ANALYZING ADVERSARIAL ROBUSTNESS OF VISION TRANSFORMERS AGAINST SPATIAL AND SPECTRAL ATTACKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision Transformers have emerged as a powerful architecture that can outperform convolutional neural networks (CNNs) in image classification tasks. Several attempts have been made to understand robustness of Transformers against adversarial attacks, but existing studies draw inconsistent results, i.e., some conclude that Transformers are more robust than CNNs, while some others find that they have similar degrees of robustness. In this paper, we address two issues unexplored in the existing studies examining adversarial robustness of Transformers. First, we argue that the image quality should be simultaneously considered in evaluating adversarial robustness. We find that the superiority of one architecture to another in terms of robustness can change depending on the attack strength expressed by the quality of the attacked images. Second, by noting that Transformers and CNNs rely on different types of information in images, we formulate an attack framework as a tool for implementing flexible attacks, where an image can be attacked in the spectral domain as well as in the spatial domain. This attack perturbs the magnitude and phase information of particular frequency components selectively. Through extensive experiments, we find that Transformers tend to rely more on phase information and low frequency information than CNNs, and thus sometimes they are even more vulnerable under frequency-selective attacks. It is our hope that this work provides new perspectives in understanding the properties and adversarial robustness of Transformers.

1 INTRODUCTION

Convolution neural networks (CNNs) have been a dominant neural network architecture in computer vision for a long time. However, Transformer-based structures have recently emerged as another promising architecture (Dosovitskiy et al., 2021), achieving even better performance than CNNs especially in image classification.

CNNs are known to be vulnerable to adversarial attacks, i.e., an imperceptible perturbation added to an image can fool a trained CNN so that it misclassifies the attacked image (Akhtar & Mian, 2018). Investigating robustness of a model against adversarial attacks is important because not only the vulnerability issue is critical in security-sensitive applications but also such investigation can lead to better understanding of the operating mechanism of the model. Then, a naturally arising question is: how vulnerable are Transformers compared to CNNs?

The researches comparing adversarial robustness of CNNs and Transformers do not reach consistent conclusions. One group of studies claims that Transformers are more robust to adversarial attacks than CNNs (Naseer et al., 2021; Benz et al., 2021; Shao et al., 2021; Aldahdooh et al., 2021). However, another group of studies claims that the two architectures have similar levels of robustness (Bai et al., 2021; Mahmood et al., 2021; Bhojanapalli et al., 2021).

As an effort to further investigate the adversarial robustness of Transformers, we address two issues that have been unexplored previously. First, we argue that the visual quality of the attacked images needs to be considered properly when adversarial robustness is evaluated. We note that the existing studies often consider only a limited number of conditions regarding the amount of adversarial perturbation (i.e., attack strength). However, the superiority in terms of robustness between models

may change depending on the amount of perturbation. In addition, different attack methods control the amount of perturbation with different algorithmic parameters (e.g., L_∞ norm), which makes it difficult to compare their results. Moreover, the perturbations optimized for different models by different attack methods may have different visual patterns, for which algorithmic parameters may be inadequate to be a criterion of the amount of perturbation. Therefore, we suggest using image quality to represent the amount of perturbation, and also considering a wide range of image quality for comprehensive analysis. Second, noting that different models rely on different features in images, we consider a flexible attack framework to effectively attack both CNNs and Transformers. Previous studies identify that Transformers tend to rely more on low frequency features (Park & Kim, 2022; Benz et al., 2021), while CNNs focus more on high frequency features (Wang et al., 2020; Jo & Bengio, 2017). Popular gradient-based attack methods tend to perturb high frequency features in images, which may make CNNs be fooled more easily than Transformers. We formulate an attack framework, which can perturb images flexibly in both spatial and spectral domains.

Considering these issues, we conduct extensive experiments to compare adversarial robustness of off-the-shelf CNN and Transformer models. Our contributions can be summarized as follows.

- We evaluate various models of CNN and Transformer over a wide range of image quality of attacked images because the relative robustness between models sometimes changes depending on the image quality.
- We formulate an attack framework that can perturb the magnitude and phase spectra in the spectral domain and the pixel values in the spatial domain. In particular, examining frequency-selective perturbations by the attacks is the key to deeper understanding of the adversarial robustness of Transformers.
- We conduct in-depth analyses to investigate the frequency-dependent behaviours and importance of spectral information in Transformers.

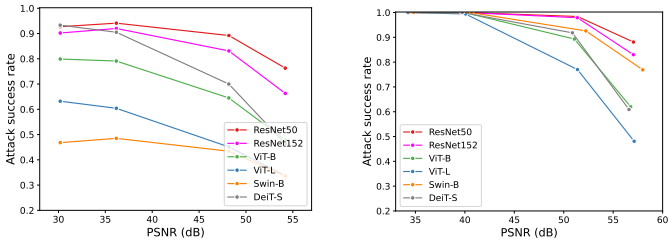
2 Related Works

2.1 Vision Transformers

The vision Transformer (ViT) has appeared as a powerful neural architecture using the self-attention mechanism (Dosovitskiy et al., 2021). Several variants of ViT have been also proposed. The Swin Transformer (Liu et al., 2021) improves the efficiency over ViT using a shifted window scheme for self-attention. To resolve the issue that Transformers require a large dataset for training, the data-efficient image Transformer (DeiT) (Touvron et al., 2021) is trained through distillation from a CNN teacher. Other variants include token-to-token ViT (Yuan et al., 2021), pyramid ViT (Wang et al., 2021), Transformer in Transformer (Han et al., 2021), cross-covariance image Transformer (Ali et al., 2021), etc.

2.2 Adversarial Attack Methods

The goal of a typical adversarial attack is to change the classification result of a model by injecting a noise-like perturbation to the image while the perturbation is kept imperceptible in order not to be detected easily. The perturbation is usually found via gradient-based optimization. The fast gradient sign method (FGSM) (Goodfellow et al., 2015) uses the sign of gradient of the classification loss. The projected gradient descent (PGD) method (Madry et al., 2018) implements a stronger attack by iteratively optimizing the perturbation. These attacks limit the L_p norm of the perturbation to control the imperceptibility. The C&W attack (Carlini & Wagner, 2017) optimizes the weighted sum of the amount of perturbation and the classification loss, which is known to be one of the strongest attacks. Frequency-domain filtering can be use to constrain perturbations only in certain frequency regions (Li et al., 2020; Guo et al., 2020; Sharma et al., 2019), which are still based on image-domain attacks. Recently, direct perturbation in the frequency domain was also tried (Wen, 2022), which is different from our method that can perturb the magnitude and phase components separately. Note that we do not intend to develop a new stronger attack in the frequency domain, but aim to formulate a unified attack framework that can perturb the pixel values, magnitude spectrum, and phase spectrum of an image separately or simultaneously in a flexible manner.



