LESS IS MORE: TRAINING ON LOW-FIDELITY IMAGES IMPROVES ROBUSTNESS TO ADVERSARIAL ATTACKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep neural networks (DNNs) have been shown to be vulnerable to adversarial attacks – subtle, perceptually indistinguishable perturbations of inputs that nonetheless change the response of the model. Since adversarial attacks are defined relative to human perception, it may be fruitful to investigate why human perception (and biological perception in general) is robust to the types of perturbations that DNNs are convincingly deceived by. In the context of vision, we hypothesize that a factor contributing to the robustness of human visual perception is our constant exposure to low-fidelity visual stimuli. To investigate the impact, vis-à-vis adversarial robustness, of exposure to low-fidelity visual stimuli, we train and evaluate object recognition DNNs on images which have been blurred and have had their color saturation reduced. We find that DNNs trained on such images can achieve high classification accuracy over a small number of classes, while becoming significantly more robust to low-magnitude adversarial attacks. Furthermore, we design a blurring module that simulates that loss of visual acuity with increasing eccentricity by selecting the intensity of Gaussian blur at each pixel based on its distance from a given fixation point. Our results indicate that using this retina-inspired blurring mechanism, instead of blurring the entire image with the same Gaussian kernel, yields better robustness while keeping the accuracy on clean data unchanged.

1 INTRODUCTION

Deep Neural Networks (DNNs), particularly their instantiation as Convolutional Neural Networks (CNNs) have improved the state-of-the-art in computer vision by leaps and bounds, however, they are vulnerable to *adversarial attacks*. An adversarial attack attempts to manipulate a DNN's response to a given image by perturbing it so subtly that to a human observer the perturbed data appears identical (if not perceptually identical, then at the very least semantically identical) to the original (Akhtar et al., 2021). The possibility of such attacks raise questions about the reliability of DNN-based systems that are becoming ever more pervasive in real-world application. Consequently, developing methods to make DNNs more robust to such semantically irrelevant perturbations has become a priority.

Several defenses have been proposed over the years to make DNNs more robust to adversarial attacks. Most of these defenses can be divided into two categories based on the philosophy behind them. The defenses in the first category include training on perturbed images, where the perturbation may be computed via an adversarial attack (adversarial training) (Madry et al., 2018; Wong et al., 2019; Zhang et al., 2019; Bai et al., 2021), or could be random noise (randomized smoothing) (Cohen et al., 2019; Fischer et al., 2020; Kumar & Goldstein, 2021), denoising the input and the intermediate activations of the DNN (Salman et al., 2020; Carlini et al., 2022), and training adversarial perturbation detectors to filter out corrupted data (Metzen et al., 2017). The second category of defenses is based on the view that DNNs are supposed to be a model of human vision and, since the existence of adversarial attacks presents a divergence between DNNs and humans, there may be some aspect of human vision fundamental to its robustness that is not modeled by DNNs. These defenses generally involve integrating computational analogues of biological process that are absent from common DNNs, such as predictive/sparse coding (Paiton et al., 2020; Bai et al., 2021), biologically constrained visual filters, nonlinearities and stochasticity (Dapello et al., 2020), foveation (Jonnalagadda et al., 2022; Luo et al., 2015; Gant et al., 2021), and non-uniform retinal sampling and cortical fixations Vuyyuru et al. (2020),into DNNs and it has been demonstrated that the resulting models are made more robust to adversarially perturbed data.

While the first category of defenses has achieved great success, we are inclined towards biologically motivated methods of inducing adversarial robustness because they are trying to bring DNNs closer to the oracle, i.e., the human visual perception, relative to which the adversarial/non-adversarial nature of an input is defined. Following this line of inquiry, in this paper, we investigate the contribution of low-fidelity visual sensing, particularly the kind that occurs in peripheral vision, to robustness against small perturbations that convincingly deceive DNNs. Unlike DNNs, which sense visual stimuli at maximum fidelity at every point in their visual field, most of the human visual field is low fidelity in that it lacks fine-grained contrast and color information. In adults, with fully developed vision, visual stimuli can be sensed with high fidelity in only a small region (less than 1% by area) of the visual field around the point of fixation. In the remainder of the visual field (the periphery), the fidelity of the sensed stimuli decreases exponentially with distance from the point of fixation (eccentricity). This phenomenon is called "foveation". Despite this limitation, humans can accurately categorize objects that appear in the periphery into high-level classes (Ramezani et al., 2019), while the presence of a small amount of noise or blurring can decimate the accuracy of an otherwise accurate DNN if this distortion was not presented during training. Therefore, we hypothesize that the experience of viewing the world at multiple levels of fidelity, perhaps even at the same instant, leads to human vision becoming invariant to low-level features, such as those that are perturbed by adversarial attacks, and instead relying on higher-level features.

Although some prior works study the impact of simulating foveation in CNNs on their adversarial robustness, the body of literature in this space remains sparse. In one of the earliest works in this area Luo et al. (2015) implement a foveation-based defense by cropping the input image using various schemes. More recently, Vuyyuru et al. (2020) proposed a defense that simulates retinal non-uniform sampling. They do this by resampling input images so that the sampling density of pixels is maximal at the point of fixation and decays progressively in regions further away from it. Although the accuracy of these biologically inspired approaches is higher than that of an unmodified model under adversarial attack, there is still a large gap in robustness between the proposed approaches and non-biological defenses, indicating that there is room for improvement.

In this paper, we propose an alternate foveation technique, that we call *Retina Blur* (referred to as R-Blur henceforth), that blurs the input image and reduces its color saturation adaptively based on the distance from a given fixation point. The effect is that regions further away from the fixation point appear more blurry and less vividly colored than regions closer to the point of fixation. Although similar adaptive blurring methods have been proposed as computational approximations for foveation (Deza & Konkle, 2021; Pramod et al., 2018; Wang & Cottrell, 2017), their impact on robustness has not yet been evaluated to the best of our knowledge. Furthermore, color sensitivity is known to decrease in the peripheral regions of the visual field (Hansen et al., 2009; Johnson, 1986), yet most of the existing methods of foveation do not account for this phenomenon. In light of recent results (Jinsi et al., 2022) indicating that CNNs benefit from being trained on blurry grayscaled images we decided to simulate this phenomenon as well in our proposed approach.

