On the unreasonable vulnerability of transformers for image restoration – and an easy fix

Shashank Agnihotri[1], Kanchana Vaishnavi Gandikota[1], Julia Grabinski[1,2], Paramanand Chandramouli[1], and Margret Keuper[1,3]

[1]Institute for Vision and Graphics, University of Siegen

[2] Fraunhofer ITWM, Kaiserslautern and IMLA, University of Offenburg

[3]Max-Planck-Institute for Informatics, Saarland Informatics Campus

Abstract

Following their success in visual recognition tasks, Vision Transformers(ViTs) are being increasingly employed for image restoration. As a few recent works claim that ViTs for image classification also have better robustness properties, we investigate whether the improved adversarial robustness of ViTs extends to image restoration. We consider the recently proposed Restormer model, as well as NAFNet and the "Baseline network" which are both simplified versions of a Restormer. We use Projected Gradient Descent (PGD) and CosPGD for our robustness evaluation. Our experiments are performed on real-world images from the GoPro dataset for image deblurring. Our analysis indicates that contrary to as advocated by ViTs in image classification works, these models are highly susceptible to adversarial attacks. We attempt to find an easy fix and improve their robustness through adversarial training. While this yields a significant increase in robustness for Restormer, results on other networks are less promising. Interestingly, we find that the design choices in NAFNet and Baselines, which were based on iid *performance, and not on robust generalization, seem to be at odds with the model robustness.*

1. Introduction

The goal of image restoration is to recover high-quality images from degraded observations. The degradation could be due to a variety of factors such as noise, blur, artifacts due to jpeg compression, raindrops, haze, and other factors. Earlier methods for image restoration [39, 6, 11, 14, 3] employed carefully chosen priors and degradation models to derive degradation-specific restoration algorithms. Yet, such methods are limited by the strength of the image prior and the accuracy in modeling or estimating the degradation operator. The past decade saw a large-scale adoption of deep learning methods to image restoration [45], which outperformed the classical approaches[31]. Recent approaches [62, 55, 50] successfully adopt novel architectures such as Transformers [52, 16] and MLP-mixers [48] for restoration.

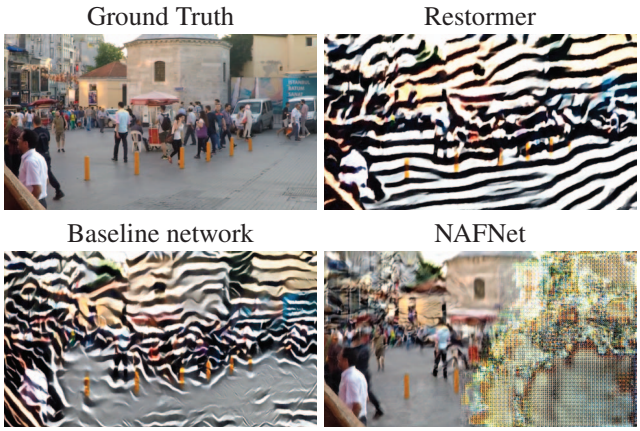

Figure 1. Comparing images reconstructed by all considered models after 5 iterations of CosPGD attack . We observe strong spectral artifacts in the reconstructed images.

Yet, CNNs, MLP-mixers as well as Transformer have been shown to be vulnerable to carefully crafted adversarial examples [32, 18]. Recent work [1, 9, 58, 17] also confirms the existence of such vulnerabilities in deep learning-based image restoration. Yet, existing works mainly analyze the robustness of CNN-based restoration methods. Conversely, with the introduction of novel network architectures such as vision Transformers [29, 16], MLP mixers [48], and improved convolutional architectures [30, 26] which outperform the earlier networks such as ResNets [21], there have been several studies on the robustness of these new architectures [5, 42, 47, 13, 1]. To the best of our knowledge, very limited works [13, 4] investigate the effect of architectural components and training recipes. Existing works focus on image classification and do not study restoration.

Thus to bridge this gap, in this work, we study the ad-

We acknowledge support by the DFG research unit 5336 - Learning to Sense.

versarial robustness of Transformer based restoration networks. These networks include, Restormer [62], and two architectures introduced in [7] the *Baseline network* and the *Non-linear Activation free Network (NAFNet)*, both obtained by simplifying the original Restormer, with modifications to the channel attention and activation functions. Further, to better understand the architectural design choices made by [7], we include an *Intermediate network* also considered by [7] which serves as a step between the Baseline network and NAFNet. This study is particularly interesting as recent works [57, 4] indicate that the choice of activation function significantly impacts adversarial robustness. We study the network robustness under standard and adversarial attacks, by considering ℓ_∞ perturbations crafted using PGD attack [32] and CosPGD attack proposed in [1] for dense prediction tasks. We conduct our experiments on dynamic deblurring using the Go-Pro dataset [35].

Our experiments reveal that under standard training settings, Transformer based restoration networks are not robust to adversarial attacks in general. As shown in Figure 1, the networks also exhibit distinct artifacts in the reconstructions under attack. The images from the Baseline network and the Restormer exhibit severe ringing artifacts [34], whereas the NAFNet reconstructs images with very strong grid and color artifacts under adversarial attack. We find that adversarial training can largely reduce the artifacts and significantly improve the robustness of all three networks. However, the recently proposed NAFNet and Baseline network fail to rival the performance of Restormer, which leads us to contemplate the importance of the architectural components necessary to achieve robust generalization.

The main contributions of this work can be summarized as follows:

- We investigate the robustness of recently proposed Transformer based architectures for image restoration, namely image deblurring.
- We analyze the quality of the restored images and the spectral artifacts introduced by models under the aforementioned adversarial attacks.
- We understand the effects of defense strategy against adversarial attacks that consequently reduce the spectral artifacts in reconstructed images.
- Lastly, we study the effect of certain architectural design choices in the recently proposed *state-of-the-art* image restoration model, NAFNet, to improve their robustness.

