*Proceedings of the 2022 Winter Simulation Conference*
*B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C.G. Corlu, L.H. Lee, E.P. Chew, T. Roeder, and P. Lendermann, eds.*

# HUMAN IMPERCEPTIBLE ATTACKS AND APPLICATIONS TO IMPROVE FAIRNESS

Xinru Hua

Department of Computer Science
Stanford University
450 Serra Mall
Stanford, CA 94305, USA

Huanzhong Xu

Institute for Computational and Mathematical Engineering
Stanford University
450 Serra Mall
Stanford, CA 94305, USA

Jose Blanchet

Department of Management Science and Engineering
Stanford University
450 Serra Mall
Stanford, CA 94305, USA

Viet Anh Nguyen

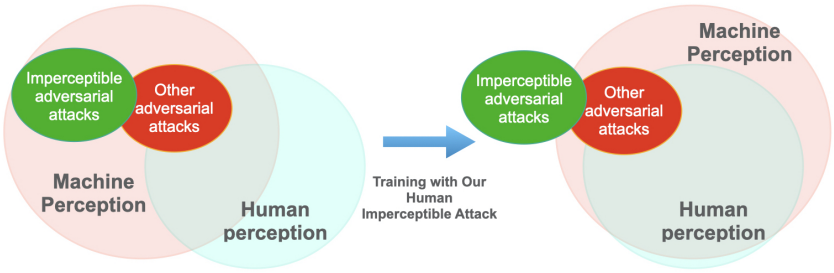
Department of Systems Engineering and Engineering Management
Chinese University of Hong Kong
William M.W. Mong Engineering Building
Shatin, N.T., HONG KONG

## ABSTRACT

Modern neural networks are able to perform at least as well as humans in numerous tasks involving object classification and image generation. However, small perturbations which are imperceptible to humans may significantly degrade the performance of well-trained deep neural networks. We provide a Distributionally Robust Optimization (DRO) framework which integrates human-based image quality assessment methods to design optimal attacks that are imperceptible to humans but significantly damaging to deep neural networks. Through extensive experiments, we show that our attack algorithm generates better-quality (less perceptible to humans) attacks than other state-of-the-art human imperceptible attack methods. Moreover, we demonstrate that DRO training using our optimally designed human imperceptible attacks can improve group fairness in image classification. Towards the end, we provide an algorithmic implementation to speed up DRO training significantly, which could be of independent interest.

## 1 INTRODUCTION

Deep learning models are making strides into our daily life with tremendous successes in diverse areas of applications, such as self-driving cars and face recognition. However, we still lack a fundamental understanding of how deep neural networks (DNNs) perceive and process information. One behavior of DNNs that we do not fully understand is how they are impacted by adversarial attacks. The potential implication of these attacks involves threats in safety and robustness. On the other hand, adversarial attacks provide a method to study the relationship between machine perception and human perception. Over the years, neural network design has been inspired by the ways in which the human brain responds to visual stimuli (Xu and Vaziri-Pashkam 2021; Voulodimos et al. 2018). Although adversarial attacks are intended for DNNs, they may cause differences to human vision systems as well (Zhou and Firestone 2019; Elsayed et al. 2018). In our work, we study adversarial attacks that are designed to primarily affect machine




(a) Adversarial training with our human imperceptible attack aligns machine perception and human perception more closely.


(b) Images from Pakistan (**Left**) and the US (**Right**)

perception and are human imperceptible. We demonstrate that, by training against our human imperceptible attacks, we can improve the attributes of the classification models, such as fairness in classification.

A formal definition of adversarial attacks on image classification (Moosavi-Dezfooli et al. 2016) is the following. Given a classifier $f$, an image $\boldsymbol{x}$, and a cost function $c$ on the image space, an optimal adversarial attack solves $\boldsymbol{\Delta}$ that can change the model's classification results via the smallest budget:

$$\min_{\boldsymbol{\Delta}} c(\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{\Delta}), \quad \text{with} \quad f(\boldsymbol{x}+\boldsymbol{\Delta}) \neq f(\boldsymbol{x}). \tag{1}$$

Traditional adversarial attack methods use $L_p$ distances as the cost function (Goodfellow et al. 2015; Madry et al. 2018; Moosavi-Dezfooli et al. 2016; Tramr et al. 2018). However, as reported in recent literature (Sharif et al. 2018; Wang et al. 2004), $L_p$ distances do not accurately measure differences in human perception. One goal of this paper is to systematically generate adversarial attacks which can mislead DNNs and the differences to the original images are imperceptible to humans.

To design adversarial attacks that humans cannot perceive, the choice of cost function $c$ in Eq. (1) is important. As mentioned before, $L_p$ cost functions may not constrain adversarial attacks to be imperceptible to humans. In our work, we constrain the attacked images to be close to the original image in two choices of cost functions which measure human perceptual distances: structural similarity index measure (SSIM) (Wang et al. 2004) and PieAPP (Prashnani et al. 2018). These human perceptual distances are introduced by works in image quality assessment (IQA) methods. By qualitative and quantitative comparison, we show that our attacked images induce less human perceptual differences to the original images than other state-of-the-art (SOTA) works. Moreover, we aim to better align machine perception with human perception by using the Distributionally Robust Optimization (DRO) framework incorporated with our attack method and we hypothesize that the aligning process reduces the models' biases. As shown in Figure 1a, the left image shows the relationship between human imperceptible attacks and other attacks that are perceived by both humans and machines. As DRO training with our imperceptible attacks discourages the model from perceiving human imperceptible perturbations, it pushes the machine perception circle away from the green circle. Thus, in the right image, machine perception and human perception align more closely, which will be later shown to make the model to perform better on underrepresented groups.

Recently, the DRO framework has been studied extensively in machine learning, because it can be used to compute the most reliable model under distributional uncertainty (Blanchet et al. 2022; Rahimian and Mehrotra 2019; Esfahani and Kuhn 2018). DRO-trained models are able to achieve uniform performance across all groups of data, even on out-of-sample data (Blanchet and Kang 2021; Volpi et al. 2018). In our work, we simulate the worst distributions using our attacked images and apply DRO training to the model with the worst distributions. As shown in Figure 1a, the benefits behind training with our proposed attack is that the models to focus more on perturbations that humans can perceive and use in classification.