We evaluate R-Blur as a preprocessing step for popular CNN models, and train them on object categorization tasks, on datasets of real world images, namely CIFAR-10 (Krizhevsky et al.) and Ecoset (Mehrer et al., 2021). We show that models augmented with R-Blur are more robust to small and moderate magnitude adversarial perturbations by as much as 50% (absolute) on moderately large ℓ_2 (0.008) and ℓ_∞ (2.0) perturbations computed via 100-step APGD (Croce & Hein, 2020), compared to the unmodified CNN, and other baselines (Vuyyuru et al., 2020; Madry et al., 2018; Cohen et al., 2019). Particularly, R-Blur outperforms biologically inspired defenses by up to 50%. For non-biological defenses, R-Blur closes the gap in robustness between biological and non-biological defenses from around 70% to 30% for moderate perturbation distances ($||\epsilon||_\infty = 0.008$).

We also show that the robustness achieved by R-Blur is certifiable using the approach from (Cohen et al., 2019) and that the certified accuracy achieved by R-Blur is competitive with that achieved by randomized smoothing (Cohen et al., 2019) for small radii. Finally, by means of ablation, we show that both blurring and color desaturation contribute to the improved robustness.

2 Retinal Blur: An Approximation for Peripheral Vision

To simulate the loss in contrast and color sensitivity of human perception with increasing eccentricity, we propose R-Blur, an adaptive Gaussian blurring and color desaturation technique. Applying R-Blur

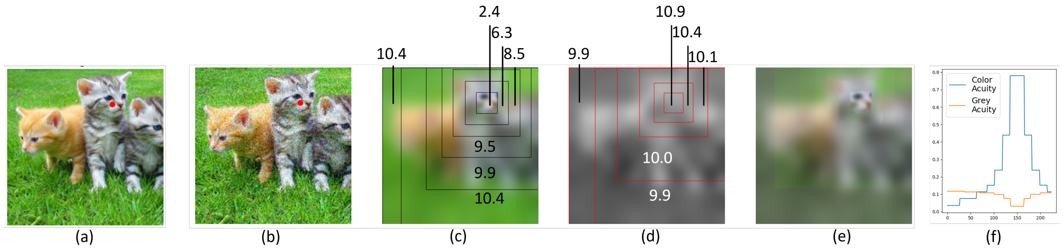


Figure 1: Given a clean input image (a) and a fixation point (shown in red at location $(50, 150)$), R-Blur adds Gaussian noise to the image and obtains (b). R-Blur then creates a colored and a grayscaled copy of the image, and applies adaptive Gaussian blurring to them. The low fidelity color and grayscale images are shown in (c) and (d). The boxes in the figure denote the regions in which the same Gaussian kernel is used and the number indicates the standard deviation of this kernel. The blurred color and gray images are combined in a pixel-wise weighted combination in which the weight of the colored and gray pixels are a function of their respective estimated acuity values. The estimated acuity values for the color and gray pixels in row 50 (the row in which the fixation point is located) of the image shown by the blue and orange curves, respectively, in (f). The final image, shown in (e), is passed to the downstream network.

to an image at a given fixation point will result in an image in which sharpness and color saturation decrease exponentially as the distance from the point of fixation, i.e. eccentricity, increases. In this section, we first provide a high-level overview of the operations performed by R-Blur and then discuss some of the more involved operations in further detail in the succeeding sub-sections.

2.1 Overview

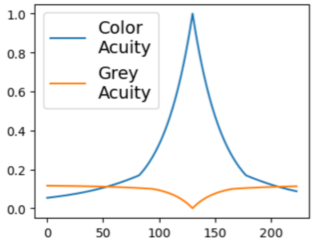
The sequence of operations performed by R-Blur, given an image and fixation point, is briefly described below and illustrated in Figure 1. First, R-Blur adds Gaussian noise to the image to simulate stochasticity in the firing rates of biological neurons and photoreceptors. Then it creates two copies of the image, one colored and the other grayscale, and estimates the acuity of color and grayscale vision at each pixel location using distributions that approximate the relationship between eccentricity and photopic and scotopic visual acuity levels in humans. R-Blur then applies *adaptive* Gaussian blurring to both image copies such that the standard deviation of the Gaussian kernel at each pixel in the color and grayscale image is a function of the estimated color and grayscale acuity, respectively, at that pixel. Finally, the two blurred images are combined in a pixel-wise weighted combination in which the weights of the colored and gray pixels are a function of their respective estimated acuity values. Now we describe some of the more involved operations in greater detail.

2.2 Eccentricity Computation

The distance of a pixel location from the fixation point, i.e. its eccentricity, determines the standard deviation of the Gaussian kernel applied to it and the combination weight of the color and gray images at this location. Our desiderata for the distance metric is that it should produce un-rotated square level sets, which are better aligned with the un-rotated square receptive fields of CNNs, and that it must be normalized to lie in $[0, 1]$ in order to facilitate downstream computation by making them invariant to the size of the visual field. In this context the visual field refers to the rectangular region over which R-Blur operates and defines the maximum image size that is expected by R-Blur. With these requirements in mind, we compute the eccentricity of the pixel at location (x_p, y_p) as

$$e_{x_p,y_p} = \frac{\max(|x_p - x_f|, |y_p - y_f|)}{W_V}, \tag{1}$$

where (x_f, y_f) and W_V represent the fixation point and the width of the visual field, respectively. In ophthalmology, the eccentricity is usually measured in degrees so one may consider e_{x_p,y_p} to be the traditional measure of eccentricity in radians divided by π.



(a)

Figure 2: The visual acuity of sharp and colorful, photopic, and gray and blurry, scotopic, vision, estimated using the method from 2.3.

