Popular Imperceptibility Measures in Visual Adversarial Attacks are Far from Human Perception

Ayon Sen[1(✉)], Xiaojin Zhu[1], Erin Marshall[2], and Robert Nowak[2]

[1] Computer Sciences Department, University of Wisconsin-Madison, Madison, USA
asen6@wisc.edu, jerryzhu@cs.wisc.edu
[2] Department of Electrical and Computer Engineering,
University of Wisconsin-Madison, Madison, USA
{limarshall,rdnowak}@wisc.edu

Abstract. Adversarial attacks on image classification aim to make visually imperceptible changes to induce misclassification. Popular computational definitions of imperceptibility are largely based on mathematical convenience such as pixel p-norms. We perform a behavioral study that allows us to quantitatively demonstrate the mismatch between human perception and popular imperceptibility measures such as pixel p-norms, earth mover's distance, structural similarity index, and deep net embedding. Our results call for a reassessment of current adversarial attack formulation.

Keywords: Adversarial machine learning · Imperceptibility · Just noticeable difference

1 Introduction

Recent visual adversarial attack research frequently uses the following formulation [10,17]. Let $\mathbf{x}_0$ be an image in an appropriate vector space, y be its true class label, θ a trained classifier, and ℓ the learner's loss function. The attacker seeks a perturbed image $\mathbf{x}$ to make the true label y seem unlikely (by maximizing the loss):

$$\begin{aligned} \max_{\mathbf{x}} \quad & \ell(\mathbf{x}, y, \theta) \\ \text{s.t.} \quad & d(\mathbf{x}, \mathbf{x}_0) \le \epsilon. \end{aligned} \tag{1}$$

The feasible set is defined by a distance function $d()$ and a threshold ϵ. A common choice for $d()$ is the infinite norm in the pixel space: $d(\mathbf{x}, \mathbf{x}_0) := \|\mathbf{x} - \mathbf{x}_0\|_\infty$, although other p-norms (especially for $p = 1, 2$) and several other measures (defined in the next section) are popular, too. An alternative formulation minimizes the distance function $d(\mathbf{x}, \mathbf{x}_0)$ subject to wrong label prediction.

Implicit in such formulations is the assumption that the feasible set defined by $d(\mathbf{x}, \mathbf{x}_0) \le \epsilon$ coincides with imperceptible perturbations as

Q. Zhu et al. (Eds.): GameSec 2020, LNCS 12513, pp. 188–199, 2020.
https://doi.org/10.1007/978-3-030-64793-3_10

observed by human inspectors [3,7,15,20,26,27]. Then perhaps the attacker can wreck havoc against the classifier without being noticed by humans. This assumption has been criticized for its over-simplification of the threat model [6,21]. Indeed, many adversarial learning researchers readily admit that popular choices of $d()$ are more of a mathematical convenience, and may not correspond well with human perception. Disconcertingly, a large number of papers keep making this assumption without verifying how good or bad the assumption really is: Out of 32 recent papers we surveyed, 27 papers (each with over 100 citations) used pixel p-norms for $d()$. Among these 27, 20% assumed p-norms are a good match to human perception without providing evidence; 50% used them because other papers did; and the rest used them without justification. Given the recent prominence of visual adversarial learning research, there is a need to quantitatively study this assumption to refine the threat model.

