Are Vision Transformers Robust to Patch-wise Perturbations?

Supplementary Material

A Training Setting Affect Model Robustness

We train ResNet18 on CIFAR10 in the standard setting [1]. To study the impact of training settings on model robustness, we train models with different input sizes (i.e., 32, 48, 64), with or without Weight Standardization and Group Normalization to regularize the training process. The foolong rate of single patch attack is reported. Especially, with our experiments, we find that Weight Standardization and Group Normalization can have a significant impact on model robustness (See Tab. 1). The two techniques are applied in BiT [3] to improve its performance. However, they are not applied to standard ViT and DeiT training settings. Hence, the robustness difference between ViT and BiT cannot be attributed to the difference between model architectures.

Note that a comprehensive study of the relationship between all factors of training and model adversarial robustness is out of the scope of this paper. We aim to point out that these factors can have an impact on model robustness to different extents. The robustness difference cannot be blindly attributed to the difference of model architectures. We need to build new fair base models to study the robustness of ResNet and ViT.

Table 1: Study of the training factors on the relation to model robustness: While the input size has minor impact on model robustness in the first tabular, Weight Standardization (WS) and Group Normalization (GN) can change model robustness significantly in the second tabular.

Model	Input Size		
ResNet18	32	48	64
Clean Accuracy	93.4	93.8	93.7
FR of Patch Attack	35.9	42.2	39.2

Model	Training Techniques			
ResNet18	No	WS	GN	WS + GN
Clean Accu	93.4	93.6	92.0	93.8
Patch Attack FR	35.9	51.3	52.6	71.1

Table 2: Fair base models. DeiT and counter-part ResNet are trained with the exact same setting. Two models of each pair achieve similar clean accuracy with comparable model sizes.

Model	Model Size	Clean Accuracy
ResNet50	25M	78.79
DeiT-small	22M	79.85
ResNet18	12M	69.39
DeiT-tiny	5M	72.18

B Natural Patch Corruption with Different Levels and Types

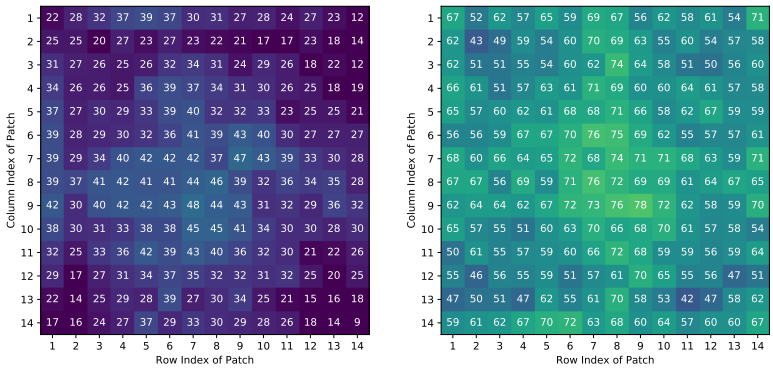
Models can show different robustness when the inputs are corrupted with different natural noise types. To better evaluate the model robustness to natural corruption, the work [2] summarizes 15 common natural corruption types. The averaged score is used as an indicator of model robustness. In this appendix section, we show more details of model robustness to different noise types. As show in Fig. 5 and 6, The FR on DeiT is lower than on ResNet. We conclude that DeiT is more robust than ResNet to natural patch corruption.

Furthermore, we also investigate the model robustness in terms of different noise levels. As shown in Fig. 7 and 8. The different colors stand for different noise level. S1-S5 corresponds to the natural corruption severity from 1 to 5. In each noise type, the left bar corresponds to ResNet variants and the right one to DeiT variants. We can observe that DeiT show lower FR in each severity level. Namely, the conclusion drawn above also holds across different noise levels.

C Gradient Visualization of Adversarial Images under Patch Attack

We first get the absolute value of gradient received by input and sum them across the channel dimension. The final values are mapped into gray image scale. We also mark the adversarial patch with a blue bounding box in the visualized gradient maps.

The adversarial patch noises with different patch sizes (i.e., P=16 and P=32) are shown on DeiT and ResNet in Fig. 9, 10, 11, and 12. In each row of these figures, we fist show the clean image and visualize the gradients of inputs as a mask on the image. Then, we show the images with patch noises on different patch positions, and the gradient masks are also shown following the corresponding adversarial images.



(a) Adversarial Patch Attack FRs on ResNet50

(b) Adversarial Patch Attack FRs on DeiT-small

Fig. 1: Patch Attack FR (in %) in each patch position is visualized on ResNet50 and DeiT-small.

D More Figures of Attention on Different Patch Sizes and Positions

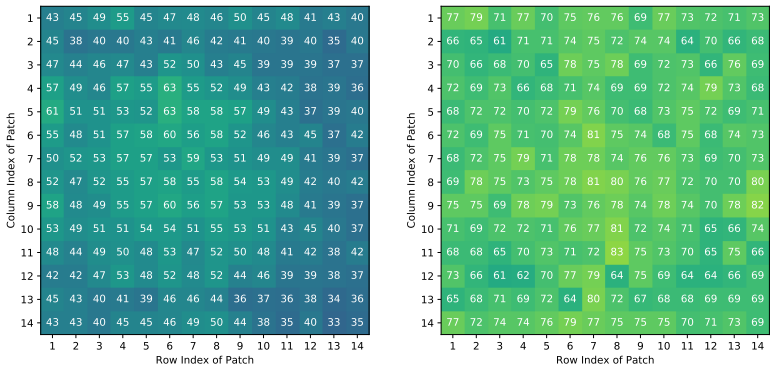
In this appendix section, we show more Attention Rollout on DeiT and Feature Map Masks on ResNet. The adversarial patch noises with different patch sizes are shown (i.e., P=16 and P=32) in Fig. 13, 14, 15, and 16. In each row of these figures, we fist show the clean image and visualize the attention as a mask on the image. Then, we show the images with patch noises on different patch positions, and the attention masks are also shown following the correspond adversarial images.

E Attention under Natural Patch Corruption and Adversarial Patch Attack

The rollout attention on DeiT and Feature Map mask on ResNet on naturally corrupted images are shown in Fig. 17, 18, 19, and 20. We can observe that ResNet treats tha corrupted patches as normal ones. On DeiT, the attention is slightly distract by naturally corrupted patches when they are in the background. However, the main attention is still on the main object of input.