2.3 Visual Acuity Estimation

We compute the visual acuity at each pixel location based on its eccentricity. The biological retina contains two types of photoreceptors. The first type, called cones, are color sensitive and give rise to high-fidelity visual perception at the fovea, while the second type, called rods, are sensitive to only to illumination but not color and give rise to low fidelity vision in the periphery. The acuity of color and grayscale vision, arising from the cones and rods, is estimated at each pixel location, (x, y), using two sampling distributions $D_R(e_{x,y})$ and $D_C(e_{x,y})$, respectively. In this work we have used

$$\mathcal{D}(e; \sigma, \alpha) = \max\left[\lambda(e; 0, \sigma), \gamma(e; 0, \alpha\sigma)\right] \tag{2}$$

$$D_C(e; \sigma_C, \alpha) = \mathcal{D}(e; \sigma_C, \alpha) \tag{3}$$

$$D_R(e; \sigma_R, \alpha, p_{max}) = p_{max}(1 - \mathcal{D}(e; \sigma_R, \alpha)), \tag{4}$$

where $\lambda(.; \mu, \sigma)$ and $\gamma(.; \mu, \sigma)$ are the PDFs of the Laplace and Cauchy distribution with location and scale parameters μ and σ, and α is a parameter used to control the width of the distribution. In this paper we set $\sigma_C = 0.12, \sigma_R = 0.09, \alpha = 2.5$ and $p_{max} = 0.12$ to match the curves of photopic and scotopic visual acuity from (vis). The resulting acuity estimates are illustrated in Figure 2a and the measured photopic and scotopic acuity curves from (vis) can be viewed at this url [1]. Unfortunately the illustration from (vis) cannot be reproduced here due to copyright reasons.

2.3.1 Quantizing the Visual Acuity Estimate

In the form stated above, we would need to create and apply as many Gaussian kernels as the distance between the fixation point and the farthest vertex of the visual field. This number can be quite large as the size of the image increases and will drastically increase the per-image computation time. In order to mitigate this issue we quantize the estimated acuity values. As a result, the locations to which the same kernel is applied no longer constitute a single pixel perimeter, but become a much wider region (see Figure 1 (c) and (d)), which allows us to apply the Gaussian kernel in these regions very efficiently using optimized implementations of the convolution operator.

To create a quantized eccentricity-acuity mapping, we do the following. We first list all the color and gray acuity values possible in the visual field by assuming a fixation point at $(0, 0)$, computing eccentricity values $e_{0,y}$ for $y \in [0, W_V]$ and the corresponding values of $\mathcal{D}_R = \{D_R(e_{0,y}) | y \in [0, W_V]\}$ and $\mathcal{D}_C = \{D_C(e_{0,y}) | y \in [0, W_V]\}$. We then compute and store the histograms, H_R and H_C, from $\mathcal{D}_R$ and $\mathcal{D}_C$, respectively. To further reduce the number of kernels we need to apply and increase the size of the region each of them is applied to, we merge the bins containing less than τ elements in each histogram with the adjacent bin to their left. Thereafter, given an image to process, we will compute the color and gray visual acuity for each pixel, determine which in which bin it falls in H_R and H_C, and assign it the average value of that bin.

2.4 Changing the Viewing Distance

The gain in fidelity resulting from bringing the target of visual perception closer to the eye can be quite easily simulated with R-Blur. This is illustrated in Figure 3. A reduction in viewing distance

[1] See Figure 14.3 at `https://nba.uth.tmc.edu/neuroscience/m/s2/chapter14.html`

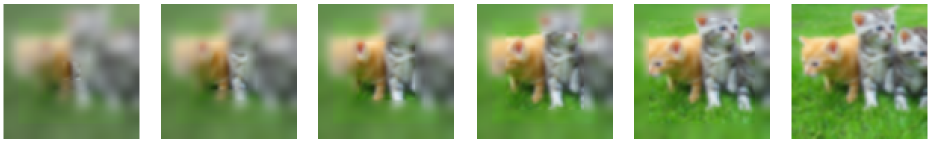


Figure 3: Illustration of increasing the viewing distance (left to right) for a given image.

can be simulated by scaling down $e_{x,y}$ prior to estimating the acuity values. However, for the results we present in this paper, we adopted a slightly different procedure. This procedure effectively drops the k lowest acuity bins and shifts the pixels assigned to them k bins ahead and the pixels that were in bins 1 through $k-1$ are now assigned to bin 1. This is implemented by simply slicing the arrays containing the quantized $D_C(e_{x,y})$ and $D_R(e_{x,y})$, instead of having to compute them again for the down-scaled $e_{x,y}$ values for each image.

Formally, we quantize $D_C(e_{x,y})$ and $D_R(e_{x,y})$ into bins. Let $D = [d_1, ..., d_n]$ represent the value assigned to each bin and let $P_i = [p_{i,1}, ..., p_{i,m_i}]$ be the pixel locations assigned to the i^{th} bin, and let P_1 and P_n correspond to points with the lowest and highest eccentricity, respectively. To reduce the viewing distance we merge bins 1 through k such that the new set of bins has values $D' = [d_1, ..., d_{n-k}]$ and the pixels corresponding to them are $P'_1 = [P_1, ..., P_k]$ and $P_{i>1} = P_{k+1}$.

2.5 Blurring and Color Desaturation

We map the estimated visual acuity at each pixel location, (x_p, y_p), to the standard deviation of the Gaussian kernel that will be applied at that location as $\sigma_{(x_p,y_p)} = \beta W_V(1 - D(e_{x,y}))$, where β is constant to control the maximum standard deviation, and $D = D_C$ for pixels in the colored image and $D = D_R$ for pixels in the grayscaled image. We then apply Gaussian kernels of the corresponding standard deviation to each pixel in the colored and grayscale image to obtain an adaptively blurred copy of each, which we combine in a pixel-wise weighted combination to obtain the final image. The weight of each colored and gray pixel is given by the normalized color and gray acuity, respectively, at that pixel. Formally, the pixel at (x_p, y_p) in the final image has value

$$v^f_{(x_p,y_p)} = \frac{v^c_{(x_p,y_p)} D_C(e_{x,y}; \sigma_C, \alpha) + v^g_{(x_p,y_p)} D_R(e_{x,y}; \sigma_C, \alpha)}{D_C(e_{x,y}; \sigma_C, \alpha) + D_R(e_{x,y}; \sigma_C, \alpha)},$$

where $v^c_{(x_p,y_p)}$ and $v^g_{(x_p,y_p)}$ are the pixel value at (x_p, y_p) in the blurred color and gray images respectively.

