# Sensitivity of Deep Convolutional Networks to Gabor Noise

Kenneth T. Co [1] Luis Muñoz-González [1] Emil C. Lupu [1]

**Abstract**

Deep Convolutional Networks (DCNs) have been shown to be sensitive to Universal Adversarial Perturbations (UAPs): input-agnostic perturbations that fool a model on large portions of a dataset. These UAPs exhibit interesting visual patterns, but this phenomena is, as yet, poorly understood. Our work shows that visually similar procedural noise patterns also act as UAPs. In particular, we demonstrate that different DCN architectures are sensitive to Gabor noise patterns. This behaviour, its causes, and implications deserve further in-depth study.

## 1. Introduction

Deep Convolutional Networks (DCNs) have enabled deep learning to become one the primary tools for computer vision tasks. However, adversarial examples–slightly altered inputs that change the model's output–have raised concerns on their reliability and security. Adversarial perturbations can be defined as the noise patterns added to natural inputs to generate adversarial examples. Some of these perturbations are *universal*, i.e. the same pattern can be used to fool the classifier on a large fraction of the tested dataset (Moosavi-Dezfooli et al., 2017; Khrulkov & Oseledets, 2018). As shown in Fig. 1, it is interesting to observe that such Universal Adversarial Perturbations (UAPs) for DCNs contain structure in their noise patterns.

Results from (Co et al., 2018) together with our results here suggest that DCNs are sensitive to procedural noise perturbations, and more specifically here to Gabor noise. Existing UAPs have some visual similarities with Gabor noise as in Figure 2. Convolutional layers induce a prior on DCNs to learn local spatial information (Goodfellow et al., 2016), and DCNs trained on natural image datasets, such as ImageNet, learn convolution filters that are similar

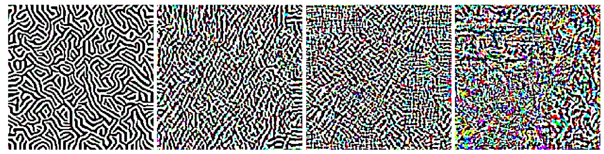

*Figure 1.* UAPs generated for VGG-19 targeting specific layers using singular vector method (Khrulkov & Oseledets, 2018).

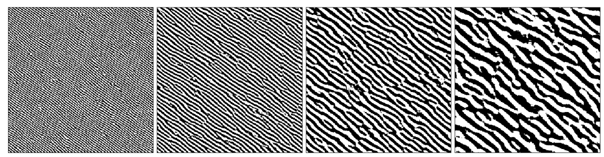

*Figure 2.* Gabor noise with normalized variance spectrums (Neyret & Heitz, 2016) and decreasing frequency from left to right.

in appearance to Gabor kernels and colour blobs (Yosinski et al., 2014; Olah et al., 2017). Gabor noise is a convolution between a Gabor kernel[2] and a sparse white noise. Thus, we hypothesize that DCNs are sensitive to Gabor noise, as it exploits specific features learned by the convolutional filters.

In this paper we demonstrate the sensitivity of 3 different DCN architectures (*Inception v3*, *ResNet-50*, and *VGG-19*), to Gabor noise on the ImageNet image classification task. We empirically observed that even random Gabor noise patterns can be effective to generate UAPs. Understanding this behaviour is important, as the generation and injection of Gabor noise is computationally inexpensive and, therefore, can become a threat to the security and reliability of DCNs.

## 2. Background

Compared to standard adversarial examples, UAPs reveal more general features that the DCN is sensitive to. In contrast, adversarial perturbations generated for specific inputs, though less detectable in many cases, can "overfit" and evade only on inputs they were generated for (Zhou et al., 2018). Previous approaches to generate UAPs use knowledge of the model's learned parameters. Moosavi-Dezfooli et al. (2017) use the DeepFool algorithm (Moosavi-Dezfooli et al., 2016) iteratively over a set of images to construct a

[1]Department of Computing, Imperial College London, United Kingdom. Contact: Kenneth T. Co <k.co@imperial.ac.uk>, Luis Muñoz-González <l.munoz@imperial.ac.uk>, Emil C. Lupu <e.c.lupu@imperial.ac.uk>.

*Accepted to the Workshop on Identifying and Understanding Deep Learning Phenomena within the $36^{th}$ International Conference on Machine Learning*, Long Beach, California. 

[2]A kernel (or filter) in image processing refers to a mask or small matrix used for image convolution.

UAP. A different approach is proposed in (Mopuri et al., 2018), where UAPs are computed using Generative Adversarial Nets (GANs).

Khrulkov & Oseledets (2018) proposed the singular vector method to generate UAPs targeting specific layers of DCNs, learning a perturbation $s$ that maximises the $L_p$-norm of the differences in the activations for that specific layer, $f_i$:

$$\arg\max_s \|f_i(x) - f_i(x+s)\|_p, \quad \|s\|_q = \varepsilon$$

where the $L_q$-norm of $s$ is constrained to $\varepsilon$. This can approximated using the Jacobian for that layer:

$$\|f_i(x) - f_i(x+s)\|_p \approx \|J_i(x) \cdot s\|_p.$$

The solution $s$ that maximizes this is the $(p, q)$-singular vector can be computed with the power method (Boyd, 1974). Then, $s$ is effective to generate UAPs targeting a specific layer in the DCN. The solutions obtained with this method for the first layers of DCNs (see Fig. 1) resemble the Gabor noise patterns shown in Fig. 2.