F Fooling Rates of Each Patch on ResNet50 and DeiT-small

The FRs in different patch positions of DeiT are similar, while the ones in ResNet are center-clustered. A similar pattern can also be found on DeiT-small and ResNet50 in Fig. 1.



(a) FRs of ResNet18 on Corner-biased Data

(b) FRs of DeiT-tiny on Center-biased Data

Fig. 2: Patch Attack FR (in %) in each patch position is visualized on ResNet18 and DeiT-tiny on biased data.

G Fooling Rates of Each Patch on ResNet and DeiT on Corner-biased Data

In the coner-biased image set, the FR on ResNet is still center-clustered, as shown in Fig. 2a.

H Fooling Rates of Each Patch on ResNet and DeiT on Center-biased Data

In the center-biased image set, the FR on DeiT is still similar on different patch postions, as shown in Fig. 2b.

I Transferability of Adversarial Patches across Images, Models, and Patch Positions

As shown in Tab. 3, the adversarial patch noise created on a given image hardly transfer to other images. When large patch size is applied, the patch noises on DeiT transfer slightly better than the ones on ResNet.

The transferbility of adversrial noise between Vision Transformer and ResNet has already explored in a few works. They show that the transferability between them is remarkablely low. As shown in Tab. 4, the adversarial patch noise created on a given image does not transfer to other models.

When they are transfered to another patch, the adversarial patch noises are still highly effective. However, the transferability of patch noise can be low, when the patch is not aligned with input patches. The claim on the patch noise with size of 112 is also true, as shown in Tab. 5.

Table 3: Transferability of adversarial patch across images

Models	ResNet50	DeiT-small	ResNet18	DeiT-tiny
across images (Patch Size=16)	3.5	2.1	3.4	6.4
across images (Patch Size=112)	8.1	13.4	10.6	21.5

Table 4: Transferability of adversarial patch across models

	Patch Size=16			
Models	ResNet50	DeiT-small	ResNet18	DeiT-tiny
ResNet50	-	0.3	0.16	2.2
DeiT-small	0.04	-	0.09	1.79
ResNet18	0.09	0.22	-	1.9
DeiT-tiny	0.04	0.13	0.06	-
	Patch Size=112			
Models	ResNet50	DeiT-small	ResNet18	DeiT-tiny
ResNet50	-	5.25	8	11.75
DeiT-small	5.5	-	9.25	12.25
ResNet18	5.75	5	-	12
DeiT-tiny	5.5	5	9.25	-

J More Settings and Visualization of Adversarial Examples with Imperceptible Noise

In the standard adversarial attack, the artificial noise can be placed anywhere in the image. In our adversarial patch attack, we conduct experiments with different patch sizes, which are multiple times the size of a single patch. The robust accuracy under different attack patch sizes is reported in Tab. . We can observe that DeiT is more vulnerable than ResNet under imperceptible attacks.

The clean images and the adversarial images created on different models are shown in Fig. 3. The adversarial perturbations created with imperceptible patch attack are imperceptible for human vision.

K Visualization of Adversarial Patch Noise

Besides reporting the FRs, we also visualize the adversarial patch perturbation created on ResNet and DeiT. The adversarial patch perturbation are shown in Fig. 4a and 4c. We are not able to recognize any object in the target class.

Following Karmon et al. 's LaVAN, we enhance the attack algorithm where we place the patch noise on different patch positions in different images in each

Table 5: Transferability of adversarial patch across patch positions

Model	ResNet50	DeiT-small	ResNet18	DeiT-tiny
across positions (0, 4)	6.25	5.25	11.25	12.75
across positions (0, 16)	5.75	**34.5**	11.5	**54**
across positions (0, 64)	6	22	9.5	30.75
across positions (4, 0)	6.5	5.75	9.75	12.5
across positions (16, 0)	7.25	**35**	10.25	**54**
across positions (64, 0)	5.5	18.25	9.25	31
across positions (4, 4)	6	4.75	8.5	13.5
across positions (16, 16)	4.5	**18.5**	9	**33**
across positions (64, 64)	6	9.75	8.25	17.5

Table 6: Adversarial Patch Attack with Imperceptible Perturbation . FRs are reported in percentage.

Model	PatchSize=16	PatchSize=32	PatchSize=112	PatchSize=224
ResNet50	2.9	20.9	98.3	100
DeiT-small	4.1	38.7	100	100
ResNet18	3.1	26.0	99.1	100
DeiT-tiny	11.2	46.8	100	100

attack iteration. From the visualization of the created noise in Fig. 4b and 4d, we can recognize the object/object parts of the target class on both ResNet and DeiT. In this section, we conclude that the recognizability of adversarial patch noise is dependent more on attack algorithms than the model architectures.

(a) Clean Images

(b) Adversarial Examples on ResNet18

(c) Adversarial Examples on DeiT-tiny

(d) Adversarial Examples on ResNet50

(e) Adversarial Examples on DeiT-small

Fig. 3: Visualization of Adversarial Examples with Imperceptible Patch Noise: The adversarial images with patch noise of size 112 in the left-upper corner of the image are visualized. Please Zoom in to find the subtle difference.

(a) Patch Noise on ResNet50 under the 1st Setting

(b) Patch Noise on ResNet50 under the 2nd Setting

(c) Patch Noise on DeiT-small under the 1st Setting

(d) Patch Noise on DeiT-small under the 2nd Setting

Fig. 4: Visualization of Adversarial Patch Perturbations under different Settings: In the 1st setting, the patch noise is created to fool a single classification in a given patch position. The goal in the 2nd setting to mislead the classifications of a set of images at all patch positions.