2. Related Work

Transformers for Image Restoration The past decade saw significant improvements in image restoration, largely owing to the adoption of deep networks trained on large datasets of clean and degraded images. While earlier restoration networks largely adopted CNN-based architectures, subsequent works also explored the use of attention mechanisms inside CNNs [63, 36, 46]. We refer [45] for a detailed survey on deep learning approaches to restoration. More recently, vision Transformers [29, 16] are increasingly adopted for several restoration tasks. While [27, 62, 55, 12, 56] adopt Transformers for generic restoration tasks, a few works focus on specific restoration tasks by including such as deblurring [49], deraining [28], dehazing [20, 44], removing degradations due to adverse weather conditions [51]. These networks typically employ encoder-decoder-based architectures with Transformer blocks combined with convolutions.

Adversarial Robustness of Image Restoration. While the adversarial robustness of deep networks for image recognition is extensively studied, a few works also study the robustness of image restoration networks to adversarial attacks. [9, 10, 61] evaluate adversarial robustness of deep learning-based image super-resolution. While [10] propose adversarial regularization, [61] propose frequency domain adversarial example detection, combined with random frequency masking to improve robustness. [17] evaluate adversarial robustness of deblurring networks with and without the knowledge of the blur operator, and introduce targeted attacks on restoration. In [8], the adversarial robustness of image-to-image translation models is studied, including restoration tasks, and adversarial training and different transformation-based defenses are evaluated. Yan et al. [58] investigate the robustness of image denoising to zero-mean adversarial perturbations and propose training with clean and adversarial samples to improve robustness. Yu et al. [60] investigate adversarial robustness of deep learning-based rain removal, and study the effect of architecture and training choices on robustness. Yet, these works do not focus on the more recent Transformer based restoration networks. With the notable exception of [1], where they simply benchmark the adversarial performance of the image restoration networks recently proposed by [7].

Robustness of Transformers & other modern architectures. Recently, Vision Transformers (ViTs) [16, 29] have been successfully applied to image recognition, outperforming the older ResNets. Follow-up works modified training schemes and architectures leading to better performing CNN architectures such as ConvNext [30], and hybrid models combining components of ViTs and CNNs [2]. Following the introduction of these novel architectures, several works examined the robustness properties of these models. [43, 5, 42, 37] suggest Transformers have better adversarial robustness than CNNs. However, [33] shows

that vision Transformers are also as vulnerable as CNNs under strong attacks. [4] show that CNNs can achieve similar adversarial robustness as Transformers when trained using similar training recipes, yet Transformers still outperform CNNs on out-of-distribution generalization. [47] benchmark for robustness dependent on the network architecture. They find that Transformers are best suited against adversarial attacks while being extremely vulnerable to common corruptions [22] and system noise. Conversely, CNNs are more robust against common corruptions and system noise while being weakest against adversarial attacks. Further, they show that MLP-Mixers are not the best and also not the worst for both cases.

In their work, [38] benchmark the robustness of state-of-the-art Transformers and CNN architectures and show that CNNs using ConvNext architecture can be at least as robust as Transformers for image recognition. Meanwhile [13] analyzes the effect of different architectural components such as patches, convolution, activation, and attention, and demonstrates that ConvNexts have better adversarial robustness than ResNets. [57] observe that smooth activation functions improve adversarial training as they enable better gradient updates to compute harder adversarial examples. Subsequent works [4, 13] also confirm improvement in robustness when GELU [23] activation functions are used in adversarial training. While [4] attribute significant robustness gains in Transformers to the self-attention mechanism, [53] identify other architectural components, including, the use of patches, larger kernels, reducing activation and normalization layers which when incorporated into CNNs lead to out of distribution robustness at least on par with Transformers without the use of attention.

In contrast, our work focuses on the investigation of the robustness of several recent Transformer based restoration models and shows interesting effects of adversarial attacks that can be attributed to different building modules of such models.

3. Methodology

Following, we describe the attack framework used and the defense strategy used to combat the vulnerabilities of the architectures exposed by the adversarial attacks.

3.1. Attack Framework

Let $\mathbf{x}$ denote the ground-truth image, which is corrupted by a possibly non-linear degradation operator $\mathbf{A}$, resulting in an observation $\mathbf{y}^{\text{clean}}$, which can be expressed as

$$\mathbf{y}^{\text{clean}} = \mathbf{A}(\mathbf{x}). \tag{1}$$

Let $\mathcal{G}_\theta$ be a (Transformer-based) neural network parameterized by θ trained to recover $\mathbf{x}$ from $\mathbf{y}^{\text{clean}}$. In this work, we are interested in studying the stability of $\mathcal{G}_\theta$ to adversarial attacks that aim to degrade its performance through visually imperceptible changes to the inputs [18, 32]. We evaluate the robustness against attacks using additive perturbations δ with ℓ_p-norm constraints. We generate the adversarial perturbations based on two powerful attack methods CosPGD [1] developed for dense prediction tasks, and PGD attack [32], both of which we detail in the following. The objective of the attack is to maximize the deviation of the network output from the ground truth as measured by a loss function L, subject to ℓ_p norm constraints on the perturbation:

$$\underset{\delta}{\text{maximize}}\ L(\mathcal{G}_\theta(\mathbf{y}^{\text{clean}} + \delta),\ \mathbf{x})\ \text{ s.t. } \|\delta\|_p \leq \epsilon. \tag{2}$$

Please refer to Section A for details of the attack formulations.

