

Unadversarial Examples: Designing Objects for Robust Vision

Hadi Salman*
hadi.salman@microsoft.com
Microsoft Research

Andrew Ilyas*
ailyas@mit.edu
MIT

Logan Engstrom*
engstrom@mit.edu
MIT

Sai Vemprala
saihv@microsoft.com
Microsoft Research

Aleksander Madry
madry@mit.edu
MIT

Ashish Kapoor
akapoor@microsoft.com
Microsoft Research

Abstract

We study a class of realistic computer vision settings wherein one can influence the design of the objects being recognized. We develop a framework that leverages this capability to significantly improve vision models’ performance and robustness. This framework exploits the sensitivity of modern machine learning algorithms to input perturbations in order to design “robust objects,” i.e., objects that are explicitly optimized to be confidently detected or classified. We demonstrate the efficacy of the framework on a wide variety of vision-based tasks ranging from standard benchmarks, to (in-simulation) robotics, to real-world experiments. Our code can be found at <https://git.io/unadversarial>.

1 Introduction

Performing reliably on unseen or shifting data distributions is a difficult challenge for modern computer vision systems. For example, slight rotations and translations of images suffice to reduce the accuracy of state-of-the-art classifiers [ETT+19; ALG+19; KMF18]. Similarly, models that attain near human-level performance on benchmarks exhibit significantly degraded performance when faced with even mild image corruptions and transformations [HD19; KSH+19]. Furthermore, when an adversary is allowed to modify inputs directly, standard vision models can be manipulated into predicting arbitrary outputs (cf. *adversarial examples* [BCM+13; SZS+14]). While robustness interventions and additional training data can improve out-of-distribution behavior, they do not fully close the gap between model performance on standard heldout data and on corrupted/otherwise unfamiliar data [TDS+20; HBM+20]. The situation is worse still when test-time distribution is under- or mis-specified, which occurs commonly in practice.

How can we change this state of affairs? We propose a new approach to image recognition in the face of unforeseen corruptions or distribution shifts. This approach is rooted in a reconsideration of the problem setup itself. Specifically, we observe that in many situations, a system designer not only trains the model used to make predictions, but also controls, to some degree, the inputs that are fed into that model. For example, a drone operator seeking to train a landing pad detector can modify the surface of the landing pad; and, a roboticist training a perception model to recognize a small set of custom objects can slightly alter the texture or design of these objects.

We find that such control over inputs can be leveraged to drastically improve our ability to tackle computer vision tasks. In particular, it allows us to turn the input-sensitivity of modern vision systems from a weakness into a strength. That is, instead of optimizing inputs to *mislead* models (as in adversarial examples), we can alter inputs to *reinforce* correct behavior, yielding what we refer to as “unadversarial examples.” Indeed, we show that even a simple gradient-based algorithm can successfully construct such unadversarial examples for a variety of vision settings and demonstrate that, by optimizing objects for vision systems (rather than only vice-versa), we can significantly improve both in-distribution performance *and* robustness to unforeseen data shifts and corruptions.

*Equal contribution.

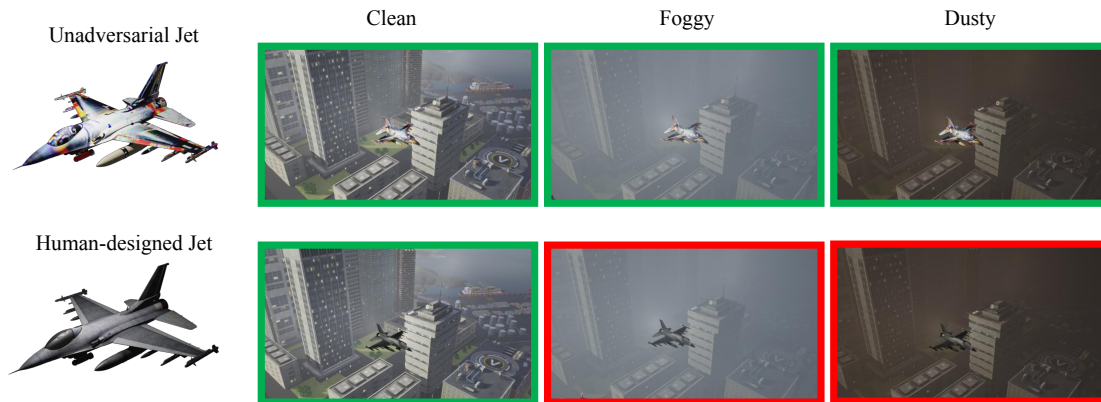


Figure 1: In this paper, we demonstrate that optimizing objects (e.g., the pictured jet) for pre-trained neural networks can significantly boost performance and robustness on computer vision tasks.