3 Evaluation

3.1 Experimental Setup

We use three datasets containing real-world images, namely CIFAR-10 (Krizhevsky et al.), and two subsets of Ecoset (Mehrer et al., 2021) with 10 and 100 classes, referred to henceforth as Ecoset-10 and Ecoset-100. To create Ecoset-10 and Ecoset-100, we order the classes in the Ecoset dataset by the number of images that each class is assigned to and then select the top 10 and top 100 classes respectively. The training/validation/test splits of CIFAR-10, Ecoset-10 and Ecoset-100 are 45K/5K/10K, 48K/859/1K and 470.6K/5K/1K, respectively.. During training we used random horizontal flipping and padding + random cropping, as well as AutoAugment (Cubuk et al., 2018) for CIFAR-10 and RandAugment for Ecoset-10/100. All Ecoset images were resized to 224 × 224 pixels. We used variants of ResNet (He et al., 2016) as our image classification models. Specifically, for CIFAR-10 we use a Wide-Resnet (Zagoruyko & Komodakis, 2016) model with 22 convolutional layers and a widening factor of 4, and for Ecoset-10/100 we use ResNet-18, in which the number of channels in each layer has been doubled. On each dataset we train models with five configurations: (1) no modification (ResNet), (2) fast adversarial training (Wong et al., 2019) with $||\epsilon||_\infty = 0.008$ (AT), (3) adding Gaussian noise with mean 0 and standard deviation σ to the images during training (G-Noise) similar to (Cohen et al., 2019), (4) R-Blur preprocessing, and (5) Retinal Warping (R-Warp) preprocessing proposed by (Vuyyuru et al., 2020), which simulates foveation by resampling

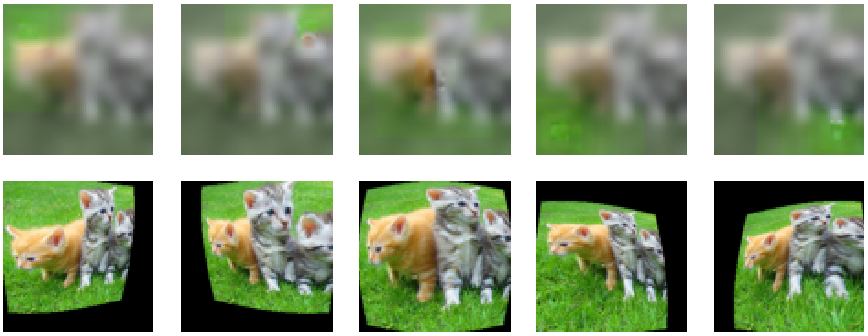


Figure 4: An example of the set of images produced by the five-point fixation procedure for R-Blur (top) and R-Warp (bottom)

input images such that the sampling density of pixels is maximal at the point of fixation and decays progressively in regions further away from it. The value of σ in (3) was set to 0.062 for CIFAR-10 and 0.125 for Ecoset-10/100. While training models with R-Blur and R-Warp, we split each batch into sub-batches of 32 images, and for each sub-batch we randomly sample a fixation point and use it to apply R-Blur or R-Warp to the images in that sub-batch. During inference, we average the logits for 5 predefined fixation points corresponding to the 4 corners of the image and the center. These points are $(h/6, w/6)$, $(h/6, w - w/6)$, $(h/2, w/2)$, $(h - h/6, w/6)$, $(h - h/6, w - w/6)$, where h and w are the height and width of the image. An example of these fixations is shown in Figure 4. While training the R-Blur model, we also set the viewing distance uniformly at random using the procedure described in 2.4. For R-Blur and G-Noise, we add Gaussian noise *during training only*, no noise is added during inference. For CIFAR-10, Ecoset-10 and Ecoset-100, we train 5, 2, and 1 models respectively per configuration. The numbers in the Table 1 are averages and the error bars in Figure 5 represent 95% CI.

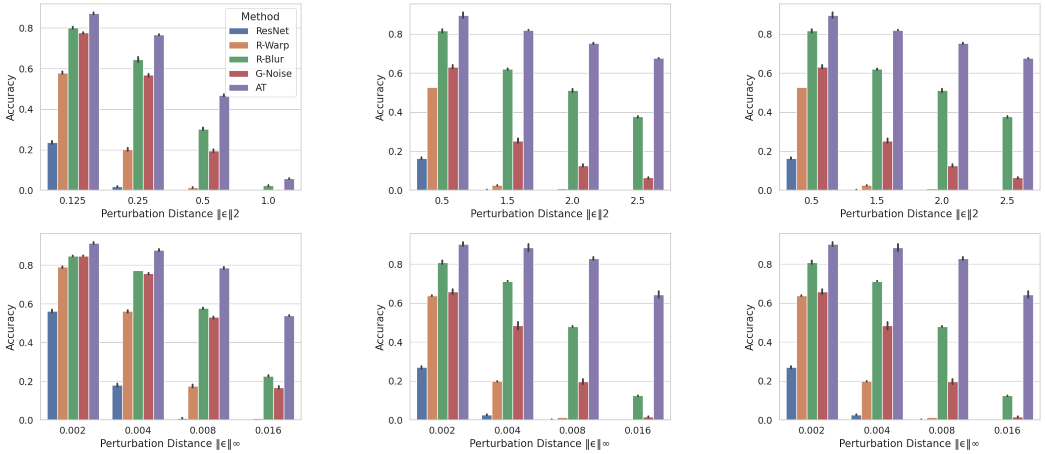
3.2 Accuracy Against Adversarial Attacks

The accuracy achieved by R-Blur and baseline models against 100-step APGD attack Croce & Hein (2020) for various ℓ_2 and ℓ_∞ bounds is shown in Figure 5. APGD is a state-of-the-art white-box attack, that uses projected gradient descent with adaptive step sizes to find adversarial perturbations. As expected, the accuracy of the unmodified model (ResNet) decreases rapidly as the perturbation distance increases. Note that even at relatively large perturbations of ℓ_2-norm 2.5 and ℓ_∞-norm 0.008, the model with R-Blur maintains an accuracy close to 50% for the most complex dataset in our evaluation, Ecoset-100. Moreover, at the maximum perturbation distance, both R-Warp and G-Noise have almost 0% accuracy on Ecoset-100, while R-Blur achieves more than 10% accuracy which is 10 times larger than chance. If we look at biological defenses (R-Warp), the accuracy of R-Warp decreases rapidly and falls below 20% on the second smallest perturbation size and then to almost 0% as the perturbation increases. In contrast, R-Blur has a higher accuracy than R-Warp against adversarial perturbations and thus, R-Blur improves the robustness of DNNs compared to existing biologically inspired, foveation-based, defenses (R-Warp) by as much as 50% (absolute) on moderately large ℓ_2 and ℓ_∞ perturbations, respectively.

