Exploring Adversarial Robustness of Vision Transformers in the Spectral Perspective -Supplementary Material-

Gihyun Kim Juyeop Kim Jong-Seok Lee
Yonsei University, Republic of Korea
{kkh9314,juyeopkim,jong-seok.lee}@yonsei.ac.kr

A. Results using Other Image Quality Metrics

While we use PSNR as the image quality metric in the main paper, results using other perceptual image quality metrics are shown here. We choose the multi-scale structural similarity index measure (MS-SSIM) [5], which is advanced form of SSIM that is one of the most popular perceptual metrics, and the mean deviation similarity index (MDSI) [3], which was shown to perform the best in [2], and learned perceptual image patch similarity (LPIPS) [6], which uses the features from a pre-trained network for classification. The results are shown in Figures 1-6. Note that a larger value of MDSI or LPIPS indicates a lower quality level. Overall, the trends remain similar to those shown in Figures 3 and 4 of the main paper.

B. Comparison of Models Pre-trained on ImageNet-1k

In Figures 7 and 8, we show the results for models pretrained on ImageNet-1k from [4]. The results show similar trends to those shown in Figures 3 and 4 of the main paper.

C. Combinations of Employed Perturbations

We additionally conduct experiments by employing multiple perturbations at the same time. In particular, we consider employing both δ_{mag} and δ_{phase} (“mag+phase” attack) and all of δ_{mag}, δ_{phase}, and δ_{pixel} (“mag+phase+pixel” attack). The results are in Figures 9. It is observed that the mag+phase attack tends to be similar to or weaker than the single component attacks for ResNets or Transformers, respectively; it seems that using more variables to be optimized makes optimization more challenging. And, perturbing all components (i.e., the magnitude+phase+pixel attack) does not improve the attack performance compared to the pixel attack for all models.

D. Average Distribution of Distortion

The distributions of the distortion by the phase attack over different frequency regions, averaged over all images, are shown in Figure 10. The trends are the same to those observed in Figure 5 of the main paper.

E. Change in Attention Map after Attack

We examine how much the attacks change the image regions attended by Transformers. The attention maps are obtained by the rollout method [1] for each pair of the original and attacked images, and the Pearson correlation coefficient is computed between them. Figure 11 plots the histogram of the correlation coefficients for DeiT-S under the magnitude, phase, and pixel attacks, and Figure 12 shows example attention maps. The attacked images show PSNR of about 50 dB on average for each of the three attacks. The average correlation coefficients are 0.873, 0.886, and 0.918 for the magnitude, phase, and pixel attacks, respectively. The phase attack causes larger deviations in the attention patterns than the magnitude and pixel attacks, which seems to lead to higher vulnerability to the phase attack as shown in Section 4.2 of the main paper.

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, 2020. 1

[2] Sergey Kastryulin, Dzhamil Zakirov, and Denis Prokopenko. PyTorch Image Quality: Metrics and measure for image quality assessment, 2019. Open-source software available at https://github.com/photosynthesis-team/piq. 1