Specifically, in this paper, we develop methods for constructing unadversarial stickers/patches that boost the performance of the deep learning-based image classifiers operating on the corresponding objects. We then demonstrate the efficacy of these methods on both standard benchmarks (CIFAR, ImageNet) and robustness-based benchmarks (ImageNet-C, CIFAR-C) while also comparing them to a broad set of baselines, including QR codes. To further highlight the practicality of our framework, we (a) extend our methods to designing the texture of three-dimensional objects (rather than patches); (b) deploy unadversarial examples in a simulated drone setting; and (c) ensure that the performance improvement yielded by the objects we design actually transfer to the physical world.

2 Motivation and Approach

While vision models tend to perform well on held-out data drawn from the same distribution as the training data, out-of-distribution inputs can severely degrade this performance. For example, models behave unreliably under distribution shifts induced by new data collection procedures [RRS+19; EIS+20; TE11], synthetic corruptions [HD19; KSH+19], spatial transformations [ETT+19; FF15], as well as under other types of shift.

Given a fixed type of distribution shift, a standard approach to increasing model robustness is to explicitly train on or regularize with data from the corresponding anticipated test distribution [KSH+19]. For example, Engstrom et al. [ETT+19] find that vision models trained on worst-case rotations and translations end up being fairly robust to rotation and translation-based distribution shifts. However, this approach is not without shortcomings—for example, Kang et al. [KSH+19] find that training CIFAR classification models that are robust to JPEG-compression in this manner requires a significant sacrifice in natural accuracy. Recent works make similar observations in the context of other distribution shift mechanisms like ℓ_p adversaries [TSE+19; SZC+18; RXY+19] or texture swapping [GRM+19].

These observations give rise to a more general question: given that performing reliably in the face of constrained, well-specified distribution shifts is already a difficult challenge, how can we attain robustness to broad, unforeseen distribution shifts?

2.1 Leveraging more controlled vision settings

Consider the vision tasks of detecting a landing pad from a drone, or classifying manufacturing components from a factory robot. In both these tasks, reliable in-distribution performance is a necessity; still, a number of possible distribution shifts may occur at deployment time. For example, the drone might approach the landing pad at an atypical angle, or have a view obstructed by snow, smoke, or rain. Similarly, the factory robot may encounter objects in unfamiliar poses, or using a low-quality/noisy camera.

At first glance, dealing with these issues seems to require tackling the difficult problem of general distribution shift robustness discussed earlier in this section. However, there is in fact a critical distinction

between the scenarios considered above and vision tasks in their full generality. In particular, in these scenarios and many others, the system designer has control over not only the model that is used but also the physical objects that the model operates on. As we will demonstrate, the designer can use this capability to modify these objects to majorly boost the model’s ability to solve the problem at hand.

For instance, the designer of the drone’s landing algorithm could, in addition to training a detection model, also paint the landing pad bright yellow. A machine learning model trained to detect this custom landing pad might then be more effective than a model trained to detect a standard grey pad, especially in low-visibility conditions. Still, the particular choice to paint the landing pad yellow is rather ad hoc, and likely rooted in the way *humans* recognize objects. Meanwhile, an abundance of prior work (e.g., [JBZ+19; GRM+19; JLT18; IST+19]) demonstrates that humans and machine learning models tend to use different sets of features to make their decisions. This suggests that rather than relying on human priors, we should instead be asking: *how can we build objects that are easily detectable by machine learning models?*

2.2 Unadversarial examples

The task of making inputs *less* recognizable by computer vision systems has been a focus of research in *adversarial examples*. Adversarial examples are small, carefully constructed perturbations to natural images that can induce arbitrary (mis)behavior from machine learning models [BCM+13; SZS+14]. These perturbations are typically constructed as the result of an optimization problem that maximizes the loss of a machine learning model with respect to the input, i.e., by solving the optimization problem

$$\delta_{adv} = \arg \max_{\delta \in \Delta} L(f_{\theta}(x + \delta), y), \quad (1)$$

where f_{θ} is a parameterized model (e.g., a neural network with weights θ); x is a natural input; y is the corresponding correct label; L is the loss function used to train θ (e.g., cross-entropy loss) and Δ is a class of permissible perturbations (e.g., norm-bounded perturbations: $\Delta = \{\delta : \|\delta\|_p \leq \epsilon\}$ for some small $\epsilon > 0$). Adversarial perturbations are typically crafted via projected gradient descent (PGD) [Nes03] in input space, a standard iterative first-order optimization method—prior work in adversarial examples has shown that even a few iterations of PGD suffice to completely change the prediction of many state-of-the-art machine learning systems [MMS+18].