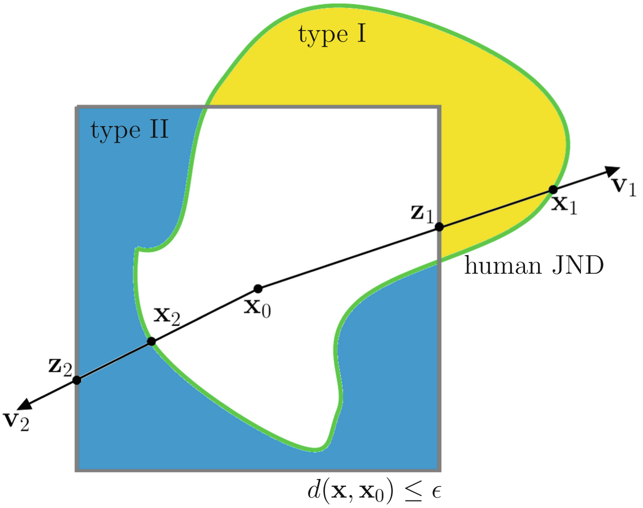


Fig. 1. Mismatches between human perception and distance function d

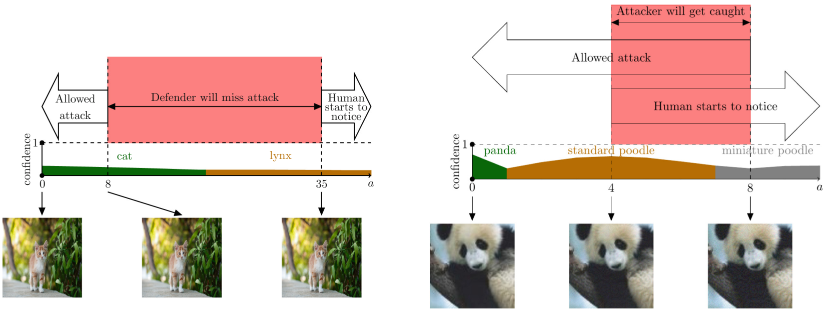
What is the harm if $d()$ and human perception differ? Consider the image space around image $\mathbf{x}_0$ in Fig. 1. The feasible set $\{\mathbf{x} : d(\mathbf{x}, \mathbf{x}_0) \leq \epsilon\}$ is the region within the gray contour, while the human imperceptibility region is within the green contour: intuitively, any image in this region looks like $\mathbf{x}_0$ to an average human (precise definition below).

- The yellow region is type I error: humans perceive images there, e.g. $\mathbf{x}_1$, the same as $\mathbf{x}_0$ but the feasible set by definition thinks otherwise. Type I errors are dangerous because it lets the machine's guard down: the machine does not even consider $\mathbf{x}_1$ to be a valid attack (while $\mathbf{x}_1$ may in fact change the label prediction), *and* human inspection will not notice the attack.

– The blue region is type II error: humans perceive images there, e.g. $\mathbf{z}_2$, as noticeably different from $\mathbf{x}_0$ but the feasible set thinks they cannot be distinguished. Type II errors waste the machine's resources by defending against fictitious threats.

Both types of error have occurred in practice, as shown in Fig. 2. In both examples we used $d(\mathbf{x}, \mathbf{x}_0) = \|\mathbf{x} - \mathbf{x}_0\|_\infty$ and $\epsilon = 8$ (out of pixel value 0–255), as is commonly used in adversarial machine learning [1,31].

Our main contribution is a new human experiment design that allows us to quantitatively gauge the mismatch between human perception and popular imperceptibility measures $d()$, specifically pixel p-norms, EMD, 1-SSIM, and DNN representation p-norms. Our results call for a reassessment of the adversarial attack formulation (1) vis-à-vis real threats.



(a) type I error

(b) type II error

Fig. 2. (a) $\mathbf{x}_0$=cat photo, $\mathbf{v}$=M_RGB_Box (see Sect. 3), $\mathbf{x} = \mathbf{x}_0 + a\mathbf{v}$. $d(\mathbf{x}, \mathbf{x}_0) = \|\mathbf{x} - \mathbf{x}_0\|_\infty$ and $\epsilon = 8$ as in the literature [1,31]; this corresponds to $a = 8$. On the other hand, our experiments showed that human JND is not until $a = 35$. The images produced by $a \in (8, 35)$ are type I errors: a machine defender will not consider them, and humans cannot tell them apart from $\mathbf{x}_0$. Critically, Inception V3 classifier [25] will classify $a \geq 20$ as lynx, meaning images in $a \in [20, 35)$ are dangerous attacks. (b) $\mathbf{x}_0$=panda photo, $\mathbf{v}$=FGSM [7] attack direction. Again d is the infinite norm, and $\epsilon = 8$. Along this direction humans are good at detecting changes: our experiments showed that human JND happens at $\|\mathbf{x} - \mathbf{x}_0\|_\infty = 4$ already. An attack produced by FGSM with $\|\mathbf{x} - \mathbf{x}_0\|_\infty \in \{4, 5, \ldots, 8\}$ will get caught. Therefore, the specific FGSM attack will likely be detected by humans. The issue on the surface may look like an inappropriate ϵ threshold used by FGSM, but keep in mind that along different directions $\mathbf{v}$ the human JND threshold can vary, and there may not be a correct global ϵ. The root cause is an inappropriate $d()$ used by adversarial attacks.

2 Study Overview

Our study is designed to facilitate human experiments. Given a natural image $\mathbf{x}_0$, consider an arbitrary direction $\mathbf{v}$ as shown in Fig. 1. The ray centered at $\mathbf{x}_0$ in the direction $\mathbf{v}$ is parametrized as $\{\mathbf{x} := \mathbf{x}_0 + a\mathbf{v} \mid a \geq 0\}$ with a scalar parameter a. Larger a leads to more changes to $\mathbf{x}_0$. We expect to find a threshold value a_v for direction $\mathbf{v}$, above which an average human inspector will notice the difference between $\mathbf{x}_0$ and $\mathbf{x}_0 + a_v\mathbf{v}$. These are images $\mathbf{x}_1, \mathbf{x}_2$ in Fig. 1 for directions $\mathbf{v}_1, \mathbf{v}_2$, respectively.

