# Appendix

We provide additional insights into our study through various sections in the appendix, which include results for both versions of the VMamba models [29]: *v2* (reported in the main paper) and *v0*. Results that do not specify the VMamba version should be assumed to refer to *v2* by default.

In Section A, we provide detailed results on *Information Drop* experiments. Section A.1 explores information drop along the scanning direction. Sections A.2 and A.3 offer additional detailed results on Salient and Non-Salient Patch Drop and Random Patch Drop, respectively. Section A.4 presents further results on patch shuffling. Section B provides results on *Image Corruptions*. In Section B.1, robustness against global corruptions such as common corruption, and out-of-distribution datasets is evaluated. Additionally, in Section B.2, we provide model calibration results on in-distribition and out-of-distribution datasets. In Section B.3, we report results on fine-grained corruptions across different models. In Section B.4, we report image corruption results for the task of object detection and semantic segmentation. Finally, we expand our analysis of *Adversarial Attacks* across models in Section C. We present results for white-box attacks, transfer-based black-box attacks, and frequency-based attacks.

## A. Robustness against Information Drop

### A.1. Information Drop along the Scanning Axis

We investigate the models' behavior when information is dropped along the scanning directions. We consider three settings: (1) linearly increasing the amount of information dropped in each patch along the scanning direction, with the most information dropped in patches that are traversed last by the scanning operation, (2) dropping most of the information in the center of the scanning directions while preserving most of the information in patches that are traversed at the beginning and the end, and (3) sequentially dropping patches along the scanning directions. In Figure 5 we show qualitative samples for showing information drop along scanning directions. Figure 6 we report results across the first experiment setting across all the scanning directions. In Figure 7 we report results for the second experiment setting. Furthermore, in Figure 8 we report results across the third experiment setting across all the scanning directions.

### A.2. Random Patch Drop

In Table 6 we expand the random patch drop experiments from the main paper (Table 1) to patch sizes 56×56 and 224×224. Furthermore, in Figure 9, 10, 11, 12, 13, and 14 we expand our analysis on random patch drop to several other models.

### A.3. Salient and Non-Salient Patch Drop

In Figure 15 and 16, we report results on salient and non-salient patch drop of information.

### A.4. Patch Shuffling

In Figure 17, we expand the patch shuffling experiment from the main paper to several CNN and transformer-based architectures.

## B. Robustness against Image Corruptions

### B.1. Robustness against Global Corruptions

In Table 7 relative corruption error is reported across models discussed in the main paper. In Figure 18 and 19, we present the relative corruption error and mean corruption error (mCE) of various models subjected to all 19 corruption methods applied to the ImageNet dataset. The common corruption consists of various types of synthetic corruptions to assess the models' robustness against image distortions like noise, blur, and compression artifacts. In Table 8 we report results on ***ImageNetV2*** [42], ***ImageNet-A*** [20], ***ImageNet-R*** [19], and ***ImageNet-S*** [11]. These datasets are designed to evaluate the robustness and generalization capabilities of models trained on the original ImageNet dataset. ImageNetV2 [42] tests the models' performance on previously unseen images that follow the same distribution as the training data, while ImageNet-A [20] contains naturally occurring adversarial examples that are difficult for models to classify correctly. ImageNet-R [19] consists of images with different artistic renditions to test the models' ability to generalize to different visual domains and styles, and ImageNet-S [11] consists of sketch-based images.

### B.2. Model Calibration

Model calibration assesses how well a model's predicted confidence aligns with its actual accuracy. For example, if a model predicts a confidence level of 70% for its predictions, a well-calibrated model should have an actual accuracy close to 70%. To quantify this alignment, we use the Expected Calibration Error (ECE). ECE involves dividing predictions into $M$ bins based on their confidence levels (e.g., 60%-70%, 70%-80%). For each bin, the average accuracy and confidence are computed, and the ECE is the weighted average of the differences between these values. Calibration can also be evaluated visually using reliability diagrams, which plot predicted confidence against actual accuracy; a well-calibrated model should show points near the diagonal. Additionally, confidence histograms reveal the distribution of prediction confidences. Evaluation is performed on both in-distribution data (e.g., ImageNet, ImageNetV2) and out-of-distribution data (e.g., ImageNet-R, ImageNet-S, ImageNet-A), with $M = 15$ bins used in all experiments.

In Figure 20, 21, and 22, we report the ECE plot the reliability diagrams for Tiny, Small, and Base version of different architectures, respectively. Figure 23, we report the ECE score. We observe that while ViT models report a low ECE error on in-distribution datasets, the error increases significantly for out-of-distribution datasets. An exception is the Base model, which can be attributed to the ViT-B model being pretrained on a larger dataset (ImageNet-21k) comapared to other models. For the hybrid-based VSSM model MambaVision we observe a contrasting behaviour with ViT models, while the ECE score on in-distribution datasets is relatively high, it improves significantly on out-of-distribution datasets. The improvement in ECE for MambaVision on out-of-distribution datasets suggests that the hybrid approach of combining vision transformers with other model components may enhance calibration performance in these scenarios.

### B.3. Robustness against Fine-grained Corruptions

In Table 9 *(left)* we report results on ImageNet-E dataset across several CNNs, Transformers, and VSSM-based models, and in Table 9 *(right)*, we report results on ImageNet-B dataset.

### B.4. Robustness Evaluation for Object Detection and Semantic Segmentation

In Table 10, we report AP scores of different architectures on COCO-DC dataset. On COCO-DC, we observe that VMamba models' high performance on clean images does not translate to color and texture background variations. Swin-S model achieves the highest average AP score of $56.70$ across the background variations, followed by score of $56.20$ by VMamba-S model. Figure 24 shows the mIoU, mAcc, and aAcc scores on ADE-20K after applying common corruptions to the dataset. Similarly, Figure 25 displays the mAP, APs, APm, and APl scores on the COCO-C dataset. All the scores are averaged across all severity levels of each corruption.

## C. Robustness against Adversarial Attacks

In this section, we expand our analysis on adversarial attacks and their transferability across different models. We include more model families, such as DeiT, DenseNet, and VGG. Figures 26, 27, and 28 present the robust accuracy of various models under both white-box and black-box settings for FGSM, PGD, and MIFGSM attacks, respectively. All adversarial examples are crafted with a perturbation budget of $\epsilon = \frac{8}{255}$. For PGD and MIFGSM attacks, we use 20 iterations to craft the adversarial examples.

In Table 11, we evaluate the robustness of VMamba and MambaVision against frequency-specific adversarial attacks crafted using Projected Gradient Descent (PGD). These perturbations are constrained to designated frequency bands through a discrete cosine transform (DCT) mask filter [34]. Results in Tab. 11 (*left*) demonstrate that VMamba and MambaVision maintain robustness above $90\%$ for low-frequency perturbations up to $\epsilon = 16$, indicating strong resilience similar to the ConvNext and Swin transformer counterparts. ViT models exhibit the most performance drop under low-frequency adversarial attacks. For high-frequency adversarial attacks (Tab. 11 (*middle*)), the robustness of all models decreases more rapidly with increasing perturbation strength, although ViT-based models show the highest robustness. Finally, for standard attacks where the complete frequency range is used to generate adversarial examples, VSSM models display higher robustness compared to other models, including ConvNext, ViT, and Swin.

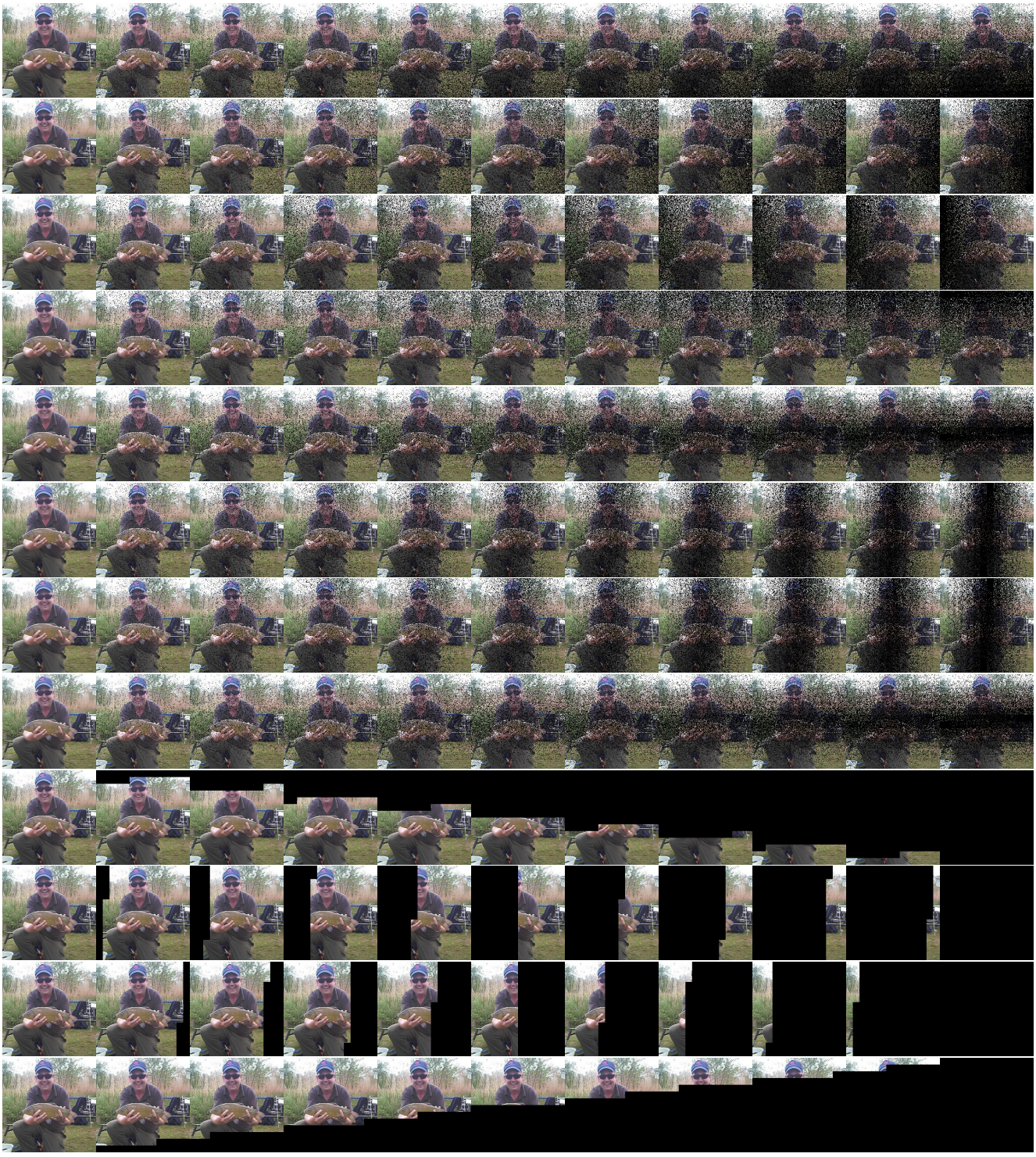

Figure 5. Information drop along scanning direction: The top four rows represent linearly increasing information drop along the four scanning directions at patch size $14 \times 14$. The center four rows represent linearly increasing information drop till the center along the four scanning directions at patch size $14 \times 14$. The bottom four rows represent sequentially dropping patches along the four scanning directions at patch size $14 \times 14$.

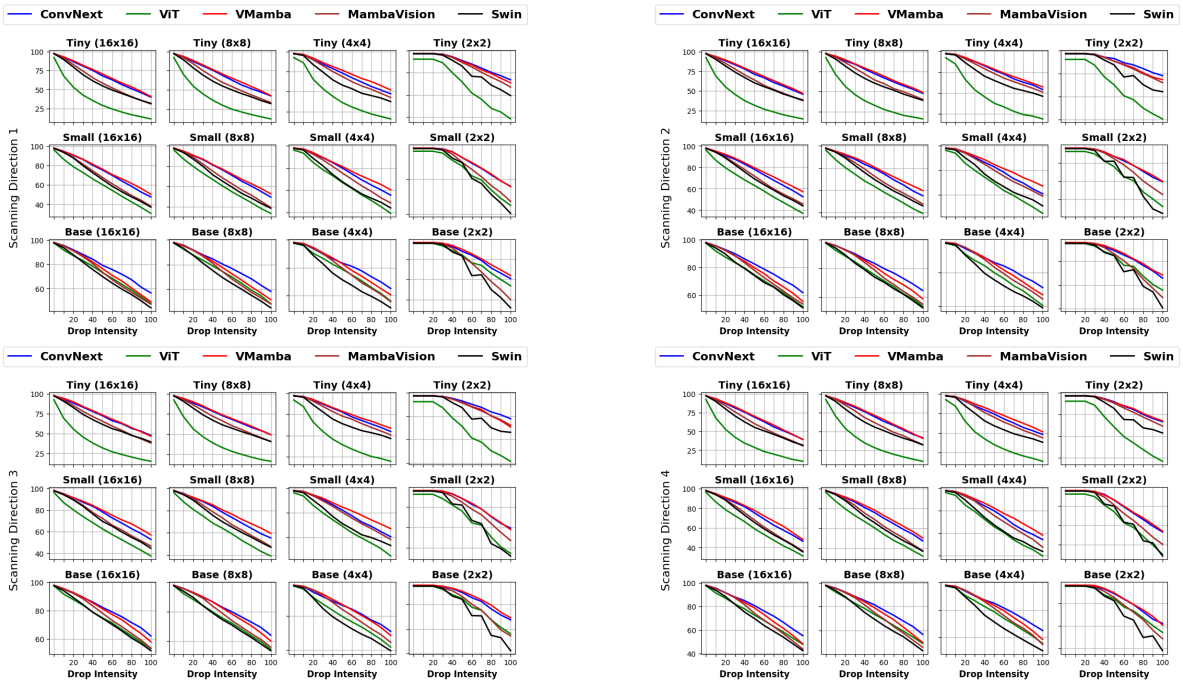


Figure 6. Information drop of Tiny and Small family of models along the scanning direction: the image is split into a sequence of fixed-size non-overlapping patches of size 16x16, 8x8. 4x4, and 2x2. We report results of linearly increasing the number of pixels dropped from each patch to the maximum threshold (*Drop Intensity*) along the scanning direction. Top row shows results for top-to-bottom *(Direction 1)* and left-to-right direction*(Direction 2)*. Bottom row shows results for right-to-left *(Direction 3)* and bottom-to-top direction*(Direction 4)*.

