Evaluating the Adversarial Robustness of a Foveated Texture Transform Module in a CNN

Anonymous Author(s)
Affiliation
Address
email

Abstract

Biologically inspired mechanisms such as foveation and multiple fixation points have previously been shown to help alleviate adversarial examples (Reddy et al., 2020). By mimicking the effects of visual crowding present in human vision, foveated, texture-based computations may provide another route for increasing adversarial robustness. Previous statistical models of texture rendering (Portilla & Simoncelli, 2000; Gatys et al., 2015) paved the way for the development of a Foveated Texture Transform (FTT) module which utilizes localized texture synthesis in foveated receptive fields (Deza et al., 2017). The FTT module was added to a VGG-11 CNN architecture and ten random initializations were trained on 20-class subsets of the Places and EcoSet datasets for scene and object classification respectively. The trained networks were attacked using Projected Gradient Descent (PGD) and the adversarial accuracies were calculated at multiple epochs to evaluate changes in robustness as the networks trained. The results indicate that the FTT module significantly improved adversarial robustness for scene classification, especially when the validation loss was at a minimum. However, the FTT module did not provide a statistically significant increase in adversarial robustness for object classification. Furthermore, we do not find a trade-off between accuracy and robustness (Tsipras et al., 2018) for the FTT module suggesting a benefit of using foveated, texture-based distortions in the latent space during learning compared to non-perturbed latent space representations. Finally, we investigate the nature of latent space distortions with additional controls that probe other directions in the latent space that are not texture-based.

1 Introduction

Investigating the representational advantage of visual crowding is a theme of on-going work (Rosenholtz et al., 2019). While there are a handful of models that simulate crowding via texture-based summary synthesis models (Freeman & Simoncelli, 2011; Rosenholtz et al., 2012; Deza et al., 2017; Wallis et al., 2019), there is still no general consensus to understand if this spatially-adaptive computation serves a representational goal – or if it is a limitation of the human visual system – though some efforts have been made to find an answer (Cheung et al., 2016; Deza & Konkle, 2020; Reddy et al., 2020; Han et al., 2020). Indeed, crowding is a phenomena that is mainly noticeable in the periphery (away from the fovea) of the visual field, where image features are pooled together giving rise to a locally "crowded" representation of the visual stimulus (Figure 1C).

In an effort to investigate the representational goal of crowding, we designed and tested a fully-differentiable, foveated visual system that performs localized, texture-based computations (Balas et al., 2009; Wallis et al., 2016, 2019; Vacher et al., 2020; Ziemba & Simoncelli, 2021) – the same that give rise to the effects of crowding – and test its general accuracy and adversarial robustness on

Submitted to 3rd Workshop on Shared Visual Representations in Human and Machine Intelligence (SVRHM 2021) of the Neural Information Processing Systems (NeurIPS) conference.

