Robust Universal Adversarial Perturbations

Changming Xu [1] **Gagandeep Singh** [1 2]

Abstract

Universal Adversarial Perturbations (UAPs) are imperceptible, image-agnostic vectors that cause deep neural networks (DNNs) to misclassify inputs with high probability. In practical attack scenarios, adversarial perturbations may undergo transformations such as changes in pixel intensity, scaling, etc. before being added to DNN inputs. Existing methods do not create UAPs robust to these real-world transformations, thereby limiting their applicability in practical attack scenarios. In this work, we introduce and formulate UAPs robust against real-world transformations. We build an iterative algorithm using probabilistic robustness bounds and construct UAPs robust to transformations generated by composing arbitrary sub-differentiable transformation functions. We perform an extensive evaluation on the popular CIFAR-10 and ILSVRC 2012 datasets measuring our UAPs' robustness under a wide range common, real-world transformations such as rotation, contrast changes, etc. We further show that by using a set of primitive transformations our method generalizes well to unseen transformations such as fog, JPEG compression, etc. Our results show that our method can generate UAPs up to 23% more robust than state-of-the-art baselines.

1. Introduction

Deep neural networks (DNNs) have achieved impressive results in many application domains such as natural language processing (Abdel-Hamid et al., 2014; Brown et al., 2020), medicine (Esteva et al., 2017; 2019), and computer vision (Simonyan and Zisserman, 2014; Szegedy et al., 2016). Despite their performance, they can be fragile in the face of adversarial perturbations: small imperceptible changes added to a correctly classified input that make a DNN misclassify. While there is a large amount of work on generating adversarial perturbations (Szegedy et al., 2013; Goodfellow et al., 2014; Madry et al., 2017; Carlini and Wagner, 2017; Xiao et al., 2018a; Dong et al., 2018; Croce and Hein, 2019; Wang et al., 2019; Zheng et al., 2019; Andriushchenko et al., 2019; Tramèr et al., 2020), these works depend upon unrealistic assumptions about the power of the attacker: the attacker knows the DNN input in advance, generates input-specific perturbations in real-time and *exactly* combines the perturbation with the input before being processed by the DNN. Thus, we argue that these threat models are not realizable in many real-world applications.

Practically feasible adversarial perturbations. In this work, we consider a more practical adversary to reveal real-world vulnerabilities of state-of-the-art DNNs. We assume that the attacker (i) does not know the DNN inputs in advance, (ii) can only transmit additive adversarial perturbations, and (iii) their transmitted perturbations are susceptible to modification due to real-world effects. Examples of attacks in our threat model include adding stickers to the cameras for fooling image classifiers (Li et al., 2019b) or transmitting perturbations over the air for deceiving audio classifiers (Li et al., 2019a). Note that our threat model is distinct from directly generating adversarial examples (i.e. creating physical adversarial objects (Athalye et al., 2018)) which require access to the original input. In Appendix A, we further discuss how our threat model compares to simultaneously transforming the input.

The first two requirements in our threat model can be fulfilled by generating Universal Adversarial Perturbations (UAPs) (Moosavi-Dezfooli et al., 2017). Here the attacker can train a single adversarial perturbation that has a high probability of being adversarial on all inputs in the training distribution. However, as our experimental results show, the generated UAPs need to be combined with the DNN inputs precisely, otherwise they fail to remain adversarial. In practice, changes to UAPs are likely due to real-world effects. For example, a sticker applied to a camera can undergo changes in contrast due to weather conditions or a transmitted perturbation in audio can change due to noise in the transmission channel. This non-robustness reduces the efficacy of attacks created with existing methods (Moosavi-Dezfooli et al., 2017; Shafahi et al., 2020; Li et al., 2019b;a).

[1]Department of Computer Science, University of Illinois Urbana-Champaign, Champaign, USA [2]VMWare, California, USA. Correspondence to: Changming Xu <cxu23@illinois.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

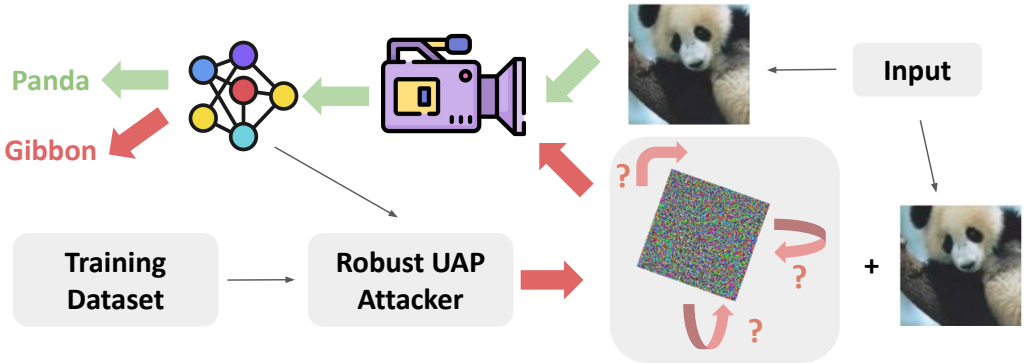

Figure 1. Robust UAP Threat Model: Input Agnostic + Robust to Transmission Transformation

This work: Robust UAPs. To overcome the above limitation, we propose the concept of robust UAPs: perturbations that have a high probability of remaining adversarial on inputs in the training distribution even after applying a set of real-world transformations. The optimization problem in generating robust UAPs (Moosavi-Dezfooli et al., 2017) is made challenging as we are looking for perturbations that are adversarial for a set of inputs as well as to a set of potentially unknown transformations applied to the perturbations. To address this challenge, we make the following main **contributions**:

- We introduce *Robust UAPs* and formulate their generation as an optimization problem. We separate our threat model into two scenarios depending on whether the transformation set is known apriori.
- We design a new method, `RobustUAP`, for constructing robust UAPs. Our method is general and constructs UAPs robust to any transformations generated by composing arbitrary sub-differentiable transformation functions. We provide an algorithm for computing provable probabilistic bounds on the robustness of our UAPs against many practical transformations. We show that in the vision domain we can use a set of primitive transforms (adapted from Modas et al. (2022)) to create *Universally* Robust UAPs.
- We perform an extensive evaluation of `RobustUAP` on state-of-the-art models for the CIFAR-10 (Krizhevsky et al., 2009) and ILSVRC 2012 (Deng et al., 2009) datasets. We compare the robustness of our UAPs under compositions of challenging real-world transformations, such as rotation, contrast change, etc. We show that on both datasets, the UAPs generated by `RobustUAP` are significantly more robust, achieving up to 23% more robustness, than the UAPs generated from the baselines. Furthermore, we show that RobustUAP significantly outperforms UAP in a real-world wireless setting.

Our work is complementary to the development of real-world attacks (Li et al., 2019a;b) in various domains, which require modeling how the universal perturbations change during transmission. `RobustUAP` can improve the efficiency of such attacks by constructing perturbations more robust to real-world transformations than existing methods. Our results using primitive transformations in vision suggest that we can forego domain-specific modeling in other domains given a good set of primitives for that domain. Finally, our preliminary results on adversarial training with robust UAPs suggest that robustness against practical adversaries such as robust UAPs can be achieved without sacrificing as much accuracy as standard adversarial training.

2. Background

In this section, we provide necessary background definitions and notation used in the rest of our work. For the remainder of the paper, let $\mu \subset \mathbb{R}^d$ be the input data distribution, $\mathbf{x} \in \mu$ be an input point with the corresponding true label $y \in \mathbb{R}$, and $f : \mathbb{R}^d \to \mathbb{R}^{d'}$ be our target classifier. We define $f_k(\mathbf{x})$ to be the kth element of $f(\mathbf{x})$ and allow $\hat{f}(\mathbf{x}) = \arg\max_k f_k(\mathbf{x})$ to directly refer to the classification label. We use $\mathbf{v}$ to reference input-specific perturbations and $\mathbf{u}$ to reference UAPs, $\mathbf{v_r}$ and $\mathbf{u_r}$ refer to their robust variants. We provide formal definitions in Appendix B.

Adversarial Examples and Perturbations. An *adversarial example* is a misclassified data point that is *close* (in some norm) to a correctly classified data point (Goodfellow et al., 2014; Madry et al., 2017; Carlini and Wagner, 2017). In this paper, we consider examples $\mathbf{x}'$ generated as $\mathbf{x}' = \mathbf{x} + \mathbf{v}$ where $\mathbf{v}$ is an *adversarial perturbation*.

Universal Adversarial Perturbations. UAPs are single vector, input-agnostic perturbations (Moosavi-Dezfooli et al., 2017). They differ from traditional adversarial attacks, which create perturbations dependent on each input sample. To measure UAP performance, we introduce the notion of universal adversarial success rate (ASR_U), which measures the probability that a perturbation $\mathbf{u}$ when added to $\mathbf{x}$, sampled from μ, causes a change in classification under f. Thus

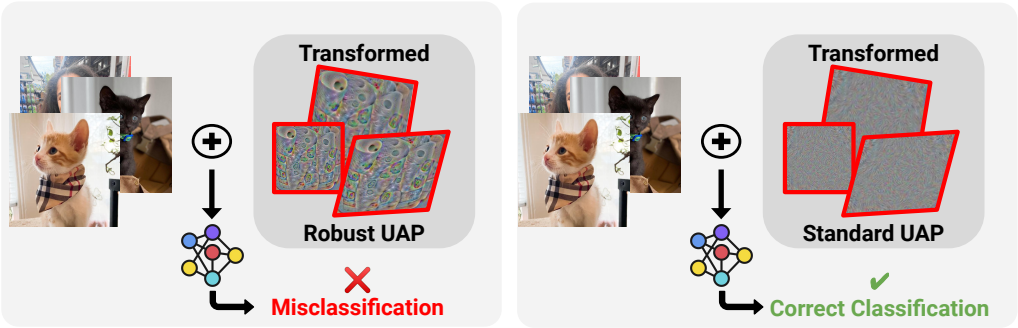

Figure 2. Robust UAPs (left) cause a classier to misclassify on *most* of the data distribution even after transformations are applied on them. Standard UAPs (right) are not robust to transformations and have a low probability of remaining UAPs after transformation.

a perturbation, u, is a UAP given two conditions: its ASR_U is greater than a given threshold, γ, and its norm is small (additive perturbations with small l_p-norm do not affect the semantic content of the image). We pose the construction of UAPs as an expectation minimization problem:

$$\arg\max_{u} \mathop{\mathbb{E}}_{\mathbf{x}\sim\mu} [\delta(\hat{f}(\mathbf{x}+\mathbf{u}), \hat{f}(\mathbf{x}))] \text{ s.t. } ||\mathbf{u}||_p < \epsilon \quad (1)$$

where $\delta(a, b) = 0$ if $a = b$ and 1 otherwise.

3. Robust Universal Adversarial Perturbations

In this section, we first define our notion of transformation sets and neighborhoods in order to define robust UAPs. Here, when we are referencing transformation sets as the ones applied during transmission, if these are unknown, we detail our method for overcoming this in Section 4.1. Formal definitions of all terms can be found in Appendix B.

Transformation Sets and Neighborhoods. We define a transformation set, T, as all transforms, τ, which can be made by composing from a predefined set of bijective sub-differentiable transformation functions. The neighborhood, $N_T(\mathbf{v})$, of a point $\mathbf{v}$ is all points, $\mathbf{v}'$ reachable from $\mathbf{v}$ using transformations from T and which still satisfy $||\mathbf{v}'||_p < \epsilon$.

Example 3.1. *Let T be all transformations represented by a rotation of $\pm 30°$ and scaling of up to a factor of 2, in this case one $\tau \in T$ could be {rotation of $8°$ and scaling a factor of 1.2} in that order and $N_T(\mathbf{v})$ would include any point obtained by applying a transformation $\tau \in T$ on $\mathbf{v}$ which still satisfies $||\tau((v))||_p < \epsilon$.*

Robust UAPs. In order to define robust UAPs we introduce robust universal adversarial success rate. The *robust universal adversarial success rate*, ASR_R, measures the probability that a neighbor of $\mathbf{u_r}$ is also an UAP on μ, i.e. after transformation it maintains high universal ASR above some threshold γ.

Definition 3.2. A *robust UAP*, $\mathbf{u_r}$, is one which *most* points within a neighborhood of $\mathbf{u_r}$ when added to *most* points in μ fool the classifier, f. $\mathbf{u_r}$ satisfies $||\mathbf{u_r}||_p < \epsilon$ and $\text{ASR}_R(f, \mu, T, \gamma, \mathbf{u_r}) > \zeta$, where ζ is the minimum robust UAP probability threshold.

