Appendix

A. Experimental Setup

Image Preprocessing. Following [28, 9], we directly resize images—ignoring aspect ratio and without cropping—to the dimension expected by each network as input, normalizing intensities to the appropriate range. This size is 384×384 for most models.

Adversarial Perturbations. For both FGSM and PGD, we compute the gradients with respect to input pixels of the cross-entropy loss against the correct image label, and then use the sign of these gradients to update the images. We clip the updated image intensities to lie within the valid range after each update. The step sizes and overall L_∞ norm of one gray level are translated to the expected intensity normalization of each model.

Spatial Adversarial Attacks. We provide additional details for the spatial adversarial grid attack that we employ in Sec. 3.6. We perturb each image with a discrete set of spatial transformations. The attack is considered successful if any of the transformed images is incorrectly classified by the corresponding model model. The same fixed set of $2511 (9 \times 9 \times 31)$ transformations was used for all images and all models. The perturbation set corresponds to the vertices of a grid, which is defined by an outer product of sampled values of three parameters: horizontal translation, vertical translation, and rotation. The samples are equally spaced within each parameter's range. Engstrom et al. [10] use 5 values each for horizontal and vertical translation. We use a denser set of 9 values, so as to explore the space of translations at a finer resolution, and a range of $[-16, 16]$ pixels as the translation ranges, so as to span the largest patch size (32×32) used by any of the ViT models. For rotations, we follow [10] using 31 values in the range $[-30°, 30°]$. When rotating and translating the images, we use bilinear interpolation and fill regions that lie outside the bounds of the original image with zeros (black pixels).

Restricted Attention. We pass masks to the Transformer attention layers to evaluate the effect of restricted attention. After the initial embedding layer, the image is transformed to a flat sequence of patches and a CLS token. To compute the pair-wise mask between all entries of that sequence, we consider the spatial location of the patches, and mask out pairs whose distance (on the patch grid, which is of size $384/16 \times 384/16$ for the models we consider) along the $x-$ or $y-$ axis is greater than the restricted distance. Our mask always allows attention between the CLS token and all patches.

B. Raw Accuracy Values

Table 3 shows the raw accuracy values corresponding to the results shown in Figs. 2 and 6.

C. ImageNet-C Detailed Results

The ImageNet-C benchmark [19] includes 15 types of synthetically generated corruptions, grouped into 4 categories: 'noise', 'blur', 'weather' and 'digital'. Each corruption type has five levels of severity, resulting in 75 distinct corruptions. The benchmark also includes an 'extra' group with additional corruptions. The results in Fig. 2 are averaged across 95 distinct corruptions (75 from the corruption groups, and 20 from the 'extra' group). In this section we provide more detailed results.

Corruption Groups. In Fig. 10 we show accuracy for each of the corruption groups: 'noise', 'blur', 'weather' and 'digital'. The results for each corruption group are averaged across all corruption types in the group, and over all severity levels.

	ILSVRC-2012	ImageNet-C	ImageNet-R	ImageNet-A	Conflict Stimuli
Models trained on ILSVRC-2012					
ResNet-50x1	76.80%	46.14%	21.45%	4.15%	23.91%
ResNet-101x1	78.00%	50.24%	23.00%	6.28%	24.49%
ViT-B/32	73.37%	48.77%	20.23%	3.80%	40.28%
ViT-B/16	77.91%	52.22%	21.90%	7.00%	38.26%
ResNet-50x3	80.00%	51.09%	23.62%	7.15%	24.27%
ViT-L/32	71.18%	47.73%	19.14%	3.31%	41.13%
ViT-L/16	76.50%	49.31%	17.87%	6.68%	23.30%
ResNet-101x3	80.30%	53.36%	24.47%	9.07%	31.05%
ResNet-152x4	80.40%	54.46%	25.82%	11.64%	39.14%
Models trained on ImageNet-21k					
ResNet-50x1	80.10%	52.97%	26.82%	9.92%	30.96%
ResNet-101x1	82.40%	58.08%	30.37%	15.73%	35.54%
ViT-B/32	81.27%	62.79%	31.79%	16.29%	48.45%
ViT-B/16	83.98%	65.83%	37.99%	26.65%	46.58%
ResNet-50x3	84.10%	59.87%	34.04%	20.40%	35.64%
ViT-L/32	81.03%	65.05%	34.33%	19.04%	58.26%
ViT-L/16	85.12%	70.03%	40.64%	28.12%	48.27%
ResNet-101x3	84.50%	63.07%	35.93%	24.97%	42.42%
ViT-H/14	85.11%	71.13%	39.99%	31.00%	45.93%
ResNet-152x4	84.80%	65.31%	39.67%	30.01%	45.34%
Models trained on JFT-300M					
ResNet-50x1	79.00%	51.26%	35.52%	11.35%	29.84%
ResNet-101x1	81.60%	59.02%	45.45%	18.77%	46.57%
ViT-B/32	80.72%	62.76%	43.17%	12.07%	58.99%
ViT-B/16	84.14%	67.70%	53.02%	25.49%	50.29%
ResNet-50x3	84.90%	64.16%	55.41%	31.15%	38.21%
ViT-L/32	84.40%	70.94%	55.38%	22.81%	61.91%
ViT-L/16	87.13%	76.08%	65.28%	44.00%	53.65%
ResNet-101x3	86.30%	69.15%	62.64%	40.21%	48.12%
ViT-H/14	88.33%	78.74%	69.91%	55.85%	50.44%
ResNet-152x4	87.30%	73.41%	68.52%	51.23%	67.39%

Table 3. **Raw Accuracy.** Accuracy of ViT and ResNet models on different datasets. ImageNet-C [19], ImageNet-R [18], and ImageNet-A [21] are designed to evaluate robustness in the presence of "natural corruptions", "naturally occurring distribution shifts", and "natural adversarial examples", respectively. Conflict Stimuli [12] is designed to evaluate the degree to which a model is biased towards relying on texture over shape for image classification. For ImageNet-C the accuracy is averaged across all corruption types (including corruptions in the 'extra' group), and over all severity levels.

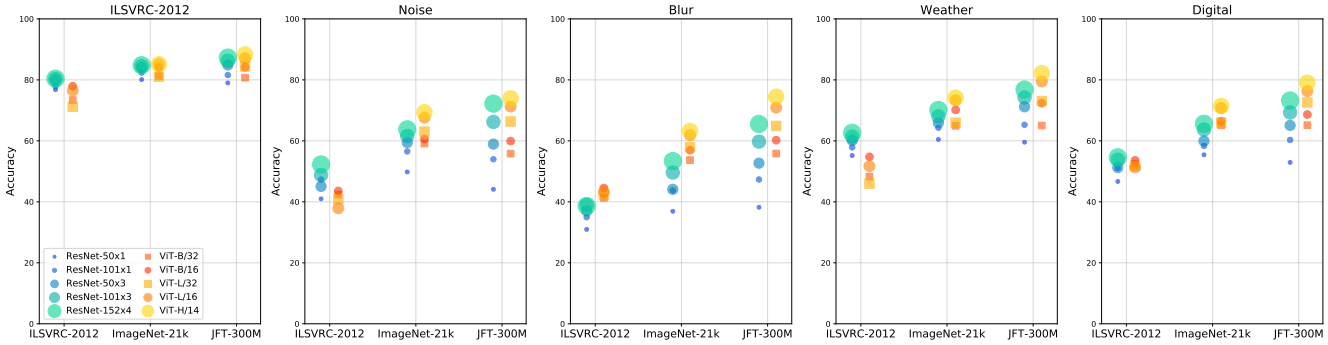

Figure 10. **ImageNet-C Corruption Groups**. Accuracy of ViT and ResNet models on ILSVRC-2012 (clean) and the different corruption groups in ImageNet-C. The accuracy for each group is averaged across all corruption types in the group, and over all severity levels.