However none of these works highlight the interesting visual patterns that manifest from these UAPs. In contrast, we show that procedural noise can generate UAPs targeting DCNs in a systematic and efficient way.

## 3. Gabor Noise

Gabor noise is the convolution of a sparse white noise and a Gabor kernel, making it a type of *Sparse Convolution Noise* (Lagae et al., 2009; 2010). The Gabor kernel $g$ with parameters $\{\kappa, \sigma, \lambda, \omega\}$ is the product of a circular Gaussian and a harmonic function

$$g(x, y) = \kappa e^{-\pi\sigma^2(x^2+y^2)} \cos\left[2\pi\lambda(x\cos\omega + y\sin\omega)\right]$$

where $\kappa$ and $\sigma$ are the magnitude and width of the Gaussian, and $\lambda$ and $\omega$ are the frequency and orientation of the Harmonic (Lagae et al., 2010). The value of the Gabor noise at point $(x, y)$ is given by

$$G(x, y) = \sum_i w_i\, g(x - x_i, y - y_i; \kappa_i, \sigma_i, \lambda_i, \omega_i)$$

where $(x_i, y_i)$ are the coordinates of sparse random points and $w_i$ are random weights.

Gabor noise is an expressive noise function and has exponentially many parameterizations to explore. To simplify the analysis, we choose anisotropic Gabor noise, where the Gabor kernel parameters and weights are the same for each $i$. This results in noise patterns that have uniform orientation and thickness. We also normalize the variance spectrum of the Gabor noise using the algorithm in (Neyret & Heitz, 2016) to achieve min-max oscillations within the pattern.

## 4. Experiments

For our experiments we use the validation set from the ILSVRC2012 ImageNet image classification task (Russakovsky et al., 2015) with 1,000 distinct categories. We use 3 pre-trained ImageNet DCN architectures from `keras.applications`: Inception v3 (Szegedy et al., 2016), ResNet-50 (He et al., 2016), and VGG-19 (Simonyan & Zisserman, 2014).

Inception v3 take input images with dimensions $299 \times 299 \times 3$ while the other two networks take images with dimensions $224 \times 224 \times 3$. The kernel size $\kappa = 23$ is fixed so that the Gabor kernels will fill the entire image regardless of the distribution of points. The number of points $i$ distributed will be proportional to the image dimensions, which is independent of the Gabor kernel parameters. The resulting Gabor noise parameters we control are $\Theta = \{\sigma, \omega, \lambda\}$. We test the sensitivity of the models with 1,000 random Gabor noise perturbations generated from uniformly drawn parameters $\Theta$ with $\sigma, \lambda \in [1.5, 9]$ and $\omega \in [0, \pi]$.

We evaluate our Gabor noise on 5,000 random images from the validation set with an $\ell_\infty$ norm constraint of $\varepsilon = 12$ on the noise. The choice of $\frac{12}{256} \approx 0.047$ is consistent with other attacks on ImageNet-scale models with less than 5% perturbation magnitude. To provide a baseline, we also measure the sensitivity of the models to 1,000 uniform random noise perturbations from $\{-\varepsilon, \varepsilon\}^{D\times D\times 3}$ where $D$ is the image's side length. This is useful for showing that the sensitivity to Gabor noise is not trivial.

### 4.1. Metrics

Given model output $f$, input $x \in X$, perturbation $s$, and small $\varepsilon > 0$, we define the **universal sensitivity** of a model on perturbation $s$ over $X$ as

$$\frac{1}{|X|}\sum_{x\in X} \|f(x) - f(x+s)\|_\infty, \quad \|s\|_\infty = \varepsilon.$$

The norm constraint on $s$ ensures that the perturbation is small. For this paper, we choose $\infty$-norm as it is straightforward to impose for Gabor noise perturbations and is often used in the adversarial machine learning literature. For classification tasks, it is also useful to consider the **universal evasion** rate of a perturbation $s$ over $X$

$$\frac{|\{x \in X : \arg\max f(x) \neq \arg\max f(x+s)\}|}{|X|}.$$

This corresponds to the definition that an adversarial perturbation is a small change that alters the predicted output label. Note that we are not interested in the ground truth labels for $x$ or $x+s$. We focus instead on how small changes to the input result in large changes to the model's *original predictions*.