As we start to apply DNN models in daily applications, fairness has become more crucial, especially the question of whether the models perform equally well on the data from underrepresented groups. Recent paper reveals that current datasets do not have a uniform distribution on images from all geographical groups and models may infer biases from the unbalanced data. In both of the two popular open-source

datasets: ImageNet and Open Images, approximately half of the images are collected from 2 countries: the United States and Great Britain (Shankar et al. 2017). Moreover, DNNs are suspected to learn spurious features to help classification and the spurious features are learned from the majority groups (de Vries et al. 2019; Khani and Liang 2021). Both works studying classification fairness (Shankar et al. 2017; de Vries et al. 2019) group images by the country where images are collected, so we follow the convention and collect our ImageNet geo-location dataset, where each image has its country information. Figure 1b shows two images of our dataset from the class grocery store, grocery, food market, market. The image from Pakistan is misclassified and the image from the US is not. We assume that humans are more fair in classifying images from different countries. By DRO training algorithms with our proposed adversarial attack method, we prevent the models from learning biases that humans cannot perceive and do not use, and become fairer in image classification task. Finally we show that DRO training with our adversarial attack reduces more biases from the model, when compared to DRO with the Projected Gradient Descent (PGD) attack method (Madry et al. 2018) and PerC attack method (Zhao et al. 2020).

Our work's contributions are summarized as follows:

1. We connect human perceptual distance PieAPP with the DRO framework to generate adversarial attacks that are imperceptible to humans and attack classification models successfully. We use methods in the human vision learning area to show that our attacks are less perceptible to humans than other SOTA imperceptible attacks. We add a confidence parameter to our algorithm, so our method with high confidence is the most successful method against two defense methods.
2. We collect a dataset from ImageNet (Russakovsky et al. 2015), a real-world dataset, with country information, and design a general framework to quantize biases in classification models. We design two hypothesis tests to test whether a model has biases in classification significantly and compare if our training method significantly reduce biases than other methods.
3. We provide an algorithmic implementation of independent interest which can speed up DRO training and sample models efficiently.

This paper unfolds as follows. In Section 2, we conduct a comprehensive literature review. In Section 3, we introduce our adversarial attack method that computes optimal imperceptible attacks. We also show numerical comparison results and image comparisons. In Section 4, we solve the DRO problem and compare fairness in DRO trained models with two other adversarial attack methods.

## 2 RELATED WORK

**Adversarial attacks.** Since the seminal work of Goodfellow et al. (2015), there is a surge of papers on adversarial attacks (Carlini and Wagner 2017; Madry et al. 2018; Moosavi-Dezfooli et al. 2016; Kurakin et al. 2018; Dong et al. 2018). Some papers deploy different distances: Wasserstein distance (Wong et al. 2019) or the human perceptual distance (Zhao et al. 2020; Laidlaw et al. 2021). There are a number of adversarial attacks with different mechanism: sparse adversarial attacks (Zhu et al. 2021; Su et al. 2019), spatial perturbations (Zeng et al. 2019; Aydin et al. 2021), contour region attack (Na et al. 2021), and black-box attack methods (Guo et al. 2019; Ilyas et al. 2018). A comprehensive review for the adversarial attack methods can be found in a recent review paper (Akhtar et al. 2021).

**Adversarial attack and human vision.** Despite the fact that adversarial attacks are designed towards DNNs, two recent papers discover that attacks also have influences on human vision (Zhou and Firestone 2019; Elsayed et al. 2018). Zhao et al. (2018) generate adversarial attacks that are semantically meaningful, that can be perceived by humans. Madry et al. (2018) also reports that $L_2$ based attacks can be large enough to cause misclassification by humans.

**Human perceptual distance.** We need distance functions to measure differences in human perception to truly constraint adversarial attacks in human perception. Image quality assessment (IQA) methods study how to measure human perceptual distance. Traditional IQA methods include SSIM (Wang et al. 2004),

MS-SSIM (Wang et al. 2003) FSIM (Zhang et al. 2011), and PSNR (Hor and Ziou 2010). DNN-based IQA methods include DISTS (Ding et al. 2022), PieAPP (Prashnani et al. 2018), LPIPS (Zhang et al. 2018), and PIM (Bhardwaj et al. 2020).

**Distributionally Robust Optimization (DRO).** As people care more about models' robustness in the extreme circumstances, DRO framework gain a lot of interests recently. There have been a number of theoretical work on DRO and Optimal Transport (Kuhn et al. 2019; Kuhn et al. 2019). Especially, Blanchet and Murthy (2019) proves the useful strong duality results enabling to solve the DRO problem. Sinha et al. (2018) first introduces combining DRO framework and adversarial attacks.

**Fairness.** Many recent papers discover unfairness in image classification and object detection models (de Vries et al. 2019; Buolamwini and Gebru 2018). Specifically, these papers point out that neural network models discriminate against underrepresented groups. One possible explanatory factor of unfairness is that the open-source datasets are unbalanced (Shankar et al. 2017). Yang et al. (2020), Gong et al. (2012) starts to fix the datasets by collecting data that are representative among all demographics. In the natural language processing community, recent works discover that word embedding models learn biases from data (Bolukbasi et al. 2016; Caliskan et al. 2017). A comprehensive review on fairness in machine learning can be found in Mehrabi et al. (2021).

## 3 METHOD

We consider the following DRO problem which finds the model that minimizes the expected loss under the worst-case distribution:

$$\min_{\theta} \sup_{P:D(P,P_0)\leq\delta} \mathbb{E}_P[\ell(\theta;X,Y)], \tag{2}$$

where $\theta$ is the model parameter and $\ell$ is the pre-specified loss function. The distribution $P_0$ is the empirical distribution of the joint random vector $(X,Y)$ constructed from the training data, $D$ is a distance metric between probability distributions, and $\delta$ is the size of the distributional uncertainty. Similar to (Sinha et al. 2018), we choose the Wasserstein distance as our metric $D$. Specifically, let $c((\boldsymbol{x},y),(\boldsymbol{x}',y'))$ denote the cost function to measure the similarity between two samples $(\boldsymbol{x},y)$ and $(\boldsymbol{x}',y')$, and $\Gamma(P,P_0)$ denote the set of all joint distributions of $(X,Y)$ and $(X',Y')$ with marginals $P$ and $P_0$, then the metric $D$ is given by

$$D(P,P_0) = \inf_{\gamma\in\Gamma(P,P_0)} \mathbb{E}_\gamma[c((X,Y),(X',Y'))].$$