In order to construct robust UAPs, we can pose the following expectation maximization problem:

$$\arg\max_{\mathbf{u_r}} \mathop{\mathbb{E}}_{\mathbf{u_r'}\in N_T(\mathbf{u_r}), \mathbf{x}\sim\mu} [\delta(\hat{f}(\mathbf{x}+\mathbf{u_r'}), \hat{f}(\mathbf{x}))] \text{ s.t. } ||\mathbf{u_r}||_p < \epsilon \quad (2)$$

Here $I : \mathbb{R}^d \to \mathbb{R}$ denotes an indicator function. Taking the expected value over $\mathbf{x} \sim \mu$ maximizes the UAP condition for the transformed perturbation $\mathbf{u_r'}$ while $\mathbf{u_r'} \in N_T(\mathbf{u_r})$ maximizes over the neighborhood. The composition of these conditions in Equation 2 makes it computationally harder than maximizing over only the transformation set, as in EOT (Athalye et al., 2018), or than maximizing over only the data distribution, as in Equation 1.

4. Generating Robust UAP

In this section we discuss how we deal with both known and unknown transformation sets, then describe our approach for optimizing Equation 2. As it would be computationally prohibitive to precisely compute the expected value, we estimate the expected value per batch, $\hat{\mathbf{x}} \subset \mu$, and random set of transformations sampled from T, $\hat{\tau} \subset T$. We then get a Lagrangian relaxation of Equation 2 (see in Appendix C). We describe our two threat models and some intuitive baselines for optimizing Equation 2. We then present our new algorithm, `RobustUAP`.

4.1. Known vs Unknown Transformation Sets

In the above section, we have assumed that the transformations applied during transmission is known to the attacker and used to train the UAP. However, in a real-world attack scenario the attacker may not know precisely what transformations its perturbation will undergo. In such scenarios, they may want their attack to be robust to unseen perturbations. In this case, we propose generating robust UAPs using

a set of primitive transformations. For the image domain, we draw from existing work in the data augmentation area. In their paper, **PRI**mitives of **M**aximum **E**ntropy (PRIME), Modas et al. (2022) define three primitive transformations: spectral, spatial, and color. Using random combinations of these transformations to train, they find that they are able to generalize well to unseen transformations such as frost, JPEG compression, motion blur, etc. We can use these primitive transformations to generate robust UAPs and we show that in practice this generates robust UAPs which generalize well to a variety of unseen transforms. In other domains, we hope that this work helps to inspire finding similar primitive transformation sets.

4.2. Baseline Algorithms

We propose two baseline algorithms for generating robust UAPs. The first method is momemtum based Stochastic Gradient Descent (SGD). We can directly solve Equation 2 using gradient descent. The second baseline is leveraging the standard UAP algorithm from Moosavi-Dezfooli et al. (2017), but instead of computing an adversarial perturbation at each point, we compute a robust adversarial perturbation at each point. More details about both of these baseline algorithms can be found in Appendix G. Both of these algorithms can be seen as naively combining the EoT and UAP algorithms, in the next section we describe `RobustUAP` our algorithm which takes a more principled approach at robust UAP generation.

4.3. Robust UAP Algorithm

The baseline algorithms have two fundamental limitations: (i) they rely on random sampling over the symbolic transformation region, but the sampling strategy does not explicitly try to maximize the robustness of the generated UAP over the entire symbolic region, and (ii) they do not estimate robustness on unsampled transformations. As a result, the baselines yield suboptimal UAPs (as confirmed by our experiments below). To overcome these limitations, we create a method to compute probabilistic bounds for expected robustness on an entire symbolic region. We leverage this method for approximating expected robustness. We make a simplifying assumption that $N_T(\mathbf{u_r})$ has a well-defined, sampleable probability density function (PDF) as we cannot bound robustness for arbitrary transformations. Our experiments show that even though our assumptions do not hold for all the transformation sets considered in this work, they significantly improve the robustness of our generated UAPs. Our approximation of the expected robustness relies on the following theorem:

Theorem 4.1. *Given a perturbation $\mathbf{u_r}$, a neural network f, a finite set of inputs $\mathbf{X}$, a set of transformations T, and minimum universal adversarial success rate $\gamma \in \mathbb{R}$. Let $p(\gamma) = P_{\mathbf{u'_r} \sim N_T(\mathbf{u_r})}(ASR_U(f, \mathbf{X}, \mathbf{u'_r}) > \gamma)$. For $i \in 1 \ldots n$, let $\mathbf{u_r^i} \sim N_T(\mathbf{u_r})$ be random variables with a well defined PDF and $I : \mathbb{R}^d \to \mathbb{R}$ be the indicator function, let*

$$\hat{p}_n(\gamma) = \frac{1}{n} \sum_{i=1}^{n} I(ASR_U(f, \mathbf{X}, \mathbf{u_r^i}) > \gamma) \tag{3}$$

For accuracy level, $\psi \in (0, 1)$, and confidence, $\phi \in (0, 1)$, where $(0, 1)$ is the open interval between 0 and 1. If $n \geq \frac{1}{2\psi^2} \ln \frac{2}{\phi}$ then

$$P(|\hat{p}_n(\gamma) - p(\gamma)| < \psi) \geq 1 - \phi \tag{4}$$

The proof of this theorem can be found in Appendix D. Theorem 4.1 states that with enough samples from the neighborhood of a perturbation, $\mathbf{u_r}$, the adversarial success rate of $\mathbf{u_r}$ on the entire neighborhood is arbitrarily close to the adversarial success rate of $\mathbf{u_r}$ on sampled transformations with probability greater than $1 - \phi$.

Leveraging Theorem 4.1, we create `EstRobustness` which given accuracy, ψ, and confidence, ϕ, returns the ASR_R on a finite set of inputs with probabilistic robustness guarantees under the assumptions of Theorem 4.1. The pseudocode for `EstRobustness` is in Appendix J.

Our algorithm: `RobustUAP`. We leverage Theorem 4.1 and `EstimateRobustness` to develop `RobustUAP`, the pseudocode for which is seen in Algorithm 1. Similar to the SGD baseline, we approximate the expectation in Equation 2 in batches. We first sample transformations from the PDF of the neighborhood. We set the number of transformations, n, based on Theorem 4.1 to satisfy the desired confidence level and accuracy. For each gradient step, we compute the mean loss over the current batch and set of sampled transforms (line 8). For each set of batch

Algorithm 1 Robust UAP Algorithm

1: Initialize $\mathbf{u_r} \leftarrow 0, n \leftarrow \lceil \frac{1}{2\psi^2} \ln \frac{2}{\phi} \rceil$
2: **repeat**
3: **for** $\mathbf{B} \subset \mathbf{X}$ **do**
4: For $i = 1 \ldots n$ sample $\tau_i \sim T$
5: **if** $\texttt{ER}(f, \mathbf{B}, T, \gamma, \mathbf{u_r}, \psi, \phi) < \zeta$ **then**
6: $\Delta \mathbf{u_r} \leftarrow 0$
7: **repeat**
8: Compute $L_{\mathbf{B},\tau} = \frac{1}{|\mathbf{B}| \times n} \sum_{i=1}^{|\mathbf{B}|} \sum_{j=1}^{n} L[f(\mathbf{B_i} + \tau_j(\mathbf{u_r} + \Delta \mathbf{u_r})), f(\mathbf{B_i})]$
9: $\Delta \mathbf{u_r} = \mathcal{P}_{p,\epsilon}(\Delta \mathbf{u_r} + \alpha \text{sign}(\nabla L_{\mathbf{B},\tau}))$
10: **until** $\texttt{ER}(f, \mathbf{B}, T, \gamma, \mathbf{u_r} + \Delta \mathbf{u_r}, \psi, \phi) < \zeta$
11: $\mathbf{u_r} \leftarrow \mathcal{P}_{p,\epsilon}(\mathbf{u_r} + \Delta \mathbf{u_r})$
12: **end if**
13: **end for**
14: **until** $\texttt{ER}(f, \mathbf{X}, T, \gamma, \mathbf{u_r}, \psi, \phi) < \zeta$

Table 1. Robust ASR of `RobustUAP` compared to the three baselines.

DATASET	TRANSFORMATION SET	STANDARD UAP	SGD	STANDARD UAP_RP	ROBUST UAP
ILSVRC 2012	$R(20)$	0.0%	69.9%	2.9%	**93.2%**
	$T(2,2)$	35.9%	96.1%	38.8%	**97.1%**
	$Sc(5), R(5), B(5, 0.01)$	22.3%	85.4%	43.7%	**96.1%**
	$R(10), T(2,2), Sh(2), Sc(2), B(2, 0.001)$	0.0%	63.1%	2.9%	**86.4%**
CIFAR-10	$R(30), B(2, 0.001)$	0.0%	64.1%	2.9%	**75.7%**
	$R(2), Sh(2)$	42.7%	88.3%	52.4%	**96.1%**
	$R(10), T(2,2), Sh(2), Sc(2), B(2, 0.001)$	0.0%	58.3%	7.8%	**79.6%**

and sampled transformations, instead of making a single gradient update like SGD, we use Projected Gradient Descent (PGD) to iteratively compute a more robust update to the universal perturbation and end only when the estimated robustness on the batch satisfies a given threshold (line 10). At the end of each epoch, we check the robustness across the entire training set and transformation space using `EstRobustness` (`ER`) and stop when we have reached the desired performance (line 14).

5. Evaluation

We empirically evaluate our method `RobustUAP` and three baseline approaches (`SGD`, `StandardUAP_RP`, `StandardUAP` (Moosavi-Dezfooli et al., 2017)) on popular models from the vision domain. We show that `RobustUAP` is more robust on both uniform random noise and compositions of real-world transformations such as rotation, scaling, etc. We did not have the hardware to print high resolution transparent stickers so we could not produce real-world results in the vision domain. We show that training `RobustUAP` on a set of primitive transforms results in a universally robust UAP which generalizes well to unseen transformations allowing for successful attacks without the need for domain specific modeling.

Experimental evaluation. We consider two popular image recognition datasets: CIFAR-10(Krizhevsky et al., 2009) and ILSVRC 2012(Deng et al., 2009). We evaluate on a pretrained VGG16 (Simonyan and Zisserman, 2014) and Inception-v3 (Szegedy et al., 2016) network on CIFAR-10 and ILSVRC 2012 respectively. For both we evaluate on a random subset (1000 images) for the test set. All experiments were performed on a desktop PC with a GeForce RTX(TM) 3090 GPU and a 16-core Intel(R) Core(TM) i9-9900KS CPU @ 4.00GHz.

We report the results for l_2-norm with $\epsilon = 100$ for ILSVRC 2012 and $\epsilon = 10$ for CIFAR-10. These values were chosen based on the values presented by the original UAP paper (Moosavi-Dezfooli et al., 2017). We use an image normalization function given by our pretrained models and thus scaled our ϵ values accordingly. We note that the ϵ-values are significantly smaller than the image norms resulting in imperceptible perturbations that do not affect the semantic content of the image. Due to the hardness of the optimization problem, for the same norm value, the effectiveness of a UAP is less than input-specific perturbations; however, crafting input-specific perturbations requires making unrealistic assumptions about the power of the attacker as mentioned in the introduction and therefore we do not consider them part of our threat model which aims to generate practically feasible perturbations. We use $\psi = 0.05$ and $\phi = 0.05$ resulting in $n = 738$ for generating samples for our `RobustUAP` algorithm as well as reporting robust ASR in our evaluation. The UAPs are trained on 2,000 images, other parameters for evaluation are given in Appendix K. Error bars/variances are reported in Appendix AE.

5.1. Robustness to Random Noise

We first generate UAPs robust against uniform random noise. The results and future discussion can be found in Appendix O. Since our neighborhood has a well-defined PDF we get robustness guarantees from `EstimateRobustness`, in the following sections we consider semantic and unknown transformations and do not have the same guarantees.

5.2. Robustness to Semantic Transformations

Next, we consider transformation sets generated by composing five popular semantic transformations in existing literature (Athalye et al., 2018; Balunović et al., 2019): brightness/contrast, rotation, scaling, shearing, and translation.