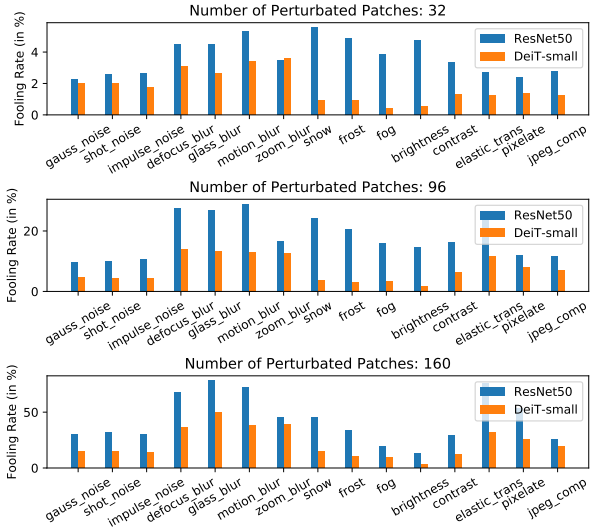


Fig. 5: Comparison of ResNet50 and Deit-small on Naturally Corrupted Patches

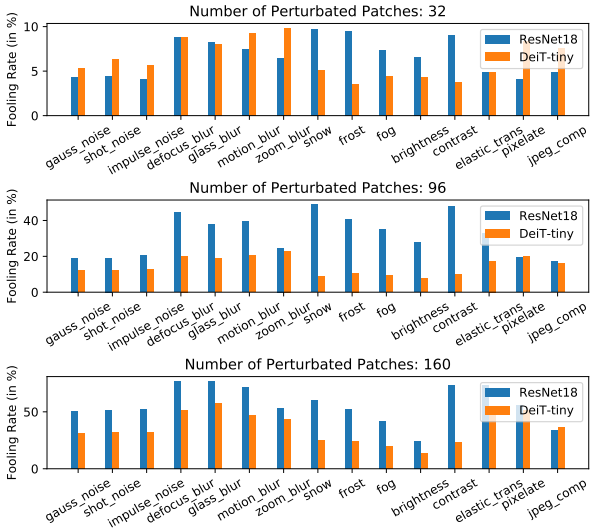


Fig. 6: Comparison of ResNet18 and Deit-tiny on Naturally Corrupted Patches

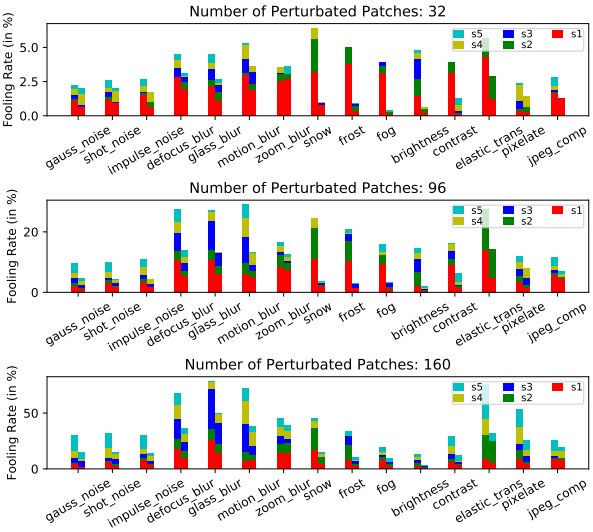


Fig. 7: Comparison of ResNet50 and Deit-small on Patches Corrupted with Different Levels

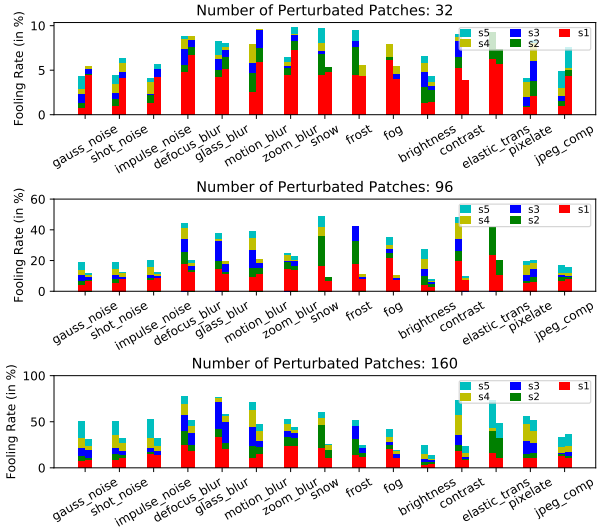


Fig. 8: Comparison of ResNet18 and Deit-tiny on Patches Corrupted with Different Levels