By (Blanchet and Murthy 2019, Theorem 1), the DRO problem Eq. (2) is equivalent to

$$\min_{\theta} \inf_{\lambda\geq 0} \lambda\delta + \mathbb{E}_{P_0}[\phi_\lambda(\theta;X_0,Y_0)], \tag{3}$$

where the robust surrogate loss $\phi_\lambda$ is defined by

$$\phi_\lambda(\theta;\boldsymbol{x}_0,y_0) = \sup_{\boldsymbol{x}} \ \ell(\theta;\boldsymbol{x},y_0) - \lambda c((\boldsymbol{x},y_0),(\boldsymbol{x}_0,y_0)). \tag{4}$$

Here we use a separable ground cost $c((\boldsymbol{x},y),(\boldsymbol{x}',y')) = c_0(\boldsymbol{x},\boldsymbol{x}') + \infty\cdot\mathbb{1}\{y\neq y'\}$, which penalizes infinitely the discrepancy between the image labels $y$ and $y'$, and $c_0$ measures the dissimilarity between the images $\boldsymbol{x}$ and $\boldsymbol{x}'$. In this paper, we focus on the PieAPP model for the ground cost with $c_0(\boldsymbol{x},\boldsymbol{x}') = \text{PieAPP}(\boldsymbol{x},\boldsymbol{x}')$. The PieAPP measure (Prashnani et al. 2018) uses a deep neural network to measure the visual differences in terms of human judgment between images. The PieAPP can be applied on images of any size larger than $64\times 64$. Further, PieAPP is unique in its novel pairwise preference probability: when compared to a reference image, PieAPP encodes the probability that humans think one image is more similar than another image. Pairwise comparison is more robust because humans may have clear preferences between all pairs of images but cannot assign scores to all images. Another advantage is that PieAPP does not depend on any existing architectures or pretrained models, as opposed to LPIPS and DISTS.

Given a fixed training image $(\boldsymbol{x}_0,y_0)$, we solve problem (4) to obtain the adversarial attack.

(a) The predictions are: 0, 2, 2 (b) The predictions are: 1, 4, 4 (c) The predictions are: 4, 9, 9

Figure 2: **Left**: Original image, **Middle**: Our one-step attack, **Right**: One-step PGD attack ($L_2$). We show three examples in MNIST that are successfully attacked by both methods. Although all the attacks have the same $L_2$ distance to the original image, our attack does not change the structure of the numbers and PGD attack method makes the images to resemble the images of predicted labels (2,4,9).

### 3.1 SSIM-based Attack

In this subsection, we show a teaser example of $c_0 = 1 - \text{SSIM}$. The structural similarity index measure (SSIM) (Wang et al. 2004) is a reward function on two grayscale images that captures structural similarity in the two images. Because $\text{SSIM} \leq 1$, we use $1 - \text{SSIM}$ as a cost function. We derive a one-step solution when using this $c_0$ function, and compare with one-step PGD attack in Figure 2.

### 3.2 PieAPP-based Attack

Now we focus on choosing PieAPP as our human perceptual distance $c_0$. We attack RGB images of size $3 \times 299 \times 299$ and use gradient descent to solve problem (4), described in Algorithm 1. We incorporate a confidence parameter $a \geq 0$ in our early-stop mechanism to enhance the strength of our attacks, shown at line 5, so they can attack the defended images successfully. At line 10, we truncate $\boldsymbol{x}_{\text{adv}}$ to have the same precision as an RGB image. In our experiments, we choose a ResNet-50 model pretrained on ImageNet (He et al. 2016) as the model $\theta$, cross-entropy loss as $\ell$, $N = 100$, $\varepsilon = 0.1$, $\lambda = 1$ and confidence $a = \{0, 1, 5\}$.

**Algorithm 1** Attack an image

**Input:** image $\boldsymbol{x}$, label $y$, classification model $\theta$, loss function $\ell$, confidence $a$, cost function $c_0$, number of iterations $N$, step size $\varepsilon$

**Output:** adversarial image $\boldsymbol{x}_{\text{adv}}$

1: **Initialize:** $\boldsymbol{x}_{\text{adv}} \leftarrow \boldsymbol{x}$
2: **for** $k = 1, 2, \ldots, N$ **do**
3: $\quad logits = \theta(\boldsymbol{x}_{\text{adv}})$ ▷ $logits_i$ is the logits before softmax for class $i$
4: $\quad$ **if** $\max_i logits_{i \neq y} - logits_y > a$ **then**
5: $\quad\quad$ **return** $\boldsymbol{x}_{\text{adv}}$
6: $\quad$ **end if**
7: $\quad \boldsymbol{\Delta} = \frac{\partial \ell(\theta; \boldsymbol{x}_{\text{adv}}, y)}{\partial \boldsymbol{x}_{\text{adv}}} - \lambda \frac{\partial c_0(\boldsymbol{x}, \boldsymbol{x}_{\text{adv}})}{\partial \boldsymbol{x}_{\text{adv}}}$
8: $\quad \boldsymbol{x}_{\text{adv}} \leftarrow \boldsymbol{x}_{\text{adv}} + \varepsilon \boldsymbol{\Delta}$
9: $\quad$ Validate $\boldsymbol{x}_{\text{adv}}$ ▷ Validate $\boldsymbol{x}_{\text{adv}}$ as an RGB image
10: **end for**
11: **return** $\boldsymbol{x}_{\text{adv}}$

We compare our method with PGD ($L_2$) with 100 iterations and early-stop mechanism. The formulation is $\boldsymbol{x}^0 = \boldsymbol{x}_0$, $\boldsymbol{x}^n = \boldsymbol{x}^{n-1} + \varepsilon \nabla_{\boldsymbol{x}} \ell(\theta; \boldsymbol{x}^{n-1}, y_0) / \|\nabla_{\boldsymbol{x}} \ell(\theta; \boldsymbol{x}^{n-1}, y_0)\|_2$. We also compare with two latest SOTA methods on human imperceptible attacks, NPTM (Laidlaw et al. 2021) and PerC (Zhao et al. 2020). Specifically, we compare with NPTM (PPGD) and NPTM (LPA), the two main methods in the NPTM paper, and with PerC_AL, the faster and less perceptible method in PerC paper. Other methods, for example, FGSM (Goodfellow et al. 2015), C&W (Carlini and Wagner 2017), DDN (Rony et al. 2019), PGD ($L_\infty$), StAdv (Xiao et al. 2018), and ReColorAdv (Laidlaw and Feizi 2019), are compared in the PerC and NPTM

paper, so we do not repeat the comparison. Different from PerC and NPTM, our attack method directly solves the inner optimization problem (4). PerC_AL alternates between the two goals of attacking the image successfully and minimizing the perceptual distance, while our method combines the two goals in a single step. NPTM (PPGD) and NPTM (LPA) require an extra projection step, while our method does not.