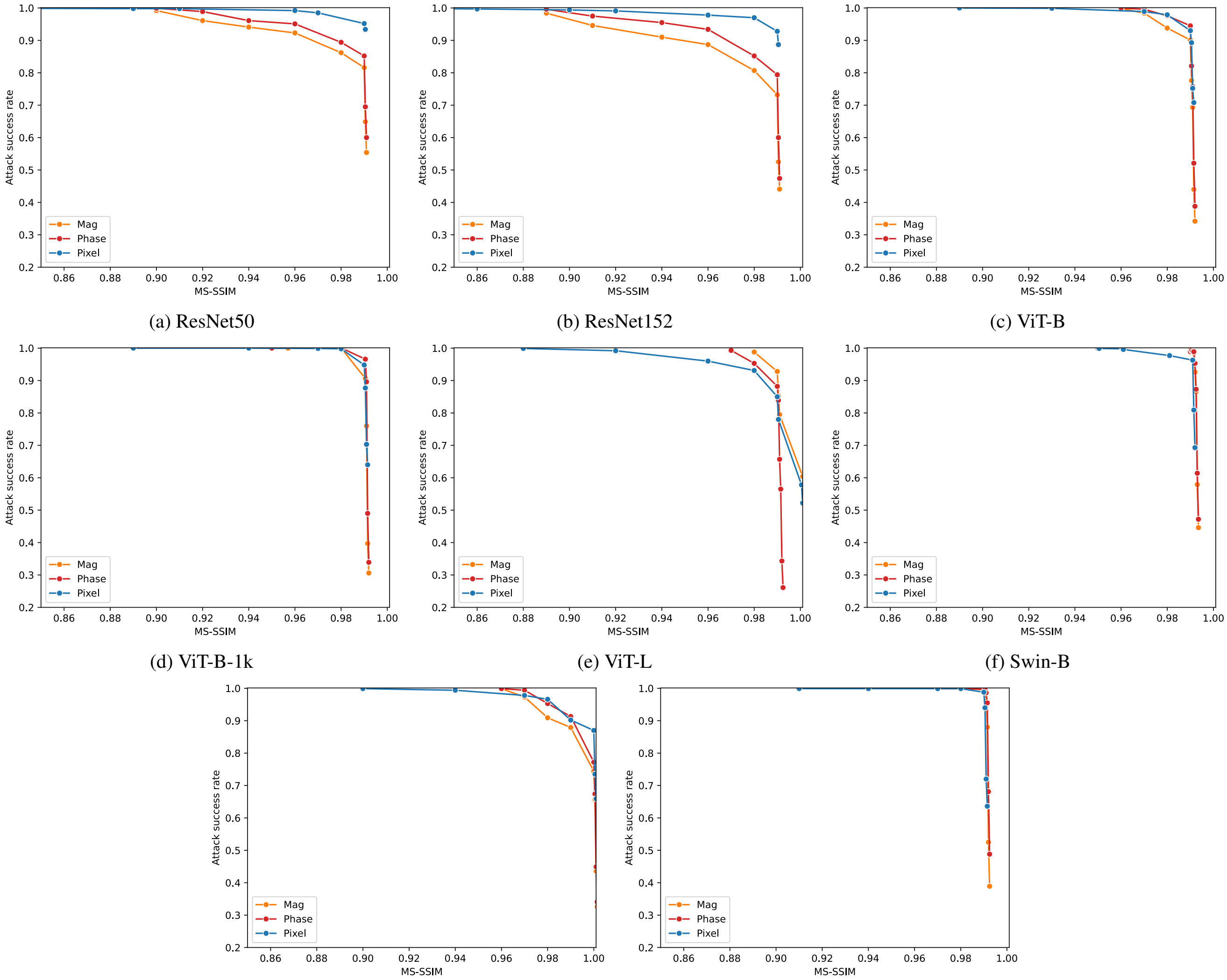
[3] Hossein Ziaei Nafchi, Atena Shahkolaei, Rachid Hedjam, and Mohamed Cheriet. Mean deviation similarity index: Efficient and reliable full-reference image quality evaluator. *IEEE Access*, 4:5579–5590, 2016. 1

[4] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming

Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019. 1

[5] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *Proceedings of the Asilomar Conference on Signals, Systems & Computers*, volume 2, pages 1398–1402, 2003. 1

[6] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 1



(a) ResNet50

(b) ResNet152

(c) ViT-B

(d) ViT-B-1k

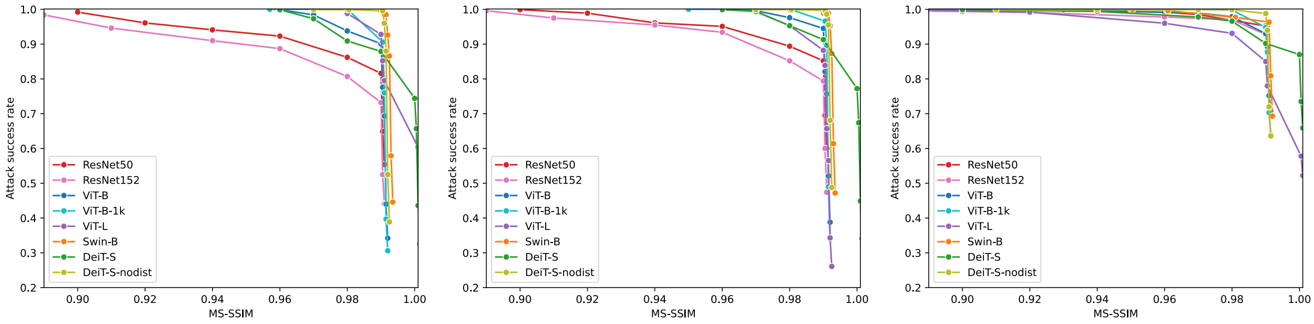
(e) ViT-L

(f) Swin-B

(g) DeiT-S

(h) DeiT-S with no distillation

Figure 1. Comparison of different attacks for each model using MS-SSIM.

