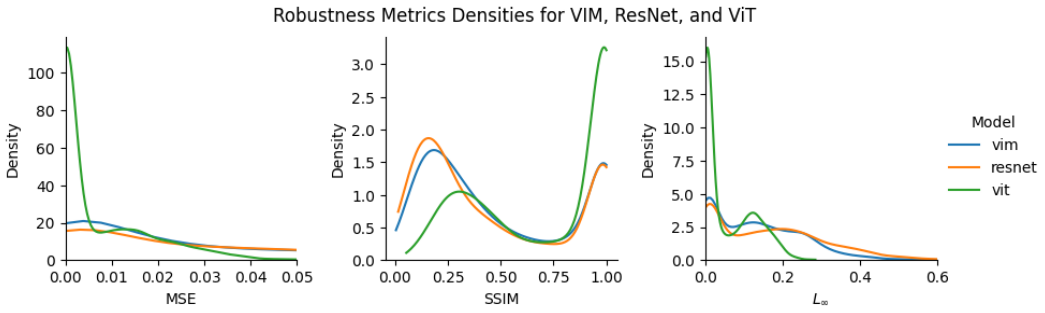

Figure 1: Baseline distributions of MSE, SSIM and L^∞ metrics for each model for minimum pertubation FGSM attack on ImageNette-Val. MSE and L^∞ are cropped for better visualization.

The baseline distributions in Figure 1 suggest that ResNet is the more robust model in its original form, as its metric distributions are the most skewed towards requiring significant image differences to cause misclassification. Conversely, ViT appears to be much more sensitive, requiring minimal changes to lead to misclassification. This indicates that the robustness of our models may be inversely proportional to their representational power.

5.2 Disjoint Fine-tuning

To prepare for the transfer attack we fine tuned two copies of each base model with the CelebTrainA and CelebTrainB dataset, the evolution of training metrics can be seen in figure 2, for a less cluttered view and final figures see appendix A.1.

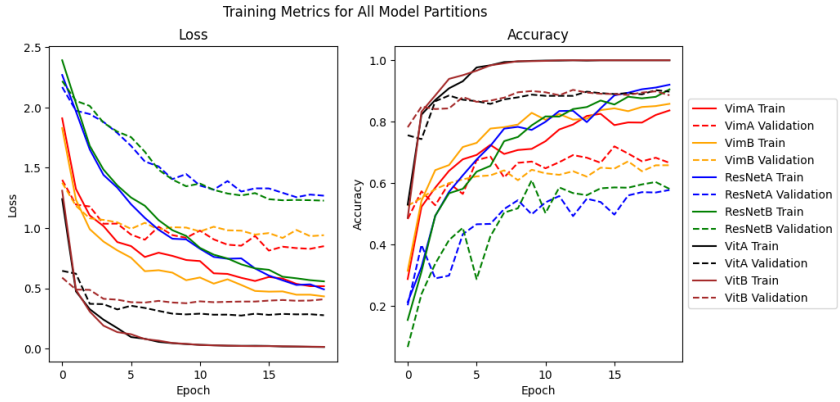

Figure 2: Training metrics during fine tuning of all models on CelebTrainA and CelebTrainB respectively.

5.3 Transfer Attack

Using adversarial examples generated for the models fine-tuned on disjoint datasets, we follow the procedure outlined in 4.3 and obtain the results seen in table 3.

Table 3: Transfer Attack accuracy and the base test accuracy for the different models.

Metric / Model	VimA	VitA	ResNetA
Base Accuracy	68.50	91.29	56.06
Transfer Attack Accuracy	12.41	20.09	38.02

Despite their differing representational powers, it is evident that ResNet, despite its lower baseline accuracy, is the least susceptible to transfer attacks. In contrast, ViT and ViM exhibit a significantly higher rate of misclassification when subjected to adversarial attack.

5.4 Adversarial Data Augmentation

Implementing the procedure as described in section 4.4, we obtain the training metrics of which can be seen on figure 3.