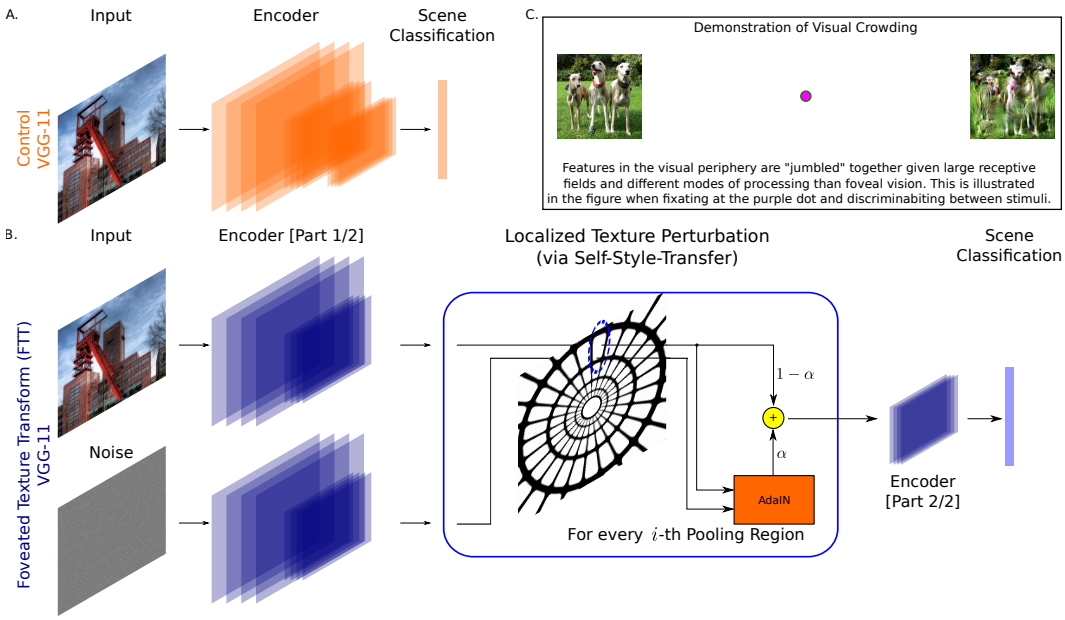

Figure 1: A schematic of the two main networks used in our experiments. We use a regular VGG-11 as our control system (A.) that performs an image classification task; and a foveated, texture-based VGG-11 where foveation happens in the latent space of the network (B.). (C.) Example of the visual crowding phenomena generated through foveated, texture-based synthesis models (Freeman & Simoncelli, 2011). Our transform approximates this computation in the latent space of a VGG-11.

a scene and object classification task. Our hypothesis is that foveated, texture-based systems will have a lower baseline accuracy than a non-foveated system but greater adversarial robustness, thus potentially inferring a representational goal of crowding. We will empirically show that crowding simulated through this texture-based perturbation – that is implemented with the aid of a stylization operator on Gaussian noise in the latent space (Huang & Belongie, 2017) – regularizes the learned representation of the image leading to an increase in adversarial robustness contingent on the dataset (scenes vs. objects).

2 Design of the Foveated Texture Transform (FTT)

Current models of texture-based foveation are two-stage such as those introduced in Deza & Konkle (2020). While these models succeed at probing the emergent representations of a particular type of computation such as adaptive Gaussian blurring that simulate a loss of visual acuity, or adaptive texture-based distortions that simulate crowding; a limitation – or advantage depending on the hypothesis being tested – is that they are not end-to-end differentiable by design. Within the context of evaluating adversarial robustness, this is a limitation and only few groups have succeeded in testing foveated models (Luo et al., 2016; Reddy et al., 2020; Jonnalagadda et al., 2021). These non-differentiable two-stage models are generally defined by a *fixed* first-stage transform $f(\circ)$, and a *learnable* second-stage transform $g(\circ)$ such that the general perceptual system S that receives an image I is defined as:

$$S = g(f(I)). \tag{1}$$

where f is an image-to-image transform as done in Deza & Konkle (2020) resembling the form $f = \mathcal{D}(\mathcal{E}_\Sigma(\circ))$ with the foveated encoder $\mathcal{E}_\Sigma$; and decoder $(\mathcal{D} \sim \mathcal{E}^{-1})$. Critically, in this paper we will remove $\mathcal{D}$ and make S two-stage by partitioning the VGG-11 encoder in two parts $(f' : \mathcal{E}^{(1)}, g' : \mathcal{E}^{(2)})$ and embedding the foveation transform $(\circ_\Sigma)$ in the latent space between these two stages. Thus our foveated network takes the following shape (similar to Eq. 1) $S = \mathcal{E}^{(2)}(\mathcal{E}^{(1)}_\Sigma(\circ))$. By de-coupling the foveated stylization function $(\mathcal{S}_\Sigma)$ from the encoder we can now define the VGG-11 Foveated Texture Transform (FTT) as end-to-end differentiable where the encoders $(\mathcal{E}^{(1)}, \mathcal{E}^{(2)})$ are learnable