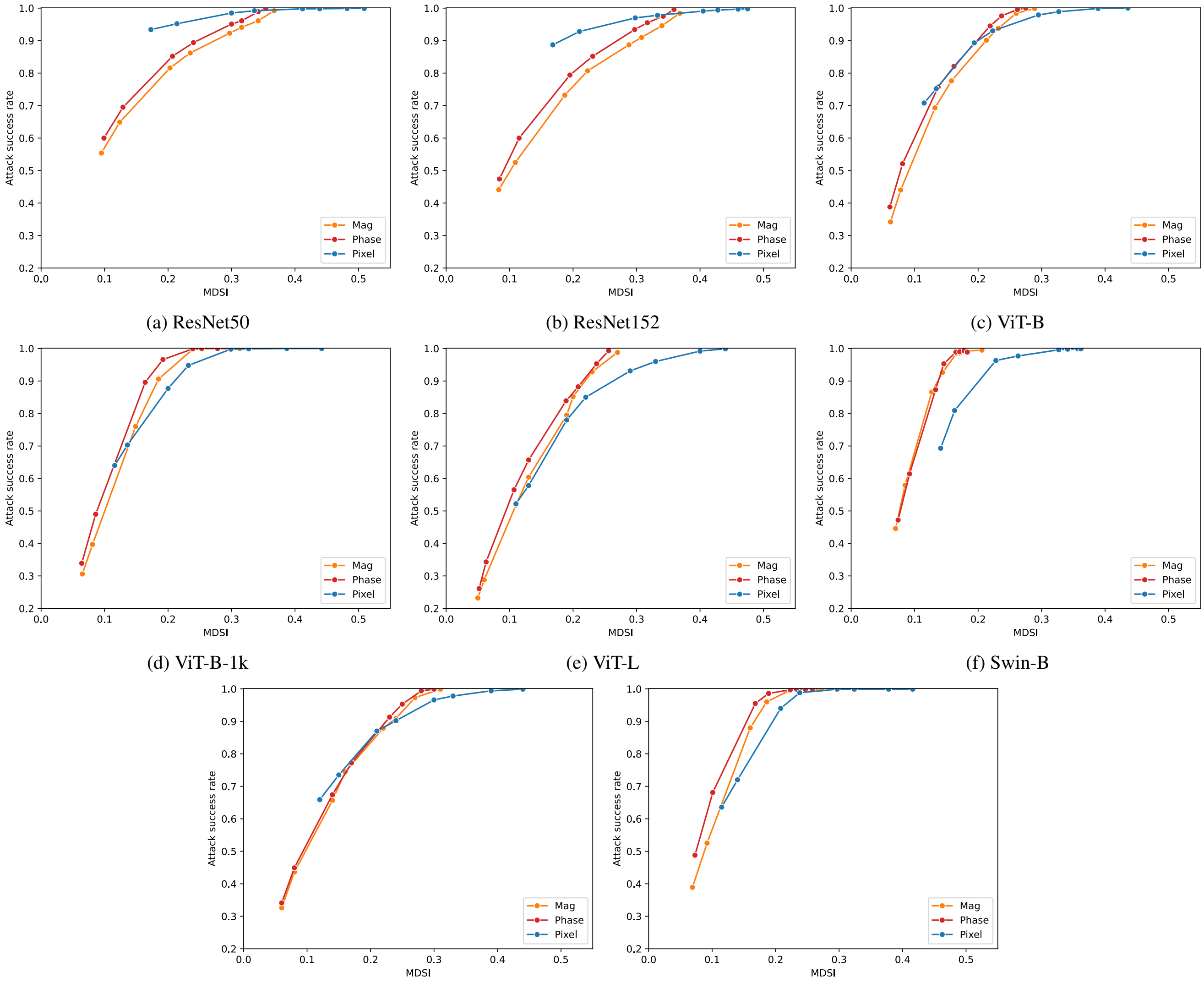


(a) Magnitude attack

(b) Phase attack

(c) Pixel attack

Figure 2. Comparison of different models for each attack type using MS-SSIM.



