RenderAttack: Hundreds of Adversarial Attacks Through Differentiable Texture Generation

Dron Hazra
University of Cambridge
dh700@cam.ac.uk

Alex Bie
MATS
alexbie98@gmail.com

Mantas Mazeika
Center for AI Safety

Xuwang Yin
Center for AI safety

Andy Zou
Center for AI Safety

Dan Hendrycks
Center for AI Safety

Max Kaufmann
UK AI Safety Institute

Abstract

A longstanding problem in adversarial robustness has been defending against attacks beyond standard ℓ_p threat models. However, the space of possible non-ℓ_p attacks is vast, and existing work has only developed a small number of attacks, due to the manual effort required to design and implement each individual attack. Building on recent progress in differentiable material rendering, we propose RenderAttack, a scalable framework for developing large numbers of structurally diverse, non-ℓ_p adversarial attacks. RenderAttack leverages vast, existing repositories of hand-designed image perturbations in the form of *procedural texture generation graphs*, converting them to differentiable transformations amenable to gradient-based optimization. In this work, we curate 160 new attacks and introduce the ImageNet-RA benchmark. In experiments, we find that ImageNet-RA poses a challenge for existing robust models and exposes new regions of attack-space. By comparing state-of-the-art models and defenses, we identify promising directions for future work in ensuring robustness to a wide range of test-time adversaries.

1 Introduction

Recent work in adversarial robustness has greatly improved defenses against imperceptible ℓ_p attacks [Wang et al., 2023, Fort and Lakshminarayanan, 2024]. However, relatively little progress has been made against attacks beyond the standard ℓ_p threat model. This is concerning, as real-world adversaries are often not bound by perceptibility constraints [Gilmer et al., 2018]. For example, jailbreaking attacks on multimodal AI agents can make use of large perturbations to input images [Bailey et al., 2023], highlighting the importance of studying and improving robustness to the vast set of non-ℓ_p attacks.

Studying robustness beyond the ℓ_p threat model is challenging. This is because the space of attacks is far greater, requiring considerable effort to even set up evaluations. Structurally diverse and visually interesting attacks are hard to make, requiring manual labor and bespoke implementations. As a result, the research community has only developed a small number of non-ℓ_p attacks, limiting our ability to study the space of attacks and develop defenses.

To address this problem, we propose RenderAttack, a framework for generating large numbers of diverse, high-quality attacks. We leverage progress in computer graphics on differentiable textures along with repositories of hand-designed procedural texture graphs to create vast quantities of diverse, visually interesting, non-ℓ_p adversarial attacks. While many prior works propose individual attacks or a handful of attacks (e.g., [Xiao et al., 2018, Bhattad et al., 2019, Kaufmann et al., 2019]),

AdvML-Frontiers'24: The 3nd Workshop on New Frontiers in Adversarial Machine Learning@NeurIPS'24, Vancouver, CA.