Now, an adversarial attack feasible set in (1) is defined by distance measure $d()$ and threshold ϵ. Our primary interest is the appropriateness of $d()$ compared to human perception. ϵ is a nuisance parameter; fortunately, we do not need to know its value. The **key insight** is that, if $d()$ correctly models human perception, then under this measure the distance

$$d(\mathbf{x}_0 + a_v\mathbf{v}, \mathbf{x}_0) \tag{2}$$

is a constant for all directions $\mathbf{v}$. In other words, the "just noticeably different" images by humans form a sphere around $\mathbf{x}_0$ under the correct $d()$. Conversely, we may summarize how far off some $d()$ is from human perception by the *condition number*

$$\kappa(d) := \frac{\max_{\mathbf{v}} d(\mathbf{x}_0 + a_v\mathbf{v}, \mathbf{x}_0)}{\min_{\mathbf{v}} d(\mathbf{x}_0 + a_v\mathbf{v}, \mathbf{x}_0)}. \tag{3}$$

The larger $\kappa(d)$ is, the worse $d()$ is. The smallest possible value of $\kappa(d)$ is 1. It is analogous to the ratio of major vs. minor axes for an ellipsoid. Note $\kappa(d)$ is center-image $\mathbf{x}_0$ dependent.

We will empirically estimate $\kappa(d)$ for popular $d()$'s. Because this involves human experiments, practically we can only consider a finite, small number of center images $\mathbf{x}_0$. Furthermore, for each $\mathbf{x}_0$ we can only consider a small number of directions $V = \{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$. From these, we obtain an empirical estimate of condition number

$$\hat{\kappa}(d) := \frac{\max_{\mathbf{v} \in V} d(\mathbf{x}_0 + a_v\mathbf{v}, \mathbf{x}_0)}{\min_{\mathbf{v} \in V} d(\mathbf{x}_0 + a_v\mathbf{v}, \mathbf{x}_0)}, \tag{4}$$

where max and min only go over the directions in V. Clearly, this is an underestimate: $\hat{\kappa}(d) \leq \kappa(d)$. If our measured $\hat{\kappa}(d)$ is large (and thus $\kappa(d)$ potentially even larger), we conclude that $d()$ is inappropriate.

2.1 Human Just Noticeable Difference (JND)

We define the just noticeable difference (JND) [5,32] with respect to a center image $\mathbf{x}_0$ and direction $\mathbf{v}$ as the image $\mathbf{x}_0 + a_v\mathbf{v}$ where an average human observer starts to perceive a difference. Equivalently, human JND is characterized by the scalar a_v. We discuss how to empirically measure human JND in Sect. 3.

2.2 Popular Imperceptibility Measures $d()$

Pixel p-Norm. For any $p \in [0, \infty]$ it measures the amount of perturbation by $\|\mathbf{x} - \mathbf{x}_0\|_p := \left(\sum_{i=1}^{d} |x_i - x_{0,i}|^p\right)^{1/p}$. We define the 0-norm to be the number of nonzero elements.

Earth Mover's Distance (EMD). Also known as Wasserstein distance, it is a distance function defined between two probability distributions on a given metric space. The metric computes the minimum cost of converting one distribution to the other one. EMD has been used as a distance metric in the image space also, e.g. for image retrieval [19]. Given two images $\mathbf{x}_0$ and $\mathbf{x}$, EMD is calculated as $EMD(\mathbf{x}_0, \mathbf{x}) = \inf_{\gamma \in \Gamma(\mathbf{x}_0, \mathbf{x})} \int_{\mathbb{R}\times\mathbb{R}} |a - b| d\gamma(a, b)$. Here, $\Gamma(\mathbf{x}_0, \mathbf{x})$ is the set of joint distributions whose marginals are $\mathbf{x}_0$ and $\mathbf{x}$ (treated as histograms), respectively.

