TESTING ROBUSTNESS AGAINST UNFORESEEN AD-VERSARIES

Anonymous authors

Paper under double-blind review

Abstract

Most existing adversarial defenses only measure robustness to L_p adversarial attacks. Not only are adversaries unlikely to exclusively create small L_p perturbations, adversaries are unlikely to remain fixed. Adversaries adapt and evolve their attacks; hence adversarial defenses must be robust to a broad range of *unforeseen attacks*. We address this discrepancy between research and reality by proposing a new evaluation framework called ImageNet-UA. Our framework enables the research community to test ImageNet model robustness against attacks not encountered during training. To create ImageNet-UA's diverse attack suite, we introduce a total of four novel adversarial attacks. We also demonstrate that, in comparison to ImageNet-UA, prevailing L_{∞} robustness assessments give a narrow account of adversarial robustness. By evaluating current defenses with ImageNet-UA, we find they provide little robustness to unforeseen attacks. We hope the greater variety and realism of ImageNet-UA enables development of more robust defenses which can generalize beyond attacks seen during training.

1 INTRODUCTION

Neural networks perform well on many datasets (He et al., 2016) yet can be consistently fooled by minor adversarial distortions (Goodfellow et al., 2014). The research community has responded by quantifying and developing adversarial defenses against such attacks (Madry et al., 2017), but these defenses and metrics have two key limitations.

First, the vast majority of existing defenses exclusively defend against and quantify robustness to L_p -constrained attacks (Madry et al., 2017; Cohen et al., 2019; Raff et al., 2019; Xie et al., 2018). Though real-world adversaries are not L_p constrained (Gilmer et al., 2018) and can attack with diverse distortions (Brown et al., 2017; Sharif et al., 2019), the literature largely ignores this and evaluates against the L_p adversaries already seen during training (Madry et al., 2017; Xie et al., 2018), resulting in optimistic robustness assessments. The attacks outside the L_p threat model that have been proposed (Song et al., 2018; Qiu et al., 2019; Engstrom et al., 2017; Evtimov et al., 2017; Sharif et al., 2016) are not intended for general defense evaluation and suffer from narrow dataset applicability, difficulty of optimization, or fragility of auxiliary generative models.

Second, existing defenses assume that attacks are known in advance (Goodfellow, 2019) and use knowledge of their explicit form during training (Madry et al., 2017). In practice, adversaries can deploy *unforeseen attacks* not known to the defense creator. For example, online advertisers use attacks such as perturbed pixels in ads to defeat ad blockers trained only on the previous generation of ads in an ever-escalating arms race (Tramèr et al., 2018). However, current evaluation setups implicitly assume that attacks encountered at test-time are the same as those seen at train-time, which is unrealistic. The reality that future attacks are unlike those encountered during training is akin to a train-test distribution mismatch—a problem studied outside of adversarial robustness (Recht et al., 2019; Hendrycks & Dietterich, 2019)—but now brought to the adversarial setting.

The present work addresses these limitations by proposing an evaluation framework ImageNet-UA to measure robustness against unforeseen attacks. ImageNet-UA assesses a defense which may have been created with knowledge of the commonly used L_{∞} or L_2 attacks with six diverse attacks (four of which are novel) distinct from L_{∞} or L_2 . We intend these attacks to be used at *test-time* only and not during training. Performing well on ImageNet-UA thus demonstrates generalization to a diverse set of distortions not seen during defense creation. While ImageNet-UA



Figure 1: Adversarially distorted chow chow dog images created with old attacks and our new attacks. The JPEG, Fog, Snow, and Gabor adversarial attacks are visually distinct from previous attacks, result in distortions which do not obey a small L_p norm constraint, and serve as unforeseen attacks for the ImageNet-UA attack suite.

does not provide an exhaustive guarantee over all conceivable attacks, it evaluates over a diverse unforeseen test distribution similar to those used successfully in other studies of distributional shift (Rajpurkar et al., 2018; Hendrycks & Dietterich, 2019; Recht et al., 2019). ImageNet-UA works for ImageNet models and can be easily used with our code available at https://github.com/anon-submission-2020/anon-submission-2020.

Designing ImageNet-UA requires new attacks that are strong and varied, since real-world attacks are diverse in structure. To meet this challenge, we contribute four novel and diverse adversarial attacks which are easily optimized. Our new attacks produce distortions with occlusions, spatial similarity, and simulated weather, all of which are absent in previous attacks. Performing well on ImageNet-UA thus demonstrates that a defense generalizes to a diverse set of distortions distinct from the commonly used L_{∞} or L_2 .

With ImageNet-UA, we show weaknesses in existing evaluation practices and defenses through a study of 8 attacks against 48 models adversarially trained on ImageNet-100, a 100-class subset of ImageNet. While most adversarial robustness evaluations use only L_{∞} attacks, ImageNet-UA reveals that models with high L_{∞} attack robustness can remain susceptible to other attacks. Thus, L_{∞} evaluations are a narrow measure of robustness, even though much of the literature treats this evaluation as comprehensive (Madry et al., 2017; Qian & Wegman, 2019; Schott et al., 2019; Zhang et al., 2019). We address this deficiency by using the novel attacks in ImageNet-UA to evaluate robustness to a more diverse set of unforeseen attacks. Our results demonstrate that L_{∞} adversarial training, the current state-of-the-art defense, has limited generalization to unforeseen adversaries, and is not easily improved by training against more attacks. This adds to the evidence that achieving robustness against a few train-time attacks is insufficient to impart robustness to unforeseen test-time attacks (Jacobsen et al., 2019; Jordan et al., 2019; Tramèr & Boneh, 2019).

In summary, we propose the framework ImageNet-UA to measure robustness to a diverse set of attacks, made possible by our four new adversarial attacks. Since existing defenses scale poorly to multiple attacks (Jordan et al., 2019; Tramèr & Boneh, 2019), finding defense techniques which generalize to unforeseen attacks is crucial to create robust models. We suggest ImageNet-UA as a way to measure progress towards this goal.

2 RELATED WORK

Adversarial robustness is notoriously difficult to correctly evaluate (Papernot et al., 2017; Athalye et al., 2018a). To that end, Carlini et al. (2019a) provide extensive guidance for sound adversarial robustness evaluation. By measuring attack success rates across several distortion sizes and using a broader threat model with diverse differentiable attacks, ImageNet-UA has several of their recommendations built-in, while greatly expanding the set of attacks over previous work on evaluation.



Figure 2: Randomly sampled distortions and adversarially optimized distortions from our new attacks, targeted to the target class in red. Stochastic average-case versions of our attacks affect classifiers minimally, while adversarial versions are optimized to reveal high-confidence errors. The snowflakes in Snow decrease in intensity after optimization, demonstrating that lighter adversarial snowflakes are more effective than heavy random snowfall at uncovering model weaknesses.

We are only aware of a few prior works which evaluate on unforeseen attacks in specific limited circumstances. Wu et al. (2020) evaluate against physically-realizable attacks from Evtimov et al. (2017) and Sharif et al. (2016), though this limits the threat model to occlusion attacks on narrow datasets. Outside of vision, Pierazzi et al. (2020) proposes constraining attacks by a more diverse set of problem-space constraints in diverse domains such as text and malware or source code generation; however, even in this framework, analytically enumerating all such constraints is impossible.

Within vision, prior attacks outside the L_p threat model exist, but they lack the general applicability and fast optimization of ours. Song et al. (2018) and Qiu et al. (2019) attack using variational autoencoders and StarGANs, respectively, resulting in weaker attacks which require simple image distributions suitable for VAEs and GANs. Engstrom et al. (2017) apply Euclidean transformations determined by brute-force search. Zhao et al. (2019) use perceptual color distances to align human perception and L_2 perturbations. Evtimov et al. (2017) and Sharif et al. (2016) attack stop signs and face-recognition systems with carefully placed patches or modified eyeglass frames, requiring physical object creation and applying only to specific image types.

3 NEW ATTACKS FOR A BROADER THREAT MODEL

There are few diverse, easily optimizable, plug-and-play adversarial attacks in the current literature; outside of Elastic (Xiao et al., 2018), most are L_p attacks such as L_∞ (Goodfellow et al., 2014), L_2 (Szegedy et al., 2013; Carlini & Wagner, 2017), L_1 (Chen et al., 2018). We rectify this deficiency with four novel adversarial attacks: JPEG, Fog, Snow, and Gabor. Our attacks are differentiable and fast, while optimizing over enough parameters to be strong. We show example adversarial images in Figure 1 and compare stochastic and adversarial distortions in Figure 2.

Our novel attacks provide a range of *test-time* adversaries visually and semantically distinct from L_{∞} and L_2 attacks. Namely, they cause distortions with large L_{∞} and L_2 norm, but result in images that are perceptually close to the original. These attacks are intended as unforeseen attacks not used during training, allowing them to evaluate whether a defense can generalize from L_{∞} or L_2 to a more varied set of distortions than current evaluations. Though our attacks are not exhaustive, performing well against them already demonstrates robustness to occlusion, spatial similarity, and simulated weather, which are absent from previous evaluations.

Our attacks create an adversarial image x' from a clean image x with true label y. Let model f map images to a softmax distribution, and let $\ell(f(x), y)$ be the cross-entropy loss. Given a target class $y' \neq y$, our attacks attempt to find a valid image x' such that (1) the attacked image x' is obtained by applying a distortion (of size controlled by a parameter ε) to x, and (2) the loss $\ell(f(x'), y')$ is minimized. An unforeseen adversarial attack is a white- or black-box adversarial attack unknown to the defense designer which does not change the true label of x according to an oracle or human.

3.1 FOUR NEW UNFORESEEN ATTACKS

JPEG. JPEG applies perturbations in a JPEG-encoded space of compressed images rather than raw pixel space. More precisely, JPEG compression is a linear transform JPEG which applies color-space conversion, the discrete cosine transform, and then quantization. Our JPEG attack imposes the L_{∞} -constraint

$$\|\mathsf{JPEG}(x) - \mathsf{JPEG}(x')\|_{\infty} \leq \varepsilon$$

on the attacked image x'. We optimize z = JPEG(x') under this constraint to find an adversarial perturbation in the resulting frequency space. The perturbed frequency coefficients are quantized, and we then apply a right-inverse of JPEG to obtain the attacked image x' in pixel space. We use ideas from Shin & Song (2017) to make this differentiable. The resulting attack is conspicuously distinct from L_p attacks.

Fog. Fog simulates worst-case weather conditions. Robustness to adverse weather is a safety critical priority for autonomous vehicles, and Figure 2 shows Fog provides a more rigorous stress-test than stochastic fog (Hendrycks & Dietterich, 2019). Fog creates adversarial fog-like occlusions by adversarially optimizing parameters in the diamond-square algorithm (Fournier et al., 1982) typically used to render stochastic fog effects.

This algorithm starts with random perturbations to the four corner pixels of the image. At step t, it iteratively perturbs pixels at the centers of squares and diamonds formed by those pixels perturbed at step t-1. The perturbation of a step t pixel is the average of the neighboring step t-1 perturbations plus a parameter value which we adversarially optimize. We continue this process until all pixels have been perturbed; the outcome is a fog-like distortion to the original image.

Snow. Snow simulates snowfall with occlusions of randomly located small image regions representing snowflakes. Because the distortions caused by snowflakes are not differentiable in their locations, we instead place occlusions representing snowflakes at randomly chosen locations and orientations and adversarially optimize their intensities. This choice results in a fast, differentiable, and strong attack. Compared to synthetic stochastic snow (Hendrycks & Dietterich, 2019), our adversarial snow is faster and includes snowflakes at differing angles. Figure 2 shows adversarial snow exposes model weaknesses more effectively than the easier stochastic, average-case snow.

Gabor. Gabor spatially occludes the image with visually diverse Gabor noise Lagae et al. (2009). Gabor noise is a form of band-limited anisotropic procedural noise which convolves a parameter mask with a *Gabor kernel* which is a product of a Gaussian kernel and a harmonic kernel. We choose the Gabor kernel randomly and adversarially optimize the parameters of the mask starting from a sparse initialization. We apply spectral variance normalization (Co et al., 2019) to the resulting distortion and add it to the input image to create the attack.