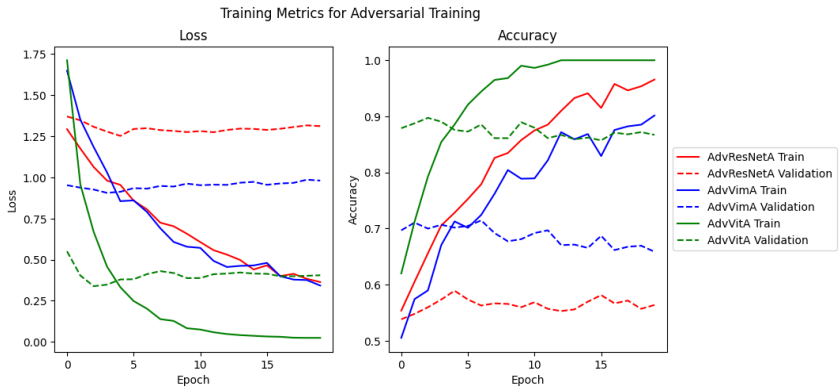

Figure 3: Training metrics during adversarial training of each suffix-A model.

Here we see that validation accuracy remains near unaffected under fine-tuning on the adversarial dataset, but with a slight decrease in accuracy.

5.5 Robustness Evolution

Performing a proper statistical analysis of the metrics is beyond the scope of this project. Descriptive statistics for the robustness metrics of each model are available in Table 6 in Appendix A.3. A visual overview of the metric distributions for the base model, adversarially trained model, and blur-preprocessed variant is shown in Figure 4.

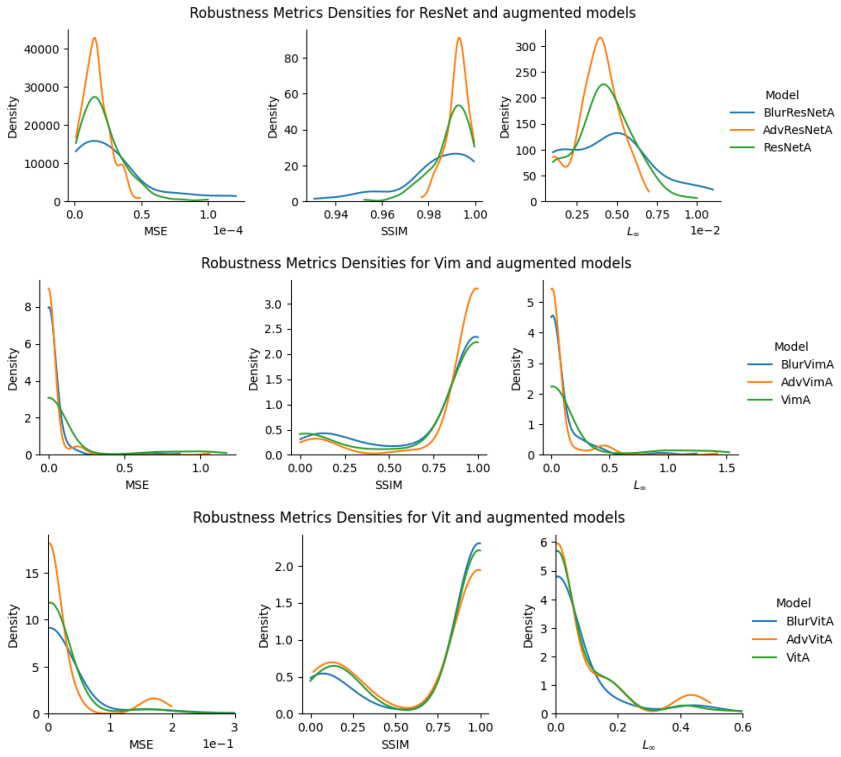

Figure 4: Kernel density estimation plots of the robustness metrics for the each model type and its robustness augmentation attempts.

Compared to the baseline distribution in Figure 1, the SSIM metric has become heavily right-skewed, indicating that the fine-tuning process compromised the base model's inherent robustness.

Interestingly, adversarial training shifts the distributions of all metrics towards lower image differences, except for SSIM in VitA, suggesting our adversarial training implementation is counterproductive, allowing for lower perturbation adversarial examples.

Blur pre-processing appears effective only for ResNet, shifting all metric distributions towards larger required perturbations.

Note that some adversarial generation attempts failed due to exceeding the maximum perturbation limit in our code. Failure rates for each model are listed in Table 8 in Appendix A.3.

To provide qualitative insights, we display the loss landscape and decision boundary around a single training example, in the plane spanned by the FGSM attack vector (v2) and a random direction (v1). See Figure 5 for ViM model visualizations, and Appendix A.2 for the other two.

Unfortunately these visualizations do not reveal a flattening of the loss landscape as we would have hoped to achieve after adversarial training, nor does the blur guarantee the sample image to be in a locally flat region.