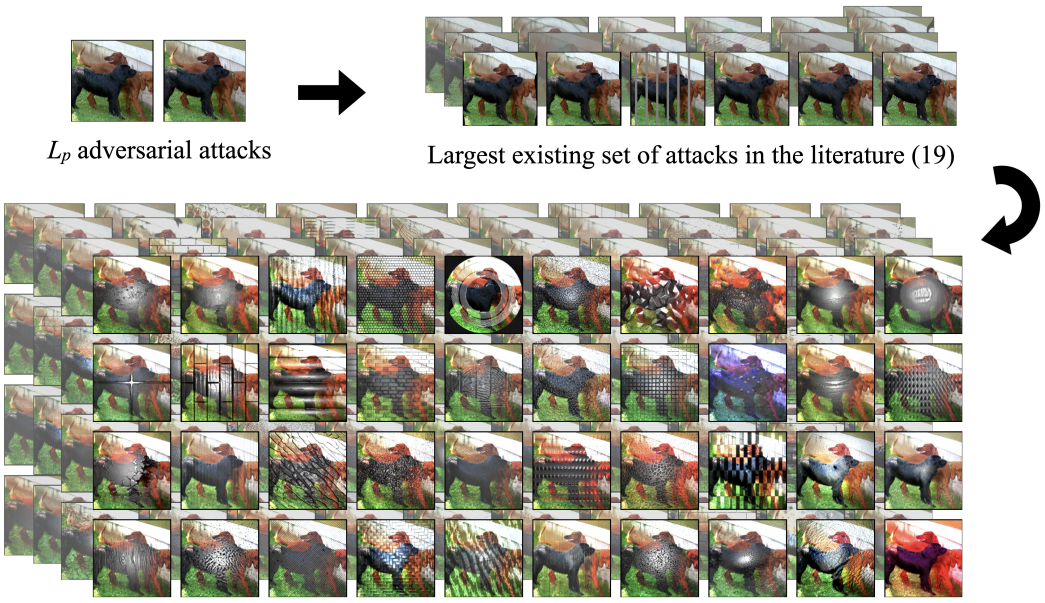


Figure 1: **ImageNet-RA is the largest set of adversarial attacks in the literature.** By leveraging vast repositories of high-quality procedural textures, we develop $10\times$ more attacks than Kaufmann et al. [2019], which is the current largest set of attacks available in the literature. Pictured above are attacks from ImageNet-RA (high severity). Visually, our attacks exhibit a high degree of variation.

we leverage our framework to introduce 160 new attacks. Using these attacks, we introduce the ImageNet-RA benchmark, which provides three levels of severity calibrated to challenge current and future models while preserving semantic information. Compared to the largest existing suite of attacks, this represents a nearly tenfold increase in the number of attacks.

We analyze the statistics of our new attacks and characterize their performance in extensive evaluations against a large suite of models. Our results reveal various factors that contribute to improved robustness on ImageNet-RA and demonstrate that our attacks cover a broader range of attack space compared to prior work. This suggests that ImageNet-RA could serve as a valuable tool for work requiring multiple diverse attacks, including research multi-attack robustness and robustness to unforeseen adversaries. We hope that these results and our new attacks enable future work on studying and improving robustness to attacks beyond the standard ℓ_p threat model.

2 Related work

Moving beyond ℓ_p-based robustness evaluations. Most of the classical work in the adversarial robustness literature of previous work in adversarial robustness has been on the ℓ_p-ball, either directly measuring robustness to ℓ_p-based attacks [Croce and Hein, 2020], or trying to place bounds on the ℓ_p robustness of classifiers [Moosavi-Dezfooli et al., 2016, Weng et al., 2018]. Many previous works have pointed out the need to go further than these ℓ_p-based evaluations of robustness, highlighting the need to (1) provide a much more comprehensive multi-attack measure of robustness [Dai et al., 2023, 2024, Maini et al., 2020], and (2) testing robustness against adversaries which are not available during training[Kaufmann et al., 2019, Laidlaw et al., 2020]. However, most of these works create their own adversarial attacks [Kaufmann et al., 2019, Xiao et al., 2018, Dai et al., 2023], which is highly time consuming [Kaufmann et al., 2019]. Instead of relying solely on a small pool of researcher efforts, we leverage the work of Shi et al. [2020] to create a pipeline for turning non-differentiable procedural textures into differentiable adversarial attacks, allowing us to create much larger sets of adversarial attacks than what is currently available in the literature.