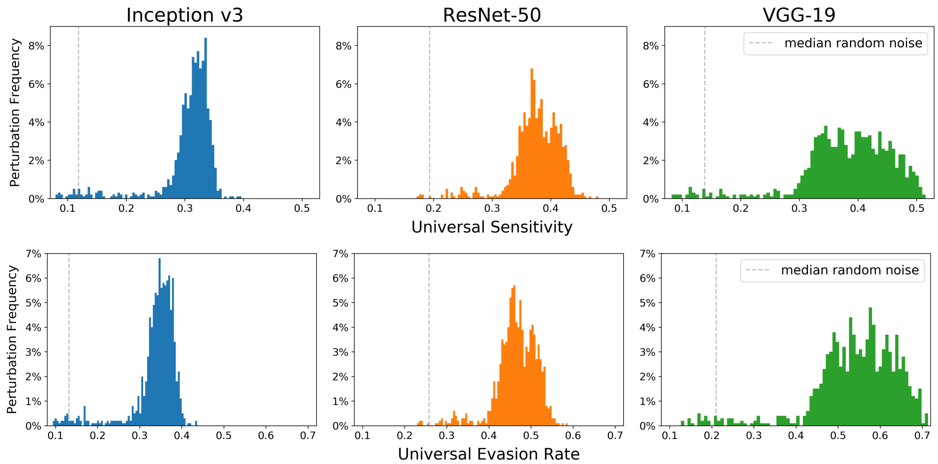


*Figure 3.* Histogram of 1,000 Gabor noise perturbations' (top) universal sensitivity and (bottom) universal evasion over 5,000 inputs.

It is worth using both the universal sensitivity and the universal evasion metrics, as the former gives a continuous measure of the sensitivity, while the latter tells us on how much of the dataset that perturbation changes the decision of the model.

### 4.2. Sensitivity to Gabor Noise

Our results show that the order from least to most sensitive models are Inception v3, ResNet-50, and then VGG-19. This is not surprising as the validation accuracies of these models also appear in the same order. Overall, our experiments show that the three models are significantly more sensitive to the Gabor noise than random noise. The universal sensitivity and evasion rates of random noise have very small variance and their values are clustered around their medians. Table 1 shows how close the quartiles of random noise's are for VGG-19.

Inception v3 is also insensitive to random noise, but has a moderate sensitivity to Gabor noise. ResNet-50 appears to be more sensitive to the random noise than VGG-19, but VGG-19 is more sensitive to Gabor noise than ResNet-50. This implies that when comparing models higher sensitivity to one type of perturbation does not imply the same relationship for another type of perturbation.

The results in Fig. 3 suggest that across the three models a random Gabor noise is likely to affect the model outputs on a third or more of the input dataset. From the histograms, the Gabor noise perturbations appear to centre around relatively high modes for both metrics. As an example, the first quartile of Gabor noise, as seen in Table 1, has 49.3% universal evasion, i.e. about 75% of the Gabor noise perturbations change VGG-19's decision on about half or more of the input dataset. For the remainder of this analysis we focus on VGG-19 as it is the most sensitive model. Similar figures and statistics for the other two models are in the appendix.

*Table 1.* Sensitivity (%) metric quartiles of Gabor and random noise perturbations on VGG-19.

| | Universal Sensitivity | | Universal Evasion | |
|---|---|---|---|---|
| Quartile | Gabor | Random | Gabor | Random |
| 1st | **34.2** | 13.8 | **49.3** | 20.8 |
| 2nd | **39.1** | 13.9 | **55.5** | 21.0 |
| 3rd | **43.7** | 13.9 | **61.5** | 21.3 |

**"Best" Parameters.** Taking the top 10 perturbations that VGG-19 is most sensitive to, we see that the other two models are also very sensitive to these noise patterns. The ranges of the universal evasion rate for these are 69.7% to 71.4% for VGG-19, 50.7% to 53.4% for ResNet-50, and 37.9% to 39.4% for Inception v3. These values are all above the 3rd quartile for each of these models, showing its generalizability to the other models.

In Fig. 5 we see a strong correlation ($\geq 0.74$) between the universal sensitivity and evasion rates across models. This further suggests that strong perturbations transfer across these models. We also see a weak correlation between $\lambda$ and the sensitivity and evasion rates for Inception v3, though there appears to be none between $\lambda$ and the sensitivity values for ResNet50.

The universal evasion rate of the perturbations appears to be insensitive to its Gaussian width $\sigma$ and orientation $\omega$. However, the sensitivity for small $\lambda < 0.3$ appears to fall below the average, suggesting that below a certain value the Gabor noise does not affect the model's decision. Interestingly, $\lambda$ corresponds to the width or thickness of the bands in the image. Examples of Gabor noise perturbations can be seen in the appendix.

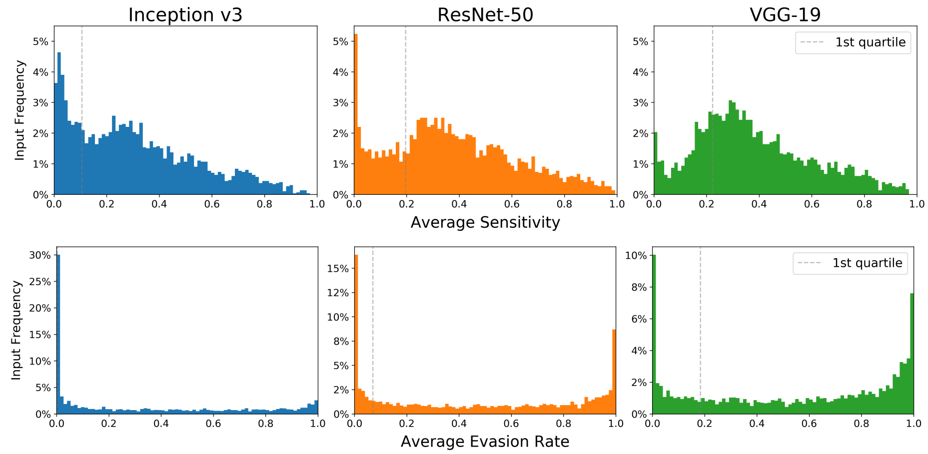


*Figure 4.* Histogram of 5,000 inputs' (top) average sensitivity and (bottom) average evasion over 1,000 Gabor noise perturbations.