3.2. Architectures: from Restormer to NAFNet

We evaluate the adversarial robustness of *Restormer* [62], a Transformer based architecture for image restoration and two architectures introduced in [7] by modifying the Restormer architecture. Restormer [62] has a UNet [40] like encoder-decoder architecture, using multi-head channel-wise attention modules, gated linear units [15] and depth-wise convolutions in the feed-forward network. This network achieved state-of-the-art performance in image restoration at the time of its publication. The authors in [7] investigate whether it is possible to retain the performance of Restormer, with a simplified architecture. After a thorough ablation study, they propose a simplified *Baseline* network that improved upon the *SotA* performance. The Baseline network utilizes GELU activations [23] and replaces multi-headed self-attention in [62] with a channel attention module [25]. Without loss in i.i.d. performance, they further simplify this architecture by removing activation functions altogether, replacing GELU with a *simple gate* which performs element-wise product of feature maps, and replacing the channel attention by a *simplified channel attention* without activation functions. The resulting network is referred to as a Nonlinear Activation-Free Network (NAFNet). In contrast to [7] who focus on performance with clean inputs, we analyze the adversarial robustness of these networks, which also allows us to evaluate the effect of different activation functions and attention mechanisms on the robustness of restoration transformers. In Figure 1, we observe that NAFNet has significantly different artifacts in the reconstructed images compared to Restormer and the Baseline network. One might simply hypothesize that these strange artifacts which appear to be the cumulative effect of aliasing and color mixing are due to the use of 'Simple Gate' in place of a non-linear activation function like GELU. To confirm this hypothesis we additionally consider an *Intermediate network*, from [7]. In this *Intermediate network* we replace the *channel attention* in the baseline network with the

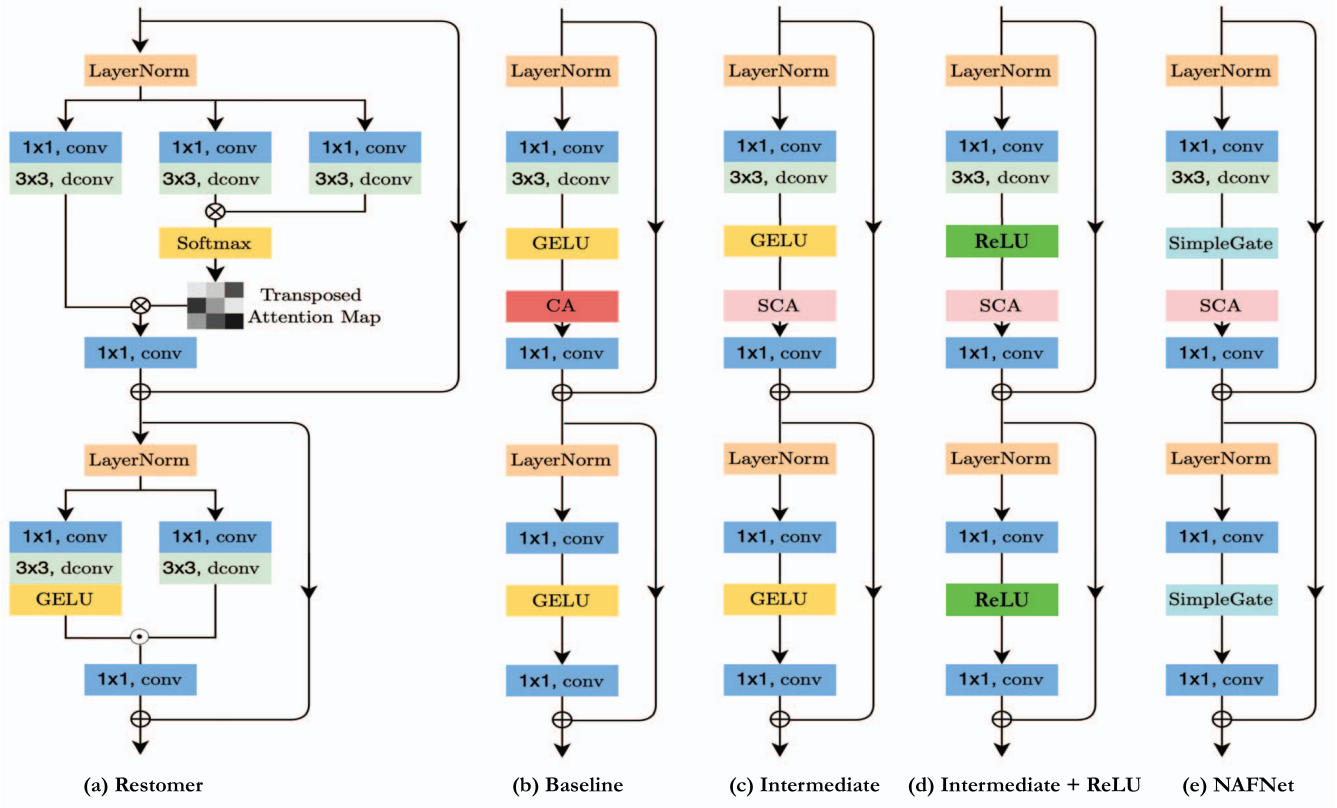

Figure 2. Modified visualization of repeating blocks of the architectures from [7] including the considered *Intermediate network* from [7] (please refer to (c)) and *Intermediate + ReLU network* (please refer to (d)).

simplified channel attention but retain the GELU activation. Additionally, to better understand the role of non-linear activation functions in this context, we consider an architecture the same as the *Intermediate network* but with ReLU activations instead of GELU. In Figure 2, we modify the visualization by [7], to present the repeating blocks of all the considered architectures in our work.

3.3. Defenses

As discussed in Section 1, we observe in Figure 1 that all considered architectures are vulnerable to adversarial attacks. Prior work [18, 32, 19] has shown that adversarial training is an effective defense against adversarial attacks. Thus we use adversarial training as a defense strategy.

Adversarial Training. We use the FGSM attack as proposed by [18] to generate adversarial samples during training. Adversarial training can be hypothesized as a min-max problem, where we try to find perturbations for the samples such that the loss is maximized while training the network on these samples to minimize the loss of the model over training iterations. PGD attack is essentially a multi-step extension of FGSM attack, and thus the loss that FGSM attack attempts to maximize remains the same. Additionally, the attack step of FGSM is also the same as described in Section 3.1, with one notable difference being that in the case of an FGSM attack, the attack step size α is equal to the permissible perturbation size of ϵ.

