A EXPERIMENTAL SETUP

A.1 DATASET

We used the ImageNet 1k (Deng et al., 2009) dataset for training. ImageNet1K contains 1,000 classes and the number of training and validation images are 1.28 million and 50,000, respectively. We validate the effectiveness of our models in the different datasets proposed in the Brain-Score (Schrimpf et al., 2020a) competition.

A.2 CUSTOM SCHEDULER

The proposed learning rate scheduler is based on Jeddi et al. (2020) and is formulated as $LR = 0.00012 \times e - 0.0004$ for $e = 1$ and $LR = \frac{0.00002}{2^{e-2}}$ for $1 < e <= 6$. As shown in Figure 10, we start with a small learning rate and then it is smoothly increased for one epoch. We empirically found that fine-tuning the transformer for more than 1 epoch resulted in an under-fitting behavior of the adversarial robustness. After this first epoch, the learning rate is reduced very fast so that model performance converges to a steady state, without having too much time to overfit on the training data.

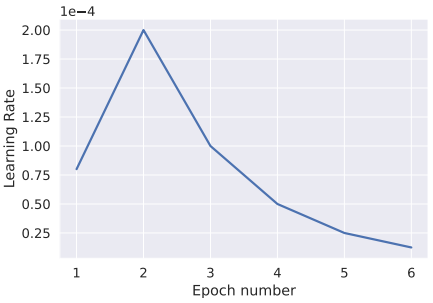

Figure 10: Custom scheduler used for training the Vision Transformer.

A.3 TRAINING SETUP

We used a pretrained CrossViT-18† (Chen et al., 2021) downloaded from the timm library that is adversarially trained via a fast gradient sign method (FGSM) attack and random initialization (Wong et al., 2020). We opted for this strategy, known as "Fast Adversarial Training" as it allows a faster iteration in comparison with other common approaches (*e.g.* adversarial training with the PGD attack). In particular, all experiments used $\epsilon = 2/255$ and step size $\alpha = 1.25\epsilon$ as proposed originally in (Wong et al., 2020). However, in contrast to the previous method, we follow a 5 epoch fine-tuning approach with a custom learning rate scheduler in order to avoid underfitting. We optimize our networks with Adaptive Moment Estimation (Adam *a la* Kingma & Ba (2014)) and employed mixed precision for faster training. All input images were preprocessed with resizing to 256×256 followed by standard random cropping and horizontal mirroring. In the case of our best performing model (#991), we additionally incorporated a random grayscale transformation ($p = 0.25$) and a set of hard rotation transformations of (0°, 90°, 180°, 270°) – implicitly aiding for rotational invariance – due to the characteristics of images appearing in the behavioral benchmark of Rajalingham et al. (2018). All our experiments were run locally on a GPU-Tesla V-100. Each adversarial training of a vision transformer took around 48 hours.

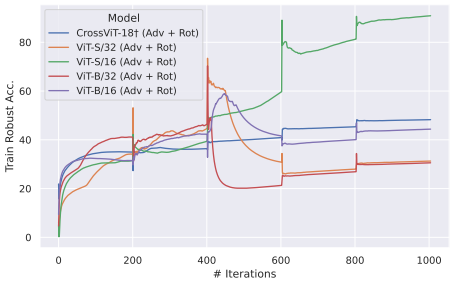

Figure 11: Training robust acc. of each Vision Transformer model (Adv + Rot). We clearly observed that ViT-S/16 has over-fitted during training.

B ADDITIONAL ASSESSMENT OF CROSSVIT-18†-BASED MODELS

B.1 COMMON CORRUPTION BENCHMARKS

We also looked into how adversarial training would affect the performance of the different sets of neural networks to common corruptions that are *not* adversarial. To do this, we ran our models and benchmarked them to the ImageNet-C dataset (Hendrycks & Dietterich, 2019).

One would have expected Brain-Aligned models like our adversarially-trained + rotationally invariant CrossViT to also present strong robustness to common corruptions. To our surprise, this was not the case as seen in Table 4. This is a puzzling result, though there have been several bodies of work suggesting that adversarial robustness and common corruptions robustness are independent phenomena (Laugros et al., 2019), however Kireev et al. (2021) have proved otherwise contingent on the l_∞ radius [6] – but now see Li et al. (2022).