3.2 IMPROVING EXISTING ATTACKS

Elastic modifies the attack of Xiao et al. (2018); it warps the image by distortions x' = Flow(x, V), where $V : \{1, \ldots, 224\}^2 \to \mathbb{R}^2$ is a vector field on pixel space, and Flow sets the value of pixel (i, j) to the bilinearly interpolated original value at (i, j) + V(i, j). We construct V by smoothing a vector field W by a Gaussian kernel (size 25×25 , $\sigma \approx 3$ for a 224×224 image) and optimize W under $||W(i, j)||_{\infty} \leq \varepsilon$ for all i, j. The resulting attack is suitable for large-scale images.

The other three attacks are L_1, L_2, L_∞ attacks, but we improve the L_1 attack. For L_∞ and L_2 constraints, we use randomly-initialized projected gradient descent (PGD), which applies gradient descent and projection to the L_∞ and L_2 balls (Madry et al., 2017). Projection is difficult for L_1 , and previous L_1 attacks rely on computationally intensive methods for it (Chen et al., 2018; Tramèr & Boneh, 2019). We replace PGD with the Frank-Wolfe algorithm (Frank & Wolfe, 1956), which



Figure 3: Accuracies of L_2 and Elastic attacks at different distortion sizes against a ResNet-50 model adversarially trained against L_2 at $\varepsilon = 9600$ on ImageNet-100. At small distortion sizes, the model appears to defend well against Elastic, but large distortion sizes reveal that robustness does not transfer from L_2 to Elastic.

optimizes a linear function instead of projecting at each step (pseudocode in Appendix D). This makes our L_1 attack more principled than previous implementations.

4 ImageNet-UA: MEASURING ROBUSTNESS TO UNFORESEEN ATTACKS

We propose the framework ImageNet-UA and its CIFAR-10 analogue CIFAR-10-UA to measure and summarize model robustness while fulfilling the following desiderata: (1) defenses should be evaluated against a broad threat model through a diverse set of attacks, (2) defenses should exhibit generalization to attacks not exactly identical to train-time attacks, and (3) the range of distortion sizes used for an attack must be wide enough to avoid misleading conclusions caused by overly weak or strong versions of that attack (Figure 3).

The ImageNet-UA evaluation framework aggregates robustness information into a single measure, the mean Unforeseen Adversarial Robustness (mUAR). The mUAR is an average over six different attacks of the Unforeseen Adversarial Robustness (UAR), a metric which assesses the robustness of a defense against a specific attack by using a wide range of distortion sizes. UAR is normalized using a measure of attack strength, the ATA, which we now define.

Adversarial Training Accuracy (ATA). The Adversarial Training Accuracy ATA(A, ε) estimates the strength of an attack A against adversarial training (Madry et al., 2017), one of the strongest known defense methods. For a distortion size ε , it is the best adversarial test accuracy against Aachieved by adversarial training against A. We allow a possibly different distortion size ε' during training, since this can improve accuracy, and we choose a fixed architecture for each dataset.

For ImageNet-100, we choose ResNet-50 for the architecture, and for CIFAR-10 we choose ResNet-56. When evaluating a defense with architecture other than ResNet-50 or ResNet-56, we recommend using ATA values computed with these architectures to enable consistent comparison. To estimate ATA(A, ε) in practice, we evaluate models adversarially trained against distortion size ε' for ε' in a large range (we describe this range at this section's end).

UAR: Robustness Against a Single Attack. The UAR, a building block for the mUAR, averages a model's robustness to a single attack over six distortion sizes $\varepsilon_1, \ldots, \varepsilon_6$ chosen for each attack (we describe the selection procedure at the end of this section). It is defined as

$$\mathsf{UAR}(A) := 100 \times \frac{\sum_{k=1}^{6} \mathsf{Acc}(A, \varepsilon_k, M)}{\sum_{k=1}^{6} \mathsf{ATA}(A, \varepsilon_k)},\tag{1}$$

where $Acc(A, \varepsilon_k, M)$ is the accuracy $Acc(A, \varepsilon_k, M)$ of a model M after attack A at distortion size ε_k . The normalization in (1) makes attacks of different strengths more commensurable in a stable way. We give values of $ATA(A, \varepsilon_k)$ and ε_k for our attacks on ImageNet-100 and CIFAR-10 in Tables 4 and 5 (Appendix B), allowing computation of UAR of a defense against a single attack with six adversarial evaluations and no adversarial training.

mUAR: Mean Unforeseen Attack Robustness. We summarize a defense's performance on ImageNet-UA with the mean Unforeseen Attack Robustness (mUAR), an average of UAR scores



Figure 4: UAR for adv trained defenses (row) against attacks (col) on ImageNet-100. Defenses from L_{∞} to Gabor were trained with $\varepsilon = 32, 4.8k$, 612k, 2, 16, 8192, 8, and 1.6k.



Figure 5: UAR (L_{∞}) and mUAR for L_{∞} -trained models at different distortion sizes. Increasing distortion size in L_{∞} -training improves UAR (L_{∞}) but hurts the mUAR, suggesting models heavily fit L_{∞} at the cost of generalization.

for the L_1 , Elastic, JPEG, Fog, Snow, and Gabor attacks:

$$\mathsf{mUAR} := \frac{1}{6} \Big[\mathsf{UAR}(L_1) + \mathsf{UAR}(\text{Elastic}) + \mathsf{UAR}(\text{JPEG}) + \mathsf{UAR}(\text{Fog}) + \mathsf{UAR}(\text{Snow}) + \mathsf{UAR}(\text{Gabor}) \Big].$$

Our measure mUAR estimates robustness to a broad threat model containing six unforeseen attacks at six distortion sizes each, meaning high mUAR requires generalization to several held-out attacks. In particular, it cannot be achieved by the common practice of engineering defenses to a single attack, which Figure 4 shows does not necessarily provide robustness to different attacks.

Our four novel attacks play a crucial role in mUAR by allowing us to estimate robustness to a sufficiently large set of adversarial attacks. As is customary when studying train-test mismatches and distributional shift, we advise against adversarially training with these six attacks when evaluating ImageNet-UA to preserve the validity of mUAR, though we encourage training with *other* attacks.

Distortion Sizes. We explain the ε' values used to estimate ATA and the choice of $\varepsilon_1, \ldots, \varepsilon_6$ used to define UAR. This calibration of distortion sizes adjusts for the fact (Figure 3) that adversarial robustness against an attack may vary drastically with distortion size. Further, the relation between distortion size and attack strength varies between attacks, so too many or too few ε_k values in a certain range may cause an attack to appear artificially strong or weak according to UAR.

We choose distortion sizes between ε_{\min} and ε_{\max} as follows. The minimum distortion size ε_{\min} is the largest ε for which the adversarial accuracy of an adversarially trained model at distortion size ε is comparable to that of a model trained and evaluated on unattacked data (for ImageNet-100, within 3 of 87). The maximum distortion size ε_{\max} is the smallest ε which either reduces adversarial accuracy of an adversarially trained model at distortion size ε below 25 or yields images confusing humans (adversarial accuracy can remain non-zero in this case).

As is typical in recent work on adversarial examples (Athalye et al., 2018b; Evtimov et al., 2017; Dong et al., 2019; Qin et al., 2019), our attacks can be perceptible at large distortion sizes. We make this choice to reflect perceptibility of attacks in real world threat models per Gilmer et al. (2018).

For ATA, we evaluate against models adversarially trained with ε' increasing geometrically from ε_{\min} to ε_{\max} by factors of 2. We then choose ε_k as follows: We compute ATA at ε increasing geometrically from ε_{\min} to ε_{\max} by factors of 2 and take the size-6 subset whose ATA values have minimum ℓ_1 -distance to the ATA values of the L_{∞} attack in Table 4 (Appendix B.1). For example, for Gabor, $(\varepsilon_{\min}, \varepsilon_{\max}) = (6.25, 3200)$, so we compute ATAs at the 10 values $\varepsilon = 6.25, \ldots, 3200$. Viewing size-6 subsets of the ATAs as vectors with decreasing coordinates, we select ε_k for Gabor corresponding to the vector with minimum ℓ_1 -distance to the ATA vector for L_{∞} .

	Clean Accuracy	L_{∞}	L_2	mUAR		Clean Accuracy	L_{∞}	L_2	mUAR
Normal Training	86.7	7.3	17.2	14.0	Normal Training	86.7	7.3	17.2	14.0
$L_{\infty} \varepsilon = 1$	86.2	46.4	54.2	30.7	$L_2 \varepsilon = 150$	86.6	38.0	49.4	27.1
$L_{\infty} \varepsilon = 2$	85.5	59.8	64.4	36.9	$L_2 \varepsilon = 300$	85.9	49.7	60.1	33.3
$L_{\infty} \varepsilon = 4$	83.9	72.1	73.6	42.3	$L_2 \varepsilon = 600$	84.7	61.9	71.6	40.0
$L_{\infty} \varepsilon = 8$	79.8	82.6	72.0	42.2	$L_2 \varepsilon = 1200$	82.3	72.9	82.0	46.8
$L_{\infty} \varepsilon = 16$	74.5	89.1	60.0	37.5	$L_2 \varepsilon = 2400$	76.8	79.6	88.5	50.7
$L_{\infty} \varepsilon = 32$	70.8	88.1	41.9	31.8	$L_2 \varepsilon = 4800$	68.3	80.4	87.7	50.5

Table 1: Clean Accuracy, UAR, and mUAR scores for models adv trained against L_{∞} and L_2 attacks.
L_{∞} training, the most popular defense, provides less robustness than L_2 training. Comparing the
highest mUAR achieved to individual UAR values in Figure 4 indicates a large robustness gap.

Table 2: Clean Accuracy, UAR, and mUAR scores for models jointly trained against (L_{∞}, L_2) . Joint training does not provide much additional robustness.

	Clean Accuracy	L_{∞}	L_2	mUAR
$L_{\infty} \varepsilon = 1, L_2 \varepsilon = 300$	86.1	50.3	60.2	33.6
$L_{\infty} \varepsilon = 2, L_2 \varepsilon = 600$	85.1	62.8	72.5	41.0
$L_{\infty} \varepsilon = 4, L_2 \varepsilon = 1200$	81.3	72.9	81.2	46.9
$L_{\infty} \varepsilon = 8, L_2 \varepsilon = 2400$	76.5	80.0	87.3	50.8
$L_{\infty} \varepsilon = 16, L_2 \varepsilon = 4800$	68.4	81.5	87.9	50.9

5 NEW INSIGHTS FROM ImageNet-UA

We use ImageNet-UA to assess existing methods for adversarial defense and evaluation. First, ImageNet-UA reveals that L_{∞} trained defenses fail to generalize to different attacks, indicating substantial weakness in current L_{∞} adversarial robustness evaluation. We establish a baseline for ImageNet-UA using L_2 adversarial training which is difficult to improve upon by adversarial training alone. Finally, we show non-adversarially trained models can still improve robustness on ImageNet-UA over standard models and suggest this as a direction for further inquiry.

5.1 EXPERIMENTAL SETUP

We adversarially train 48 models against the 8 attacks from Section 3 and evaluate against targeted attacks. We use the CIFAR-10 and ImageNet-100 datasets for ImageNet-UA and CIFAR-10-UA. ImageNet-100 is a 100-class subset of ImageNet-1K (Deng et al., 2009) containing every tenth class by WordNet ID order; we use a subset of ImageNet-1K due to the high compute cost of adversarial training. We use ResNet-56 for CIFAR-10 and ResNet-50 from torchvision for ImageNet-100 (He et al., 2016). We provide training hyperparameters in Appendix A.

To adversarially train against attack A, at each mini-batch we select a uniform random (incorrect) target class for each training image. For maximum distortion size ε , we apply targeted attack A to the current model with distortion size $\varepsilon' \sim \text{Uniform}(0, \varepsilon)$ and take a SGD step using only the attacked images. Randomly scaling ε' improves performance against smaller distortions.

We train on 10-step attacks for attacks other than Elastic, where we use 30 steps due to a harder optimization. For L_p , JPEG, and Elastic, we use step size $\varepsilon/\sqrt{\text{steps}}$; for Fog, Gabor, and Snow, we use step size $\sqrt{0.001/\text{steps}}$ because the latent space is independent of ε . These choices have optimal rates for non-smooth convex functions (Nemirovski & Yudin, 1978; 1983). We evaluate on 200-step targeted attacks with uniform random (incorrect) target, using more steps for evaluation than training per best practices (Carlini et al., 2019b).

Figure 4 summarizes ImageNet-100 results. Full results for ImageNet-100 and CIFAR-10 are in Appendix E and robustness checks to random seed and attack iterations are in Appendix F.