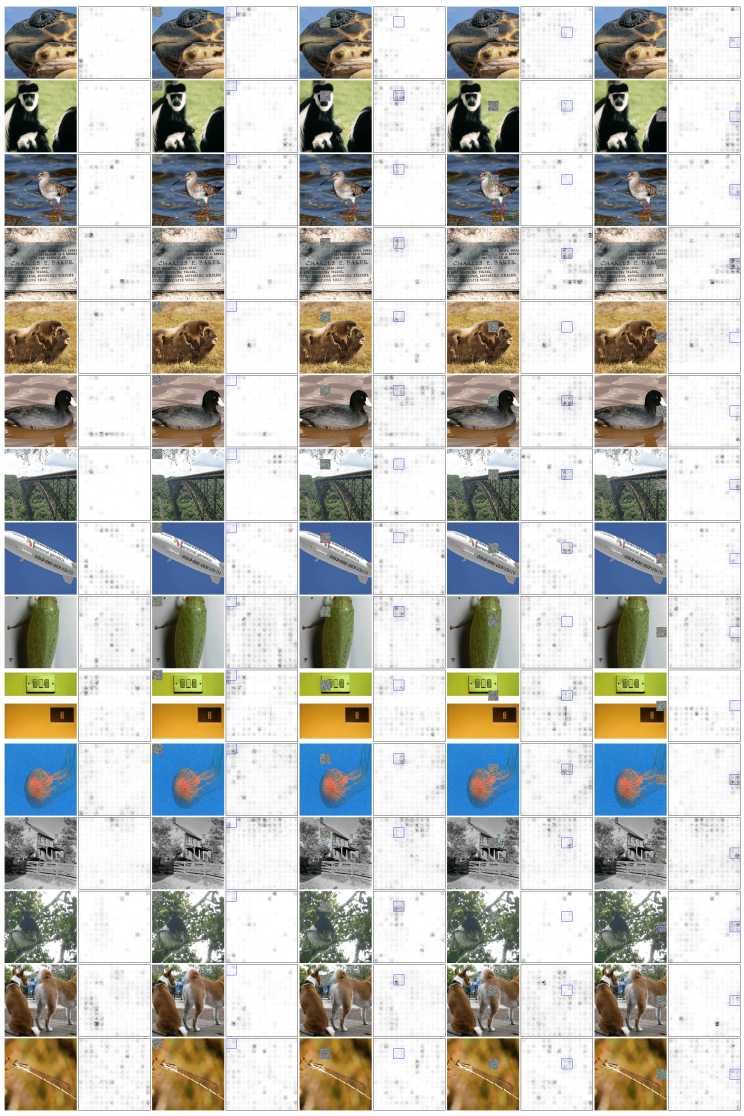


Fig. 9: Gradient Visualization on DeiT-small with Attack Patch size of 32

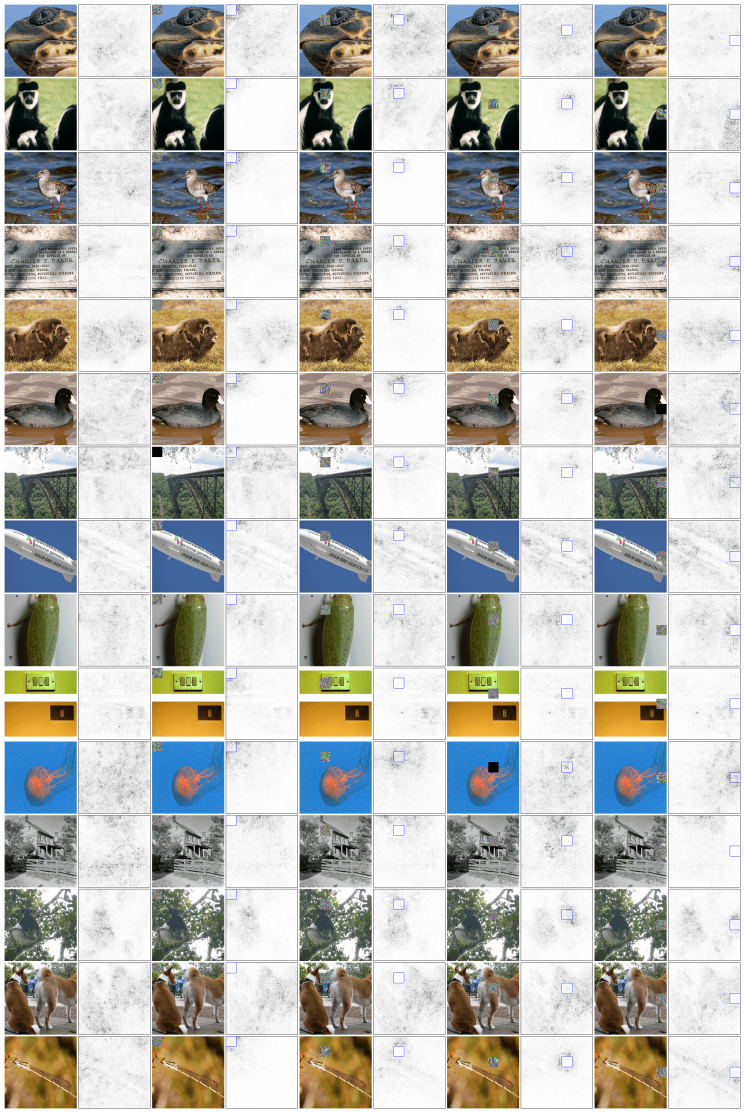


Fig. 10: Gradient Visualization on ResNet50 with Attack Patch size of 32

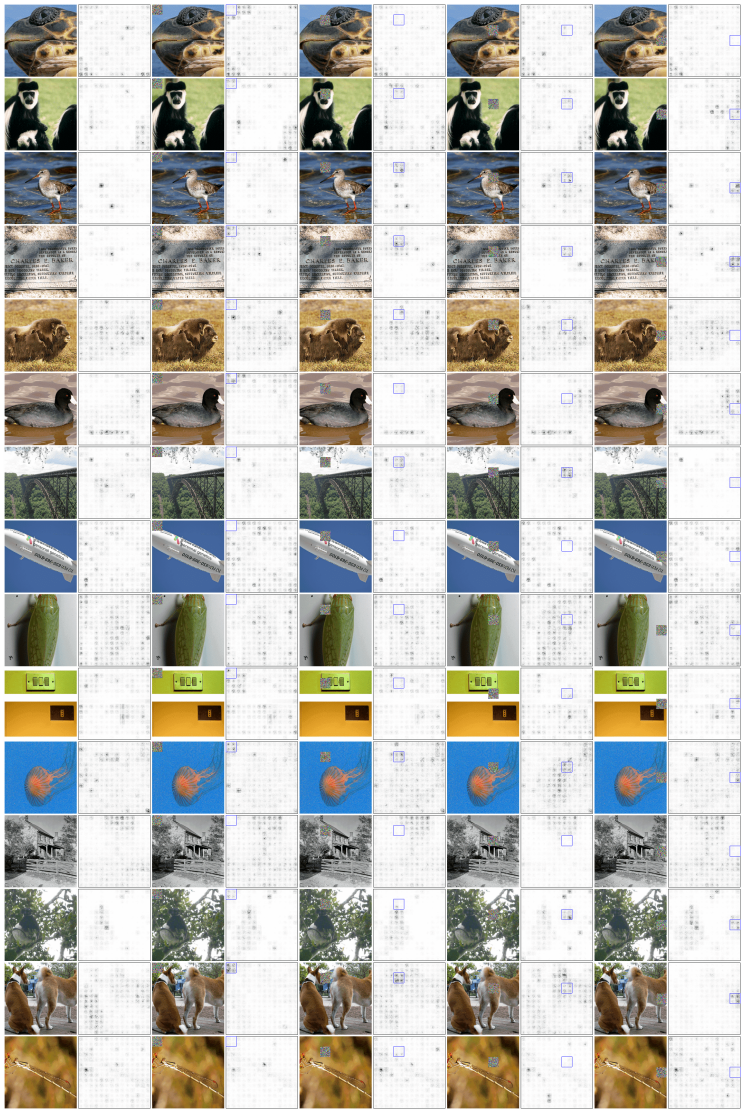