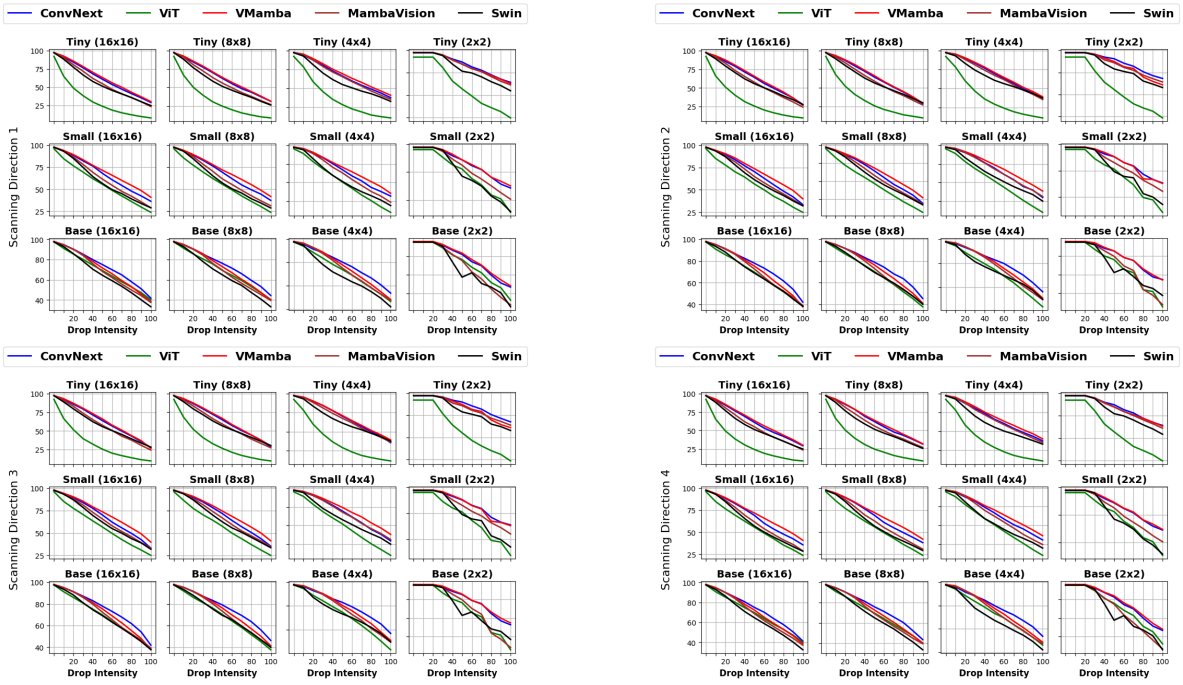


Figure 7. Information drop of Tiny and Small family of models along the scanning direction: the image is split into a sequence of fixed-size non-overlapping patches of size 16x16, 8x8. 4x4, and 2x2. We report results of linearly increasing the number of pixels dropped from each patch to the maximum threshold (*Drop Intensity*) at the center of the scanning direction and then again linearly decreased till the end. The top row shows results for top-to-bottom *(Direction 1)* and left-to-right direction*(Direction 2)*. The bottom row shows results for right-to-left *(Direction 3)* and bottom-to-top direction*(Direction 4)*.

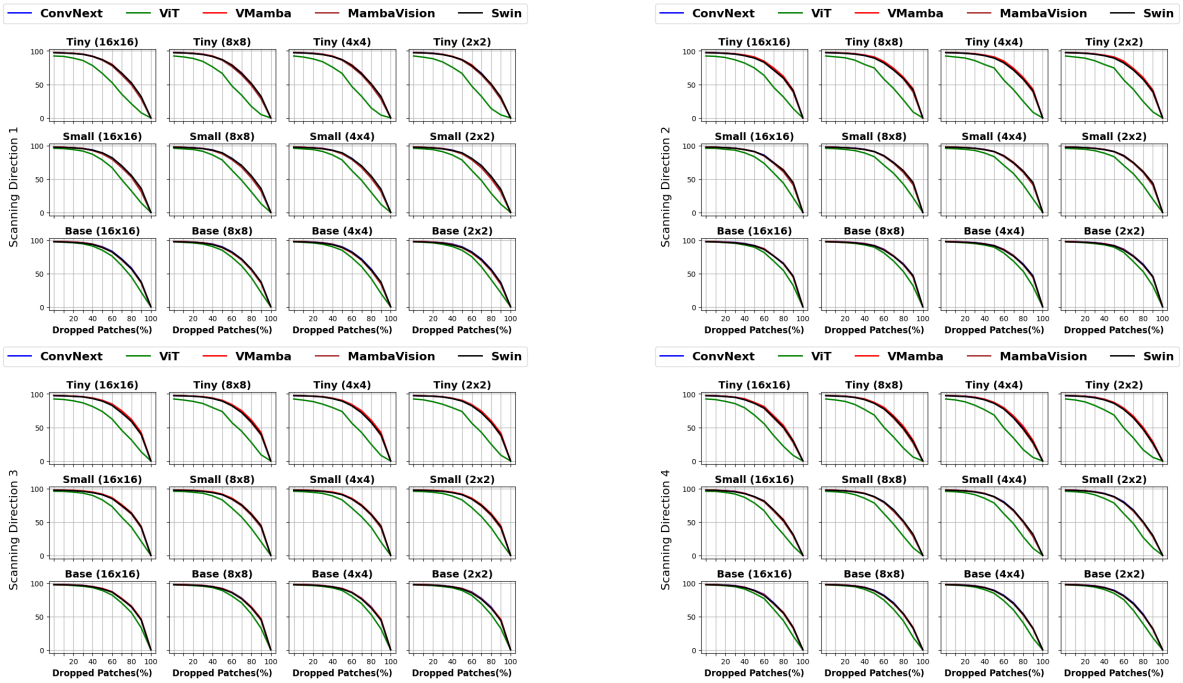


Figure 8. Information drop of Tiny and Small family of models along the scanning direction: the image is split into a sequence of fixed-size non-overlapping patches of size 16x16, 8x8. 4x4, and 2x2. We report results of sequentially dropping patches along the scanning direction. The top row shows results for top-to-bottom *(Direction 1)* and left-to-right direction*(Direction 2)*. The bottom row shows results for right-to-left *(Direction 3)* and bottom-to-top direction*(Direction 4)*.

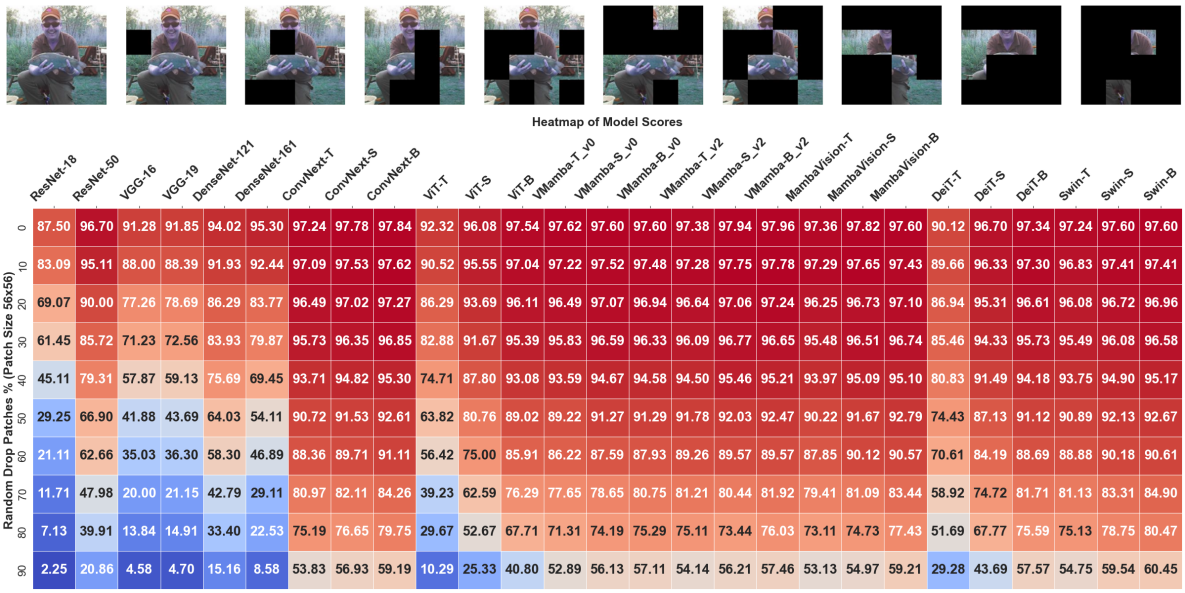


Figure 9. Top-1 classification accuracy of various architectures under random patch drop occlusions, using $56 \times 56$ patch size.

Table 6. Top-1 classification accuracy reported random patch drop occlusion using $16 \times 16$, $8 \times 8$, $4 \times 4$ and $1 \times 1$ patch sizes.

| 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 $16 \times 16$ *(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 $8 \times 8$ *(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 |
| Patch Size $4 \times 4$ *(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 |
| 30.11 | 84.75 | 88.74 | 90.51 | 29.62 | 82.23 | 90.04 | 86.34 | 87.17 | 90.48 | 80.16 | 85.39 | 89.79 | 92.51 | 92.97 | 94.68 |
| 12.02 | 64.23 | 77.49 | 81.82 | 9.25 | 65.28 | 77.00 | 72.09 | 75.43 | 77.77 | 57.80 | 57.14 | 71.68 | 84.20 | 86.17 | 90.71 |
| 6.49 | 33.19 | 47.91 | 70.73 | 3.63 | 43.92 | 56.17 | 51.01 | 55.23 | 53.83 | 29.05 | 24.13 | 44.27 | 72.07 | 77.69 | 85.35 |
| 2.91 | 13.34 | 19.71 | 54.95 | 1.58 | 21.73 | 32.19 | 27.17 | 33.21 | 29.41 | 11.75 | 8.59 | 23.15 | 57.71 | 65.13 | 78.48 |
| 1.46 | 5.17 | 6.67 | 34.97 | 0.70 | 8.96 | 15.24 | 10.51 | 15.82 | 11.57 | 4.00 | 2.84 | 10.25 | 41.57 | 47.95 | 69.11 |
| 0.75 | 2.07 | 2.23 | 17.37 | 0.41 | 3.69 | 6.01 | 3.07 | 5.41 | 3.40 | 1.41 | 0.90 | 3.89 | 26.49 | 29.31 | 54.59 |
| 0.40 | 0.85 | 0.90 | 6.21 | 0.24 | 1.45 | 1.91 | 0.83 | 1.28 | 0.63 | 0.50 | 0.39 | 1.05 | 12.17 | 12.45 | 35.95 |
| 0.21 | 0.33 | 0.44 | 1.70 | 0.17 | 0.55 | 0.71 | 0.23 | 0.30 | 0.24 | 0.29 | 0.17 | 0.23 | 3.48 | 2.47 | 14.47 |
| 0.13 | 0.20 | 0.25 | 0.40 | 0.19 | 0.24 | 0.25 | 0.19 | 0.21 | 0.22 | 0.19 | 0.13 | 0.14 | 0.64 | 0.41 | 1.55 |
| Patch Size $1 \times 1$ *(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 |
| 54.40 | 83.43 | 87.12 | 89.68 | 47.56 | 76.30 | 85.64 | 85.87 | 89.55 | 89.73 | 80.38 | 86.96 | 90.11 | 74.40 | 85.10 | 85.14 |
| 37.39 | 69.17 | 76.25 | 79.47 | 27.12 | 61.71 | 75.93 | 71.95 | 77.99 | 79.11 | 57.76 | 69.25 | 76.01 | 48.34 | 63.30 | 62.51 |
| 26.21 | 53.81 | 62.42 | 68.38 | 15.67 | 47.69 | 65.85 | 56.63 | 64.54 | 66.81 | 39.91 | 50.91 | 57.74 | 33.44 | 45.94 | 49.32 |
| 17.80 | 38.14 | 45.85 | 58.50 | 9.02 | 34.97 | 55.16 | 43.11 | 52.29 | 54.38 | 27.55 | 35.95 | 40.89 | 25.06 | 32.73 | 39.90 |
| 11.85 | 26.13 | 31.79 | 47.28 | 5.11 | 23.24 | 43.27 | 32.61 | 41.20 | 42.08 | 18.56 | 24.83 | 27.92 | 18.84 | 23.16 | 31.09 |
| 7.19 | 17.81 | 21.99 | 35.81 | 2.78 | 14.30 | 31.58 | 23.08 | 30.97 | 29.57 | 11.42 | 16.06 | 19.34 | 12.75 | 15.41 | 22.05 |
| 3.94 | 11.18 | 15.01 | 21.95 | 1.39 | 7.82 | 19.59 | 13.39 | 21.11 | 17.70 | 5.97 | 9.18 | 11.85 | 6.69 | 9.58 | 11.99 |
| 1.79 | 4.47 | 8.49 | 9.15 | 0.61 | 3.51 | 10.46 | 4.65 | 9.64 | 6.79 | 2.19 | 3.19 | 4.87 | 2.15 | 4.15 | 4.05 |
| 0.62 | 0.99 | 2.91 | 2.62 | 0.29 | 1.23 | 3.13 | 0.79 | 1.47 | 1.23 | 0.71 | 0.63 | 1.08 | 0.61 | 1.08 | 0.95 |