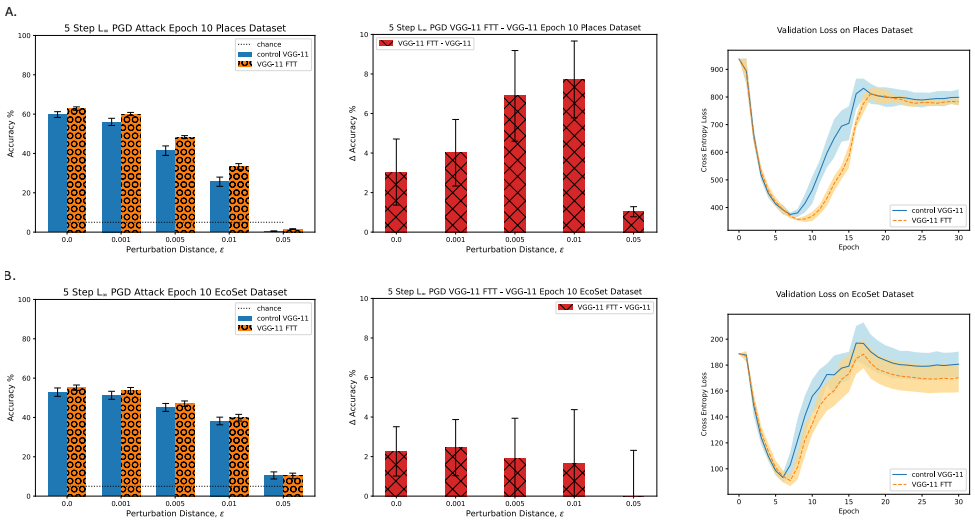

Figure 2: Our results suggest that VGG-11 FTT network shows greater robustness to L_∞ PGD attacks compared to the control VGG-11 network – however this result is only noticeable for scene (A.) but not object (B.) classification. Notably the observed increase in adversarial robustness for scene classification is highest around epoch 10 when the validation loss of the model is near a minimum (A.). Additional accuracy-robustness tradeoff curves can be found in Appendix B.

via:

$$S = \mathcal{E}^{(2)}(\mathcal{S}_\Sigma(\mathcal{E}^{(1)}(I), \mathcal{E}^{(1)}(N))). \tag{2}$$

It is worth noting that the stylization function we use $\mathcal{S}(n, m)$ was introduced in Huang & Belongie (2017) and consists of aligning the channel-wise mean and standard deviation of the hidden feature activations in a deep neural network between a content image m and style image n and can also be used as a feed-forward texture synthesis model when the content image is noise. Its foveated version ($\mathcal{S}_\Sigma$) was introduced in Deza et al. (2017) that consists of the same operation applied over each log-polar receptive field in the latent space. A diagram comparing our VGG-11 FTT model to the control VGG-11 can be seen in Figure 1.

3 Implications of FTT on Adversarial Robustness

Both the control VGG-11 and VGG-11 FTT networks were trained for 60 epochs on a 20 class subset of the Places (Zhou et al., 2017) and EcoSet (Mehrer et al., 2021) datasets for scene and object classification respectively. Ten unique, random initializations of each network were trained to provide a large enough sample size for statistical analyses of the results. Appendix A contains more information on the training hyperparameters used.

After training, 5-step untargeted L_∞ PGD attacks (Carlini & Wagner, 2017) were applied to the control VGG-11 and VGG-11 FTT networks on both datasets at epochs 10, 20, and 60 across a range of $\epsilon \in \{0, 0.001, 0.005, 0.01, 0.05\}$ with a step size of $\frac{\epsilon}{3}$ using the Foolbox Python library (Rauber et al., 2017). As seen in Figure 2A, for scene classification the VGG-11 FTT network has a higher baseline accuracy and adversarial accuracies than the control VGG-11 network. The difference in the network performances is largest at $\epsilon = 0.01$ for the PGD attacks performed at epoch 10. Importantly, at epoch 10 the validation loss for both networks is at a minimum. At higher epochs, the differential advantage of the FTT module decreases but is still apparent (Figure 7; Appendix). On the other hand, for object classification the FTT module does not provide an appreciable increase in adversarial accuracy across all epochs investigated (Figure 2B, Figure 8; Appendix).

4 Contributions of FTT to Image Representation

In addition to the FTT, another module with a similar transform to the FTT was developed to determine whether stochasticity or texture is the primary contributing factor to the increase in