From adversarial examples to unadversarial objects. The goal of this work is to modify the design of objects so that they are more easily recognizable by computer vision systems. If we could specify every pixel of every image that a model encounters at test time, we could draw on the effectiveness of adversarial examples, and construct image perturbations (using PGD) that *minimize* the loss of the system, e.g.,

$$\delta_{unadv} = \arg \min_{\delta \in \Delta} L(\theta; x + \delta, y). \quad (2)$$

In our setting of interest, however, having such fine-grained access to the test inputs is unrealistic (presumably, if we had precise control over every pixel in the input, we could just directly encode the ground-truth label directly in it). Instead, we have *limited* control over some physical objects; these objects are in turn captured within image inputs, along with many signals that are out of our control, such as camera artifacts, weather effects, or background scenery.

It turns out that we can still draw on techniques from adversarial examples research in this limited-control setting. Specifically, a recent line of work [KGB17; SBB+16; EEF+18a; AEI+18] concerns itself with constructing *robust adversarial examples* [AEI+18], i.e., physically realizable objects that act as adversarial examples when introduced into a scene in any one of a variety of ways. For example, Sharif et al. [SBB+16] design glasses frames that cause facial recognition models to misclassify faces, Athalye et al. [AEI+18] design custom-textured 3D models that are misclassified by state-of-the-art ImageNet classifiers from many angles and viewpoints, and [BMR+18] design adversarial patches: stickers that can be placed anywhere on objects causing them to be misclassified. In this paper, we leverage the techniques developed in the above line of work to construct *robust un-adversarial objects*—physically realizable *objects* optimized to minimize (rather than maximize) the loss of a target classifier. In the next section, we will more concretely discuss our methods for generating unadversarial objects, then outline our evaluation setup.

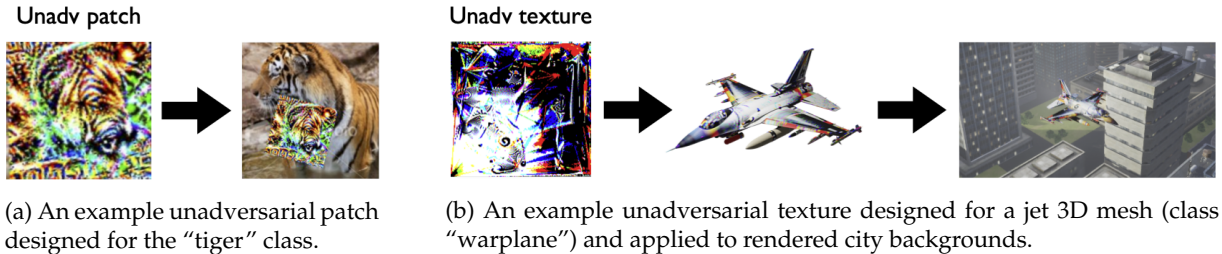


Figure 2: Examples of the two considered methods for constructing unadversarial objects.

2.3 Constructing unadversarial objects

In the previous section, we identified a class of scenarios where a system designer can not only control the machine learning model being deployed but also, to some extent, the objects that model operates on. In these settings, we motivated so-called *unadversarial examples* as a potential way to boost models’ overall performance and robustness to distribution shifts. In this section, we present and illustrate two concrete algorithms for constructing unadversarial examples: unadversarial patches and unadversarial textures. In the former, we design a sticker or “patch” [BMR+18] that can be placed on the object; in the latter, we design the 3D texture of the object (in a similar manner to the texture-based adversarial examples of Athalye et al. [AEI+18]). Example results from both techniques are shown in Figure 2. For simplicity, we will assume that the task being performed is image classification, but the techniques are directly applicable to other tasks as well. In all cases, we require access to a pre-trained model for the dataset of interest.