For the non-biological defenses (G-Noise and AT), R-Blur achieves a higher accuracy than G-Noise, on all the three datasets. Although the accuracy of R-Blur is lower than that of AT, it does noticeably close the gap in robustness between existing biological defenses (R-Warp) and AT and takes us closer in our goal of developing a biologically-inspired defense against adversarial attacks.

3.3 Accuracy on Clean Data

We evaluate R-Blur and baseline models on the unperturbed test set of each dataset and present the accuracy of each model in Table 1. Considering non-biological defenses, AT achieves much higher accuracy than R-Blur on all the datasets. R-Blur, however, achieves a higher accuracy than G-Noise for the more complex Ecoset-10 and Ecoset-100 datasets, while being slightly better on CIFAR-10.



(a) CIFAR-10 (b) Ecoset-10 (c) Ecoset-100

Figure 5: R-Blur outperforms biologically inspired defenses (R-Warp) by up to 50% (absolute). Moreover, for non-biological defenses, R-Blur improves robustness over G-Noise by up to 30% (absolute). For AT, we close the gap in robustness between existing biological defenses and AT from around 70% to 30% for moderate perturbation distances ($||\epsilon||_\infty = 0.008$).

Method	CIFAR-10	Ecoset-10	Ecoset-100
ResNet	94.7	93.8	87.4
AT	93.7	92.1	83.9
GNoise	91.3	80.6	74.8
R-Blur (ours)	90.4	87.9	75.7
R-Warp	93.1	93.4	88.0

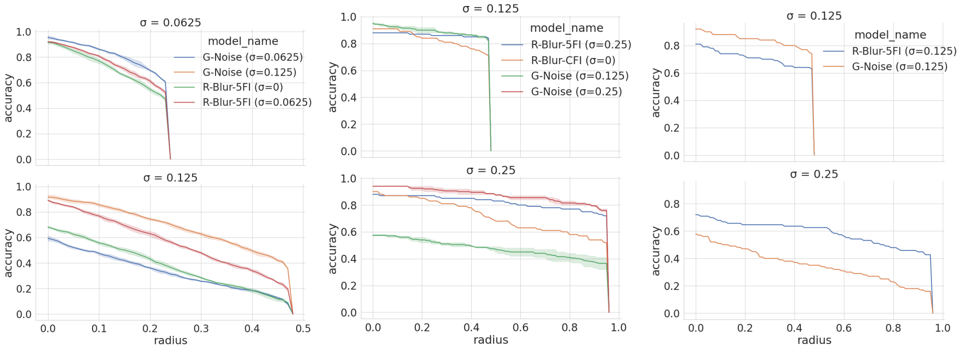
Table 1: Top-1 classification accuracy of our approach (R-Blur) and baselines on various datasets.

This indicates that the adaptive blurring mechanism in R-Blur counteracts the negative impact of Gaussian noise to some extent while retaining its robustness-related benefits. Comparing to R-Warp, R-Blur achieves significantly lower accuracy on clean data, which is not altogether surprising given that, as shown in Figure 4, the images processed by R-Warp, unlike R-Blur, retain a lot of fine-grained details and all of the color information – R-Warp is essentially simulating a large viewing distance in which most of the image lies close to the center rather than the periphery of the visual field.

3.4 Certified Adversarial Robustness

The results presented in the previous section although are a very strong indication of the robustness of our model, they do not constitute a formal guarantee that some other, better, method of producing adversarial perturbations will not be able to drastically reduce the accuracy of our model. To make sure that the gains in robustness we have observed above are indeed reliable, we use the randomized smoothing approach of Cohen et al. (2019) to certify the robustness of our model. This entails perturbing the input image with a *very* large number of Gaussian noise samples, predicting each perturbed input using the model, and using a hypothesis test to determine if the model's prediction is indeed stable and correct in the region spanned by the noise samples.

The plots in Figure 6 show the certified accuracy at several ℓ_2 norm radii for our model and the G-Noise model on different datasets. We compare with G-Noise here because this was the setup used in (Cohen et al., 2019). These plots indicate that the model achieves the accuracy indicated by the curve 99.9% of the time, if the data used to generate these plots (200 CIFAR images, and 100 images from Ecoset-10 and Ecoset-100) is perturbed within an ℓ_2 corresponding to the radius values on the x-axis. We see that R-Blur achieves a high-level of certified robustness on all three datasets, with the certified accuracy being close to the ones observed in Figure 5, indicating that our earlier results were a faithful representation of R-Blur's robustness. Furthermore, it can be seen, that the trends from



(a) CIFAR-10 (b) Ecoset-10 (c) Ecoset-100

Figure 6: The certified accuracy at various ℓ_2-norm radii of R-Blur and G-Noise models. The value of σ in the legend denotes the standard deviation of noise added during training only, where as the σ on top of the plots refers to the standard deviation of the noise used to compute the certified accuracy. R-Blur achieves a high-level of certified robustness on all three datasets, with the certified accuracy being close to the ones observed in Figure 5. The certified robustness of R-Blur degrades slower than that of G-Noise at higher radii.

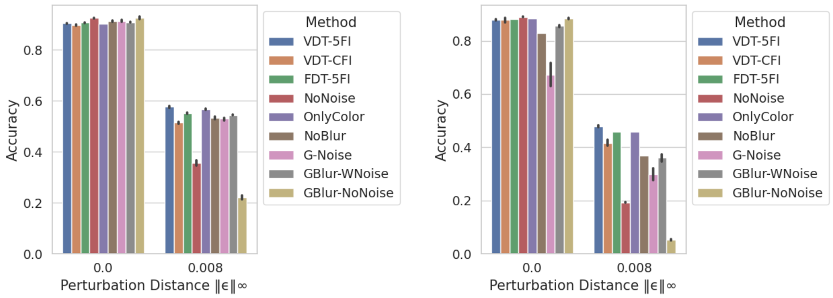
Figure 5 are mirrored here as well, in that the certified robustness of R-Blur degrades slower than that of G-Noise at higher radii thus indicating that G-Noise would need to be trained with a much larger amount of noise to achieve the robustness that R-Blur achieves with a fraction of the noise, or even none at all (see the plots corresponding to R-Blur ($\sigma = 0$)).