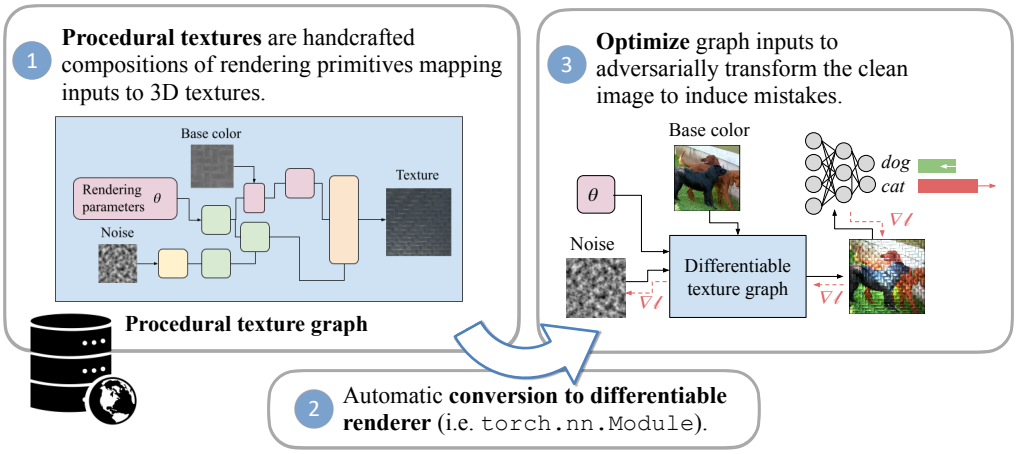


Figure 2: **Our attacks leverage hand-designed, differentiable texture graphs.** We depict our RenderAttack framework for designing attacks in the above figure. Our attacks leverage hand-designed texture graphs available online. We then use the DiffMatt library to turn these texture graphs into differentiable operators that apply realistic, optimizable textures to input images. This enables the use of high-quality textures for adversarial attacks.

Differentiable rendering for generating adversarial attacks. Rendering is the process of converting a description of some scene in terms of objects and lighting sources into an image of that scene. Differentiable renderers can be used to backpropagate through the rendering process—allowing underlying scenes to be optimized with respect to properties of the output images. Previous works, such as those by Liu et al. [2018] and Jain et al. [2019], utilize such differentiable renderers to create adversarial attacks. However, these works focus solely on spatial transformations of 3D objects, or simple lighting changes. In contrast we leverage the work of Shi et al. [2020] to turn hand-designed procedural textures into end-to-end differentiable graphs, into which we can optimize the latent noise variables to create an adversarial attack. Compared to these previous works, our attacks allow variations across whatever features designers parameterize: including color, textures, and reflective properties—leading to much higher variety of adversaries.

3 RenderAttack

3.1 The need for new measures of adversarial robustness

Several existing works [Kaufmann et al., 2019, Dai et al., 2024, 2023] have pointed out the limitations of the classic ℓ_p adversarial robustness [Madry et al., 2019], namely that:

1. Adversaries may be *unforeseen*: it is unlikely that we will have train-time access to the adversaries to which we need to be robust to at deployment time.
2. Adversaries may be *non-*ℓ_p: attackers do not need to constrain themselves to the ℓ_p ball.
3. Adversaries may be *diverse*: motivated attackers are likely to find and use a diverse range of worst-case inputs to circumvent various defense mechanisms.

To model these properties of real-world adversaries, we follow Kaufmann et al. [2019] and consider measuring the robustness of our classifiers f robustness against a population of adversaries, defined as:

$$\mathbb{E}_{(x,y),A\sim\mathcal{D},\mathcal{A}}\left[\min_{x_{\text{adv}}\in S_x^A}\{\mathbf{1}_{f(x_{\text{adv}})=y}\}\right] \tag{1}$$

where $\mathcal{D}$ is the data distribution, $\mathcal{A}$ is our distribution of adversaries, and S_x^A is the set of potential adversarial examples set for a particular $A \in \mathbf{Dom}(\mathcal{A})$.

Under this definition, asking for attacks to be *unforeseen* places the constrains that any defence mechanism must not make use of information from the distribution $\mathcal{A}$ during training. To create *non-*ℓ_p and *diverse* adversaries, we must select a suitable adversarial population $\mathcal{A}$—this is what our work focuses on.

3.2 Differentiable procedural materials for adversarial attacks

Procedural textures are textures which are generated by mapping randomly generated noise to output texture, rather than being defined by a fixed artist-drawn image. They are widely used in a variety of computer graphics applications, offering low storage costs, more realistic results and flexible resolution. It is most common to represent procedural textures as *procedural material graphs*, which apply a series of image transformations on input noise to produce the desired texture In Shi et al. [2020], the authors leverage the key property that these transforms are often differentiable to map these graphs into `torch.nn.Module`. We use this technique to turn a set of material graphs into visually diverse, highly optimizable and effective adversarial attacks.

We now describe our method for generating adversarial attacks using differentiable material graphs. For our purposes, we will encapsulate the evaluation and rendering of a material graph $\mathcal{G}$ as a differentiable function $M_{\mathcal{G}}$ that takes input noises σ and produces an output texture image. To corrupt a clean image x, we blend it with the output colour map of $\mathcal{G}$ and then render the resulting texture:

$$x' = A_{\mathcal{G}}(x, \sigma). \tag{2}$$

$A_{\mathcal{G}}$ is the combination of texture generation, and blending with the image—outputting a fully corrupted image x'. See Appendix B for details.