We use a variety of different compositions to show that our algorithm works under different conditions, and base our parameters for the transformations on (Balunović et al., 2019). For our experiments, $R(\theta)$ corresponds to rotations with angles between $\pm\theta$; $T(x, y)$, to translations of $\pm x$ horizontally and $\pm y$ vertically; $Sc(p)$ to scaling the image between $\pm p\%$; $Sh(m)$ to shearing by shearing factor between $\pm m\%$; and $B(\alpha, \beta)$ to changes in contrast between $\pm\alpha\%$ and brightness between $\pm\beta$. Further details about these transformations can be seen in Appendix E. We consider compositions of different subsets and ranges of these transformations shown in Table 1 including composing all transformations together. The hardness of generating robust

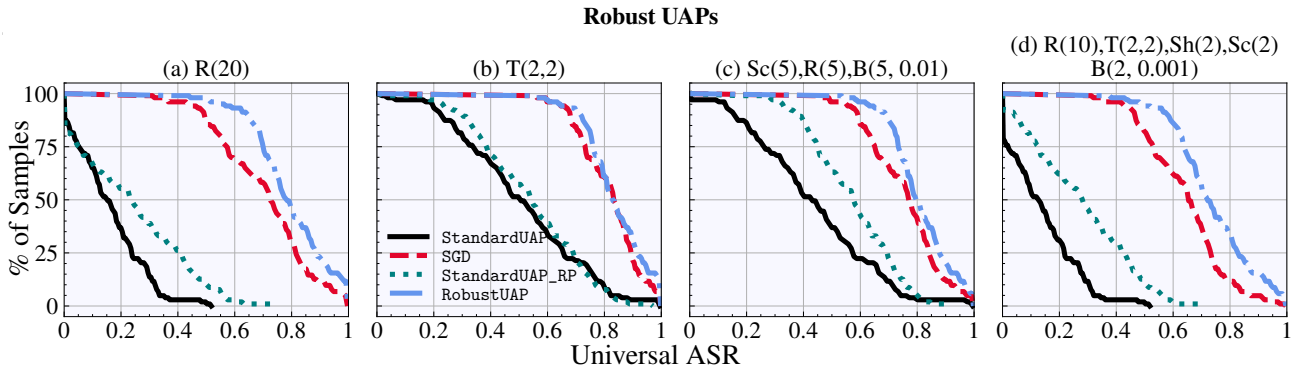

Figure 3. For each method, a point (x, y) in the corresponding line represents the percentage of sampled UAPs ($y\%$) with Universal ASR $> x$ for the different semantic transformations on ILSVRC.

UAPs depends on the effect that the transformation set has on the UAP (i.e. random noise has a relatively small effect compared to rotation). The hardness also increases with the number of transformations in the composition as well as the range of parameters for each individual transformation. For example, generating robust UAPs is harder for the composition shown in the first and last row for ILSVRC 2012 in Table 1 compared to the second and third row. The same is true for generating a UAP robust to uniform random noise.

Robust ASR ($\mathbf{ASR_R}$). Figure 3 shows performance of UAPs obtained by applying 738 randomly sampled transformations to the original UAPs generated by different methods on ILSVRC, similar graphs for CIFAR-10 can be found in Appendix M. The `RobustUAP` algorithm outperforms all others in each case, we observe that for these harder transformation sets `StandardUAP` loses its effectiveness completely. In Table 1 we compare robust universal adversarial success rate ASR_R with $\gamma = 0.6$, in other words, we are finding the percentage of sampled neighbors of the perturbation that are still UAPs with 60% effectiveness on the testing set. We provide average ASR_U scores as well as ASR_R for different γ levels in Appendix N.

Our `RobustUAP` algorithm achieves at least 53.4% higher robust ASR when compared to the standard UAP algorithm on both datasets and all transformation sets. Furthermore, our `RobustUAP` algorithm significantly outperforms both robust baseline approaches. Except for the $T(2, 2)$ case which we observe to be the easiest, `RobustUAP` achieves at least 11.6% performance gain over the baselines. `SGD` is the best performing baseline and achieves high robust ASR on relatively easier transformation sets performing within 1% of `RobustUAP` on $T(2, 2)$. On harder transformation sets, this gap widens considerably, see Table 1.

5.3. Universally Robust UAPs

Using the set of primitive transformations discussed by PRIME, we generate robust UAPs on ILSVRC using the same parameters as above. For each sampled transform, we randomly apply three transformations from identity, spectral, spatial, and color. This means that we can get multiple of the same transformation or even no transformation. We follow the setup from PRIME for the parameters of each transformation type. Table 2 shows the robust ASR when training a `RobustUAP` on PRIME, Affine ($R(10), T(2, 2), Sh(2), Sc(2), B(2, 0.001)$), and Fog transformation sets and how they perform on common corruptions (Modas et al., 2022; Kar et al., 2022; Hendrycks and Dietterich, 2018). Although prime does not inherently contain any specific affine or common corruption in its training it has generally high robustness (>58.3%) against all transformation sets tested. We observe that training on the target transformation set does bring higher robustness than training on PRIME (i.e. Affine-trained robust UAP has best performance on Gaussian, Contrast, Affine while Fog-trained robust UAP has best performance on fog); however, we find that PRIME has much better performance on unseen transformations (i.e. Fog-trained or Affine-trained robust UAP on JPEG). Our results suggest that a set of good primitive transformations is sufficient for generating universally robust UAPs that generalize well to unseen transformations.

Table 2. Robust ASR (%) of `RobustUAP` trained on PRIME, Affine ($R(10)$, $T(2, 2)$, $Sh(2)$, $Sc(2)$, $B(2, 0.001)$), and Fog when applied to Prime, Affine, and common corruption transforms

	Evaluation Corruption Set																
Train Set	Noise					Blur				Weather				Digital			
	Prime	Aff.	Gaus.	Shot	Imp.	Defo.	Glass	Moti.	Zoom	Snow	Fog	Frost	Bright	Contr.	Elast.	Pixel	JPEG
Prime	**68.4**	58.3	72.1	81.3	**88.6**	**66.5**	**75.2**	**81.0**	74.6	**77.8**	78.8	**65.3**	**85.3**	90.4	**74.2**	**69.2**	**73.3**
Affine	10.1	**86.4**	**91.2**	**93.2**	85.4	45.1	31.4	76.1	**92.4**	65.2	70.1	50.1	80.1	**94.1**	39.1	30.5	37.3
Fog	1.5	0.1	10.2	11.3	9.3	15.1	7.6	10.1	5.5	69.5	**95.2**	21.3	10.1	12.6	18.4	2.8	3.9

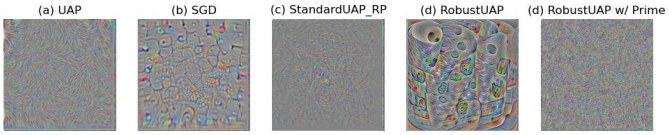

Figure 4. Comparison of UAPs on ILSVRC 2012 generated with (a) `StandardUAP`, (b) `RobustUAP`, (c) `Standard UAP_RP`, (d) `RobustUAP`, and (e) `RobustUAP` generated on Prime. Note that the size of each perturbation is ≥ 99.7 (bound is 100).

5.4. Visualization

We visualize UAPs generated with our three robust algorithms on the same transformation set and dataset in Figure 4. We further visualize UAPs generated with different algorithms transformed randomly from $R(10), T(2,2), Sh(2), Sc(2), B(2, 0.001)$ and added to images in ILSVRC 2012 in Appendix AD. Our robust UAPs have a similar level of imperceptibility to standard UAPs and robust UAPs affect the model classification after transformation with high probability, unlike standard UAPs.

5.5. Transferability of Robust UAPs

We evaluate the transferability of `RobustUAP`. Previous works on UAPs (Moosavi-Dezfooli et al., 2017) show that UAPs are transferable across different models. Here, we will evaluate whether robust UAPs exhibit the same behavior for robustness. The robust UAPs studied here are generated with `RobustUAP` on $R(10), T(2,2), Sh(2), Sc(2), B(2, 0.001)$ for ILSVRC-2012 with $\gamma = 0.6$. We use a variety of models: Inception-v3 (Szegedy et al., 2016), MobileNet (Howard et al., 2017), Inception-v3 trained to be robust on $R(20)$ (InceptionR20), Inception-v3 trained to be robust on horizontal flips (InceptionHF), and ViT (Dosovitskiy et al., 2020). Table 3 shows us that our robust UAPs are transferable between different architectures. Our results show that robust UAPs transfer their robustness properties between architectures and models. Ignoring ViT, on all of the Inception and MobileNet models, the generated UAPs maintain at least 65% robust ASR when transferred to each other. ViT maintains at least 32% robustness when transferred to or from the other convolution-based models. In Appendix U we check the transferability between different domains (ImageNet-C (Hendrycks and Dietterich, 2018) and ILSVRC-2012).

5.6. Robustness against Robust UAPs

Traditional methods for robustness, such as adversarial training, focus on being robust in scenarios where the attacker is powerful (i.e. PGD), but with this comes a significant tradeoff in accuracy. In a preliminary study of practical robustness, we train for robustness against practical attacks such as Robust UAP, while maintaining high accuracy and faster training times. We perform our experiments on CIFAR-10, with a VGG16 model architecture, and with our most challenging transformation set $(R(10), T(2,2), Sh(2), Sc(2), B(2, 0.001))$. We use a batch-wise variant of our robust UAP algorithm to do adversarial training. With this training method, our model obtains a Robust ASR of 0.4% and an accuracy of 90.1%. In contrast, standard adversarial training obtains a Robust ASR of 0.2% and an accuracy of 83.5%. Our training method obtains almost the same practical robustness without a significant reduction in accuracy. Further, our adversarial training method with robust UAP takes 12 minutes while standard adversarial training takes 48 minutes, which is one indication that our proposed training method is more efficient. We believe that further study can improve the robustness, accuracy, and efficiency of this type of training allowing more practically robustness while sacrificing less accuracy.

5.7. `RobustUAP` vs `SGD` Performance.

Previous sections highlight `SGD` as the most competitive algorithm to `RobustUAP` in terms of performance. In Appendix W, we report the runtimes for the different algorithms and observe that `SGD` runs four times as fast as `RobustUAP`. Although this seems to suggest that `SGD` is more efficient, we further investigate restricting the runtime of `RobustUAP`. We first add results to the ILSVRC 2012 part of Table 1 by also computing `RobustUAP` performance when limited to the same amount of time that `SGD` takes. Table 4 shows that `RobustUAP` outperforms `SGD` even when its compute time is limited with up to 9% more robustness on our most challenging transformation $R(10), T(2,2), Sh(2), Sc(2), B(2, 0.001)$. Further-

Table 3. Robust ASR when UAP is learned on source model and transfered to target model.

	Target Model				
Source Model	Inception	MobileNet	InceptionR20	InceptionHF	ViT
Inception	**86.4**%	65.2%	75.2%	78.5%	35.1%
MobileNet	74.3%	**86.2**%	67.3%	68.6%	38.3%
InceptionR20	80.1%	67.3%	**81.3**%	73.1%	32.0%
InceptionHF	77.8%	70.9%	75.8%	**83.8**%	34.6%
ViT	41.2%	32.4%	43.2%	39.7%	**88.5**%

Table 4. Robust ASR of RobustUAP restricted to the same amount of compute time as SGD.

TRANSFORMATION SET	SGD	ROBUST UAP	RESTRICTED ROBUST UAP
$R(20)$	69.9%	**93.2%**	72.9%
$T(2,2)$	96.1%	**97.1%**	96.9%
$Sc(5), R(5), B(5, 0.01)$	85.4%	**96.1%**	86.3%
$R(10), T(2,2), Sh(2), Sc(2), B(2, 0.001)$	63.1%	**86.4%**	72.0%

more, in Appendix X, we measure SGD's performance when varying the number of iterations and we observe that SGD's performance flatlines and does not reach the performance of RobustUAP even when allowed to run for longer. Finally, we note that since UAPs only need to be computed a single time and can be done in advance, the runtime considerations are not as big of a factor for most practical use cases.

5.8. Robust UAPs in Real-World Wireless Setting

We show the practical applicability of our attack in the wireless domain. ML-based methods have become increasingly popular for both communication and sensing applications (Liu et al., 2021; Ayyalasomayajula et al., 2020) in the wireless domain. We show that we can successfully attack a real-world system, FIRE (Liu et al., 2021), by using the robust UAP algorithm. FIRE uses an end-to-end ML based approach and its model is a variant of variational autoencoders (VAEs). FIRE is deployed at a base station and aims to predict the downlink channel without client feedback which is an important task for 5G. In this attack scenario, we propose an attacker transmit a UAP over the air to disrupt FIRE's ability to perform downlink channel prediction when deployed at a base station. While our attacker knows the model that FIRE is using it does not have real-time access to the base station or the client and thus could not perform a standard adversarial attack. Furthermore, transmitted signals undergo changes due to a variety of noise and transformation before being received at the base station and the attacker must take this into account. We consider Carrier Frequency Offset (CFO) and hardware detection delay to model the transformations applied on the UAPs during transmission. Using an attacker with a single transmission antenna we can attack a base station running FIRE which has 4 antennas. While moving around a client in a variety of different locations we record the drop in Signal-to-Noise Ratio (SNR) of the predicted channel. Our robust UAPs are able to produce a median drop of 4.06 dB which is twice the drop compared to Gaussian noise (1.84 dB). Furthermore, when using a standard UAP we find the performance to be similar to Gaussian noise (1.91 dB). This reinforces our claim that while UAPs address some of the issues with standard adversarial perturbations, transformations in the real-world severely degrade the effectiveness of UAPs rendering them unsuitable for real-world attacks. These results further show that Robust UAPs can improve upon this and provide significant performance loss against real-world systems.