Unadversarial patches (cf. Algorithm 1 in Appendix A). To train unadversarial patches (cf. Figure 2a), in addition to the pre-trained model, we require sample access to image-label pairs from the dataset of interest. At each iteration, we sample an image-label pair (x, y) from a training set, and place the patch corresponding to class y onto the image with random orientation and position¹. Since placing the patch is an affine transformation, after each iteration we can compute the gradient of the model’s loss with respect to the pixels in the patch, and take a negative gradient step on the patch parameters. The algorithm terminates when the model’s loss on sticker-boosted images plateaus, or after a fixed number of iterations.

Unadversarial textures (cf. Algorithm 2 in Appendix A). To train unadversarial *textures* (cf. Figure 2b), we do not require sample access to the dataset, but instead a set of 3D meshes for each class of objects that we would like to augment, as well as a set of background images that we can use to simulate sampling a scene (these can be images from the dataset of interest, solid-color backgrounds, random patterns, etc.).

For each 3D mesh, our goal is to optimize a 2D texture which improves classifier performance when mapped onto the mesh. At each iteration, we sample a mesh and a random background; we then use a 3D renderer (Mitsuba [NVZ+19]) to map the corresponding texture onto the mesh. We overlay the rendering onto a random background image, and then feed the resulting composed image into the pre-trained classifier, with the label being that of the sampled 3D mesh. Since rendering is typically non-differentiable, we use a linear approximation of the rendering process (cf. Athalye et al. [AEI+18]) in order to compute (this time approximate) gradients of the model’s loss with respect to the utilized texture. From there, we apply the same SGD algorithm as we did for the patch case.

3 Experimental Evaluation

In order to determine the effectiveness of our proposed framework, we evaluate against a suite of computer vision tasks. We briefly outline the experimental setup of each task below, and show that unadversarial objects consistently improve the performance and robustness of the vision systems tested. For a more detailed account of each experimental setup, see Appendix C; code for reproducing our experimental results is available at <https://git.io/unadversarial>.

¹We allow the patch to be placed anywhere as a matter of convenience: ideally we would only be applying the patch onto the main object itself, but this would require bounding box data that we do not have for most classification datasets.

3.1 Clean data and synthetic corruptions

We first test whether unadversarial examples improve the performance of image classifiers on benchmark datasets. Using the algorithm described in Section 2.3, we construct unadversarial patches of varying size for pre-trained ResNet-50 classifiers on the CIFAR [Kri09] and ImageNet [RDS+15] datasets. For evaluation, we add these patches at random positions, scales, and orientations to validation set images (see Appendix C for the exact protocol). As shown in Figure 4a, the pre-trained ImageNet classifier is more consistently more accurate on the augmented ImageNet images. For example, an unadversarial patch 20 times smaller than ImageNet images boosts accuracy by 26.3% (analogous results for CIFAR are given in Appendix D).



Figure 3: Clean (left) and corresponding corrupted (right) ImageNet images augmented with an unadversarial patch—we use such images to evaluate the efficacy of unadversarial patches in Section 3.1.