While training, to avoid over-fitting to adversarial samples, and enable the model to make reasonable reconstructions on unperturbed samples we use the training regime similar to [19] and use only 50% of the sample in the training batch to generate perturbed adversarial samples and use the other 50% samples unperturbed. Thus, the effective learning objective is as described by Equation 3.

$$\underset{\theta}{\text{minimize}} \sum_{i} L(\mathcal{G}_\theta(\mathbf{y}^{\text{clean}_i}),\ \mathbf{x}_i) + \sum_{j} L(\mathcal{G}_\theta(\mathbf{y}^{\text{adv}_j}),\ \mathbf{x}_j) \tag{3}$$

where the indices i and j correspond to the examples from the clean and adversarial batch splits, and FGSM adversarial examples are generated as:

$$\mathbf{y}^{\text{adv}_j} = \phi^r(\mathbf{y}^{\text{clean}_j} + \phi^\epsilon(\epsilon \cdot \text{sign}\nabla_{\mathbf{y}_j} L(\mathcal{G}_\theta(\mathbf{y}_j), \mathbf{x}_j))) \tag{4}$$

4. Experiments

In this work on image restoration, we focus on reconstructing deblurred images using a few recently proposed image restoration networks.

4.1. Experimental Setup

Networks. We consider Restormer proposed by [62], and Baseline network and NAFNet proposed by [7] with width 32. For understanding the design choices that lead to NAFNet producing reconstructed images with significantly different spectral artifacts than the other considered networks, we also consider an *Intermediate network* and *Intermediate + ReLU*. This *Intermediate network* with width 32 has also been considered by [7] when discussing design choices to arrive from the Baseline network to NAFNet. These networks are similar to the Baseline, except it has the "simplified channel attention" as proposed by [7] rather than the "channel attention" used in the Baseline network. We visualize all the considered architectures in Figure 2.

Dataset. For our experiments we use the GoPro image deblurring dataset[35]. This dataset consists of 3 214 real-world images with realistic blur and their corresponding ground truth (deblurred images) captured using a high-speed camera. The dataset is split into 2 103 training images and 1 111 test images.

Metrics. We report the PSNR and SSIM scores of the reconstructed images w.r.t. to the ground truth images, averaged over all images. PSNR stands for Peak Signal-to-Noise ratio, a higher PSNR indicates a better quality image or an image closer to the image to which it is being compared. SSIM stands for Structural similarity[54]. A higher SSIM score corresponds to better higher similarity between the reconstruction and the ground-truth image.

Training Regimes. For Restormer and its adversarial training counterpart ('+ADV') we follow the training procedure used by [62] except due to computational limitations we do not train on the last recommended patch size 384. For the Baseline network, NAFNet, and its counterparts we follow the training regime used by [7].

Adversarial Training. We used FGSM [18] adversarial training for efficiency. The maximum allowed perturbation for the adversaries is set to $\epsilon = \frac{8}{255}$. We use '+ADV' after the model name to denote that the model has been trained with FGSM adversarial training.

Adversarial Attacks. We consider PGD and CosPGD attacks. Following the procedure by [1], we use $\epsilon \approx \frac{8}{255}$, α(attack step size)$= 0.01$. We consider attack iterations $\in \{5, 10, 20\}$ for our attacks. We use MSE loss for generating adversarial samples for all networks.

4.2. Results

The good performance of image restoration models on unperturbed samples is indubitably essential for possible real-world applications. However, the generalization ability of these models to perturbed samples has to be better understood for their reliability in safety-critical applications such as medical imaging, autonomous driving, etc. To this effect, we study the performance of the considered networks on both clean (unperturbed) and adversarial (perturbed) samples. Further, to overcome the observed shortcomings of these models, we harden them using adversarial training.

Table 1. Performance of the different considered networks and their counterparts on clean (unperturbed) GoPro test images. While NAFNet has highest PSNR value, Restormer is slightly better in terms of SSIM. All models slightly suffer from adversarial training when evaluated on clean data, which is to be expected.

Architecture	PSNR	SSIM
Restormer	31.99	**0.9635**
+ ADV	30.25	0.9453
Baseline	32.48	0.9575
+ ADV	30.37	0.9355
NAFNet	**32.87**	0.9606
+ ADV	29.91	0.9291

As observed in Figure 1, under adversarial attack both Restormer and Baseline network induce ringing-like artifacts in the restored images. However, NAFNet introduces aliasing like grid artifacts and color mixing in the restored images. We report the performance of three networks along with adversarial training over clean images in Table 1. Further, to study the generalization ability of these networks we adversarially attack the networks and report the findings in Table 2.

With standard training protocol, Restormer is marginally more robust in comparison to the Baseline network with fewer attack iterations, however, this difference reduces as the number of attack iterations increases. With adversarial training using FGSM adversarial examples, we observe improvement in the robustness of all three networks. Interestingly, the gain in performance of Restormer when trained with FGSM is significantly better than that of the Baseline network and NAFNet. This indicates that Restormer has a much higher potential of being generalizable than both the Baseline network and NAFNet. This raises doubts over the claims by [7] regarding the Baseline network and NAFNet having "comparable or better performance" to the recent *state-of-the-art* image restoration models. Their claim holds true for clean samples, however with just slight perturbation ($\epsilon = \frac{8}{255}$), the performance of their proposed models drops significantly. Contrary to this, *Intermediate+ReLU* is significantly more robust, across attack iterations. We discuss this further in Section 5.1.

At first, one might overlook this shortcoming, however, when considering safety-critical real-world applications like those in the medical domain for deblurring MRI images, or in autonomous driving, such shortcomings could be very hazardous. This is further highlighted in Figure 3 as we observe that both the Restormer and the Baseline network introduce ringing artifacts in the reconstructed images, however, NAFNet introduces very strong aliasing and

Table 2. Comparison of performance of the considered models against CosPGD and PGD attacks with various attack strengths. Attack strength increases with the number of attack iterations (itrs). Note that *Intermediate + ReLU* achieves reasonably robust results entirely without adversarial training. Please refer to Table A1for further results.