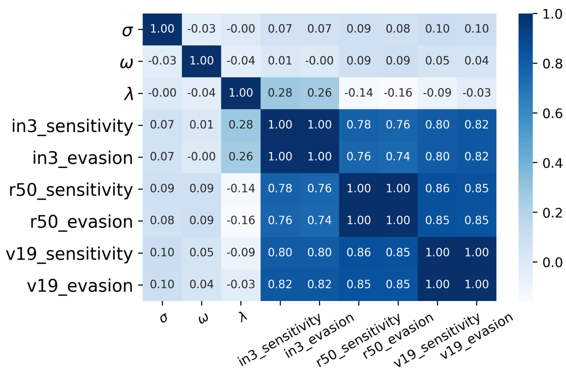


*Figure 5.* Correlation matrix of Gabor noise parameters and metrics for each model. Inception v3, ResNet-50, and VGG-19 are referred to as "in3", "r50", and "v19" respectively.

**Sensitivity of Inputs.** The model's sensitivity could vary across the input dataset, meaning that the model's predictions is stable on some inputs while more susceptible to small perturbations on others. To measure this, we look at the sensitivity of single inputs over all perturbations.

Given a set of perturbations $s \in S$, we define the **average sensitivity** of a model on input $x$ over $S$ as

$$\frac{1}{|S|} \sum_{s \in S} \|f(x) - f(x+s)\|_\infty,$$

and the **average evasion rate** on $x$ over $S$ as

$$\frac{|\{s \in S : \arg\max f(x) \neq \arg\max f(x+s)\}|}{|S|}.$$

The bimodal distribution of the average evasion rate in Fig. 4 shows that for each model there are two large subsets of the data: One that is very sensitive and another that is very insensitive. The remaining data points are somewhat uniformly spread in the middle. Note that for Inception v3, there is a much larger fraction of data points whose prediction is not affected by Gabor perturbations. The distribution for the average sensitivity appears to have similar shape, but with more inputs in the 0-20% range for Inception v3. The dataset is far less sensitive against random noise with upwards of 60% of the dataset being insensitive to that noise across all models.

## 5. Conclusion

The results show that the tested DCN models are sensitive to Gabor noise for a large fraction of the inputs, even when the parameters of the Gabor noise are chosen at random. This hints that it may be representative of patterns learned at the earlier layers as Gabor noise appears visually similar to some UAPs targeting earlier layers in DCNs (Khrulkov & Oseledets, 2018).

This phenomenon has important implications on the security and reliability of DCNs, as it can allow attackers to craft inexpensive black-box attacks. On the defender's side, Gabor noise patterns can also be used to efficiently generate data for adversarial training to improve DCNs robustness. However, both the sensitivity exploited and the potential to mitigate it require a more in-depth understanding of the phenomena at play. In future work, it may be worth analyzing the sensitivity of hidden layer activations across different families of procedural noise patterns and to investigate techniques to reduce the sensitivity of DCNs to perturbations.

## Acknowledgements

Kenneth Co is partially supported by the Data Spartan research grant DSRD201801. Example code is available at `https://github.com/kenny-co/procedural-advml`

## References

Boyd, D. W. The Power Method for Lp Norms. *Linear Algebra and its Applications*, 9:95–101, 1974.

Co, K. T., Muñoz-González, L., and Lupu, E. C. Procedural Noise Adversarial Examples for Black-Box Attacks on Deep Convolutional Networks. *arXiv preprint arXiv:1810.00470*, 2018.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press, 2016.

He, K., Zhang, X., Ren, S., and Sun, J. Deep Residual Learning for Image Recognition. In *Procs. Conf. on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Khrulkov, V. and Oseledets, I. Art of Singular Vectors and Universal Adversarial Perturbations. In *Procs. Conf. on Computer Vision and Pattern Recognition*, pp. 8562–8570, 2018.

Lagae, A., Lefebvre, S., Drettakis, G., and Dutré, P. Procedural Noise using Sparse Gabor Convolution. *ACM Trans. on Graphics*, 28(3):54, 2009.

Lagae, A., Lefebvre, S., Cook, R., DeRose, T., Drettakis, G., Ebert, D. S., Lewis, J. P., Perlin, K., and Zwicker, M. A Survey of Procedural Noise Functions. In *Computer Graphics Forum*, volume 29, pp. 2579–2600, 2010.

Moosavi-Dezfooli, S.-M., Fawzi, A., and Frossard, P. Deepfool: a Simple and Accurate Method to Fool Deep Neural Networks. In *Procs. Conf. on Computer Vision and Pattern Recognition*, pp. 2574–2582, 2016.