Fig. 11: Gradient Visualization on DeiT-tiny with Attack Patch size of 32

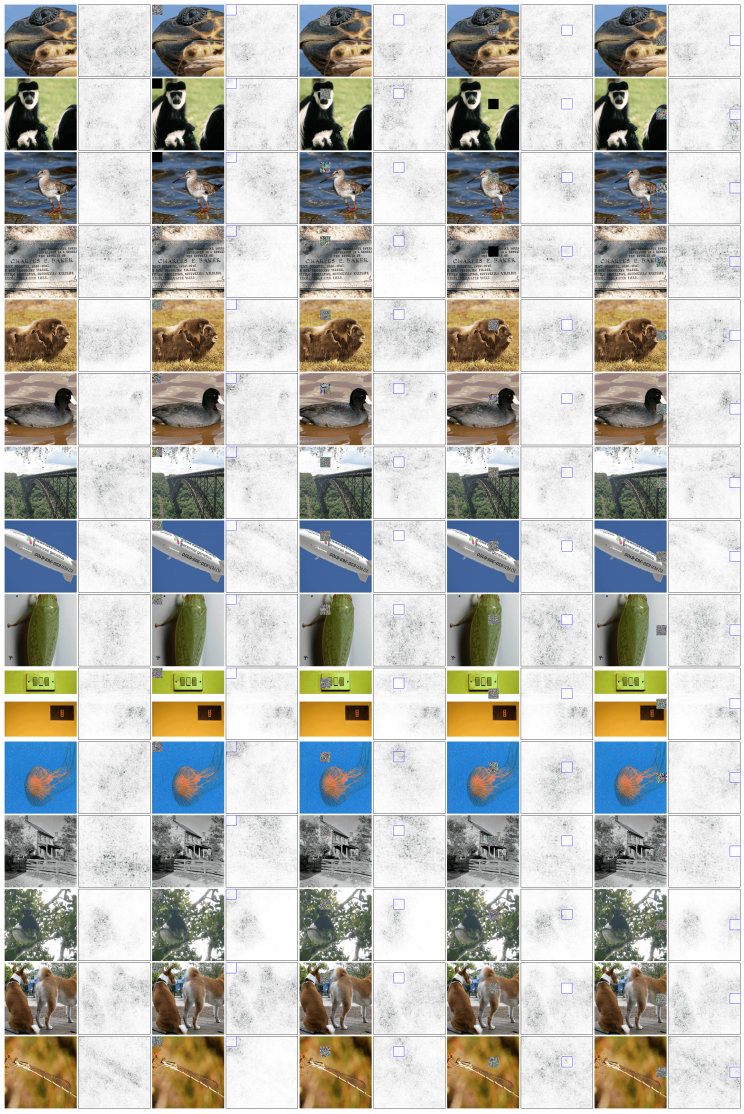


Fig. 12: Gradient Visualization on ResNet18 with Attack Patch size of 32

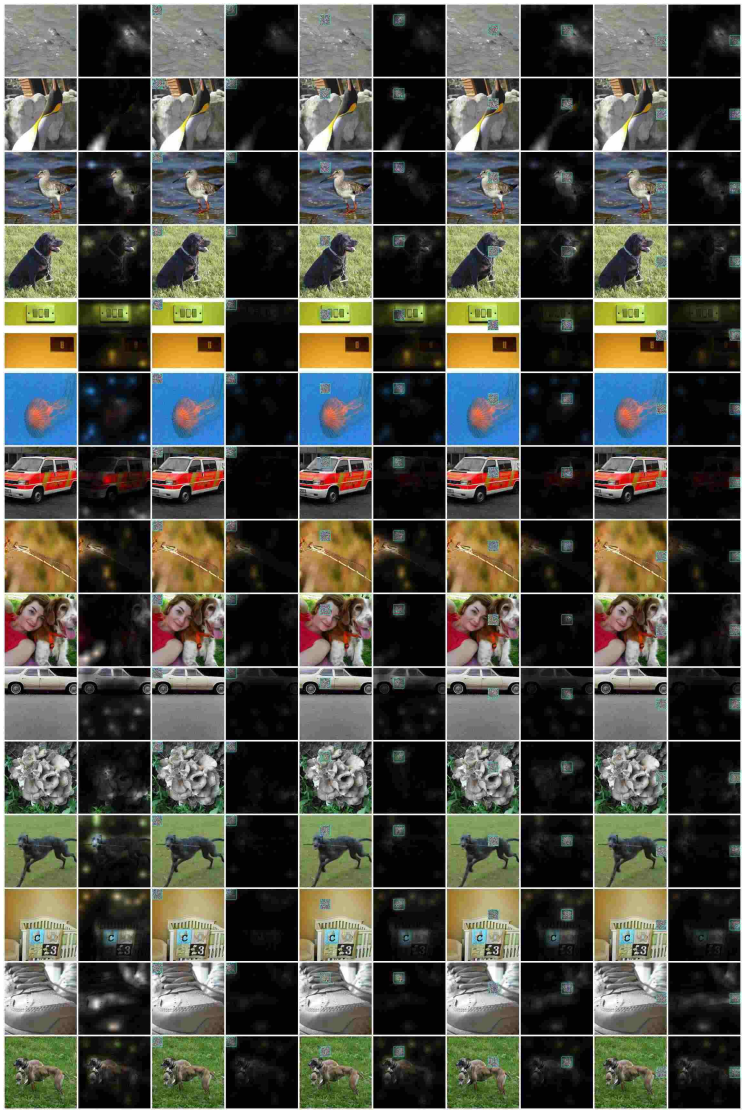


Fig. 13: Rollout Attention on DeiT-small with Attack Patch size of 32 on Adversarial Images