5.2 ImageNet-UA REVEALS WEAKNESSESS IN L_{∞} TRAINING AND TESTING

We use ImageNet-UA to reveal weaknesses in the common practices of L_{∞} robustness evaluation and L_{∞} adversarial training. We compute the mUAR and UAR (L_{∞}) for models trained against the L_{∞} attack with distortion size ε and show results in Figure 5. For small $\varepsilon \leq 4$, mUAR and Table 3: Non-adversarial defenses can noticeably improve ImageNet-UA performance. ResNeXt-101 $(32 \times 8d)$ + WSL is trained on approximately 1 billion images Mahajan et al. (2018). Stylized ImageNet is trained on a modification of ImageNet using style transfer Geirhos et al. (2019). Patch Gaussian augments using Gaussian distortions on small portions of the image Lopes et al. (2019). AugMix mixes simple random augmentations of the image Hendrycks et al. (2020). These results suggest that ImageNet-UA performance may be achieved through non-adversarial defenses.

	Clean Acc.	L_{∞}	L_2	L_1	Elastic	JPEG	Fog	Snow	Gabor	mUAR
SqueezeNet	84.1	5.2	11.2	14.9	25.9	1.9	20.1	9.8	4.4	12.8
ResNeXt-101 (32×8d)	95.9	2.5	5.5	20.7	26.5	1.8	14.1	12.4	5.3	13.4
ResNeXt-101 $(32 \times 8d)$ + WSL	97.1	3.0	5.7	28.3	29.4	1.9	26.2	20.3	8.0	19.0
ResNet-18	91.6	2.7	8.2	13.5	22.6	1.8	20.3	9.5	4.2	12.0
ResNet-50	94.2	2.7	6.6	20.1	24.9	1.8	15.8	11.9	4.9	13.2
ResNet-50 + Stylized ImageNet	94.6	2.9	7.4	22.8	26.0	1.8	16.2	12.5	8.1	14.6
ResNet-50 + Patch Gaussian	93.6	4.5	10.9	27.4	28.2	1.8	23.9	10.5	5.2	16.2
ResNet-50 + AugMix	95.1	6.1	13.4	34.3	38.8	1.8	28.6	24.7	11.1	23.2

 $UAR(L_{\infty})$ increase together with ε . For larger $\varepsilon \geq 8$, $UAR(L_{\infty})$ continues to increase with ε , but the mUAR decreases, a fact which is not apparent from L_{∞} evaluation.

The decrease in mUAR while $UAR(L_{\infty})$ increases suggests that L_{∞} adversarial training begins to heavily fit L_{∞} distortions at the expense of generalization at larger distortion sizes. Thus, while it is the most commonly used defense procedure, L_{∞} training may not lead to improvements on other attacks or to real-world robustness.

Worse, L_{∞} evaluation against L_{∞} adversarial training at higher distortions indicates higher robustness. In contrast, mUAR reveals that L_{∞} adversarial training at higher distortions in fact hurts robustness against a more diverse set of attacks. Thus, L_{∞} evaluation gives a misleading picture of robustness. This is particularly important because L_{∞} evaluation is the most ubiquitous measure of robustness in deep learning (Goodfellow et al., 2014; Madry et al., 2017; Xie et al., 2018).

5.3 LIMITS OF ADVERSARIAL TRAINING FOR ImageNet-UA

We establish a baseline on ImageNet-UA using L_2 adversarial training but show a significant performance gap even for more sophisticated existing adversarial training methods. To do so, we evaluate several adversarial training methods on ImageNet-UA and show results in Table 1.

Our results show that L_2 trained models outperform L_{∞} trained models and have significantly improved absolute performance, increasing mUAR from 14.0 to 50.7 compared to an undefended model. The individual UAR values in Figure 7 (Appendix E.1) improve substantially against all attacks other than Fog, including several (Elastic, Gabor, Snow) of extremely different nature to L_2 .

This result suggests pushing adversarial training further by training against multiple attacks simultaneously via *joint adversarial training* (Jordan et al., 2019; Tramèr & Boneh, 2019) detailed in Appendix C. Table 2 shows that, despite using twice the compute of L_2 training, (L_{∞}, L_2) joint training only improves the mUAR from 50.7 to 50.9. We thus recommend L_2 training as a baseline for ImageNet-UA, though there is substantial room for improvement compared to the highest UARs against individual attacks in Figure 4, which are all above 80 and often above 90.

5.4 ImageNet-UA ROBUSTNESS THROUGH NON-ADVERSARIAL DEFENSES

We find that methods can improve robustness to unforeseen attacks without adversarial training. Table 3 shows mUAR for SqueezeNet (Iandola et al., 2017), ResNeXts (Xie et al., 2016), and ResNets. For ImageNet-1K models, we mask 900 logits to predict ImageNet-100 classes.

A popular defense against average case distortions (Hendrycks & Dietterich, 2019) is Stylized ImageNet (Geirhos et al., 2019), which modifies training images using image style transfer in hopes of making networks rely less on textural features. Table 3 shows it provides some improvement on ImageNet-UA. More recently, Lopes et al. (2019) propose to train against Gaussian noise applied to small image patches, improving the mUAR by 3% over the ResNet-50 baseline. The second largest mUAR improvement comes from training a ResNeXt on approximately 1 billion images (Mahajan et al., 2018). This three orders of magnitude increase in training data yields a 5.4% mUAR increase over a vanilla ResNeXt baseline. Finally, Hendrycks et al. (2020) create AugMix, which randomly mixes stochastically generated augmentations. Although AugMix did not use random nor adversarial noise, it improves robustness to unforeseen attacks by 10%.

These results imply that defenses not relying on adversarial examples can improve ImageNet-UA performance. They indicate that training on more data only somewhat increases robustness on ImageNet-UA, unlike many other robustness benchmarks (Hendrycks & Dietterich, 2019; Hendrycks et al., 2019) where more data helps tremendously (Orhan, 2019). While models with lower clean accuracy (e.g., SqueezeNet and ResNet-18) have higher UAR(L_{∞}) and UAR(L_2) than many other models, there is no clear difference in mUAR. Last, these non-adversarial defenses have minimal cost to accuracy on clean examples, unlike adversarial defenses. Much remains to explore, and we hope non-adversarial defenses will be a promising avenue toward adversarial robustness.

6 CONCLUSION

This work proposes a framework ImageNet-UA to evaluate robustness of a defense against *unforeseen attacks*. Because existing adversarial defense techniques do not scale to multiple attacks, developing models which can defend against attacks not seen at train-time is essential for robustness. Our results using ImageNet-UA show that the common practice of L_{∞} training and evaluation fails to achieve or measure this broader form of robustness. As a result, it can provide a misleading sense of robustness. By incorporating our 4 novel and strong adversarial attacks, ImageNet-UA enables evaluation on the diverse held-out attacks necessary to measure progress towards robustness more broadly.

REFERENCES

- Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*, 2018a.
- Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In Jennifer Dy and Andreas Krause (eds.), Proceedings of the 35th International Conference on Machine Learning, volume 80 of Proceedings of Machine Learning Research, pp. 284–293, Stockholmsmssan, Stockholm Sweden, 10–15 Jul 2018b. PMLR. URL http://proceedings.mlr.press/v80/athalye18b.html.
- Tom B. Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *CoRR*, abs/1712.09665, 2017. URL http://arxiv.org/abs/1712.09665.
- Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In 2017 *IEEE Symposium on Security and Privacy (SP)*, pp. 39–57. IEEE, 2017.
- Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian G Goodfellow, and Aleksander Madry. On evaluating adversarial robustness: Principles of rigorous evaluations. 2019a.
- Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian J. Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. *CoRR*, abs/1902.06705, 2019b. URL http://arxiv.org/abs/1902.06705.
- Pin-Yu Chen, Yash Sharma, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. EAD: Elastic-net attacks to deep neural networks via adversarial examples. In *Thirty-second AAAI conference on artificial intelligence*, 2018.
- Kenneth T. Co, Luis Muñoz-González, and Emil C. Lupu. Sensitivity of deep convolutional networks to Gabor noise. CoRR, abs/1906.03455, 2019. URL http://arxiv.org/abs/ 1906.03455.
- Jeremy M. Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing. *CoRR*, abs/1902.02918, 2019. URL http://arxiv.org/abs/1902.02918.

- 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, pp. 248–255. IEEE, 2009.
- Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019.
- Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. A rotation and a translation suffice: Fooling CNNs with simple transformations. *arXiv preprint arXiv:1712.02779*, 2017.
- Ivan Evtimov, Kevin Eykholt, Earlence Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, and Dawn Xiaodong Song. Robust physical-world attacks on deep learning models. 2017.
- Alain Fournier, Don Fussell, and Loren Carpenter. Computer rendering of stochastic models. Commun. ACM, 25(6):371–384, June 1982. ISSN 0001-0782. doi: 10.1145/358523.358553. URL http://doi.acm.org/10.1145/358523.358553.
- Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.
- Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bygh9j09KX.
- Justin Gilmer, Ryan P. Adams, Ian J. Goodfellow, David Andersen, and George E. Dahl. Motivating the rules of the game for adversarial example research. *ArXiv*, abs/1807.06732, 2018.
- Ian J. Goodfellow. A research agenda: Dynamic models to defend against correlated attacks. *ArXiv*, abs/1903.06293, 2019.
- Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.
- Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch SGD: Training ImageNet in 1 hour. arXiv preprint arXiv:1706.02677, 2017.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pp. 630–645. Springer, 2016.
- Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.
- Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *arXiv preprint arXiv:1907.07174*, 2019.
- Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. AugMix: A simple data processing method to improve robustness and uncertainty. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.
- Forrest N. Iandola, Matthew W. Moskewicz, Khalid Ashraf, Song Han, William J. Dally, and Kurt Keutzer. Squeezenet: AlexNet-level accuracy with 50x fewer parameters and <1mb model size. *ArXiv*, abs/1602.07360, 2017.
- Jrn-Henrik Jacobsen, Jens Behrmannn, Nicholas Carlini, Florian Tramr, and Nicolas Papernot. Exploiting excessive invariance caused by norm-bounded adversarial robustness, 2019.
- Matt Jordan, Naren Manoj, Surbhi Goel, and Alexandros G. Dimakis. Quantifying perceptual distortion of adversarial examples. *arXiv e-prints*, art. arXiv:1902.08265, Feb 2019.

- Ares Lagae, Sylvain Lefebvre, George Drettakis, and Philip Dutré. Procedural noise using sparse Gabor convolution. ACM Trans. Graph., 28(3):54:1–54:10, July 2009. ISSN 0730-0301. doi: 10. 1145/1531326.1531360. URL http://doi.acm.org/10.1145/1531326.1531360.
- Raphael Gontijo Lopes, Dong Yin, Ben Poole, Justin Gilmer, and Ekin Dogus Cubuk. Improving robustness without sacrificing accuracy with patch gaussian augmentation. ArXiv, abs/1906.02611, 2019.
- 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.
- Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (eds.), *Computer Vision – ECCV 2018*, pp. 185–201, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01216-8.
- Arkadi Nemirovski and D Yudin. On Cezari's convergence of the steepest descent method for approximating saddle point of convex-concave functions. In *Soviet Math. Dokl*, volume 19, pp. 258–269, 1978.
- Arkadi Nemirovski and D Yudin. Problem Complexity and Method Efficiency in Optimization. Intersci. Ser. Discrete Math. Wiley, New York, 1983.
- A. Emin Orhan. Robustness properties of Facebook's ResNeXt WSL models. *ArXiv*, abs/1907.07640, 2019.
- Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM* on Asia conference on computer and communications security, pp. 506–519. ACM, 2017.
- Fabio Pierazzi, Feargus Pendlebury, Jacopo Cortellazzi, and Lorenzo Cavallaro. Intriguing properties of adversarial ml attacks in the problem space. In 2020 IEEE Symposium on Security and Privacy (SP), pp. 1308–1325. IEEE Computer Society, 2020. doi: 10.1109/SP40000. 2020.00073. URL https://doi.ieeecomputersociety.org/10.1109/SP40000. 2020.00073.
- Haifeng Qian and Mark N. Wegman. L₂-nonexpansive neural networks. In International Conference on Learning Representations (ICLR), 2019. URL https://openreview.net/forum? id=ByxGSsR9FQ.
- Chongli Qin, James Martens, Sven Gowal, Dilip Krishnan, Krishnamurthy Dvijotham, Alhussein Fawzi, Soham De, Robert Stanforth, and Pushmeet Kohli. Adversarial robustness through local linearization, 2019.
- Haonan Qiu, Chaowei Xiao, Lei Yang, Xinchen Yan, Honglak Lee, and Bo Li. Semanticadv: Generating adversarial examples via attribute-conditional image editing. *ArXiv*, abs/1906.07927, 2019.
- Edward Raff, Jared Sylvester, Steven Forsyth, and Mark McLean. Barrage of random transforms for adversarially robust defense. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6528–6537, 2019.
- P. Rajpurkar, R. Jia, and P. Liang. Know what you don't know: Unanswerable questions for SQuAD. In Association for Computational Linguistics (ACL), 2018.
- Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, 2019.
- L. Schott, J. Rauber, W. Brendel, and M. Bethge. Towards the first adversarially robust neural network model on MNIST. May 2019. URL https://arxiv.org/pdf/1805.09190.pdf.