Architecture	CosPGD						PGD					
	5 attack itrs		10 attack itrs		20 attack itrs		5 attack itrs		10 attack itrs		20 attack itrs	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Restormer	11.36	0.3236	9.05	0.2242	7.59	0.1548	11.41	0.3256	9.04	0.2234	7.58	0.1543
+ ADV	**24.49**	0.81	**23.48**	**0.78**	21.58	0.7317	**24.5**	0.8079	**23.5**	**0.7815**	21.58	0.7315
Baseline	10.15	0.2745	8.71	0.2095	7.85	0.1685	10.15	0.2745	8.71	0.2094	7.85	0.1693
+ ADV	15.47	0.5216	13.75	0.4593	12.25	0.4032	15.47	0.5215	13.75	0.4592	12.24	0.4026
NAFNet	8.67	0.2264	6.68	0.1127	5.81	0.0617	10.27	0.3179	8.66	0.2282	5.95	0.0714
+ ADV	17.33	0.6046	14.68	0.509	12.30	0.4046	15.76	0.5228	13.91	0.4445	12.73	0.3859
Intermediate	6.0224	0.0509	5.8166	0.0366	5.7199	0.0315	6.0225	0.0509	5.8158	0.0365	5.7173	0.0314
+ ADV	24.02	**0.8213**	22.01	0.7775	20.15	0.7286	24.02	**0.8213**	21.98	0.7770	20.15	0.7286
Intermediate + ReLU	13.87	0.4093	11.63	0.3128	10.29	0.2538	13.87	0.4094	11.62	0.3127	10.29	0.2542
+ ADV	23.90	0.8046	22.46	0.7637	**21.85**	**0.7484**	23.91	0.8046	22.47	0.7638	**21.84**	**0.7481**

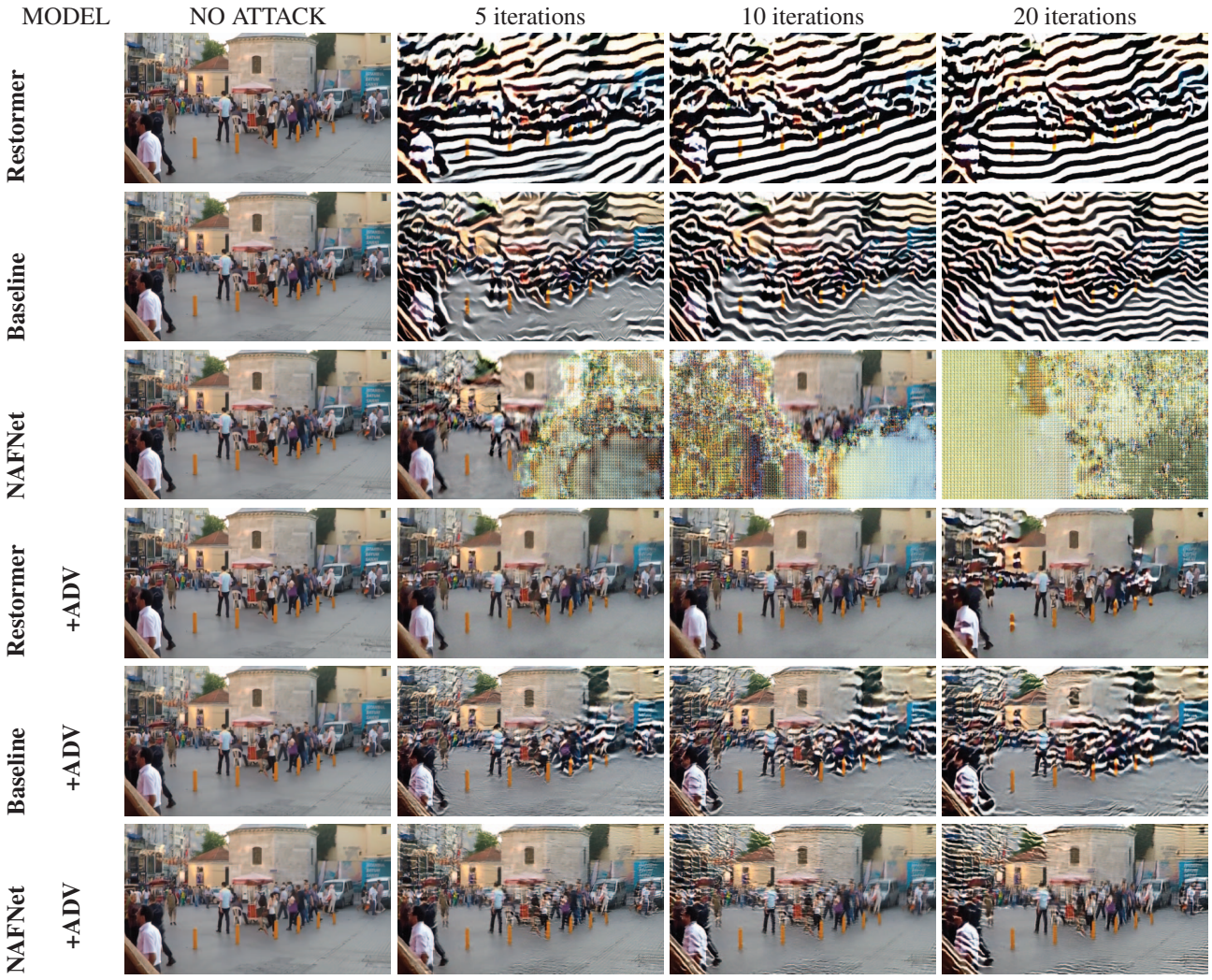

Figure 3. Images reconstructed by different models after **CosPGD attack**. See Figure A1 (Appendix B) to compare over all considered models.

color mixing that gets worse as the attack strength increases. While aliasing and color artifacts are significantly reduced with adversarial training (please refer to Figure 3), the reconstructions of NAFNet and the Baseline network are still affected by residual ringing artifacts. Interestingly, the quality of images reconstructed by Restormer after adversarial training is significantly better, as indicated by its performance in terms of PSNR and SSIM in Table 2. At a low amount of adversarial attack iterations, the artifacts present in the images reconstructed by Restormer are negligible. To ascertain that these observations are not specific to the adversarial attack itself, we visualize the images reconstructed after the PGD attack in Figure A2 and observe a similar phenomenon. This accentuates the strength of the architectural design of Restormer and casts doubts over that of the networks proposed by [7].

5. Analysis and Discussion

Following we discuss the design choices made in NAFNet and the Baseline network that constrain the performance of the network against adversarial attacks, despite employing adequate defense techniques.