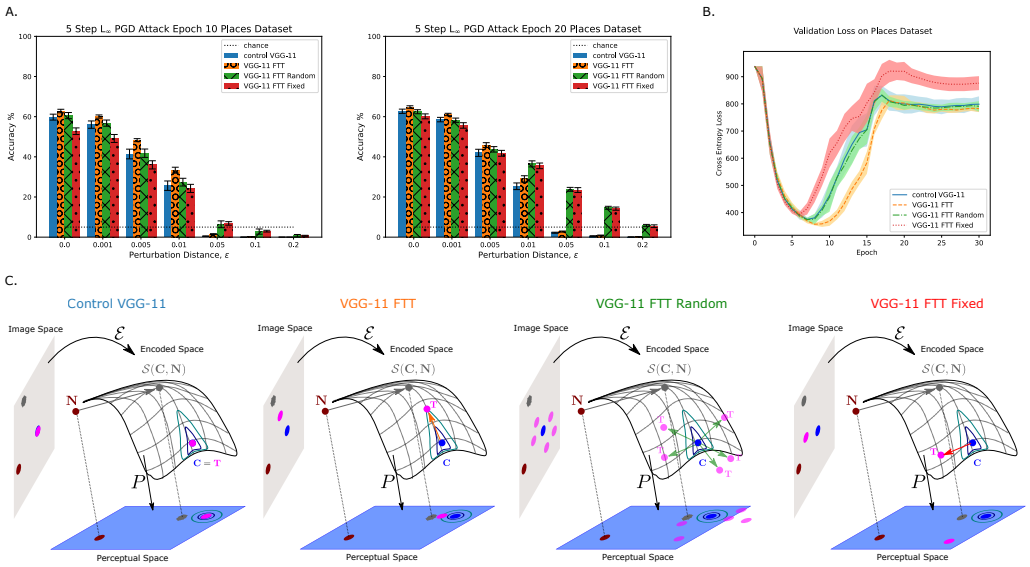

Figure 3: (A.) L_∞ PGD Attack results for VGG-11 FTT Random and Fixed networks. (B.) Cross-entropy validation loss values for all networks up to epoch 30. (C.) Visualization of the directions of the perturbation vectors for each network.

adversarial robustness (Dapello et al., 2020; Berardino et al., 2017). In this transformation the difference of the stylized noise feature vector with the content feature vector is shuffled to maintain the magnitude of the transformation while making the perturbation direction random. The (i-th) receptive field or pooling region in the FTT is defined by:

$$T_i(I, N, \alpha_i) = (1 - \alpha_i)\mathcal{E}_i^{(1)}(I) + \alpha_i \mathcal{S}\left(\mathcal{E}_i^{(1)}(I), \mathcal{E}_i^{(1)}(N)\right). \tag{3}$$

Reformulating the above equation as:

$$T_i(I, N, \alpha_i) = \underbrace{\mathcal{E}_i^{(1)}(I)}_{\text{Initial Encoding}} + \underbrace{\alpha_i \left(\mathcal{S}\left(\mathcal{E}_i^{(1)}(I), \mathcal{E}_i^{(1)}(N)\right) - \mathcal{E}_i^{(1)}(I)\right)}_{\text{Perturbation Vector}} \tag{4}$$

highlights the interpretation of the FTT as a perturbation to the encoded content in a direction defined by the difference of the stylized noise (texture-equivalent) and encoded content. The magnitude of this perturbation for each receptive field is controlled by α_i. Let

$$D(I, N) = \mathcal{S}\left(\mathcal{E}_i^{(1)}(I), \mathcal{E}_i^{(1)}(N)\right) - \mathcal{E}_i^{(1)}(I) \tag{5}$$

and recognize that $D(I, N)$ is defined by a set of matrices that can be uniquely defined by an index set $J = \{j \in \mathbb{N} | d_j \in D(I, N)\}$. Hence, a new direction with the same magnitude as $D(I, N)$ can be defined by permuting the index set and thus shuffling the elements of $D(I, N)$. The permutation of the index set can be randomly chosen each time the network performs a forward pass (Figure 3C: VGG-11 FTT Random) or be fixed for a given network (Figure 3C: VGG-11 FTT Fixed).

Ten unique, random initializations of the VGG-11 FTT Random and Fixed networks were trained with the same hyperparameters used to train the control VGG-11 and VGG-11 FTT networks. After training, 5-step L_∞ PGD attacks were performed on the networks and the adversarial accuracies reported in Figure 3A. At epoch 20, the performance of the VGG-11 FTT Random network is comparable to the control VGG-11 at low perturbation distances and retains higher accuracies as epsilon increases when compared to the control VGG-11 and VGG-11 FTT networks. However, at epoch 10 the VGG-11 FTT Random network closely matches the accuracy of the control VGG-11 network and performs significantly worse than the VGG-11 FTT network. The VGG-11 FTT Fixed network has a marginally lower accuracy than the VGG-11 FTT Random network at epoch 20 and this difference is larger at epoch 10. Interestingly, the validation loss of all of the models follow a