Individual Corruptions and Severities. Next, Figures 11-29 show detailed results for the 95 distinct corruptions in the ImageNet-C benchmark.

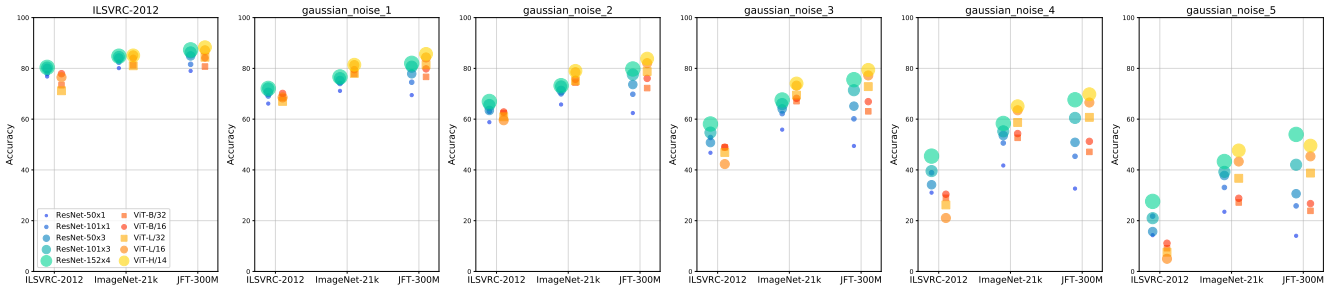

Figure 11. **ImageNet-C 'gaussian noise'**. Accuracy of ViT and ResNet models on ILSVRC-2012 (clean) and 'gaussian noise'.

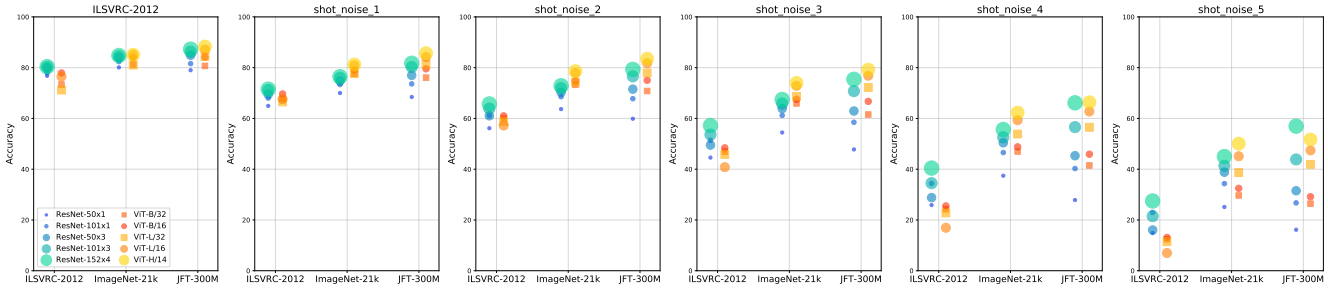

Figure 12. **ImageNet-C 'shot noise'**. Accuracy of ViT and ResNet models on ILSVRC-2012 (clean) and 'shot noise'.

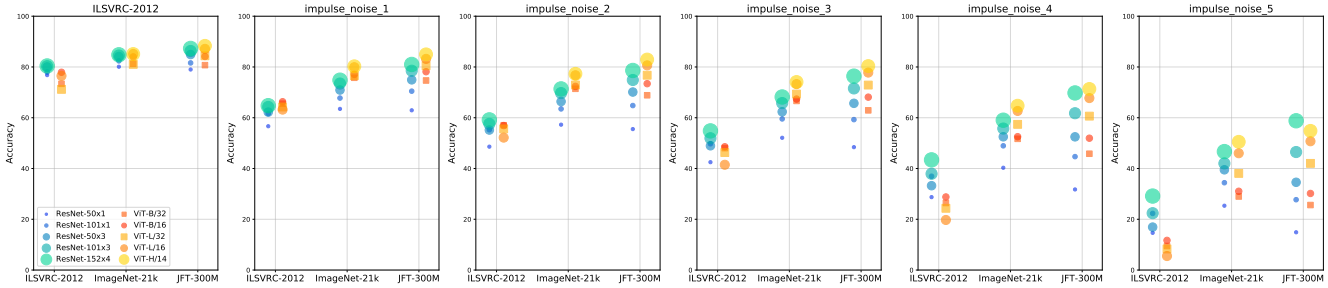

Figure 13. **ImageNet-C 'impulse noise'**. Accuracy of ViT and ResNet models on ILSVRC-2012 (clean) and 'impulse noise'.

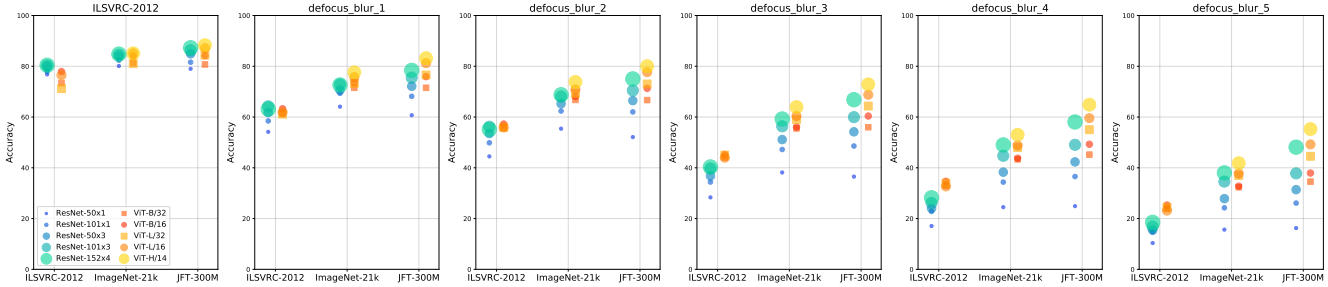

Figure 14. **ImageNet-C 'defocus blur'**. Accuracy of ViT and ResNet models on ILSVRC-2012 (clean) and 'defocus blur'.

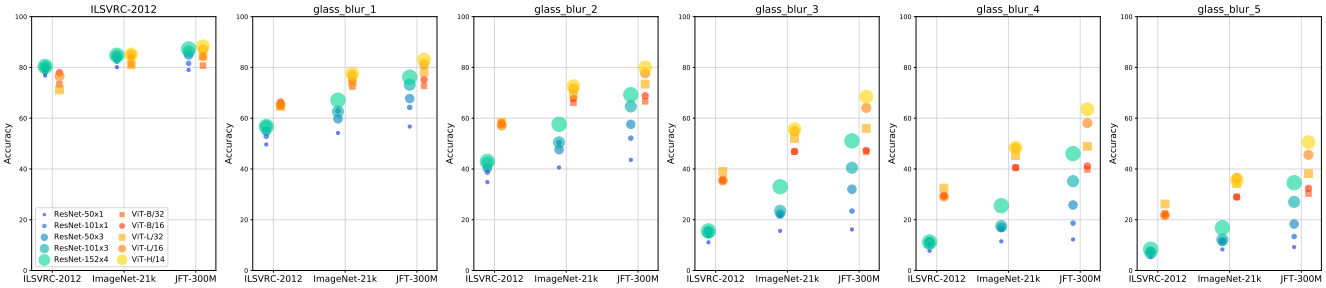

Figure 15. **ImageNet-C 'glass blur'**. Accuracy of ViT and ResNet models on ILSVRC-2012 (clean) and 'glass blur'.