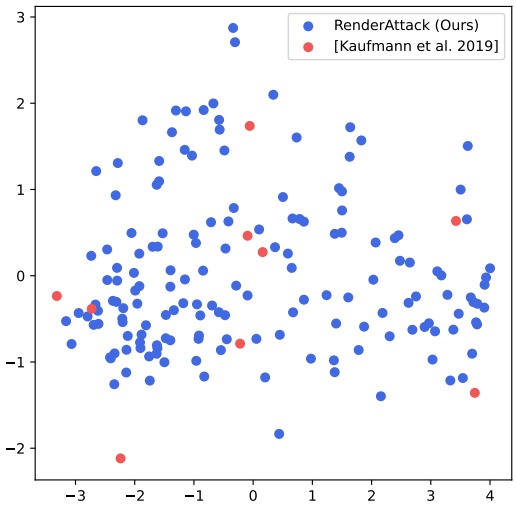


Figure 3: **Our attacks produce more diverse model behaviors than existing literature.** Compared to prior suites of non-ℓ_p attacks, the space of attacks is covered far more extensively by our new RenderAttack adversaries. Each point corresponds to an attack. For each attack, we take the corresponding vector of model accuracies across a suite of 18 models. We normalize these vectors so that each dimension has mean 0 and variance 1, then plot the first two principal components. This illustrates the space of different accuracy profiles that attacks have, independent of their relative strength.

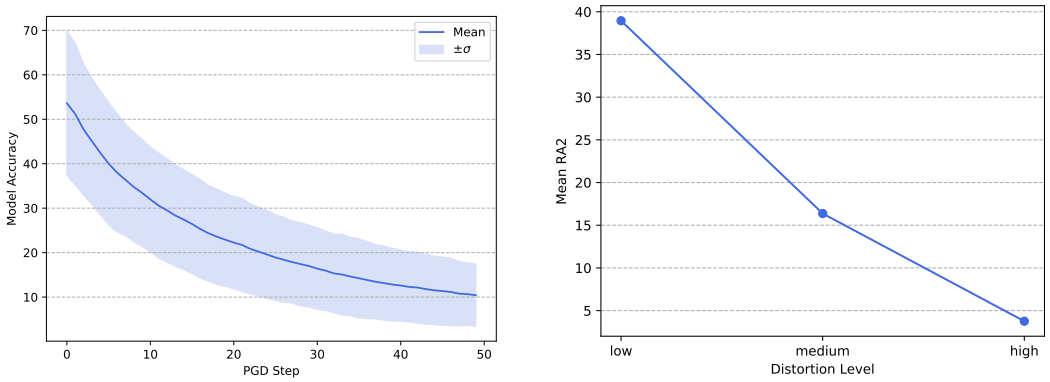
To find the worst-case versions of this corruption in x_{adv}, we will optimise our latent variables σ to maximise the loss of our model f, subject to an L_p-based constraint on our latent variables:

$$\sigma_{\text{adv}} = \underset{\sigma:\|\sigma\|_p \leq \varepsilon}{\operatorname{argmin}} \{\mathcal{L}(f(A_{\mathcal{G}}(x, \sigma)), y)\}$$
$$x_{\text{adv}} = A_{\mathcal{G}}(x, \sigma_{\text{adv}}).$$

As originally done in Madry et al. [2019] we use the popular PGD (Projected Gradient Descent) algorithm to find x_{adv}.

A scalable path towards potentially thousands of adversarial attacks. When creating a new set of diverse adversarial attacks, adversarial robustness researchers must hand-design a visually diverse and interesting set of optimizable transformations which can be applied to images [Kaufmann et al., 2019, Xiao et al., 2018, Dai et al., 2023]. This takes large amounts of creativity and effort, fundamentally limiting how many attacks can be introduced by any one paper in the literature. In contrast, our pipeline allows us to use large, community-created repositories of procedural textures and turn them into adversarial attacks. This unlocks a huge new set of potential adversaries for the community to experiment with.

160 new attacks for the adversarial robustness community. We use our generation pipeline to create 160 adversarial attacks, which is an order of magnitude more than any other works in the



(a) Performance with increasing optimisation steps.

(b) Performance with increasing attack epsilon levels.