Moosavi-Dezfooli, S.-M., Fawzi, A., Fawzi, O., and Frossard, P. Universal Adversarial Perturbations. In *Procs. Conf. on Computer Vision and Pattern Recognition*, pp. 86–94, 2017.

Mopuri, K., Ojha, U., Garg, U., and Babu, R. V. NAG: Network for Adversary Generation. In *Procs. Conf. on Computer Vision and Pattern Recognition*, pp. 742–751, 2018.

Neyret, F. and Heitz, E. *Understanding and Controlling Contrast Oscillations in Stochastic Texture Algorithms using Spectrum of Variance*. PhD thesis, LJK/Grenoble University-INRIA, 2016.

Olah, C., Mordvintsev, A., and Schubert, L. Feature Visualization. *Distill*, 2(11):e7, 2017.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet Large Scale Visual Recognition Challenge. *Int. Journal of Computer Vision*, 115(3):211–252, 2015.

Simonyan, K. and Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *Procs. Int. Conf. on Learning Representations*, 2014.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In *Procs. Conf. on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. How Transferable are Features in Deep Neural Networks? In *Advances in Neural Information Processing Systems*, pp. 3320–3328, 2014.

Zhou, W., Hou, X., Chen, Y., Tang, M., Huang, X., Gan, X., and Yang, Y. Transferable Adversarial Perturbations. In *Procs. European Conf. on Computer Vision (ECCV)*, pp. 452–467, 2018.

## A. Sensitivity to Gabor Noise

As seen in Figure 6, sensitivity metric values for random noise fall in a narrow range and are significantly smaller than the metric values of the Gabor noise. This is further shown when comparing the quartiles of the universal evasion and sensitivity in Tables 2 and 3.

Figures 9, 10, 11, 12, and 13 show some adversarial examples with the top perturbations.

*Table 2.* Sensitivity (%) metric quartiles of Gabor and random noise perturbations on Inception v3.

| | Universal Sensitivity | | Universal Evasion | |
|---|---|---|---|---|
| Quartile | Gabor | Random | Gabor | Random |
| 1st | **29.9** | 11.8 | **32.8** | 13.0 |
| 2nd | **31.8** | 11.8 | **34.9** | 13.2 |
| 3rd | **33.2** | 11.9 | **36.9** | 13.5 |

*Table 3.* Sensitivity (%) metric quartiles of Gabor and random noise perturbations on ResNet-50.

| | Universal Sensitivity | | Universal Evasion | |
|---|---|---|---|---|
| Quartile | Gabor | Random | Gabor | Random |
| 1st | **35.7** | 19.3 | **44.3** | 25.6 |
| 2nd | **37.7** | 19.3 | **46.8** | 25.8 |
| 3rd | **40.4** | 19.4 | **50.1** | 26.0 |

## B. Sensitivity of Inputs

Large part of the input dataset is insensitive to random noise as shown in Tables 4, 5, 6 and Figure 7. With about 60% of the dataset on having near 0% average evasion over the random noise perturbations for all three models.

*Table 4.* Sensitivity (%) metric quartiles of input data over perturbations on Inception v3.

| | Average Sensitivity | | Average Evasion | |
|---|---|---|---|---|
| Quartile | Gabor | Random | Gabor | Random |
| 1st | **10.6** | 1.8 | **0.3** | 0.0 |
| 2nd | **26.8** | 6.7 | **19.8** | 0.0 |
| 3rd | **45.1** | 17.0 | **65.6** | 4.1 |

*Table 5.* Sensitivity (%) metric quartiles of input data over perturbations on ResNet-50.

| | Average Sensitivity | | Average Evasion | |
|---|---|---|---|---|
| Quartile | Gabor | Random | Gabor | Random |
| 1st | **19.6** | 2.3 | **7.0** | 0.0 |
| 2nd | **34.8** | 13.3 | **45.1** | 0.0 |
| 3rd | **53.9** | 29.2 | **84.6** | 53.7 |

*Table 6.* Sensitivity (%) metric quartiles of input data over perturbations on VGG-19.

| | Average Sensitivity | | Average Evasion | |
|---|---|---|---|---|
| Quartile | Gabor | Random | Gabor | Random |
| 1st | **22.3** | 3.1 | **18.4** | 0.0 |
| 2nd | **34.2** | 10.3 | **60.2** | 0.0 |
| 3rd | **52.2** | 19.6 | **90.0** | 24.5 |