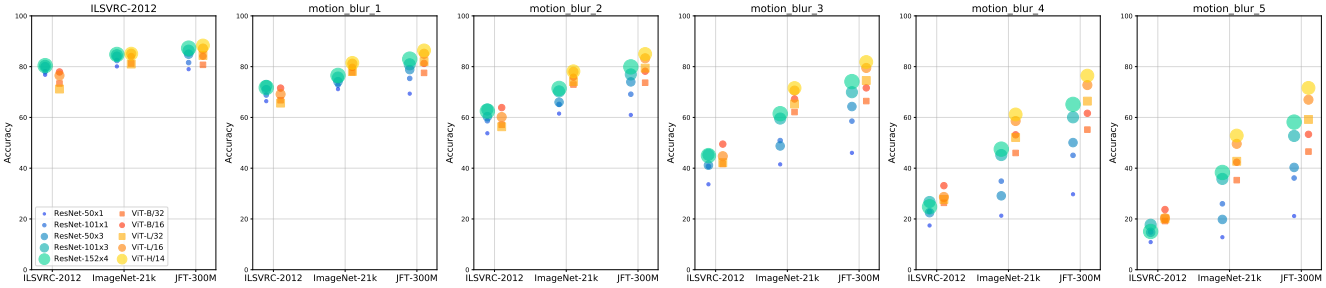

Figure 16. **ImageNet-C 'motion blur'**. Accuracy of ViT and ResNet models on ILSVRC-2012 (clean) and 'motion blur'.

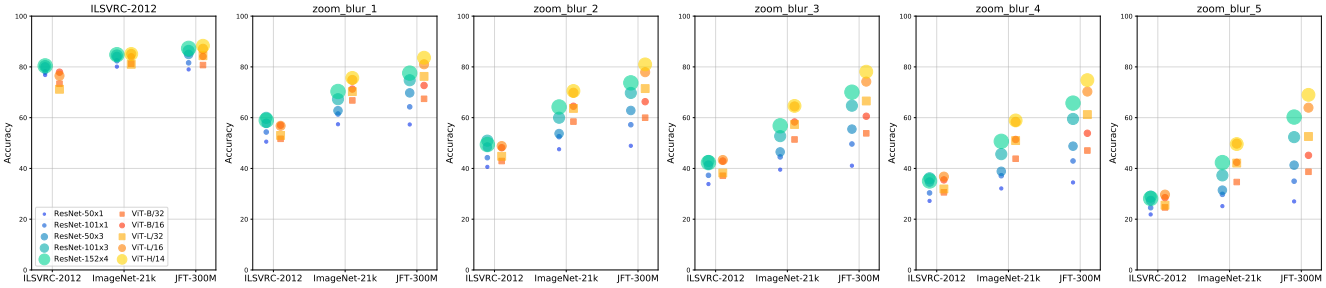

Figure 17. **ImageNet-C 'zoom blur'**. Accuracy of ViT and ResNet models on ILSVRC-2012 (clean) and 'zoom blur'.

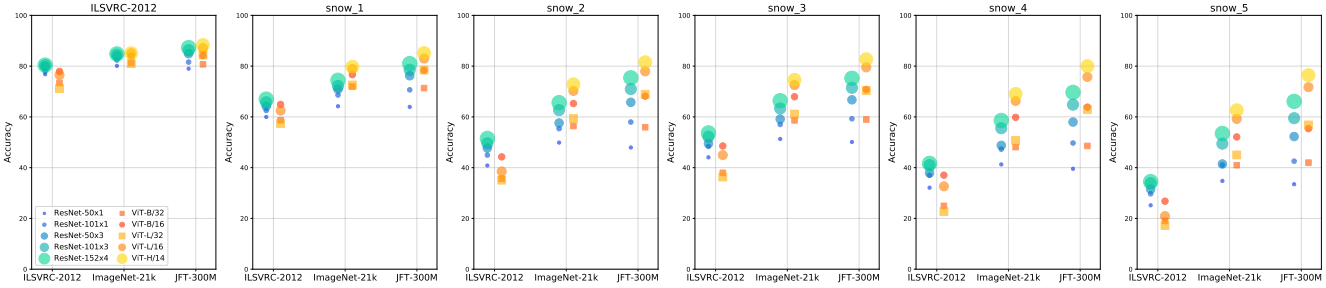

Figure 18. **ImageNet-C 'snow'**. Accuracy of ViT and ResNet models on ILSVRC-2012 (clean) and 'snow'.

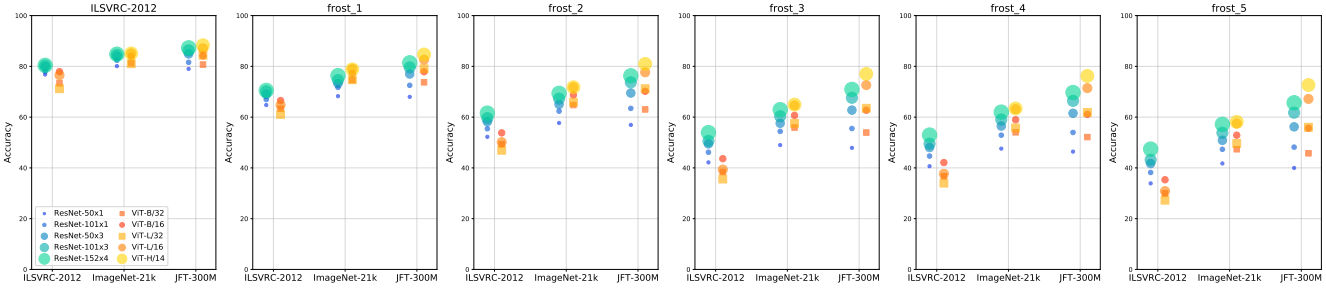

Figure 19. **ImageNet-C 'frost'**. Accuracy of ViT and ResNet models on ILSVRC-2012 (clean) and 'frost'.

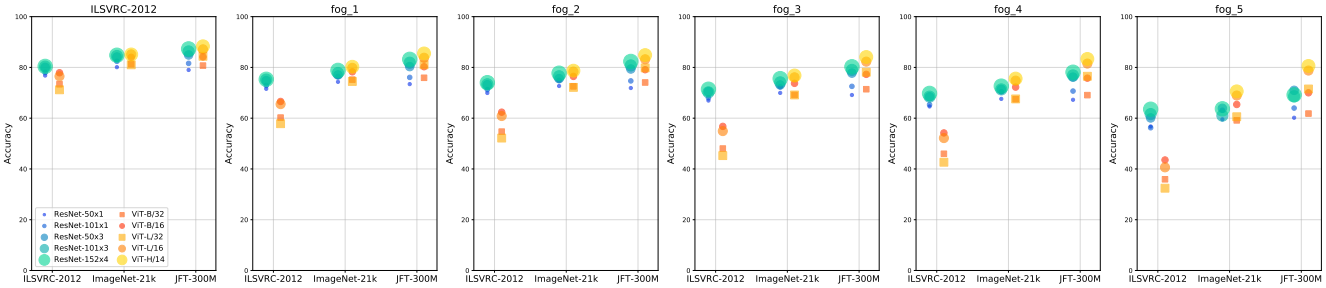

Figure 20. **ImageNet-C 'fog'**. Accuracy of ViT and ResNet models on ILSVRC-2012 (clean) and 'fog'.