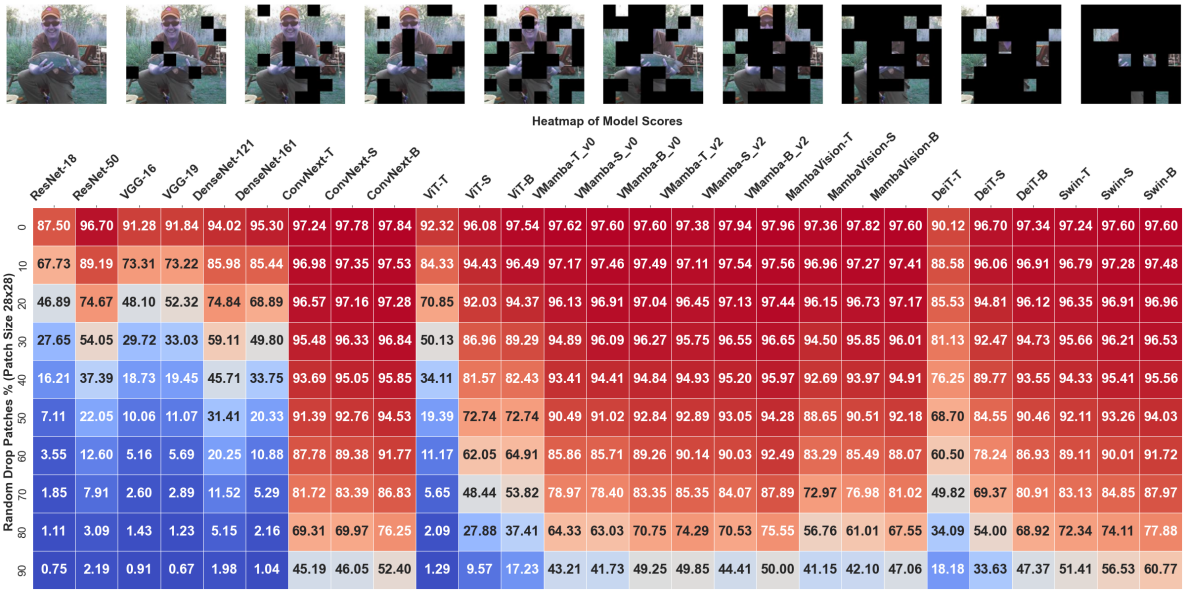


Figure 10. Top-1 classification accuracy of various architectures under random patch drop occlusions, using $28 \times 28$ patch size.

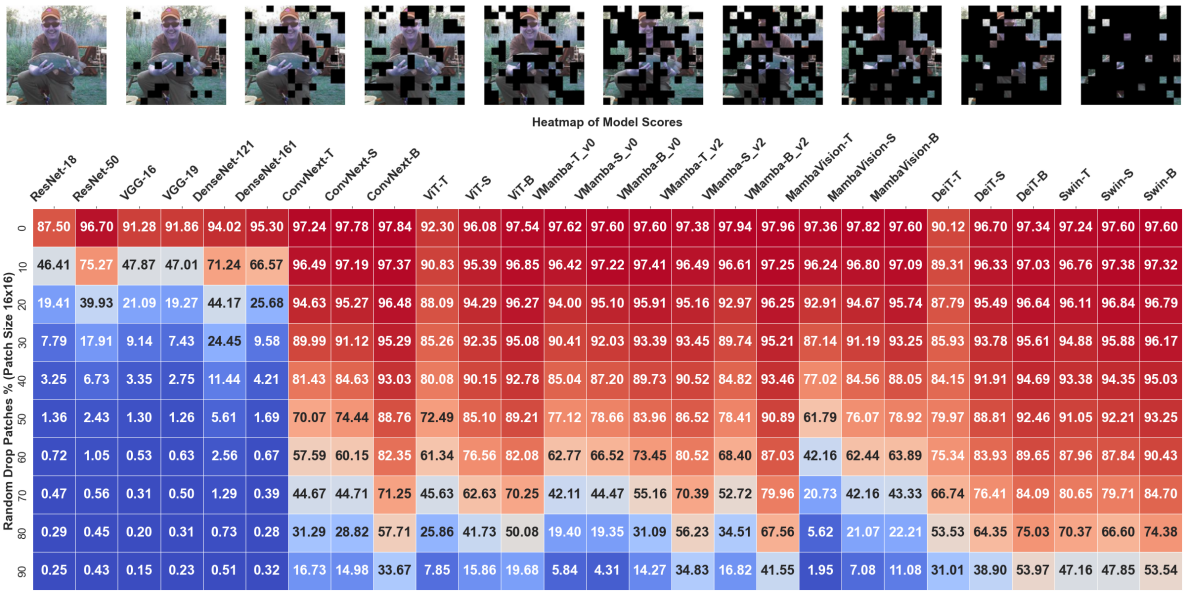


Figure 11. Top-1 classification accuracy of various architectures under random patch drop occlusions, using $16 \times 16$ patch size.

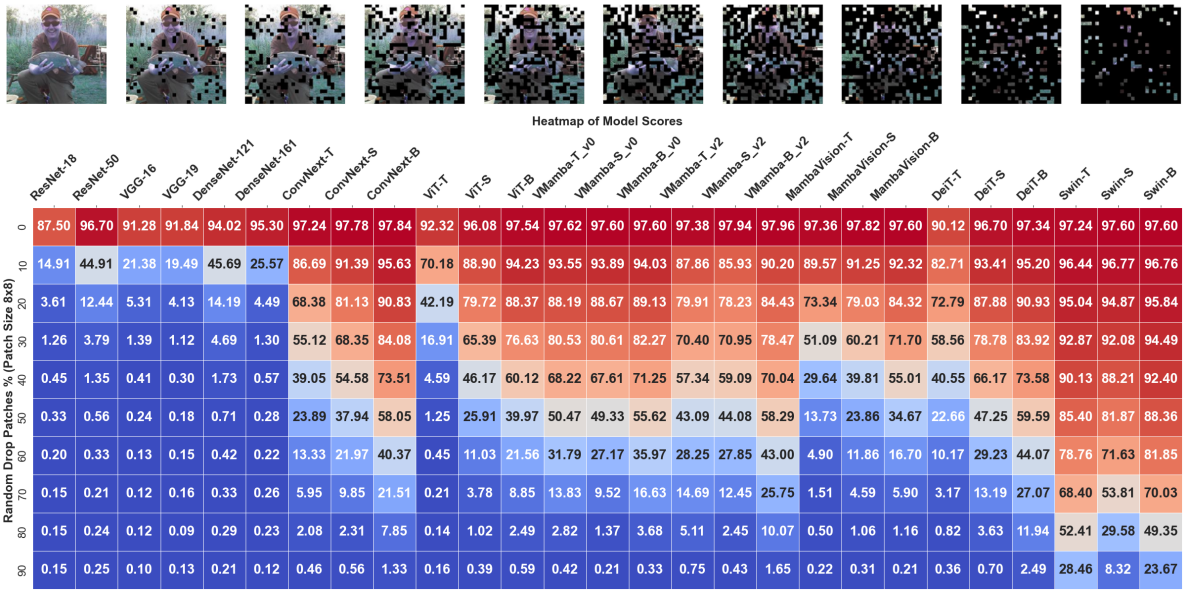


Figure 12. Top-1 classification accuracy of various architectures under random patch drop occlusions, using $8 \times 8$ patch size.

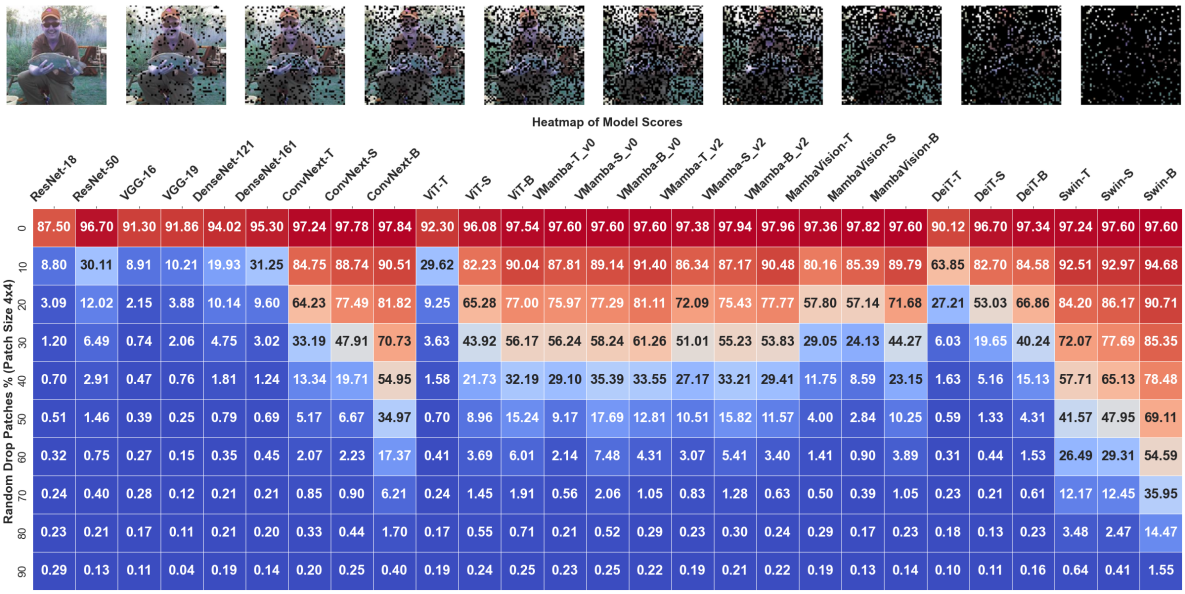


Figure 13. Top-1 classification accuracy of various architectures under random patch drop occlusions, using $4 \times 4$ patch size.

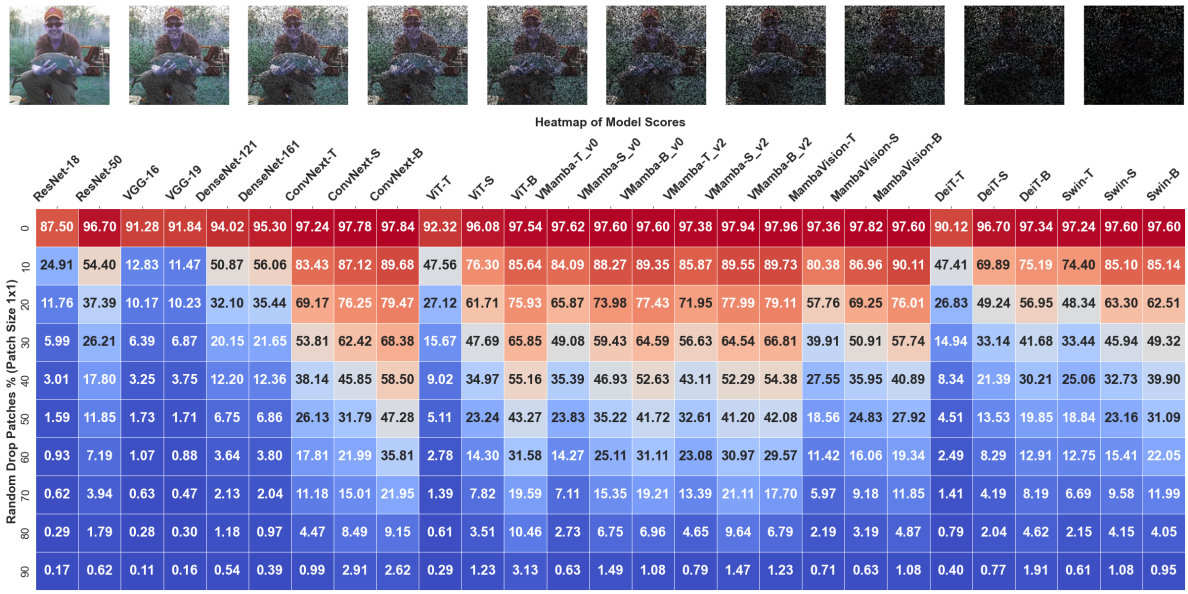


Figure 14. Top-1 classification accuracy of various architectures under random patch drop occlusions, using $1 \times 1$ patch size.