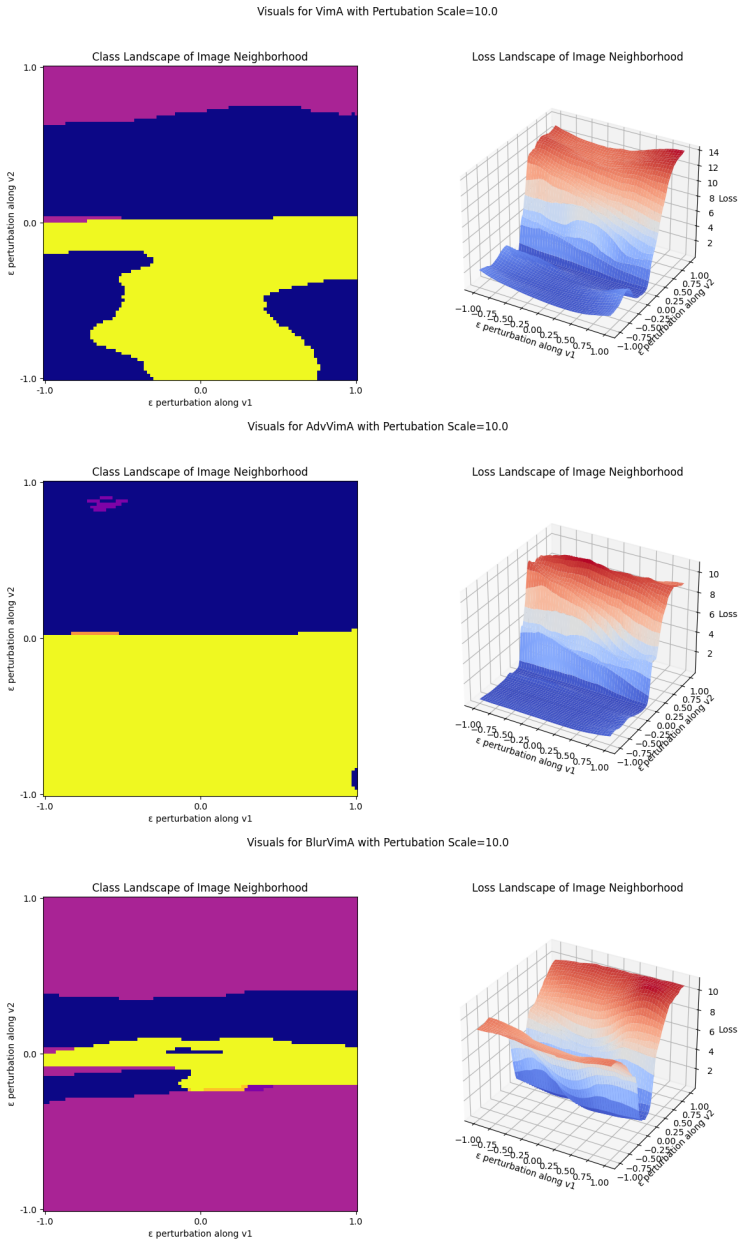

Figure 5: Decision boundary and loss landscape for Vim-based models in a neighborhood of a sample image, perturbations are made along a random unit vector (v1) and the normalized gradient of the unaugmented model (v2).

6 Conclusion

It is premature to draw definitive conclusions about these models based on our results, as the study's scope is too limited to produce generalizable findings. The substantial differences in robustness metric distributions between the baseline and our model adaptations suggest that the fine-tuning process may have compromised the models' inherent qualities, leading to unreliable results.

The question of the adversarial robustness of ViM models remains unresolved and requires a more comprehensive study. Future research should include sufficient computational resources to support adversarial training and large-scale robustness evaluations. Additionally, implementing denoised smoothing in place of blur preprocessing would be a valuable avenue for further investigation.

References

[1] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2015.

[2] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *ArXiv*, abs/2312.00752, 2023.

[3] Hadi Salman, Mingjie Sun, Greg Yang, Ashish Kapoor, and J. Zico Kolter. Denoised smoothing: A provable defense for pretrained classifiers. In *NeurIPS 2020*. ACM, September 2020.

[4] Rulin Shao, Zhouxing Shi, Jinfeng Yi, Pin-Yu Chen, and Cho-Jui Hsieh. On the adversarial robustness of vision transformers. *Trans. Mach. Learn. Res.*, 2022, 2022.