Figure 4: **Our attacks improve with more optimisation power.** In Figure 4a, we plot model accuracy as we increase number of optimization steps and in Figure 4b we plot how increasing the epsilon budget which limits the adversary improves performance. Experiments in Figure 4a are across all the ResNet50s in our baseline models, and Figure 4b is across all of our baseline models. In both cases, we see that our attacks are sensitive to optimisation power.

literature (the closest, that of Kaufmann et al. [2019], presents 19 novel attacks). We hope that these can be used both to evaluate defence techniques against a large and varied pool of adversaries, but also as a resource to help carry out more fundamental research on the adversarial robustness of models. We further note that 160 was an arbitrary cutoff point—if needed, future work can use our pipeline to generate even more novel attacks.

3.3 ImageNet-RA: a new benchmark for comparing multi-attack robustness

To demonstrate the usefulness of our attack generation pipeline, we introduce a new benchmark ImageNet-RA (ImageNet-RenderAttack), and a new metric RA2 (RenderAttack Accuracy). We construct this metric by considering the population of adversaries in Equation (1) to be a uniform choice out of the 160 adversaries we gathered using our pipeline. This leads a single number accuracy, corresponding to the average model accuracy against one of the attacks on our suite. In line with previous work [Kaufmann et al., 2019], we also select three difficulty levels for the benchmark: RA2_{low}, RA2_{med}, RA2_{high}, corresponding to three different hand-picked per-attack ϵ optimisation budgets. In this work, we focus our results around RA2_{med}, which we simply call RA2.

4 Experiments

We evaluate a suite of 18 baseline models on ImageNet-RA. These allow us to better characterize how our metric behaves, and hence find a range of promising defence techniques

4.1 How does RA2 compare to other measures of robustness?

RA2 is a measure of worst-case robustness. Benchmarks such as ImageNet-C [Hendrycks and Dietterich, 2019] measure average-case robustness. As can be seen in Table 1, ImageNet-RA behaves like a measure of worst-case robustness—adversarially trained models perform much better on ImageNet-RA than models trained against fixed augmentations. In contrast, models trained with augmentations perform better on ImageNet-C.

RA2 does not behave like ℓ_p robustness metrics. In Table 2, we show how adversarial training affects performance on ImageNet-RA. Adversarial training improves RA2, from 4.1 to the mid-20 percent range. Intuitively, one might expect that adversarial training with higher epsilon would further improve robustness to perceptible attacks, since the training perturbations are larger. However, we find that adversarial training with higher epsilons does not yield further gains in RA2. This suggests that new techniques are needed to improve robustness to RenderAttack adversaries.

To further gauge differences between our attacks and standard ℓ_p attacks, we compare RA2 to ℓ_∞ PGD accuracy across our suite of models. We find weak correlation between the two metrics ($r = 0.526$), with many models achieving high RA2 while not being ℓ_∞-robust (see Figure 7a). Instead, as can be

Model	$\text{RA}_{\text{zero}} \uparrow$ (no optimization)	mCE $\downarrow$	RA2 $\uparrow$
Resnet50	42.0	76.7	4.1
Resnet50 + AugMix	47.0	65.7	9.3
Resnet50 + DeepAug	50.8	**61.1**	7.4
Resnet50 + PixMix	**54.4**	65.8	10.4
Resnet50 + L_2, ($\varepsilon = 5$)	35.1	89.0	20.6
Resnet50 + L_∞, ($\varepsilon = 8/255$)	35.6	85.1	**24.9**

Table 1: **Common corruptions and** RA2 We compare performance on the ImageNet-C benchmark (mCE) to performance against both non-optimized and optimized versions of our attacks. We find that after optimization, RA2 behaves like worst-case measure of robustness—best performing models are those which have been adversarially trained, not those with image augmentations.

seen in Figure 7b our metric behaves more like existing measures of multi-attack robustness, showing a much stronger correlation ($r = 0.704$) with the UA2 metric of Kaufmann et al. [2019].

RenderAttack elicits a more diverse set of model behaviours. We are interested to see if our higher number of attacks leads to different sets of model behaviours across our benchmark. In Figure 3, we give each attack a *model accuracy profile* (the vector of the accuracies of our baselines on that attack), and compare whether these show more variety than the accuracy profiles of Kaufmann et al. [2019]. We perform dimensionality reduction with PCA and plot the first two principle components. This visualization shows that our attacks encompass the range of attacks proposed by Kaufmann et al. [2019] while having much denser coverage of the space. This suggests that our attacks cover a broader range of behaviour profiles across models.