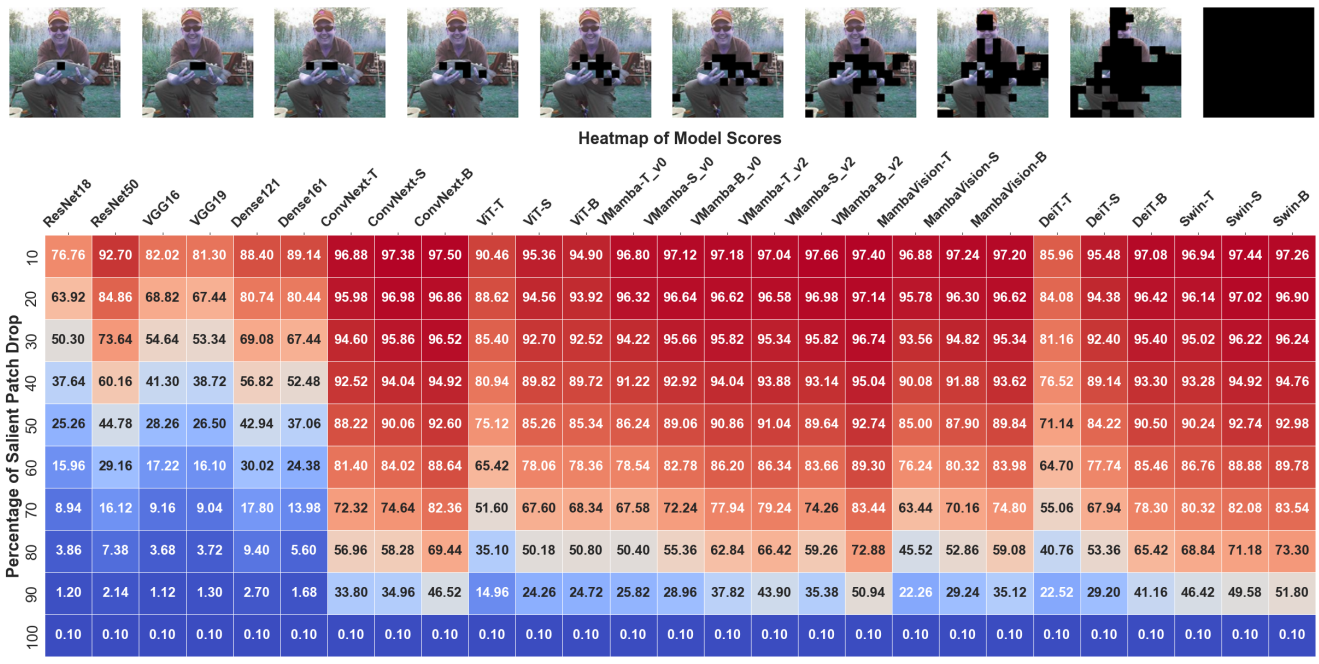


Figure 15. Top-1 classification accuracy reported under salient patch drop occlusion using $16 \times 16$ patch size.

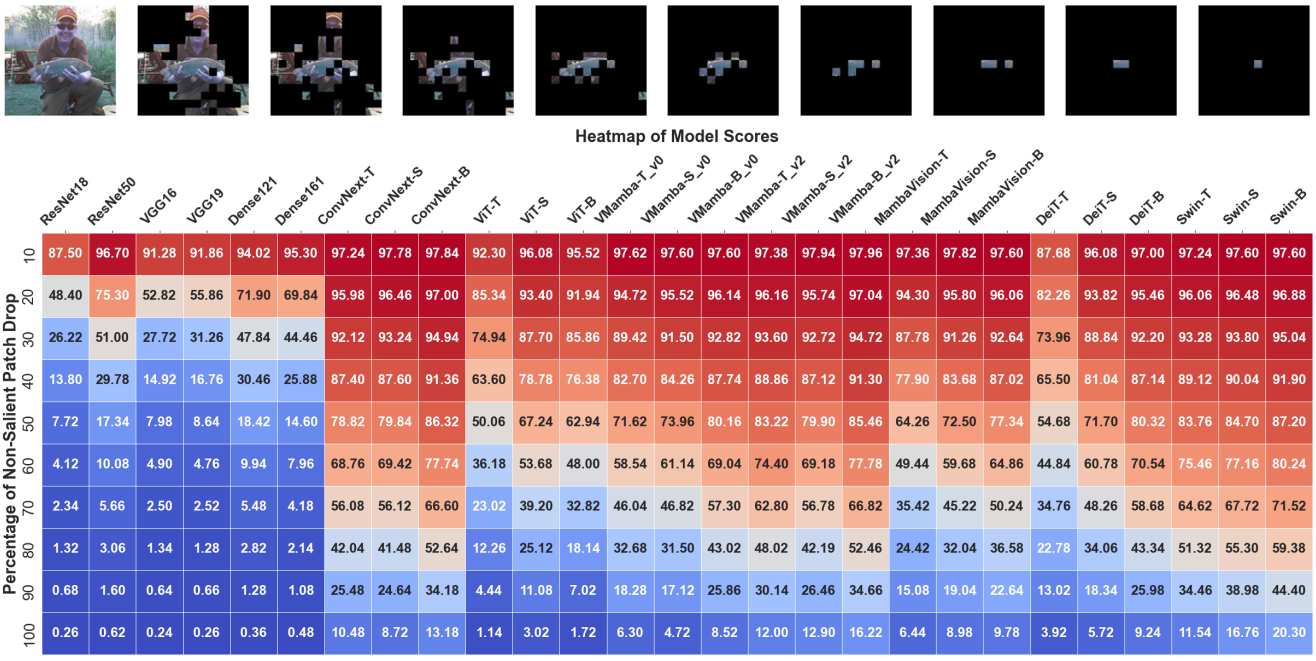


Figure 16. Top-1 classification accuracy reported under non-salient patch drop occlusion using $16 \times 16$ patch size.

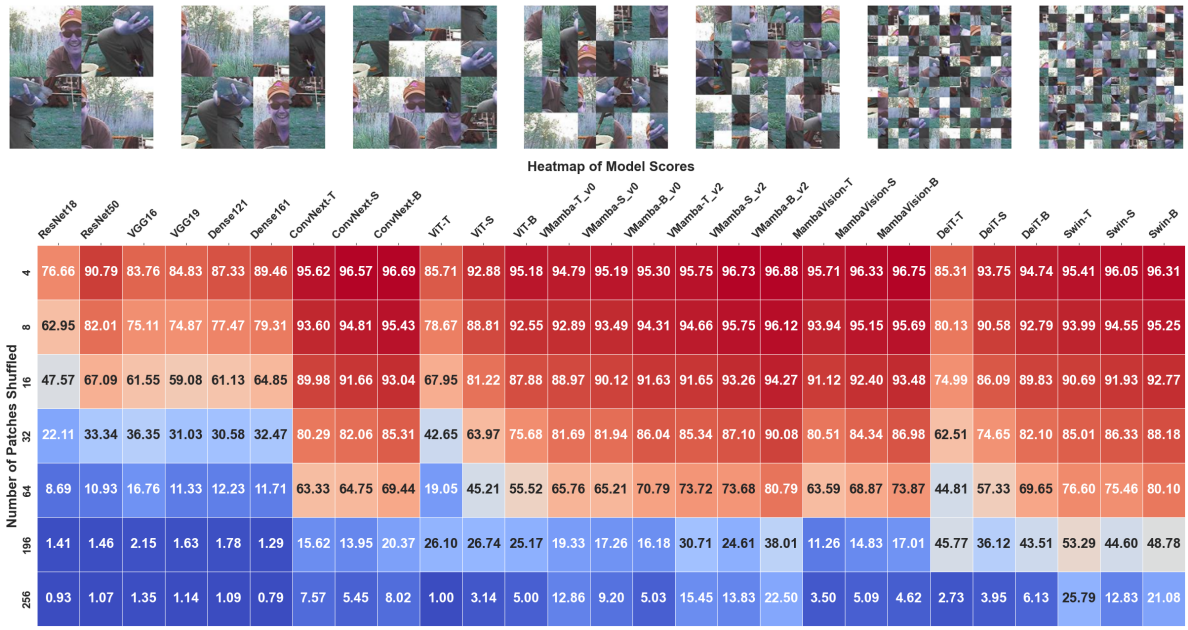


Figure 17. Top-1 classification accuracy among various architectures under increasing patch shuffling.

Table 7. Corruption Error (CE) of various architectures on ImageNet-C corruptions over multiple intensity levels. Results are relative to the CE on ResNet-50 and are evaluated on 5000 ImageNet validation set images.

| **Corruptions ↓** | 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 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Brightness | 100.00 | 68.20 | 58.63 | 52.15 | 223.14 | 101.09 | 41.67 | 61.60 | 53.06 | 52.94 | 64.39 | 53.73 | 51.30 | 75.89 | 65.35 | 63.54 |
| Contrast | 100.00 | 40.14 | 30.29 | 28.96 | 426.29 | 186.49 | 103.16 | 38.17 | 30.02 | 27.70 | 52.05 | 34.93 | 29.72 | 44.41 | 42.09 | 41.70 |
| Defocus Blur | 100.00 | 85.70 | 77.82 | 73.87 | 103.27 | 70.64 | 42.81 | 84.96 | 72.32 | 73.00 | 77.80 | 71.71 | 67.04 | 87.50 | 66.82 | 81.49 |
| Elastic Transform | 100.00 | 81.11 | 72.18 | 69.44 | 70.34 | 55.69 | 41.43 | 82.78 | 69.89 | 72.21 | 78.96 | 79.82 | 75.06 | 86.36 | 66.53 | 76.78 |
| Fog | 100.00 | 113.68 | 86.18 | 91.19 | 312.26 | 119.54 | 68.79 | 87.47 | 64.93 | 55.88 | 62.13 | 46.48 | 39.35 | 84.32 | 87.24 | 68.61 |
| Frost | 100.00 | 64.12 | 57.09 | 50.75 | 146.53 | 75.52 | 27.19 | 42.05 | 36.58 | 30.79 | 57.65 | 49.73 | 40.25 | 47.77 | 40.72 | 38.63 |
| Gaussian Blur | 100.00 | 86.69 | 84.58 | 77.75 | 107.84 | 82.50 | 55.22 | 88.91 | 77.05 | 77.21 | 79.64 | 74.80 | 70.72 | 90.59 | 81.36 | 85.37 |
| Gaussian Noise | 100.00 | 52.01 | 51.36 | 56.02 | 130.74 | 66.75 | 30.16 | 56.81 | 38.29 | 52.62 | 115.29 | 110.36 | 96.79 | 73.31 | 59.86 | 54.49 |
| Glass Blur | 100.00 | 89.71 | 85.46 | 81.43 | 98.30 | 84.73 | 66.52 | 93.59 | 86.00 | 87.28 | 90.58 | 90.13 | 89.29 | 96.53 | 86.27 | 90.24 |
| Impulse Noise | 100.00 | 47.18 | 44.71 | 45.78 | 122.95 | 61.19 | 28.27 | 44.55 | 31.93 | 39.72 | 92.95 | 84.21 | 75.73 | 63.83 | 48.63 | 48.68 |
| JPEG Compression | 100.00 | 69.79 | 59.62 | 60.96 | 131.33 | 78.90 | 43.49 | 82.29 | 65.82 | 63.30 | 74.33 | 69.63 | 62.72 | 144.27 | 88.13 | 108.22 |
| Motion Blur | 100.00 | 65.05 | 56.10 | 51.42 | 106.50 | 58.86 | 33.26 | 65.84 | 54.55 | 54.13 | 63.42 | 58.58 | 53.44 | 80.88 | 50.12 | 62.42 |
| Pixelate | 100.00 | 82.35 | 71.52 | 66.00 | 52.11 | 27.31 | 15.88 | 80.34 | 66.30 | 65.02 | 67.42 | 67.31 | 59.34 | 109.24 | 100.70 | 101.51 |
| Saturate | 100.00 | 63.44 | 53.74 | 47.61 | 250.68 | 116.28 | 48.06 | 57.26 | 47.97 | 47.07 | 60.37 | 52.12 | 48.11 | 75.29 | 62.40 | 58.34 |
| Shot Noise | 100.00 | 52.69 | 51.85 | 56.73 | 128.00 | 66.02 | 30.12 | 56.20 | 38.75 | 51.72 | 105.73 | 104.55 | 91.49 | 73.51 | 62.09 | 57.99 |
| Snow | 100.00 | 54.55 | 48.57 | 43.18 | 134.88 | 61.12 | 26.04 | 54.29 | 46.02 | 42.14 | 54.68 | 48.88 | 42.89 | 57.70 | 39.38 | 47.00 |
| Spatter | 100.00 | 39.86 | 32.09 | 29.41 | 117.51 | 48.16 | 21.03 | 33.10 | 28.59 | 24.33 | 39.28 | 29.69 | 24.92 | 32.71 | 23.79 | 22.92 |
| Speckle Noise | 100.00 | 44.78 | 40.91 | 34.08 | 132.49 | 60.64 | 26.89 | 42.92 | 30.43 | 34.07 | 84.20 | 79.68 | 64.73 | 61.92 | 48.85 | 48.80 |
| Zoom Blur | 100.00 | 82.47 | 68.22 | 64.74 | 125.18 | 84.89 | 52.90 | 82.72 | 70.04 | 70.37 | 76.07 | 71.43 | 65.52 | 93.27 | 77.34 | 77.96 |
| **mCE** | 100.00 | 67.55 | 59.52 | 56.92 | 153.7 | 79.28 | 42.26 | 65.05 | 53.08 | 53.76 | 66.58 | 50.50 | 49.61 | 77.86 | 63.04 | 64.98 |