Network	Clean Accuracy (↑)	mce (↓)	Gauss	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contrast	Elastic	Pixel	JPEG
ResNet50-Augmix	77.53	67.1	65.5	65.1	66.4	67.7	81	63.9	65.5	71.6	70.9	66.5	57.8	60.2	76.9	59.5	68.5
CrossViT-18† (Adv + Rot)	73.53	79.5	80.7	81.6	83.2	90.2	78.7	82.4	80	77.6	74	107.9	65	100.4	74.2	57.4	58.7
CrossViT-18† (Adv)	64.60	88.8	85	85.7	86.7	96.7	88	92.1	91.3	85.8	83.6	109.3	82.2	104.9	90	70.3	80.9
CrossViT-18† (Rot)	79.22	73.1	75.4	76.7	75	75.7	85.3	72.3	79.2	68.8	70.9	64.3	54.7	67.6	78.4	75.4	76.4
CrossViT-18†	**83.05**	**51**	**46.1**	**48.8**	**46.4**	**61.2**	**72.6**	**54.4**	**65**	**44.9**	**42.1**	**37.2**	**41.5**	**37**	**67.2**	**46.8**	**54.2**

Table 4: A table showing the comparison of mean corruption errors (mce)'s across CrossViT models contingent on their training regime. A ResNet50-Augmix is shown as a reference of a particularly strong model to common corruptions. Here lower scores are indicative of better robustness to the different distortion types of Hendrycks & Dietterich (2019).

B.2 IMAGENET-R

We also looked into how adversarial training would affect the performance of generalization to various abstract visual renditions. To do this, we ran our models and benchmarked them on the ImageNet-Rendition (ImageNet-R) dataset (Hendrycks et al., 2021a).

We observe that the accuracy on ImageNet-R decreases when the CrossViT is adversarially trained. However, when we combine the rotation invariance and adversarial training regimes, the accuracy on ImageNet-R becomes competitive with its pretrained version. In addition, we also appreciate that this combination does not affect the IID/OOD Gap with respect to the pretrained CrossViT.

Network	ImageNet-200 (↑)	ImageNet-R (↑)	Gap (↓)
CrossViT-18† (Adv + Rot)	90.75	41.14	**49.61**
CrossViT-18† (Adv)	85.52	35.73	49.79
CrossViT-18† (Rot)	93.89	37.35	56.54
CrossViT-18†	95.64	**45.7**	49.94

Table 5: A table showing the comparison of the accuracy on Imagenet-R dataset across CrossViT models contingent in their training regime.

B.3 CENTER KERNEL ALIGNMENT TO UNDERSTAND CROSSVIT REPRESENTATIONS

We also calculated the center kernel alignment (Kornblith et al., 2019) scores at each brain-region layer and on the Behavior and Inversion layers using a linear kernel. Besides, CKA scores were generated using the *'ImageNette'* validation dataset (Howard, 2019) which is a subset of 10 easily classified classes from ImageNet. The objective of this experiment is to understand how correlated are the variance of internal representations across the different versions of the optimized CrossViT-18†.

[6] Also see Li et al. (2022) that shows that generally robust models (robust to adversarial + commmon corruptions) have a preference for low-spatial frequency statistics.

We can see in Figure 12 that intermediate brain-region layers (IT, Behavior) tend to have similar representations across the 3 variants of CrossViT-18† (Rot. + Adv., Rot. and Adv.) based on the CKA score. In addition, we also appreciate that our best model (Crossvit-18† + Rot. + Adv.) is more correlated with their individual versions (Rot. and Adv.) than with its pretrained version.

It is also remarkable that at the penultimate layer of the largest branch (inversion layer), our best CrossViT possesses a very weak similarity with its pretrained form. This suggests that adversarial training and rotation invariance, either jointly or independently, strongly changes the representation of the final layers with respect to its pretrained version (CrossViT-18†).