Structural Similarity (SSIM). This measure is intended to be a perceptual similarity measure that quantifies image quality loss due to compression [29], and used as a signal fidelity measure with respect to humans in multiple research works [22,28]. SSIM has three elements: luminance, contrast and similarity of local structure. Given two images $\mathbf{x}_0$ and $\mathbf{x}$, SSIM is defined by $SSIM(\mathbf{x}_0, \mathbf{x}) = \left(\frac{2\mu_{\mathbf{x}_0}\mu_{\mathbf{x}} + C_1}{\mu^2_{\mathbf{x}_0} + \mu^2_{\mathbf{x}} + C_1}\right)\left(\frac{2\sigma_{\mathbf{x}_0}\sigma_{\mathbf{x}} + C_2}{\sigma^2_{\mathbf{x}_0} + \sigma^2_{\mathbf{x}} + C_2}\right)\left(\frac{\sigma_{\mathbf{x}_0\mathbf{x}} + C_3}{\sigma_{\mathbf{x}_0}\sigma_{\mathbf{x}} + C_3}\right)$. $\mu_{\mathbf{x}_0}$ and $\mu_{\mathbf{x}}$ are the sample means; $\sigma_{\mathbf{x}_0}$, $\sigma_{\mathbf{x}}$ and $\sigma_{\mathbf{x}_0\mathbf{x}}$ are the standard deviation and sample cross correlation of $\mathbf{x}_0$ and $\mathbf{x}$ (after subtracting the mean) respectively. To compute SSIM we use window size 7 without Gaussian weights. Since SSIM is a similarity score, we define $d(\mathbf{x}, \mathbf{x}_0) = 1 - SSIM(\mathbf{x}, \mathbf{x}_0)$.

Deep Neural Network (DNN) Representation. Even though DNNs are designed with engineering goals in mind, studies comparing their internal representations to primate brains have found similarities [11]. Let $\xi(\mathbf{x}) \in \mathbb{R}^D$ denote the last hidden layer representation of input image $\mathbf{x}$ in a DNN. We define $d(\mathbf{x}, \mathbf{x}_0) = \|\xi(\mathbf{x}) - \xi(\mathbf{x}_0)\|_p$ as a potential distance metric for our purpose. We use Inception V3 [25] representations with $D = 2048$.

3 Human JND Experiments

Center Images $\mathbf{x}_0$ and Perturbation Directions $\mathbf{v}$: We chose three natural images (from the Imagenet dataset [4]) popular in adversarial research: a panda [7], a macaw [16] and a cat [1] as $\mathbf{x}_0$ in our experiments. We resized the images to 299 × 299 to match the input dimension of the Inception V3 image classification network [25].

As indicated in Fig. 1, we consider $\mathbf{x}$ generated along the ray defined by a perturbation direction $\mathbf{v} \in \mathbb{R}^d$ with a perturbation scale $a > 0$. To render the image for display, we project it to the image space: $\mathbf{x} = \Pi(\mathbf{x}_0 + a\mathbf{v})$, namely, clipping pixel values to $[0, 255]$ and rounding to integers.

For each natural image $\mathbf{x}_0$ we considered 10 perturbation directions $\mathbf{v}$, see Fig. 3. Eight are specially crafted ±1-perturbation (i.e., $\mathbf{v}$ has elements -1, 0, 1) directions varying in three attributes (Table 1). Specifically, the nonzero elements v_i depend on the value of the corresponding element $x_{0,i}$ in $\mathbf{x}_0$: $v_i = 1$ if $x_{0,i} < 128$, and -1 otherwise. For ±1-perturbations $\mathbf{v}$ and integer $a \in \{1, \ldots, 128\}$ it is easy to see that the projection Π is not needed: $\mathbf{x} = \Pi(\mathbf{x}_0 + a\mathbf{v}) = \mathbf{x}_0 + a\mathbf{v}$.