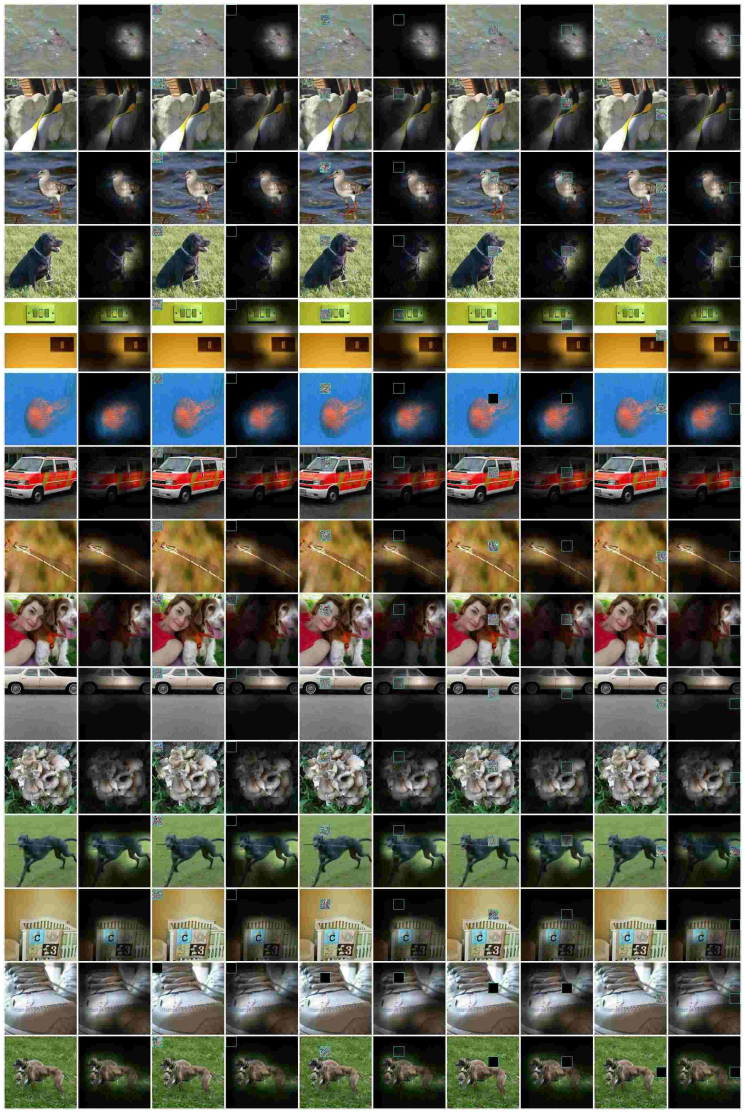


Fig. 14: Averaged Feature Maps of ResNet50 as Attention with Attack Patch size of 32 on Adversarial Images

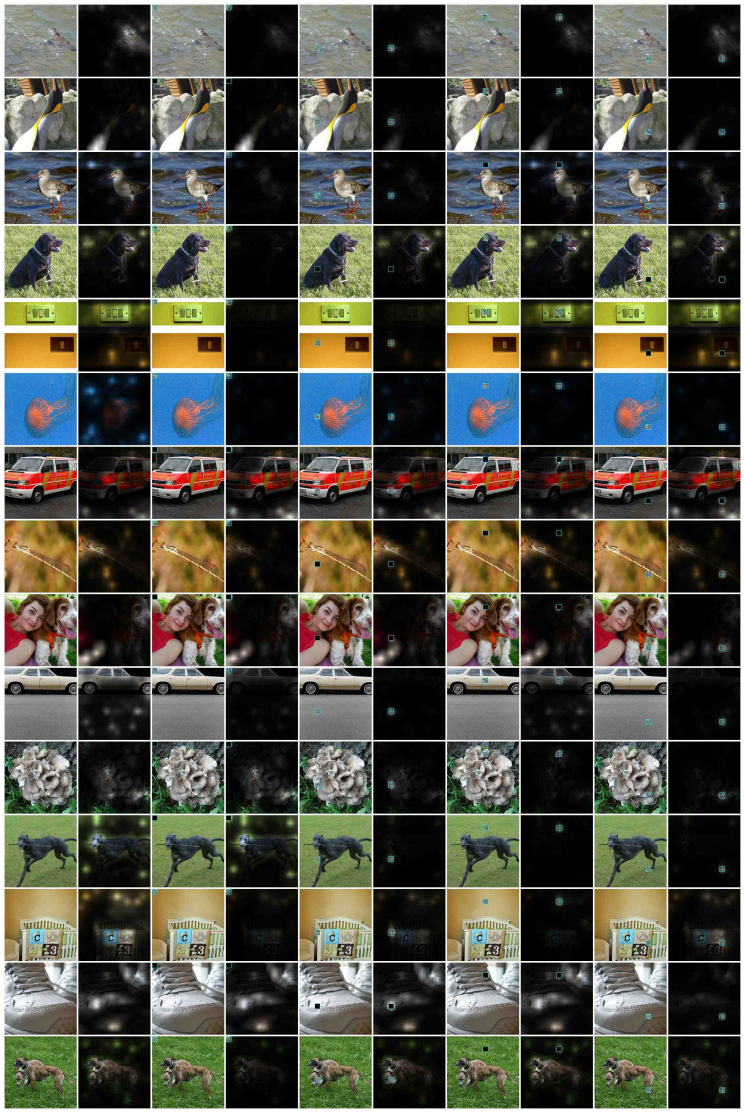


Fig. 15: Rollout Attention on DeiT-small with Attack Patch size of 16 on Adversarial Images