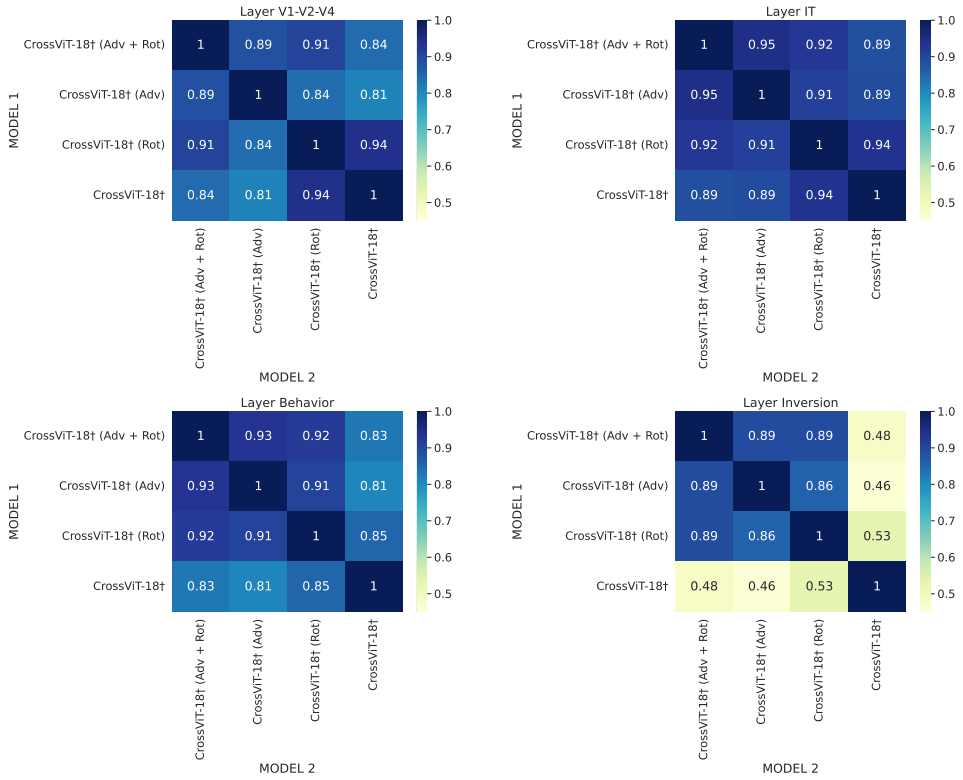

Figure 12: Similarity of representations at V1, V2, V4, Behavior and Inversion layers across the four versions of CrossViT-18† (pretrained, Adv., Adv. + Rot., Rot.). A score of 1.0 indicates highest representational similarity, while a score of 0.0 indicated lowest.

C Adversarial Attacks Experiments

C.1 Targeted Adversarial Attacks

In this experiment, we maximize the probability of a specific class ("Goldfish" targeted attack) for the 4 flavors of the CrossViT-18†. We observed that as the average "Brain-Score" increases, the models tend to resemble more accurately the samples of the target class (Figure 3). In addition, we also performed targeted attacks for different classes on the ImageNet dataset as can be seen in Figure 4. Parameters used for these experiments can be found in Table 6

C.1.1 Adversarial Robustness to PGD attacks

Results of PGD adversarial attacks on different versions of CrossViT-18† can be found in Table 7. All experiments used $\epsilon \in \{1/255, 2/255, 4/255, 6/255, 8/255, 10/255\}$ and step-size = $\frac{2.5}{\#PGD_{iterations}}$ as in the robustness Python library (Engstrom et al., 2019a).

Dataset	ϵ	Steps	Step size
ImageNet	300	500	1

Table 6: Parameters used for the targeted attacks

	$\epsilon - test(\uparrow)$					
Model	1/255	2/255	4/255	6/255	8/255	10/255
CrossViT-18† (Adv + Rot)	65.1/64.84/64.83	55.99/54.27/54.23	39.52/32/31.69	27.81/15.76/15.28	19.33/6.67/6.32	14.77/2.72/2.43
CrossViT-18† (Adv)	62.27/5.3/4.15	59.9/4.2/2.14	55.36/7.18/0.996	51.02/14.97/0.66	47.16/12.84/0.6	43.76/6.37/0.6
CrossViT-18† (Rot)	48/1.75/1.5	4.87/0/0	2.17/0/0	1.89/0/0	1.87/0/0	2.13/0/0
CrossViT-18†	48.31/14.87/6.64	44.01/5.56/1.1	41.58/1.47/0.09	40.96/0.59/0.02	40.79/0.35/0.01	40.9/0.13/0

Table 7: PGD adversarial attacks on different flavors of CrossViT-18†. Results represent adversarial accuracy at 1/10/20 PGD-iterations

D BRAIN-SCORE

D.1 METRICS

Brain-Score is a composite of multiple neural and behavioral benchmarks that score most of the artificial neural networks on how similar they are to the primate's brain mechanisms for core object recognitionSchrimpf et al. (2020a).