Our attack method is evaluated on a subset of the ImageNet, which contains real-world images and is the same dataset as (Zhao et al. 2020). Since the dataset has 1000 images and we plan to compare against four other methods, involving humans to judge every pair of images is expensive. Thus, we apply two human perceptual distances and one salient object detection network as proxies of human vision to measure the differences in human perception. Note that the two human perceptual distances do not include PieAPP, because our method minimizes PieAPP distance and we want to objectively compare our method with other methods. In our comparison table, the success rate is defined as the number of attacked images that labels change from correct to incorrect divided by the number of correctly classified images.

Figure 3 provides two examples to visually compare the quality of attacks. First, we apply two Image Quality Assessment (IQAs) methods, LPIPS (Zhang et al. 2018) and DISTS (Ding et al. 2022), to quantify the perceptual distance between two images in human vision. The numerical results are given in Table 1.

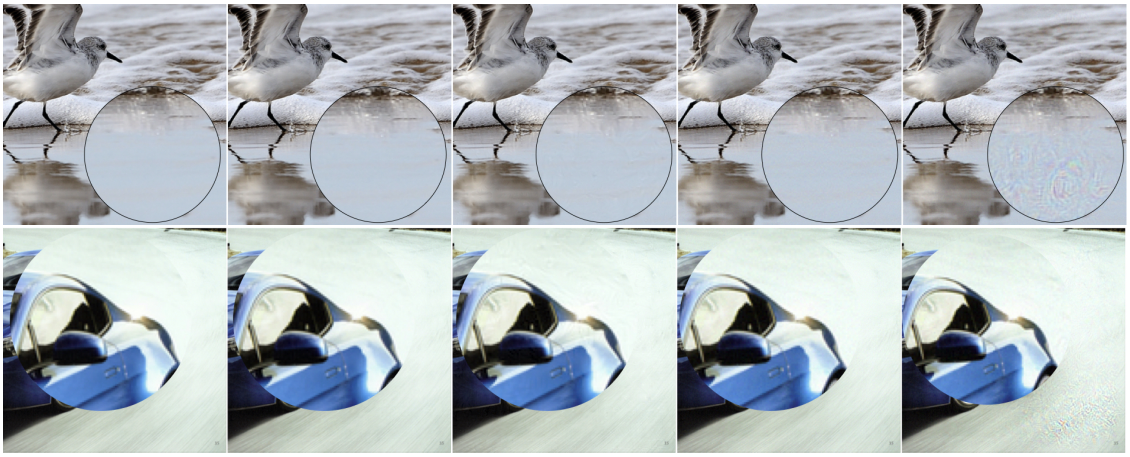



Figure 3: The comparison between the original image and adversarial attacks. We do not include PGD attack images here due to similar visual quality as ours. We zoom in the regions where we can perceive the differences.The PerC_AL images have noticeable marble effects in both images. Both LPA images have noticeable sandy noises compared with other images and both PPGD images have an area of noises.

Table 1: The distances measure the difference between attacked images and original images. The LPIPS and DISTS values are scaled by 1000. We embolden the smallest distance values in each column. Our method with $a = 0$ has the **smallest** human perceptual distances (LPIPS and DISTS), despite larger $L_p$ distances than PGD.

| **Approach** | **Success Rate (%)** | **Distance in adversarial images** | | | | |
|---|---|---|---|---|---|---|
| | | $L_1$ | $L_2$ | $L_\infty$ | LPIPS | DISTS |
| PerC_AL | 100 | 633.12 | 2.22 | 0.085 | 33.96 | 33.82 |
| PGD ($L_2$) | 100 | **592.74** | **1.56** | **0.005** | 7.82 | 8.77 |
| NPTM (PPGD) | 95.75 | 2544.21 | 6.60 | 0.115 | 81.57 | 51.08 |
| NPTM (LPA) | 99.78 | 2157.77 | 5.31 | 0.049 | 51.64 | 35.92 |
| Ours ($a = 0$) | 100 | 783.86 | 1.91 | 0.006 | **7.30** | **8.17** |
| Ours ($a = 1$) | 100 | 1965.10 | 4.45 | 0.014 | 22.24 | 21.47 |
| Ours ($a = 5$) | 100 | 3925.55 | 8.53 | 0.029 | 44.89 | 40.05 |

Then we test our adversarial attack method's effectiveness against defense methods. Without knowing which adversarial attack is applied to the image, there are generic defense methods against attacks. We test our method on two such defense methods: **JPEG compression** (Das et al. 2018; Guo et al. 2018) and **bit depth reduction** (Guo et al. 2018; He et al. 2017). The comparison in Figure 4 shows that our attack method is more successful when comparing with attack methods with similar perceptibility.

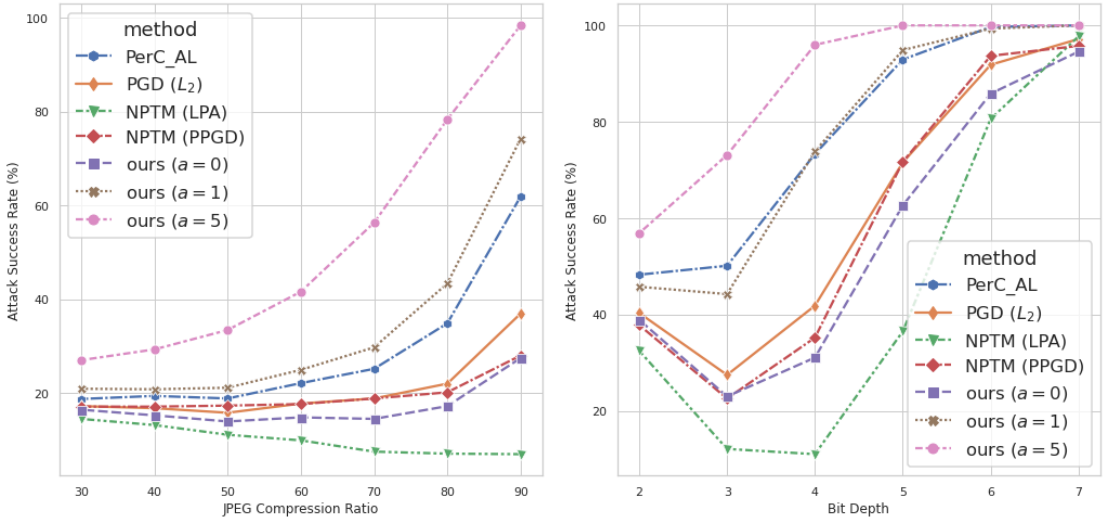



Figure 4: **Left**: JPEG compression ratio defense. **Right**: Bit depth defense. Our method with $a = 5$ has the highest attack success rate on both defense while being less perceptible than NPTM (PPGD)'s images (Table 1). Our method with $a = 1$ generates less perceptible attack images than PerC_AL and has a higher attack success rate in JPEG compression defense. Note that a higher confidence value make the adversarial attacks more perceptible, as there is a trade-off between the attacks' imperceptibility and the strength.