5.1. Analyzing Intermediate networks

First, we study the *Intermediate network* to ascertain if the spectral artifacts introduced by NAFNet in its reconstructed images were due to replacing a non-linear activation function with a *Simple Gate*. This is because the channel-wise multiplication would best explain the color mixing artifact and the inherent wrong sub-sampling during this operation and would account for the accentuated aliasing artifacts. Further to understand the influence of the non-linear activation, we also train the Intermediate network with ReLU activation, referred to as *Intermediate + ReLU*.

We report the findings on the Intermediate networks in Table A1. Here we observe that the Intermediate network performs marginally worse than even NAFNet, especially under adversarial attacks. Additionally, in Figure A1, we visualize the images reconstructed by the Intermediate network. Firstly, the clean images (unperturbed) have not been deblurred significantly. Secondly, even under mild adversarial attacks, the quality of the reconstructed images is abysmal. We observe severe checkerboard patterns, aliasing, and color mixing in all images reconstructed by the Intermediate network under adversarial attack. Thus, to better understand the performance of the Intermediate network in comparison to the Baseline network and NAFNet, we perform significantly weaker adversarial attacks. To this effect, we use the CosPGD attack but with $\epsilon \approx \frac{2}{255}$, and consider attack iterations $\in \{1, 3, 5\}$. We again use $\alpha = 0.01$.

We report the performance of the Intermediate networks in Table 3. Interestingly, we observe that after one adversarial attack iteration, the Intermediate network is significantly outperforming both the Baseline network and NAFNet. However, the Intermediate network is unable to retain this superior performance, and its performance significantly drops as we increase the attack strength (attack iterations). Additionally, in Figure 4 we observe the introduction of the same spectral artifacts for the Intermediate network as those observed in Figure A1 and Figure A2 (please refer to Section B). The intensity of the spectral artifacts increases as we increase the attack strength. This phenomenon is similar to the performance of NAFNet, which performs admirably on clean samples and under weak adversarial attacks but begins to perform significantly worse as the attack strength increases. This indicates that even smoothed activation functions in the NAFNet architecture instead of Simple Gate produce strong spectral artifacts in the reconstructed images.

Table 3. Comparison of performance of the Baseline network, NAFNet, and Intermediate networks against significantly weak CosPGD attack. For this comparison we use $\epsilon \approx \frac{2}{255}$ and $\alpha = 0.01$ and consider fewer attack steps i.e. iterations $\in \{1, 3, 5\}$

Architecture	CosPGD					
	1 attack itrs		3 attack itrs		5 attack itrs	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Baseline	21.38	0.7520	17.19	0.6356	16.99	0.6316
NAFNet	22.54	0.7883	18.80	0.6948	18.46	0.6835
Intermediate	25.14	0.8410	10.37	0.2940	8.56	0.1812
+ ADV	25.47	0.8555	25.16	0.8501	25.32	0.8555
Intermediate + ReLU	23.96	0.8112	20.96	0.7458	21.5777	0.7594
+ ADV	26.11	0.8616	25.10	0.8459	24.86	0.8413

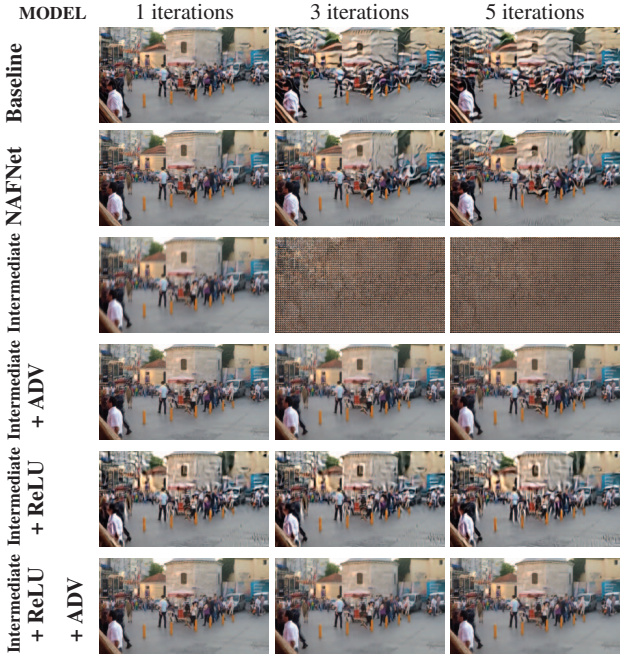

Figure 4. Comparing images reconstructed by different models after a significantly weaker **CosPGD attack** as $\epsilon \approx \frac{2}{255}$.

This is in striking contrast to using a non-smooth non-linear activation function, ReLU. Interestingly, we observe that *Intermediate+ReLU* is significantly more robust, and the degradation in its performance with attack strength is significantly lower than all considered networks, including Restormer. In Figures A1, A2 &4 we observe that the images reconstructed by *Intermediate+ReLU*, while blurry, have significantly fewer artifacts for reasonable values of ϵ.

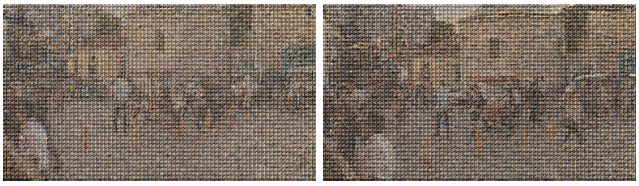

Figure 5. Two different randomly chosen images reconstructed by *Intermediate + ReLU* after 5 iterations of CosPGD attack with significantly higher $\epsilon \approx \frac{20}{255}$. We observe strong spectral artifacts similar to *Intermediate network* in the recovered images.

Under adversarial attacks, the images reconstructed by *Intermediate+ReLU* do not have spectral artifacts similar to *Intermediate network* or NAFNet, but more similar to Restormer and the Baseline. It is only at severely higher $\epsilon \approx \frac{20}{255}$ (refer Fig. 5) that spectral artifacts similar to those produced by *Intermediate network* appear in the reconstructed images from *Intermediate+ReLU*. Thus, the smoothening of feature maps by the conjunction of Simplified Channel Attention and GeLU, and Simple Gate could be attributed to the introduction of some peculiar spectral artifacts and loss in robustness.Using a non-smoothed non-linear activation function like ReLU appears to be an effective mitigation technique.