- 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 23rd ACM SIGSAC Conference on Computer and Communications Security*, 2016.
- Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. A general framework for adversarial examples with objectives. *ACM Transactions on Privacy and Security (TOPS)*, 22(3): 1–30, 2019.
- Richard Shin and Dawn Song. JPEG-resistant adversarial images. In NIPS 2017 Workshop on Machine Learning and Computer Security, 2017.
- Yang Song, Rui Shu, Nate Kushman, and Stefano Ermon. Constructing unrestricted adversarial examples with generative models. In *NeurIPS*, 2018.
- 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.
- Florian Tramèr and Dan Boneh. Adversarial training and robustness for multiple perturbations. *arXiv e-prints*, art. arXiv:1904.13000, Apr 2019.
- Florian Tramèr, Pascal Dupré, Gili Rusak, Giancarlo Pellegrino, and Dan Boneh. Ad-versarial: Defeating perceptual ad-blocking. *CoRR*, abs/1811.03194, 2018. URL http://arxiv.org/abs/1811.03194.
- Tong Wu, Liang Tong, and Yevgeniy Vorobeychik. Defending against physically realizable attacks on image classification. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=H1xscnEKDr.
- Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially transformed adversarial examples. *arXiv preprint arXiv:1801.02612*, 2018.
- Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. *arXiv preprint arXiv:1812.03411*, 2018.
- Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5987–5995, 2016.
- Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7472–7482, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/ v97/zhang19p.html.
- Zhengyu Zhao, Zhuoran Liu, and Marisa Larson. Towards large yet imperceptible adversarial image perturbations with perceptual color distance. *ArXiv*, abs/1911.02466, 2019.

A TRAINING HYPERPARAMETERS

For ImageNet-100, we trained on machines with 8 NVIDIA V100 GPUs using standard data augmentation (He et al., 2016). Following best practices for multi-GPU training (Goyal et al., 2017), we ran synchronized SGD for 90 epochs with batch size 32×8 and a learning rate schedule with 5 "warm-up" epochs and a decay at epochs 30, 60, and 80 by a factor of 10. Initial learning rate after warm-up was 0.1, momentum was 0.9, and weight decay was 10^{-4} . For CIFAR-10, we trained on a single NVIDIA V100 GPU for 200 epochs with batch size 32, initial learning rate 0.1, momentum 0.9, and weight decay 10^{-4} . We decayed the learning rate at epochs 100 and 150.

B CALIBRATION OF ImageNet-UA AND CIFAR-10-UA

B.1 CALIBRATION FOR ImageNet-UA

Calibrated distortion sizes and ATA values are in Table 4.

B.2 CALIBRATION FOR CIFAR-10-UA

The ε calibration procedure for CIFAR-10 was similar to that used for ImageNet-100. We started with small ε_{\min} values and increased ε geometrically with ratio 2 until adversarial accuracy of an adversarially trained model dropped below 40. Note that this threshold is higher for CIFAR-10 than ImageNet-100 because there are fewer classes. The resulting ATA values for CIFAR-10 are shown in Table 5.

C JOINT ADVERSARIAL TRAINING

Our joint adversarial training procedure for two attacks A and A' is as follows. At each training step, we compute the attacked image under both A and A' and backpropagate with respect to gradients induced by the image with greater loss. This corresponds to the "max" loss of Tramèr & Boneh (2019). We train ResNet-50 models for (L_{∞}, L_2) , (L_{∞}, L_1) , and $(L_{\infty}, \text{Elastic})$ on ImageNet-100.

Table 6 shows training against (L_{∞}, L_1) is worse than training against L_1 at the same distortion size and performs particularly poorly at large distortion sizes. Table 7 shows joint training against

Table 4: Calibrated distortion sizes and ATA values for different distortion types on ImageNet-100.

Attack	ε_1	ε_2	ε_3	ε_4	ε_5	ε_6	ATA_1	ATA_2	ATA_3	ATA_4	ATA_5	ATA ₆
L_{∞}	1	2	4	8	16	32	84.6	82.1	76.2	66.9	40.1	12.9
L_2	150	300	600	1200	2400	4800	85.0	83.5	79.6	72.6	59.1	19.9
L_1	9562.5	19125	76500	153000	306000	612000	84.4	82.7	76.3	68.9	56.4	36.1
Elastic	0.25	0.5	2	4	8	16	85.9	83.2	78.1	75.6	57.0	22.5
JPEG	0.062	0.125	0.250	0.500	1	2	85.0	83.2	79.3	72.8	34.8	1.1
Fog	128	256	512	2048	4096	8192	85.8	83.8	79.0	68.4	67.9	64.7
Snow	0.0625	0.125	0.25	2	4	8	84.0	81.1	77.7	65.6	59.5	41.2
Gabor	6.25	12.5	25	400	800	1600	84.0	79.8	79.8	66.2	44.7	14.6

Table 5: Calibrated distortion sizes and ATA values for ResNet-56 on CIFAR-10

Attack	ε_1	ε_2	ε_3	ε_4	ε_5	ε_6	ATA_1	ATA_2	ATA ₃	ATA_4	ATA_5	ATA ₆
L_{∞}	1	2	4	8	16	32	91.0	87.8	81.6	71.3	46.5	23.1
L_2	40	80	160	320	640	2560	90.1	86.4	79.6	67.3	49.9	17.3
L_1	195	390	780	1560	6240	24960	92.2	90.0	83.2	73.8	47.4	35.3
JPEG	0.03125	0.0625	0.125	0.25	0.5	1	89.7	87.0	83.1	78.6	69.7	35.4
Elastic	0.125	0.25	0.5	1	2	8	87.4	81.3	72.1	58.2	45.4	27.8

	UAR_{L_∞}	UAR_{L_1}
$L_{\infty} \varepsilon = 2, L_1 \varepsilon = 76500$	48	66
$L_{\infty} \varepsilon = 4, L_1 \varepsilon = 153000$	51	72
$L_{\infty} \varepsilon = 8, L_1 \varepsilon = 306000$	44	62
$L_1 \varepsilon = 76500$	50	70
$L_1 \varepsilon = 153000$	54	81
$L_1 \varepsilon = 306000$	59	87

Table 6: UAR scores for L_1 -trained models and (L_{∞}, L_1) -jointly trained models. At each distortion size, L_1 -training performs better than joint training.

Table 7: UAR scores for L_{∞} - and Elastic-trained models and (L_{∞} , Elastic)-jointly trained models. No jointly trained model matches a Elastic-trained model on UAR vs. Elastic.

	UAR_{L_∞}	UAR _{Elastic}
$L_{\infty} \varepsilon = 4$, Elastic $\varepsilon = 2$	68	63
$L_{\infty} \varepsilon = 8$, Elastic $\varepsilon = 4$	35	65
$L_{\infty} \varepsilon = 16$, Elastic $\varepsilon = 8$	69	43
Elastic $\varepsilon = 2$	37	68
Elastic $\varepsilon = 4$	36	81
Elastic $\varepsilon = 8$	31	91

 $(L_{\infty}, \text{Elastic})$ also performs poorly, never matching the UAR score of training against Elastic at moderate distortion size ($\varepsilon = 2$).

D THE FRANK-WOLFE ALGORITHM

We chose to use the Frank-Wolfe algorithm for optimizing the L_1 attack, as Projected Gradient Descent would require projecting onto a truncated L_1 ball, which is a complicated operation. In contrast, Frank-Wolfe only requires optimizing linear functions $g^{\top}x$ over a truncated L_1 ball; this can be done by sorting coordinates by the magnitude of g and moving the top k coordinates to the boundary of their range (with k chosen by binary search). This is detailed in Algorithm 1.

E FULL EVALUATION RESULTS

E.1 FULL EVALUATION RESULTS AND ANALYSIS FOR IMAGENET-100

We show the full results of all adversarial attacks against all adversarial defenses for ImageNet-100 in Figure 6. These results also include L_1 -JPEG and L_2 -JPEG attacks, which are modifications of the JPEG attack applying L_p -constraints in the compressed JPEG space instead of L_∞ constraints. Full UAR scores are provided for ImageNet-100 in Figure 7.

E.2 FULL EVALUATION RESULTS AND ANALYSIS FOR CIFAR-10

We show the results of adversarial attacks and defenses for CIFAR-10 in Figure 8. We experienced difficulty training the L_2 and L_1 attacks at distortion sizes greater than those shown and have omitted those runs, which we believe may be related to the small size of CIFAR-10 images. Full UAR values for CIFAR-10 are shown in Figure 9.

F ROBUSTNESS OF OUR RESULTS

F.1 REPLICATION

We replicated our results for the first three rows of Figure 6 with different random seeds to see the variation in our results. As shown in Figure 10, deviations in results are minor.