5.9. Additional Experiments

In Appendix P, we show how our robust UAPs compare to standard UAPs on the non-robust universal ASR metric. In Appendix Q, we evaluate our methods on ResNet18 (He et al., 2015) and MobileNet (Howard et al., 2017) for CIFAR-10 and ILSVRC 2012 respectively as well as showing results on MNIST (Deng, 2012) and Fashion MNIST (Xiao et al., 2017). The results follow the same trends as those reported in Table 1. To address additional real-world transformations, we investigate fog perturbations from (Kar et al., 2022) in Appendix R. In Appendix Y, we show that our methods work well for targetted robust UAPs. In Appendix Z, we show that RobustUAP has good data efficiency and obtains good performance with less data. The recent popularity of transformer based models has also led us to show that our methods work on transformer based networks, results in Appendix AA. In Appendix AB, we find that traditional model robustness does not seem to effect ability to create robust UAPs. In Appendix V, we compare standard UAP to Robust UAP performance on models trained for single adversarial perturbation robustness. In Appendix S, we provide additional data on using scaling and hue transformation sets. In Appendix T, we show how using different transformation sets in training affects robustness on an unseen transformation set. Finally, In Appendix AC, we perform an ablation study on optimization strategy and show that SGD outperforms other popular optimizers.

6. Related Work

UAP Algorithms. Most works focusing on UAPs (Moosavi-Dezfooli et al., 2017; Mopuri et al., 2018; Zhang et al., 2020a; Khrulkov and Oseledets, 2018; Akhtar et al., 2018; Hendrik Metzen et al., 2017; Zhang et al., 2020b) generate singular vectors and do not consider perturbation robustness. Bahramali et al. (2021) introduces a perturbation generator model (PGM) for the wireless domain which creates UAPs. They show that both adversarial training and noise subtracting defenses used in the wireless domain are highly effective in mitigating the effects of a single vector UAP attack; they further show that their method of generating a set of UAPs is an effective way for an attacker to circumvent

these defenses. Although PGM provides a method for efficiently sampling unique UAPs, it is not robust to real-world transformations. In contrast, our method enables efficient sampling of UAPs that are robust to transformations.

Robust Adversarial Examples. The following papers introduce notions of robustness under different viewpoints and environmental conditions for constructing realizable adversarial examples. This is a different threat model compared to the additive perturbations discussed in this paper. Luo et al. (2018) constructs adversarial examples which minimize human detectability, further introducing the idea of robustness for adversarial examples. They show that their attacks are robust against jpeg compression. Their work demonstrates a successful physical attack under stable conditions and poses. Hu (2022) proposes using lasers to disrupt street sign classification; their method is focused on singular objects viewed from different viewpoints rather than universality over different signs. Zhong et al. (2022) creates shadows which are adversarial for sign classification and robust to image transformations, but their method is not universal. Athalye et al. (2018) introduce Expectation over Transformation (EOT) and use it to print real-world objects which are adversarial given a range of physical and environmental conditions. Sharif et al. (2016), Eykholt et al. (2018), Feng et al. (2023), Casper et al. (2022), Komkov and Petiushko (2021), Brown et al. (2017), Wu et al. (2020), Thys et al. (2019), and Hu et al. (2021) all create printable physical adversarial patches which are robust to a set of domain specific transformations.

Robust Adversarial Perturbations. Li et al. (2019a) generates music which prevents voice assistants from picking up its wake word. Li et al. (2019b) generates a targeted adversarial sticker which changes an image classifier's classification from one pre-specified class to another. Both of these methods rely on specific use cases and are tailored towards generating adversaries coming from strict distributions, e.g. (Li et al., 2019a) generates guitar music while (Li et al., 2019b) generates a small grid of dots. Li et al. (2023) proposes an inaudible perturbation for attacking speech recognition; their method models transformations in the audio domain. These works build on algorithms akin to our baseline approaches and are limited in scope to domain specific transformations. Our work provides a framework for improving robustness against a wide range of transformations in diverse domains and can be leveraged for improving the effectiveness of these attacks.

7. Limitations

We outline the limitations of our work. Firstly, we note that our methods do not have a way to address non-differentiable transformations. We hope that future work leverages methods such as REINFORCE which do not have a dependence on differentiability (Williams, 1992). We also note that in some threat models (where the source and perturbation are simultaneously transformed) we do not support non-distrbutive transformations. Secondly, `RobustUAP` is less effective against models trained with standard adversarial training since to have a UAP one needs to generate standard adversarial examples. However, currently, robustly trained models are seldom used since adversarial attacks are hard to realize (i.e. UAPs) and these models come with a significant reduction in accuracy. Our preliminary research suggests that defending against practical attacks such as robust UAP does not come with the same tradeoff, allowing for more practical robustness while retaining similar accuracy.

8. Conclusion

In this paper, we demonstrate that standard UAPs fail to be universally adversarial under transformation. We propose a new method, `RobustUAP`, to generate robust UAPs based upon obtaining probabilistic bounds on UAP robustness across an entire transformation space. We show that `RobustUAP` works for both known and unknown transformation sets. Our experiments provide empirical evidence that our principled approach generates UAPs that are more robust than those from the existing/baseline methods. Our preliminary work suggests that robustness against practical adversaries such as robust UAPs may require much less tradeoff with accuracy and we hope that inspires research into robustness against practical attacks.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. We understand that our proposed methods could cause harm/expose vulnerabilities of existing deployed ML methods. We hope that by showing the existence of robust UAPs in safety and security critical applications of ML we will spawn further research into practical robustness against real-world implementable attacks.

References

Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu. Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on audio, speech, and language processing*, 22(10): 1533–1545, 2014.

DC Agarwal. *Tensor Calculus and Riemannian Geometry*. Krishna Prakashan Media, 2013.

Naveed Akhtar, Jian Liu, and Ajmal Mian. Defense against universal adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3389–3398, 2018.

Cesare Alippi. *Intelligence for embedded systems*. Springer, 2014.

Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search, 2019.

Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International conference on machine learning*, pages 284–293. PMLR, 2018.

Roshan Ayyalasomayajula, Aditya Arun, Chenfeng Wu, Sanatan Sharma, Abhishek Rajkumar Sethi, Deepak Vasisht, and Dinesh Bharadia. Deep learning based wireless localization for indoor navigation. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, pages 1–14, 2020.

Alireza Bahramali, Milad Nasr, Amir Houmansadr, Dennis Goeckel, and Don Towsley. Robust adversarial attacks against dnn-based wireless communication systems. *arXiv preprint arXiv:2102.00918*, 2021.

Mislav Balunović, Maximilian Baader, Gagandeep Singh, Timon Gehr, and Martin Vechev. Certifying geometric robustness of neural networks. *Advances in Neural Information Processing Systems 32*, 2019.

Philipp Benz, Chaoning Zhang, Tooba Imtiaz, and In So Kweon. Double targeted universal adversarial perturbations. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

Philipp Benz, Soomin Ham, Chaoning Zhang, Adil Karjauv, and In So Kweon. Adversarial robustness comparison of vision transformer and mlp-mixer to cnns. *arXiv preprint arXiv:2110.02797*, 2021.

Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. IEEE, 2017.

Stephen Casper, Max Nadeau, Dylan Hadfield-Menell, and Gabriel Kreiman. Robust feature-level adversaries are interpretability tools. *Advances in Neural Information Processing Systems*, 35:33093–33106, 2022.

Herman Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, pages 493–507, 1952.

Francesco Croce and Matthias Hein. Minimally distorted adversarial examples with a fast adaptive boundary attack, 2019.

Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670*, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. A guide to deep learning in healthcare. *Nature medicine*, 25(1):24–29, 2019.

Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1625–1634, 2018.

Weiwei Feng, Nanqing Xu, Tianzhu Zhang, Baoyuan Wu, and Yongdong Zhang. Robust and generalized physical adversarial attacks via meta-gan. *IEEE Transactions on Information Forensics and Security*, 2023.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Kaiming He, X Zhang, S Ren, and J Sun. Deep residual learning for image recognition." computer vision and pattern recognition (2015). *Google Scholar There is no corresponding record for this reference*, pages 770–778, 2015.

Jan Hendrik Metzen, Mummadi Chaithanya Kumar, Thomas Brox, and Volker Fischer. Universal adversarial perturbations against semantic image segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 2755–2764, 2017.

Dan Hendrycks and Thomas G Dietterich. Benchmarking neural network robustness to common corruptions and surface variations. *arXiv preprint arXiv:1807.01697*, 2018.

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

Chengyin Hu. Adversarial laser spot: Robust and covert physical adversarial attack to dnns. *arXiv preprint arXiv:2206.01034*, 2022.

Yu-Chih-Tuan Hu, Bo-Han Kung, Daniel Stanley Tan, Jun-Cheng Chen, Kai-Lung Hua, and Wen-Huang Cheng. Naturalistic physical adversarial patch for object detectors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7848–7857, 2021.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28:2017–2025, 2015.

Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678, 2014.

Oğuzhan Fatih Kar, Teresa Yeo, Andrei Atanov, and Amir Zamir. 3d common corruptions and data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18963–18974, 2022.

Valentin Khrulkov and Ivan Oseledets. Art of singular vectors and universal adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8562–8570, 2018.

Stepan Komkov and Aleksandr Petiushko. Advhat: Real-world adversarial attack on arcface face id system. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 819–826. IEEE, 2021.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images, 2009.

Juncheng Li, Shuhui Qu, Xinjian Li, Joseph Szurley, J. Zico Kolter, and Florian Metze. Adversarial music: Real world audio adversary against wake-word detection system. In *Proc. Neural Information Processing Systems (NeurIPS)*, pages 11908–11918, 2019a.

Juncheng Li, Frank R. Schmidt, and J. Zico Kolter. Adversarial camera stickers: A physical camera-based attack on deep learning systems. In *Proc. International Conference on Machine Learning, ICML*, volume 97, pages 3896–3904, 2019b.

Xinfeng Li, Chen Yan, Xuancun Lu, Zihan Zeng, Xiaoyu Ji, and Wenyuan Xu. Inaudible adversarial perturbation: Manipulating the recognition of user speech in real time. *arXiv preprint arXiv:2308.01040*, 2023.

Zikun Liu, Gagandeep Singh, Chenren Xu, and Deepak Vasisht. Fire: enabling reciprocity for fdd mimo systems. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*, pages 628–641, 2021.

Bo Luo, Yannan Liu, Lingxiao Wei, and Qiang Xu. Towards imperceptible and robust adversarial example attacks against neural networks. In *Thirty-second aaai conference on artificial intelligence*, 2018.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Apostolos Modas, Rahul Rade, Guillermo Ortiz-Jiménez, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Prime: A few primitives can boost robustness to common corruptions. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXV*, pages 623–640. Springer, 2022.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017.

Konda Reddy Mopuri, Aditya Ganeshan, and R Venkatesh Babu. Generalizable data-free objective for crafting universal adversarial perturbations. *IEEE transactions on pattern analysis and machine intelligence*, 41(10):2452–2465, 2018.

Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. *arXiv preprint arXiv:2205.07460*, 2022.

Ali Shafahi, Mahyar Najibi, Zheng Xu, John Dickerson, Larry S Davis, and Tom Goldstein. Universal adversarial training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5636–5643, 2020.

Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 acm sigsac conference on computer and communications security*, pages 1528–1540, 2016.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

Simen Thys, Wiebe Van Ranst, and Toon Goedemé. Fooling automated surveillance cameras: adversarial patches to attack person detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019.

Florian Tramèr, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. *CoRR*, abs/2002.08347, 2020.

Shiqi Wang, Yizheng Chen, Ahmed Abdou, and Suman Jana. Enhancing gradient-based attacks with symbolic intervals, 2019.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.*, 8(3–4):229–256, may 1992. ISSN 0885-6125. doi: 10.1007/BF00992696. URL https://doi.org/10.1007/BF00992696.