4.2 How do different defence techniques affect RA2?

Model scale improves RA2**.** Some prior work has found that larger models tend to be more robust [Mao et al., 2022]. We check whether this property also holds for our attacks in Figure 5. We evaluate RA2 across the ConvNeXt-V2 family of models, finding that larger models indeed perform better on ImageNet-RA. This corroborates the findings of earlier works in the adversarial robustness literature and suggests that this is a consistent property that extends to non-ℓ_p attacks.

Training	Train ε	Clean Acc.	RA2
Standard	-	**75.9**	4.1
L_2	1	67.2	17.3
	3	62.8	20.8
	5	56.1	20.6
L_∞	$2/255$	69.1	24.3
	$4/255$	63.9	**26.6**
	$8/255$	54.5	24.9

Table 2: ℓ_p **training.** We evaluate a range of ResNet-50 models trained against ℓ_p adversaries on ImageNet-RA. Adversarial training on ℓ_p attacks provides limited benefit, and training on larger perturbations is not an effective way to improve robustness to our diverse, perceptible attacks. This suggests that new methods are needed to improve performance on ImageNet-RA.

The best-performing model is not adversarially trained. We show RA2 for all models in Figure 6. The best-performing model is DinoV2-Large with registers. This is surprising, as this model was not explicitly adversarially trained, but is instead the product of large-scale self-supervised pretraining objective—this is in contrast with classical adversarial benchmarks, where models trained outside of robustness do not perform well. However, we can see in Table 2, adversarial training still largely improves on standard models. Unifying pre-training and adversarial training might be a promising avenue for high performance against such a diverse set of adversaries.

5 Conclusion

We present ImageNet-RA, a robustness benchmark with 160 novel, hand-designed attacks—by far the largest number of attacks in the literature. Our RenderAttack framework leverages differentiable rendering techniques and online collections of hand-designed texture graphs to create large numbers of high-quality attacks. We hope these new attacks foster future work on studying robustness beyond the standard ℓ_p threat model and improving defence techniques.

References

Luke Bailey, Euan Ong, Stuart Russell, and Scott Emmons. Image hijacks: Adversarial images can control generative models at runtime. *arXiv preprint arXiv:2309.00236*, 2023.

Anand Bhattad, Min Jin Chong, Kaizhao Liang, Bo Li, and David A Forsyth. Unrestricted adversarial examples via semantic manipulation. *arXiv preprint arXiv:1904.06347*, 2019.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020.

Sihui Dai, Saeed Mahloujifar, Chong Xiang, Vikash Sehwag, Pin-Yu Chen, and Prateek Mittal. Multirobustbench: Benchmarking robustness against multiple attacks. In *International Conference on Machine Learning*, pages 6760–6785. PMLR, 2023.

Sihui Dai, Chong Xiang, Tong Wu, and Prateek Mittal. Position paper: Beyond robustness against single attack types. *arXiv preprint arXiv:2405.01349*, 2024.

Stanislav Fort and Balaji Lakshminarayanan. Ensemble everything everywhere: Multi-scale aggregation for adversarial robustness. *arXiv preprint arXiv:2408.05446*, 2024.

Justin Gilmer, Ryan P Adams, Ian Goodfellow, David Andersen, and George E Dahl. Motivating the rules of the game for adversarial example research. *arXiv preprint arXiv:1807.06732*, 2018.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019.

Lakshya Jain, Varun Chandrasekaran, Uyeong Jang, Wilson Wu, Andrew Lee, Andy Yan, Steven Chen, Somesh Jha, and Sanjit A Seshia. Analyzing and improving neural networks by generating semantic counterexamples through differentiable rendering. *arXiv preprint arXiv:1910.00727*, 2019.

Max Kaufmann, Daniel Kang, Yi Sun, Steven Basart, Xuwang Yin, Mantas Mazeika, Akul Arora, Adam Dziedzic, Franziska Boenisch, Tom Brown, et al. Testing robustness against unforeseen adversaries. *arXiv preprint arXiv:1908.08016*, 2019.

Cassidy Laidlaw, Sahil Singla, and Soheil Feizi. Perceptual adversarial robustness: Defense against unseen threat models. *arXiv preprint arXiv:2006.12655*, 2020.