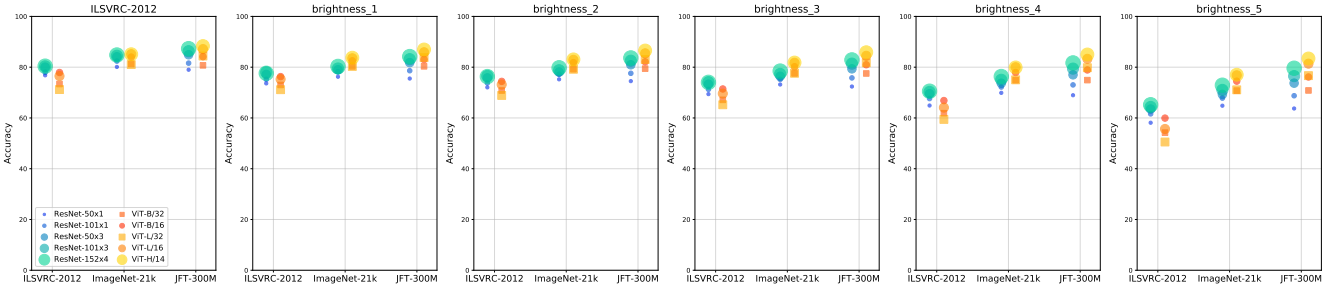

Figure 21. **ImageNet-C 'brightness'**. Accuracy of ViT and ResNet models on ILSVRC-2012 (clean) and 'brightness'.

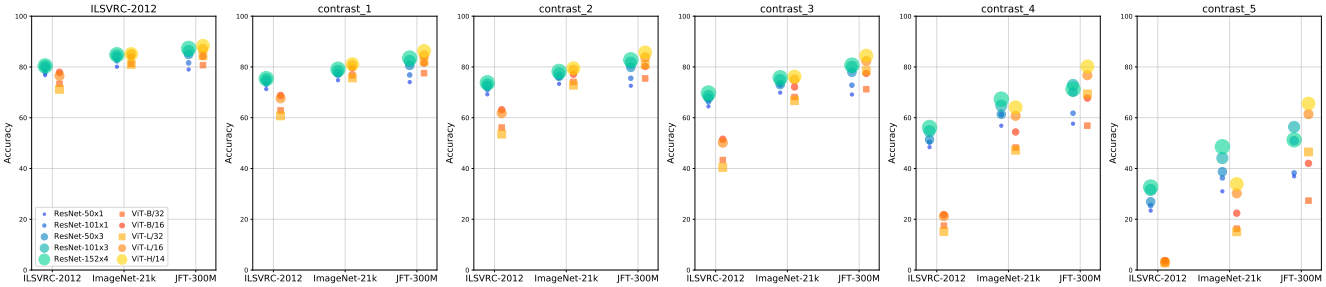

Figure 22. **ImageNet-C 'contrast'**. Accuracy of ViT and ResNet models on ILSVRC-2012 (clean) and 'contrast'.

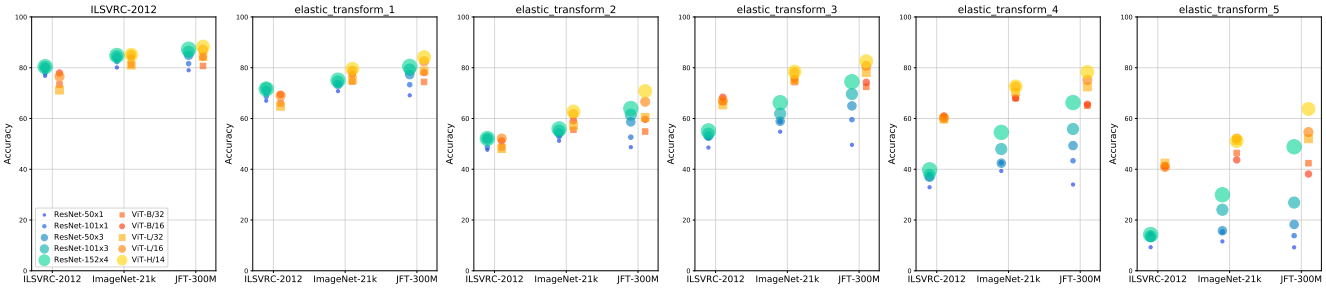

Figure 23. **ImageNet-C 'elastic transform'**. Accuracy of ViT and ResNet models on ILSVRC-2012 (clean) and 'elastic transform'.

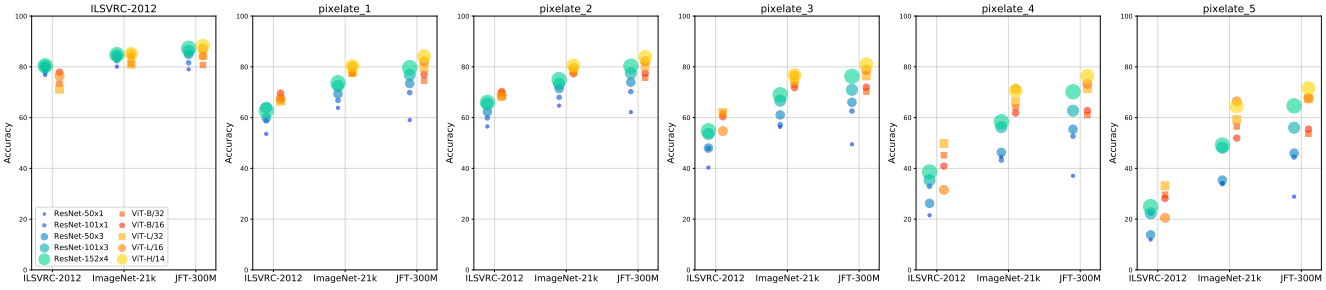

Figure 24. **ImageNet-C 'pixelate'**. Accuracy of ViT and ResNet models on ILSVRC-2012 (clean) and 'pixelate'.

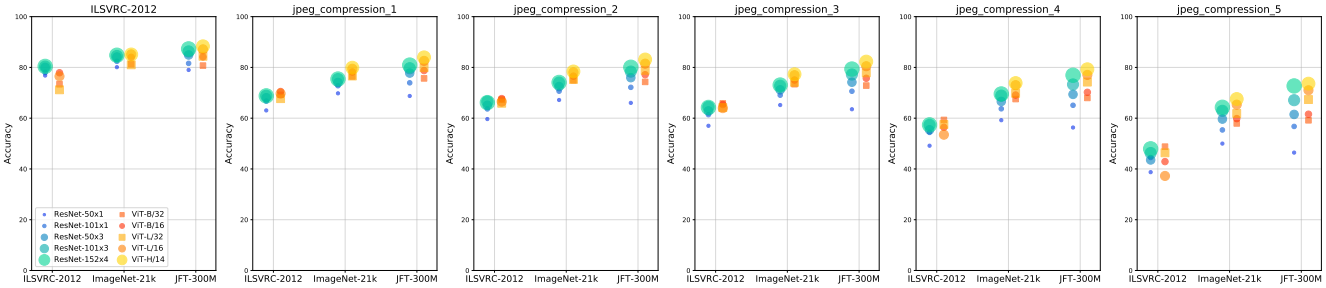

Figure 25. **ImageNet-C 'jpeg compression'**. Accuracy of ViT and ResNet models on ILSVRC-2012 (clean) and 'jpeg compression'.