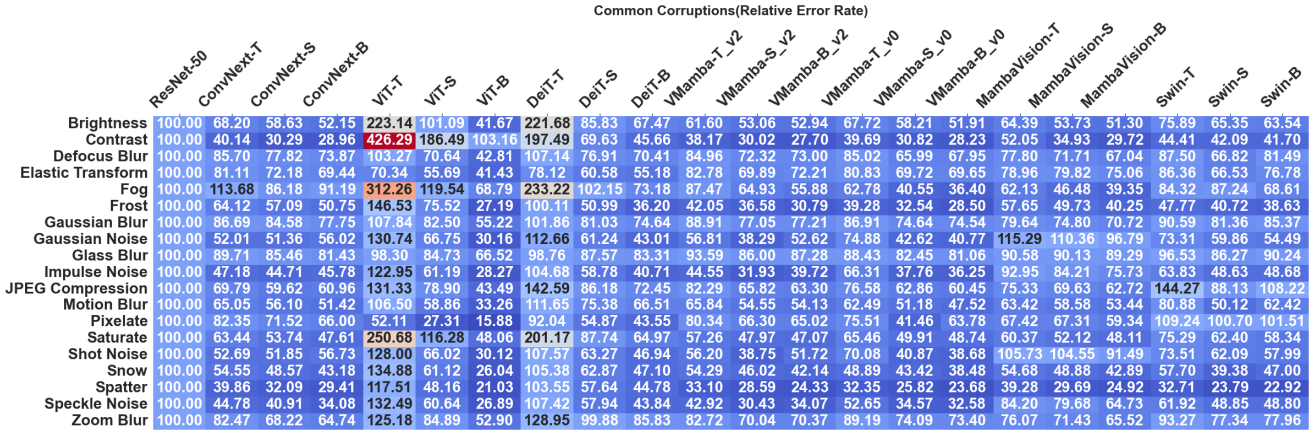


Figure 18. Corruption Error(CE) values for different corruptions and architectures on ImageNet-C, with error rates relative to ResNet-50.

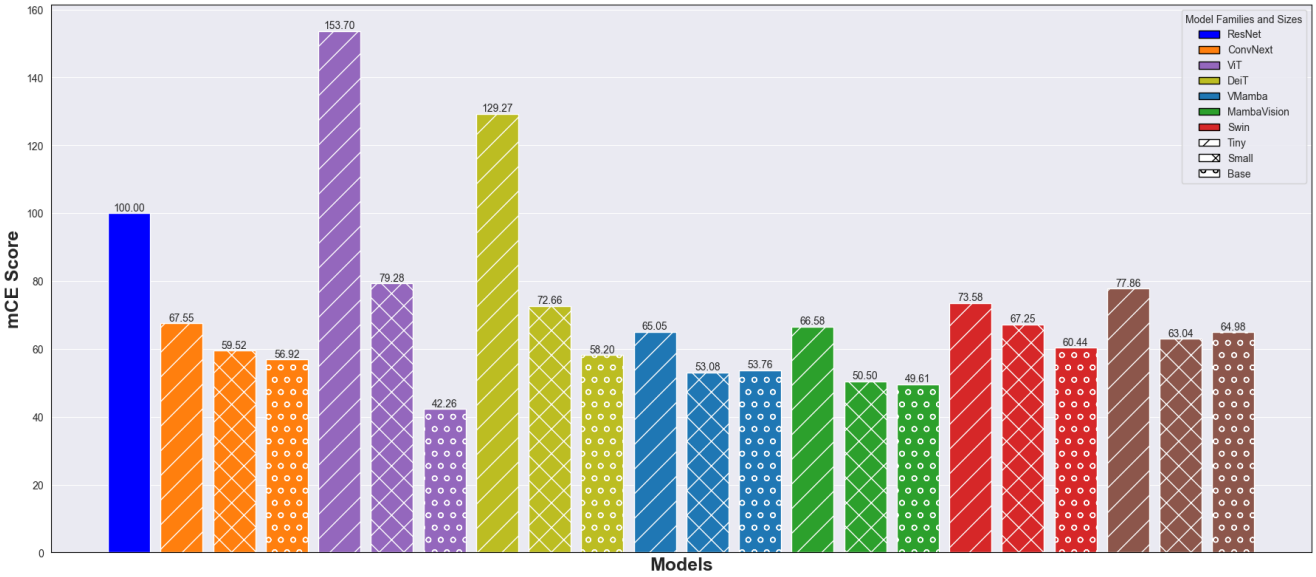


Figure 19. mCE for different corruptions and architectures on ImageNet-C, with error rates relative to ResNet-50.

Table 8. Comparison in domain generalization setting. Models trained on ImageNet are evaluated on datasets with domain shifts.

| Model ↓ | ImageNet | ImageNetV2 | ImageNet-S | ImageNet-A | ImageNet-R | Average |
|---|---|---|---|---|---|---|
| ConvNext-T | 81.87 | 70.67(-11.20) | 33.96(-47.91) | 10.48(-71.39) | 32.53(-49.34) | 36.90(-44.97) |
| ViT-T | 75.35 | 63.05(-12.30) | 20.88(-54.47) | 3.31(-72.04) | 20.29(-55.06) | 26.88(-48.47) |
| Swin-T | 80.91 | 69.31(-11.60) | 29.24(-51.67) | 8.93(-71.98) | 28.52(-52.39) | 34.00(-46.91) |
| VMamba-T*(v0)* | 81.92 | 70.82(-11.10) | 33.01(-48.91) | 13.56(-68.36) | 32.11(-49.81) | 37.37(-44.55) |
| VMamba-T*(v2)* | 82.28 | 71.16(-11.12) | 33.99(-48.29) | 12.08(-70.20) | 32.05(-50.23) | 37.32(-44.96) |
| MambaVision-T | 82.10 | 71.50(-10.60) | 33.49(-48.61) | 13.39(-68.71) | 32.28(-49.82) | 37.66(-44.44) |
| ConvNext-S | 82.82 | 72.07(-10.75) | 37.16(-45.66) | 14.43(-68.39) | 35.57(-47.25) | 39.81(-43.01) |
| ViT-S | 81.40 | 69.98(-11.42) | 32.77(-48.63) | 13.09(-68.31) | 31.14(-50.26) | 36.74(-44.66) |
| Swin-S | 82.90 | 71.62(-11.28) | 31.97(-50.93) | 15.72(-67.18) | 31.93(-50.97) | 37.81(-45.09) |
| VMamba-S*(v0)* | 83.15 | 72.89(-10.26) | 38.05(-45.10) | 17.28(-65.87) | 37.05(-46.10) | 41.32(-41.83) |
| VMamba-S*(v2)* | 83.48 | 73.01(-10.47) | 36.98(-46.50) | 16.45(-67.03) | 35.75(-47.73) | 40.55(-42.93) |
| MambaVision-S | 83.22 | 72.62(-10.60) | 35.53(-47.69) | 15.97(-67.25) | 33.96(-49.26) | 39.52(-43.70) |
| ConvNext-B | 83.75 | 73.68(-10.07) | 38.23(-45.52) | 18.16(-65.59) | 36.66(-47.09) | 41.68(-42.07) |
| ViT-B | 84.40 | 73.84(-10.56) | 43.01(-41.39) | 24.09(-60.31) | 41.03(-43.37) | 45.49(-38.91) |
| Swin-B | 83.08 | 72.09(-10.99) | 32.62(-50.46) | 17.95(-65.13) | 33.23(-49.85) | 38.97(-44.11) |
| VMamba-B*(v0)* | 83.48 | 72.97(-10.51) | 38.24(-45.24) | 18.88(-64.60) | 37.33(-46.15) | 41.86(-41.62) |
| VMamba-B*(v2)* | 83.76 | 73.22(-10.54) | 38.53(-45.23) | 18.35(-65.41) | 35.99(-47.77) | 41.52(-42.24) |
| MambaVision-B | 83.96 | 73.84(-10.12) | 36.69(-47.27) | 21.69(-62.27) | 35.72(-48.24) | 41.98(-41.98) |

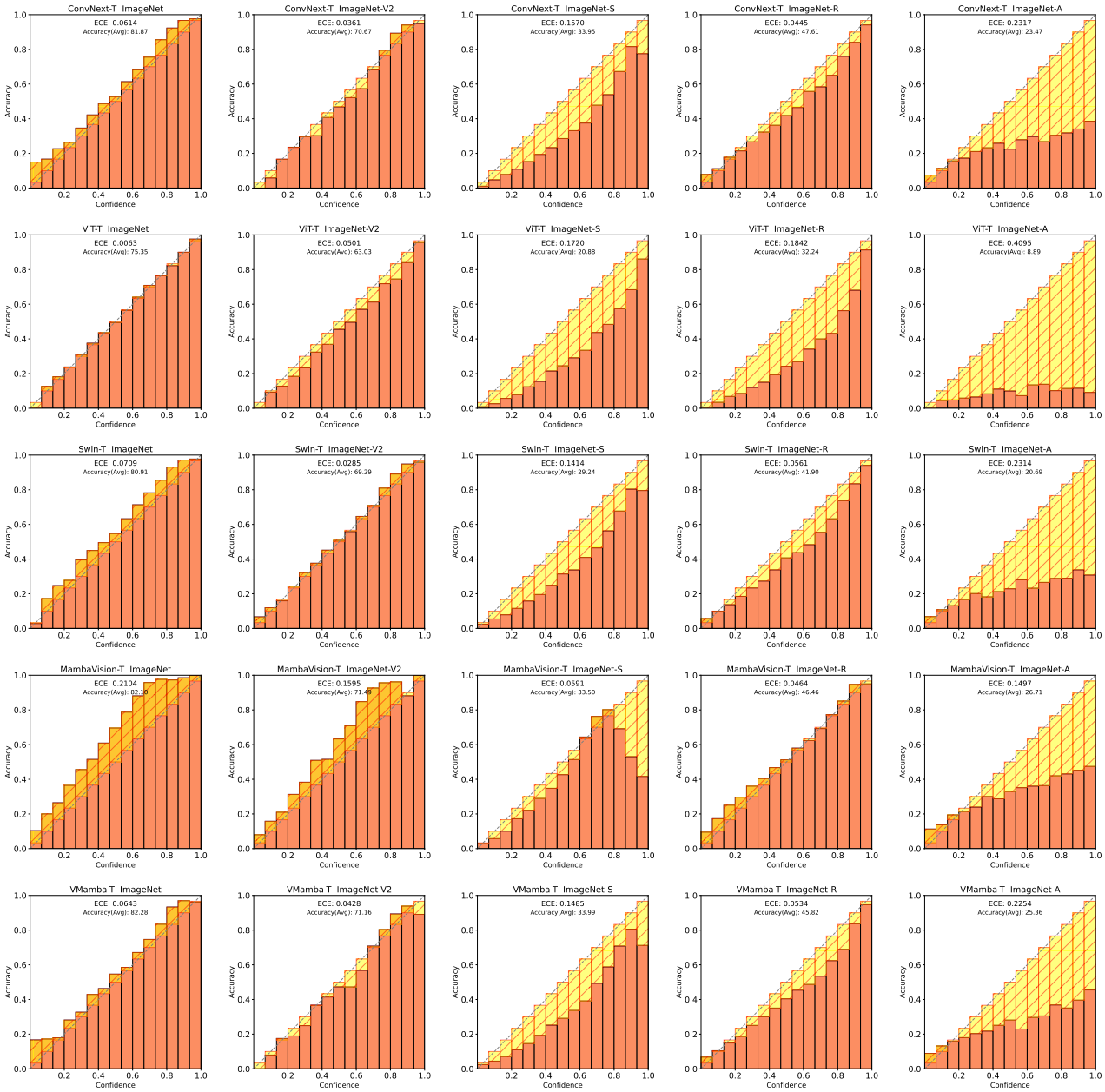


Figure 20. Caliberation Results: Reliability diagrams and ECE on ConvNext-T, ViT-T, Swin-T, VMamba-T, and MambaVision-T across ImageNet, ImageNet-V2, ImageNet-S, ImageNet-R, and ImageNet-A.