## 4 GEOGRAPHICAL FAIRNESS IN CLASSIFICATION

The nature of training DNNs, which is minimizing the summation of the losses evaluated on the training samples, means that the models may favor features from the majority group and their performance may degrade on the underrepresented groups. Recent works reveal that two open-source large image datasets, such as ImageNet and Open Images, are severely unbalanced in the geographical location, and many object recognition systems do not perform equally well across different geographical groups (Shankar et al. 2017; de Vries et al. 2019). Our collected ImageNet geo-location dataset contains 43.2% images from the US and 12.8% images from the UK. The top five countries that the most images come from are the US, the UK, Canada, Australia, and Germany. This leads to our concern that image classification models train better on the images from higher income level countries, so we want to study and alleviate models' biases related to the images' income levels. We assume human are rather fair in classifying images of different income levels. Since we can successfully attack a classifier by human imperceptible attacks, the classifier utilizes features that humans cannot see and do not use, which tend to be features causing biases. By adversarial training with the human imperceptible perturbations, we reduce the model's dependence on these imperceptible features. In that way, the model performs more equally well across geographical groups.

### 4.1 Bias Metric

We first introduce a general method to measure a model's biases on images from countries with different income levels. Let $\mu$ be a probability model defined on the space $\mathscr{X} \times \mathscr{Y}$ of image and label pairs, and dataset $\mathscr{D}$ is a sample of $N$ independent and identically distributed (image, label) pairs, each following distribution $\mu$. Let $(\boldsymbol{x}, y)$ be one test sample independent of the training set $\mathscr{D}$ following distribution $\mu$

and $Q$ be the joint distribution of the $N+1$ samples from $\mu$. We write $\mathbb{E}$ as the expectation under $Q$. We use $\mathbb{1}(\boldsymbol{x}, y; \theta)$ to represent a Bernoulli event that model $\theta$ classifies correctly the input image $\boldsymbol{x}$ in class $y$ or not. For each image $\boldsymbol{x}$, we can access its country information. For this country, we can get its income level and define as $\mathrm{g}(\boldsymbol{x})$. For a fair classifier and a random out-of-sample image, the classification accuracy should be *independent* of the income level of the image. We choose the following condition to test:

$$\mathbb{E}[\mathbb{1}(\boldsymbol{x}, y; \theta(\mathscr{D}))] = \mathbb{E}[\mathbb{1}(\boldsymbol{x}, y; \theta(\mathscr{D}))|\mathrm{g}(\boldsymbol{x})].$$

This means the classification accuracy is uncorrelated with the income level of the image, which is a necessary condition for the independence to hold. Given this necessary condition, we group the images by the country where they are from and each country is associated with a corresponding income level. Then we study the correlation between the income level and the classification accuracy of each country.

We test $\theta$ on $\mathscr{D}$ and compute the accuracy by groups. We denote the groups as $\{\mathbf{g}_i, \mathbf{p}_i\}$, where $\mathbf{g}_i$ is the per capita GDP of $i$th country in log scale and $\mathbf{p}_i$ is the accuracy of classifying the images in $i$th country. We assume that the error in accuracy of each country is negatively related to the country's number of images, so we write a diagonal covariance matrix as $\mathbf{\Sigma}_{ii} = 1/\sqrt{n_i}$, where $n_i$ is the number of images in the $i$th country. Then we run generalized least squares with $\mathbf{\Sigma}$ on data $\{\mathbf{g}_i, \mathbf{p}_i\}$ and obtain a linear estimator $\mathbf{p} = \beta\mathbf{g} + \varepsilon$. We use the slope $\beta$ as our metric, since it captures how severely the accuracy is linearly correlated with the income. In Section 4.3, we introduce two hypothesis tests to test and compare the significance of the biases. We use a significance level of 0.05 in the statistical tests.

### 4.2 DRO Algorithms

We design Algorithm 2 to generate an augmented dataset $\mathscr{D}_{\text{rob}}$ and Algorithm 3 to approximate the solution to the DRO problem (3). To conduct hypothesis testing, we run Algorithm 3 50 times to sample 50 values of $\theta$'s based on $\mathscr{D}_{\text{rob}}$ and the randomness of $\theta$ comes from stochastic gradient descent.

**Algorithm 2** Generate an adversarial dataset $\mathscr{D}_{rob}$

**Input:** initial model $\theta_0$, learning rate $\alpha$, dataset $\mathscr{D} = \{x_i, y_i\}_{i=1,\dots N}$, number of steps $T_1$
**Output:** robust dataset $\mathscr{D}_{\text{rob}} = \{x_i, y_i, P_i\}_{i=1,\dots M}$

**Initialize:** $\theta = \theta_0, \mathscr{D} = \{x_i, y_i, P_i\}_{i=1,\dots N}$ with $P_i = 1$
**for** $k = 1, 2, \dots, T_1$ **do**
    Sample $\{x_i, y_i, P_i\}_{i=1,\dots N}$ proportionally to the weights $P_i$ with replacement from dataset $\mathscr{D}$
    **for** $i = 1, 2, \dots, N$ **do**
        $\theta \leftarrow \theta - \alpha P_i \nabla_\theta \ell(\theta; x_i, y_i)$
        Input $\theta, x_i, y_i$ to Algorithm 1 to generate attack $\{x_i', y_i\}$
        Append $\{x_i', y_i, P_i\}$ to dataset $\mathscr{D}$ with weight $P_i = (k-1)N + i$
    **end for**
**end for**
**return** dataset $\mathscr{D}_{rob}$

The intuition behind Algorithm 2 is that the outer loop chooses batches of size $N$ and the batches are sampled biased towards recent iterations. In turn, the adversarial examples are added in the inner loop corresponding to the current optimization model parameters, which are updated according to standard stochastic gradient descent. The overall result is similar to a two-time-scale stochastic approximation algorithm (Borkar 1997), which will be analyzed in future work. Compared with the algorithm in (Volpi et al. 2018), our approach's benefit is that we can sample approximately-solved DRO models efficiently.