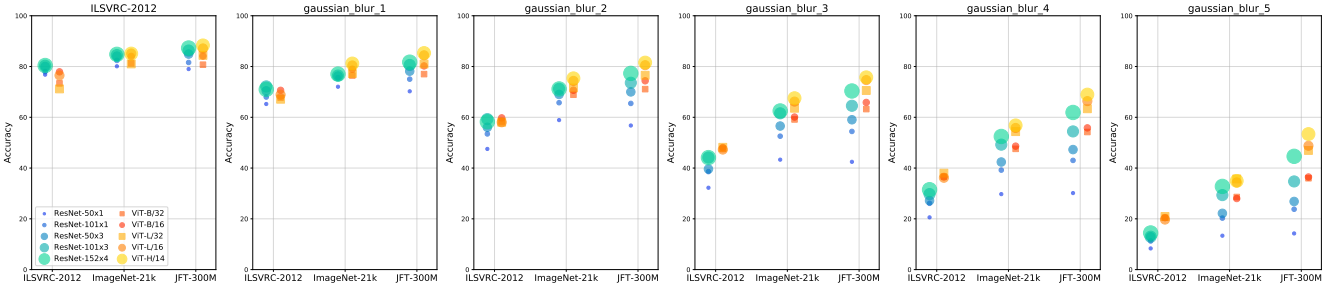

Figure 26. **ImageNet-C 'gaussian blur'**. Accuracy of ViT and ResNet models on ILSVRC-2012 (clean) and 'gaussian blur'.

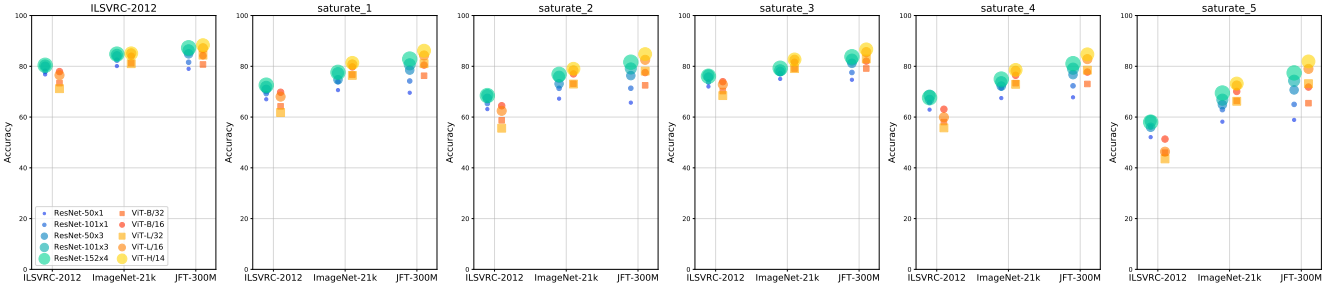

Figure 27. **ImageNet-C 'saturate'**. Accuracy of ViT and ResNet models on ILSVRC-2012 (clean) and 'saturate'.

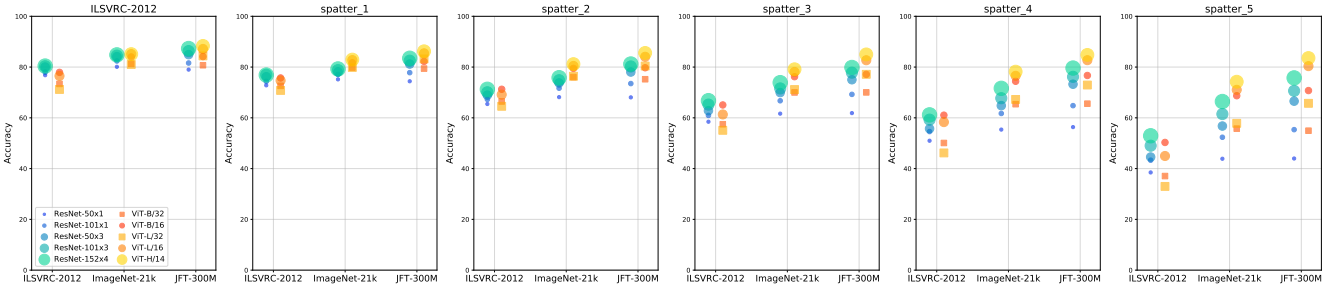

Figure 28. **ImageNet-C 'spatter'**. Accuracy of ViT and ResNet models on ILSVRC-2012 (clean) and 'spatter'.

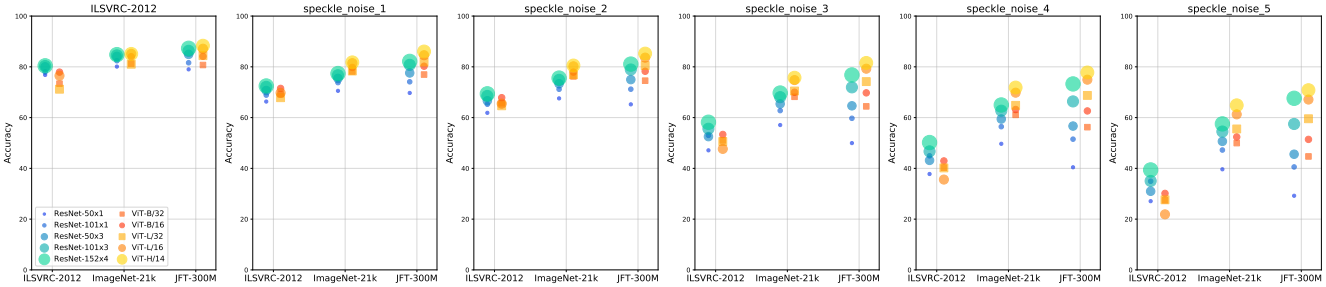

Figure 29. **ImageNet-C 'speckle noise'**. Accuracy of ViT and ResNet models on ILSVRC-2012 (clean) and 'speckle noise'.

D. Robustness Scaling on ImageNet-21k

As we pointed out in Sec. 3, the size of the pretraining dataset has a fundamental effect on the model's robustness, especially for ViTs. We compared the scaling of robustness on various benchmarks when all models are trained on ImageNet-21k in Fig. 30. With this smaller pretraining set, scaling up the ViT models does not offer better gains compared to scaling up ResNets in most cases, with ImageNet-C being an exception.

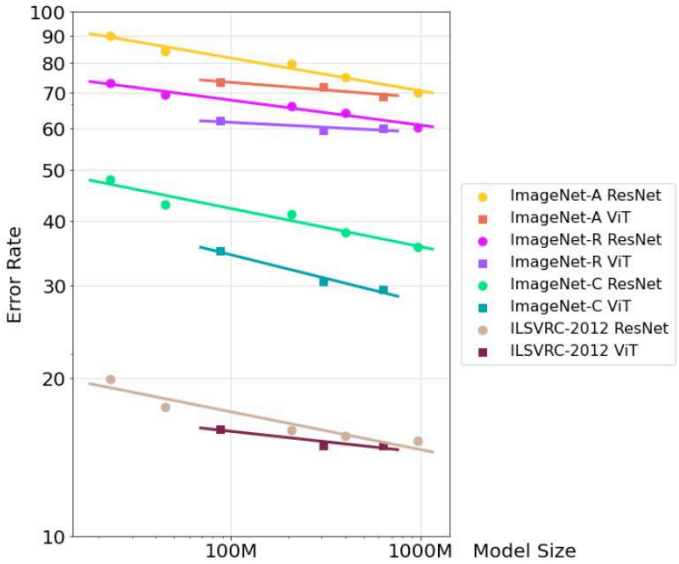

Figure 30. **Performance of ViT and ResNet Models on Different Datasets as a Function of the Number of Model Parameters**. All models are pre-trained on ImageNet-21k and fine-tuned on ILSVRC-2012. We observe that this pretraining dataset is not large enough for the ViT models to exhibit better robustness scaling.