3.5 Ablation Study

Having established that R-Blur indeed improves the robustness of image recognition model, we examine, by way of ablation, how much each component of R-Blur contributes towards this improvement as shown in Figure 7. Note that NoNoise, NoBlur, and OnlyColor include Variable Distance Training with 5-fixation inferrence (VDT-5FI). The most significant contributor to robustness was the addition of noise during training (VDT-5FI v. NoNoise), followed by blurring (VDT-5FI v. NoBlur). We note here that adaptive nature of the blur was a major contributor as well since applying uniform Gaussian blurring (GBlur-WNoise) with noise achieved significantly lower robustness – equal to the model without blur (NoBlur). The next most significant factor was evaluating with five fixation points (VDT-5FI) which improved robustness significantly compared to a single fixation point in the center of the image (VDT-CFI). This indicates that multiple fixations and saccades are important when the image is hard to recognize and present a promising direction for future work to explore. Furthermore, we note that not adaptively desaturating the colors (OnlyColor) reduces the robustness slightly, but noticeably. We find that training with variable viewing distances (VDT-CFI) improves robustness compare to training with fixed viewing distances (FDT-CFI). To summarize, all the biologically-motivated components of R-Blur contribute towards improving the adversarial robustness of object recognition DNNs from close to 0% to around 65% ($||\epsilon||_\infty = 0.004$ for Ecoset-10)

4 Related Work

Non-biological defenses: These defenses make up the bulk of proposed defenses and can be broadly classified into three categories, namely adversarial training, certified robustness and detection algorithms. Adversarial training algorithms (Madry et al., 2018; Zhang et al., 2019; Rebuffi et al., 2021; Bai et al., 2021; Wong et al., 2019) train models with a modified loss function such that in each training iteration instead of being shown a natural image, the model is shown an adversarially perturbed image in which the perturbation is computed by using Projected Gradient Descent to *maximize* the classification loss. On the other Certified robustness approaches (Cohen et al., 2019; Fischer et al., 2020; Kumar & Goldstein, 2021; Li et al., 2019) propose defenses that are acompanied by provable guarantees of the form: with probable $1 - \delta$, where δ is small, the model's output will not change if a given image, x, is perturbed by a perturbation having norm at most ϵ. Finally, detection



(a) CIFAR-10

(b) Ecoset-10

Figure 7: The results of an ablation study that show the effect, on clean accuracy and robustness to ℓ_∞ APGD attack, of each component of R-Blur. We can see that removing the biologically-motivated components of R-Blur, namely adaptive blur (VDT-5FI v. NoBlur), adaptive color desaturation (VDT-5FI v. OnlyColor) and noise (VDT-5FI v. NoNoise), deteriorates the adversarial robustness.

algorithms (Metzen et al., 2017) are designed to detect if an adversarial example has been perturbed or not, so that it may be potentially excluded from the dataset. The main shortcoming of these non-biologically motivated defenses is that by not being cognizant of the biological basis of robust vision they are excluding, from the hypothesis class under consideration, a large set of potentially very effective and elegant approaches for defending against adversarial attacks.

Biologically inspired defenses: These defenses involve integrating computational analogues of biological process that are absent from common DNNs, such as predictive/sparse coding (Paiton et al., 2020; Bai et al., 2021), biologically constrained visual filters, nonlinearities and stochasticity (Dapello et al., 2020), foveation (Jonnalagadda et al., 2022; Luo et al., 2015; Gant et al., 2021), and non-uniform retinal sampling and cortical fixations Vuyyuru et al. (2020),into DNNs and it has been demonstrated that the resulting models are made more robust to adversarially perturbed data.

Retina Models for Robustness: Some prior works have studied the impact of simulating foveation in CNNs on their adversarial robustness, the body of literature in this space remains rather sparse. In one of the earliest works in this area Luo et al. (2015) implement a foveation-based defense by cropping the input image using various schemes. Vuyyuru et al. (2020) proposed a defense that simulates the non-uniform sampling of the visual stimuli that occurs in the retina. They do this by resampling input images so that the sampling density of pixels is maximal at the point of fixation and decays progressively in regions further away from it. More recently, Gant et al. (2021) proposed a neural network modifies the textures in the image such that the modified image appears identical to human observers when viewed in the peripheral visual. They report improvement in adversarial robustness when DNNs are trained on images to which this transform has been applied. While we do not implement and evaluate their approach in this paper, we note that they report 40% accuracy against perturbation of ℓ_∞ norm 0.005 by a 5-step PGD attack, while we achieve an accuracy close to 50% at ℓ_∞ norm 0.008 computed by a 100-step APGD attack.

Although the accuracy of these biologically inspired approaches is higher than that of an unmodified model under adversarial attack, there is still a large gap in robustness between the proposed approaches and non-biological defenses, indicating that there is room for improvement.

5 Conclusion

In this paper, we propose a defense technique called R-Blur based on the biological foveation mechanism to make DNNs more robust to adversarial attacks. Adversarial attacks add human imperceptible perturbation to images to manipulate DNN's response. Since the existence of adversarial attacks presents a divergence between DNNs and humans, we ask if some aspect of human vision is fundamental to its robustness that is not modeled by DNNs. To this end, we propose a foveation technique R-Blur, that blurs the input image and reduces its color saturation adaptively based on the distance from a given fixation point. We evaluate R-Blur and other baseline models against 100-step APGD attacks on two datasets containing real world images. R-Blur outperforms other biologically inspired defenses by up to 50%. Moreover, for non-biological defenses, R-Blur closes the gap in robustness between biological defenses and non-biological defenses from around 70% to 30% for moderate perturbation distances ($||\epsilon||_\infty = 0.008$). Moreover, the robustness achieved by R-Blur is certifiable using the approach from (Cohen et al., 2019) and that the certified accuracy achieved by R-Blur is competitive with that achieved by randomized smoothing (Cohen et al., 2019) for small radii.