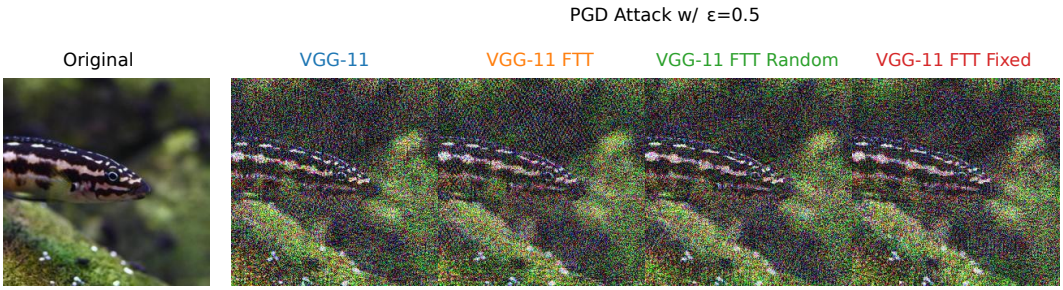

Figure 4: Visualization of the adversarial examples for the VGG-11, VGG-11 FTT, VGG-11 FTT Random, VGG-11 FTT Fixed networks at epoch 10 for an exaggerated $\epsilon = 0.5$ on an image of an aquarium from the Places dataset.

similar behavior as seen in Figure 3B. Hence, the VGG-11 FTT Random network performs better when the validation loss is *not* at a minimum in contrast to the VGG-11 FTT network.

Finally, we inspect if the direction of the perturbation in the latent space causes perceptual differences in the adversarial perturbations for each network. We can observe in Figure 4 that there are not any significant differences in terms of the visual structure of the PGD attacks across neural network models.

5 Discussion

The results indicate that the FTT significantly increased adversarial robustness for scene classification across all epochs investigated (Table 1, Table 2; Appendix B). In particular the FTT showed the largest advantage for scene classification when the validation loss was near a minimum around epoch 10. Additionally, there is no trade-off between accuracy and robustness (Tsipras et al., 2018) when the FTT module is added to the VGG-11 network for both scene and object classification. The lack of trade-off showcases a benefit of using foveated, texture-based distortions in the latent space instead of non-perturbed latent space representations during learning.

In contrast to the performance of the FTT on scene classification, the FTT did not provide a statistically significant increase in adversarial robustness for object classification (Table 3, Table 4; Appendix B). The difference in performance between scene and object classification could be due to the lack of features in the periphery in object classification images. Additionally, scenes tend to be in natural settings and contain repetitive textures while objects are usually made by humans and lack repeating elements. The lack of repeating elements results in a poor representation when texture-based computations are performed as shown in Gatys et al. (2015). Previous psychophysical work has also shown that humans can recognize scenes accurately with texture-based representation (Renninger & Malik, 2004), whereas the same has not been shown for objects, suggesting that texture-based representations are able to encapsulate the information necessary for scene but not object classification.

Moreover, the observed reduction in adversarial robustness at higher epochs for the VGG-11 FTT network is likely a result of overfitting. As the network trains, the decision boundaries in the latent space are expected to become increasingly complex and more closely fit the training data. Thus, an adversarial attack would be more likely to succeed on a network that is overfitting the data, since a small perturbation to the latent space representation of the image has a higher probability of crossing the decision boundary (Fawzi et al., 2017; Shamir et al., 2021). On the Places dataset, the VGG-11 FTT Random and Fixed networks do not suffer from this issue and in fact have higher adversarial robustness at higher epochs (Figure 3A, Table 2; Appendix B). The permutation of the perturbation direction in the latent space (Figure 3C) helps prevent overfitting since the randomness of the direction forces the decision boundaries to be a looser, more generalizable fit of the data. Notably the difference in adversarial robustness for the VGG-11 Random and Fixed network is negligible suggesting the permutation itself and not the randomness of the permutation is the primary contributing factor to the increase in adversarial robustness. Overall, this suggests that different types of latent-space perturbations (texture and non-texture based) may implicitly act as regularizers of learned visual representations.

References

Benjamin Balas, Lisa Nakano, and Ruth Rosenholtz. A summary-statistic representation in peripheral vision explains visual crowding. *Journal of vision*, 9(12):13–13, 2009.

Alexander Berardino, Johannes Ballé, Valero Laparra, and Eero P Simoncelli. Eigen-distortions of hierarchical representations. *arXiv preprint arXiv:1710.02266*, 2017.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. *arXiv preprint arXiv:1608.04644*, 2017.

Brian Cheung, Eric Weiss, and Bruno Olshausen. Emergence of foveal image sampling from learning to attend in visual scenes. *arXiv preprint arXiv:1611.09430*, 2016.

Joel Dapello, Tiago Marques, Martin Schrimpf, Franziska Geiger, David D Cox, and James J DiCarlo. Simulating a primary visual cortex at the front of cnns improves robustness to image perturbations. *BioRxiv*, 2020.

Arturo Deza and Talia Konkle. Emergent properties of foveated perceptual systems. *arXiv preprint arXiv:2006.07991*, 2020.