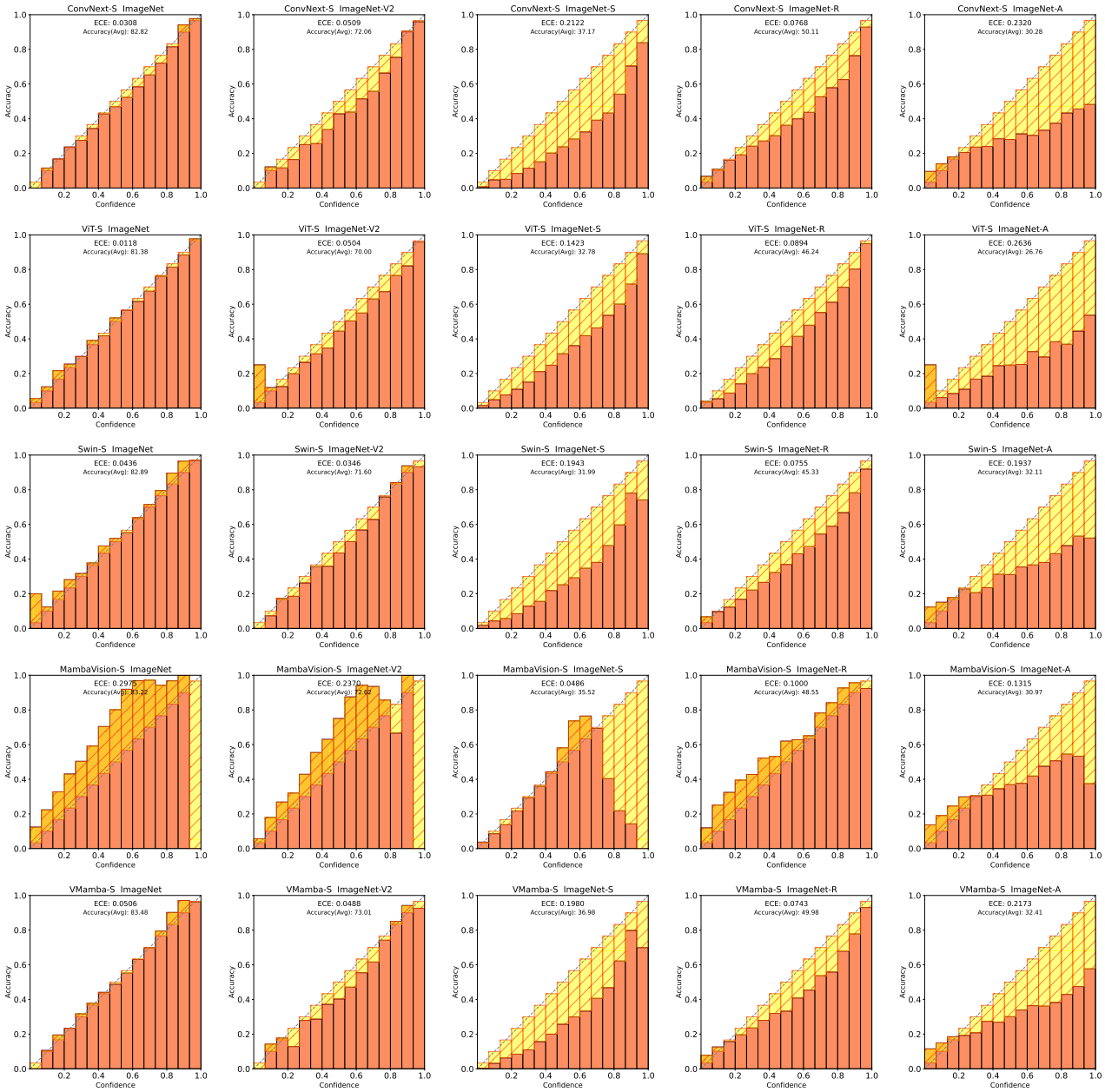


Figure 21. Caliberation Results: Reliability diagrams and ECE on ConvNext-S, ViT-S, Swin-S, VMamba-S, and MambaVision-S across ImageNet, ImageNet-V2, ImageNet-S, ImageNet-R, and ImageNet-A.

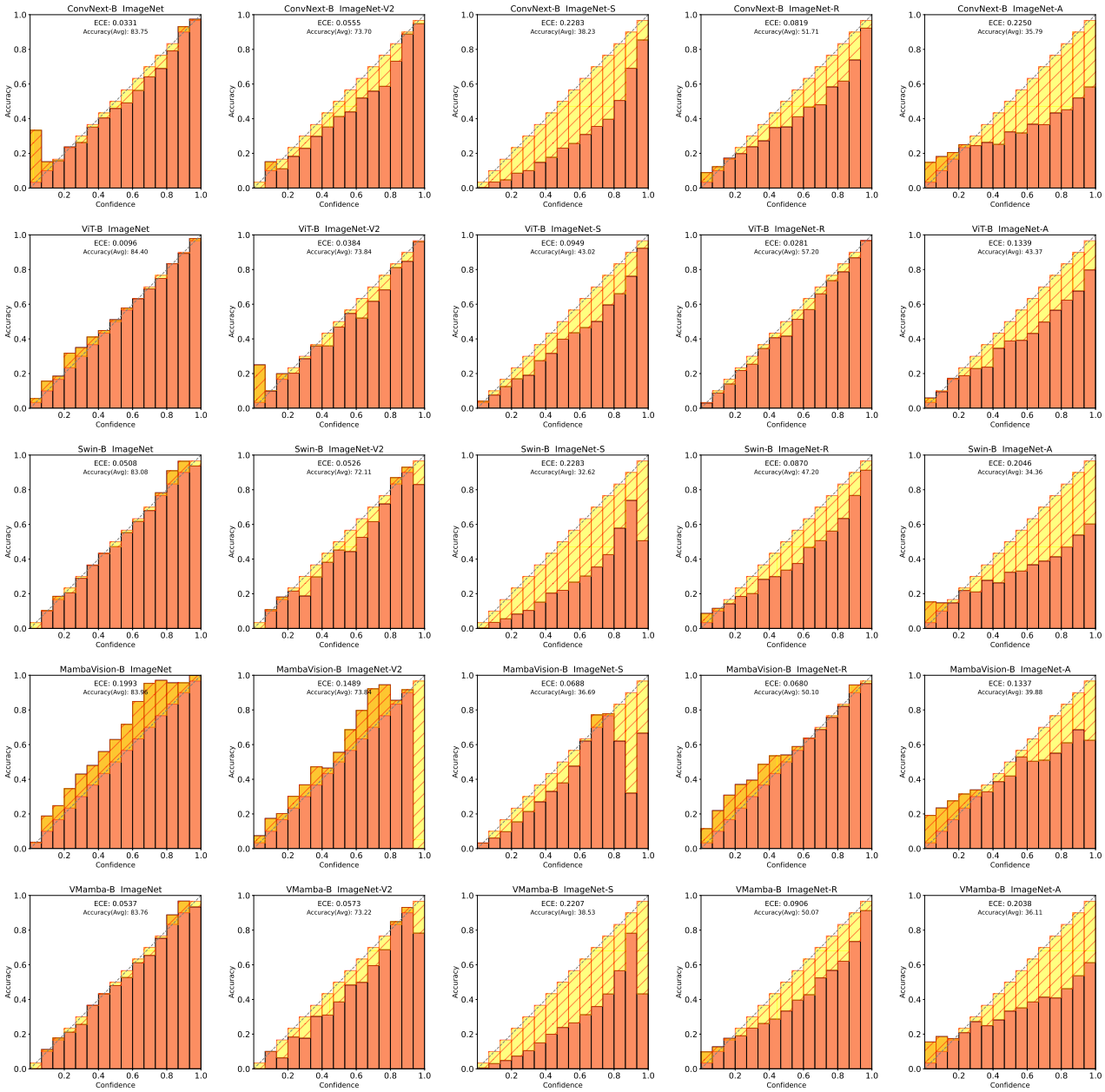


Figure 22. Caliberation Results: Reliability diagrams and ECE on ViT-B, Swin-B, VMamba-B, and MambaVision-B across ImageNet, ImageNet-V2, ImageNet-S, ImageNet-R, and ImageNet-A.

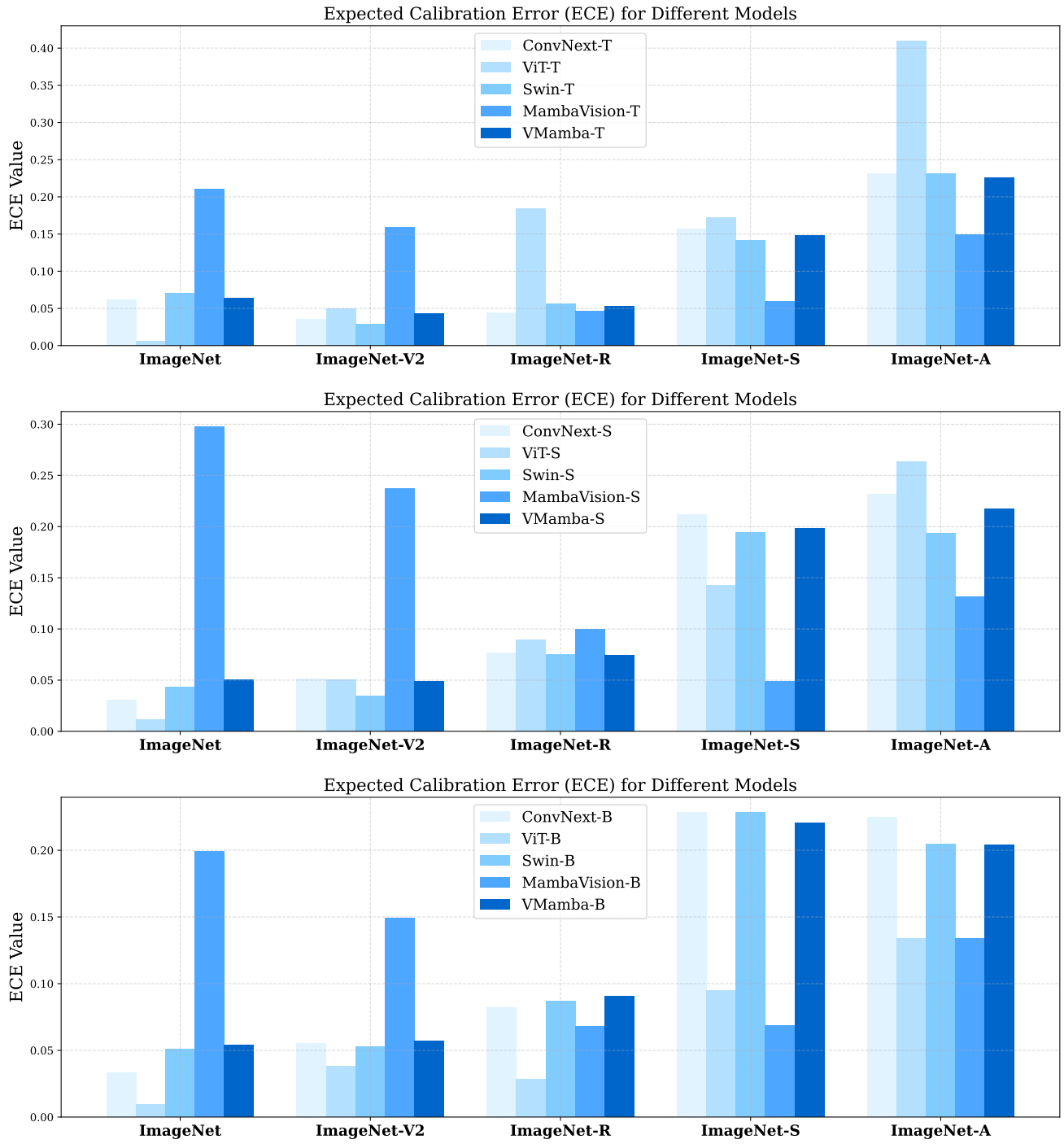


Figure 23. ECE error across classification models across ImageNet, ImageNet-V2, ImageNet-S, ImageNet-R, and ImageNet-A.