(a) FGSM (b) PGD

Figure 1: PSNR vs. attack success rate for CNNs and Transformers under the FGSM and PGD attacks.

2.3 Adversarial Robustness of Transformers

As mentioned in the introduction, two conflicting conclusions exist in the literature regarding the relative adversarial robustness of CNNs and Transformers. In one group of studies, it is claimed that Transformers are more robust than CNNs. Benz et al. (2021); Shao et al. (2021); Aldahdooh et al. (2021); Paul & Chen (2022) commonly make a conclusion that Transformers are more robust against gradient-based attacks including FGSM, PGD, and C&W because CNNs rely on high frequency information while Transformers do not. In addition, Benz et al. (2021) claim that the shift-invariant property of CNNs might be a reason for their lower robustness. Another group of studies argues that Transformers are as vulnerable to attacks as CNNs. Mahmood et al. (2021) find that ViTs are not advantageous over the ResNet and big transfer (BiT) models (Kolesnikov et al., 2020) in terms of robustness against various attack methods such as FGSM, PGD, and C&W. Bhojanapalli et al. (2021) show that CNNs and Transformers are similarly vulnerable against various natural and adversarial perturbations. Bai et al. (2021) attempt to compare CNNs and Transformers on a common training setup, from which it is concluded that they have similar adversarial robustness.

The aforementioned studies usually do not consider the trade-off characteristics between vulnerability and attack strength in an extensive manner over a wide range of image quality degradation. In addition, most of the studies consider only pixel-domain perturbations. These issues are dealt with in our work.

3 Method

In this section, we present the two issues considered in our comparison of adversarial robustness of CNNs and Transformers, and explain how we perform the comparison.

3.1 Consideration of Image Quality

In many popular attack methods, the attack strength can be controlled by certain parameters (e.g., L_∞ norm of perturbation in FGSM, the balancing parameter between the amount of distortion and the change of the classification loss in C&W). As an attack becomes strong, the target model becomes more vulnerable, but the image distortion may become perceptible. In the previous studies, such a trade-off relationship has not been fully considered in benchmarking the adversarial robustness of Transformers. However, depending on the considered attack strength, the superiority of one model to another in terms of robustness may vary. Figure 1 shows an example of this issue, which shows the attack success rate (ASR) on CNNs and Transformers with respect to the attack strength (represented as the image quality in peak signal-to-noise ratio (PSNR)) for the FGSM and PGD attacks (see Appendix for the experimental setup).

Overall, the ResNet models tend to be more vulnerable than the Transformers in both attacks by showing higher ASR, which is consistent with the results in (Benz et al., 2021; Shao et al., 2021; Aldahdooh et al., 2021). When we have a look at the details, additional observations can be also made. For instance, in Figure 1a, DeiT-S is more robust than both ResNet models over 40 dB but they show similar vulnerability at 30-35 dB.

In addition, in Figure 1b, all models are similarly vulnerable showing almost 100% of ASR for low PSNR values, but their vulnerability becomes different after about 40 dB. These results demonstrate that it is important to examine a wide range of image quality in order to better understand the vulnerability of different models.

While the attack strength and image quality are generally compatible in the context of adversarial attack, we consider the latter as a more proper way due to the following reasons. (1) The attack strength does not consider visual perception of the perturbation unlike image quality. PSNR is defined in the signal point of view but already shows moderately high correlation with human perception (Athar & Wang, 2019), and thus has potential as an image quality metric for robustness evaluation. Furthermore, many perceptual image quality metrics have been developed in literature, which can be used to measure the perceptual amount of perturbation. (2) At the same attack strength, the perturbations optimized by different attack methods on different models may have visually different patterns, which need to be quantified appropriately. As a hypothetical example of an L_∞ norm-constrained attack, consider that a perturbation has a nonzero value at a single pixel location, while another perturbation has nonzero values at many pixel locations over the whole image. Both cases satisfy the constraint, but they should be treated differently, which can be accomplished using an image quality metric. (3) As aforementioned, different attack methods use different algorithmic parameters to control the attack strength, which cannot be easily compared with each other. This problem is readily resolved if an image quality metric is used. (4) We consider an attack framework perturbing different components (i.e., magnitude, phase, and pixel). It is observed that the perturbation patterns are significantly different depending on the target component. An image quality metric is required in order to compare the attack strengths in different components properly.

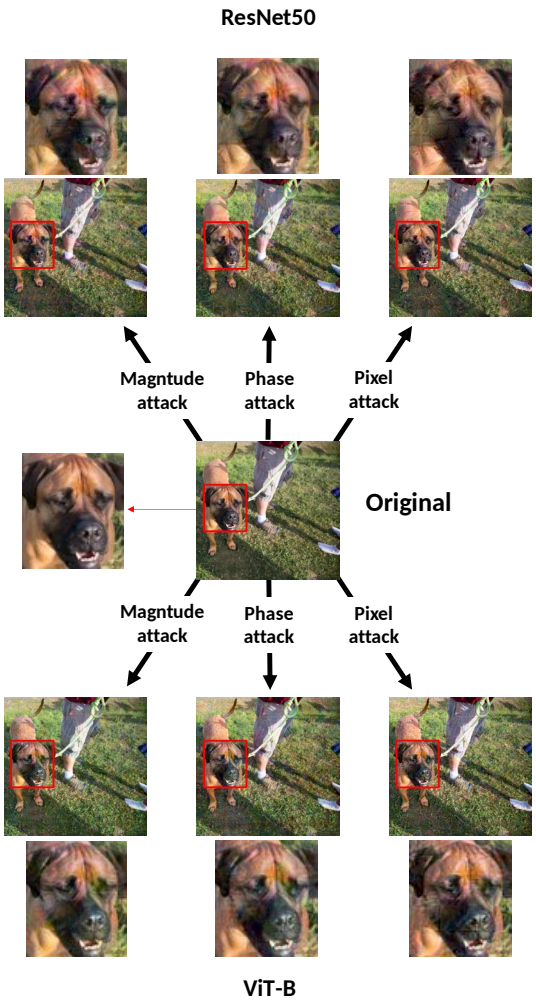


Figure 2: Example case of our attack perturbing the magnitude, phase, and pixel values, respectively, for ResNet50 and ViT-B.