Arturo Deza, Aditya Jonnalagadda, and Miguel Eckstein. Towards metamerism via foveated style transfer. *arXiv preprint arXiv:1705.10041*, 2017.

Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, Pascal Frossard, and Stefano Soatto. Classification regions of deep neural networks. *arXiv preprint arXiv:1705.09552*, 2017.

Jeremy Freeman and Eero P Simoncelli. Metamers of the ventral stream. *Nature neuroscience*, 14(9): 1195–1201, 2011.

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. *arXiv preprint arXiv:1505.07376*, 2015.

Yena Han, Gemma Roig, Gad Geiger, and Tomaso Poggio. Scale and translation-invariance for novel objects in human vision. *Scientific reports*, 10(1):1–13, 2020.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1501–1510, 2017.

Aditya Jonnalagadda, William Wang, and Miguel P Eckstein. Foveater: Foveated transformer for image classification. *arXiv preprint arXiv:2105.14173*, 2021.

Yan Luo, Xavier Boix, Gemma Roig, Tomaso Poggio, and Qi Zhao. Foveation-based mechanisms alleviate adversarial examples. *arXiv preprint arXiv:1511.06292*, 2016.

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), 2021.

Javier Portilla and Eero P Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *International journal of computer vision*, 40(1):49–70, 2000.

Jonas Rauber, Wieland Brendel, and Matthias Bethge. Foolbox: A python toolbox to benchmark the robustness of machine learning models. In *Reliable Machine Learning in the Wild Workshop, 34th International Conference on Machine Learning*, 2017. URL `http://arxiv.org/abs/1707.04131`.

Manish V Reddy, Andrzej Banburski, Nishka Pant, and Tomaso Poggio. Biologically inspired mechanisms for adversarial robustness. *arXiv preprint arXiv:2006.16427*, 2020.

Laura Walker Renninger and Jitendra Malik. When is scene identification just texture recognition? *Vision research*, 44(19):2301–2311, 2004.

Ruth Rosenholtz, Jie Huang, Alvin Raj, Benjamin J Balas, and Livia Ilie. A summary statistic representation in peripheral vision explains visual search. *Journal of vision*, 12(4):14–14, 2012.

Ruth Rosenholtz, Dian Yu, and Shaiyan Keshvari. Challenges to pooling models of crowding: Implications for visual mechanisms. *Journal of vision*, 19(7):15–15, 2019.

Adi Shamir, Odelia Melamed, and Oriel BenShmuel. The dimpled manifold model of adversarial examples in machine learning. *arXiv preprint arXiv:2106.10151*, 2021.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

Jonathan Vacher, Aida Davila, Adam Kohn, and Ruben Coen-Cagli. Texture interpolation for probing visual perception. *arXiv preprint arXiv:2006.03698*, 2020.

Thomas SA Wallis, Matthias Bethge, and Felix A Wichmann. Testing models of peripheral encoding using metamerism in an oddity paradigm. *Journal of vision*, 16(2):4–4, 2016.

Thomas SA Wallis, Christina M Funke, Alexander S Ecker, Leon A Gatys, Felix A Wichmann, and Matthias Bethge. Image content is more important than bouma's law for scene metamers. *ELife*, 8: e42512, 2019.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

Corey M Ziemba and Eero P Simoncelli. Opposing effects of selectivity and invariance in peripheral vision. *Nature Communications*, 12(1):1–11, 2021.

A Training Dynamics and Convergence of Loss Function

The networks were trained using Stochastic Gradient Descent (SGD) with a weight decay of 5×10^{-4} and momentum of 0.9. To ensure convergence of the loss function, a learning rate scheduler was used. The initial learning rate of 10^{-3} was set to 5×10^{-4} at epoch 15 and 2.5×10^{-4} at epoch 30. A batch size of 16 was used, the images were resized to 256x256 and color normalized. The number of training and validation images were the same for both datasets to ensure consistency in the training dynamics.