Hsueh-Ti Derek Liu, Michael Tao, Chun-Liang Li, Derek Nowrouzezahrai, and Alec Jacobson. Beyond pixel norm-balls: Parametric adversaries using an analytically differentiable renderer. *arXiv preprint arXiv:1808.02651*, 2018.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks, 2019. URL `https://arxiv.org/abs/1706.06083`.

Pratyush Maini, Eric Wong, and Zico Kolter. Adversarial robustness against the union of multiple perturbation models. In *International Conference on Machine Learning*, pages 6640–6650. PMLR, 2020.

Xiaofeng Mao, Gege Qi, Yuefeng Chen, Xiaodan Li, Ranjie Duan, Shaokai Ye, Yuan He, and Hui Xue. Towards robust vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12042–12051, June 2022.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016.

Liang Shi, Beichen Li, Miloš Hašan, Kalyan Sunkavalli, Tamy Boubekeur, Radomir Mech, and Wojciech Matusik. Match: Differentiable material graphs for procedural material capture. *ACM Transactions on Graphics (TOG)*, 39(6):1–15, 2020.

Zekai Wang, Tianyu Pang, Chao Du, Min Lin, Weiwei Liu, and Shuicheng Yan. Better diffusion models further improve adversarial training. In *International Conference on Machine Learning*, pages 36246–36263. PMLR, 2023.

Tsui-Wei Weng, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, Dong Su, Yupeng Gao, Cho-Jui Hsieh, and Luca Daniel. Evaluating the robustness of neural networks: An extreme value theory approach. *arXiv preprint arXiv:1801.10578*, 2018.

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

A Additional Results

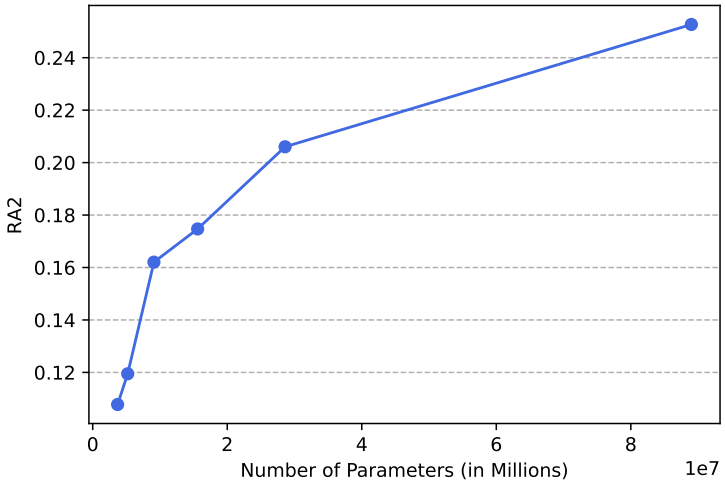


Figure 5: We measure performance of ConvNext-V2 models of varying sizes, observing that larger models obtain higher performance on ImageNet-RA.

B Further Details on Attack Generation

Formally, a procedural material graph $\mathcal{G}$ is modelled as a directed acyclic graph with *generator* or *filter* nodes. Generator nodes take no input and create spatial textures, like noise or structural patterns, which we refer to with σ. This initial data is passed through filter nodes, which perform operations on such as HSL edits, interpolation, or blurs on their inputs and pass their output along the edges of the graph. This process terminates at the final output nodes, which produce per-pixel parameter maps of a spatially-varying bidirectional reflectance distribution function (SV-BRDF). We encapsulate graph evaluation with a function $T_{\mathcal{G}}$. We follow the DiffMat implementation and use four output maps: *albedo/color*, *normals*, *roughness* and *metallicity*. These maps can then be rendered under some lighting conditions to produce the final output image, which we encapsulate as R.

In our work, we blend the color map produced by our material graph with our clean image x with some blend function B, and then render the result. Hence our process for generating an adversarial example is

$$(x, \sigma) \xrightarrow{T} (x, \dots) \xrightarrow{B} (\tilde{x}, \dots) \xrightarrow{R} x_{adv}$$

We use "soft light" where $B(a, b) = (1 - 2b)a^2 + 2ba$. We tested other blend modes, and we found "soft light" tended to visually preserve the most information.