[5] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[6] Papers with Code. Image classification on imagenet. `https://paperswithcode.com/sota/image-classification-on-imagenet`. Accessed: 2024-05-23.

[7] 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*, abs/2401.09417, 2024.

A Appendix / supplemental material

A.1 Training Metrics

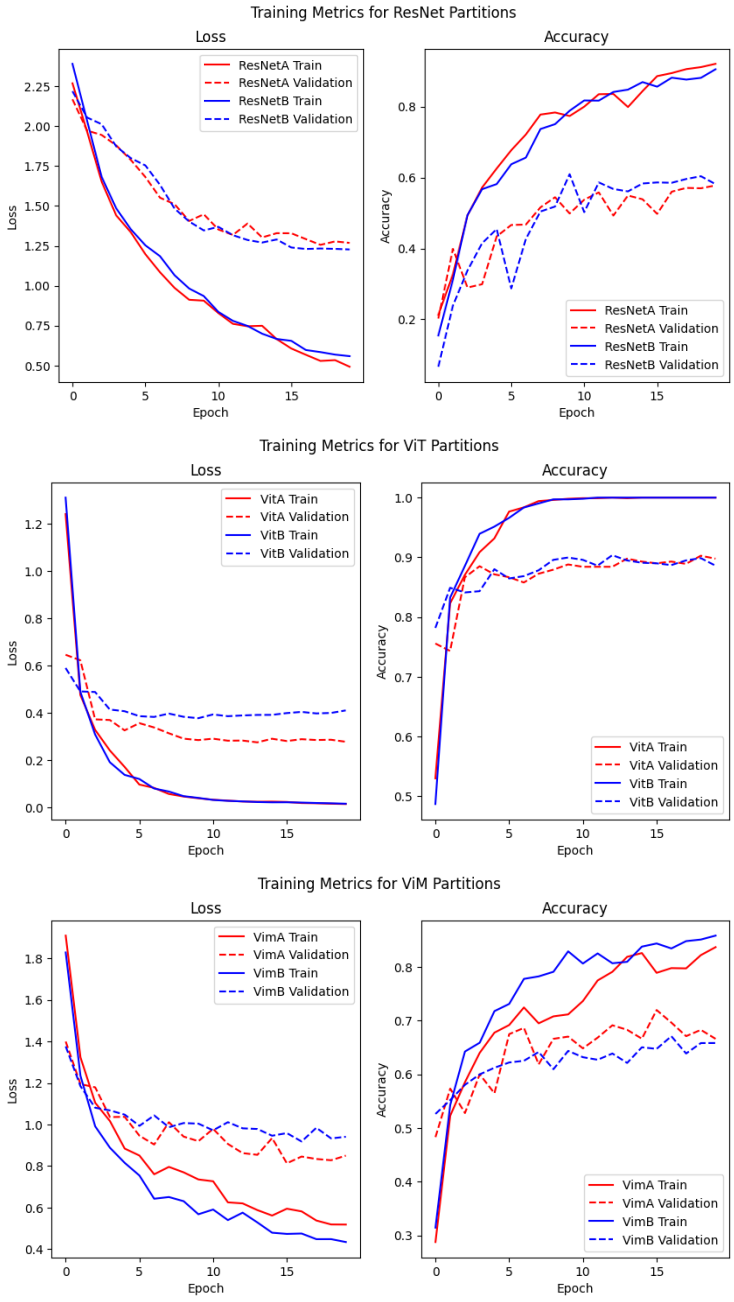

Figure 6: Training metrics during fine tuning of base models on the CelebTrainA and CelebTrainB datasets.

Table 4: Overview of fine-tuned models, suffix indicates the training partition it was fine-tuned on.

Name	Validation Accuracy	Test Accuracy
ResNetA	57.76	56.06
ResNetB	58.13	55.19
VitA	89.78	91.29
VitB	88.61	89.54
VimA	66.60	68.50
VimB	65.82	72.13
AdvResNetA	56.36	57.55
AdvVitA	86.70	87.89
AdvVimA	65.85	64.06

Table 5: Overview of fine-tuning hyper parameters.