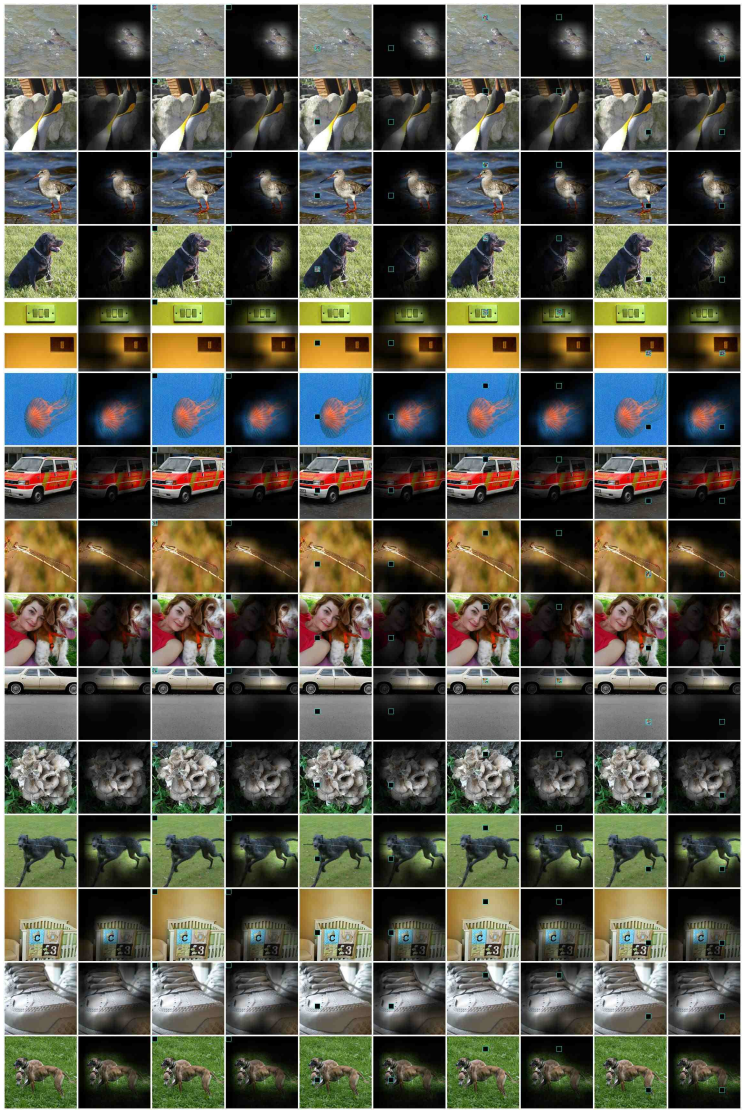


Fig. 16: Averaged Feature Maps of ResNet50 as Attention with Attack Patch size of 16 on Adversarial Images

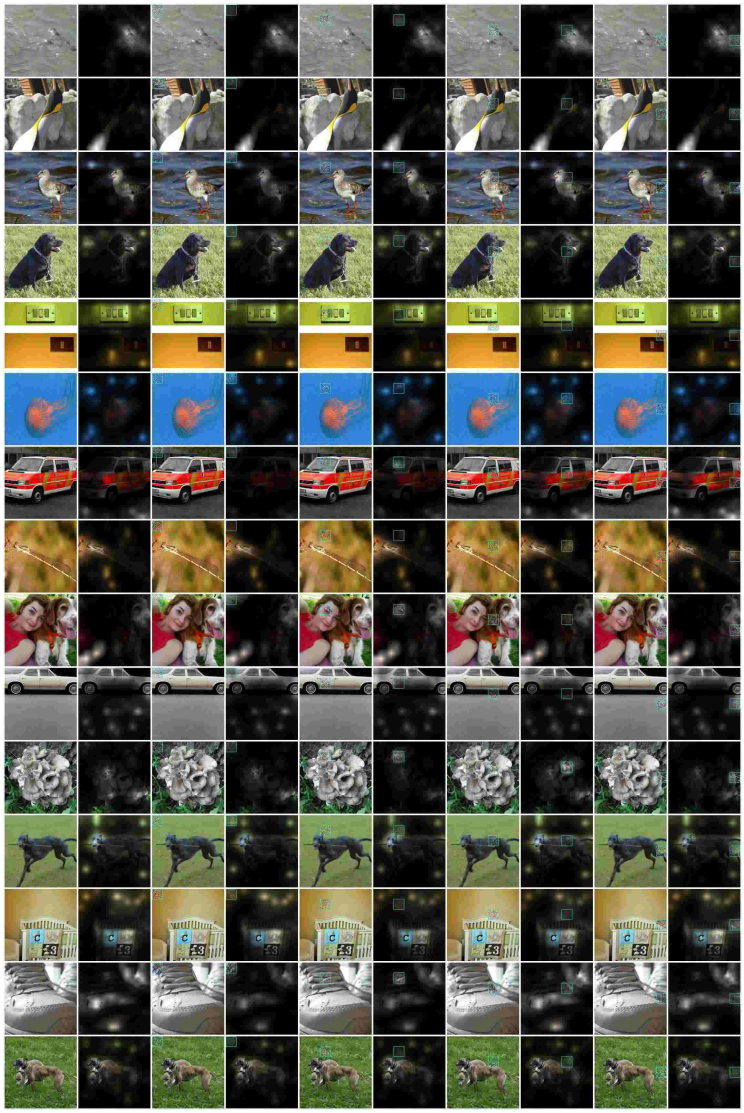


Fig. 17: Rollout Attention on DeiT-small with Attack Patch size of 32 on Corrupted Images

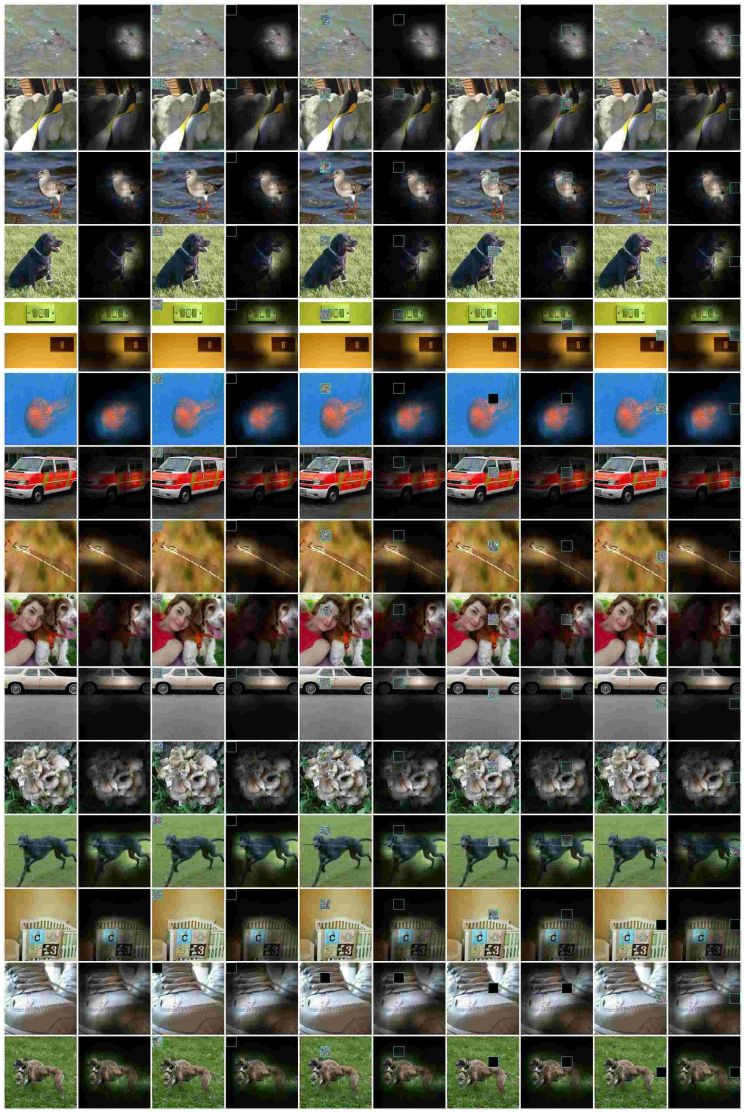


Fig. 18: Averaged Feature Maps of ResNet50 as Attention with Attack Patch size of 32 on Corrupted Images