(a) ResNet50

(b) ResNet152

(c) ViT-B

(d) ViT-B-1k

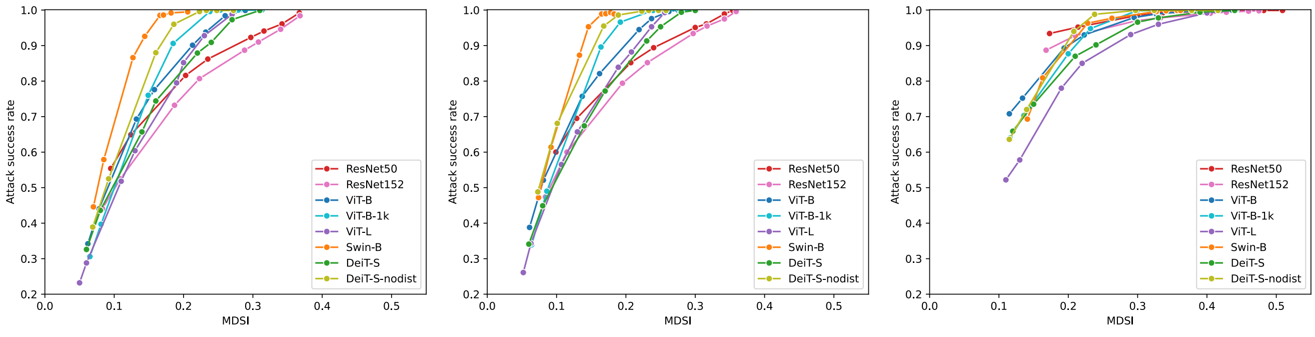
(e) ViT-L

(f) Swin-B

(g) DeiT-S

(h) DeiT-S with no distillation

Figure 3. Comparison of different attacks for each model using MDSI.