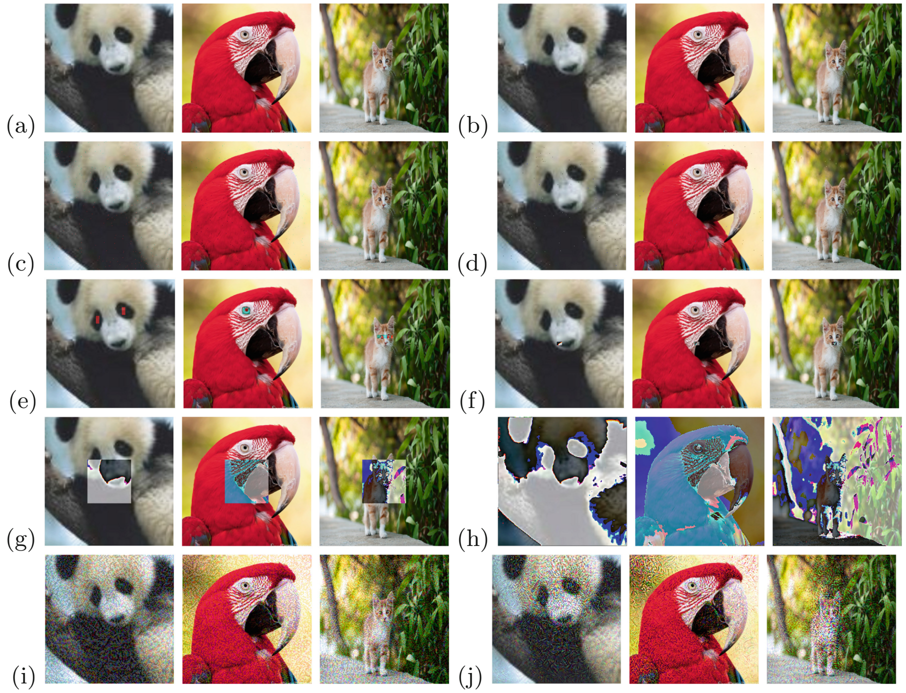


Fig. 3. All 10 perturbation directions $\mathbf{v}$ with severe perturbation scale $a = 128$. (a) S_Red_Box: the red channel of the center pixel. (b) S_Red_Dot: a randomly selected red channel. (c) M_Red_Dot: 288 randomly selected red channels. (d) M_RGB_Dot: all three color channels of 96 randomly selected pixels ($s = 3 \times 96 = 288$). (e) M_Red_Eye: 288 red channels around the eyes of the animals. (f) M_RGB_Box: all colors of a centered 8×12 rectangle. (g) L_RGB_Box: all colors of a centered 101×101 rectangle. (h) X_RGB_Box: all dimensions. (i) FGSM. (j) PGD.

The remaining two perturbation directions are adversarial directions. We used Fast Gradient Sign Method (FGSM) [7] and Projected Gradient Descent (PGD) [14] to generate two adversarial images $\mathbf{x}^{FGSM}, \mathbf{x}^{PGD}$ for each $\mathbf{x}_0$, with Inception V3 as the victim network. All attack parameters are set as suggested in the methods' respective papers. PGD is a directed attack and requires a target label; we choose gibbon (on panda) and guacamole (on cat) following the papers, and cleaver (on macaw) arbitrarily. We then define the adversarial perturbation directions by $\mathbf{v}^{FGSM} = 127.5(\mathbf{x}^{FGSM} - \mathbf{x}_0)/\|\mathbf{x}^{FGSM} - \mathbf{x}_0\|_2$ and $\mathbf{v}^{PGD} = 127.5(\mathbf{x}^{PGD} - \mathbf{x}_0)/\|\mathbf{x}^{PGD} - \mathbf{x}_0\|_2$. We use the factor 127.5 based on a pilot study to ensure that changes between consecutive images in the adversarial perturbation directions are not too small or too big.

Table 1. Naming convention for perturbation directions $\mathbf{v}$

# Dimensions changed	S = 1, M = 288 	
	L = 30603, X = 268203
	(mnemonic: garment size)

Color channels affected	Red = only the red channel of a pixel RGB = all three channels of a pixel

Shape of perturbed pixels	Box = a centered rectangle
	Dot = scattered random dots
	Eye = on the eye of the animal

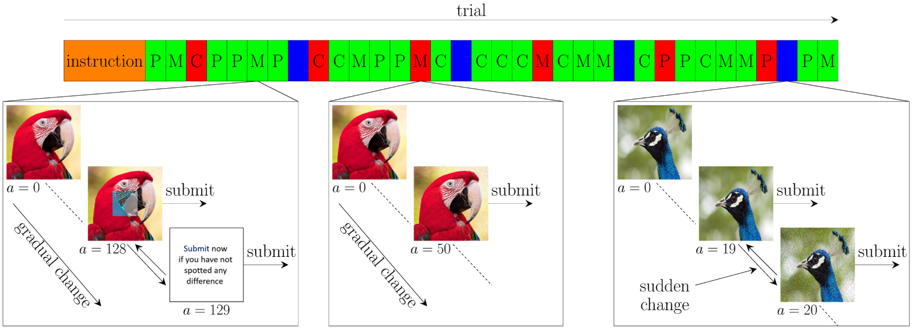


Fig. 4. Experiment procedure. The green, red and blue cells denote ±1-perturbation, adversarial, and guard trials, respectively.The letters P, M and C denote the panda, macaw and cat $\mathbf{x}_0$, respectively. (Color figure online)

Experimental Procedure : See Fig. 4. Each participant was first presented with instructions and then completed a sequence of 34 trials, of which 30 were ±1-perturbation or adversarial trials, and 4 were guard trials. The order of these trials was randomized then fixed (see figure). During each trial the participants were presented with an image $\mathbf{x}_0$. They were instructed to increase (decrease) perturbations to this image by using right/left arrow keys or buttons. Moving right (left) incremented (decremented) a by 1, and the subject was then presented with the new perturbed image $\mathbf{x} = \Pi(\mathbf{x}_0 + a\mathbf{v})$. We did not divulge the nature of the perturbations $\mathbf{v}$ beforehand, nor the current perturbation scale a the participant had added to $\mathbf{x}_0$ at any step of the trial. **The participants were instructed to submit the perturbed image x when they think it became just noticeably different from the original image** $\mathbf{x}_0$. The participants had to hold $\mathbf{x}_0$ in memory, though they could also go all the way left back to see $\mathbf{x}_0$ again. We hosted the experiment using the NEXT platform [9,23].