Zuxuan Wu, Ser-Nam Lim, Larry S Davis, and Tom Goldstein. Making an invisibility cloak: Real world adversarial attacks on object detectors. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 1–17. Springer, 2020.

Chaowei Xiao, Bo Li, Jun yan Zhu, Warren He, Mingyan Liu, and Dawn Song. Generating adversarial examples with adversarial networks. In *Proc. International Joint Conference on Artificial Intelligence (IJCAI-18)*, pages 3905–3911, 2018a.

Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially transformed adversarial examples. *arXiv preprint arXiv:1801.02612*, 2018b.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Chaoning Zhang, Philipp Benz, Tooba Imtiaz, and In So Kweon. Understanding adversarial examples from the mutual influence of images and perturbations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14521–14530, 2020a.

Chaoning Zhang, Philipp Benz, Tooba Imtiaz, and In-So Kweon. Cd-uap: Class discriminative universal adversarial perturbation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6754–6761, 2020b.

Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6848–6856, 2018.

Tianhang Zheng, Changyou Chen, and Kui Ren. Distributionally adversarial attack. *Proc. AAAI Conference on Artificial Intelligence*, 33:2253–2260, 07 2019.

Yiqi Zhong, Xianming Liu, Deming Zhai, Junjun Jiang, and Xiangyang Ji. Shadows can be dangerous: Stealthy and effective physical-world adversarial attack by natural phenomenon. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15345–15354, 2022.

Appendix

A. UAP Transformation vs. UAP + Input Transformation

In this section, we discuss how our threat model of perturbations which may undergo transformation compares to a threat model where the source input is also transformed. Let's consider two cases, one in which the source image and target image are perturbed separately (i.e. $f(\tau_1(v) + \tau_2(x))$) and one in which the source image and UAP are perturbed simultaneously (i.e. $f(\tau_2(\tau_1(v)+x))$). In the first case, we argue that since the universal attacker is naïve about the input this case is similar to the dynamic we consider in the paper (i.e. $f(\tau(v) + x)$) as in either case our attacker has no knowledge of what the input is. If $\tau_2(x)$ is in-distribution for f then a properly trained robust UAP would have a high probability of being robust to $\tau_2(x)$. If $\tau_2(x)$ is out-of-distribution for f then f acting on $\tau_2(x)$ is inherently less robust and thus would with high probability be worse when attacked with v. For the second case, if we assume that transformations are distributive or associative then we have $f(\tau_2(\tau_1(v)) + tau_2(x))$ which is just the first case (while two perturbations is not the same as one, the composition is functionally similar). While this will not hold for all transformations it holds for the transformations considered in this paper (affine transformations, additive noise, etc.), we leave non-distributive, non-associative transformations in the second case for future work. Using ILSVRC and our Inception-v3 model we experiment with both cases drawing tau_1, tau_2 from $R(10), T(2, 2), Sh(2), Sc(2), B(2, 0.001)$. For case 1, we get a robust ASR of 91.2% and for case 2, we get a robust ASR of 87.3%. These numbers are actually both higher than our original result of 86.4% as our model is not inherently robust to transformation so although the transformations are not adversarially chosen they still degrade the accuracy of the model helping our attacker.

Furthermore, if applying UAPs/RobustUAPs to transformed standard images the question shifts more to whether or not the original model is robust to transformation rather than if the original model is robust to UAPs. Using ILSVRC and our Inception-v3 model we draw τ from R(10), T(2, 2), Sh(2), Sc(2), B(2, 0.001). Let ASR_P be the ASR over the test dataset when applying a randomly drawn τ to perturbed inputs (i.e. $f(\tau(x + u))$). We find that standard UAP has an ASR_P of 75.6% and RobustUAP has an ASR_P of 91.7% (Table 1 gives Robust ASR of 0.0% and 86.4% for UAP and RobustUAP respectively). This shows that RobustUAP retains a performance increase over UAP even when transforming the perturbed image.

B. Definitions

In this section, we will formally define the terms used in the main body of the paper. We first start with adversarial examples.

Definition B.1. Given a correctly classified point $\mathbf{x}$, a distance function $d(\cdot, \cdot) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, and bound $\epsilon \in \mathbb{R}$, $\mathbf{x}'$ is an *adversarial example* iff $d(\mathbf{x}', \mathbf{x}) < \epsilon$ and $\hat{f}(\mathbf{x}') \neq y$.

We distinguish between adversarial examples and perturbations. An *adversarial perturbation* added to the point it is attacking is an *adversarial example*, $x' = x + v$. We can construct v by solving the following optimization problem:

$$\arg\min_{v} ||v||_p \text{ s.t. } \hat{f}(x + v) \neq \hat{f}(x) \tag{5}$$

Here, we are looking for the smallest v such that f's classification changes from the original output (assuming f correctly classified x).

Next, we define the adversarial success rate which measures whether or not a given perturbation is adversarial.

Definition B.2. Given a datapoint x, and perturbation v, *adversarial success*, AS, is defined as

$$\text{AS}(f, x, v) = 1 - \delta(\hat{f}(x + v), \hat{f}(x)) \tag{6}$$

Here, $\delta(i, j)$ refers to the Kronecker Delta function (Agarwal, 2013), formally,

$$\delta(i, j) = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases} \tag{7}$$

With AS, we measure whether v changes f's classification of x.

Using the definition of adversarial success we can now define universal adversarial success rate.

Definition B.3. Given a data distribution μ, and perturbation $\mathbf{u}$, *universal adversarial success rate*, ASR_U, for $\mathbf{u}$, is

$$\text{ASR}_U(f, \mu, \mathbf{u}) = \underset{\mathbf{x}\sim\mu}{P}(\hat{f}(\mathbf{x}+\mathbf{u}) \neq \hat{f}(\mathbf{x})) \tag{8}$$

Using Definition B.3, we formally define a UAP.

Definition B.4. A *universal adversarial perturbation* is a vector $\mathbf{u} \in \mathbb{R}^d$ which, when added to almost all datapoints in μ causes the classifier f to misclassify. Formally, given γ, a bound on universal ASR, and l_p-norm with corresponding bound ϵ, $\mathbf{u}$ is a UAP iff $\text{ASR}_U(f, \mu, \mathbf{u}) > \gamma$ and $||\mathbf{u}||_p < \epsilon$.

Now, we move onto definitions pertaining to robust UAPs. We start by formally defining transformation sets and neighborhoods.

Definition B.5. A *transformation*, τ, is a composition of bijective sub-differentiable transformation functions. A *transformation set*, T, is a set of distinct transformations. A point $\mathbf{v}'$ is in the *neighborhood* $N_T(\mathbf{v})$, of $\mathbf{v}$, if there is a transform in T that maps $\mathbf{v}$ to $\mathbf{v}'$ and $||\mathbf{v}'||_p < \epsilon$. Formally,

$$\mathbf{v}' \in N_T(\mathbf{v}) \iff \exists \tau \in T \text{ s.t. } \tau(\mathbf{v}) = \mathbf{v}' \tag{9}$$

Finally, we formally define robust universal adversarial success rate.

Definition B.6. Given a data distribution μ, transformation set T, universal ASR level γ, bound ϵ on l_p-norm, and perturbation $\mathbf{u_r}$, *robust universal adversarial success rate*, ASR_R, is defined as,

$$\text{ASR}_R(f, \mu, T, \gamma, \mathbf{u_r}) = \underset{\mathbf{u_r}'\sim N_T(\mathbf{u_r})}{P}(\text{ASR}_U(f, \mu, \mathbf{u_r}') > \gamma \wedge ||\mathbf{u_r}'||_p < \epsilon) \tag{10}$$

C. Lagrangian Relaxation of Robust UAP Optimization Problem

We can approximate Equation 2 using Lagrangian relaxation to get the following optimization objective.

$$\frac{1}{|\hat{\mathbf{x}}| \times |\hat{\tau}|} \sum_{i=1}^{|\hat{\mathbf{x}}|} \sum_{\mathbf{v}'}^{\mathbf{u_r}' \in N_T(\mathbf{u_r})} L[f(\hat{\mathbf{x}}_\mathbf{i} + \mathbf{u_r}'), f(\hat{\mathbf{x}}_\mathbf{i})] - \lambda ||\mathbf{u_r}||_p \tag{11}$$

D. Proof of Theorem 4.1

This proof relies heavily on similar proofs provided by Chernoff (1952) and by Alippi (2014). We refer to the reader to these texts for further details. In our proof, we show how to adapt universal ASR to these proofs.

Proof. The bound on n is derived via the Chernoff inequality applied to $\hat{p}_n(\gamma)$ and $\mathbb{E}[\hat{p}_n(\gamma)] = p(\gamma)$ (Chernoff, 1952; Alippi, 2014). Equation 4 holds since computing universal ASR is Lebesgue measurable over the data distribution and since we assume $N_T(\mathbf{u_r})$ has a well defined PDF. □

E. Semantic Transformations

In this section, we discuss the semantic transformations used in the paper. Brightness and contrast can be represented via *bias* (β) and *gain* ($\alpha > 0$) parameters respectively. Formally, if $\mathbf{x}$ is the original image, then the transformed image, $\mathbf{x}'$, can be represented as

$$\mathbf{x}' = \alpha\mathbf{x} + \beta \tag{12}$$

Rotation, scaling, shearing, and translation are all affine transformations acting on the coordinate system, c, of the images instead of the pixel values, $\mathbf{x}$. In order to recover the pixel values and differentiate over the transformation, we will need

sub-differentiable interpolation, see Appendix F. For finite dimensions, affine transformations can be represented as a linear coordinate map where the original coordinates are multiplied by an invertible augmented matrix and then translated with additional bias vector. Below, we give the general form for an affine transformation given augmented matrix $\mathbf{A}$, bias matrix $\mathbf{b}$, and input coordinates c. We can compute the output coordinates, c', as

$$\begin{bmatrix} \mathbf{c}' \\ 1 \end{bmatrix} = \begin{bmatrix} [ccc|c] & \mathbf{A} & & \mathbf{b} \\ 0 & \dots & 0 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ 1 \end{bmatrix} \tag{13}$$

Below, we give the augmented matrix $\mathbf{A}$ and additional bias matrix $\mathbf{b}$ for rotation, scaling, shearing, and translation.

Rotation, $R(\theta)$, by θ degrees:

$$\mathbf{A} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}, \mathbf{b} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{14}$$

Scaling, $Sc(p)$, by $p\%$:

$$\mathbf{A} = \begin{pmatrix} 1 + \frac{p}{100} & 0 \\ 0 & 1 + \frac{p}{100} \end{pmatrix}, \mathbf{b} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{15}$$

Shearing, $Sh(m)$, by shear factor $m\%$:

$$\mathbf{A} = \begin{pmatrix} 1 & 1 + \frac{m}{100} \\ 0 & 1 \end{pmatrix}, \mathbf{b} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{16}$$

Translation, $T(x, y)$, by x pixels horizontally and y pixels vertically:

$$\mathbf{A} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \mathbf{b} = \begin{pmatrix} x \\ y \end{pmatrix} \tag{17}$$

F. Interpolation

Affine transformations may change a pixel's integer coordinates into non-integer coordinates. Interpolation is typically used to ensure that the resulting image can be represented on a lattice (integer) pixel grid. For this paper, we will be using bilinear interpolation, a common interpolation method which achieves a good trade-off between accuracy and efficiency in practice and is commonly used in literature (Xiao et al., 2018b; Balunović et al., 2019). Let $x_{i,j}$, $x'_{i,j}$ represent the pixel value at position i, j for the original and transformed image respectively. Let ${c'}^x_{i,j}$, ${c'}^y_{i,j}$ represent the x-coordinate and y-coordinate of the pixel at i, j after transformation. We define our transformed image by summing over all pixels $n, m \in [1 \dots H] \times [1 \dots W]$ where H and W represent the height and width of the image.

$$x'_{i,j} = \sum_{n}^{H} \sum_{m}^{W} x_{n,m} \max(0, 1 - |{c'}^x_{i,j} - m|) \max(0, 1 - |{c'}^y_{i,j} - n|) \tag{18}$$

This interpolation can be computed for each channel in the image. While interpolation is typically not differentiable, in order to generate adversarial examples using standard techniques we need a differentiable version of interpolation. (Jaderberg et al., 2015) introduces differentiable image sampling. Their method works for any interpolation method as long as the (sub-)gradients can be defined with respect to $x, {c'}_{i,j}$. For bilinear interpolation this becomes,

$$\frac{\partial x'_{i,j}}{\partial x_{n,m}} = \sum_{n}^{H} \sum_{m}^{W} \max(0, 1 - |{c'}^x_{i,j} - m|) \max(0, 1 - |{c'}^y_{i,j} - n|) \tag{19}$$

$$\frac{\partial x'_{i,j}}{\partial {c'}^x_{i,j}} = \sum_{n}^{H} \sum_{m}^{W} x_{n,m} \max(0, 1 - |{c'}^y_{i,j} - n|) \begin{cases} 1 & \text{if } m \geq |{c'}^x_{i,j} - m| \\ -1 & \text{if } m < |{c'}^x_{i,j} - m| \\ 0 & \text{otherwise} \end{cases} \tag{20}$$

G. Baseline Algorithms

G.1. Stochastic Gradient Descent

The first baseline directly solves Equation 2 using gradient descent. Since we are solving a constrained optimization problem, we cannot use gradient descent directly. Instead, we can solve the Lagrangian-relaxed form of the problem (adding a weighted norm term to the minimization problem), as in (Carlini and Wagner, 2017; Athalye et al., 2018), using a momentum based Stochastic Gradient Descent (SGD). Shafahi et al. (2020) suggests that this is an effective method for generating standard UAPs. Our SGD algorithm is in Appendix H. In order to implement it, we replace the δ-function with a loss function, L. We iteratively converge towards the inner expectation by computing it in batches, and towards the outer expectation by sampling a large number of transformations. Given that we would like to estimate on a batch, $\hat{\mathbf{x}} \subset \mu$, and a random set of transformations sampled from T, $\hat{\tau} \subset T$, we can approximate using Equation 11.