						Adversarial accuracy					
- 1.0			0.8		. 0.6		- 0.4		0.2	0.0	
0 0	000000	0 0 0 0 0 0	н о о о о о о	00000	H H O O O O O O O O O O O O O O O O O O	н н о о о о о о о о о о о о	111000			0 0 0 0 0 0 0 m a 0	
0	000000	0 0 0 0 0	0 0 0 0 0 0 0	0 0 0 0 0 0 0	00000000		0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	000000000000000000000000000000000000000			
000	2 2 0 0 0	2 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	3 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0				0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0				
1 0	6 18 19 10 18 19 10 10 10 10 10 10 10 10 10 10 10 10 10	0 3 0 8 11 1 7 22 3 8 35 1	4 2 0 2 2 0 4 1 0 0 4 0 0 4 0 0 2 0 2 0 2 0 2 0 2 0 2 0	1 1 0 3 1 0 3 2 0 3 2 0 8 1 2 8 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	9 1 0 9 1 0 9 1 0 9 3 0 9 3 0 8 1 3 2 5 1 3 2 5 1 3 2 5 1 3 2 5 1 3 2 5 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1	0 3 0 3 16 2 5 34 1 7 36 1 4 36 1 2 38 2	2 1 0 7 4 0 9 2 0 9 8 1 9 8 1 2 9 9 8 1 2 9 9 8 9 8 9 8 9 3 5 3 9 2 9 3 2 2 9 3 2 2	1 9 1 418 7 7 301 7 2 27 10 2 25 6 2 25 125 125 125 125 125 125 125 125 125	3 4 3 11 1 7 4 4 1 1 7 4 4 1 1 7 4 4 1 1 7 7 6 4 5 1 1 1 7 1 7 1 7 1 7 1 7 1 7 1 7 1 7 1	
1 32 7	0 68 29 9 73 45 9 74 56 1 72 65 8 66 59 8 66 59 8 60 56	8 60 20 9 67 3 8 72 46 5 72 5 0 68 58 0 58 56	3 50 1 5 5 9 2 7 64 30 6 7 9 8 66 4 9 0 55 4 9 0 56 4 1 0 56 30 0 5 0 5 0 1	2 44 1 5 51 11 5 59 20 7 59 20 64 29 1 67 48 1 67 48 1 67 48 1 67 48	040 2 5214 55214 55214 75920 75933 76733 76733 76733 76633 96653	L 40 E 5 5 2 14 5 5 2 14 5 5 2 14 5 6 5 3 5 6 3 3 5 6 3 4 5 5 6 3 4 5 5 6 4 1 6 7 5 6 7 6 7 6 7 6 7 6 7 6 7 6 7 6 7 7 6 7 7 7 6 7 7 7 7	5 52 20 5 61 30 5 68 4: 5 69 5: 5 63 5: 5 64 5: 5 66 5	7 47 11 0 57 16 1 65 2 1 65 2 0 72 44 5 69 56 6 56 6 56 7 65 66 7 65 66 7 65 67 7 44 40	714: 7554 7455 7455 7455 57055 57055 57055 57055 57055 57055 57155 57155 57155 57155 57155	0 16 10 10 10 10 10 10 10 10 10 10 10 10 10	
õ		~~~~~~			*******						
0	000000	0000000	0000000		000000000		0000000		0 0 0 0 0 0 0 m M		
0	0 0 0 0 0 N	00000	0000000		000000000		0000	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0		
0	0 0 0 1 4 1 0 0 0 1	M H O O O O	н о о о о о о о л о о о о о о	4 5 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0				0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	1 0 0 1 7 0 1 7 0 1 7 0 1 7 0 1 7 0 1 57 4 1 57 5 1		
-	8 4 9 12 5 37 7 64 1 9 63 4	5 3 9 6 9 6 9 6 9 6 9 6 9 6 9 6 9 6 9 6 9	9 2 7 3 7 3 7 3 7 3 0 12 1 1 2 3 1 2 3 5 3 5 5 5 3 5 5 5 3 5	5 1 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2	+ 1 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2	+ 1 1 2 1 2 1 2 1 2 1 2 2 2 4 2 1 3 2 2 4 2 1 3 2 2 4 2 1 3 1 3	0 2 1 5 2 14 2 2 14 9 32 9 32 9 32 1 2 2 2 9 1 2 2 2 2	8 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1	6 22 9 56 1 0 72 4 0 80 6 7 73 6 5 70 6 5 70 6 3 67 6 2 64 5 2 64 5 2 64 5 2 64 5 9 61 5	7 2 3 3 5 3 5 3 5 5 3 5 5 5 5 5 5 5 5 5 5	
12	74 2 81 5 82 7 79 7 79 7 74 7 70 6	62 1 77 3 81 6 81 7 81 7 76 7 68 6	41 561 775 776 696	34 511 692 785 787 787 787	22 - 22 - 33 - 33 - 33 - 33 - 33 - 33 -	24 34 48 48 48 74 78 78 78 77 77 75 77 75 72 63 20 73 75 77 75 77 75 70 75 70 75 70 75 70 75 70 75 70 75 70 70 70 70 70 70 70 70 70 70 70 70 70	391 562 714 765 755 7155 7155 7155	19 25 31 40 52 1 52 1 52 1 52 54 5 54 5 54 5 54 5 5	82 6 84 7 83 8 83 8 83 8 81 7 81 7 73 7 78 7 78 7 78 7 78 7 78 7 78 7	21 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	
00	0 0 0 0 0 0 0 0 0 0 1 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0		ооооооооооооооооооооооооооооооооооооо	
0	000000	о о о о о о о о о о о о о о						0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0			
0	000000	000044	0000000	0000000	00000000	000000000000	0000000	2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	00000000000		
0	000000	001100	0000000	000000	0000000		1 1 1 1 1 0 0 0	3 3 451 921 921 9715 9715 6726 6726 55556 55556 7575	4 H O O O O O H 4	2 2 0 6 10 0 14 0 0 14 0 0 14 0 0 0 14 0 0 0 14 0 0 0 0	
4 15	447 3471 043 333 525 529	3 46 1 4 47 1 2 45 1 7 40 9 30 9 23	833 936 936 039 837 837 132 024	728 933 933 933 933 933 933 933 933 933 93	424 727 833 935 634 129 029	4 24 8 30 9 35 9 35 9 35 9 35 9 35 9 35 7 17 7 17	035 242 247 350 1 350 1 342 1 038 1 038	3 69 3 5 78 5 6 82 7 6 84 7 6 84 7 9 78 7 9 78 7 9 78 7 6 57 5 6 57 5 1 50 5	4491 5521 2491 144 043 839 840 9471 9471 9502	4 5 1 1 5 6 5 2 2 5 6 0 2 2 6 5 5 5 3 3 7 5 5 6 3 7 5 5 6 3 7 5 6 5 5 3 7 5 6 6 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5	
0	000000	0 0 0 0 0 0 V V V 0 0 4	0000000	0000000	000000000	00000000000000000000000000000000000000		000000000000000000000000000000000000000			
0	лнооо 2 1000	0 0 0 0 4 1	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0		0000000000000	00000000000000	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0		0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	, , , , , , , , , , , , , , , , , , ,	
0	6 3 2 7 2 15 6 31 6 31 2 4 1	8 1 6 9 6 23 8 38 8 38 3 47 1	9 1 8 1 1 3 2 6 7 12 2 21 1 33	4 0 0 1 8 1 8 1 2 7 2 7 7 11	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	5 2 915 851 1733 9787 9787 2717 2717	00000000000000000000000000000000000000	7 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2		
9 47	4753 4785 3796 9766 3716	3712 3764 3795 3795 1786 6746 8666	2 66 1 3 71 2 2 75 4 2 77 5 9 75 5 7 7 3 6 0 68 6	3 65 1 3 73 2 3 73 2 2 75 4 0 76 5 8 75 5 7 74 5	2591 2651 3691 3733 3733 2764 2765 7756	1581 2671 2671 2753 2753 2753 2753 2755 2753 2755 2755	5784 6836 6836 4837 4838 4838 1807 8767 8767 4737	154 2581 2611 2611 2611 4601 5491 638 6381 126	3 73 3 3 74 3 2 73 3 3 73 3 2 73 3 2 73 3 3 74 3 2 73 3 2 77 3 3 77 3 2 77 3 3 77 3 2 77 3 3 77 4 77 4		
0	000000	0 0 0 0 0 m	0000000	00004	<u>7 7 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8</u>	0000047000	0000000	© © © © © © © © © © © © © © © © © © ©	000000000000000000000000000000000000000		
0		0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 1 48 4 1 48 4 2 55 4	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0				0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	
0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0		4 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 1 0 3 2 2 28 3 69 6 3 69 6	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0					
0	001004	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 4 2 2 4 5 2 2 4 5 2 2 4 5 2 2 4 5 2 2 4 5 2 2 4 5 2 5 1 2 4 5 2 5 1 2 4 5 5 1 2 4 5 5 1 2 4 5 5 1 2 4 5 5 2 5 1 2 4 5 5 2 5 1 2 4 5 5 2 5 1 2 4 5 5 2 5 1 2 4 5 5 2 5 1 2 4 5 5 2 5 1 2 4 5 5 5 5 1 2 4 5 5 5 5 1 2 4 5 5 5 5 1 2 4 5 5 5 5 1 2 4 5 5 5 5 1 2 4 5 5 5 5 5 1 2 4 5 5 5 5 1 2 4 5 5 5 5 5 1 2 4 5 5 5 5 1 2 4 5 5 5 5 1 2 4 5 5 5 5 1 2 4 5 5 5 5 1 2 4 5 5 5 5 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	7 0 7 14 8 43 1 6 69 5 8 76 7 7 75 7	1 0 8 0 8 3 0 24 8 73 6 8 73 6 6 75 7 7 75 7	2 0 6 2 6 34 6 34 9 75 6 9 75 7 7 75 7 7 75 7 2 71 6	0000000000000			• • • • • • • • • • • • • • • • • • •	
0	8 5 616 831 4331 2251 6231	5 2 5 7 9 24 3 56 2 3 68 5 8 66 6	1 1 8 5 118 2 47 1 8 71 4 8 71 4 0 69 6	7 36 6 57 2 6 57 2 1 72 4 2 78 6 0 79 7 9 79 7 8 77 7	417 041 9642 3817 3817 3817 7777 7777	5481 0632 2754 2754 3817 3817 3827 3827 3827 3827 7777 7777 7777 777	4 1 0 0 0 1 4 1 0 0 0 0 1 1 4 1 0 0 0 0	000000000000			
0 20	6 60 2 9 69 4 9 73 5 4 67 5 4 67 5 6 47 3 6 47 3	3 48 1 8 66 3 2 75 5 2 75 5 1 80 7 7 76 7 8 68 6	1 44 1 8 63 2 1 75 5 3 80 7 3 80 7 3 80 7 8 80 7 1 71 7 1 71 7	380658327582758386608080808080808080808080808080808080	3765 4817 5837 5848 5848 4838 4838 4838 7777 7777 8787	5837 5848 5848 5848 5848 6858 6858 848 4848 4	233 - 7431 7472 8522 8522 8522 8522 8522 216 -	918 918 97 55 77 73 73 73 73	0 8 5 3 2 3 7 4 9 8 2 10 110		
0	000000	000000		000000	0 0 0 0 0 0 0 0 m	000000000000000000000000000000000000000	00000000	4 4 M V N H	000000000000000000000000000000000000000		
0			0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	• • • • • • • • • • • • • • • • • • •	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0				
0	7 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	2 0 9 2 4 25 7 62	1 0 3 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	5 3 4 18 1 51 1 51 2 73 1 0 77 5 9 76 6 8 76 6	3 0 6 3 6 3 6 24 6 24 2 61 2 762 7 7 7 7 7 7 7 7 7 7 6 7	0 1 6 1 6 2 7 8 2 7 8 2 7 8 2 7 8 2 7 8 2 7 8 2 7 8 2 7 8 7 9 6 7 7 9 7 8 7 9 7 8 7 1 6 7 7 6 7 7 7 6 7 1 1 1 1	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0				
1	14 66 1 14 77 3 13 79 5 18 70 4 10 59 2 1 34	1 48 3 73 1 14 80 4 12 81 7 16 76 7 8 68 6	1 22 9 48 2 68 2 2 76 4 2 76 4 1 75 6	36 82 36 84 7 36 85 8 34 84 8 34 84 8 31 81 8 31 81 8 31 81 8 37 9 7 9 7 9 7	55742 568256847 568475858847 55858585858585858585858585858585858585	5 79 4 56 82 6 5 84 7 5 85 84 7 5 85 84 7 5 85 8 5 85 8 5 85 8 5 8 8 5 8 8 5 8 8 5 8 8 5 8 8 5 8 8 7 8 7 8	8 11 1 25 6 41 6 45 88 31 16 12 16 12	0 0 0 0 0 0 0 1 7 1 0 0 0 0 0 0 0 0 0 0	0199700000	1111100104-0104-01011111	
70 2	868 858 848 797 737 737 737 737	85 8 85 8 85 8 85 8 85 8 82 8 77 7 77 7 77 7	84 7 84 7 84 8 84 8 84 8 80 7 7 8 7 8 7 8 7 8 7 8	868 878 868 868 818 818 807 787	86 86 86 86 86 86 86 86 86 86 86 86 86 8	86 88 86 86 86 86 86 86 86 86 86 86 86 8	82 9 84 7 83 7 82 7 82 7 70 4 70 4	74 2 57 1 57 1 57 1 57 1 57 1 57 1 57 1 57 1	64 42 33 31 38 38 38 38 58 58 59 59 59 59 59 59 59 59 59 59		
0 0	0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 4 10 0 0 0 0 0 1	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	00000000000000000000000000000000000000	
0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 m <u>5</u>	2 0 7 14 9 64 7 73 5 62	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0				
1 0	11 13 11 48 22 69 77 59 12 8	1 2 5 22 1 60 11 76 1 76 16 75 8 67 8 67	8 0 8 14 5 4 14 5 45 7 4 60 7 4 60 7 4 60 7 4 60	33 54 84 75 85 83 84 83 84 83 84 83 180 9 79 8 77 8 77	1 15 22 49 25 74 25 82 26 83 26 83 26 83 26 83 26 83 21 81 7 77	916 838 8477 8477 8477 8477 8477 9382 979 979 979 979	3 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	сососососососососососососососососососо	
20	84 7 84 8 84 8 79 7 73 6 73 6 73 6 73 6	81 9 84 7 84 8 84 8 84 8 82 8 77 7 77 7 77 7 68 6	67 1 76 2 81 6 82 7 79 7 79 7 71 6	868 856 856 858 848 818 818 807 787	85 86 86 86 86 84 84 84 84 84 84 84 84 84 84 84 84 84	82 82 82 82 82 82 82 82 82 82 82 82 82 8	64 1 76 80 80 80 79 70 53 27	23 9 6 4 w 4 2 6 4 w 4 2 6 4 m 4	35 16 8 8 8 8 8 8 8 8 8 8 14 117 17 22 21	27 27 27 27 27 27 27 27 27 27 27 27 27 2	
0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	5 m 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	00000000000000000000000000000000000000	0 0 0 0 0 0 0 0 0 0 0 <u>0 0 0 0</u>				
0 0	5 0 5 0 2 6 2 6 2 0 2 0 2 0	5 0 14 9 55 32 55 32 71 57 5	13 0 32 3 34 12 75 64 1 75 69 4 70 67	1 0 2 0 2 0 8 2 8 2 18 17	1 0 1 0 3 0 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 6 30	1 0 1 0 2 2 0	2 0 9 1 7 1 1 0 2 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0		
5	56 35 72 52 75 61 54 49 54 49 13 24 13 24 21 8	5936 7657 3071 3077 7675 38680	77 53 31 68 32 77 32 77 33 81 33 79 78 78 72 71	50 14 50 24 57 37 73 53 74 62 75 68 75 68	13 10 52 16 54 28 73 46 2 77 62 2 77 70 77 70 75 71 76 71	5517 5526 7138 7138 7138 7138 7138 7138 7138 7170 7170 7170 7170 7170 7170 7170 717	18 14 18 25 13 35 19 25 19 25 19 25 14 4	9 2 2 2 2 2 4 4 4 5 5 5 5 5 2 7 4 4 5 5 5 5 2 7 7 9 4 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7	39 11 22 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5	2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	
19	80 81 72 58 37 37	80 82 82 83 83 83 83 83 83 83 83 83 83 83 83 83	83 84 8 84 8 84 8 84 8 84 8 84 8 79 1 79 1 79 1	75 77 80 80 77 77 77	72 75 79 79 81 81 81 81 77	77 79 81 82 82 82 82 81 77 77 72 72	72 6 77 6 75 6 66 6 47 3 31	62 54 54 54 54 54 55 25 25 25 25 25 22 22 22 22 22 22	66 50 50 45 45 41 41 41 41 36 36		
0 0	0 0 2 0 3 0 0 0	0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 4 0 7	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0			00000000000000000000000000000000000000	
000	18 2 71 18 79 48 73 50 73 50 83 30 83 5	34 1 77 40 30 69 76 73 58 67	14 0 32 1 55 6 57 20 59 35 71 46 71 46 71 46	7 0 19 0 56 13 74 41 74 50 74 50 74 50	2 0 2 0 0 2 0 0 2 0 0 2 0 0 2 0 0 2 0 0 2 0 0 2 0 0 2 0 2 0 2 0 2 0 2 0 0 2 0 0 2 0	5 0 25 1 25 1 35 1 35 1 35 34 35 34 35 34 35 34 35 34 35 34 35 55 35 34 35 55 35 55 35 55 35 55 35 55 35 55 35 55 35 55 35 55 36 1 37 55 36 1 37 55 36 1 37 5 36 1 37 5 37 5 37 5 37 5 37 5 37 5 37 5 37 5	4 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0		0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	ноооооооооооооооооооооооооооооооооооо	
57 11	86 81 85 83 84 83 79 78 73 71 6 59 60	85 78 85 82 84 83 82 82 82 82 77 77 58 68 6	82 64 84 74 84 74 84 80 83 81 83 81 83 81 78 77 78 77 72 71 6	83 58 84 71 84 79 84 79 83 81 83 81 79 78 79 78	80 44 83 59 84 74 84 74 84 74 84 80 83 82 81 80 78 77 77 77	8151 8363 8474 8481 8480 8480 7979 7979 7375 7375 7375 7375	79 44 82 58 82 67 81 71 78 67 73 54 73 54 73 54	69 16 68 17 59 12 50 9 50 9 24 2 24 2 24 2 9 1 9 1 9 1 0 2	75 33 64 17 51 10 53 10 57 13 57 13 58 15 59 15 50 16 56 15	68 24 68 24 7 25 6 6 23 23 24 7 7 25 6 6 7 26 23 24 7 7 25 6 6 7 25 6 6 7 25 6 6 7 25 6 6 7 25 6 6 7 25 6 6 7 25 6 6 7 25 6 6 7 25 6 6 7 25 6 6 7 25 6 6 7 25 6 6 7 25 6 6 7 25 6 6 7 25 6 6 7 25 7 25	
•	0000-0	00007				<u> </u>	0000000			2° % -	
0 0	0 0 22 0 59 6 67 34 52 40	0 0 0 1 4 0 56 6 51 28	2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0000 10000 10000	84000000000000000000000000000000000000	20000000000000000000000000000000000000	0 0 0 0 0 0 0 0 0 0 0 0 0 0	<u> </u>	<u> </u>	<u>م م</u> م م م م م م م م م م م م م م م م م	
1 0	70 14 81 50 82 74 79 76 74 73 70 69	53 3 74 22 81 56 81 74 76 74 58 67	24 1 41 3 62 11 71 28 72 43 72 53 72 53 69 59	28 1 47 3 68 14 77 42 77 42 77 68 77 68	12 0 26 1 50 4 70 15 79 67 76 71	15 0 27 1 47 3 5612 7534 77534 77534 77571 7771 7771 7771 7165	14 0 25 1 49 7 49 7 39 5 22 2	00000000000000000000000000000000000000	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	v v v v v v v v v v v v v v v v v v v	
25	84 85 83 79 74 71	82 84 82 82 82 77 68	71 78 81 82 77 77 71	75 88 88 79 77	64 75 81 83 83 81 73 77	66 79 83 82 79 82 77 73 75 73	63 77 78 69 69 58	04 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	22 24 23 24 23 24 23 24 23 24 23 24 23 24 22 24 22 24 22 24 22 24 22 24 22 24 22 22	8 8 7 5 2 5 3 3 3 9 4 6	
g -8	2 - 85 2 - 85 8 - 84 8 - 84 5 - 75 2 - 75	0 - 87 0 - 85 0	4 - 86 5 - 86 0 - 84 0 - 84 0 - 79 0 - 79	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	2 87 86 88 86 86 88 86 8	8 80 80 80 80 80 80 80 80 80 80 80 80 80	5 - 87 5 - 87 5 - 87 5 - 84 5 - 84 5 - 78 7 8 7 8 7 8 7 8 7 8 7 8 7 8 7 8 7 8	8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8	5 - 85 5 - 85 5 - 85 5 - 85 8 - 75 8	2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	
rainin	6 = 10 6 = 10 7 = 10	e = 15 e = 30 e = 60 = 120 = 120 = 2400	562.4 1912 8250. 7650. 15300(1200(1200(0312 0.062 = 0.12 = 0.2 ε = 0.2 G ε = 2 G ε = 2	6 8 = 6 8 = 6 8 = 8 = 1 8 = 1	f = 12 f = 25 f = 204; f = 204	$\varepsilon = 0.2$ $\varepsilon = 0.2$ ic $\varepsilon = 1$ ic $\varepsilon = 1$ ic $\varepsilon = 1$	<pre>c = 12 c = 25 c = 25 c = 51 c = 102 c = 102 c = 102 c = 103 c = 1</pre>	$\varepsilon = 10^{-2}$ $\varepsilon = 10^{-2}$ $\varepsilon = 2^{-2}$ $\varepsilon = 10^{-2}$ $\varepsilon = 2^{-10}$ $\varepsilon = 2^{-10}$ ε	0.0312 0.062 = 0.12 = 0.2 ≤ = 0. w ε = 1 w ε = 1 v ε = 1	
rmal t	 	L2 L2 L2 L2 L2 L2 L2 E L2 E L2 E	$[\epsilon = 9 \\ L_1 \epsilon = 3 \\ L_1 \epsilon = 1 \\ 1 \epsilon = 6 \\ 1 \epsilon = 6 $	5 ∈ = C +JPEG € = L.«-JPEG €	L2-JPE L2-JPE 2-JPEG 2-JPEG 2-JPEG - JPEG 4	JPEG - JPEG. - JPEG. - PEG.ε PEG.ε - PEG.ε - 26.ε = 1 - 26.ε = 1	astic ٤ Elast Elast Elast Elast Elast Elast	Fog Fog Fog Fog Fog Fog Fog Fog	abor ∉ abor ∉ Gabor Gabor jabor ∉ jabor ∉ iabor ∉ bor €	k ε = C w ε = C Snov ε Sno Sno Sno Sno	
No			2 2 777			17 17 17 17 17 17 17 17 17 17 17 17 17 1				5 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	