3.2 Our Attack

Different from the existing works, we formulate a unified attack framework to attack images both spatially and spectrally. The following considerations motivate us to use the framework. First, the previous studies point out that CNNs tend to rely on high frequency information in images, which is why they appear more vulnerable than Transformers (Benz et al., 2021; Shao et al., 2021). In other words, the popular attack methods injecting high frequency noise are disadvantageous to CNNs. Our attack framework tries to alleviate such an inherent bias by enabling to flexibly perturb images in both spatial location-selective and frequency-selective manners. Second, we aim to analyze the mechanisms of Transformers in various viewpoints through the results of attacks in different domains.

The Fourier transform of an image X can be written by

$$\mathcal{F}\{X\} = M \cdot e^{j\phi}, \tag{1}$$

where M and ϕ are the magnitude and phase spectra, respectively. The image is attacked by the combination of multiplicative magnitude perturbation[1] δ_{mag}, additive phase perturbation δ_{phase}, and additive pixel perturbation δ_{pixel} as follows:

$$\tilde{X}' = \mathcal{F}^{-1}\Big\{\mathrm{clip}_{0,\infty}(M \otimes \delta_{\mathrm{mag}}) \cdot e^{j(\phi+\delta_{\mathrm{phase}})}\Big\} + \delta_{\mathrm{pixel}}, \tag{2}$$

$$X' = \mathrm{clip}_{0,1}(\tilde{X}'), \tag{3}$$

where $\mathcal{F}^{-1}$ is the inverse Fourier transform, $\otimes$ is the element-wise multiplication, and $\mathrm{clip}_{a,b}(x)$ clips each element of x within a and b. Here, we assume that the pixel values are normalized within 0 and 1. Note that δ_{mag} and δ_{phase} are kept to be symmetric in order to ensure the resulting image after the inverse Fourier transform to have real-valued pixel values.

We consider various combinations of employed perturbations, i.e., single component attacks employing only one among δ_{mag}, δ_{phase}, and δ_{pixel}, denoted as "magnitude attack," "phase attack," and "pixel attack," respectively, the attack employing both δ_{mag} and δ_{phase}, denoted as "mag+phase attack," and the attack employing all perturbations, denoted as "mag+phase+pixel attack."

The process to optimize the perturbations is inspired by the C&W attack (Carlini & Wagner, 2017). In other words, we minimize the L_2 difference (i.e., mean squared error (MSE)) between the original and attacked images and maximize the cross-entropy (CE) loss. Thus, the loss function of our attack is given by

$$\mathrm{Loss} = \lambda \cdot \mathrm{MSE}(X', X) - \mathrm{CE}(f(X'), y), \tag{4}$$

where λ is a parameter balancing MSE and CE, $f(\cdot)$ is the model, and y is the ground truth.

Figure 2 shows an example case, where each of the magnitude, phase, and pixel components is perturbed. It can be observed that attacking different components induce different distortion patterns in the image. The distortion pattern also varies depending on the target model.

4 Experiments

4.1 Setup

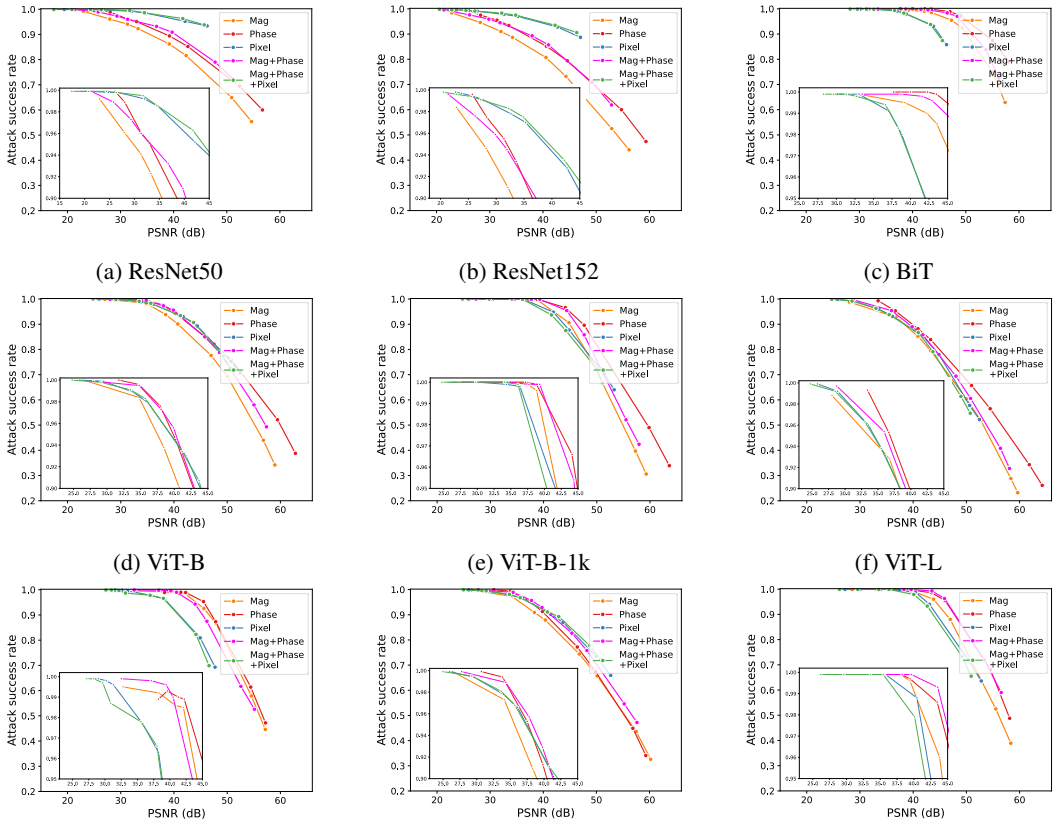
We aim to benchmark the adversarial robustness of the pre-trained models that serve as off-the-shelf solutions in general image classification applications. We consider ResNet50, ResNet152, and BiT as CNNs, and ViT-B/16, ViT-L/16, DeiT-S, and Swin-B as Transformers. Here, S, B and L mean small, base, and large, and /16 means the patch size. ResNet50 and ResNet152 are from the torchvision models (Paszke et al., 2019) trained on ImageNet-1k (Russakovsky et al., 2015). BiT based on ResNet152x4, ViT-B, ViT-L, DeiT-S, and Swin-B are from the timm module (Wightman, 2019), which are pretrained on ImageNet-21k (Deng et al., 2009) and finetuned on ImageNet-1k. ViT trained on ImageNet-1k and DeiT-S without distillation are also considered.