G.2. Standard UAP Algorithm with Robust Adversarial Perturbations

For our second baseline, we leverage the standard UAP algorithm from Moosavi-Dezfooli et al. (2017) (see Appendix I for the algorithm). The standard UAP algorithm takes each input, $\mathbf{x_i}$, computes the smallest additive change, $\Delta\mathbf{u}$, to the current perturbation, $\mathbf{u}$, that would make $\mathbf{u} + \Delta\mathbf{u}$ an adversarial perturbation for $\mathbf{x_i}$. Intuitively, over time the algorithm will approach a perturbation that works on most inputs in the training dataset. We modify this approach by computing robust adversarial perturbations rather than standard adversarial perturbations. At each point $\mathbf{x_i}$, we compute the smallest additive change, $\Delta\mathbf{u_r}$, to the current robust adversarial perturbation, $\mathbf{u_r}$, that would make $\mathbf{u_r} + \Delta\mathbf{u_r}$ a robust adversarial perturbation for $\mathbf{x_i}$. We search for robust adversarial perturbations by optimizing the expectation that a point in the neighborhood of $\mathbf{v_r}$ is adversarial while restricting the perturbation to an l_p norm of ϵ.

H. SGD Algorithm

Our SGD UAP algorithm is based on standard momentum based SGD while optimizing over the objective proposed in 2, the algorithm details can be seen in Algorithm 2.

Algorithm 2 Stochastic Gradient Descent UAP Algorithm

Initialize $\mathbf{u_r} \leftarrow 0, \Delta\mathbf{u_r} \leftarrow 0$
repeat
  for $\mathbf{B} \in \mathbf{X}$ **do**
    Sample $\hat{t} \subset T$
    $\Delta\mathbf{u_r} \leftarrow \alpha\Delta\mathbf{u_r} - \frac{\nu}{|\hat{\mathbf{x}}|\times|\hat{t}|}\sum_{i=1}^{|\hat{\mathbf{x}}|}\sum_{j=1}^{|\hat{t}|}\nabla L[f(\hat{\mathbf{x}}_\mathbf{i} + \hat{t}_j(\mathbf{u_r})), f(\hat{\mathbf{x}}_\mathbf{i})]$
    Update the perturbation with projection:
    $\mathbf{u} \leftarrow \mathcal{P}_{p,\epsilon}(\mathbf{u_r} + \Delta\mathbf{u_r})$
  end for
until $ASR_R(f, \mathbf{X}, T, \gamma, \mathbf{u_r}) < \zeta$

I. Iterative UAP Algorithm

Moosavi-Dezfooli et al. (2017) introduces an iterative UAP algorithm, the algorithm can be seen in Algorithm 3.

J. Estimate Robustness Algorithm

In this section, we give our algorithm for estimating the robustness of a UAP.

K. Experiment Parameters

In our experiments, we have capped all algorithms at 5 epochs or if they have achieved an ASR_R of 0.95. The UAPs are trained with the same transformation set that they are evaluated on. For algorithms running PGD internally, we have capped the number of iterations to 40.

Algorithm 3 Iterative Universal Perturbation Algorithm (Moosavi-Dezfooli et al. (2017))

Initialize $\mathbf{u} \leftarrow 0$
repeat
  for $\mathbf{x_i} \in \mathbf{X}$ **do**
    if $\hat{f}(\mathbf{x_i} + \mathbf{u}) = \hat{f}(\mathbf{x_i})$ **then**
      Compute minimal adversarial perturbation:
      $\Delta\mathbf{u} \leftarrow \arg\min_{\mathbf{r}} ||\mathbf{r}||_2$ s.t. $\hat{f}(\mathbf{x_i} + \mathbf{u} + \mathbf{r}) \neq \hat{f}(\mathbf{x_i})$
      Update the perturbation with projection:
      $\mathbf{u} \leftarrow \mathcal{P}_{p,\epsilon}(\mathbf{u} + \Delta\mathbf{u})$
    end if
  end for
until $ASR_U(f, \mathbf{X}, \mathbf{u}) < \gamma$

Algorithm 4 EstRobustness

Draw $n = \lceil \frac{1}{2\psi^2} \ln \frac{2}{\phi} \rceil$ samples $\tau_i \sim T$
Return $\hat{p}_n(\gamma) = \frac{1}{n} \sum_i^n I(ASR_U(f, \mathbf{X}, \tau_i(\mathbf{u_r})) > \gamma)$

L. Further Evaluation of Uniform Noise

A table of results for uniform random noise can be seen in Table 5.

Table 5. Robust ASR with uniform random noise, $\gamma = 0.8$.

DATASET	TRANSFORMATION SET	STANDARD UAP	SGD	STANDARD UAP_RP	ROBUST UAP
ILSVRC 2012	$U(0.1)$	81.6%	94.2%	91.3%	**99.0%**
	$U(0.3)$	10.7%	68.9%	42.7%	**96.1%**
CIFAR-10	$U(0.1)$	66.0%	98.1%	96.1%	**100%**
	$U(0.3)$	5.8%	96.1%	47.6%	**100%**

M. UAP performance against semantic transformations on CIFAR-10

In this section, we show similar results as shown in ILSVRC-2012 but on CIFAR-10. Here, we again see that RobustUAP outperforms all other baselines.

N. Average ASR_U and ASR_R with different γ's

We provide additional metrics computed on the same set of transformations, datasets, and models as in Table 1. In Table 6, we present the Average ASR_U rather than ASR_R. The average shows us that our `RobustUAP` algorithm creates UAPs which after transformation on average are better UAPs than all other algorithms. We observe that the average shows us that even standard UAPs aren't completely ineffective after transformation they just have a very low chance of being highly effective.

In Table 7, we present ASR_R computed at $\gamma = [0.5, 0.7]$ rather than $\gamma = 0.6$. This table shows a similar story to above, and shows that our algorithm produces better results under a variety of success thresholds.

O. Robustness to Random Noise

First, we show that our algorithm generates UAPs robust against uniform random noise. Here our neighborhood is defined as an L_∞ ball of radius ϵ around the perturbation. $U(\epsilon)$ represents noise drawn uniformly from such a ball. Figure 6 shows the performance of each algorithm. For example, the `RobustUAP` algorithm achieves a ASR_U of 0.9 greater than 97% of the time under $U(0.1)$ on CIFAR-10, where all other algorithms achieve 0.9 at most 30% of the time. `RobustUAP`

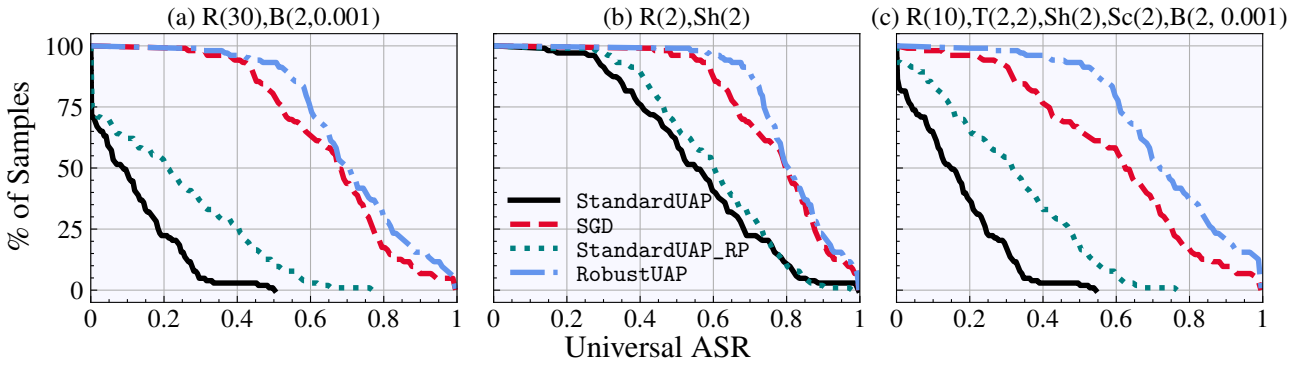

Figure 5. For each method, a point (x, y) in the corresponding line represents the percentage of sampled UAPs ($y\%$) with Universal ASR $> x$ for the different semantic transformations on CIFAR-10.

Table 6. Average Universal ASR of our Robust UAP algorithms and the standard UAP (Moosavi-Dezfooli et al., 2017) method.

DATASET	TRANSFORMATION SET	STANDARD UAP	SGD	STANDARD UAP_RP	ROBUST UAP
ILSVRC 2012	$R(20)$	16.3%	71.5%	24.7%	**81.3%**
	$T(2,2)$	52.6%	82.6%	55.4%	**85.4%**
	$Sc(5), R(5), B(5, 0.01)$	44.9%	76.3%	58.5%	**82.2%**
	$R(10), T(2,2), Sh(2), Sc(2), B(2, 0.001)$	13.6%	64.8%	29.0%	**75.3%**
CIFAR-10	$R(30), B(2, 0.001)$	9.9%	66.8%	22.2%	**73.4%**
	$R(2), Sh(2)$	57.1%	78.8%	61.2%	**82.9%**
	$R(10), T(2,2), Sh(2), Sc(2), B(2, 0.001)$	16.2%	61.2%	32.6%	**76.4%**

outperforms all other algorithms for both noise sizes. `StandardUAP` has a lower mean and higher variance in universal ASR and is much less robust to transformation. A table of Robust ASR results for $\gamma = 0.8$ can be seen in Appendix L. Our Robust ASR results are guaranteed to be ± 0.05 from the actual result with a probability of 95%. For example, we estimate that `RobustUAP` has ASR_R of 96.1% for U(0.3), we are guaranteed that the true robustness is $> 91.1\%$ with a probability of 95%. Note that we get robustness guarantees from `EstRobustness` as our neighborhood has a well-defined PDF.

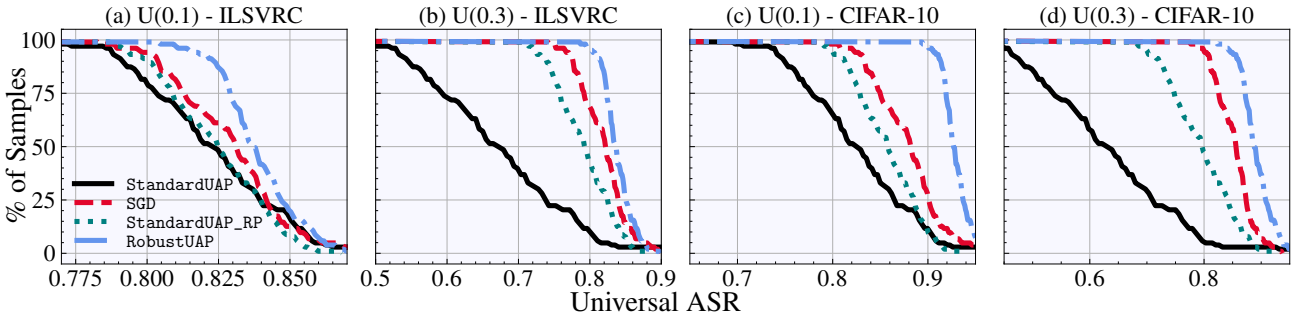

Figure 6. For each method, a point (x, y) in the corresponding line represents the percentage of sampled UAPs ($y\%$) with Universal ASR $> x$ for $U(0.1)$ and $U(0.3)$ on ILSVRC and CIFAR-10.