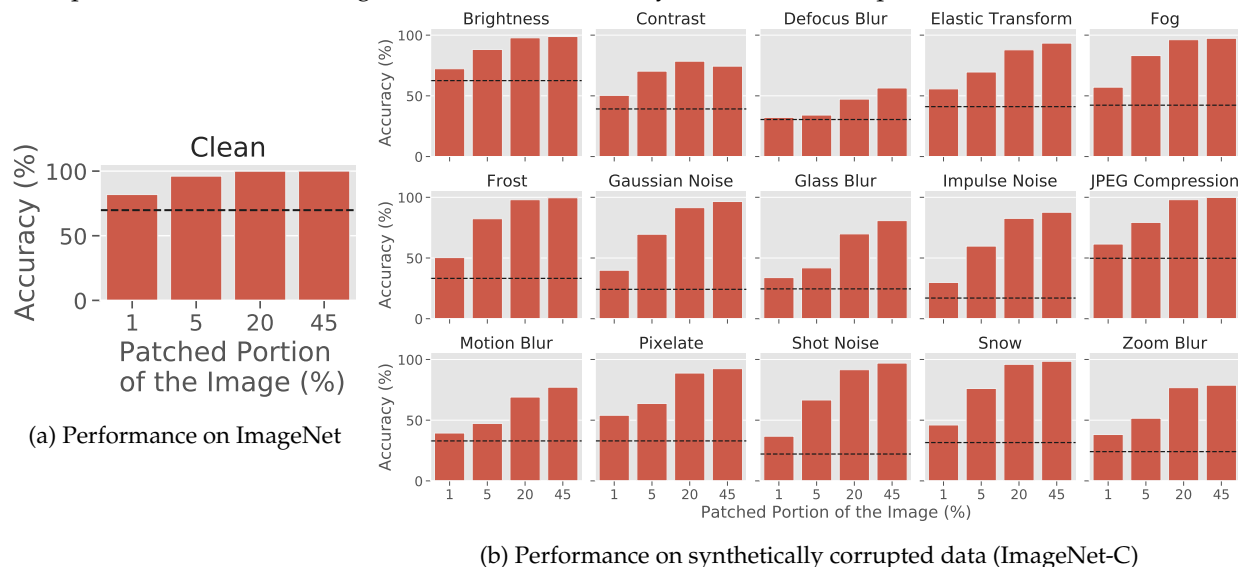


Figure 4: Accuracy on (a) clean ImageNet images and (b) synthetically corrupted ImageNet-C images as a function of patch size (given as a percentage of image area). In (b), each bar denotes the average accuracy over the five severities in ImageNet-C, and the horizontal dashed lines report the accuracy on the original (non-patched) datasets. Unadversarial patches consistently boost performance for both clean and corrupted images, with accuracy monotonically increasing with patch size. The patches were trained without any corruptions or non-standard data augmentation in-the-loop (we train with the same augmentations that the pre-trained model itself was trained with).

Robustness to synthetic corruptions. Next, we use the CIFAR-C and ImageNet-C datasets [HD19] (consisting of the CIFAR and ImageNet validation sets corrupted in 15 systematic ways) to see whether the addition of unadversarial patches to images confers any robustness to image corruptions.

We use the same patches and evaluation protocol that we used when looking at clean data (to ensure a fair evaluation, we apply corruptions to boosted images only *after* the unadversarial patches have been applied). As a consequence, at test time neither model nor patch has been exposed to any image corruptions beyond standard data augmentation. As a result, this experiment tests the ability for unadversarially boosted images to withstand completely unforeseen corruptions; we also avoid any potential biases from training on (and thus “overfitting” to [KSH+19]) a specific type of corruption. The results (cf. Figure 4b for

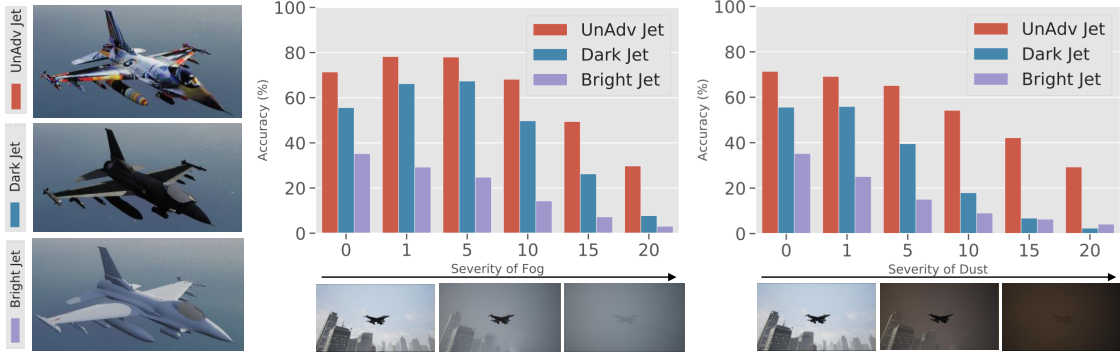


Figure 5: The jet unadversarial example task. We show example conditions under which we evaluate the objects, along with aggregate statistics for how well an ImageNet classifier classifies the objects in different conditions. We find that the classifiers perform consistently better on the unadversarial jet texture over the standard jet texture in both standard and distributionally shifted conditions. We also give a baseline of a white jet with a lighter texture because of the poorly visibility inherent in the simulator; we find it performed worse than even the standard jet.