Algorithm 1 Pseudocode for the Frank-Wolfe algorithm for the L_1 attack.

1: Input: function f, initial input $x \in [0, 1]^d$, L_1 radius ρ , number of steps T. 2: **Output:** approximate maximizer \bar{x} of f over the truncated L_1 ball $B_1(\rho; x) \cap [0, 1]^d$ centered at x. 3: 4: $x^{(0)} \leftarrow \text{RandomInit}(x)$ {Random initialization} 5: for t = 1, ..., T do 6: $g \leftarrow \nabla f(x^{(t-1)})$ {Obtain gradient} for k = 1, ..., d do 7: $s_k \leftarrow$ index of the coordinate of g by with k^{th} largest norm 8: 9: end for 10: $S_k \leftarrow \{s_1, \ldots, s_k\}.$ 11: {Compute move to boundary of [0, 1] for each coordinate.} 12: 13: for i = 1, ..., d do 14: if $g_i > 0$ then 15: $b_i \leftarrow 1 - x_i$ 16: else 17: $b_i \leftarrow -x_i$ end if 18: end for 19: 20: $M_k \leftarrow \sum_{i \in S_k} |b_i|$ {Compute L_1 -perturbation of moving k largest coordinates.} $k^* \leftarrow \max\{k \mid M_k \le \rho\}$ {Choose largest k satisfying L_1 constraint.} 21: 22: {Compute \hat{x} maximizing $g^{\top}x$ over the L_1 ball.} 23: for i = 1, ..., d do 24: if $i \in S_{k^*}$ then 25: 26: $\hat{x}_i \leftarrow x_i + b_i$ else if $i = s_{k^*+1}$ then 27: $\hat{x}_i \leftarrow x_i + (\rho - M_{k^*}) \operatorname{sign}(g_i)$ 28: 29: else 30: $\hat{x}_i \leftarrow x_i$ end if 31: 32: end for $x^{(t)} \leftarrow (1 - \frac{1}{t})x^{(t-1)} + \frac{1}{t}\hat{x}$ {Average \hat{x} with previous iterates} 33: 34: **end for** 35: $\bar{x} \leftarrow x^{(T)}$

F.2 CONVERGENCE

We replicated the results in Figure 6 with 50 instead of 200 steps to see how the results changed based on the number of steps in the attack. As shown in Figure 11, the deviations are minor.



Figure 7: UAR scores for adv. trained defenses (rows) against distortion types (columns) for ImageNet-100.

Ö.
÷
ż
\mathbf{A}
H
Ú
ц
0
б
Ŀ
G
ğ
ğ
Ę
ĕ
.Ħ
ra
Y t
Ξ,
ia
ar
SI
Š
ġ
10
US
Ē
õõ
) a
É
E
Ę
3
$\overline{}$
Č
ta
at
al
·Ξ
sa
G
<u>Z</u>
б
Ŀ
ž
<u>୍</u> ରୁ '
LL 2
5
₹C
4
÷.
e
E.
:50
щ





Figure 9: UAR scores on CIFAR-10. Displayed UAR scores are multiplied by 100 for clarity.



Figure 10: Replica of the first three block rows of Figure 6 with different random seeds. Deviations in results are minor.