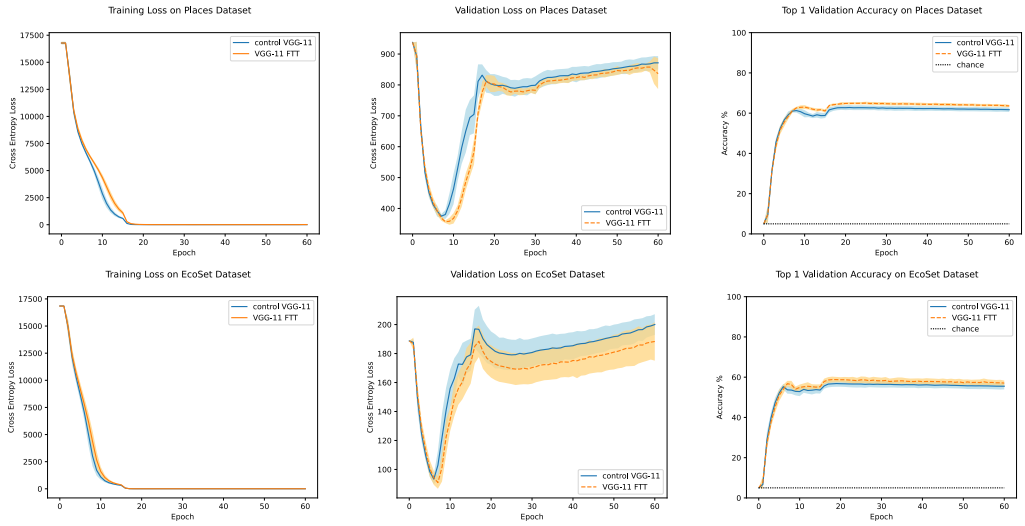

Figure 5: Cross entropy training/validation loss and accuracy as a function of epoch for both VGG-11 and VGG-11 FTT networks trained on subsets of the Places (top) and EcoSet (bottom) datasets.

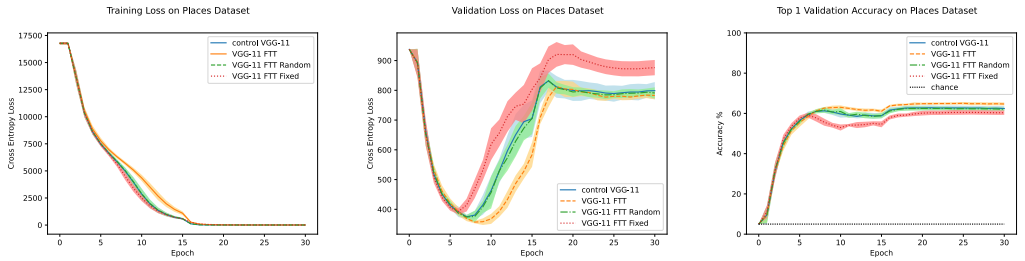

Figure 6: Cross entropy training/validation loss and accuracy as a function of epoch (up to epoch 30) for control VGG-11, VGG-11 FTT, VGG-11 FTT Random, VGG-11 FTT Fixed networks trained on subsets of the Places dataset.

B Adversarial Robustness as a Function of Learning Dynamics

Epoch	$\epsilon = 0$	$\epsilon = 0.001$	$\epsilon = 0.005$	$\epsilon = 0.01$	$\epsilon = 0.05$
10	2.6×10^{-5}	5.6×10^{-6}	8.5×10^{-8}	5.9×10^{-8}	1.1×10^{-10}
20	4.2×10^{-5}	4.4×10^{-6}	2.2×10^{-5}	7.6×10^{-5}	9.1×10^{-5}
30	1.6×10^{-5}	4.0×10^{-6}	3.7×10^{-6}	4.9×10^{-5}	1.2×10^{-5}
60	3.4×10^{-8}	6.6×10^{-7}	3.4×10^{-5}	1.6×10^{-5}	9.8×10^{-3}

Table 1: p-values from two-tailed t-test on the control VGG-11 and VGG-11 FTT networks accuracies with respect to perturbation distance ϵ and epoch number for the Places dataset.

Epoch	VGG-11	VGG-11 FTT	VGG-11 FTT Random	VGG-11 FTT Fixed
10	0.9572	1.227	1.508	1.422
20	1.081	1.243	3.656	3.549
30	1.046	1.190	3.569	3.479
60	0.9616	1.050	NA	NA

Table 2: Area under the curve estimates for accuracies with respect to perturbation distance ϵ and epoch number for the Places dataset for control VGG-11, VGG-11 FTT, VGG-11 FTT Random, and VGG-11 FTT Fixed.