Variant	Epochs	Optimizer	Learning Rate	Loss Function	Batch Size
Regular	20	Adam	1e-2	Cross Entropy	128
Adversarial	20	Adam	1e-3	Cross Entropy	128

A.2 Visualizations

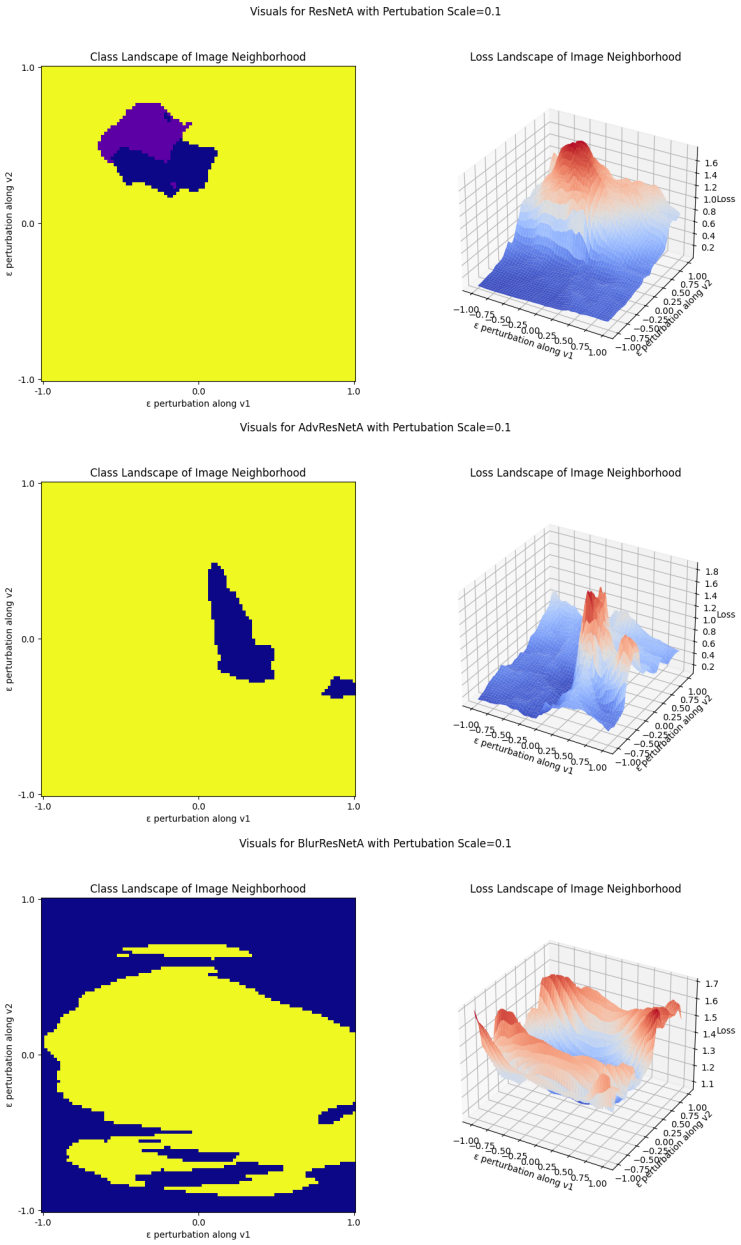

Figure 7: Decision boundary and loss landscape for ResNet-based models in a neighborhood of a sample image, perturbations are made along a random unit vector (v1) and the normalized gradient of the unaugmented model (v2).