To evaluate adversarial robustness of the models, the image dataset from the NeurIPS 2017 Adversarial Challenge (Wightman, 2017) is used. For BiT, the images are resized to 480×480 to match the input dimension.

To optimize the perturbations, we use the Adam optimizer with a fixed learning rate of 5×10^{-3} and a weight decay parameter of 5×10^{-6}. The maximum number of iterations is set to 1000. We also set a termination condition that the optimization stops when the loss in Eq. (4) does not improve for five consecutive iterations. We vary the value of λ as $\{1, 10^3, 5 \times 10^3, 10^4, 5 \times 10^4, 10^5, 5 \times 10^5, 10^6\}$.

We report the results using PSNR as the image quality metric. The results using other metrics, mean deviation similarity index (MDSI) (Nafchi et al., 2016) and learned perceptual image patch similarity (LPIPS) (Zhang et al., 2018), are shown in Appendix.

[1]We also tried an additive magnitude perturbation but it was not optimized well because the magnitude spectrum has values over a wide range.



(a) ResNet50 (b) ResNet152 (c) BiT

(d) ViT-B (e) ViT-B-1k (f) ViT-L

(g) Swin-B (h) DeiT-S (i) DeiT-S with no distillation

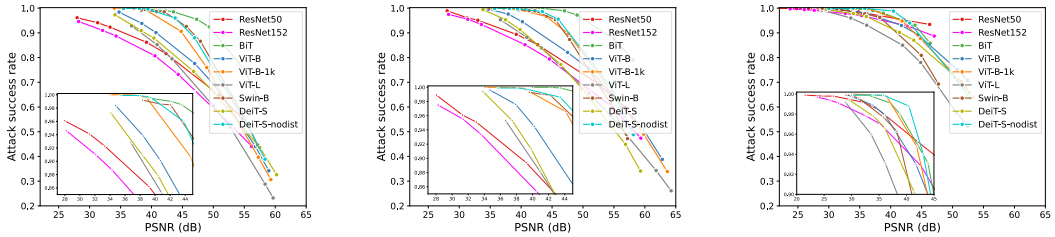
Figure 3: Comparison of different attacks for each model. The range for high attack success rates is enlarged for better visualization.

4.2 Results

4.2.1 Comparison of Attacks

We first compare different attacks in our attack framework in order to show the effectiveness of attacking in the spectral domain. Figure 3 shows the results in terms of ASR with respect to PSNR. All the attacks can achieve (almost) 100% of ASR at the lower extremes of PSNR for all models. However, their relative effectiveness varies depending on the model. For ResNet50 and ResNet152, the pixel attack appears stronger than the magnitude attack and the phase attack, whereas the phase attack is the strongest for ViTs and Swin-B. This shows that the flexible frequency domain attack overcomes the limitation of the pixel attack perturbing mainly high frequency information, and becomes a strong attack on the Transformers. DeiT-S is more vulnerable to the pixel attack as in the ResNet models, probably because it is trained using distillation from the CNN teacher, which is supported by the observation that DeiT-S without distillation is more vulnerable to the phase attack as in ViTs and Swin-B. 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. Thus, we consider only the single component attacks in the following analyses. Another observation is that the phase attack is mostly stronger than the magnitude attack. Further analysis on this is presented later.

4.2.2 Comparison of Vulnerability of Models



(a) Magnitude attack (b) Phase attack (c) Pixel attack

Figure 4: Comparison of different models for each attack type. The range for high attack success rates is enlarged for better visualization.

Figure 4 compares ASR of the models with respect to PSNR for each attack type. Most importantly, when ASR is relatively high, the Transformers are similar to or more robust than the ResNet models under the pixel attack (as in the case of PGD in Figure 1b) but more vulnerable to the magnitude and phase attacks. Again, this is due to the flexible perturbations in the frequency domain, which will be further analyzed later. When PSNR is high (over 45-50 dB), these trends do not hold any more, i.e., some Transformers (ViTs and DeiT-S) become similarly robust to the ResNet models under the magnitude and phase attacks, and the ResNet models become more vulnerable to all Transformers under the pixel attack. When Swin-B is compared to ViT-B and ViT-L, the former shows higher vulnerability than the latter for the magnitude and phase attacks. When the model size is concerned, larger models (ViT-L and ResNet152) are more robust than smaller models (ViT-B and ResNet50) under all attacks. While Bhojanapalli et al. (2021) also observed a similar trend using pixel-domain attacks (FGSM and PGD), we find that the same also holds for the attacks in the spectral domain. When ViT-B and ViT-B-1k are compared, ViT-B is more robust than ViT-B-1k, but becomes more vulnerable when PSNR becomes high, especially for the pixel attack. Bhojanapalli et al. (2021) noted that training with a larger dataset is beneficial for robustness; we additionally find that the benefit of a larger dataset is even more prominent for the magnitude and phase attacks than for the pixel attack, but the benefit disappears when the amount of perturbation is small. DeiT-S behaves more like ResNet than the other Transformers due to the distillation using CNN. It is interesting to see that BiT, which has the ResNet structure, shows the highest vulnerability to the magnitude and phase attacks and is one of the most vulnerable models under the pixel attack. Mahmood et al. (2021) experimented the C&W attack, which can be considered to be similar to our pixel attack, but all models recorded 100% of ASR and thus no comparative conclusion could be made; in contrast, our result considering a wide range of image quality for the pixel attack reveals the high vulnerability of BiT.