Epoch	$\epsilon = 0$	$\epsilon = 0.001$	$\epsilon = 0.005$	$\epsilon = 0.01$	$\epsilon = 0.05$
10	1.7×10^{-2}	9.6×10^{-3}	2.7×10^{-2}	4.9×10^{-2}	7.5×10^{-1}
20	1.1×10^{-2}	1.1×10^{-2}	5.3×10^{-3}	5.9×10^{-2}	9.8×10^{-1}
60	7.4×10^{-3}	2.2×10^{-2}	6.7×10^{-2}	1.2×10^{-1}	6.4×10^{-1}

Table 3: p-values from two-tailed t-test on the control VGG-11 and VGG-11 FTT networks accuracies with respect to perturbation distance ϵ and epoch number for the Ecoset dataset.

Epoch	VGG-11	VGG-11 FTT
10	1.422	1.474
20	1.644	1.695
60	1.642	1.673

Table 4: Area under the curve estimates for control VGG-11 and VGG-11 FTT accuracies with respect to perturbation distance ϵ and epoch number for the Ecoset dataset.

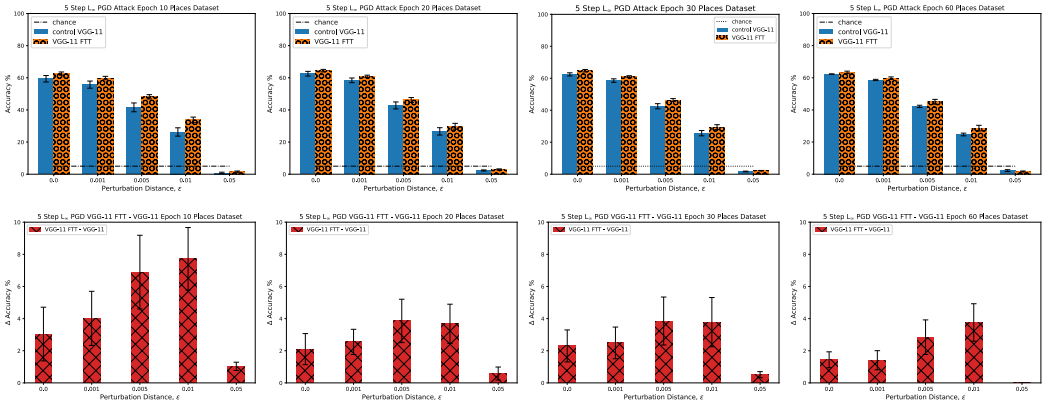

Figure 7: 5-step L_∞ PGD attacks on control VGG-11 and VGG-11 FTT for the Places dataset (scene classification). Top row shows the performance of both models and bottom row shows the difference in accuracy across multiple epochs (10, 20, 30 and 60) to illustrate the robustness of the results independent of learning dynamics.

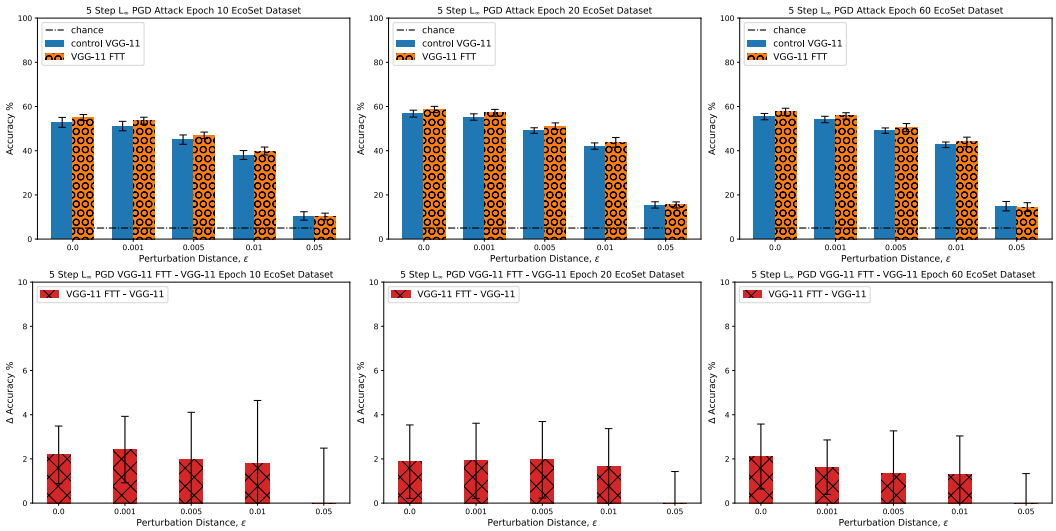

Figure 8: 5-step L_∞ PGD attacks on control VGG-11 and VGG-11 FTT networks for the EcoSet dataset (object classification). Top row shows the performance of both models and bottom row shows the difference in accuracy.