REFERENCES

Visual processing: Eye and retina (section 2, chapter 14) neuroscience online: An electronic textbook for the neurosciences: Department of neurobiology and anatomy - the university of texas medical school at houston. URL https://nba.uth.tmc.edu/neuroscience/m/s2/chapter14.html.

Naveed Akhtar, Ajmal Mian, Navid Kardan, and Mubarak Shah. Advances in adversarial attacks and defenses in computer vision: A survey. *IEEE Access*, 9:155161–155196, 2021.

Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. Recent advances in adversarial training for adversarial robustness. *arXiv preprint arXiv:2102.01356*, 2021.

Nicholas Carlini, Florian Tramer, J Zico Kolter, et al. (certified!!) adversarial robustness for free! *arXiv preprint arXiv:2206.10550*, 2022.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pp. 1310–1320. PMLR, 2019.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pp. 2206–2216. PMLR, 2020.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018.

Joel Dapello, Tiago Marques, Martin Schrimpf, Franziska Geiger, David Cox, and James J DiCarlo. Simulating a primary visual cortex at the front of cnns improves robustness to image perturbations. *Advances in Neural Information Processing Systems*, 33:13073–13087, 2020.

Arturo Deza and Talia Konkle. Emergent properties of foveated perceptual systems, 2021. URL https://openreview.net/forum?id=2_Z6MECjPEa.

Marc Fischer, Maximilian Baader, and Martin Vechev. Certified defense to image transformations via randomized smoothing. *Advances in Neural Information Processing Systems*, 33:8404–8417, 2020.

Jonathan M Gant, Andrzej Banburski, and Arturo Deza. Evaluating the adversarial robustness of a foveated texture transform module in a cnn. In *SVRHM 2021 Workshop@ NeurIPS*, 2021.

Thorsten Hansen, Lars Pracejus, and Karl R Gegenfurtner. Color perception in the intermediate periphery of the visual field. *Journal of vision*, 9(4):26–26, 2009.

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*, pp. 770–778, 2016.

Omisa Jinsi, Margaret M Henderson, and Michael J Tarr. Why is human vision so poor in early development? the impact of initial sensitivity to low spatial frequencies on visual category learning. *bioRxiv*, 2022.

MARY A Johnson. Color vision in the peripheral retina. *American journal of optometry and physiological optics*, 63(2):97–103, 1986.

Aditya Jonnalagadda, William Yang Wang, B.S. Manjunath, and Miguel Eckstein. Foveater: Foveated transformer for image classification, 2022. URL https://openreview.net/forum?id=mqIeP6qPvta.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). URL `http://www.cs.toronto.edu/~kriz/cifar.html`.

Aounon Kumar and Tom Goldstein. Center smoothing: Certified robustness for networks with structured outputs. *Advances in Neural Information Processing Systems*, 34:5560–5575, 2021.

Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin. Certified adversarial robustness with additive noise. *Advances in neural information processing systems*, 32, 2019.

Yan Luo, Xavier Boix, Gemma Roig, Tomaso Poggio, and Qi Zhao. Foveation-based mechanisms alleviate adversarial examples. *arXiv preprint arXiv:1511.06292*, 2015.

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

Johannes Mehrer, Courtney J Spoerer, Emer C Jones, Nikolaus Kriegeskorte, and Tim C Kietzmann. An ecologically motivated image dataset for deep learning yields better models of human vision. *Proceedings of the National Academy of Sciences*, 118(8):e2011417118, 2021.

Jan Hendrik Metzen, Tim Genewein, Volker Fischer, and Bastian Bischoff. On detecting adversarial perturbations. In *International Conference on Learning Representations*, 2017. URL `https://openreview.net/forum?id=SJzCSf9xg`.

Dylan M Paiton, Charles G Frye, Sheng Y Lundquist, Joel D Bowen, Ryan Zarcone, and Bruno A Olshausen. Selectivity and robustness of sparse coding networks. *Journal of vision*, 20(12):10–10, 2020.

RT Pramod, Harish Katti, and SP Arun. Human peripheral blur is optimal for object recognition. *arXiv preprint arXiv:1807.08476*, 2018.

Farzad Ramezani, Saeed Reza Kheradpisheh, Simon J Thorpe, and Masoud Ghodrati. Object categorization in visual periphery is modulated by delayed foveal noise. *Journal of Vision*, 19(9): 1–1, 2019.

Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Andrei Calian, Florian Stimberg, Olivia Wiles, and Timothy A Mann. Data augmentation can improve robustness. *Advances in Neural Information Processing Systems*, 34:29935–29948, 2021.

Hadi Salman, Mingjie Sun, Greg Yang, Ashish Kapoor, and J Zico Kolter. Denoised smoothing: A provable defense for pretrained classifiers. *Advances in Neural Information Processing Systems*, 33:21945–21957, 2020.

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.

Manish Reddy Vuyyuru, Andrzej Banburski, Nishka Pant, and Tomaso Poggio. Biologically inspired mechanisms for adversarial robustness. *Advances in Neural Information Processing Systems*, 33: 2135–2146, 2020.

Panqu Wang and Garrison W Cottrell. Central and peripheral vision for scene recognition: A neurocomputational modeling exploration. *Journal of vision*, 17(4):9–9, 2017.

Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations*, 2019.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *British Machine Vision Conference*, 2016.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pp. 7472–7482. PMLR, 2019.