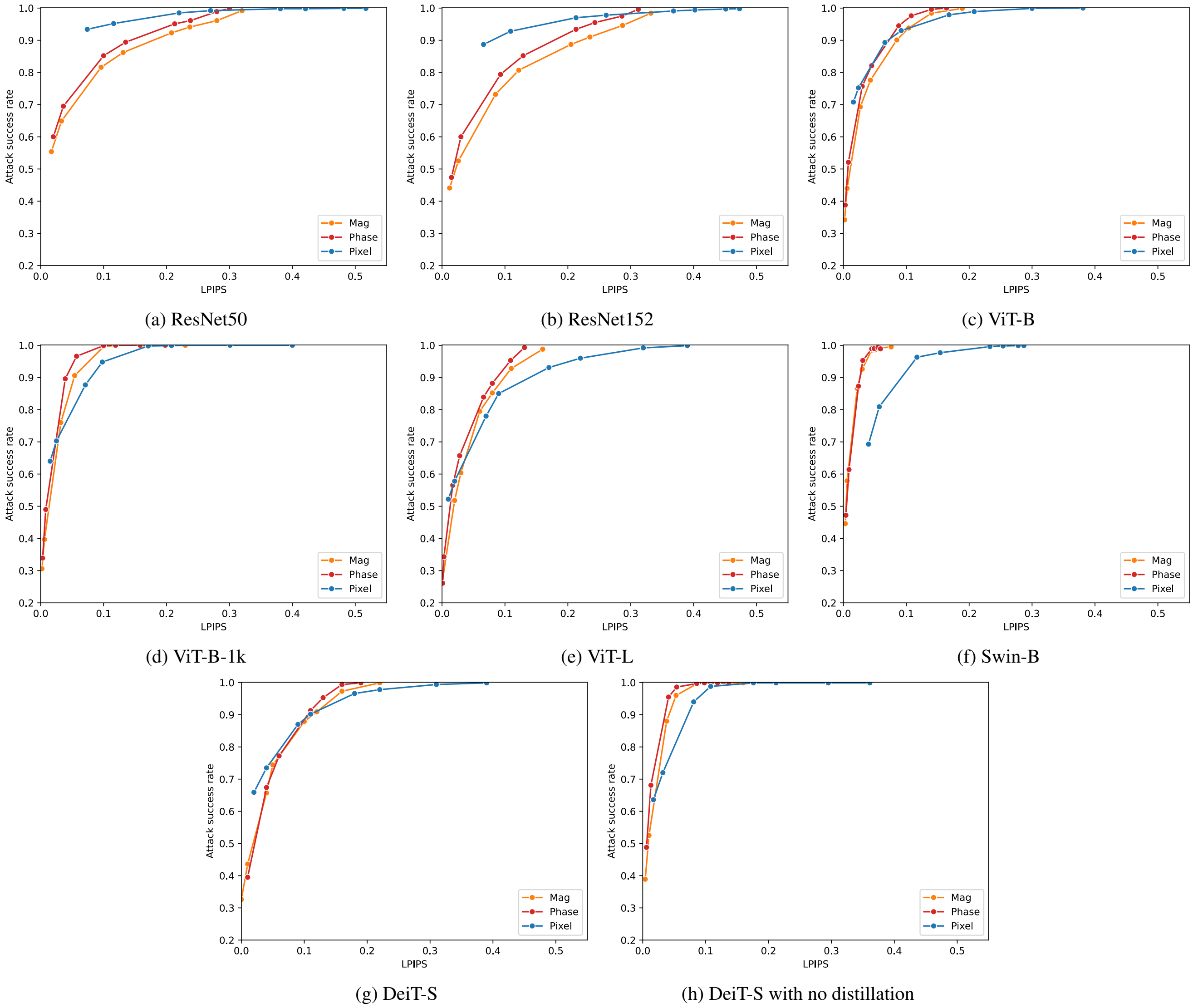


(a) Magnitude attack

(b) Phase attack

(c) Pixel attack

Figure 4. Comparison of different models for each attack type using MDSI.



(a) ResNet50

(b) ResNet152

(c) ViT-B

(d) ViT-B-1k

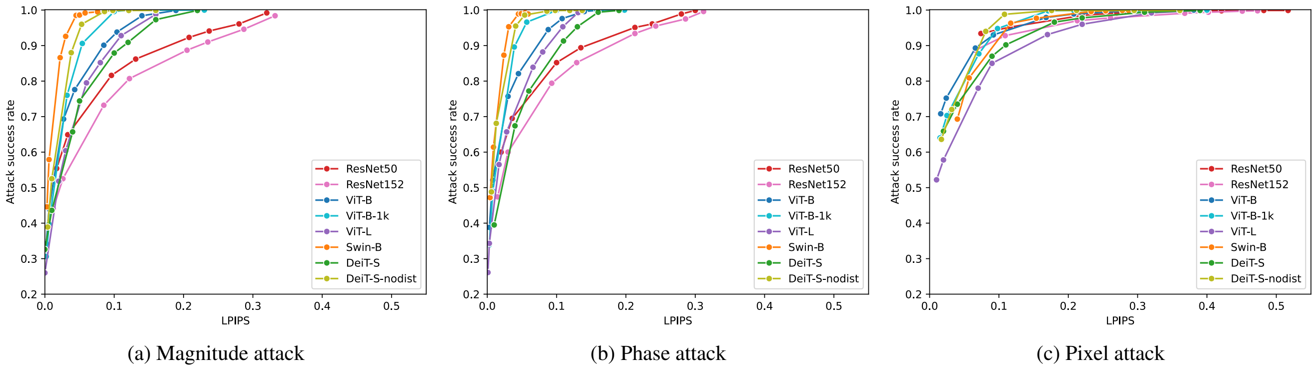
(e) ViT-L

(f) Swin-B

(g) DeiT-S

(h) DeiT-S with no distillation

Figure 5. Comparison of different attacks for each model using LPIPS.