In the same direction, the Brain-Score competition was held for 4 months from December 21 to March 22. The objective was to evaluate models that engage with the whole ventral visual stream. These models were evaluated in 33 neuronal and behavioral benchmarks related to activity in macaque visual cortical areas V1, V2, V4, and IT and human psychophysical performance in a set of object classification tasks. The metrics used in the evaluation are the followings:

Neural predictivity: Measures how well the responses to given images in a model area predict the responses of a neuronal population of the corresponding area in the macaque brain. First, the model responses are mapped to the neuronal recordings using a linear transformation (PLS regression with 25 components) on a training set of images. Then the model's predictivity is determined for held-out images by computing the Pearson correlation coefficient between the model's predictions and the neuronal responses.

Single-neuron property distribution similarity: Measures whether single neurons in a model area are functionally similar to single-neurons in the corresponding monkey brain area. This is done by comparing the distribution of single-neuron response properties between the model area and the brain area using a similarity score (using the KS distance).

Behavioral consistency: Measures the behavioral similarity between the model and humans in core object recognition tasks. This metric does not measure the overall accuracy of the model but whether it can predict the patterns of successes and failures of humans in a set of object recognition tasks. Model's and humans' behavioral accuracies are first transformed to a d'statistic and then compared using the Pearson correlation coefficient.

D.1.1 SELECTING BEST-BRAINSCORE LAYERS

Best performing layers on each vision transformer were selected by a brute-force approach. We evaluate each layer of the vision transformer models on each brain region and behavior dataset and select the layer that got the best score on the public benchmarks (in order to avoid overfitting) proportioned by Brain-Score organization. After this step, the "Adv + Rot" & pretrained versions of each transformer are submitted to the competition fixing best performing layers (See Table 8). We achieved our highest score at the time of our 4th submission, which was the lowest number of submissions in the competition (the winner of the competition performed nearly 60 submissions). All our results reflect the private scores obtained by each vision transformer model.

Additionally to the experiments on CrossViT-18†, we also evaluate the brain-scores on vanilla Vision transformers that can be seen in Table 9.

Model	V1	V2	V4	IT	Behavior
CrossViT-18†	blocks.1.blocks.1.0.norm1	blocks.1.blocks.1.0.norm1	blocks.1.blocks.1.0.norm1	blocks.1.blocks.1.4.norm2	blocks.2.revert_projs.1.2
ViT-S/16	blocks.1.mlp.act	blocks.3.attn.proj	blocks.3.norm2	blocks.9.norm1	pre_logits
ViT-S/16	blocks.1.mlp.act	blocks.3.attn.proj	blocks.3.norm2	blocks.9.norm1	pre_logits
ViT-S/32	blocks.1.mlp.act	blocks.10.norm1	blocks.2.mlp.act	blocks.10.norm1	pre_logits
ViT-B/16	blocks.1.mlp.act	blocks.6.norm2	blocks.2.mlp.act	blocks.8.norm1	pre_logits
ViT-B/32	blocks.1.mlp.act	blocks.6.norm2	blocks.2.mlp.act	blocks.11.norm1	pre_logits

Table 8: Layers selected for each brain region on each vision transformer.

	ImageNet(↑)	Brain-Score(↑)					
Description	Validation Acc. (%)	Avg	V1	V2	V4	IT	Behavior
ViT-S/16	81.40	0.445	0.527	0.295	0.454	0.449	0.498
ViT-S/32	75.99	0.415	0.531	0.271	0.422	0.423	0.426
ViT-B/16	84.53	0.451	0.522	0.317	0.398	0.487	0.529
ViT-B/32	80.72	0.440	0.553	0.311	0.413	0.418	0.505
ViT-S/16 (Adv + Rot)	50.44	0.443	0.506	0.332	0.470	0.496	0.409
ViT-S/32 (Adv + Rot)	55.20	0.457	0.512	0.347	0.433	0.485	0.508
ViT-B/16 (Adv + Rot)	67.25	0.486	0.536	0.332	0.470	0.496	0.598
ViT-B/32 (Adv + Rot)	53.01	0.457	0.524	0.357	0.417	0.472	0.515

Table 9: ImageNet accuracy, Brain-Scores of each brain area & Behavior benchmark evaluated on vanilla vision transformers. The spearman rank correlation between the validation accuracy and the average Brain-Score is -0.28 suggesting an *inverse* correlation between clean ImageNet accuracy and Brain-Score (Schrimpf et al., 2020a).