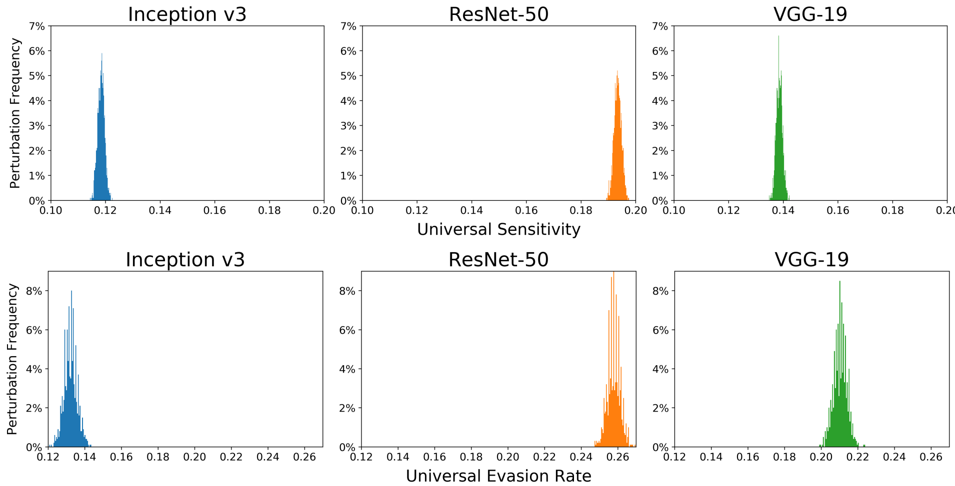


*Figure 6.* Histogram of 1,000 random noise perturbations based on (top) universal sensitivity and (bottom) universal evasion rate.

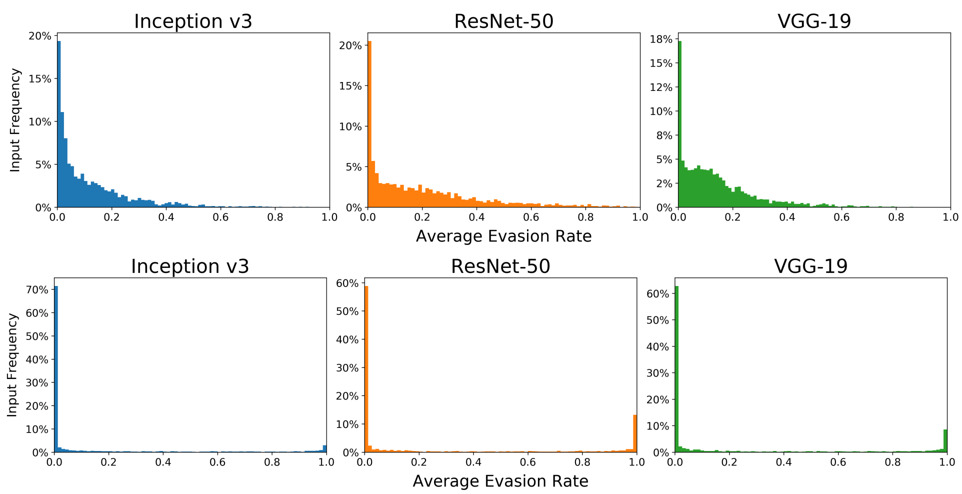


*Figure 7.* Histogram of 5,000 inputs based on (top) average sensitivity and (bottom) average evasion rate over random noise perturbations.

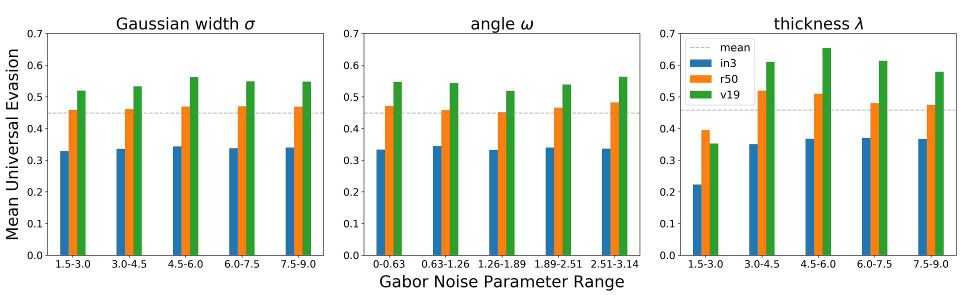


*Figure 8.* Mean universal evasion rate of Gabor noise perturbations grouped according to parameter values.

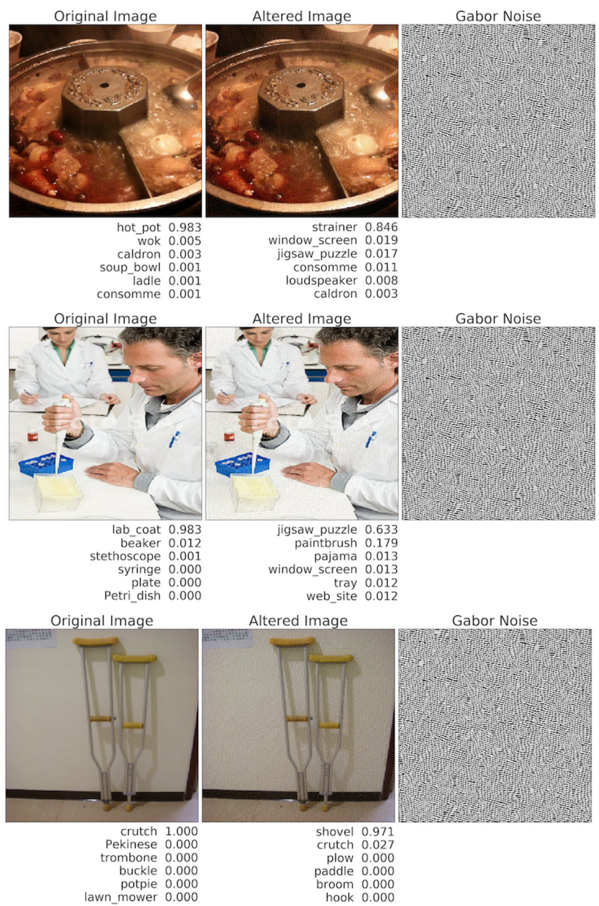


*Figure 9.* Gabor noise parameters $\Theta_6 = \{4.78, 1.81, 2.93\}$ on Inception v3.