In a ±1-perturbation trial, the perturbation direction $\mathbf{v}$ is one of the eight ±1-perturbations. We allowed the participants to vary a within $\{0, 1, \ldots, 128\}$ to avoid value cropping. If a participant was not able to detect any change even after $a = 128$, then they were encouraged to "give up".

In an adversarial trial, the perturbation direction is $\mathbf{v}^{FGSM}$ or $\mathbf{v}^{PGD}$. We allowed the participants to increment a indefinitely, though no one went beyond $a = 80$.

The guard trials were designed to filter out participates who clicked through the experiment without performing the task. In a guard trial, we showed a novel fixed natural image (not panda, macaw or cat) for $a < 20$. Then for $a \geq 20$, a highly noisy version of that image is displayed. An attentive participant should readily notice this sudden change at $a = 20$ and submit it. We disregarded guard trials in our analysis.

Participants and Data Inclusion Criterion : We enrolled 68 participants using Amazon Mechanical Turk [2] master workers. A master worker is a person who has consistently displayed a high degree of success in performing a wide range of tasks. All participants used a desktop, laptop or a tablet device; none used a mobile device where the screen would be too small. On average the participants took 33 minutes to finish the experiment. Each participant was paid \$5. As mentioned before, we use guard trials to identify inattentive participants. While the change happens at exactly $a = 20$ in a guard trial, our data indicates a natural spread in participant submissions around 20 with sharp decays. We speculate that the spread was due to keyboard/mouse auto repeat. We set a range for an acceptable guard trial if a participant submitted $a \in \{18, 19, 20, 21, 22\}$. A participant is deemed inattentive if any one of the four guard trials was outside the acceptable range. Only $n = 42$ out of 68 participants survived this stringent inclusion condition. All our analyses below are on these 42 participants.

4 Results

For each center image $\mathbf{x}_0$ and perturbation direction $\mathbf{v}$, the jth participant ($j = 1 \dots n$) gave us their individual JND threshold scale parameter $a_v^{(j)}$. That is, the image $\mathbf{x}^{(j)} = \Pi(\mathbf{x}_0 + a_v^{(j)}\mathbf{v})$ is the one participant j thinks has just-noticeable-difference to $\mathbf{x}_0$ along direction $\mathbf{v}$.

Because our participants can sometimes choose to "give up" if they did not notice a change, we have *right censored data* on a_v. All we know from a given-up trial is that $a \geq 129$, but not what larger a value will cause the participant to noticed a difference. For example, many participants failed to notice a difference along the S_Red_Box and S_Red_Dot perturbation directions, thus many a_v's in those directions (50.8% and 51.6% respectively) were censored. A total of 13.2% a_v's were censored along all directions.

4.1 Qualitative Assessment

Recall if a distance measure $d()$ is a good match to human perception, then by (2) along any direction $\mathbf{v}$ the human JND image $\mathbf{x} = \mathbf{x}_0 + a_v\mathbf{v}$ has the same $d(\mathbf{x}, \mathbf{x}_0)$. We present box plots to qualitatively assess the different $d()$'s in Figs. 5 and 6. We selectively show only one center image $\mathbf{x}_0$ for each of the measures for

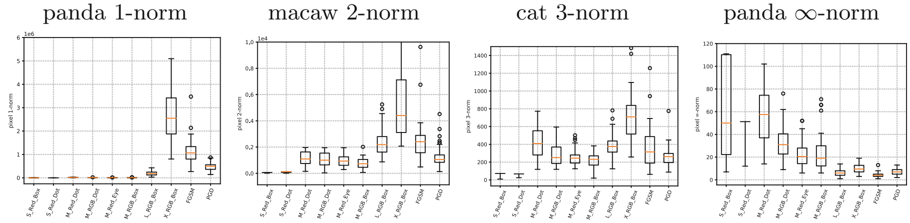


Fig. 5. Participant JND $\mathbf{x}$'s pixel norm $\|\mathbf{x} - \mathbf{x}_0\|_p$. Within a plot, each vertical box is for a perturbation direction $\mathbf{v}$. The box plot depicts the median, quartiles, and outliers.