E Image Synthesis Experiments

E.1 Standard & Robust Stimuli

We used publicly available transformer models from timm library which were trained adversarially ($\epsilon = 2/255$ and step size $\alpha = 1.25$) as in (Wong et al., 2020) coupled with a set of hard rotation462 transformations of (0°, 90°, 180°, 270°) as proposed in this paper. In order to synthesize the standard and robust images, we used the penultimate layer (norm layer) in all of our vision transformer models except in the case of the CrossViT-18† versions in which we used the penultimate layer of the largest branch for all variations. Parameters used in these experiments can be seen in Table 10.

Constraint	ϵ	Step-size	Iterations
l_2	1000	1	10000

Table 10: Parameters used for standard & robust stimuli by feature inversion

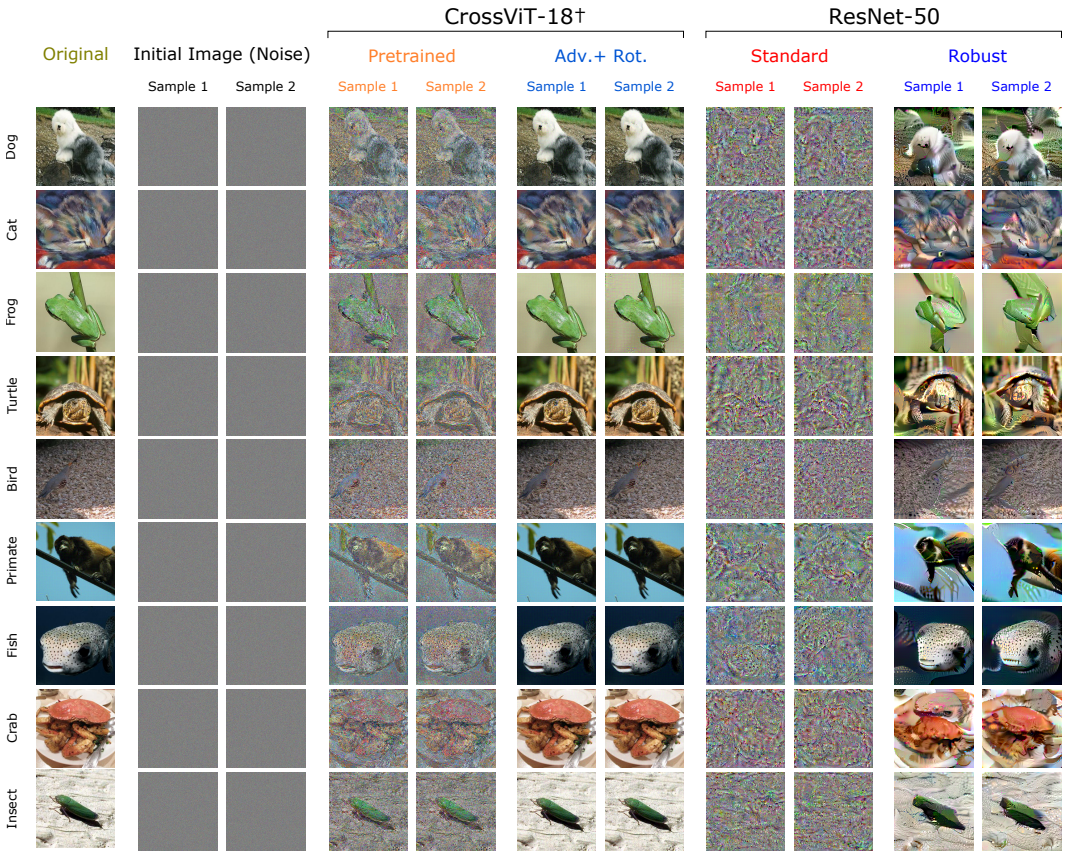

Figure 13: An extended version of Figure 7

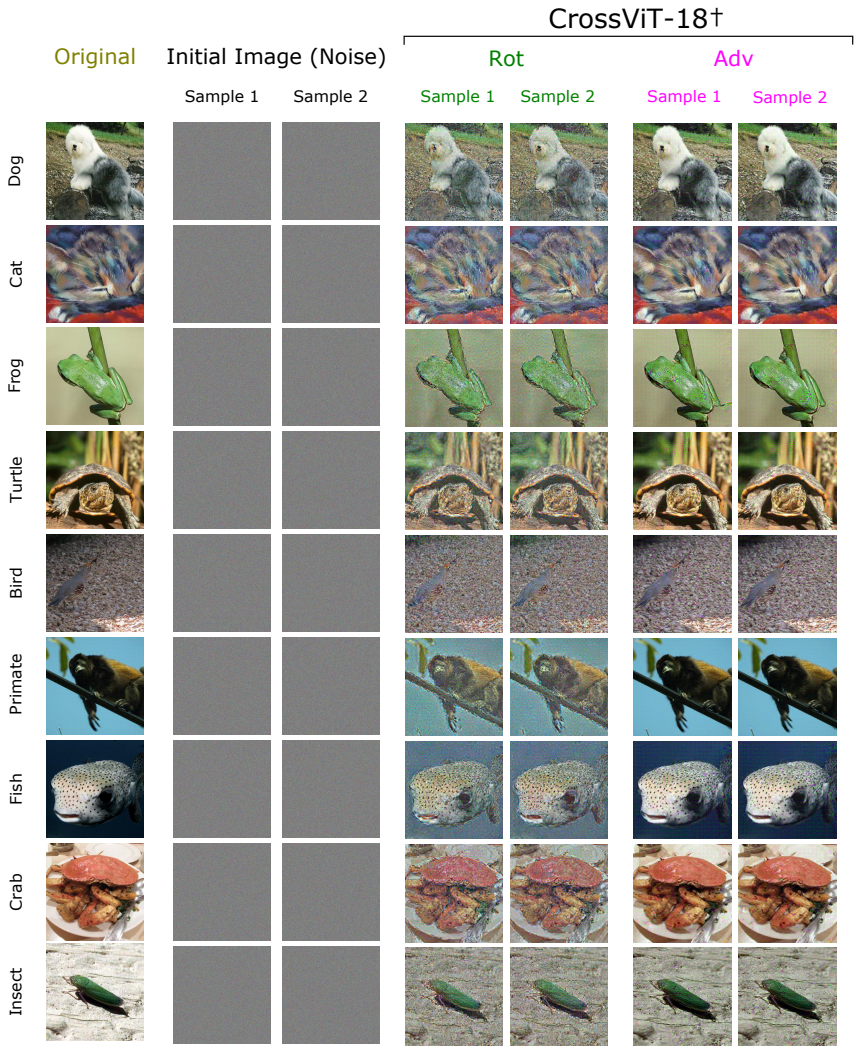

Figure 14: Feature Inversion for CrossViT-18† (Adv) & CrossViT-18† (Rot).