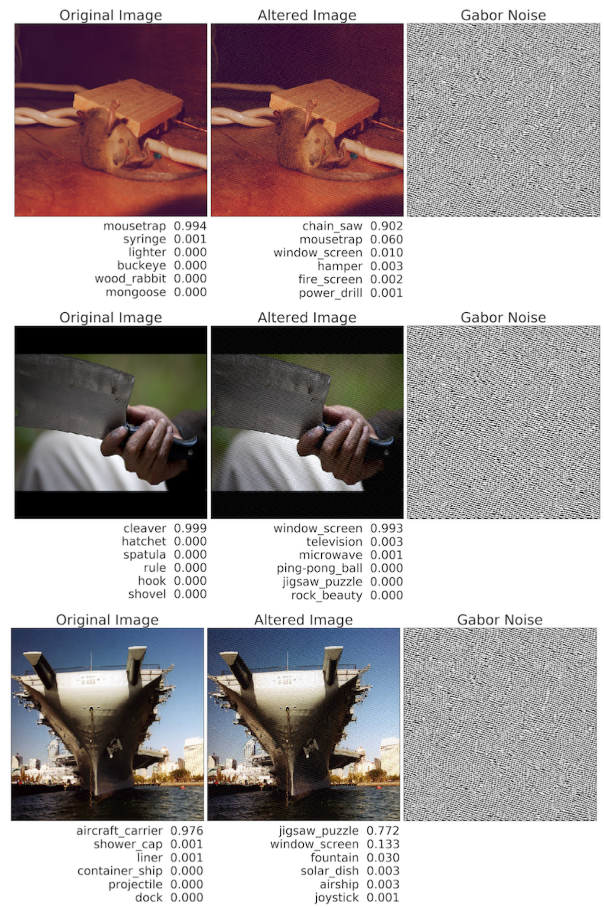


*Figure 10.* Gabor noise parameters $\Theta_{709} = \{7.92, 1.85, 3.12\}$ on Inception v3.

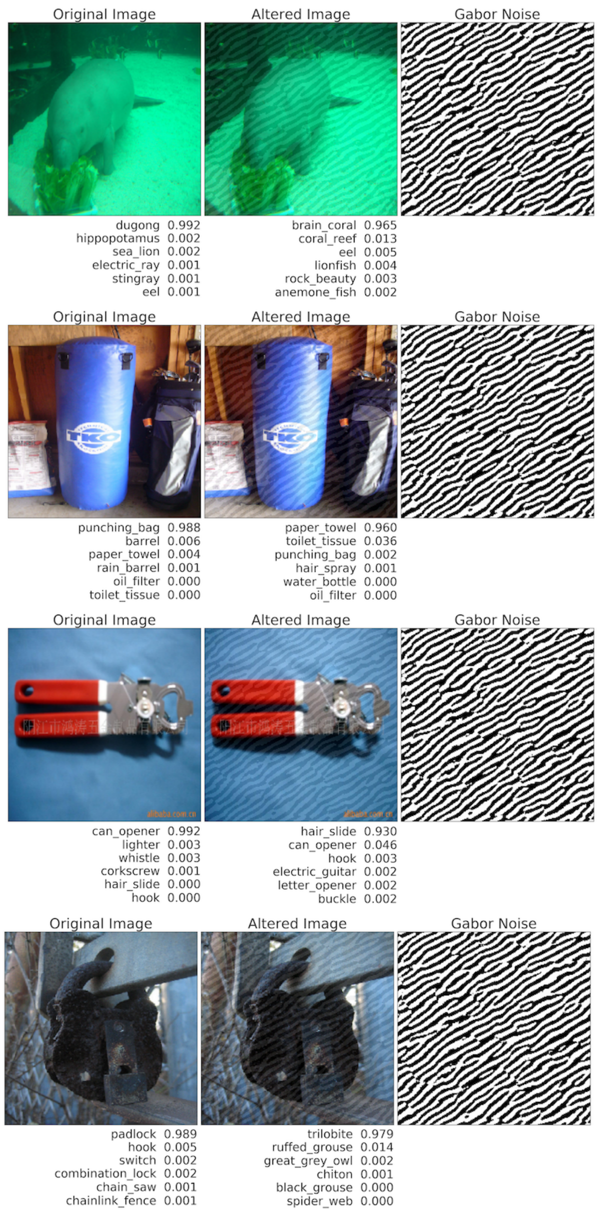


*Figure 11.* Gabor noise parameters $\Theta_{119} = \{6.69, 1.02, 8.45\}$ on ResNet-50.