Additionally, as reported in Table A1 we observe the adversarial robustness of both the *Intermediate network* and *Intermediate+ReLU* significantly increases after FGSM training, and is comparable to Restormer. This significant improvement in adversarial performance is also visible at lower ϵ attacks, please refer to Table 3 and visually shown in Figure 4. Thus, as observed before, adversarial training is a fix to reduce artifacts, even for the *Intermediate network*.

5.2. Superiority of Restormer

In their work, [7] attempt to reduce model complexity while retaining the performance of the Restormer. However, as shown in our work this significantly degrades the generalization ability of the consequent models. As larger models tend to have a better trade-off between robustness and accuracy [22, 24], the reduced model capacity in the Baseline and NAFNet could contribute to the reduced robustness. While reducing model complexity is certainly important and desirable, to maintain robustness it requires a more careful and systematic pruning of networks [59, 41, 24] than simply dropping components. Apart from the model's complexity in terms of the number of parameters, the attention mechanism itself could be crucial for robustness.

While the Restormer uses a multi-headed self-attention mechanism, both the Baseline network and NAFNet use variants of channel-attention (NAFNet uses the simplified channel-attention proposed by [7]). As shown by [4], the self-attention module of vision Transformers significantly aids the Transformer based models to improve their robustness. Additionally, it helps the model better utilize defense strategies such as additional training, distillation, etc. A similar phenomenon is observed in Table 2, as Restormer, a vision transformer-based model with a multi-headed self-attention module is able to better utilize adversarial training compared to the Baseline network and NAFNet.

Limitations. Adversarial training and design choices like the use of smoothed or non-smoothed activation functions against using Simple Gates certainly have a significant impact on the performance of the considered image restoration models. However, there still is a considerable gap in the clean performance of the considered models. While the fixes work in increasing adversarial robustness and removal of spectral artifacts the images are far from ideal restoration. As observed, the restored images after the fixes are significantly blurry. This is a limitation of this work, as this work was focused on the removal of spectral artifacts and better adversarial robustness. This work is a step towards finding a fix and not an absolute fix.

6. Conclusion

Despite recent methods outperforming baselines for various vision tasks, for a method to have a significant contribution to real-world applications, it must be reliable and robust. Thus in this work, we highlight this shortcoming of recently proposed Transformer based image restoration models. While the models proposed by [7] perform satisfactorily for image deblurring on non-perturbed samples, they fail to generalize when slight adversarial perturbations are added to the blurred images. We acknowledge that the reduction in model complexity compared to Restormer is a step in the right direction, however, in this case, it comes at the expense of model robustness. Therefore, we additionally employ adversarial training in an attempt to fix this shortcoming while also improving the quality of the reconstructed images. We observe that adversarial training is able to reduce the spectral artifacts and also results in significant improvements in the adversarial robustness of the image restoration models. However, the extent of the improvement varied with the architectural design decisions. Thus lastly, we investigate the design decisions that might lead to the occurrence of spectral artifacts and loss in robustness for the methods and find an interesting ablation concerning the type of activation functions used when downsampling.

References

[1] Shashank Agnihotri, Steffen Jung, and Margret Keuper. Cospgd: a unified white-box adversarial attack for pixel-wise prediction tasks, 2023.

[2] 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. *Advances in neural information processing systems*, 34:20014–20027, 2021.

[3] D. Babacan, M. N. Molina, R. Do, and A. K. Katsaggelos. Bayesian blind deconvolution with general sparse image priors. In *European Conference on Computer Vision*, pages 341–355, 2012.

[4] 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, pages 26831–26843. Curran Associates, Inc., 2021.

[5] 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*, pages 10231–10241, 2021.

[6] A. Buades, B. Coll, and J.-M. Morel. A non-local algorithm for image denoising. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 60–65 vol. 2, 2005.

[7] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *European Conference on Computer Vision*, pages 17–33. Springer, 2022.

[8] J. Choi, H. Zhang, J. Kim, C. Hsieh, and J. Lee. Deep image destruction: Vulnerability of deep image-to-image models against adversarial attacks. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 1287–1293. IEEE Computer Society, 2022.

[9] Jun-Ho Choi, Huan Zhang, Jun-Hyuk Kim, Cho-Jui Hsieh, and Jong-Seok Lee. Evaluating robustness of deep image super-resolution against adversarial attacks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[10] Jun-Ho Choi, Huan Zhang, Jun-Hyuk Kim, Cho-Jui Hsieh, and Jong-Seok Lee. Adversarially robust deep image super-resolution using entropy regularization. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, November 2020.

[11] P. L. Combettes and J. C. Pesquet. Image restoration subject to a total variation constraint. *IEEE Trans. Image Process.*, 13:1213–1222, 2004.

[12] Marcos V Conde, Ui-Jin Choi, Maxime Burchi, and Radu Timofte. Swin2sr: Swinv2 transformer for compressed image super-resolution and restoration. In *European Conference on Computer Vision*, pages 669–687. Springer, 2022.

[13] Francesco Croce and Matthias Hein. On the interplay of adversarial robustness and architecture components: patches, convolution and attention. In *New Frontiers in Adversarial Machine Learning Workshop at ICML*, 2022.

[14] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on image processing*, 16(8):2080–2095, 2007.

[15] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *International conference on machine learning*, pages 933–941. PMLR, 2017.

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[17] Kanchana Vaishnavi Gandikota, Paramanand Chandramouli, and Michael Moeller. On adversarial robustness of deep image deblurring. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3161–3165, 2022.

[18] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2015.

[19] Jindong Gu, Hengshuang Zhao, Volker Tresp, and Philip Torr. Segpgd: An effective and efficient adversarial attack for evaluating and boosting segmentation robustness, 2022.

[20] Chun-Le Guo, Qixin Yan, Saeed Anwar, Runmin Cong, Wenqi Ren, and Chongyi Li. Image dehazing transformer with transmission-aware 3d position embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5812–5820, 2022.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[22] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations, 2019.