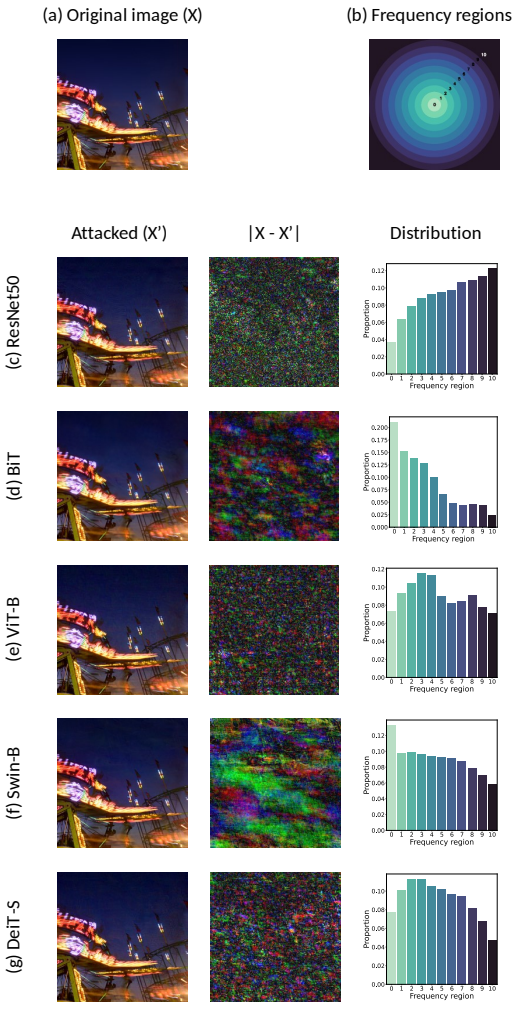
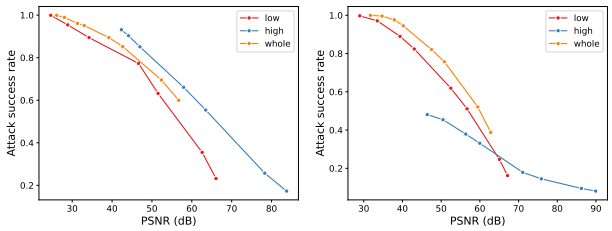
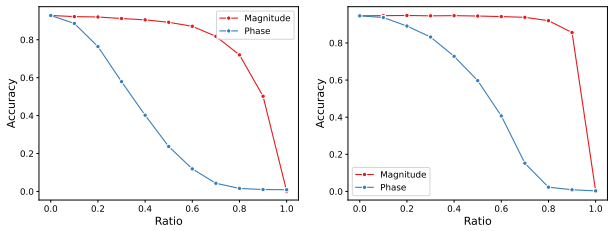


Figure 5: Example of the attacked image, distortion in the pixel domain (magnified by $\times 20$), and distribution of the distortion over different frequency regions.



(a) ResNet50 (b) ViT-B

Figure 6: Results of the phase attack when the perturbation is restricted to reside only in the low or high frequency band. The case without restriction is also shown as 'whole.'



(a) ResNet50 (b) ViT-B

Figure 7: Classification accuracy with respect to the ratio of reduction in the magnitude or phase spectrum.

4.3 Analysis

We further investigate the particular vulnerability of the Transformers to the attacks in the spectral domain.

4.3.1 Frequency Analysis of Perturbations

We examine the distortion due to the attack in the spectral viewpoint. For this, the Fourier transform is applied to the difference between the attacked image and the original image, and the magnitudes in different frequency regions are obtained, where the frequency regions are defined as shown in Figure 5b. Figures 5c to 5g show the attacked image, distortion in the pixel domain, and magnitude distribution under the phase attack. The averaged results over images are shown in Appendix. In the case of ResNet50, the distortion is concentrated on the high frequency regions, whereas low frequency regions are mainly distorted in the other models. Since CNNs and Transformers rely more on high and low frequency information, respectively (Park & Kim, 2022; Benz et al., 2021; Wang et al., 2020; Jo & Bengio, 2017), the attack effectively injects perturbations in such vulnerable frequency regions. Consequently, the distortion pattern significantly differs according to those properties. Unlike ResNet50, the distortion in BiT is extremely concentrated in the low frequency regions, which explains why BiT is the most vulnerable under the phase attack in Figure 4b.

4.3.2 Frequency-Restricted Attacks

We examine the case where the perturbation is applied to only a limited frequency band. Figure 6 compares the phase attack when the phase perturbation is only in the low frequency band (regions 1 and 2 in Figure 5b) or the high frequency band (region 10 in Figure 5b). For ResNet50, it is more effective to perturb only the high frequency band because high frequency perturbations for which ResNet50 becomes particularly vulnerable can be optimized more easily by the restriction. However, for ViT-B, perturbing only the high frequency band is the least effective and the case without restriction implements the strongest attack, because low-intermediate frequency regions need to be perturbed for an effective attack as shown in Figure 5e.

4.4 DEPENDENCE ON MAGNITUDE AND PHASE

In most results above, the magnitude attack appears to be weaker than the phase attack. We investigate this phenomenon further.

4.4.1 SENSITIVITY TO REDUCED MAGNITUDE AND PHASE

We design an experiment where the magnitude or phase spectrum is reduced gradually, i.e., $M' = M \times (1 - r)$ or $\phi' = \phi \times (1 - r)$, where $r \in \{0, 0.1, ..., 0.9, 1\}$, and the classification accuracy is measured. Note that no attack is applied in this experiment. The results are shown in Figure 7. Both ResNet50 and ViT-B are more sensitive to phase reduction than magnitude reduction, and in particular, ViT-B achieves 85.6% of accuracy even though the magnitude is reduced by 90%. These imply that the corruption of the magnitude spectrum is less critical than that of the phase spectrum, which explains the severer vulnerability of the models to the phase attack.

4.4.2 MAGNITUDE-PHASE RECOMBINATION

Inspired by Chen et al. (2021), we conduct an experiment where the magnitude component of one image and the phase component of another image are recombined in the frequency domain and the classification result of this recombined image (after inverse Fourier transform) from a model is tested. For all possible image pairs, Table 1 shows the proportion of the images that are classified as the class of the magnitude image or phase image, or none of the two classes. For ResNet50, most of the cases (97.85%) do not follow either the classes of the magnitude or the phase, while for the rest, more images follow the phase classes (2.04%) than the magnitude classes (0.11%). For ViT-B, however, a considerable amount of recombined images (33.92%) are classified as the class of the phase images, while only 0.26% of the images follow the classes of the magnitude images. These results highlight the relative importance of the phase information, providing an explanation on the particular vulnerability of the Transformers to the phase attack and the higher strength of the phase attack than the magnitude attack for the Transformers.