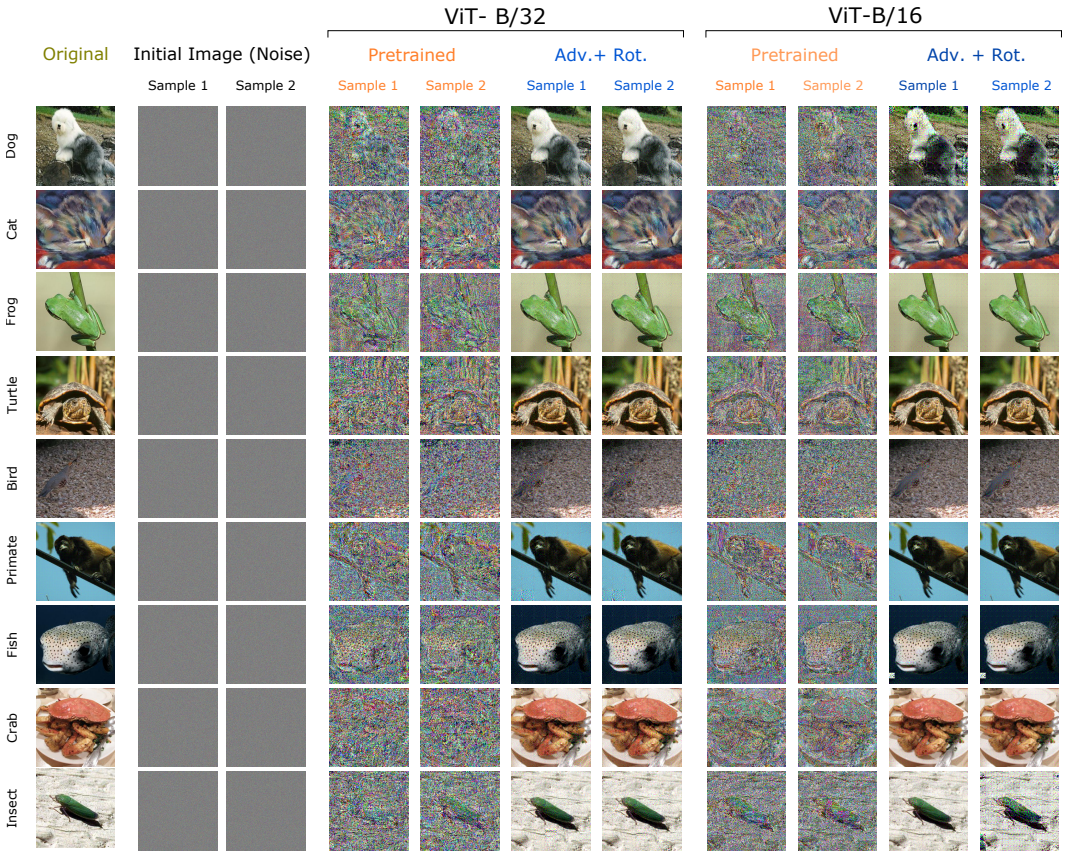

Figure 15: Feature Inversion for Vanilla Vision Transformers ViT-B/32 & ViT-B/16.

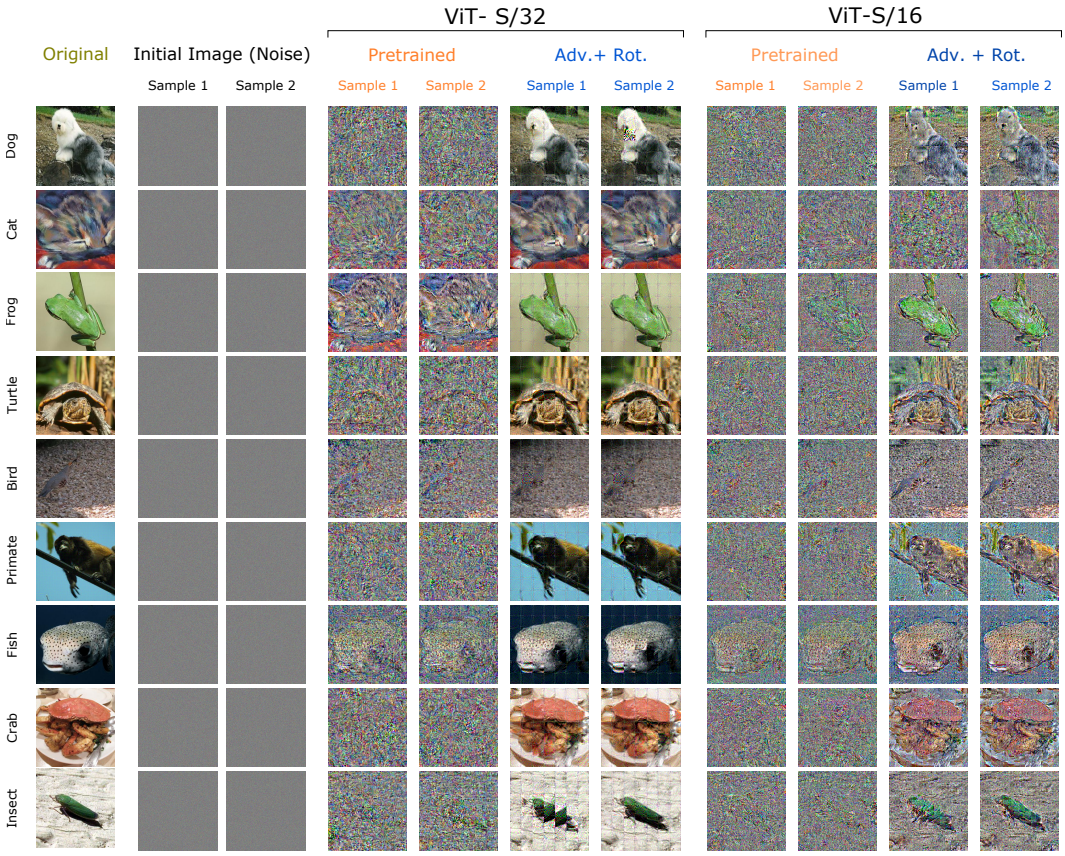

Figure 16: Feature Inversion for Vanilla Vision Transformers ViT-S/32 & ViT-S/16.