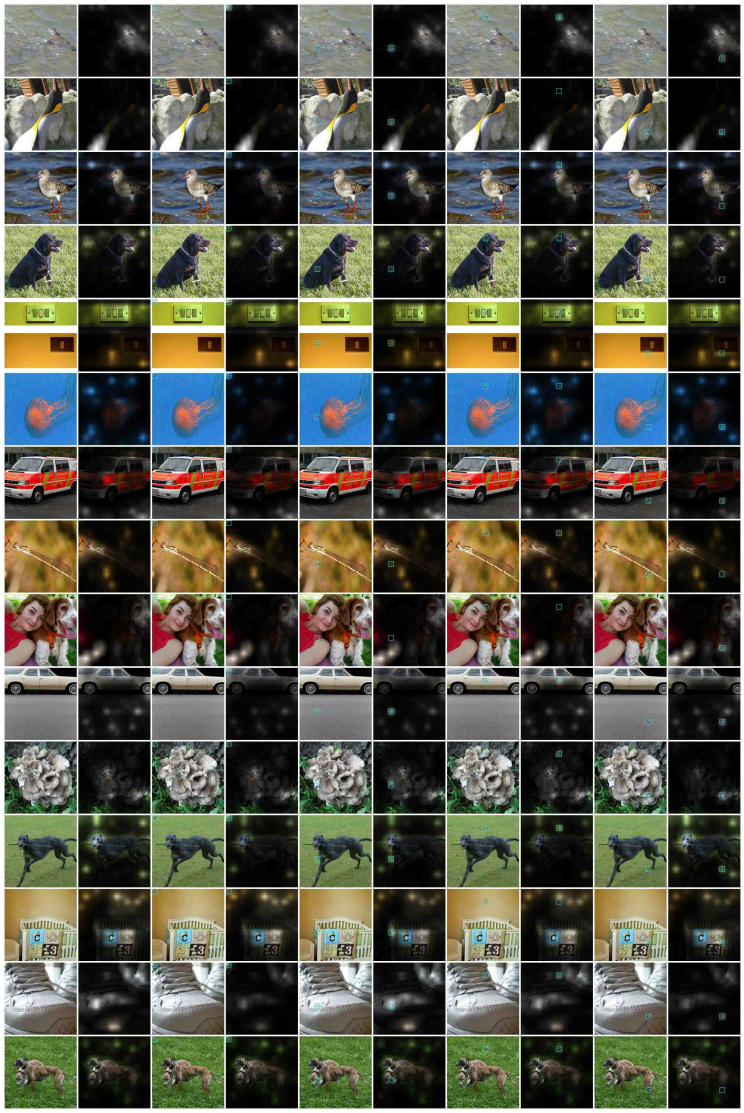


Fig. 19: Rollout Attention on DeiT-small with Attack Patch size of 16 on Corrupted Images

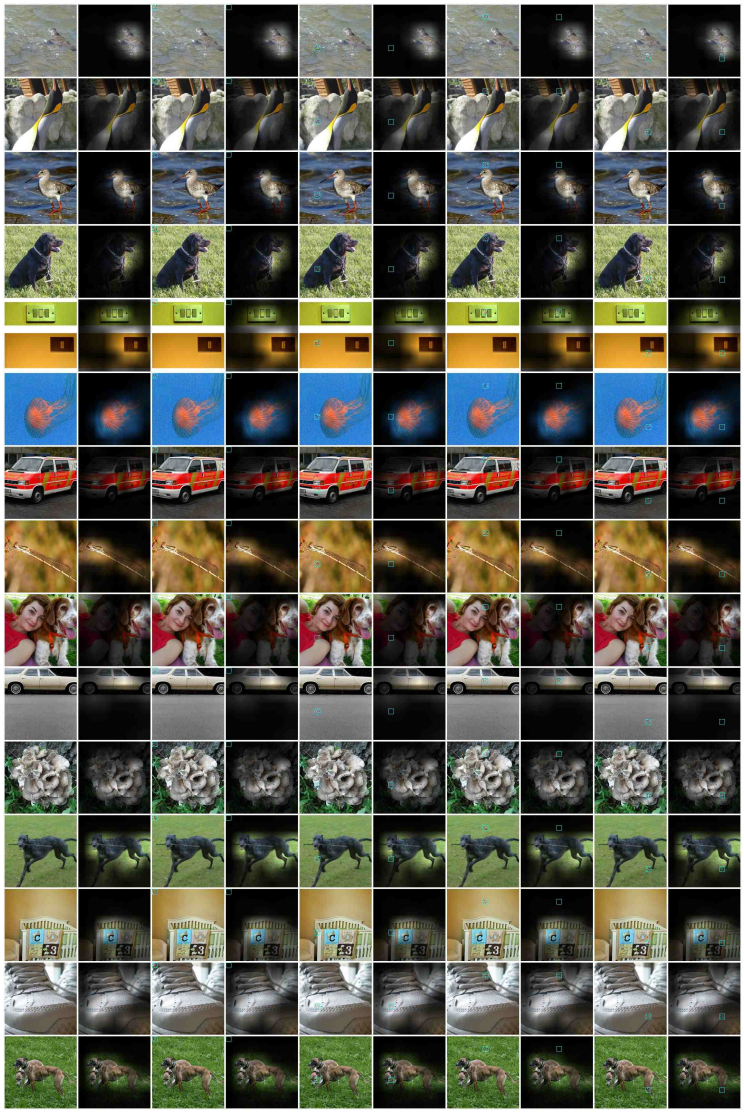


Fig. 20: Averaged Feature Maps of ResNet50 as Attention with Attack Patch size of 16 on Corrupted Images

References

1. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
2. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. In: International Conference on Learning Representations (ICLR) (2019)
3. Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., Houlsby, N.: Big transfer (bit): General visual representation learning. In: European Conference on Computer Vision (ECCV) (2020)