[23] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus), 2016.

[24] J Hoffmann, S Agnihotri, Tonmoy Saikia, and Thomas Brox. Towards improving robustness of compressed cnns. In *ICML Workshop on Uncertainty and Robustness in Deep Learning (UDL)*, 2021.

[25] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[26] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *European Conference on Computer Vision (ECCV)*, pages 491–507, Cham, 2020. Springer International Publishing.

[27] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021.

[28] Yuanchu Liang, Saeed Anwar, and Yang Liu. Drt: A lightweight single image deraining recursive transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 589–598, 2022.

[29] 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*, pages 10012–10022, 2021.

[30] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

[31] Xiao-Guang Lv, Yong-Zhong Song, Shun-Xu Wang, and Jiang Le. Image restoration with a high-order total variation minimization method. *Applied Mathematical Modelling*, 37(16):8210–8224, 2013.

[32] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks, 2019.

[33] 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 (ICCV)*, pages 7838–7847, October 2021.

[34] Ali Mosleh, J. M. Pierre Langlois, and Paul Green. Image deconvolution ringing artifact detection and removal via psf frequency analysis. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 247–262, Cham, 2014. Springer International Publishing.

[35] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, July 2017.

[36] Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen. Single image super-resolution via a holistic attention network. In *European Conference on Computer Vision*, pages 191–207. Springer, 2020.

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

[38] Francesco Pinto, Philip H. S. Torr, and Puneet K. Dokania. An impartial take to the cnn vs transformer robustness contest. In *European Conferece on Computer Vision (ECCV)*, pages 466–480, Cham, 2022. Springer Nature Switzerland.

[39] William Hadley Richardson. Bayesian-based iterative method of image restoration. *JoSA*, 62(1):55–59, 1972.

[40] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *18th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241. Springer, 2015.

[41] Vikash Sehwag, Shiqi Wang, Prateek Mittal, and Suman Jana. Hydra: Pruning adversarially robust neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 19655–19666. Curran Associates, Inc., 2020.

[42] Rulin Shao, Zhouxing Shi, Jinfeng Yi, Pin-Yu Chen, and Cho-Jui Hsieh. On the adversarial robustness of vision transformers. In *NeurIPS ML Safety Workshop*, 2022.

[43] Rulin Shao, Zhouxing Shi, Jinfeng Yi, Pin-Yu Chen, and Cho-Jui Hsieh. On the adversarial robustness of vision transformers. *Transactions on Machine Learning Research*, 2022.

[44] Yuda Song, Zhuqing He, Hui Qian, and Xin Du. Vision transformers for single image dehazing. *IEEE Transactions on Image Processing*, 32:1927–1941, 2023.

[45] Jingwen Su, Boyan Xu, and Hujun Yin. A survey of deep learning approaches to image restoration. *Neurocomputing*, 487:46–65, 2022.

[46] Maitreya Suin, Kuldeep Purohit, and A. N. Rajagopalan. Spatially-attentive patch-hierarchical network for adaptive motion deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[47] Shiyu Tang, Ruihao Gong, Yan Wang, Aishan Liu, Jiakai Wang, Xinyun Chen, Fengwei Yu, Xianglong Liu, Dawn Song, Alan Yuille, Philip H.S. Torr, and Dacheng Tao. Robustart: Benchmarking robustness on architecture design and training techniques. *https://arxiv.org/pdf/2109.05211.pdf*, 2021.

[48] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34:24261–24272, 2021.

[49] Fu-Jen Tsai, Yan-Tsung Peng, Yen-Yu Lin, Chung-Chi Tsai, and Chia-Wen Lin. Stripformer: Strip transformer for fast image deblurring. In *European Conference on Computer Vision*, pages 146–162. Springer, 2022.

[50] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxim: Multi-axis mlp for image processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5769–5780, 2022.

[51] Jeya Maria Jose Valanarasu, Rajeev Yasarla, and Vishal M Patel. Transweather: Transformer-based restoration of images degraded by adverse weather conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2353–2363, 2022.

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[53] Zeyu Wang, Yutong Bai, Yuyin Zhou, and Cihang Xie. Can CNNs be more robust than transformers? In *The Eleventh International Conference on Learning Representations*, 2023.

[54] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[55] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings*

of the IEEE/CVF conference on computer vision and pattern recognition, pages 17683–17693, 2022.

[56] Jie Xiao, Xueyang Fu, Man Zhou, Hongjian Liu, and Zheng-Jun Zha. Random shuffle transformer for image restoration. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 38039–38058. PMLR, 23–29 Jul 2023.

[57] Cihang Xie, Mingxing Tan, Boqing Gong, Alan Yuille, and Quoc V Le. Smooth adversarial training. *arXiv preprint arXiv:2006.14536*, 2020.

[58] Hanshu Yan, Jingfeng Zhang, Jiashi Feng, Masashi Sugiyama, and Vincent Y. F. Tan. Towards adversarially robust deep image denoising. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 1516–1522, 7 2022.

[59] Shaokai Ye, Kaidi Xu, Sijia Liu, Hao Cheng, Jan-Henrik Lambrechts, Huan Zhang, Aojun Zhou, Kaisheng Ma, Yanzhi Wang, and Xue Lin. Adversarial robustness vs. model compression, or both? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 111–120, 2019.

[60] Yi Yu, Wenhan Yang, Yap-Peng Tan, and Alex C Kot. Towards robust rain removal against adversarial attacks: A comprehensive benchmark analysis and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6013–6022, 2022.

[61] Jiutao Yue, Haofeng Li, Pengxu Wei, Guanbin Li, and Liang Lin. Robust real-world image super-resolution against adversarial attacks. In *Proceedings of the 29th ACM International Conference on Multimedia*, page 5148–5157, New York, NY, USA, 2021. Association for Computing Machinery.

[62] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022.

[63] Shangchen Zhou, Jiawei Zhang, Wangmeng Zuo, and Chen Change Loy. Cross-scale internal graph neural network for image super-resolution. *Advances in neural information processing systems*, 33:3499–3509, 2020.