_						Adversarial accuracy		<u>.</u>		
1.0			9.0 -		- 0.6		- 0.4		- 0.2	0.0
0	00100	11000	н н н о о о о	н н о н о о о о		111110000000		000000000000	000000000000	0 - 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0	2 1 1 0 0 0	210000	н о о о о о о о	0000000		000000000000000000000000000000000000000	5 7 7 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	000000000000000000000000000000000000000	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	
0 0	1 0 2 2 0 2 2 0 1 1 3 2 1 3 1 3	0 0 0 5 1 0 5 8 1	0 0 0 0 1 0 0 2 1 0 2 2 0 2 3	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	1 0 2 1 0 4 6 2 1 0 2 1 0 2 1 0 2 1 0 2 1 0 2 2 1 2 2 2 2	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	2 0 2 0 8 2 0 8 2 0 8 2 0 2 0 2 0 2 0 2	1 0 9 1 9 1 1 44
10 1	33 7 47 14 59 23 65 38 61 44 56 41	24 4 35 7 50 14 60 26 61 38 55 42	18 2 23 3 32 6 39 9 46 15 49 21 46 25	113 2 118 3 223 4 324 6 48 15 53 21 49 20	113 2 114 2 119 2 255 4 49 15 57 30	10 2 13 2 16 2 23 4 53 25 53 25 53 25 53 25 53 25 53 25 53 25 53 25	26 5 34 11 49 21 54 29 55 35 54 38 54 38	17 2 23 3 30 6 39 9 50 20 58 44 11 58 44 11 58 44 11 54 44 30	44 12 58 23 61 33 62 39 57 35 55 31 56 32 56 32 56 32 56 32 56 32 56 32 56 32 56 32	37 6 8 65 18 8 8 55 18 8 8 7 6 8 8 7 6 8 8 7 6 7 6 6 4 7 7 7 7 7 7 7 7 7 6 6 7 8 8 7 6 7 8 8 7 6 7 8 8 7 6 7 8 8 7 7 6 8 8 7 7 7 8 8 7 7 6 8 8 7 7 7 8 8 7 7 6 8 8 7 7 7 8 8 7 7 6 8 8 7 7 6 8 8 7 7 7 8 8 7 7 7 8 8 7 7 7 8 8 7 7 7 8 8 7 7 8 7 7 7 8 8 7 7 8 7 7 7 8 8 7 7 8 7 7 7 8 8 7 7 8 7 7 7 7
64 37	80 67 80 73 79 76 75 73 75 73 68 66 63 61	78 61 79 69 76 73 76 72 69 68 69 68	73 53 75 58 76 65 76 67 72 66 69 64 69 64	73 47 74 51 76 58 76 65 74 68 74 68 71 68 71 68	71 445 74 45 75 55 76 60 77 66 77 66 77 68 70 67 70 67	70 44 73 49 74 53 77 65 77 66 77 66 76 69 77 69 70 68 64 60 64 60	74 55 75 62 76 68 76 69 72 68 66 63 63 60	7649 7959 8164 8168 817570 7570 7570 6665 6665 5351 4745	80 71 80 75 79 75 77 71 77 71 77 71 76 70 76 70 75 71 75 71	85 75 85 88 85 88 85 84 85 84 75 75 75 75
0 0	1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	и и и и и и и и и и и и и и и и и и и	2 12 2 12 2 12 2 12 2 12	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	⁴ ⁴ ⁴ ⁴ ⁴ ⁴ ⁴ ⁴ ⁴ ⁴
0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	о о о о о о о о 	л о о о о о о л л о о о о о	0 T 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	ннн н о о о о н н н н о о о о	0 0 0 0 0 0 0 0 0 8 6 8 5 30 2 35 30 2	0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	500 500 500 500 500 500 500 500
0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	н 0 0 0 0 0 л н 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	2 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 23 6 36 18 36 18 28 12 24 12 24 12	1 0 4 1 13 1 48 5 48 5 78 42 78 42 76 78 76 78 76 78 75 77 75 77 75 77	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
2 1	2 16 2 2 16 2 7 41 5 8 65 18 8 69 46 9 64 52	1 4 1 5 8 1 7 20 2 5 45 6 461 19 7 61 31	8 3 1 2 4 1 8 7 1 8 7 1 5 16 2 8 27 4 8 27 4 7 39 8 7 39 8 7 39 8	2 1 2 1 2 3 1 1 2 3 1 1 2 3 1 1 2 3 1 2 3 1 1 2 3 1 1 1 1	5 2 0 5 3 1 5 3 1 5 3 1 5 3 1 5 3 1 3 5 6 5 1 5 2 5 0 1 5 2	2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0	5 4 1 5 7 2 8 18 4 9 29 7 9 29 7 8 38 13 8 38 13 7 37 18	<pre>2 1 2 2 1 3 2 1 3 2 0 3 2</pre>	1 32 7 1 65 25 2 79 54 2 81 78 2 81 79 9 78 79 9 78 79 9 78 79 7 76 76 7 76 76 7 76 76 7 76 76 7 76 76 7 70 69	1 3 1 3 2 2 3 3 3 3 1 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
18	75 3 81 6 82 7 79 7 74 7 70 6	65 2 78 4 81 6 81 7 76 7 76 7	47 1 61 2 73 3 78 5 78 5 78 6 75 6 75 6	41 55 551 55 78 55 78 77 78 77	28 6 54 1 70 3 80 6 80 6 77 7 76 7	31 4111 5411 5411 7542 79 79 78 77 77 77 72 57 72 57	45 1 73 45 75 5 72 6 72 6 72 6 72 6 72 6	24 32 32 54 54 50 52 55 54 54 54 54 54 54 54 54 54 54 54 54	837 848 838 838 838 818 818 797 797 7577	55 0 1 55 0 2 56 2 56 2 50 0 50
0 0	0 0 1 0 0 1 1 1 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	н о н о о о о о о о о о о о о	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1		ооооооооооооооооооооооооооооооооооооо
0 0	о о о о о о о о о о о о п п п п п п п п	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 H 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0	н о о о о о о о о о о о о о о о	л н о о о о о о о о о о о	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 1 1 8 2 58 38 58 38 66 64 66 64 49 47	н о о о о о о о о о о о о о о о о о о о	ооонононн - ₄ , , , , , , , , , , , , , , , , , , ,
0 0	0	000000	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	H 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0		0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	2 1 0 6 0 5 32 6 5 32 6 5 60 30 5 7 3 68 5 7 3 68 5 66 66 5 66 66 5 5 5 5 5 5 5 5 5 5 5 5	нооооооооо инооооооно	0 0 0 1 1 1 1 1 0 0 4 0 0 0 0 0 0 0 0 0
1 <u></u> 6 (0 23 6 19 0 15 9 6 3 9	619 520 516 512 76 76	8 10 9 12 8 12 8 12 9 9 9 9 8 12 8 12 8 12 8 12 8 12 8 12 8 12 8 12	8 11 1 13 9 13 6 12 7 10 2 10	0 12 3811 0 13 114 114 114 114 114 114 114 114 114	8 8 10 0 8 8 10 0 8 8 10 0 8 8 11 8 8 10 0 9 14 8 11 1 8 11 1 1 1 1 1 1 1 1 1 1 1 1	3 15 7 22 8 24 9 25 9 21 4 19	8 52 1 2 71 4 5 80 6 5 82 7 9 7 8 7 9 7 8 7 0 7 0 7 1 5 2 5 1 5 2 5 1 5 2 5	125 331 331 331 721 619 619 619 619 6521 5221 8311	65 32 89 39 89 39 89 45 1 89 45 1 86 45 1 14 45 2 16 43 2 16 44 16 16 16 16 16 16 16 16 16 16 16 16 16
703	78 6 76 5 76 5 78 5 58 2 58 2 58 3	77 765 755 694 613 512	74 6 74 6 72 6 69 6 50 3 50 3 50 3	75 5 75 5 73 6 73 6 73 6 73 6 73 6 73 6 73 6 73 6	75 75 75 75 75 75 75 75 75 75 75 75 75 7	75 66 67 75 66 67 75 75 75 75 75 75 75 75 75 7	77 76 75 75 75 75 75 75 75 75 75 75 75 76 75 76 76 76 77	85 7 87 8 86 8 86 8 86 8 86 8 86 8 70 7 70 7 70 7 70 7 70 7 81 8 81 8 81 8 81 8 81 8 81 8 81 8 81	71999999999999999999999999999999999999	78 78 78 78 78 78 78 78 78 78 78 78 78 7
0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	ооооооооооооооооооооооооооооооооооооо
0 0	9 4 0 3 10 0 2 18 1 6 34 4 5 44 12 3 46 16	2 2 0 5 5 0 8 12 0 7 24 2 8 39 5 3 49 12	2 1 0 3 4 0 3 8 0 8 14 0 2 3 1 1 35 4	8 1 0 2 1 0 3 4 0 3 10 0 9 14 1 7 13 1	4 1 0 3 1 0 3 2 0 4 4 0 6 1 1 1 4 2 1 2 2 1 2 2 1 2 2 1 2 2 1	2 1 0 6 0 0 2 1 0 2 2 0 0 2 8 0 9 14 1 8 17 1 8 17 1 4 14 1 4 14 1	7 4 0 9 19 0 8 52 4 1 74 30 9 78 67 6 76 75 2 72 70	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	7 3 0 1 4 0 2 1 5 0 2 2 7 0 8 1 1 0 8 1 0	
30 50	34 75 3 34 79 5 33 79 6 33 79 6 79 77 6 73 71 6 73 71 6 70 68 6	34 72 3 34 76 4 33 79 5 31 78 6 76 74 6 58 67 6	32 67 2 33 71 3 33 75 4 32 78 5 79 75 5 77 74 6 70 68 6	32 66 1 33 71 2 33 73 3 33 73 3 33 73 4 33 75 4 79 75 5 79 75 5 79 75 5	81 61 1 82 66 1 83 70 2 83 74 3 83 75 4 82 75 4 77 75 6 77 75 6	31 60 1 32 63 1 32 63 1 32 68 2 32 73 3 32 75 5 79 75 5 79 75 5 77 74 6 77 74 6 74 71 5 72 69 5	35 78 4 36 83 6 34 83 7 34 83 7 34 83 8 31 80 7 77 77 74 73 7	31 57 1 32 61 1 32 62 1 32 64 2 32 64 2 32 64 2 36 6 2 36 6 2 36 50 1 56 38 56 38 71 26 71 26	33 73 3 32 74 4 32 74 4 32 73 4 32 73 4 30 73 4 73 71 4 78 71 4 76 70 70 4 76 70 70 70 70 70 70 70 70 70 70 70 70 70	74541 74582 74582 74613 74613 73633 73633 54613 54615 54615 54615
0	000000	00000m	0000000	0000 0 8 8	5 3 3 1 1 0 0 0 0 1 8 1	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	00000000	000000000000000000000000000000000000000	000000000000000000000000000000000000000	· · · · · · · · · · · · · · · · · · ·
0 0	0 0 0 0 0 0 0 0 0 0 0 0	m 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	н 0 0 0 0 0 0 н 0 0 0 0 0 0	0 0 0 0 0 0 1 0 13 3 62 53 69 65	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	
0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 8 1 1 33 15	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	1 0 0 5 0 0 0 3 0 9 18 4 1 57 35 6 74 69 6 74 69	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	2 0 0 7 1 0 0 7 1 128 5 128 5 5 3 2 2 5 3 2 2 5 3 2 2 6 8 49 6 6 8 49 7 7 3 6 4 7 7 3 6 4 1 6 9 6 5 1 6 9 6 5	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	
1 0	101 224 3811 3916 3115 2715	4 0 4 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	3 0 1 0 1 0 1 0 1 0 0 1 0 0 0 0 0 0 0 0	16 13 32 30 74 51 2 79 70 4 79 76 7 78 77 7	27 5 15 15 15 15 15 15 15 15 15 15 15 15 1	56 18 58 35 55 4 1 30 70 4 32 79 7 30 79 7 7 77 7 7 77 7 7 77 7 7 7 77 7 7 7 77 7	2 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0		000000000000000000000000000000000000000	
32 9	64 36 1 71 52 2 74 62 3 70 59 3 58 46 3 51 41 3	55 23 69 42 1 76 61 3 80 74 6 76 74 6 68 68 6	53 20 68 40 1 81 74 59 2 81 74 59 2 79 78 7 78 77 7 71 70 6	81 71 4 83 78 83 81 7 83 81 7 83 82 7 83 82 7 9 7 9 7 9 7 9 7 9 7 9 7 9 7 8	78 61 882 74 982 74 983 80 983 80 984 82 983 83 83 83 83 83 83 83 83 83 83 83 83 8	84 76 84 81 8 85 83 8 85 83 8 85 84 8 88 88 8 88 88 8 79 79 7 77 77 77 77 75 75 73 73	44 17 52 26 56 32 1 59 38 1 50 30 1 34 18 34 18 23 8	29 7 25 5 13 2 14 2 7 1 7 1 7 1 7 1 7 1 8 1 4 1	16 3 6 1 6 1 6 1 16 4 20 5 20 5 18 5 19 5	22 5 24 5 25 6 27 4 27 4 27 4 27 4 27 4 27 4 27 4 27 4