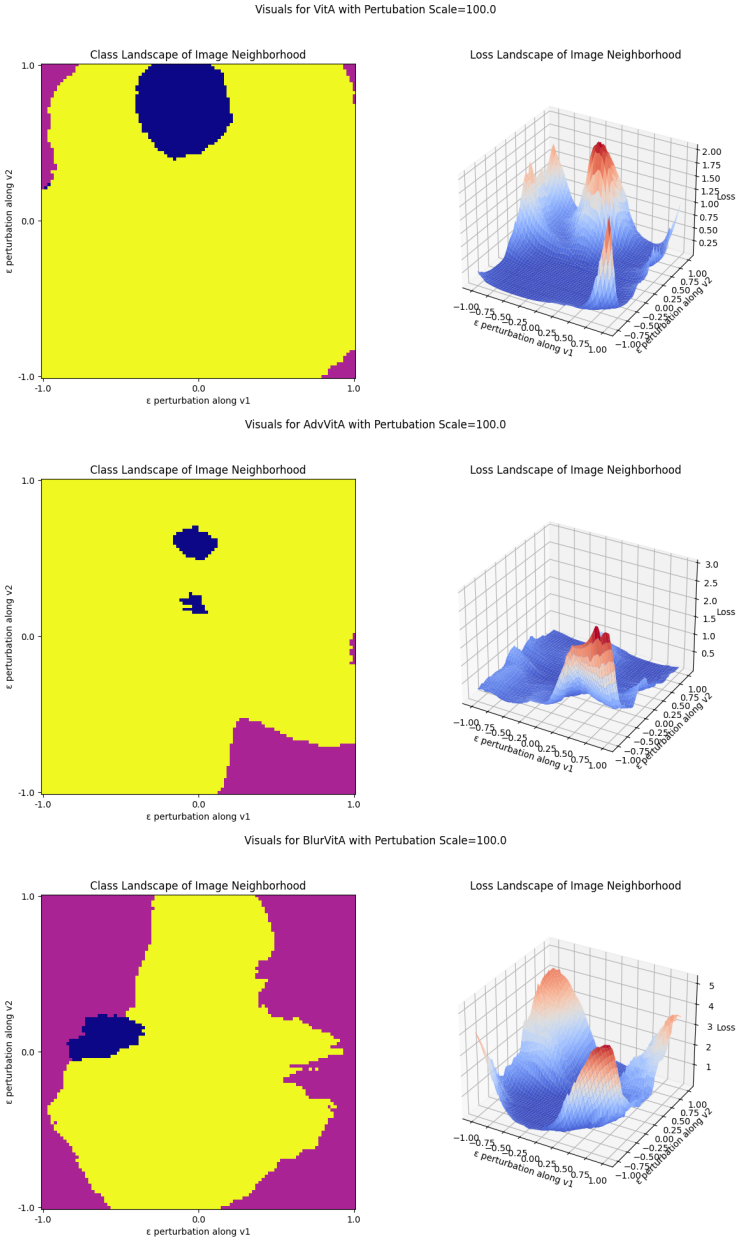

Figure 8: Decision boundary and loss landscape for ViT-based models in a neighborhood of a sample image, perturbations are made along a random unit vector (v1) and the normalized gradient of the unaugmented model (v2).

A.3 Data

Table 6: Mean and standard deviation of robustness metrics for each model and A-suffix fine-tuning variants.

Model	μ_{MSE}	σ_{MSE}	μ_{SSIM}	σ_{SSIM}	μ_{L^∞}	σ_{L^∞}
ResNet	4.2e-02	5.6e-02	0.46	0.35	0.16	0.14
ResNetA	2.1e-05	1.7e-05	0.99	8.7e-03	4.2e-03	1.7e-03
AdvResNetA	1.6e-05	1.0e-05	0.99	4.8e-03	3.8e-03	1.4e-03
BlurResNetA	2.8e-05	3.1e-05	0.98	1.7e-02	4.5e-03	2.9e-03
Vit	7.0e-03	1.1e-02	0.72	0.32	5.6e-02	6.6e-02
VitA	2.1e-02	8.5e-02	0.77	0.38	6.7e-02	0.15
AdvVitA	2.1e-02	5.0e-02	0.73	0.39	7.5e-02	0.13
BlurVitA	3.3e-02	0.11	0.79	0.38	7.9e-02	0.19
Vim	3.1e-02	4.8e-02	0.49	0.35	0.14	0.13
VimA	0.11	0.29	0.81	0.36	0.16	0.38
AdvVimA	2.2e-02	0.11	0.89	0.27	5.4e-02	0.17
BlurVimA	2.9e-02	0.12	0.81	0.33	7.6e-02	0.18

Table 7: Summary of Adversarial Datasets

Dataset Name	Number of Classes	Number of Samples	Derived From
InAdvRes	10	3717	ImageNette-Val
InAdvVit	10	3406	ImageNette-Val
InAdvVim	10	3476	ImageNette-Val
CelAdvResNetA	10	117	CelebVal
CelAdvResNetB	10	121	CelebVal
CelAdvVimA	10	113	CelebVal
CelAdvVimB	10	114	CelebVal
CelAdvVitA	10	193	CelebVal
CelAdvVitB	10	194	CelebVal

Table 8: Adversarial generation failure percentages for each model.

Model	**Fail Rate (%)**	**Model**	**Fail Rate (%)**	**Model**	**Fail Rate (%)**
ResNetA	8.23	AdvResNetA	9.13	BlurResNetA	7.83
VitA	6.93	AdvVitA	8.70	BlurVitA	6.96
VimA	6.06	AdvVimA	5.65	BlurVimA	2.17