P. Comparison on non-Robust Universal ASR metric

We compare our robust UAPs to standard UAPs on the non-robust universal ASR metric, see Table 8. All robust UAPs are generated to be robust against $R(10), T(2,2), Sh(2), Sc(2), B(2, 0.001)$. We observe that at the same l_2-norm all robust

Table 7. Robust ASR of our Robust UAP algorithms and the standard UAP (Moosavi-Dezfooli et al., 2017) method with $\gamma = [0.5, 0.7]$.

DATASET	TRANSFORMATION SET	STANDARD UAP		SGD		STANDARD UAP_RP		ROBUST UAP	
		0.5	0.7	0.5	0.7	0.5	0.7	0.5	0.7
ILSVRC 2012	$R(20)$	1.9%	0.0%	88.3%	58.3%	10.7%	1.0%	**98.1%**	**76.7%**
	$T(2,2)$	51.5%	21.4%	**100%**	84.5%	57.3%	23.3%	**100%**	**91.3%**
	$Sc(5), R(5), B(5, 0.01)$	38.8%	11.7%	96.1%	67.0%	64.1%	25.2%	**99.0%**	**87.4%**
	$R(10), T(2,2), Sh(2), Sc(2), B(2, 0.001)$	1.9%	0.0%	82.5%	38.8%	12.6%	1.0%	**95.1%**	**59.2%**
CIFAR-10	$R(30), B(2, 0.001)$	1.0%	0.0%	80.6%	43.7%	12.6%	1.0%	**93.2%**	**49.5%**
	$R(2), Sh(2)$	62.1%	22.3%	96.1%	68.9%	68.0%	30.1%	**99.0%**	**89.3%**
	$R(10), T(2,2), Sh(2), Sc(2), B(2, 0.001)$	2.9%	0.0%	67.0%	38.8%	19.4%	1.0%	**93.2%**	**55.3%**

UAPs achieve a lower universal ASR than the standard UAP algorithm. This result is not too surprising as solving the optimization problem for robust UAP is significantly more difficult. We further observe that our `RobustUAP` algorithm is the most effective in comparison to the other robust baseline approaches.

Table 8. Universal ASR of our Robust UAP algorithms and the standard UAP method.

DATASET	STANDARDUAP	SGD	STANDARDUAP_RP	ROBUSTUAP
ILSVRC 2012	**95.5%**	85.6%	82.3%	91.3%
CIFAR-10	**96.2%**	89.3%	84.0%	93.7%

Q. Additional Models and Datasets

We also provide additional data on our methods evaluated on the same transformations and datasets but on different models. In this case, we use ResNet-18 (He et al., 2015) for CIFAR-10 and MobileNet (Howard et al., 2017) for ILSVRC 2012. Results can be seen in Table 9. We observe similar performance across models suggesting that the performance of the attacks is more directly tied to transformation set and dataset. We also provide data for additional dataset/model pairs. Using the $R(10), T(2,2), Sh(2), Sc(2), B(2, 0.001)$ transformation set we provide results for CIFAR-10 (ShuffleNet (Zhang et al., 2018)), ILSVRC (CaffeNet (Jia et al., 2014)), MNIST (Deng, 2012) using a CNN with 2 convolutional layers and 2 linear layers, and Fashion-MNIST (FMNIST) (Xiao et al., 2017) with 4 linear layers (MNIST and FMNIST use $\epsilon = 5$) in Table 10

Table 9. Robust ASR on Resnet-18 for CIFAR-10 and MobileNet for ILSVRC 2012.

DATASET	MODEL	TRANSFORMATION SET	STANDARD UAP	SGD	STANDARD UAP_RP	ROBUST UAP
ILSVRC 2012	MOBILENET	$R(20)$	8.1%	71.2%	2.6%	**85.0%**
		$T(2,2)$	40.9%	98.7%	54.3%	**99.6%**
		$Sc(5), R(5), B(5, 0.01)$	16.3%	94.5%	44.3%	**96.3%**
		$R(10), T(2,2), Sh(2), Sc(2), B(2, 0.001)$	4.1%	75.7%	8.6%	**86.2%**
CIFAR-10	RESNET-18	$R(30), B(2, 0.001)$	0.9%	67.8%	6.4%	**74.9%**
		$R(2), Sh(2)$	49.9%	99.5%	49.1%	**99.8%**
		$R(10), T(2,2), Sh(2), Sc(2), B(2, 0.001)$	8.0%	70.8%	12.2%	**83.8%**

R. Common Corruptions

We also evaluate robust UAP against the 2D fog transformations in (Kar et al., 2022). We set the shift intensity of the fog to be 1 and train our robust UAPs to be robust against random fog perturbations. We observe similar results to the transformations we experiment with above. The graph of the results can be seen in Figure 7.

Table 10. Robust ASR for additional datasets/models

DATASET	MODEL	STANDARD UAP	SGD	STANDARD UAP_RP	ROBUST UAP
ILSVRC	CAFFENET	0.0%	52.1%	8.1%	**78.3%**
CIFAR10	SHUFFLENET	0.0%	76.7%	12.3%	**88.7%**
MNIST	CNN(2C,2L)	0.0%	25.2%	0.0%	**31.2%**
FMNIST	FC(4L)	0.0%	31.0%	0.0%	**35.9%**

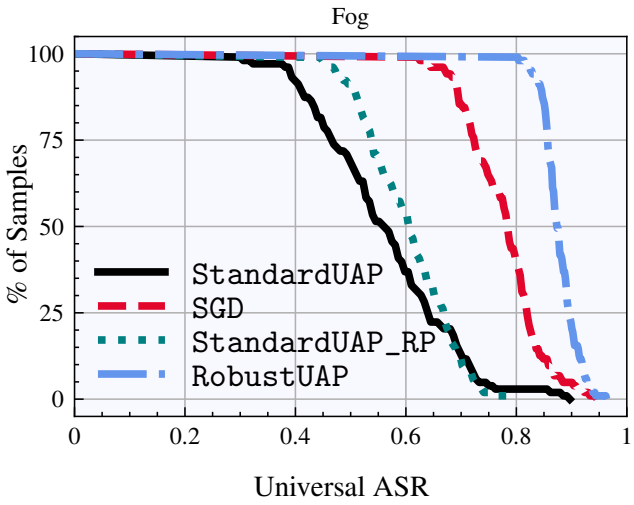

Figure 7. For each method, a point (x, y) in the corresponding line represents the percentage of sampled UAPs ($y\%$) with Universal ASR $> x$ for the different semantic transformations on ILSVRC-2012.

S. Scaling and Hue

We have further added an experiment using $Sc(20)$ and Hue as our transformation set on CIFAR-10 shown in Table 11.

Table 11. Scaling and Hue on CIFAR-10 with VGG16

TRANSFORMATION SET	STANDARD UAP	SGD	STANDARD UAP_RP	ROBUST UAP
$Sc(20)$	5.4%	85.9%	16.8%	**91.3%**
HUE	68.1% 91.3%	85.6%	**94.2%**	

T. Robust Transformation Set

We take Robust UAPs generated by [R(20)], [T(2,2)], and [Sc(5), R(5), B(5, 0.01)] and test them on R(10), T(2, 2), Sh(2), Sc(2), B(2, 0.001) for ILSVRC. We get robust ASRs of 69.5%, 24.6%, and 71.3% when training on $R(20)$, $T(2, 2)$, and $Sc(5), R(5), B(5, 0.01)$ respectively. This shows that Sc(5), R(5), B(5, 0.01) generates the most robust UAPs when considering the robustness to the R(10), T(2, 2), Sh(2), Sc(2), B(2, 0.001) transformation set.

U. Transferability to Different Domains

We experimented with Imagenet-C (Hendrycks and Dietterich, 2018) a dataset consisting of common corruptions applied to ILSVRC in Table 12. This dataset is out-of-distribution for the models we trained on ILSVRC as the images are corrupted by transformations unseen in the original dataset. RobustUAP trained on Imagenet-C obtains a robust ASR of 81.5%. We observe that we still get transferability between Imagenet-C and ILSVRC. The transferability is less than we observe across models trained on the same dataset but still substantial. We observe that transformations trained on Imagenet-C transfer

better to ILSVRC compared to the other way around (i.e. 61.3% transferring from Imagenet-C to Inception vs. 45.3% the other way). We will leave further investigation and optimization of dataset-blind robust UAPs for future work.

Table 12. Transferability between Imagenet-C and ILSVRC

DIRECTION	INCEPTION	MOBILENET	INCEPTIONR20	INCEPTIONHF	VIT
IMAGENET-C → ILSVRC	61.3	56.1	62.5	67.3	10.1
ILSVRC → IMAGENET-C	45.3	28.6	31.5	41.1	7.3

V. Robustness against Robust (to standard AP) Models

To have UAPs one must be able to at least generate standard adversarial perturbations; thus, UAP performance is generally poor when attacking models trained to be robust for standard adversarial attacks. However, note that these models (Nie et al., 2022; Croce et al., 2020) have lower accuracy and are not ideal. The goal of our paper is to highlight a practical threat model for DNNs and to inspire work in defending against such a threat model. In Section 5.6, we note that one interesting observation is that training models to be robust against Robust UAPs leads to lower training times and higher standard accuracy than when performing standard adversarial training. That being said, we also perform experiments on CIFAR-10 using the R(10), T(2, 2), Sh(2), Sc(2), B(2, 0.001) transformation set. We find an average ASR_U of 7.6%, 2.6%, and 3.1% for RobustUAP on DiffPure (Nie et al., 2022), RobustBench Rank 1, and RobustBench Rank 3 respectively (note RobustBench Rank 2 is same technique as Rank 1) (Croce et al., 2020). On the same models standard UAP gets an average ASR_U of 0.0%.

W. Algorithm Runtimes

We compare the average runtimes of the different methods on one of our most challenging $R(10), T(2,2), Sh(2), Sc(2), B(2, 0.001)$ transformation set on ILSVRC-2012 and $n = 738$. The results are in Table 13. We observe that `RobustUAP` is the slowest algorithm and `SGD` is the fastest. `RobustUAP` uses `EstimateRobustness` in each loop and thus with high n it requires much more time to compute. The extra computation enables `Robust UAP` to obtain better robustness than all baselines. On the same set of transformations and dataset we observe that one iteration of `EstimateRobustness` on the entire test set takes on average 19 minutes. When running `EstimateRobustness` in the `RobustUAP` loop, each call takes 36 seconds for a batch size of 32.

Table 13. Average Runtime for Robust UAP algorithms

ALGORITHM	TIME(MIN)
STANDARD UAP	37
SGD	32
STANDARD UAP_RP	43
ROBUST UAP	118

X. Effect of Compute Time on `SGD` vs. `RobustUAP` Performance

We further compare the performance of `SGD` and `RobustUAP` by varying the number of SGD iterations. We compute the robust ASR on ILSVRC for robustness against $R(10), T(2,2), Sh(2), Sc(2), B(2, 0.001)$. Figure 8, shows the robust ASR achieved by SGD over time, here we observe that SGD's performance flatlines after a small number of iterations and seems to be unable to surpass about 65. Here `SGD` is allowed to continue to run past where it would usually stop (at around 250 iterations), in this experiment we allow it to go to 1250 iterations which is about the same amount of time that `RobustUAP` takes to run. `RobustUAP` is able to achieve a performance of 72 even when restricted to the amount of compute time of base `SGD` (It achieves 86.4 when unrestricted). These two results in combination show that `RobustUAP` is able to find more robust UAPs than `SGD` whose performance stabilizes.