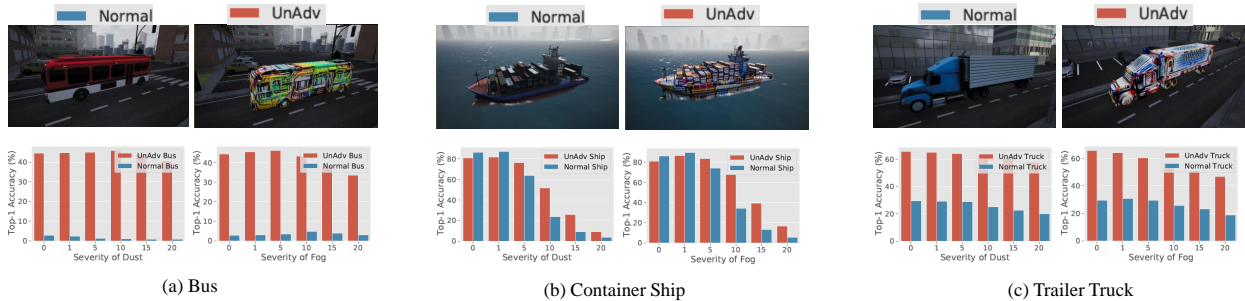


Figure 6: Additional examples reporting aggregate statistics for how well an ImageNet classifier classifies various objects in different conditions. Again, we find that the classifiers perform consistently better on the unadversarial objects texture over the standard objects.

ImageNet and Appendix D for CIFAR) indicate that unadversarial patches do improve performance across corruption types; for example, applying an unadversarial patch 5% the size of a standard ImageNet image boosts accuracy by an average of 31.7% points across corruptions ².

Baselines. We also compare our results to a variety of natural baselines; the most relevant of these is the “best loss image patch,” where we use the minimum-loss training image in place of a patch. We compare with this baseline to ensure that our method is doing something beyond this naive way to add signal to an image. The results are shown in Appendix D, along with comparisons to less sophisticated baselines, such as QR Codes and predefined random Gaussian noise patches.

3.2 Classification in 3D simulation

We now test unadversarial examples in a more practical setting: recognizing three-dimensional objects in a high-fidelity simulator. We collect a set of three-dimensional meshes corresponding to four ImageNet classes: “warplane”, “minibus”, “container ship”, and “trailer truck”, sourced from sketchfab.com. We generate a

²Since the original corruption benchmarks proposed by [HD19] are only available as pre-computed JPEGs (for which we cannot apply a patch pre-corruption) or CPU-based Python image operations (which were prohibitively slow), we re-implemented all 15 corruptions as batched GPU operations; we verified that model accuracies on our corruptions mirrored the original CPU counterparts (i.e., within 1% accuracy). For more details about our reimplementation, see our code release.

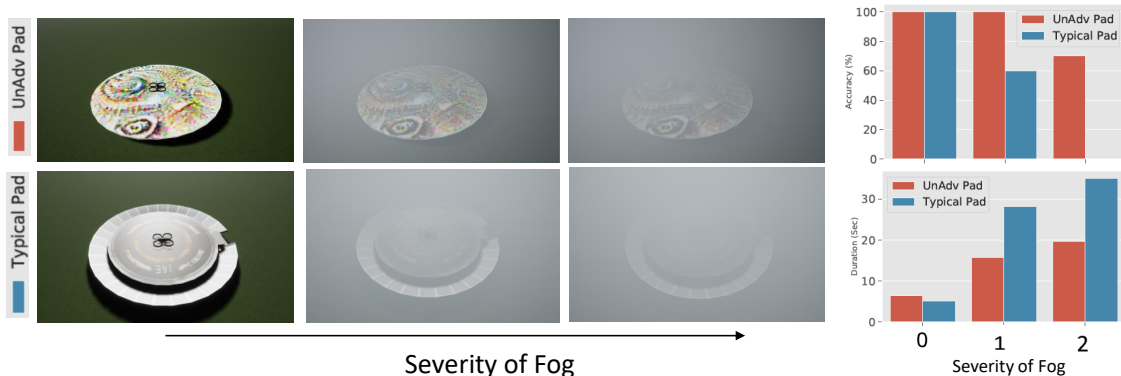


Figure 7: Drone landing task. On the left we show the unadversarial versus standard landing pads. On the right we show the results for the task when both the standard and unadversarial landing pads are used. We find that the drone consistently takes less time to land, and has a higher chance of landing correctly, when detecting the unadversarial landing pad.

texture for each object using the unadversarial texture algorithm outlined in Section 2.3, using the ImageNet validation set as the set of backgrounds for the algorithm, and a pre-trained ResNet-50 as the classifier.

To evaluate the resulting textures, we import each mesh into Microsoft AirSim, a high-fidelity three-dimensional simulator; we then test pre-trained ImageNet models’ ability to recognize each object with and without the unadversarial texture applied in a variety of surroundings. We also test each texture’s robustness to more realistic weather corruptions (snow and fog) built directly into the simulator (rather than applied as a post-processing step). We provide further detail on AirSim and our usage of it in Appendix B. Examples of the images used to evaluate the unadversarial textures, as well as our main results for one of the meshes are shown in Figure 5. We find that in both standard and simulated adverse weather conditions, the model consistently performs better on the mesh wrapped in the unadversarial texture than on the original. We present similar results for the other three meshes in Figure 6.