6 Appendix

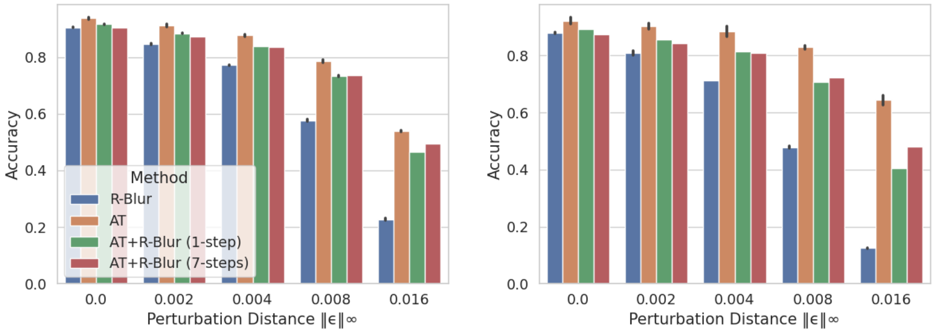
6.1 Comparison of Computation Costs

We evaluated the computational cost of the various methods used in this paper by measuring the time it takes for the model to process 1 batch of 128 224x224 Ecoset images during training and inference. The measurements are done on a machine with a 64-core Intel Xenon processor and Nvidia Tesla V100 GPU. The measured costs are presented in Table 2. We note that training cost of R-Blur is higher than R-Warp and similar to that of AT. The inference costs of R-Blur and R-Warp are higher than the other methods because the logits for 5 fixation points are computed and averaged for these methods. We believe that the computation cost of R-Blur can be reduced by a more efficient implementation that better leverages parallelism, which will be part of our future work.

Method	Training Cost (s/it)	Inference Cost (s/it)
ResNet	0.16	0.25
AT	0.31	0.25
R-Blur	0.38	1.5 (0.50)
G-Noise	0.16	0.25
R-Warp	0.20	0.625 (0.25)

Table 2: A comparison of the training and inference cost, measured as the average time taken to process one batch of 128 images, for the various configurations used in this paper. For R-Blur and R-Warp, the values in the parentheses represent the cost when only 1 fixation, and the values outside the parentheses represent the cost when 5 fixations are used.

6.2 Combining Adversarial Training and R-Blur



(a) CIFAR-10

(b) Ecoset-10

Figure 8: The accuracy of AT and trained on clean data, compared with adversarially trained R-Blur. We use two techniques for adversarially training R-Blur: 1-step fast adversarial training (Wong et al., 2019) and 7-step PGD training Madry et al. (2018).

While combining AT and R-Blur is not biologically plausible since natural stimuli normally do not contain subtle perturbations similar, nevertheless, for the sake of completeness, we have run experiments with this combination. We do not add Gaussian noise in R-Blur when training adversarially because the addition of random noise may reduce effectiveness of the adversarial perturbations used during training and thus lead to a less robust model. We use two techniques for adversarially training R-Blur: 1-step Fast Adversarial Training (FAT) (Wong et al., 2019) and 7-step PGD training Madry et al. (2018).

Figure 8 compares the performance of AT and R-Blur trained on clean data with adversarially trained R-Blur models on CIFAR-10 and Ecoset-10. We see that AT achieves significantly greater robustness than AT +R-Blur, indicating that R-Blur is reducing the effectiveness of AT. This may be because R-Blur is rendering adversarial perturbations being generated during training ineffective.

This hypothesis is supported by the fact that using 7-step PGD training with AT +R-Blur yields a more robust model, particularly on larger perturbation, than FAT. However, the improvement in robustness is nominal and comes at the cost of increasing computations by 7 fold. Investigating the reasons behind the incompatibility of AT and R-Blur, and developing techniques for adversarially training R-Blur are promising directions for future work.

6.3 Evaluations With Different Architectures

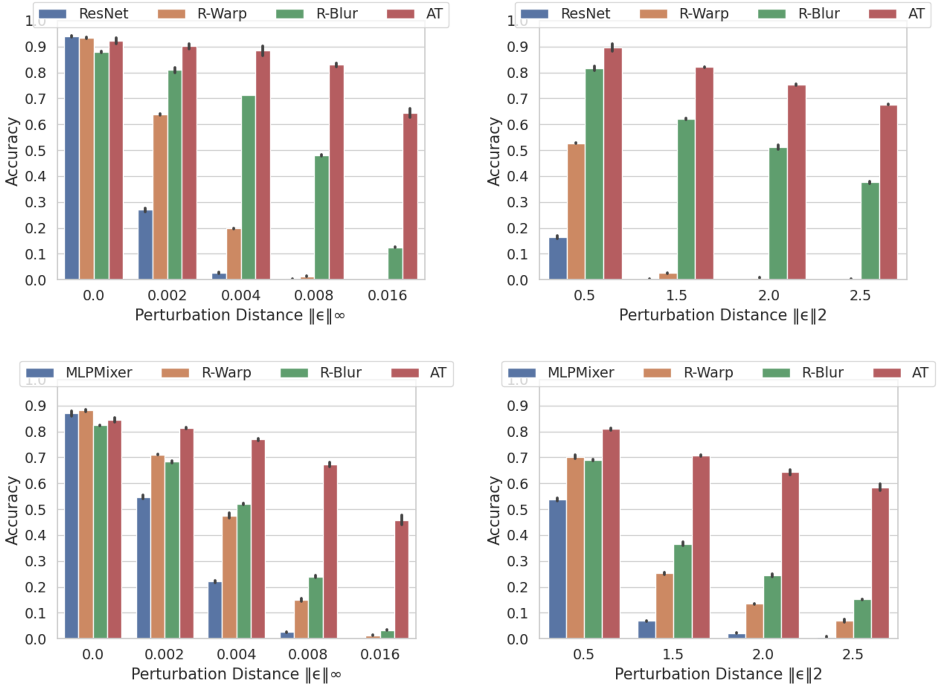


Figure 9: The accuracy of ResNets (top) and MLP-Mixers (bottom) against adversarial perturbations of various ℓ_∞ (left) and ℓ_2 (right) norms.

To demonstrate that the benefits of R-Blur are not limited to a CNNs, we trained MLP-Mixers (Tolstikhin et al., 2021) with R-Blur preprocessing and evaluated their robustness. We use the configuration of MLP-Mixer referred to as S16 in (Tolstikhin et al., 2021) and we train it with a batch size of 128 for 60 epochs using the Adam optimizer. The learning rate of the optimizer is linearly increased to 0.001 over 12 epochs and is decayed linearly to almost zero over the remaining epochs. The results are shown in Figure 9.

We observe that R-Blur significantly improves the robustness of MLP-Mixer models, and achieves greater accuracy than R-Warp at higher levels of perturbations. These results show that the robustness endowed to ResNets by R-Blur was not a chance occurence, and they further strengthen our claim that loss in fidelity due to foveation contributes to the robustness of human and computer vision.