Table 1: Proportions of magnitude-phase-recombined images that are classified to the classes of the magnitude or phase images, or some other classes.

Model	Phase (%)	Magnitude (%)	Else (%)
ResNet50	2.04	0.11	97.85
ViT-B	33.92	0.26	65.82

5 CONCLUSION

We comparatively investigated adversarial robustness of CNNs and Transformers using the unified attack framework with consideration of the relationship between the image quality and ASR. Our major findings can be summarized as follows.

- Different from the conclusions of the existing studies, we observed that Transformers can become more vulnerable than CNNs. The Transformers are more robust than the ResNets under the pixel attack, but more vulnerable under the magnitude and phase attacks. However, when the attack strength is low and the image quality becomes high, the ResNets tend to become more vulnerable than the Transformers.
- Some models including DeiT and BiT sometimes do not follow the above trend. DeiT trained by distillation from CNN behaves more similarly to ResNets. BiT appears to be more vulnerable than the other models even under the magnitude and phase attacks.
- The ResNets are sensitive to the perturbation in high frequency regions, which can be effectively injected by the pixel attack. On the other hand, the Transformers are sensitive to the perturbation in low frequency regions, which can be optimized in the frequency domain by the phase attack.
- The phase information plays more important role in classification for both ResNets and Transformers than the magnitude information, and reliance on the phase information is more prominent in Transformers.

REFERENCES

Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 6:14410–14430, 2018.

Ahmed Aldahdooh, Wassim Hamidouche, and Olivier Deforges. Reveal of vision transformers robustness against adversarial attacks. *arXiv preprint arXiv:2106.03734*, 2021.

Alaaeldin Ali, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, et al. XCIT: Cross-covariance image transformers. In *Advances in Neural Information Processing Systems*, volume 34, pp. 20014–20027, 2021.

Shahrukh Athar and Zhou Wang. A comprehensive performance evaluation of image quality assessment algorithms. *IEEE Access*, 7:140030–140070, 2019.

Yutong Bai, Jieru Mei, Alan L Yuille, and Cihang Xie. Are transformers more robust than CNNs? In *Advances in Neural Information Processing Systems*, volume 34, pp. 26831–26843, 2021.

Philipp Benz, Soomin Ham, Chaoning Zhang, Adil Karjauv, and In So Kweon. Adversarial robustness comparison of vision transformer and MLP-mixer to CNNs. In *Proceedings of the 32nd British Machine Vision Conference*, 2021.

Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10231–10241, 2021.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *Proceedings of the IEEE Symposium on Security and Privacy*, pp. 39–57, 2017.

Guangyao Chen, Peixi Peng, Li Ma, Jia Li, Lin Du, and Yonghong Tian. Amplitude-phase recombination: Rethinking robustness of convolutional neural networks in frequency domain. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 458–467, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: transformers for image recognition at scale. In *Proceedings of the 9th International Conference on Learning Representations*, 2021.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *Proceedings of the 3rd International Conference on Learning Representations*, 2015.

Chuan Guo, Jared S Frank, and Kilian Q Weinberger. Low frequency adversarial perturbation. In *Proceedings of the Uncertainty in Artificial Intelligence*, pp. 1127–1137, 2020.

Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. In *Advances in Neural Information Processing Systems*, volume 34, pp. 15908–15919, 2021.

Jason Jo and Yoshua Bengio. Measuring the tendency of CNNs to learn surface statistical regularities. *arXiv preprint arXiv:1711.11561*, 2017.

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

Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (BiT): General visual representation learning. In *Proceedings of the European Conference on Computer Vision*, pp. 491–507, 2020.

Xiu-Chuan Li, Xu-Yao Zhang, Fei Yin, and Cheng-Lin Liu. F-mixup: Attack CNNs from Fourier perspective. In *Proceedings of the 25th International Conference on Pattern Recognition*, pp. 541–548, 2020.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.

Kaleel Mahmood, Rigel Mahmood, and Marten Van Dijk. On the robustness of vision transformers to adversarial examples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7838–7847, 2021.

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.

Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. In *Advances in Neural Information Processing Systems*, volume 34, pp. 23296–23308, 2021.

Namuk Park and Songkuk Kim. How do vision transformers work? In *Proceedings of the 10th International Conference on Learning Representations*, 2022.

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.

Sayak Paul and Pin-Yu Chen. Vision transformers are robust learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 2071–2081, 2022.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115 (3):211–252, 2015.

Rulin Shao, Zhouxing Shi, Jinfeng Yi, Pin-Yu Chen, and Cho-Jui Hsieh. On the adversarial robustness of vision transformers. *arXiv preprint arXiv:2103.15670*, 2021.

Yash Sharma, Gavin Weiguang Ding, and Marcus A Brubaker. On the effectiveness of low frequency perturbations. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pp. 3389–3396, 2019.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 10347–10357, 2021.

Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8684–8694, 2020.

Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 568–578, 2021.

Zerui Wen. Fourier attack–a more efficient adversarial attack method. In *Proceedings of the 6th International Conference on Control Engineering and Artificial Intelligence*, pp. 125–130, 2022.

Ross Wightman. NIPS 2017 adversarial competition (pytorch). https://github.com/rwightman/pytorch-nips2017-adversarial, 2017.

Ross Wightman. PyTorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token ViT: Training vision transformers from scratch on ImageNet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 558–567, 2021.

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 conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

A EXPERIMENTAL ENVIRONMENTS

We run our experiments on an NVIDIA GeForce RTX 3090 GPU using CUDA 11.2. The software modules are listed in the following table.