Table 9. 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 |
| VGG16 | 84.14 | 79.62 | 62.59 | 77.27 | 83.63 | 80.16 | 70.93 | 64.15 | 91.06 | 94.10 | 88.20 | 93.70 | 74.80 | 75.00 |
| VGG19 | 83.89 | 80.15 | 63.21 | 77.80 | 81.16 | 81.01 | 70.54 | 64.91 | 92.05 | 94.50 | 87.80 | 93.30 | 77.10 | 77.80 |
| DenseNet-121 | 84.93 | 81.78 | 62.36 | 79.55 | 85.15 | 81.17 | 70.08 | 64.50 | 92.78 | 96.20 | 90.20 | 95.10 | 81.10 | 80.00 |
| DenseNet-161 | 87.48 | 85.48 | 67.72 | 82.79 | 87.22 | 84.41 | 75.00 | 70.08 | 93.04 | 97.50 | 90.70 | 94.70 | 81.10 | 80.10 |
| 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 |
| DeiT-T | 80.68 | 77.21 | 51.24 | 71.83 | 80.30 | 76.83 | 65.03 | 59.54 | 89.94 | 91.60 | 86.50 | 90.50 | 73.90 | 73.20 |
| DeiT-S | 87.52 | 84.88 | 63.03 | 80.70 | 89.13 | 85.70 | 76.75 | 71.37 | 94.14 | 98.30 | 91.40 | 95.20 | 85.70 | 84.10 |
| DeiT-B | 89.66 | 86.79 | 68.77 | 84.26 | 91.10 | 89.19 | 80.31 | 77.25 | 95.38 | 98.80 | 92.30 | 96.00 | 86.70 | 84.30 |
| VMamba-T(v0) | 90.03 | 89.31 | 70.38 | 85.59 | 91.49 | 89.59 | 82.01 | 78.63 | 95.73 | 98.00 | 91.60 | 94.90 | 86.10 | 86.70 |
| VMamba-S(v0) | 90.53 | 90.76 | 73.39 | 87.78 | 93.13 | 90.71 | 83.92 | 80.42 | 96.16 | 99.10 | 92.80 | 95.40 | 89.40 | 88.50 |
| VMamba-B(v0) | 91.75 | 90.62 | 73.90 | 88.33 | 93.29 | 91.19 | 83.96 | 81.66 | 96.00 | 98.80 | 93.50 | 96.20 | 89.70 | 88.40 |
| VMamba-T(v2) | 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(v2) | 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(v2) | 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 | 87.40 |

Table 10. Average Precision (AP) scores for different architectures on the COCO-DC dataset [35], detailing results for small (APs), medium (APm), and large objects (APl).

| **Model** | **AP** | **APs** | **APm** | **APl** | **AP** | **APs** | **APm** | **APl** | **AP** | **APs** | **APm** | **APl** | **AP** | **APs** | **APm** | **APl** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Original** | | | | **Color** | | | | **Texture** | | | | **Average** | | | |
| **ConvNext-T** | 66.2 | 41.0 | 61.3 | 71.4 | 55.0$_{(-11.20)}$ | 26.3$_{(-14.70)}$ | 49.0$_{(-12.30)}$ | 60.6$_{(-10.80)}$ | 53.5$_{(-12.70)}$ | 26.5$_{(-14.50)}$ | 47.5$_{(-13.80)}$ | 60.5$_{(-10.90)}$ | 54.25$_{(-11.95)}$ | 26.40$_{(-14.60)}$ | 48.25$_{(-13.05)}$ | 60.55$_{(-10.85)}$ |
| **Swin-T** | 66.3 | 45.1 | 61.6 | 71.6 | 55.1$_{(-11.20)}$ | 29.2$_{(-15.90)}$ | 48.0$_{(-13.60)}$ | 62.0$_{(-09.60)}$ | 53.6$_{(-12.70)}$ | 31.4$_{(-13.70)}$ | 45.2$_{(-16.40)}$ | 61.4$_{(-10.20)}$ | 54.35$_{(-11.95)}$ | 30.30$_{(-14.80)}$ | 46.60$_{(-15.00)}$ | 61.70$_{(-09.90)}$ |
| **Swin-S** | 69.1 | 43.7 | 62.3 | 75.9 | 57.4$_{(-11.70)}$ | 30.1$_{(-13.60)}$ | 49.1$_{(-13.20)}$ | 64.3$_{(-11.60)}$ | 56.0$_{(-13.10)}$ | 26.8$_{(-16.90)}$ | 45.4$_{(-16.90)}$ | 64.3$_{(-11.60)}$ | 56.70$_{(-12.40)}$ | 28.45$_{(-15.25)}$ | 47.25$_{(-15.05)}$ | 64.30$_{(-11.60)}$ |
| **VSSM-T** | 69.3 | 47.1 | 64.4 | 75.5 | 56.2$_{(-13.10)}$ | 30.9$_{(-16.20)}$ | 48.4$_{(-16.00)}$ | 63.7$_{(-11.80)}$ | 53.2$_{(-16.10)}$ | 29.2$_{(-17.90)}$ | 44.9$_{(-19.50)}$ | 61.0$_{(-14.50)}$ | 54.57$_{(-14.73)}$ | 30.05$_{(-17.05)}$ | 46.65$_{(-17.75)}$ | 62.35$_{(-13.15)}$ |
| **VSSM-S** | 70.9 | 51.7 | 64.6 | 77.1 | 56.7$_{(-14.20)}$ | 31.3$_{(-20.40)}$ | 48.9$_{(-15.70)}$ | 64.3$_{(-12.80)}$ | 55.7$_{(-15.20)}$ | 26.8$_{(-24.90)}$ | 47.4$_{(-17.20)}$ | 64.4$_{(-12.70)}$ | 56.20$_{(-14.70)}$ | 29.05$_{(-22.65)}$ | 48.15$_{(-16.45)}$ | 64.35$_{(-12.75)}$ |

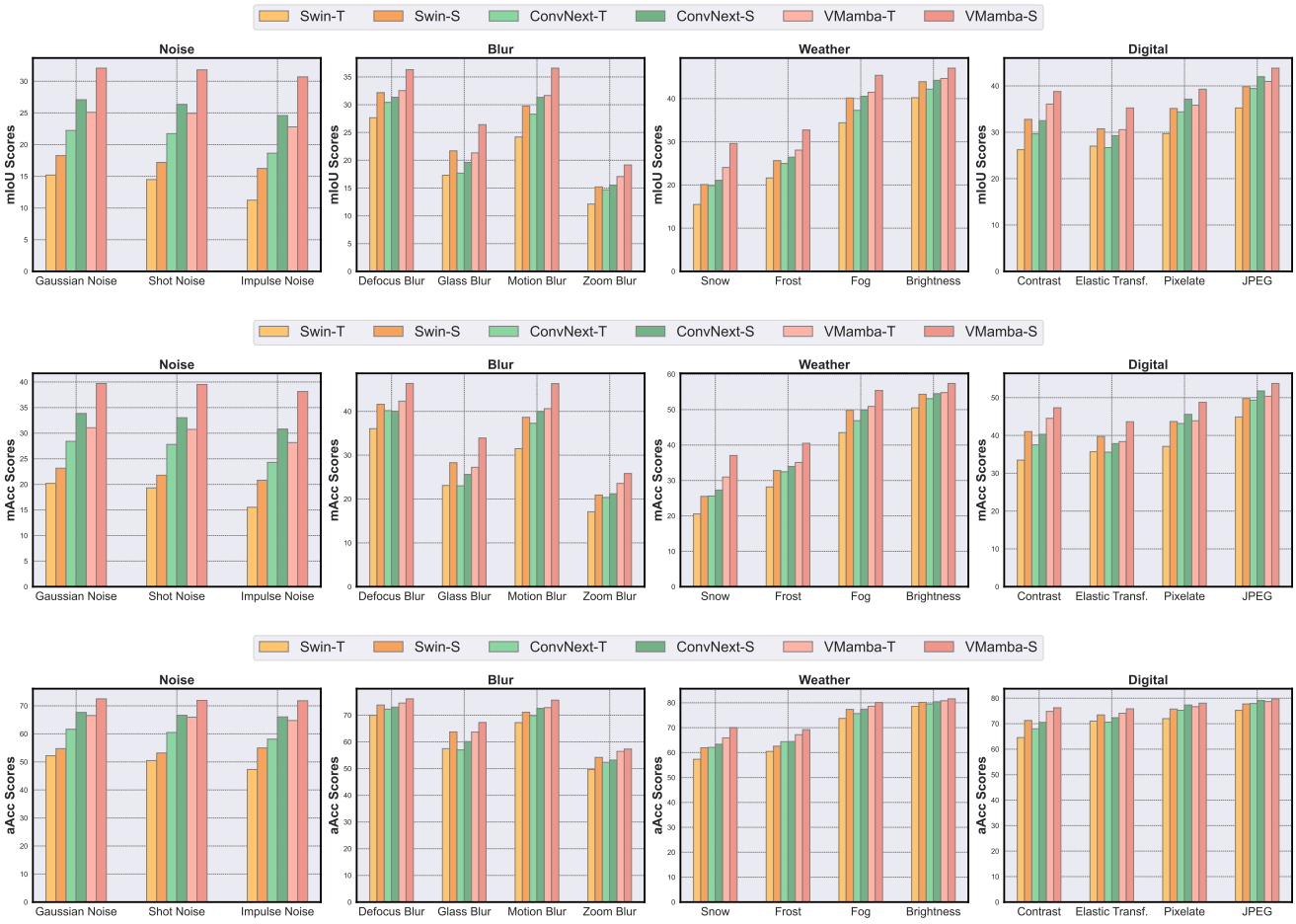


Figure 24. mIoU, mAcc, and aAcc score for different architectures on AED20k-C dataset

Table 11. Robust accuracy of various architectures against low-frequency and high-frequency-based perturbation using PGD attack at varying level of perturbation budget $\epsilon \in \{1/255, 2/255, 4/255, 8/255, 12/255, 16/255\}$.

| Filter → | **Low-Pass** | | | | | | **High-Pass** | | | | | | **All-Pass** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\epsilon$ → | 1/255 | 2/255 | 4/255 | 8/255 | 12/255 | 16/255 | 1/255 | 2/255 | 4/255 | 8/255 | 12/255 | 16/255 | 1/255 | 2/255 | 4/255 |
| ResNet-50 | 94.94 | 93.28 | 91.38 | 88.42 | 87.74 | 87.86 | 72.68 | 43.58 | 28.70 | 16.38 | 11.62 | 10.76 | 29.68 | 2.04 | 0.08 |
| ConvNext-T | 95.98 | 95.46 | 94.60 | 91.76 | 89.70 | 89.50 | 58.26 | 26.46 | 13.62 | 4.40 | 2.38 | 2.06 | 20.18 | 1.60 | 0.02 |
| ConvNext-S | 96.52 | 96.14 | 95.80 | 93.48 | 91.70 | 91.36 | 66.58 | 34.66 | 20.50 | 7.38 | 3.66 | 3.12 | 30.10 | 3.98 | 0.08 |
| ConvNext-B | 96.84 | 96.76 | 96.14 | 93.62 | 92.40 | 91.72 | 64.96 | 39.26 | 22.98 | 8.76 | 5.46 | 4.56 | 31.02 | 6.24 | 0.18 |
| ViT-T | 85.52 | 77.56 | 68.50 | 53.02 | 47.28 | 46.32 | 69.32 | 38.48 | 19.24 | 5.52 | 2.56 | 1.88 | 11.14 | 0.16 | 0.00 |
| ViT-S | 93.12 | 88.60 | 83.50 | 72.84 | 69.22 | 69.06 | 77.14 | 46.48 | 27.94 | 11.08 | 6.76 | 5.28 | 18.74 | 0.42 | 0.00 |
| ViT-B | 94.66 | 90.28 | 87.40 | 80.68 | 77.50 | 77.26 | 85.74 | 61.84 | 46.52 | 26.18 | 18.28 | 16.28 | 30.94 | 1.30 | 0.08 |
| VMamba-T | 96.46 | 96.04 | 95.22 | 93.02 | 91.70 | 91.18 | 58.12 | 32.50 | 16.46 | 5.74 | 2.92 | 2.82 | 24.00 | 5.16 | 0.28 |
| VMamba-S | 97.02 | 96.90 | 96.48 | 94.10 | 92.62 | 92.24 | 67.64 | 44.00 | 28.18 | 11.36 | 6.46 | 5.48 | 32.54 | 10.90 | 0.84 |
| VMamba-B | 97.28 | 97.00 | 96.56 | 94.28 | 92.90 | 92.24 | 66.60 | 45.44 | 27.68 | 10.36 | 6.08 | 4.78 | 33.10 | 9.88 | 0.40 |
| MambaVision-T | 96.44 | 95.98 | 95.18 | 93.02 | 92.40 | 92.54 | 60.52 | 35.32 | 22.90 | 9.00 | 4.88 | 3.86 | 18.38 | 3.04 | 0.16 |
| MambaVision-S | 96.74 | 96.26 | 95.44 | 93.56 | 92.84 | 93.10 | 67.60 | 44.56 | 33.24 | 17.48 | 11.14 | 9.24 | 23.80 | 4.46 | 0.30 |
| MambaVision-B | 96.80 | 96.24 | 95.82 | 93.86 | 93.04 | 93.22 | 70.28 | 45.68 | 33.20 | 18.36 | 12.86 | 11.24 | 28.80 | 6.40 | 0.72 |
| Swin-T | 96.24 | 95.86 | 95.24 | 93.28 | 91.94 | 91.30 | 49.12 | 23.42 | 9.78 | 2.06 | 1.10 | 0.88 | 12.60 | 1.86 | 0.00 |
| Swin-S | 96.92 | 96.60 | 96.16 | 94.26 | 93.30 | 92.70 | 60.70 | 36.50 | 20.18 | 7.48 | 3.90 | 3.30 | 25.92 | 6.74 | 0.42 |
| Swin-B | 96.82 | 96.94 | 96.24 | 94.72 | 94.10 | 93.34 | 61.04 | 40.50 | 24.68 | 9.30 | 5.82 | 4.34 | 30.14 | 9.76 | 0.70 |