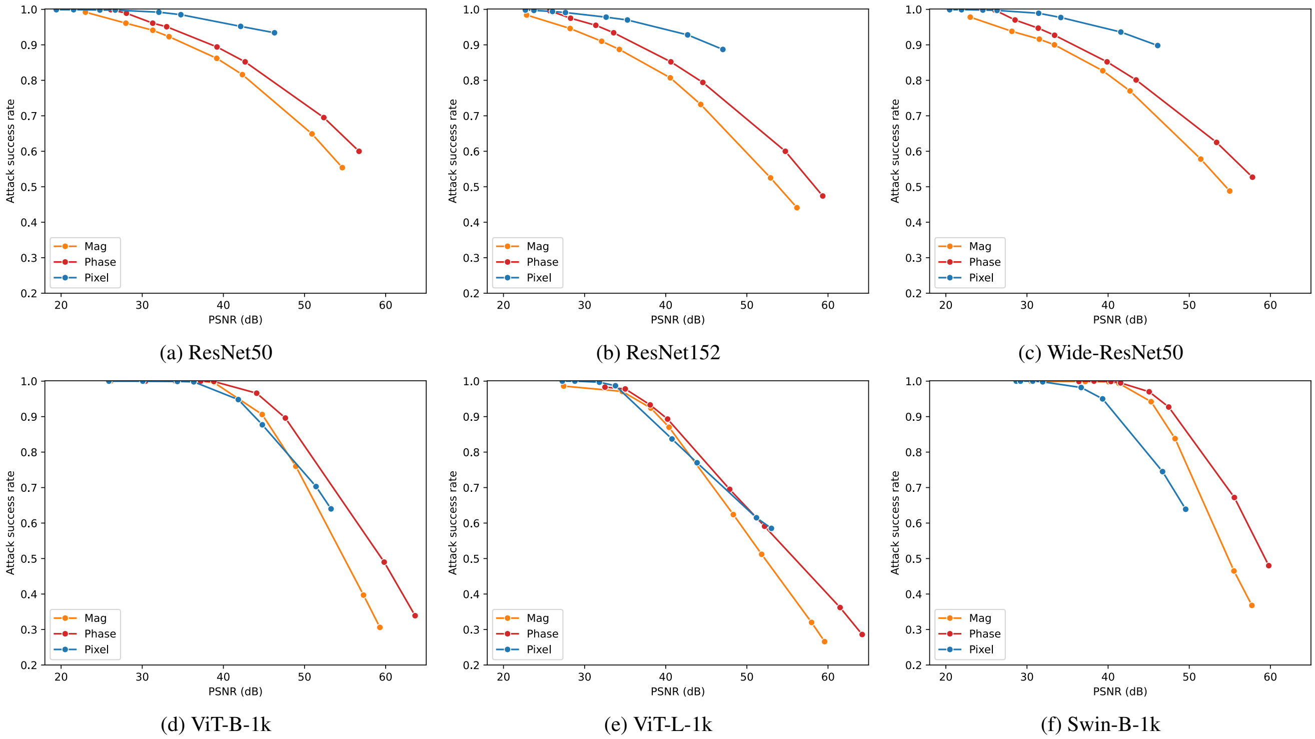


(a) Magnitude attack

(b) Phase attack

(c) Pixel attack

Figure 6. Comparison of different models for each attack type using LPIPS.



(a) ResNet50

(b) ResNet152

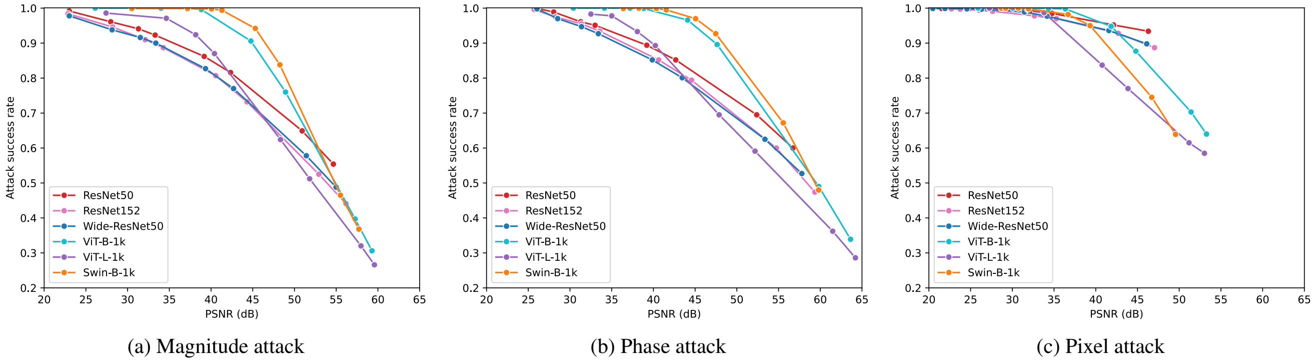
(c) Wide-ResNet50

(d) ViT-B-1k

(e) ViT-L-1k

(f) Swin-B-1k

Figure 7. Comparison of different attacks for each model pre-trained on ImageNet-1k.