Table 2: Module environments used our experiments

Module	Version
torch	1.11.0
torchvision	1.12.0
opencv-python	4.5.4.60
timm	0.5.4
piq	0.7.0
numpy	1.21.2
Pillow	8.4.0
seaborn	0.11.2
pandas	1.3.5

B EXPERIMENTAL SETUP FOR FGSM AND PGD

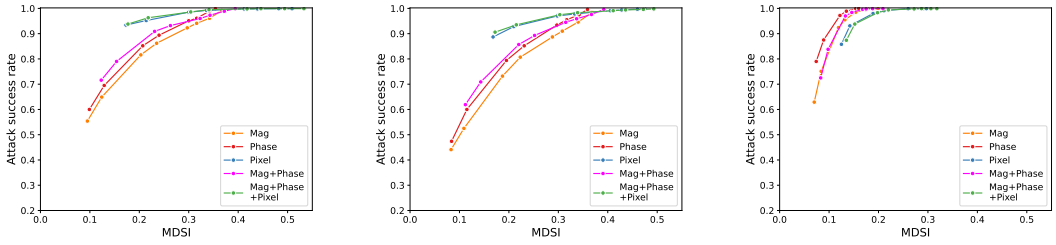
For both FGSM and PGD, the amount of perturbation is limited by the L_∞ norm $\epsilon \in \{0.1/255, 0.5/255, 1/255, 4/255, 8/255\}$. For PGD, the step size at each iteration is set to ϵ/T, where T is the number of iterations ($T = 50$).

C RESULTS USING OTHER IMAGE QUALITY METRICS

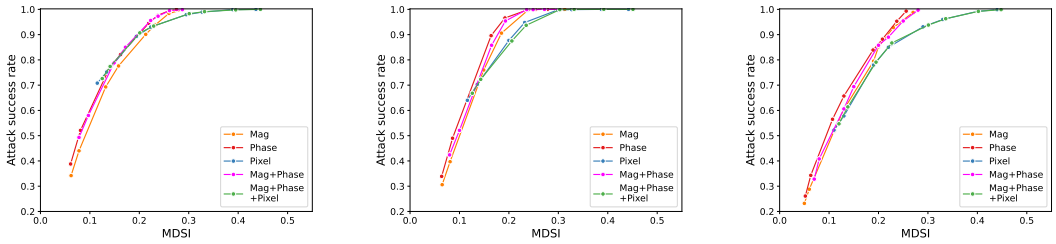
While we use PSNR as the image quality metric in the main paper, results using other perceptual image quality metrics are shown here. We choose the mean deviation similarity index (MDSI) (Nafchi et al., 2016), which was shown to perform the best in Kastryulin et al. (2019), and learned perceptual image patch similarity (LPIPS) (Zhang et al., 2018), which uses the features from a pre-trained network for classification. The results are shown in Figures 8-11. 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.

D AVERAGE DISTRIBUTION OF DISTORTION

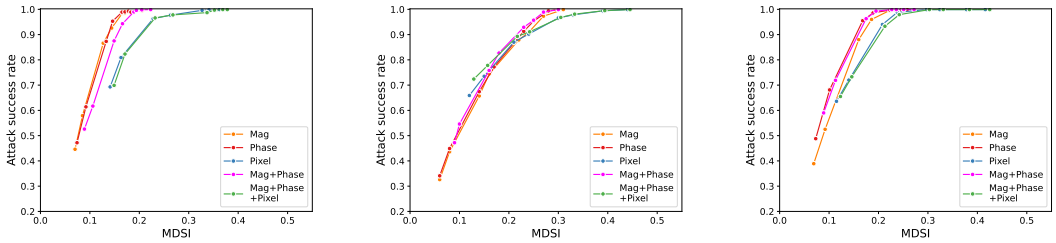
The distributions of the distortion by the phase attack over different frequency regions, averaged over all images, are shown in Figure 12. The trends are the same to those observed in Figure 5 of the main paper.



(a) ResNet50 (b) ResNet152 (c) BiT

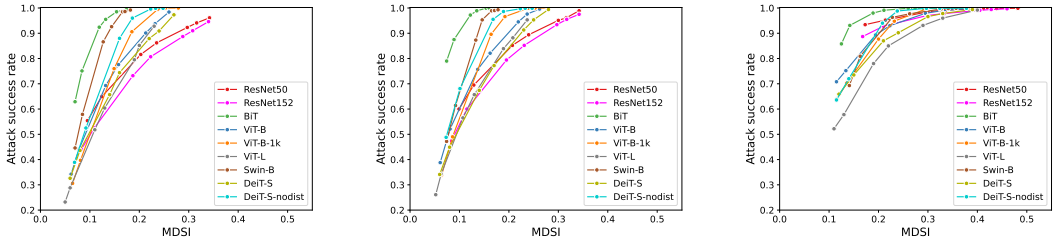


(d) ViT-B (e) ViT-B-1k (f) ViT-L



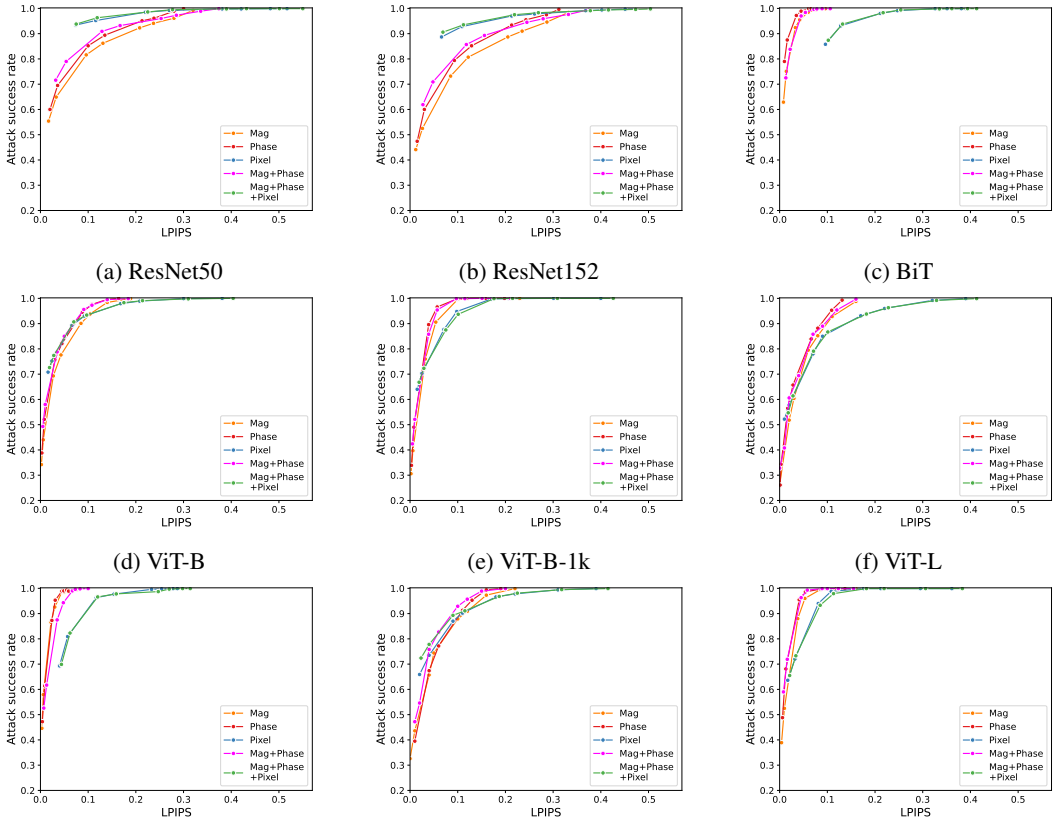
(g) Swin-B (h) DeiT-S (i) DeiT-S with no distillation