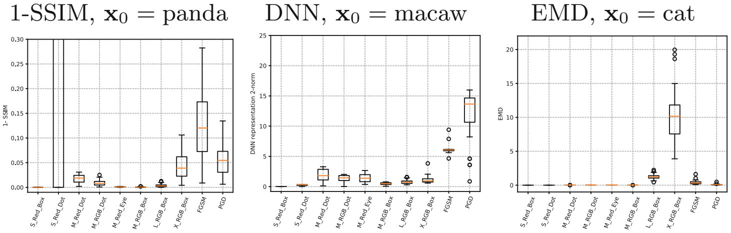


Fig. 6. Box plots of different measures ρ on human JND images.

the interest of space. We will show all plots in an extended version of this paper. The perturbation directions $\mathbf{v}$ are indicated on the x-axis. The y-axis shows the median, quartiles, and outliers of the participants' JND images, measured in the specific $d()$ indicated in the plots. The **main qualitative observation** is that none of the popular distance measure $d()$ has a flat median across the directions we tested. For example, for pixel 2-norm on $\mathbf{x}_0$=macaw, the median is 1049 and 4402 along the PGD and X_RGB_Box directions respectively. Similarly for DNN 2-norm on $\mathbf{x}_0$=macaw, the median is 1.8 and 13.6 respectively along the M_Red_Dot and PGD directions respectively. This indicates that none of these measures is a good fit to human JND.

4.2 Quantitative Assessment

Now we report the empirical estimate of condition number $\hat{\kappa}$ for each distance measure $d()$ and center image $\mathbf{x}_0$. Recall that $\hat{\kappa}(d)$ must be close to 1 for a distance measure $d()$ to have the possibility to be a good fit to human perception. Due to the large number of censored data along some directions, we estimate $\hat{\kappa}(d)$ in two ways.

- Non-censored median: We discarded all "given up" data. We then estimated the human JND distance using the median value along a direction $\mathbf{v}$. This is shown in Table 2.
- First quartile: In the second procedure, we do not discard the "given up" values but consider those distances to be infinity. Then we estimate the human

Table 2. Estimated $\hat{\kappa}(d)$ using non-censored median.

Center Image	1-norm	2-norm	3-norm	∞-norm	EMD	1 - SSIM	DNN	DNN	DNN
							1-norm	2-norm	∞-norm
panda	73853	142.6	17.8	**14**	68457	27913	476	512	575
macaw	499559	95.7	**11.9**	14.3	48210	56933	854	683	627
Cat	46460	89.7	**11.2**	14	42919	11786	389	381	355

Table 3. Estimated $\hat{\kappa}(d)$ using the first quartile.

Center Image	1-norm	2-norm	3-norm	∞-norm	EMD	1 - SSIM	DNN	DNN	DNN
							1-norm	2-norm	∞-norm
panda	84379	163	20.3	**17.1**	79777	16353	577	704	496
macaw	23752	45.9	**5.7**	21.6	23300	255502	442	355	341
Cat	31787	61.4	**7.6**	24.2	30031	68609	341	329	297

JND along a direction by the first quartile. The median would have fallen in censored values for some directions. The first quartile is a biased estimate of human JND $d(\mathbf{x}, \mathbf{x}_0)$, but has the benefit of not hitting any censored values. This is shown in Table 3.

We highlight the smallest estimated condition number $\hat{\kappa}$ in each table. All of these values are much larger than 1. This quantitatively shows that popular imperceptibility measures in visual adversarial attacks are far from human perception.

5 Discussions and Conclusion

We quantitatively show that pixel p-norms, EMD, 1 - SSIM, and DNN representation p-norms are not good matches to human perception. This paper thus calls for a rethinking of adversarial attack formulation. The closest work to ours is [21], which also conducted human experiments on adversarial attacks and human perception. That study was limited in design: they only tested pixel 0-, 2-, ∞-norms but not other p-norms or measures. Their test also relies on the knowledge of the feasible set radius ϵ, and depended on humans (mis)-categorizing a low resolution thumbnail (MNIST [13], CIFAR10 [12]). Instead, humans may notice small changes in a normal-sized image well before their categorization of the image changes. The present paper addresses these issues.

We also mention some limitations of our own work: (1) We used only three center images $\mathbf{x}_0$ in our human experiments. This is due to the fact that running human experiments is time consuming and expensive. (2) We still cannot answer "what is the correct measure $d()$", noting that computationally modeling human visual perception is still an open question in psychology [8,18,24,30]. (3) We used a "show $\mathbf{x}_0$ then perturb" experiment paradigm, while in real applications the human inspector may not have access to $\mathbf{x}_0$. (4) We limited ourselves to the visual domain. Addressing these limitations remain future work.

Acknowledgments. Acknowledgments: This work is supported in part by NSF grants 1545481, 1623605, 1704117, 1836978 and the MADLab AF Center of Excellence FA9550-18-1-0166.