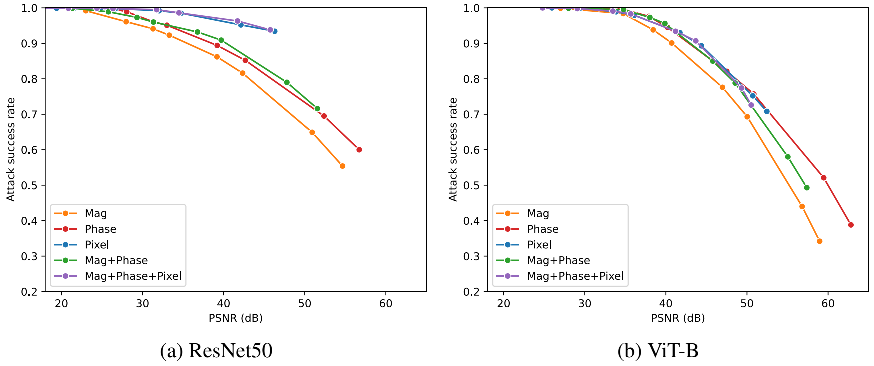


(a) Magnitude attack

(b) Phase attack

(c) Pixel attack

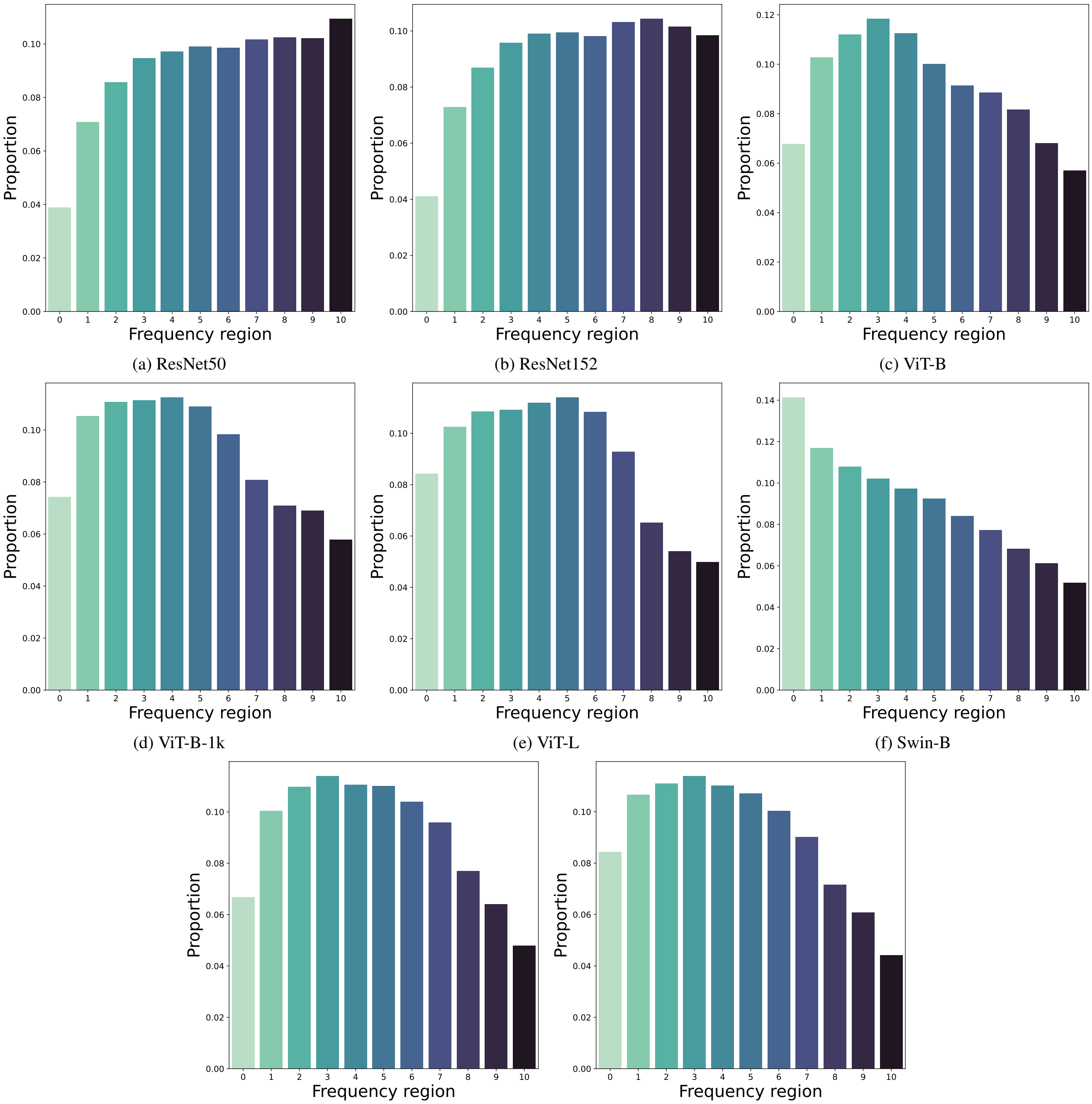
Figure 8. Comparison of different models pre-trained on ImageNet-1k for each attack type.



(a) ResNet50

(b) ViT-B

Figure 9. Comparison of various combinations of employed perturbations.



(a) ResNet50

(b) ResNet152

(c) ViT-B

(d) ViT-B-1k

(e) ViT-L

(f) Swin-B

(g) DeiT-S

(h) DeiT-S with no distillation

Figure 10. Average distributions of distortion over different frequency regions.