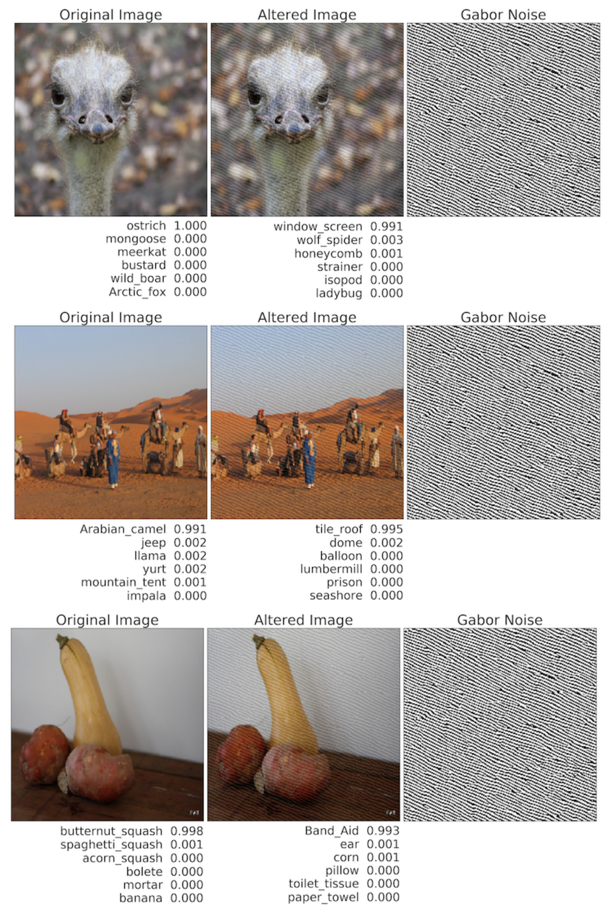


*Figure 12.* Gabor noise parameters $\Theta_{185} = \{6.10, 1.99, 3.46\}$ on ResNet-50.

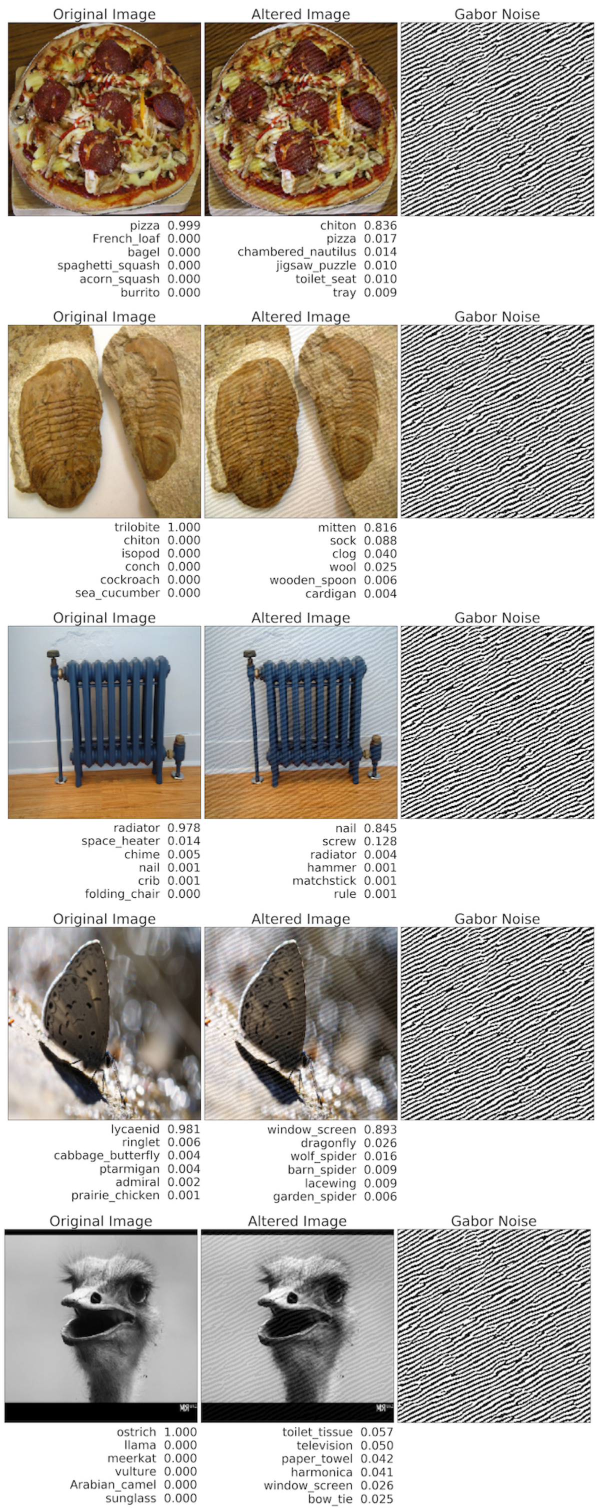


*Figure 13.* Gabor noise parameters $\Theta_{25} = \{6.29, 1.10, 4.86\}$ on VGG-19.