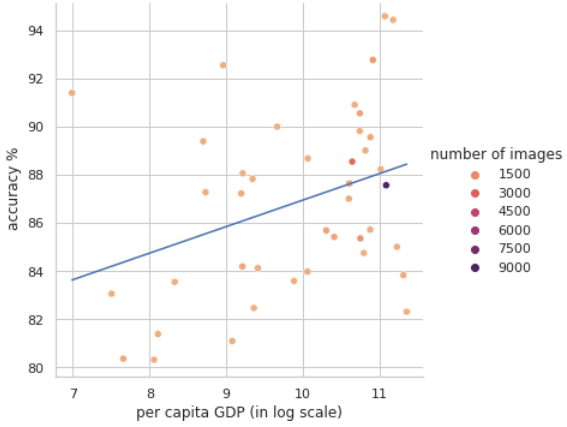
**Algorithm 3** DRO training with a given adversarial dataset

**Input:** initial model $\theta_0$, learning rate $\alpha$, robust dataset $\mathscr{D}_{\text{rob}} = \{x_i, y_i, P_i\}_{i=1,\ldots M}$, number of steps $T_2$
**Output:** DRO trained model: $\theta$

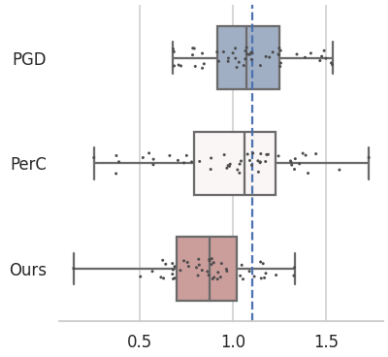
**Initialize:** $\theta = \theta_0$
**for** $k = 1, 2, \ldots, T_2$ **do**
  **for** $i = 1, 2, \ldots, M$ **do**
    Sample $\{x_i, y_i\}$ proportionally to the weights $P_i$ with replacement from dataset $\mathscr{D}_{\text{rob}}$
    Set $\theta \leftarrow \theta - \alpha P_i \nabla_\theta \ell(\theta; x_i, y_i)$
  **end for**
**end for**
**return** model $\theta$

### 4.3 Results

Given a pretrained ResNet-50 model as $\theta_0$ and computed linear regression model $\mathbf{p} = \beta_0 \mathbf{g} + \varepsilon$ (see Figure 5a), we test whether there exists a significantly non-zero linear relationship between $\mathbf{p}$ and $\mathbf{g}$ using the following null and alternative hypotheses: $H_0 : \beta_0 = 0 \quad \text{versus} \quad H_1 : \beta_0 \neq 0$. An F-test (Hahs-Vaughn and Lomax 2020) returns the computed score $F_0 = 5.392$ with the degrees of freedom $\nu_1 = 1, \nu_2 = 39$ and the probability value $\mathbb{P}(f > F_0) = 0.02554$. Thus, we can reject the null hypothesis and conclude that there exists a significant linear relationship between $\mathbf{g}$ and $\mathbf{p}$. For this pretrained model, group-fairness is *not* satisfied. After we train 50 models and compute 50 linear estimators with the PGD, PerC and our attack method respectively, we use two standard t-tests (PGD versus Ours and PerC versus Ours) to compare the distributions of $\beta$s (Sheynin 1995) to evaluate the magnitude of biases. We denote $\beta_m$ as our model's mean and $\beta'_m$ as the comparing model's mean and conduct the hypothesis testing: $H_0 : \beta_m \geq \beta'_m \quad \text{versus} \quad H_1 : \beta_m < \beta'_m$.




(a) We show a linear regression model with independent variable $\mathbf{g}$ and dependent variable $\mathbf{p}$, which are tested with the pretrained model $\theta_0$. Each dot represents one country and the color denotes the number of images in this country.




(b) Each box shows the distribution of $\beta$ obtained from each method. The blue dashed line represents $\beta_0$ of the original pretrained model. Training with our attack method reduces the biases the most significantly.

Figure 5: Geographical Fairness Results.

Against the PGD method, we compute the t-value to be 3.85 with a probability 0.00017. Against the PerC method, we compute the t-value to be 2.95 with a probability 0.00242. Thus, we can reject the null hypothesis for both competing methods, and we may conclude that the PGD and PerC trained models have higher biases than our trained model. Figure 5b illustrates the three distribution of $\beta$s, which illustrates our $\beta$s are of smaller magnitudes than the other two methods' $\beta$s. All of our algorithms are run on one

NVIDIA V100 Tensor Core GPU. Alg. 2 with $T_1 = 3$ takes about 9 hours and Alg. 3 with $T_2 = 3$ 50 times takes about 25 hours. The memory usage of Alg. 2 is the same as other adversarial training algorithms.

## 5 CONCLUSION

In this work, we present a method to generate human imperceptible adversarial attacks and design two DRO training algorithms to simulate the most difficult distributions. We show that our adversarial attack method can generate successful and least human perceptible attacks compared with other SOTA methods. For the ImageNet dataset and a model trained on it, we test the existence of inherent unfairness, such as geo-location biases. After testing two collections of models that are respectively trained by the DRO algorithm with our attack method and with the PGD attack method, our method improves fairness more significantly than the PGD method. Our hypothesis tests provide a general framework to test fairness on the space of models conditioned on datasets. The limitation of our method is that we do not have enough computational resources or data to sample datasets, so we can only condition on one dataset and randomize the models. By generating a variation of adversarial attacks, our method mitigates biases in the given dataset. We hope future work will incorporate the randomness in datasets and conduct the complete test in fairness. We also hope our work can help understand the differences between machine perception and human perception, and bridge the two areas of adversarial attacks and fairness in machine learning.

## REFERENCES


Akhtar, N., A. Mian, N. Kardan, and M. Shah. 2021. "Advances in adversarial attacks and defenses in computer vision: A survey". *IEEE Access* 9:155161–155196.

Aydin, A., D. Sen, B. T. Karli, O. Hanoglu, and A. Temizel. 2021. "Imperceptible Adversarial Examples by Spatial Chroma-Shift". In *October 20th, Proceedings of the 1st International Workshop on Adversarial Learning for Multimedia*. online, 8-14.