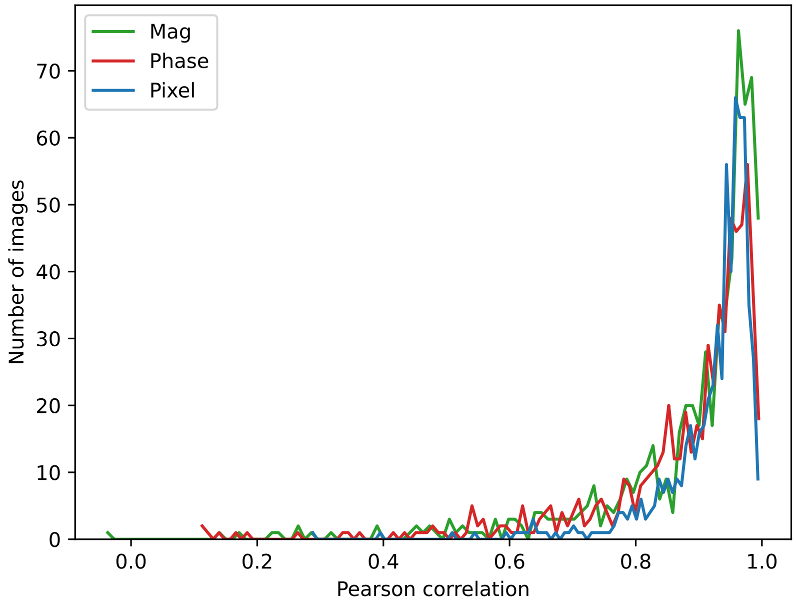
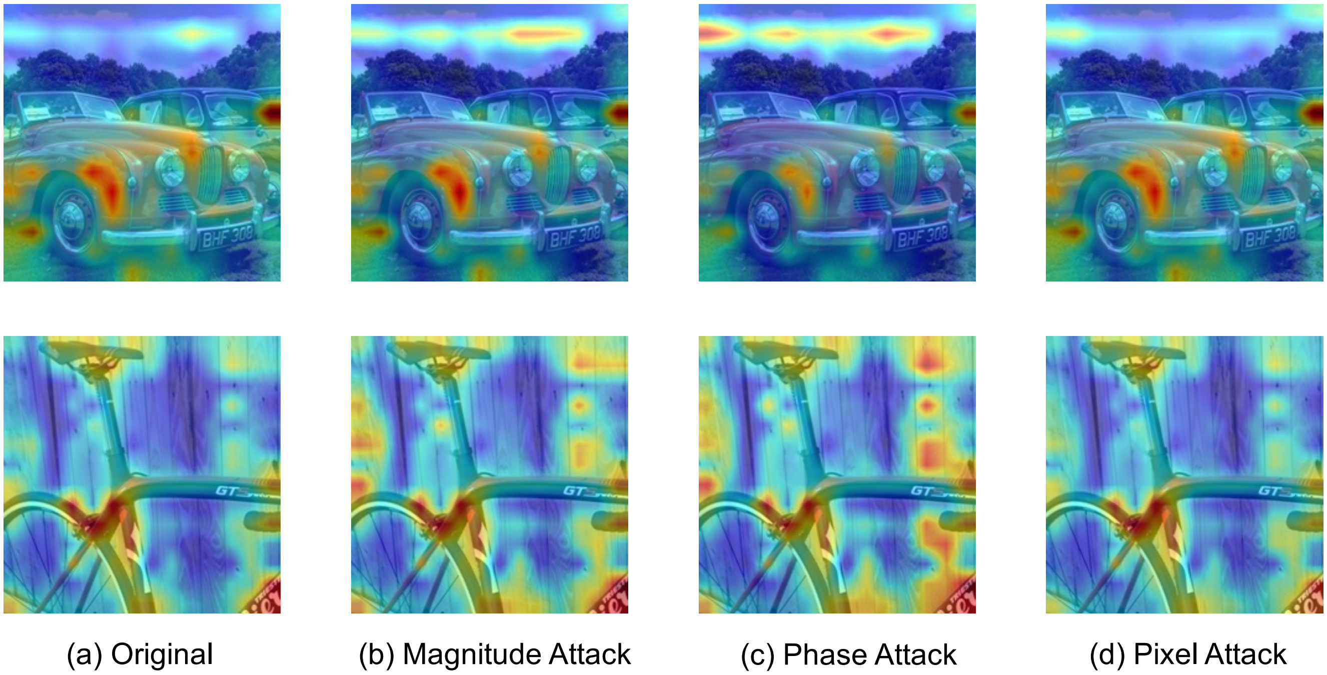


Figure 11. Histogram of the correlation coefficient between the attention maps for the original and attacked images.


(a) Original

(b) Magnitude Attack

(c) Phase Attack

(d) Pixel Attack

Figure 12. Example attention maps for the original and attacked images.