E. Adversarial Perturbations: Accuracies with Self and Cross-Over Attacks

In Tables 4 and 5, we provide a full evaluation showing that adversarial perturbations computed using ViT models fail to cause incorrect outputs with ResNet models and vice-versa. For different variants of the ViT and ResNet models trained on different amounts of data, we report accuracies under adversarial perturbations—computed using PGD and FGSM—when evaluated using the models used to compute the perturbations vs. using a different model type (ViT or ResNet).

	ViT (clean)	RN (clean)	ViT→ViT	RN→RN	ViT→RN	RN→ViT
Models trained on ILSVRC-2012						
ViT-B/16 vs. RN-101x1	77.8%	77.4%	14.3%	2.0%	75.5%	77.4%
ViT-B/32 vs. RN-50x3	73.0%	80.6%	17.2%	4.8%	79.0%	72.0%
ViT-L/16 vs. RN-101x3	75.4%	80.2%	13.0%	7.3%	78.8%	74.1%
ViT-L/32 vs. RN-152x4	71.5%	80.0%	16.0%	10.5%	79.1%	71.0%
Models trained on ImageNet-21k						
ViT-B/16 vs. RN-101x1	83.4%	82.1%	10.3%	10.5%	81.0%	82.9%
ViT-B/32 vs. RN-50x3	81.1%	84.4%	15.9%	13.0%	83.7%	80.7%
ViT-L/16 vs. RN-101x3	85.3%	85.8%	16.2%	14.4%	83.5%	84.7%
ViT-L/32 vs. RN-152x4	81.9%	84.6%	24.7%	13.9%	84.0%	82.0%
Models trained on JFT-300M						
ViT-B/16 vs. RN-101x1	85.6%	82.2%	5.4%	8.1%	79.7%	85.2%
ViT-B/32 vs. RN-50x3	81.2%	83.9%	9.5%	13.4%	82.2%	80.9%
ViT-L/16 vs. RN-101x3	86.4%	86.1%	19.7%	19.5%	84.3%	85.8%
ViT-L/32 vs. RN-152x4	86.7%	86.1%	17.6%	17.9%	85.4%	86.5%

Table 4. **Accuracy with PGD attacks.** We include the full set of results for adversarial perturbations applied to various ViT and ResNet (RN) models trained on different datasets. Shown here are accuracies on a subset of 1000 images of the ILSVRC-2012 validation set (as used in Fig. 3 and Table 2)—on the original images as well as under adversarial perturbations computed using the ViT and ResNet models with PGD, when applied to the models used to compute the perturbations themselves, and when running inference on ResNet models using perturbations computed with ViT models and vice-versa.

	ViT (clean)	RN (clean)	ViT→ViT	RN→RN	ViT→RN	RN→ViT
Models trained on ILSVRC-2012						
ViT-B/16 vs. RN-101x1	77.8%	77.4%	30.6%	14.7%	75.4%	77.2%
ViT-B/32 vs. RN-50x3	73.0%	80.6%	31.5%	22.1%	78.5%	71.9%
ViT-L/16 vs. RN-101x3	75.4%	80.2%	27.8%	23.6%	78.3%	73.7%
ViT-L/32 vs. RN-152x4	71.5%	80.0%	26.3%	33.3%	78.1%	71.2%
Models trained on ImageNet-21k						
ViT-B/16 vs. RN-101x1	83.4%	82.1%	31.3%	33.5%	80.7%	82.7%
ViT-B/32 vs. RN-50x3	81.1%	84.4%	36.7%	42.2%	83.3%	80.4%
ViT-L/16 vs. RN-101x3	85.3%	85.8%	40.5%	44.4%	83.4%	85.1%
ViT-L/32 vs. RN-152x4	81.9%	84.6%	41.6%	47.2%	84.2%	81.5%
Models trained on JFT-300M						
ViT-B/16 vs. RN-101x1	85.6%	82.2%	22.9%	30.6%	79.9%	84.4%
ViT-B/32 vs. RN-50x3	81.2%	83.9%	25.7%	40.0%	82.5%	80.4%
ViT-L/16 vs. RN-101x3	86.4%	86.1%	43.2%	48.8%	83.4%	85.3%
ViT-L/32 vs. RN-152x4	86.7%	86.1%	40.9%	52.7%	85.5%	85.7%

Table 5. **Accuracy with FGSM attacks.** Version of Table 4 with perturbations computed using FGSM instead of PGD.

F. Layer Correlation Analysis

This section includes extended results for layer correlation study. Figure 31 shows the correlation of representations across Transformer blocks for four different ViT models (ViT-B/32, ViT-B/16, ViT-L/32, ViT-L/16) and three different pre-training datasets (ILSVRC-2012, ImageNet 21k, JFT-300M). Figure 32 shows the correlation for the same models when only taking the CLS token into account. As a comparative reference, Fig. 33 shows the correlation of representations across residual network blocks for three different ResNet models (ResNet-50x3, ResNet-101x1, ResNet-101x3) and the same three different pre-training datasets (ILSVRC-2012, ImageNet 21k, JFT-300M).