Bhardwaj, S., I. Fischer, J. Ballé, and T. Chinen. 2020. "An Unsupervised Information-Theoretic Perceptual Quality Metric". In *Advances in Neural Information Processing Systems*, edited by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, 13–24: Curran Associates, Inc.

Blanchet, J., and Y. Kang. 2021. "Sample out-of-sample inference based on Wasserstein distance". *Operations Research* 69(3):985–1013.

Blanchet, J., and K. Murthy. 2019. "Quantifying distributional model risk via optimal transport". *Math. Oper. Res.* 44(2):565–600.

Blanchet, J., K. Murthy, and F. Zhang. 2022. "Optimal transport-based distributionally robust optimization: Structural properties and iterative schemes". *Mathematics of Operations Research* 47(2):1500–1529.

Bolukbasi, T., K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. 2016. "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings". In *Advances in Neural Information Processing Systems*, edited by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Volume 29: Curran Associates, Inc.

Borkar, V. S. 1997. "Stochastic approximation with two time scales". *Systems & Control Letters* 29(5):291–294.

Buolamwini, J., and T. Gebru. 2018. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification". In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, edited by S. A. Friedler and C. Wilson, 77–91. Stockholm, Sweden: PMLR.

Caliskan, A., J. J. Bryson, and A. Narayanan. 2017. "Semantics derived automatically from language corpora contain human-like biases". *Science* 356(6334):183–186.

Carlini, N., and D. Wagner. 2017. "Towards Evaluating the Robustness of Neural Networks". In *2017 IEEE Symposium on Security and Privacy (SP)*. May 22nd-26th, San Jose, USA, 39-57.

Das, N., M. Shanbhogue, S.-T. Chen, F. Hohman, S. Li, L. Chen, M. E. Kounavis, and D. H. Chau. 2018. "SHIELD: Fast, Practical Defense and Vaccination for Deep Learning Using JPEG Compression". In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. August 19th-23rd, London, United Kingdom, 196204.

de Vries, T., I. Misra, C. Wang, and L. van der Maaten. 2019. "Does Object Recognition Work for Everyone?". In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. June 16th-20th, Long Beach, USA.

Ding, K., K. Ma, S. Wang, and E. P. Simoncelli. 2022, May. "Image Quality Assessment: Unifying Structure and Texture Similarity". *IEEE transactions on pattern analysis and machine intelligence* 44(5):25672581.

Dong, Y., F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li. 2018. "Boosting Adversarial Attacks with Momentum". In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. June 18th-22nd, Salt Lake City, USA, 9185-9193.

Elsayed, G. F., S. Shankar, B. Cheung, N. Papernot, A. Kurakin, I. Goodfellow, and J. Sohl-Dickstein. 2018. "Adversarial Examples That Fool Both Computer Vision and Time-Limited Humans". In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, 39143924. Red Hook, NY, USA: Curran Associates Inc.

Esfahani, P. M., and D. Kuhn. 2018. "Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations". *Math. Program.* 171(1):115–166.

Gong, B., F. Sha, and K. Grauman. 2012. "Overcoming dataset bias: An unsupervised domain adaptation approach". In *NIPS Workshop on Large Scale Visual Recognition and Retrieval*. Dec 3rd-8th, Lake Tahoe, Nevada, USA.

Goodfellow, I. J., J. Shlens, and C. Szegedy. 2015. "Explaining and Harnessing Adversarial Examples". In *3rd International Conference on Learning Representations*. May 7th-9th, San Diego, CA, USA.

Guo, C., J. Gardner, Y. You, A. G. Wilson, and K. Weinberger. 2019. "Simple Black-box Adversarial Attacks". In *Proceedings of the 36th International Conference on Machine Learning*, edited by K. Chaudhuri and R. Salakhutdinov, 2484–2493. Long Beach, USA: PMLR.

Guo, C., M. Rana, M. Cisse, and L. van der Maaten. 2018. "Countering Adversarial Images using Input Transformations". In *International Conference on Learning Representations*. April 30th-May 3rd, Vancouver, Canada.

Hahs-Vaughn, D. L., and R. G. Lomax. 2020. *An introduction to statistical concepts*. 4th ed. Routledge.

He, K., X. Zhang, S. Ren, and J. Sun. 2016. "Deep Residual Learning for Image Recognition". In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. June 26th-July 1st, Las Vegas, USA, 770-778.

He, W., J. Wei, X. Chen, N. Carlini, and D. Song. 2017. "Adversarial example defense: Ensembles of weak defenses are not strong". In *11th USENI workshop on offensive technologies (WOOT 17)*. August 14th-July 15th, Vancouver, Canada.

Hor, A., and D. Ziou. 2010. "Image Quality Metrics: PSNR vs. SSIM". In *2010 20th International Conference on Pattern Recognition*. August 23rd-26th, NW Washington, USA, 2366-2369.

Ilyas, A., L. Engstrom, A. Athalye, and J. Lin. 2018. "Black-box Adversarial Attacks with Limited Queries and Information". In *Proceedings of the 35th International Conference on Machine Learning*, edited by J. Dy and A. Krause, 2137–2146. Stockholm, Sweden: PMLR.

Khani, F., and P. Liang. 2021. "Removing spurious features can hurt accuracy and affect groups disproportionately". In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. March 3rd-10th, online, 196-205.

Kuhn, D., P. M. Esfahani, V. A. Nguyen, and S. Shafieezadeh-Abadeh. 2019. "Wasserstein distributionally robust optimization: Theory and applications in machine learning". In *INFORMS TutORials in Operations Research*. October 21st, Seattle, USA, 130–166.

Kurakin, A., I. J. Goodfellow, and S. Bengio. 2018. "Adversarial examples in the physical world". In *Artificial intelligence safety and security*, 99–112. Chapman and Hall/CRC.

Laidlaw, C., and S. Feizi. 2019. "Functional Adversarial Attacks". In *Advances in Neural Information Processing Systems*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, Volume 32: Curran Associates, Inc.

Laidlaw, C., S. Singla, and S. Feizi. 2021. "Perceptual Adversarial Robustness: Defense Against Unseen Threat Models". In *International Conference on Learning Representations*. May 3rd-7th, online.

Madry, A., A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. 2018. "Towards Deep Learning Models Resistant to Adversarial Attacks". In *International Conference on Learning Representations*. April 30th-May 3rd, Vancouver, Canada.

Mehrabi, N., F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. 2021. "A survey on bias and fairness in machine learning". *ACM Comput. Surv. (CSUR)* 54(6):1–35.