3.3 Localization for (Simulated) Drone Landing

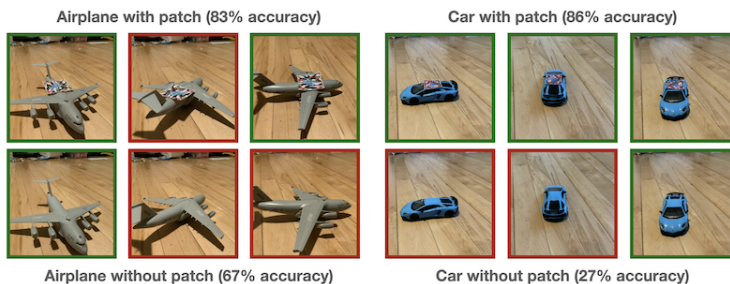
We then assess whether unadversarial examples can help outside of the classification setting. Again using AirSim, we set up a drone landing task with a perception module that receives as input an axis-aligned aerial image of a landing pad, and is tasked with outputting an estimate of the camera’s (x, y) -position relative to the pad. While this task is quite basic, we are particularly interested in studying performance in the presence of heavy (simulated) weather-based corruptions. The drone is equipped with a pretrained regression model that localizes the landing pad (described in detail in Appendix B). We optimize an unadversarial texture for the surface of the landing pad to best help the drone’s regression model in localization. Figure 7 shows an example of the landing pad localization task, along with the performance of the unadversarial landing pad compared to the standard pad. The drone landing on the unadversarial pad consistently lands both more reliably.

3.4 Physical World Unadversarial Examples

Finally, we move out of simulation and test whether the unadversarial patches that we generate can survive naturally-arising distribution shift from effects such as real lighting, camera artifacts, and printing imperfections. We use four household objects (a toy racecar, miniature plane, coffeepot, and eggnog container), and print out (on a standard InkJet printer) the adversarial patch corresponding to the label of each object. We take pictures of the toy with and without the patch taped on using an ordinary cellphone camera, and count the number of poses for which the toy is correctly classified by a pre-trained ImageNet classifier. Our results are in Table 8a, and examples of patches are in Figure 8b. Classifying both patched and unpatched images over a diverse set of poses, we find that the adversarial patches consistently improve performance even at uncommon object orientations.

Class	No Patch	Patch
“racer”	22%	83%
“eggnog”	22%	44%
“coffee pot”	39%	56%
“warplane”	67%	83%

(a) Accuracy of pre-trained ResNet-18 on photographs of real world objects with and without patches.



(b) Examples photos of the “warplane” and “racer” physical objects taken with (top) and without (bottom) an unadversarial patch.

Figure 8: Physical-world experiments. We take pictures of objects at diverse orientations while varying the presence of a patch on the object. Note that we don’t do any additional data augmentation on the patches, which are the same used in our previous ImageNet benchmark experiment.

4 Related Work

Here, we first highlight (and differentiate from) previous works using reference markers to improve recognition and localization. We will then discuss related work in adversarial robustness.

Improving computer vision with fiducial markers. In past research, vision-based precision landing was initially attempted with classical computer vision based tracking of reference designs [SMS03; LSP09; YSS+14; FZS+17], or on arrangements of fiducial markers [LZT+19] (a fiducial marker is a fixed pattern or object that is placed in a scene as a reference point for location or measure). Several types of visual fiducial markers were also proposed for robust tracking and pose estimation applications through classical vision [Fia05; GMM+14; Ols11; RMM19]. While fiducial markers such as AprilTags [Ols11] were commonly used in robotics, they suffer from limitations such as orientation uncertainty, accuracy falloff depending on viewing angles, and short detection ranges [AAB+19; WO16], which motivated research into using convolutional neural networks for landing pad detection and pose estimation [NAK+18; TNN+19]. Our work attempts to unify these perspectives by leveraging the expressivity of neural network-based systems to design robust unadversarial examples. Another key difference between fiducial markers and unadversarial objects is that the former require vision systems to be aware of their presence; in contrast, the latter are designed on top of pre-trained systems. This means that we do not require any further non-standard model training, and that we do not depend as heavily on the unadversarial example being in the field of view.