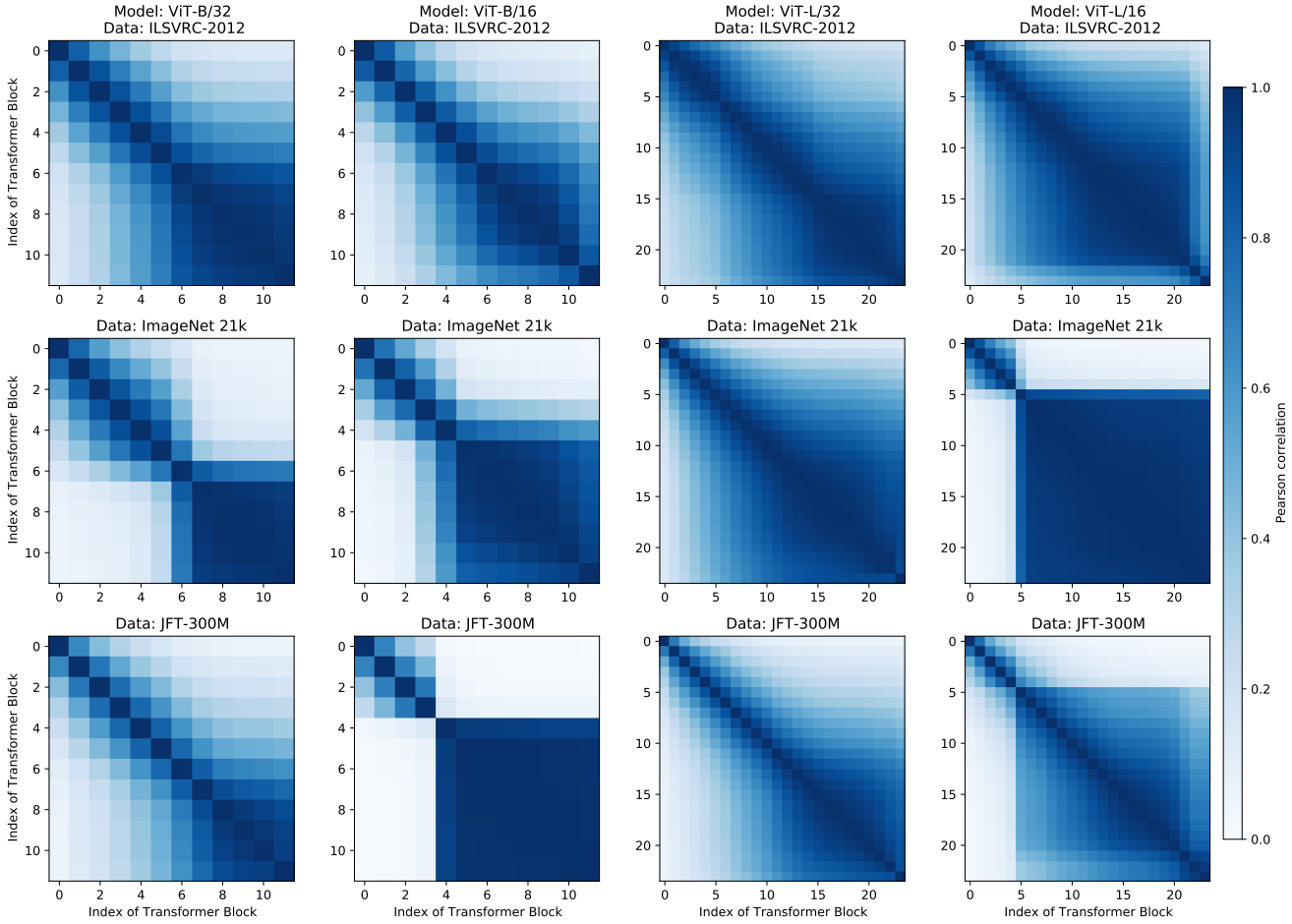

Figure 31. **Correlation of Representations Across Transformer Blocks**. We compare the representations (hidden features) after each Transformer block to those of all other blocks. We compare the similarity across representations using the absolute value of the Pearson correlation coefficient. We compare four different ViT models (ViT-B/32, ViT-B/16, ViT-L/32, ViT-L/16) and three different pre-training datasets (ILSVRC-2012, ImageNet 21k, JFT-300M). All models are fine-tuned on ILSVRC-2012 and activations are calculated on a random subset of 4096 samples from the ILSVRC-2012 validation set. White: no correlation. Dark Blue: $|\rho| = 1$.

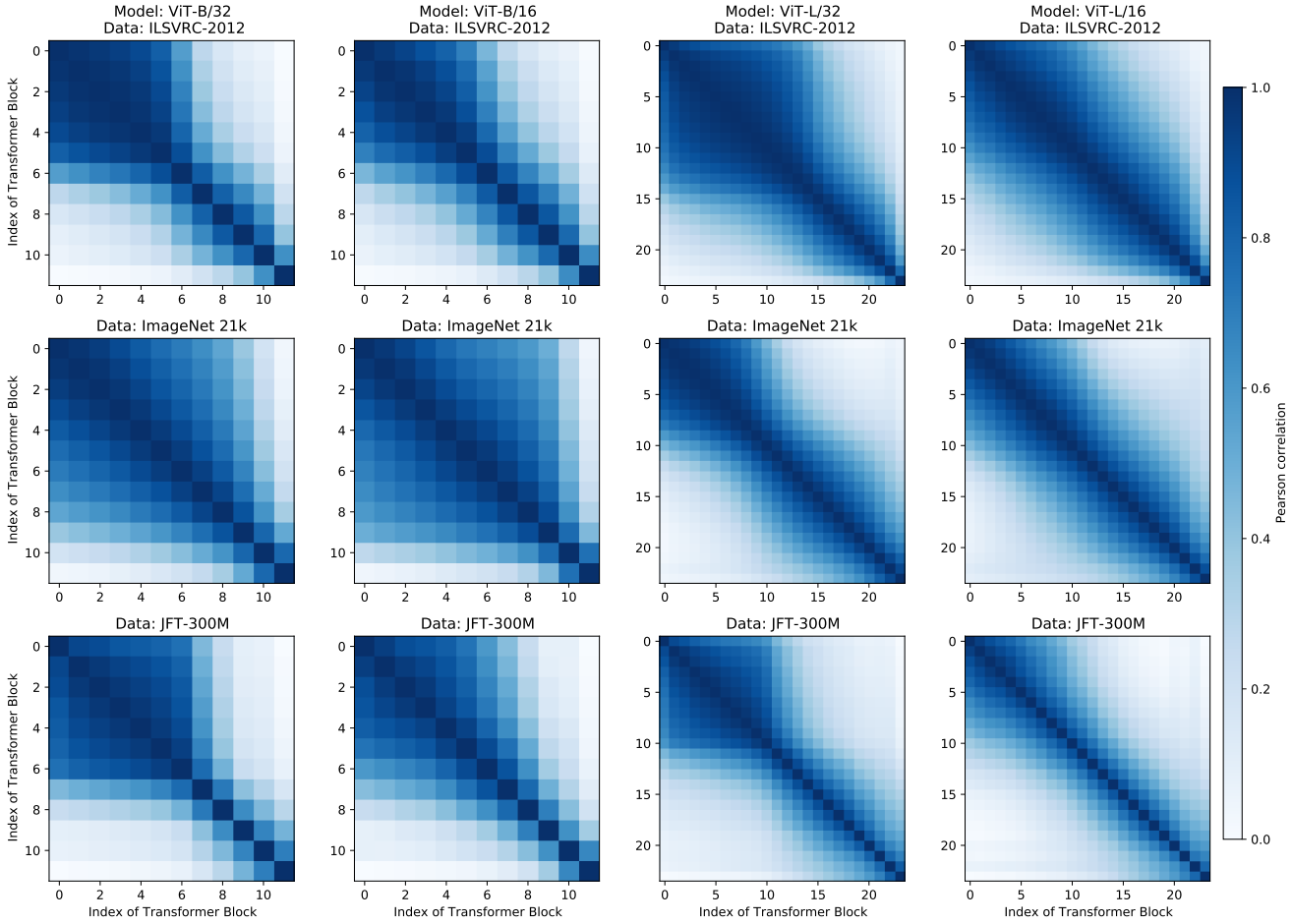

Figure 32. **Correlation of CLS Tokens Across Transformer Blocks**. We compare the representations (hidden features) of the CLS token after each Transformer block to those of all other blocks. We compare the similarity across representations using the absolute value of the Pearson correlation coefficient. We compare four different ViT models (ViT-B/32, ViT-B/16, ViT-L/32, ViT-L/16) and three different pre-training datasets (ILSVRC-2012, ImageNet 21k, JFT-300M). All models are fine-tuned on ILSVRC-2012 and activations are calculated on a random subset of 4096 samples from the ILSVRC-2012 validation set. White: no correlation. Dark Blue: $|\rho| = 1$.