References

1. Athalye, A., Carlini, N., Wagner, D.: Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. arXiv preprint arXiv:1802.00420 (2018)
2. Buhrmester, M., Kwang, T., Gosling, S.D.: Amazon's mechanical turk: a new source of inexpensive, yet high-quality, data? Perspect. Psychol. Sci. **6**(1), 3–5 (2011)
3. Carlini, N., Wagner, D.: Adversarial examples are not easily detected: bypassing ten detection methods. In: Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, pp. 3–14. ACM (2017)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
5. Fechner, G.T., Boring, E.G., Howes, D.H., Adler, H.E.: Elements of Psychophysics. Translated by Helmut E. Adler. Edited by Davis H. Howes And Edwin G. Boring, With an Introd. by Edwin G. Boring. Holt, Rinehart and Winston (1966)
6. Gilmer, J., Adams, R.P., Goodfellow, I., Andersen, D., Dahl, G.E.: Motivating the rules of the game for adversarial example research. arXiv preprint arXiv:1807.06732 (2018)
7. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples (2014). arXiv preprint arXiv:1412.6572
8. Itti, L., Koch, C.: Computational modelling of visual attention. Nat. Rev. Neurosci. **2**(3), 194 (2001)
9. Jamieson, K.G., Jain, L., Fernandez, C., Glattard, N.J., Nowak, R.: Next: a system for real-world development, evaluation, and application of active learning. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 28, pp. 2656–2664. Curran Associates Inc., Red Hook (2015)
10. Kolter, Z., Madry, A.: Adversarial robustness: theory and practice. In: Tutorial at NeurIPS (2018)
11. Kriegeskorte, N.: Deep neural networks: a new framework for modeling biological vision and brain information processing. Ann. Rev. Vis. Sci. **1**, 417–446 (2015)
12. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. Technical report, Citeseer (2009)
13. LeCun, Y.: The mnist database of handwritten digits (1998). http://yann.lecun.com/exdb/mnist/
14. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
15. Moosavi-Dezfooli, S.M., Fawzi, A., Fawzi, O., Frossard, P.: Universal adversarial perturbations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1765–1773 (2017)

16. Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: Deepfool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2574–2582 (2016)
17. Papernot, N., McDaniel, P., Sinha, A., Wellman, M.: Towards the science of security and privacy in machine learning. arXiv preprint arXiv:1611.03814 (2016)
18. Rensink, R.A.: Change detection. Ann. Rev. Psychol. **53**(1), 245–277 (2002)
19. Rubner, Y., Tomasi, C., Guibas, L.J.: The earth mover's distance as a metric for image retrieval. Int. J. Comput. Vis. **40**(2), 99–121 (2000)
20. Salamati, M., Soudjani, S., Majumdar, R.: Perception-in-the-loop adversarial examples. arXiv preprint arXiv:1901.06834 (2019)
21. Sharif, M., Bauer, L., Reiter, M.K.: On the suitability of lp-norms for creating and preventing adversarial examples. In: The Bright and Dark Sides of Computer Vision: Challenges and Opportunities for Privacy and Security (CVPR Workshop) (2018)
22. Sheikh, H.R., Sabir, M.F., Bovik, A.C.: A statistical evaluation of recent full reference image quality assessment algorithms. IEEE Trans. Image Process. **15**(11), 3440–3451 (2006)
23. Sievert, S., Ross, D., Jain, L., Jamieson, K., Nowak, R., Mankoff, R.: Next: a system to easily connect crowdsourcing and adaptive data collection. In: Proceedings of the 16th Python in Science Conference, pp. 113–119 (2017)
24. Simons, D.J., Ambinder, M.S.: Change blindness: theory and consequences. Curr. Direct. Psychol. Sci. **14**(1), 44–48 (2005)
25. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)
26. Szegedy, C., et al.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
27. Tramèr, F., Dupré, P., Rusak, G., Pellegrino, G., Boneh, D.: Adversarial: Perceptual ad blocking meets adversarial machine learning. In: Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, pp. 2005–2021 (2019)
28. Wang, Z., Bovik, A.C.: Mean squared error: love it or leave it? a new look at signal fidelity measures. IEEE Signal Process. Mag. **26**(1), 98–117 (2009)
29. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
30. Wolfe, J.M.: Visual search. Curr.Biol. **20**(8), R346–R349 (2010)
31. Yuan, X., He, P., Zhu, Q., Li, X.: Adversarial examples: attacks and defenses for deep learning. IEEE Trans. Neural Netw. Learn. Syst. **30**, 2805–2824 (2019)
32. Zhang, X., Lin, W., Xue, P.: Just-noticeable difference estimation with pixels in images. J. Vis. Commun. Image Representation **19**(1), 30–41 (2008)