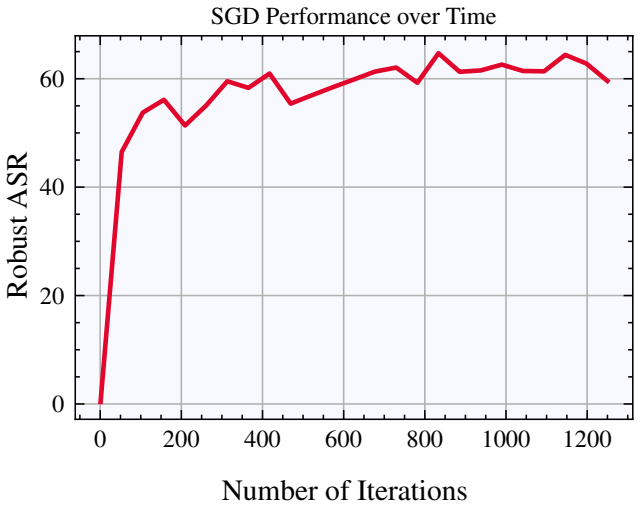

Figure 8. The Robust ASR with $\gamma = 0.6$ for SGD over time

Y. Targeted Attack

So far in this paper we have focused on untargeted attacks, i.e. attacks which aim to degrade the general performance of the model. Targeted attacks are also possible with both standard adversarial attack methods and universal adversarial perturbation methods. Here, we can simply turn our algorithm from untargeted to targeted by replacing the loss function. We would like to have target class, A, be classified as target class, B. Instead of maximizing the expected value of the cross entropy loss we can instead formulate the loss based on maximizing B while minimizing A similar to (Benz et al., 2020). For ILSVRC 2012, we randomly select a couple of target classes and perform this attack, for each of these cases, we train our robust UAP to be robust to $R(10), T(2,2), Sh(2), Sc(2), B(2, 0.001)$. Table 14 shows our results for robust ASR with $\gamma = 0.6$. We are measuring our robust ASR of turning class A into class B and observe similar results with RobustUAP being the most robust followed by SGD. It is also interesting to note that different random combinations lead to more or less success, i.e. it is easier to turn a dog into another dog than perfume into a padlock.

Table 14. Robust ASR of RobustUAP for target to target attack compared to the three baselines with $\gamma = 0.6$.

DATASET	TARGET CLASS	STANDARD UAP	SGD	STANDARD UAP_RP	ROBUST UAP
ILSVRC-2012	TOY POODLE → MALTESE DOG	42.4%	99.1%	85.6%	**99.8%**
	PERFUME → PADLOCK	0.0%	63.8%	5.1%	**76.4%**

Z. Data Efficiency

In this section, we will evaluate the data efficiency of RobustUAP. We use RobustUAP to generate UAPs robust to $R(10), T(2,2), Sh(2), Sc(2), B(2, 0.001)$ on ILSVRC-2012 with differing amounts of training data. The results can be seen in Figure 9. These results show that the algorithm is able to achieve good performance at 500 data points but continues to improve up to 4000 data points. After that it seems to stagnate.

AA. Transformer-based Models

Recently, transformers have become popular as a new architecture for deep learning models for computer vision tasks. In this section, we evaluate the effectiveness of robust UAPs against one such model, ViT (Dosovitskiy et al., 2020). Benz et al. (2021) has shown that standard UAPs are still effective against transformer based architectures. In Table 15 we can see that we get similar results compared to our results on Inception and MobileNet. This shows that our methods work against transformer based models as well.

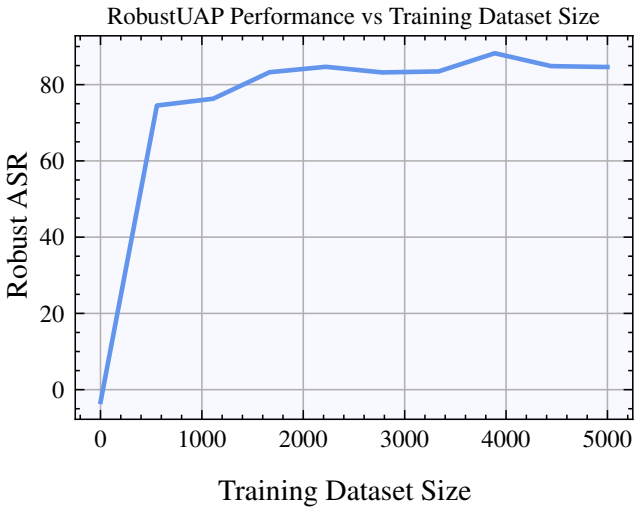

Figure 9. Robust ASR with $\gamma = 0.6$ for `RobustUAP` with differing amounts of training data

Table 15. Robust ASR of `RobustUAP` compared to the three baselines for ViT

Dataset	Model	Transformation Set	Standard UAP	SGD	Standard UAP_RP	Robust UAP
ILSVRC-2012	ViT	$R(10), T(2,2), Sh(2), Sc(2), B(2, 0.001)$	2.0%	72.1%	12.9%	**88.5%**

AB. Robust UAPs against Robustly Trained Networks

In this section, we are interested in seeing whether training networks to be robust against the same transformations that the UAP is trying to be robust against is helpful. For this, we trained two new Inception-v3 networks. Because of time limitations, we started with our base Inception-v3 network and fine-tuned it using data augmentations. For the first network InceptionR20, we augmented the data by adding random rotations within 20 degrees. For the second network InceptionHF, we augmented the data by adding horizontal flips. We then crafted UAPs robust against rotations and flips on InceptionR20 and InceptionHF respectively. The results can be seen in Table 16. We can compare the $R(20)$ results to those from our normal inception network. We postulate that since the network has received some additional robustness training it is harder to attack, and thus we should see slightly lower robustness scores. However, it seems that training the network to be robust to $R(20)$ does not significantly effect the ability to create robust UAPs. The horizontal flips seems like it might be too easy of a transformation as even standard UAP performs quite well for robust ASR.

Table 16. Robust ASR of `RobustUAP` compared to the three baselines for robust networks.

Dataset	Model	Transformation Set	Standard UAP	SGD	Standard UAP_RP	Robust UAP
ILSVRC-2012	InceptionR20	$R(20)$	6.3%	72.4%	10.2%	**81.3%**
	InceptionHF	HF	81.3%	99.5%	89.7%	**99.6%**

AC. Ablation on optimization strategy

In this section, we study the effect of using different optimizers in addition to SGD. We use a variety of standard PyTorch optimizers, Adam, Adamax, Adagrad, and RMSProp. We formulate the optimization problem in the same way but instead use these algorithms in order to optimize our perturbation. We compute these results on ILSVRC-2012 with Inception-v3 and use $R(10), T(2,2), Sh(2), Sc(2), B(2, 0.001)$ as the transformation set and with $\gamma = 0.6$. The results can be seen in Table 17. We see that the optimization strategy has some affect on the results and that SGD performs the best. We also

found that SGD performed marginally faster than the rest of the approaches.

Table 17. Comparison of different optimization strategies.

OPTIMIZER	ASR_R
SGD	**63.1%**
ADAM	59.7%
ADAMAX	60.1%
ADAGRAD	62.3%
RMSPROP	58.3%

AD. Visualization of Robust UAPs vs UAP vs RobustUAP w/ Primatives

We visualize UAPs generated with different algorithms transformed randomly from $R(10)$, $T(2,2)$, $Sh(2)$, $Sc(2)$, $B(2, 0.001)$ and added to images in ILSVRC 2012 in Figure 10.

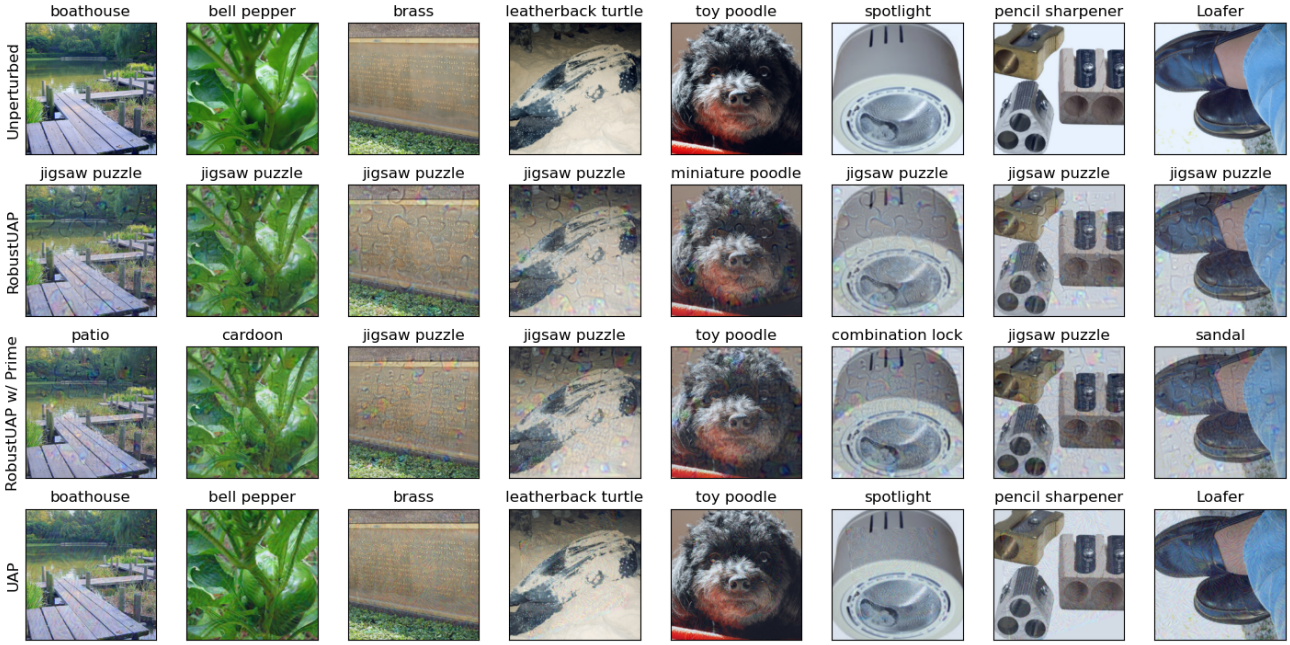

Figure 10. Examples of perturbed images with labels. The top row is unperturbed ILSVRC 2012 test set images, the second row has a randomly transformed robust UAP added to it, the third row has a randomly transformed robust UAP trained with Prime added to it, and the bottom row has a randomly transformed standard UAP added to it. Labels calculated using Inception-v3.

AE. Error Bars

In this section, we will provide error bars/standard deviations for results reported in the paper.

AE.1. Table 1

First, we report the standard deviations for Table 1 which we obtain by learning each UAP 10 times then evaluating them on each their respective dataset/transformation set combinations.

AE.2. Table 2

First, we report the standard deviations for Table 2 which we obtain by learning each UAP 10 times then evaluating them on each their respective dataset/transformation set combinations.

AE.3. Remaining Values

We find that the standard deviations are pretty similar across both tables reported and in some testing of the remaining results. For time reasons we have left the remaining standard deviations out as we don't find them informative. We are happy to provide these numbers for any results in the main body or appendix of the paper.

Table 18. Standard Deviation of Robust ASR values reported in Table 1

DATASET	TRANSFORMATION SET	STANDARD UAP	SGD	STANDARD UAP_RP	ROBUST UAP
ILSVRC 2012	$R(20)$	1.1%	7.2%	10.2%	1.5%
	$T(2,2)$	11.4%	1.8%	6.3%	2.3%
	$Sc(5), R(5), B(5, 0.01)$	10.1%	5.2%	7.4%	3.1%
	$R(10), T(2,2), Sh(2), Sc(2), B(2, 0.001)$	0.0%	8.6%	4.9%	2.8%
CIFAR-10	$R(30), B(2, 0.001)$	0.2%	7.5%	9.0%	5.2%
	$R(2), Sh(2)$	11.3%	5.5%	8.2%	0.9%
	$R(10), T(2,2), Sh(2), Sc(2), B(2, 0.001)$	1.8%	6.8%	5.1%	3.7%

Table 19. Robust ASR (%) of `RobustUAP` trained on PRIME, Affine ($R(10)$, $T(2,2)$, $Sh(2)$, $Sc(2)$, $B(2,0.001)$), and Fog when applied to Prime, Affine, and common corruption transforms

	Evaluation Corruption Set																
Train Set			Noise			Blur				Weather				Digital			
	Prime	Aff.	Gaus.	Shot	Imp.	Defo.	Glass	Moti.	Zoom	Snow	Fog	Frost	Bright	Contr.	Elast.	Pixel	JPEG
PRIME	3.7	4.6	6.1	3.2	4.6	1.3	3.8	4.9	6.7	2.4	7.5	4.3	1.4	5.2	3.1	4.6	2.7
Affine	9.7	5.9	3.0	2.7	5.2	8.7	7.6	5.2	0.7	11.2	6.9	6.6	5.2	0.6	3.8	9.6	6.8
Fog	3.1	5.2	4.7	6.6	4.6	1.9	8.4	3.5	5.1	17.8	1.6	12.1	4.0	3.6	7.3	1.1	12.6