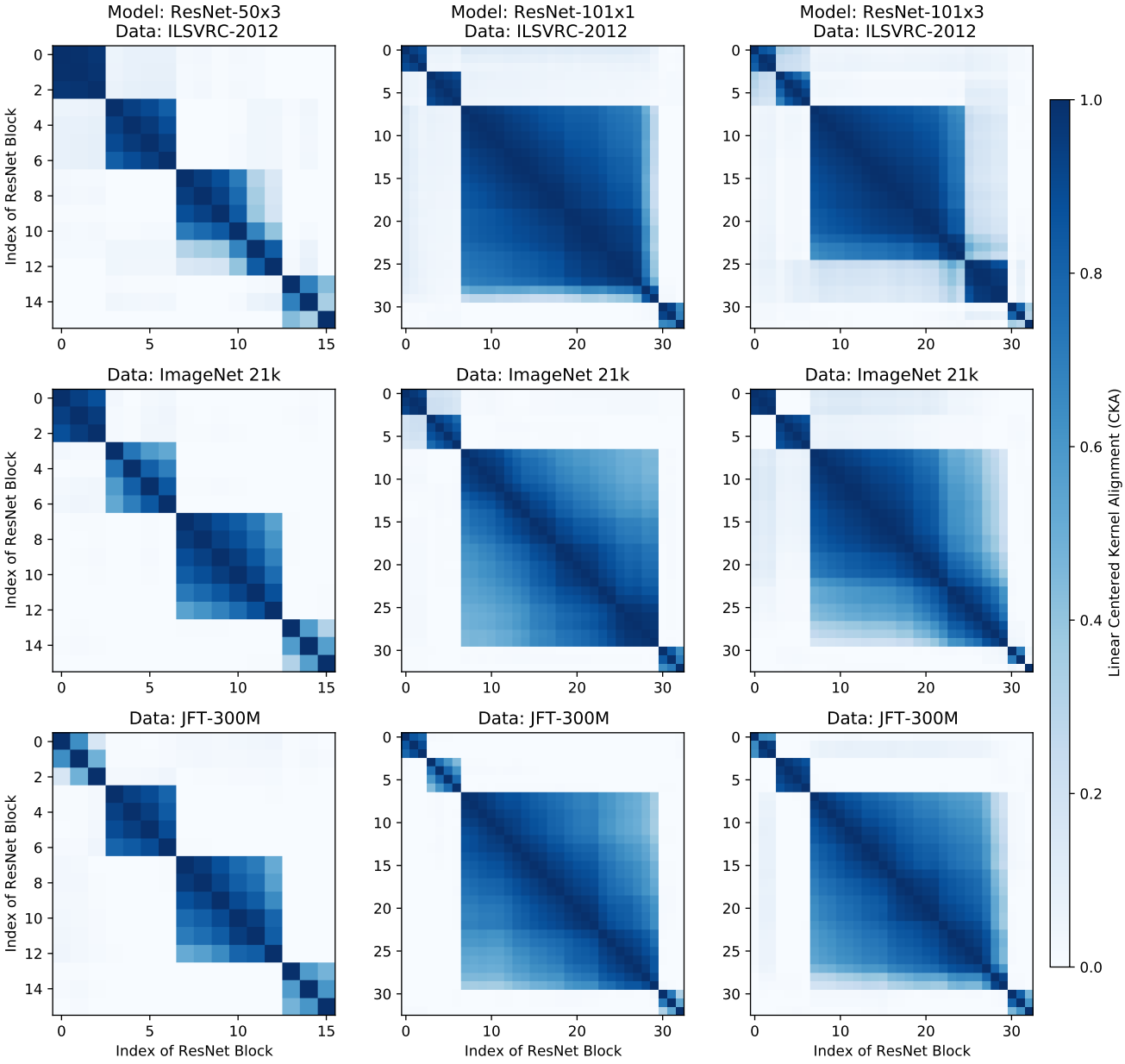

Figure 33. **Correlation of Representations Across Residual Network Blocks**. We compare the representations (hidden features) after each residual block to those of all other blocks in a residual network. We compare the similarity across representations using Linear Centered Kernel Analysis. This allows for the comparison of representations across stages which are of different shape. We compare three different ResNet models (ResNet-50x3, ResNet-101x1, ResNet-101x3) and three different pre-training datasets (ILSVRC-2012, ImageNet 21k, JFT-300M). The factor 'x' represents a multiplier on the number of channels for the ResNet models. All models are fine-tuned on ILSVRC-2012 and activations are calculated on a random subset of 4096 samples from the ILSVRC-2012 validation set.

G. Lesion Study

This section includes extended results for lesion study. Figure 34 shows an evaluation of ViT models when individual blocks, MLP layers or Self-Attention layers are removed from the model after training. We compare four different ViT models (ViT-B/32, ViT-B/16, ViT-L/32, ViT-L/16) and three different pre-training datasets (ILSVRC-2012, ImageNet 21k, JFT-300M). All models are fine-tuned on ILSVRC-2012.

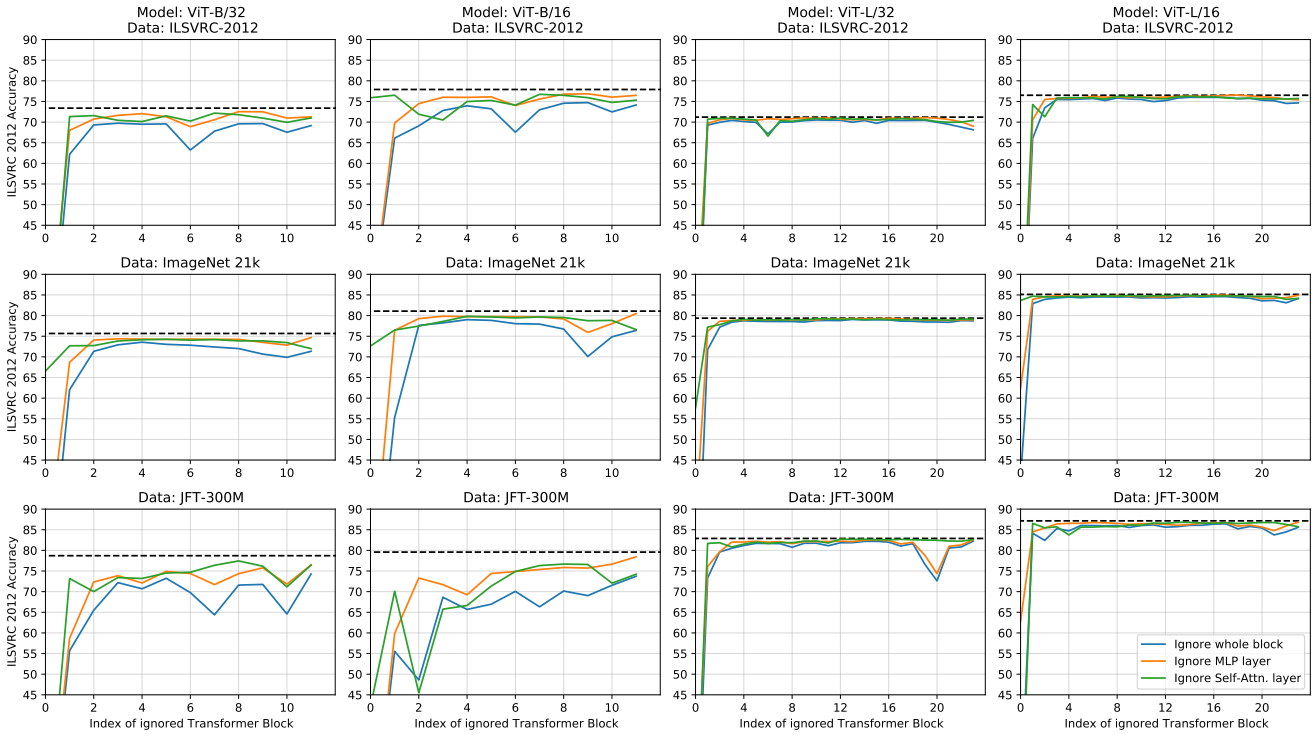

Figure 34. **Lesion Study**. Evaluation of ViT models when individual blocks, MLP layers or Self-Attention layers are removed from the model after training. We compare four different ViT models (ViT-B/32, ViT-B/16, ViT-L/32, ViT-L/16) and three different pre-training datasets (ILSVRC-2012, ImageNet 21k, JFT-300M). All models are fine-tuned on ILSVRC-2012.