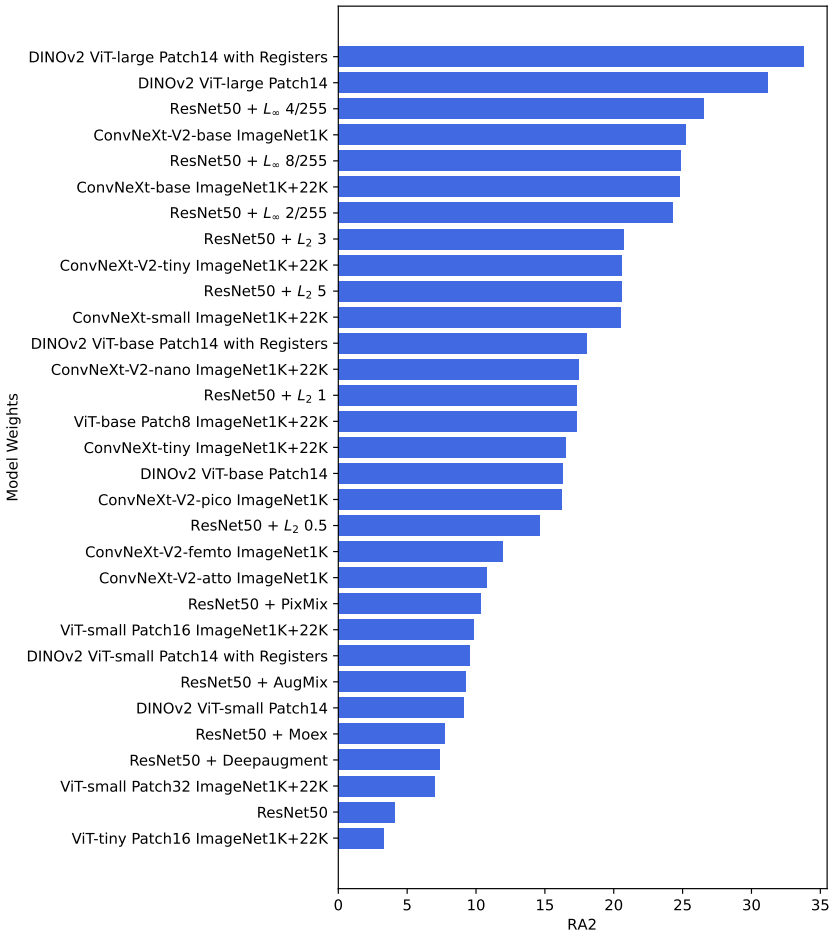
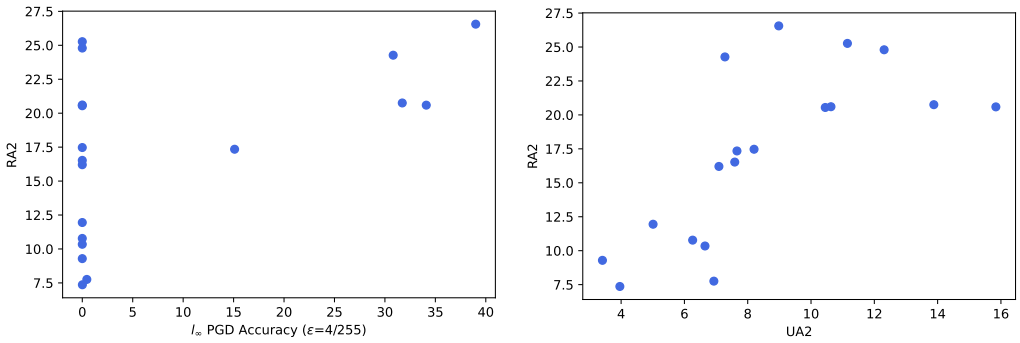


Figure 6: Here, we show RA2 for each model in our evaluations. All values are percentages.



(a) Correlation between PGD accuracy and RA2

(b) Correlation between RA2 and UA2

Figure 7: **Performance on** ImageNet-RA **is a broad measure of robustness.** We plot the correlation of model performance on ImageNet-RA and existing works in the literature (L_∞ attacks ($r = 0.526$) and ImageNet-UA ($r = 0.704$). Each point corresponds to a different model. Our results show that it behaves like other general measures of robustness.