Moosavi-Dezfooli, S., A. Fawzi, and P. Frossard. 2016. "DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks". In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. June 26th-July 1st, Las Vegas, USA, 2574-2582.

Na, H., G. Ryu, and D. Choi. 2021. "Adversarial Attack Based on Perturbation of Contour Region to Evade Steganalysis-Based Detection". *IEEE Access* 9:122308–122321.

Prashnani, E., H. Cai, Y. Mostofi, and P. Sen. 2018. "PieAPP: Perceptual Image-Error Assessment Through Pairwise Preference". In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. June 18th-22nd, Salt Lake City, USA, 1808-1817.

Rahimian, H., and S. Mehrotra. 2019. "Distributionally robust optimization: A review". *arXiv preprint arXiv:1908.05659*.

Rony, J., L. G. Hafemann, L. S. Oliveira, I. B. Ayed, R. Sabourin, and E. Granger. 2019. "Decoupling Direction and Norm for Efficient Gradient-Based L2 Adversarial Attacks and Defenses". In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. June 16th-20th, Long Beach, USA, 4317-4325.

Russakovsky, O., J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. 2015. "ImageNet Large Scale Visual Recognition Challenge". *International Journal of Computer Vision (IJCV)* 115(3):211–252.

Shankar, S., Y. Halpern, E. Breck, J. Atwood, J. Wilson, and D. Sculley. 2017. "No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World". In *NIPS 2017 workshop: Machine Learning for the Developing World*. December 8th-9th, Long Beach, USA.

Sharif, M., L. Bauer, and M. K. Reiter. 2018. "On the Suitability of Lp-Norms for Creating and Preventing Adversarial Examples". In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. June 18th-22nd, Salt Lake City, USA.

Sheynin, O. 1995. "Helmert's Work in the Theory of Errors". *Archive for History of Exact Sciences* 49(1):73–104.

Sinha, A., H. Namkoong, and J. Duchi. 2018. "Certifiable Distributional Robustness with Principled Adversarial Training". In *International Conference on Learning Representations*. April 30th-May 3rd, Vancouver, Canada.

Su, J., D. V. Vargas, and K. Sakurai. 2019. "One pixel attack for fooling deep neural networks". *IEEE Trans. Evol. Comput.* 23(5):828–841.

Tramr, F., A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel. 2018. "Ensemble Adversarial Training: Attacks and Defenses". In *International Conference on Learning Representations*. April 30th-May 3rd, Vancouver, Canada.

Volpi, R., H. Namkoong, O. Sener, J. Duchi, V. Murino, and S. Savarese. 2018. "Generalizing to Unseen Domains via Adversarial Data Augmentation". In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, 53395349. Red Hook, NY, USA: Curran Associates Inc.

Voulodimos, A., N. Doulamis, A. Doulamis, E. Protopapadakis, and D. Andina. 2018, jan. "Deep Learning for Computer Vision: A Brief Review". *Intell. Neuroscience* 2018.

Wang, Z., A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. 2004. "Image quality assessment: from error visibility to structural similarity". *IEEE transactions on image processing* 13(4):600–612.

Wang, Z., E. Simoncelli, and A. Bovik. 2003. "Multiscale structural similarity for image quality assessment". In *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003*. November 9th-12th, Pacific Grove, CA, USA, 1398-1402 Vol.2.

Wong, E., F. Schmidt, and Z. Kolter. 2019. "Wasserstein Adversarial Examples via Projected Sinkhorn Iterations". In *Proceedings of the 36th International Conference on Machine Learning*, edited by K. Chaudhuri and R. Salakhutdinov, 6808–6817. Long Beach, USA: PMLR.

Xiao, C., J.-Y. Zhu, B. Li, W. He, M. Liu, and D. Song. 2018. "Spatially Transformed Adversarial Examples". In *International Conference on Learning Representations*. April 30th-May 3rd, Vancouver, Canada.

Xu, Y., and M. Vaziri-Pashkam. 2021. "Limits to visual representational correspondence between convolutional neural networks and the human brain". *Nat. Commun.* 12(1):1–16.

Yang, K., K. Qinami, L. Fei-Fei, J. Deng, and O. Russakovsky. 2020. "Towards Fairer Datasets: Filtering and Balancing the Distribution of the People Subtree in the ImageNet Hierarchy". In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. January 27th-30th, Barcelona, Spain, 547558.

Zeng, X., C. Liu, Y.-S. Wang, W. Qiu, L. Xie, Y.-W. Tai, C.-K. Tang, and A. L. Yuille. 2019. "Adversarial Attacks Beyond the Image Space". In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4297–4306.

Zhang, L., L. Zhang, X. Mou, and D. Zhang. 2011. "FSIM: A feature similarity index for image quality assessment". *IEEE Trans. Image Process.* 20(8):2378–2386.

Zhang, R., P. Isola, A. A. Efros, E. Shechtman, and O. Wang. 2018. "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric". In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. June 18th-22nd, Salt Lake City, USA, 586-595.

Zhao, Z., D. Dua, and S. Singh. 2018. "Generating Natural Adversarial Examples". In *International Conference on Learning Representations*. April 30th-May 3rd, Vancouver, Canada.

Zhao, Z., Z. Liu, and M. Larson. 2020. "Towards large yet imperceptible adversarial image perturbations with perceptual color distance". In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. June 14th-19th, online, 1039–1048.

Zhou, Z., and C. Firestone. 2019. "Humans can decipher adversarial images". *Nat. Commun.* 10(1):1–9.

Zhu, M., T. Chen, and Z. Wang. 2021. "Sparse and Imperceptible Adversarial Attack via a Homotopy Algorithm". In *Proceedings of the 38th International Conference on Machine Learning*, edited by M. Meila and T. Zhang, 12868–12877. online: PMLR.

## AUTHOR BIOGRAPHIES

**XINRU HUA** is a Ph.D. student at the computer science department at Stanford University. Her email address is huaxinru@stanford.edu.

**HUANZHONG XU** is a Ph.D. student at the Institute of Computational and Mathematical Engineering at Stanford University. His email address is xuhuanvc@stanford.edu.

**JOSE BLANCHET** is a Professor of MS&E at Stanford University, from which he earned his doctorate degree. His email address is jose.blanchet@stanford.edu and his website is https://web.stanford.edu/ jblanche/.

**VIET ANH NGUYEN** is an assistant professor at the Chinese University of Hong Kong. His email address is nguyen@se.cuhk.edu.hk.