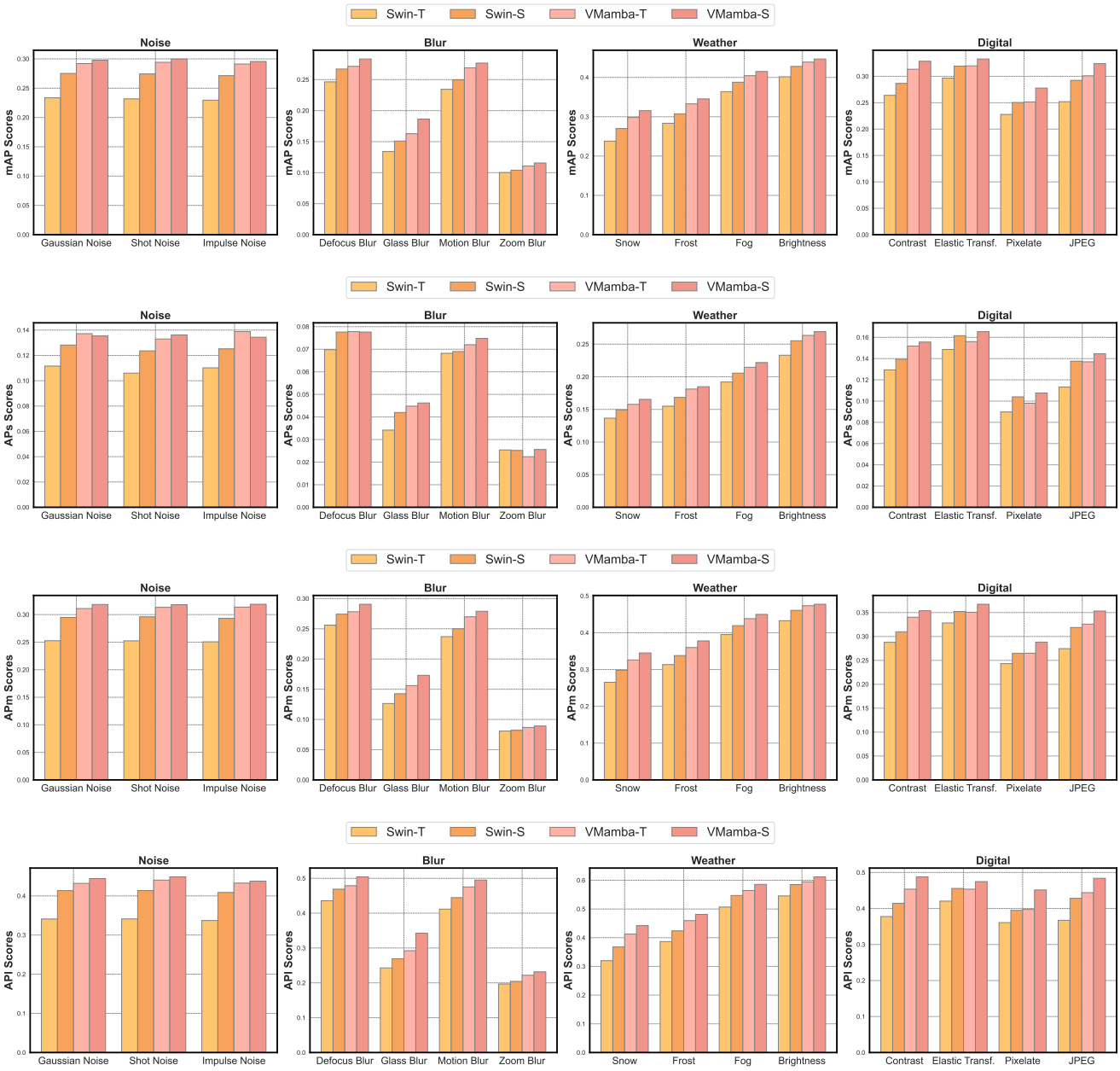


Figure 25. $mAP$, $AP_s$, $AP_m$, and $AP_l$ score for different architectures on COCO-C dataset

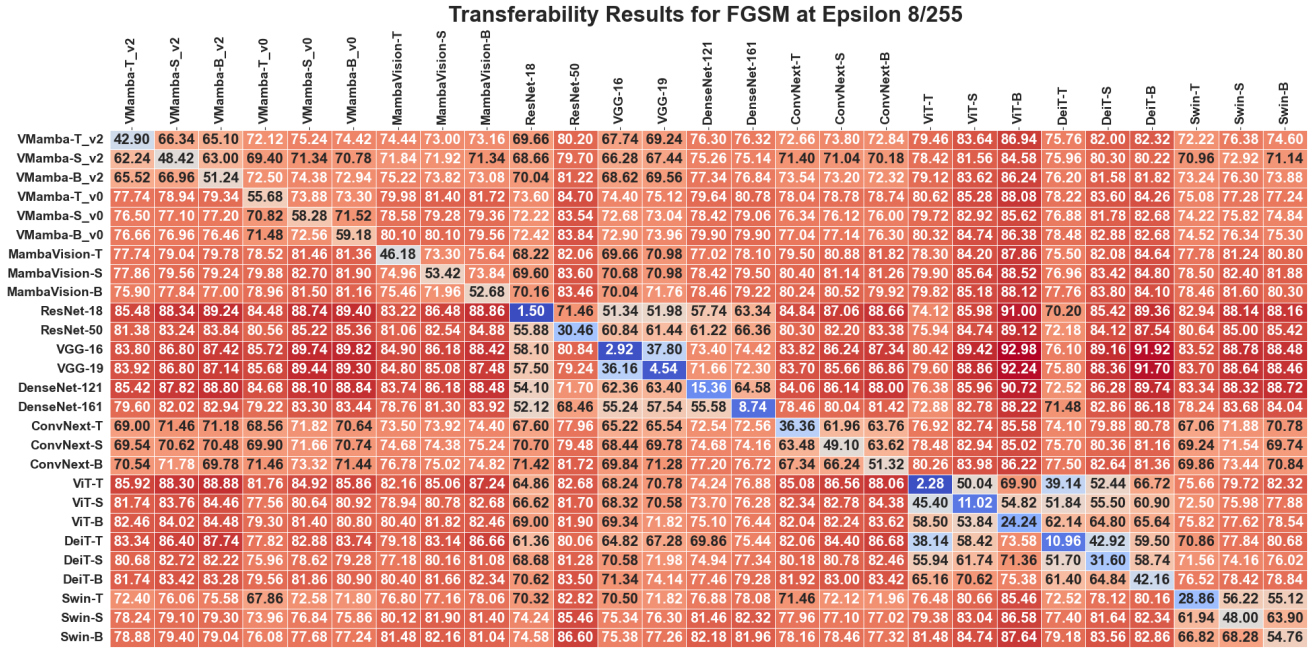


Figure 26. Robust accuracy of various architectures under white-box and black-box settings for FGSM attack. Adversarial examples are crafted at a perturbation budget $\epsilon = \frac{8}{255}$.

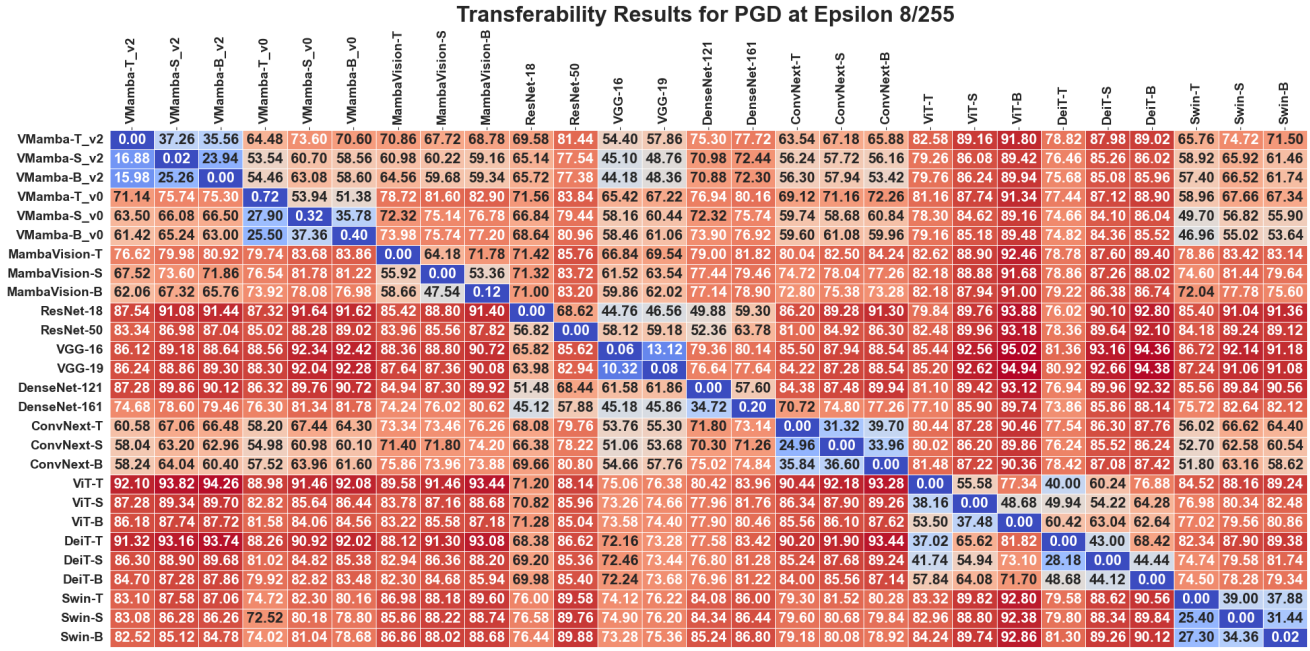


Figure 27. Robust accuracy of various architectures under white-box and black-box settings for PGD attack. Adversarial examples are crafted at a perturbation budget $\epsilon = \frac{8}{255}$.

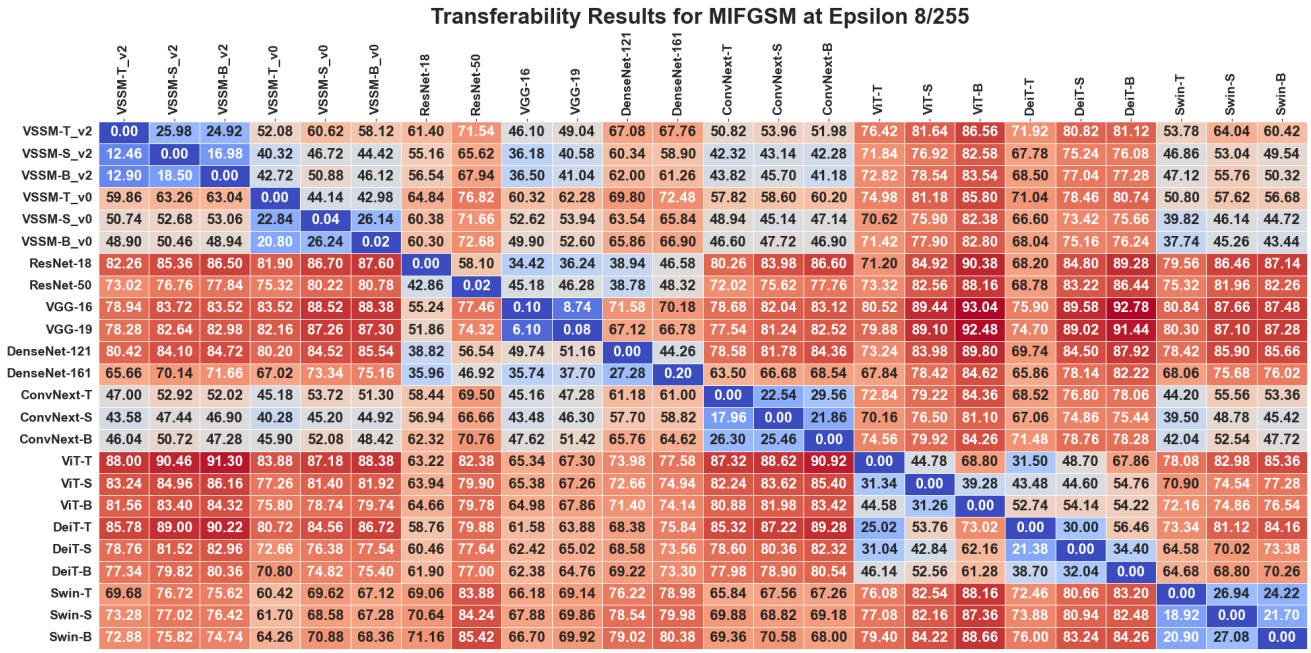


Figure 28. Robust accuracy of various architectures under white-box and black-box settings for MIFGSM attack. Adversarial examples are crafted at a perturbation budget $\epsilon = \frac{8}{255}$.