20	79 79 79 79 79 79	75 79 77 68 77 68	74 80 83 83 83 78 78 78 78 78	85 85 85 85 84 84 80 85 80 82 78 81 80 81 78 81 80 81 80 81 80 81 80 82 80 85 80 85 80 85 80 85 80 85 80 85 85 85 85 85 85 85 85 85 85 85 85 85	83 83 83 83 83 83 83 83 83 83 83 83 83 8	85 85 85 85 85 85 85 85 85 85	0 70 70 70 70 70 70 70 70 70 70 70 70 70	56 56 56 56 56 56 56 56 56 56 56 56 56 5	27 27 27 29 29 29 29 29 29 29 29 29 29 29 29 29	
00	00000	× × × × × × × × × × × × × × × × × × ×	0 0 0 0 H 4 F	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0		0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	
0 0	13 0 38 1 56 6 46 8 29 4 2 2 3 2	3 0 20 0 74 28 74 58 67 63 3	1 0 5 0 5 1 0 5 1 0 5 1 0 64 20 69 45 67 56 1	57 4 74 21 81 55 82 73 2 80 77 5 79 76 6 78 77 7	26 0 59 4 77 27 83 77 28 83 77 3 80 79 6 77 76 7	43 2 63 8 63 8 85 59 82 79 5 82 79 5 79 78 7 77 76 7 75 74 6 75 74 6 75 74 6	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0			0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
28 2	84 68 84 77 84 77 83 80 78 71 70 58 60 35	81 51 84 73 84 73 84 80 82 81 77 76 68 68	72 25 79 49 82 69 82 77 79 77 71 75 71 75	8682 8684 8684 8685 8484 8180 8180 8080 7878	8576 8682 8684 8684 8585 8484 8181 8181 7777 7877	85 79 85 83 85 84 85 84 85 84 85 84 81 80 79 79 79 79 73 73	60 13 72 28 76 44 76 49 68 34 68 34 48 13 24 3	31 1 25 1 13 0 5 0 3 0 3 0 1 0 1 0 3 1	20 1 6 0 4 0 6 0 6 0 6 1 18 1 18 1 18 1 18 1 23 2 23 2 21 2	2 2 3 2 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3
2	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	85 85 85 85 90 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	84 84 72 84 84 84 84 84 84 84 84 84 84 84 84 84	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	83 83 71 83 83 83 84 83 84 83 84 84 85 83 84 84 84 84 84 84 84 84 84 84 84 84 84	2000 200 2000 2	004 m w 4 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	
0 0	000000	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0000448	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	2 2 6 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	000000000
0 0	16 0 49 2 7013 6015 39 5 11 1	3 0 25 0 61 4 76 37 75 64 1 75 64 1	1 0 3 0 19 1 48 5 63 20 69 47 66 54 1	57 3 76 20 83 60 83 77 2 80 79 6 79 77 7 78 76 7	17 0 51 2 75 19 82 57 81 76 7 77 76 7	19 0 65 10 65 10 78 39 81 66 81 66 82 76 79 78 79 78 79 78 73 72 6	1 0 1 0 1 0 1 0 1 0 0 0			
22 1	84 72 84 81 84 81 83 82 80 77 73 68 65 45	81 53 84 76 84 81 84 81 84 81 82 81 76 76 68 68	69 20 77 43 81 66 82 75 79 76 77 75 71 70	86 83 86 85 86 85 86 85 86 85 84 84 81 81 81 81 80 79 78 78	84 72 86 81 86 84 85 85 85 85 84 84 81 81 78 78 78 77	84 70 85 78 85 82 85 82 86 84 84 84 84 84 83 83 81 81 81 81 80 79 80 79 75 75 75 75 73 73	65 16 77 36 80 57 79 58 79 58 70 39 55 16 55 16 30 4	25 1 20 0 7 0 5 0 5 0 8 0 4 0 7 1	37 37 10 10 10	2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
0 0	000000	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	2 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 H 4 4 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	000000000000000000000000000000000000000	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0			
2 0 -	10 1 24 3 39 9 31 9 31 0 3 0	10 1 26 3 50 13 55 38 56 61 4	21 1 58 18 58 18 73 43 77 64 2 76 71 4 70 67 5	4 0 6 0 6 0 11 1 24 3 39 10 52 22 54 3	2 0 4 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	3 0 3 0 9 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	5 0 11 1 1 2 0 3 0 3 0 3 0 3 0 3 0 3 0 3 0 3 0 3 0	000000000000000	4 m 2 v H H H H O H	миииноон4-000000
44 13	70 43 74 57 76 64 65 53 46 27 24 9	71 43 77 61 80 73 80 77 76 75 68 68	79 61 81 72 83 78 83 81 83 81 83 81 80 79 79 78 72 71	58 25 65 33 71 44 74 56 75 65 75 69 75 69	53 19 60 25 67 37 75 52 77 64 78 70 75 72 75 72	63 27 68 36 68 36 77 57 77 57 77 72 77 720	58 26 64 37 68 46 66 46 54 32 54 32 35 16 35 16	40 11 37 9 33 7 33 7 33 7 33 7 33 6 30 6 12 1 12 1 12 3 12 3 12 3 14 4 14 5 14 5	51 22 42 15 34 11 30 9 32 9 22 9 224 7 224 6 224 6 23 6 23 6 23 6 19 4	41 14 3 31 14 3 31 14 3 31 14 6 2 6 2 6 2 6 2 6 2 7 5 7 5 6 2 6 2 6 2 17 6 6 2 17 6 6 2 17 6 7 5 7 5 7 5 7 5 7 5 7 5 7 5 7 5 7 5 7 5
2	81 81 81 81 81 81 81 81 81 81 81 81 81 8	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8	77 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	77 77 77 77 77 77 77 77 77 77	82 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	75 76 76 76 77 77 77 76 76 76 76 76 76 76	26 22 3 3 3 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	
0 0	7 0 0	1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	000000000000000000000000000000000000000	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	
13 0	82 50 83 71 2 83 79 5 83 79 5 83 79 5 71 64 3 71 64 3 71 64 35	78 36 82 66 1 83 78 4 83 78 4 82 80 6 76 76 7 68 68 6	65 16 75 34 80 57 81 69 2 77 72 4 77 72 4 71 68 5	60 9 71 21 78 44 81 66 1 79 74 2 78 74 2 77 73 5	45 3 60 9 74 26 81 52 81 70 2 80 76 5 77 76 6 77 75 6	53 6 6413 7429 8050 8167 7975 7976 7473 7269 7269	45 4 60 12 68 24 67 29 67 29 35 6	1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	33 33 14 14 14 14 14 14 14 14 14 14 14 14 14	
58	85 84 73 88 73	85 82 82 83 82 83 83 83 83 83 84 83 84 83 84 83 84 83 84 83 84 85 84 85 85 85 85 85 85 85 85 85 85 85 85 85	82 84 88 83 73 80 73 73 80 72	8 8 8 8 8 7 8 8 8 8 8 8 8 8 8 8 8 8 8 8	80 87 7 7 7 7 7 7 7 7 7 7 7 7 7	83 88 73 73 73 73	02 28 28 28 28 28 28 28 28 28 28 28 28 28	70 53 10 10 10 10 10 10	75 57 57 57 57 57 57 57 57 57 57 57 57 5	
0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	00000000 0000000 00000000	ооооооо ниноооо ооо <u>т</u> ооо	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0		<u> </u>	<u> </u>	
2 0 (7015 8051 82742 79765 74736	53 4 73 23 31 58 31 74 2 76 73 5 86 7 6	26 1 43 3 53 13 71 31 71 31 73 46 73 46 73 54 1	29 1 48 3 88 16 77 43 77 68 1 77 68 3 76 64 2	14 0 28 1 50 4 70 17 78 46 79 67 1 76 71 4 6 71 3	17 1 30 1 67 13 7 0 5 1 7 0 5 1 7 0 5 1 7 0 5 1 8 7 1 8 7 0 2 7 6 7 1 3 1 6 4 3 1 6 4 3	28 1 28 1 29 10 29 10 24 2 20 0 24 2 20 0 24 2 20 0 24 2 20 0 24 2 25 0 25 0 25 0 25 0 25 0 25 0 25 0 25	00000000000000000000000000000000000000	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	соооооооооооооооооооооооооооооооооооо
28	84 85 84 79 74 71 71	82 84 92 84 92 84 95 95 95 95 95 95 95 95 95 95 95 95 95	79 282	75 80 80 80 77 77	25 25 25 25 25 25 25 25 25 25 25 25 25 2	64 7 7 7 9 8 8 8 8 7 7 7 7 7 9 7 7 9 7 7 9 7 7 9 7 7 9 7	63 74 78 78 75 69 59 25 59 2	4 1 1 2 2 2 2 4 0 0 0 4 0 0 0 0 0 0 0 0 0 0 0	86 88 87 89 89 89 89 89 89 89 89 89 89 89 89 89	8 4 4 2 2 3 3 4 4 4 4 4 4 4 4 4 4 4 4 4 4
19 <mark>-87</mark>	1 - 86 2 - 85 8 - 88 8 - 88 2 - 75 2 - 75	00 - 85 00 - 85 00 - 82 00 - 77 00 - 77 00 - 68	14 - 86 25 - 86 20 - 88 20 - 88 20 - 79 20 - 70 20 - 70 20 - 88 20 - 72 20 - 72 20 20 - 72 20 20 20 - 72 20 20 - 72 20 20 20 - 72 20 20	25 - 87 25 - 86 25 - 86 25 - 81 25 - 81 2 - 73 2 - 73	2 - 87 8 - 86 8 - 86 86 - 85 86 - 85 82 82 83 84 81 83 84 81 83 84 81 85 81 81 81 81 81 81 81 81 81 81 81 81 81	28 87 56 86 56 86 56 86 56 88 56 88 56 83 56 86 56 86	25 - 87 11 - 85 84 - 81 84 - 81 6 - 74 6 - 74 78	28 83 56 88 56 88 56 78 51 87 51 69 51 69 510 510 510 510 510 510 510 510 510 510	25 86 25 85 25 85 25 85 86 25 85 88 25 85 88 25 85 88 20 83 20 84 20 85 20 85 20 20 85 20 20 20 20 20 20 20 20 20 20 20 20 20	0.000 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
l trainir	L. E = L. E = L. E = L. E = L. E = E = 3	ε ε = 1 ¹ ε ε = 3(ε ε = 6(ε = 12(ε = 240 ε = 240 ε = 480	9562.4 = 1913 38250 = 7650 = 7650 15300 15300 15300	0.031: 0.065 e = 0.12 e = 0.12 5 e = 0.2 5 e = 0.2 EG e =	PEG & = 1 PEG & = 1 PEG & = 1 EG & = 1 G & = 6 G & = 6 S & = 12	$\delta \epsilon = 1$ $\delta \epsilon = 2$ $\delta \epsilon = 20$ $\delta \epsilon = 5$ $\epsilon = 105$ $\epsilon = 204$ $\epsilon = 204$	$\varepsilon = 0$. $ c \varepsilon = 0$ $ c \varepsilon = 0$ $ c \varepsilon = 1$ $ c \varepsilon = 1$ $ c \varepsilon = 1$	g = 1; g = 1; g = 2; g = 2; g = 105; e = 102; e = 102; e = 102; e = 103; e = 103;	: = = 6.1 or = 2 or = 2 or = = 5 or = = 5 or = = 12 or =	0.031: = 0.065 = 0.15 = 0.15 = 0.15 = 0.15 = 0.15 = 0.15 = 0.15 = 0.06 = 0.00 = 0.00 = 0.00 = 0.00 = 0.00 = 0.063
Vormal		575 C C C C C	$L_1 \mathcal{E} = L_1 \mathcal{E}$ $L_1 \mathcal{E} = L_1 \mathcal{E}$ $L_1 \mathcal{E} = L_1 \mathcal{E}$ $L_1 \mathcal{E} = L_1 \mathcal{E} = L_1 \mathcal{E}$ $L_1 \mathcal{E} = L_1 \mathcal{E} = L_1 \mathcal{E}$	EG ε = 	L2-JF L2-JF L2-JF L2-JPE L2-JPE L2-JPE L2-JPEG	L_1 -JPE(L_1 -JPE(L_1 -JPE(L_1 -JPEG L_1 -JPEG L_1 -JPEG L_1 -JPEG L_1 -JPEG L_1 -JPEG L_2 -JPEG L_2 -JPEG L_1 -JPEG L_2 -JPEG L_2 -JPEG L_2 -JPEG L_2 -JPEG	Elastic Elasti Ela: Ela: Ela: Ela: Elast	Fog Fog Fog Fog Fog ε Fog ε	Gabor Gabor Gab Gabor Gabor Gabor Gabor	ow ε Snow ε Snow ε Sno Sno Sno Sno
-					1				5.0	۲ v