Figure 8: Comparison of different attacks for each model using MDSI.



(a) Magnitude attack (b) Phase attack (c) Pixel attack

Figure 9: Comparison of different models for each attack type using MDSI.

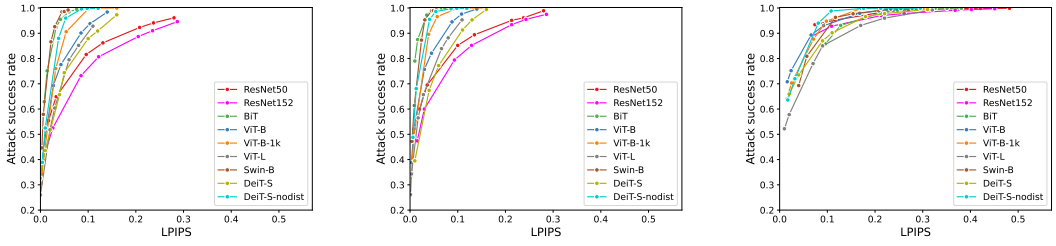


(a) ResNet50 (b) ResNet152 (c) BiT

(d) ViT-B (e) ViT-B-1k (f) ViT-L

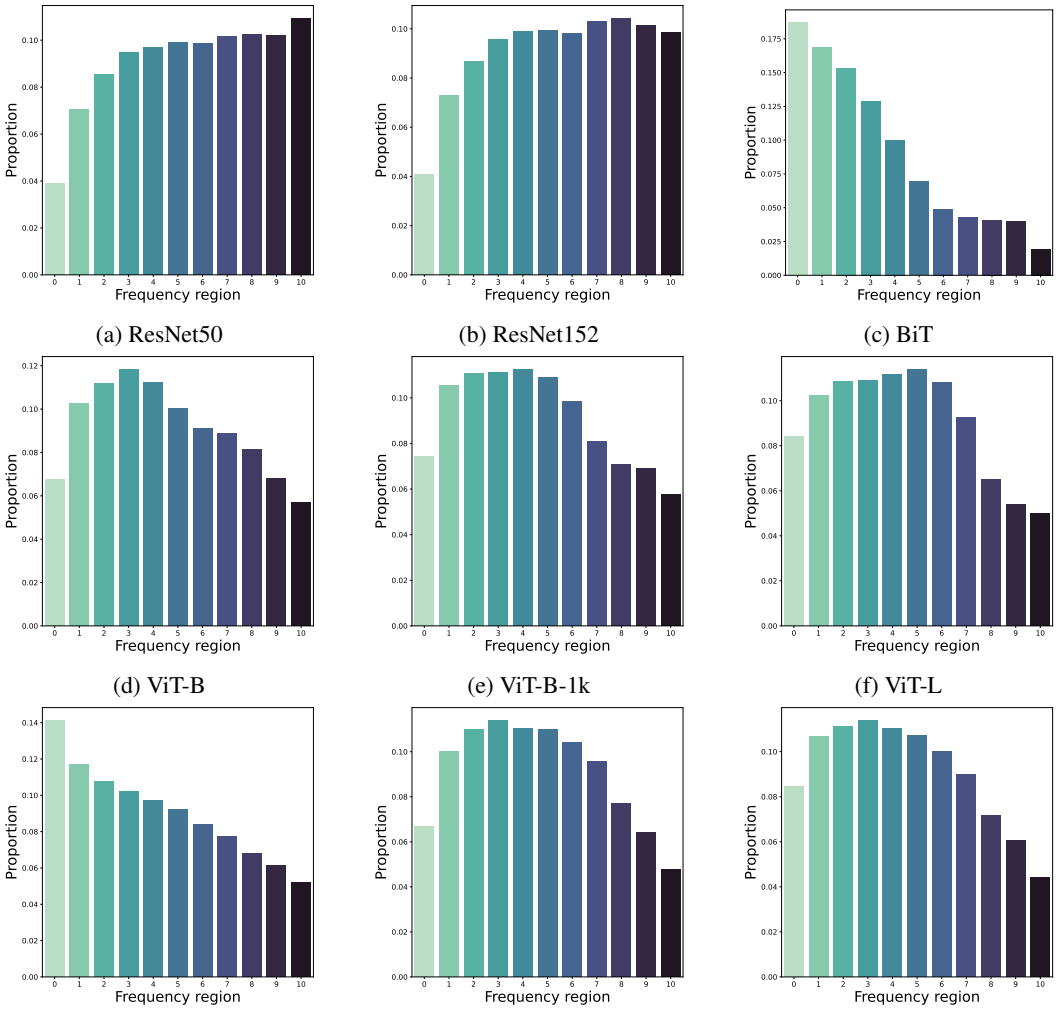
(g) Swin-B (h) DeiT-S (i) DeiT-S with no distillation

Figure 10: Comparison of different attacks for each model using LPIPS.



(a) Magnitude attack (b) Phase attack (c) Pixel attack

Figure 11: Comparison of different models for each attack type using LPIPS.



(a) ResNet50 (b) ResNet152 (c) BiT

(d) ViT-B (e) ViT-B-1k (f) ViT-L

(g) Swin-B (h) DeiT-S (i) DeiT-S with no distillation

Figure 12: Average distributions of distortion over different frequency regions.