Adversarial robustness. Our design of unadversarial examples is motivated by the success of adversarial examples, i.e., minute perturbations to the inputs of machine learning models that can induce nearly arbitrary behavior. Adversarial examples were shown to be effective in seriously hampering machine vision related tasks such as classification [SZS+14; EEF+18a; AEI+18], object detection/segmentation [EEF+18b; AMT18; LK19; XWZ+17] and visual question-answering (VQA) [XCL+18]. Furthermore, prior work has shown that adversarial examples can be constructed even without direct access to the vision system being manipulated [CZS+17; PMG16; PMG+17; IEA+18; IEM18]. Synthesized physical adversarial examples were shown to be effective in fooling person detection [TVG19], sign detection for autonomous driving [SBM+18; BHG+19]. Robotic platforms such as manipulators were also shown to be sensitive to vision based adversarial examples [MDB+17], and to specifically designed adversarial objects [WTL+19]. Additionally, a recent line of work [EGS19; NHD+19] shows that one can “reprogram” neural networks using adversarial examples, e.g., one can construct an adversarial patch that when applied, causes a CIFAR-10 classifier to operate as an MNIST classifier.

5 Discussion and Conclusions

In this work, we demonstrated that it is possible to design object alternations that boost the corresponding classifiers’ performance, even under strong and *unforeseen* distribution shift. Indeed, such resulting unadversarial objects are robust to a broad range of data shifts and corruptions, even when these were never seen in training. We view our results as a promising route towards increasing out-of-distribution robustness of computer vision models.

Domains beyond image classification. The fact that unadversarial examples and adversarial examples share the same underlying generation technique is evidence that unadversarial examples could apply to any system that is vulnerable to adversarial examples. This indicates that unadversarial examples could apply to a wide variety of systems [CMV+16; SPV20; SCJ19; KFS18; ERL+18].

Extensions The method we present in this paper does not modify the underlying training procedure at all, or modify the network used in conjunction with the object. This presents an advantage in one sense, since we can use any pre-trained vision model, but it is possible that adding additional train time augmentations could make the unadversarial (object, model) pair more robust at completing the target task. For example, we could use data augmentation or train objects on the actual distribution shift we want to be robust to if it is specifiable.

Limitations One limitation of our method is that it requires differentiability with respect to properties of the input object of interest, or a close proxy (e.g. a differentiable simulator) that can mimick the object and the environment in which the object operates.

Acknowledgements

We are grateful to Ian Engstrom for helping take the photographs for the physical-world experiments.

Work supported in part by the NSF grants CCF-1553428 and CNS-1815221, and the Microsoft Corporation. This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001120C0015.

Research was sponsored by the United States Air Force Research Laboratory and was accomplished under Cooperative Agreement Number FA8750-19-2-1000. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the United States Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

- [AAB+19] Syed Muhammad Abbas, Salman Aslam, Karsten Berns, and Abubakr Muhammad. “Analysis and Improvements in AprilTag Based State Estimation”. In: *Sensors* (2019).
- [AEI+18] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. “Synthesizing Robust Adversarial Examples”. In: *International Conference on Machine Learning (ICML)*. 2018.
- [ALG+19] Michael A Alcorn, Qi Li, Zhitao Gong, Chengfei Wang, Long Mai, Wei-Shinn Ku, and Anh Nguyen. “Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects”. In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.
- [AMT18] Anurag Arnab, Ondrej Miksik, and Philip HS Torr. “On the robustness of semantic segmentation models to adversarial attacks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 888–897.
- [BCM+13] Battista Biggio, Iginio Corona, Davide Maiorca, Blaine Nelson, Nedim Šrđić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. “Evasion attacks against machine learning at test time”. In: *Joint European conference on machine learning and knowledge discovery in databases (ECML-KDD)*. 2013.
- [BHG+19] Adith Bloor, Xin He, Christopher Gill, Yevgeniy Vorobeychik, and Xuan Zhang. “Simple physical adversarial examples against end-to-end autonomous driving models”. In: *2019 IEEE International Conference on Embedded Software and Systems (ICES)*. 2019.
- [BMR+18] Tom B. Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. *Adversarial Patch*. 2018. arXiv: [1712.09665](https://arxiv.org/abs/1712.09665) [cs.CV].
- [CMV+16] Nicholas Carlini, Pratyush Mishra, Tavish Vaidya, Yuankai Zhang, Micah Sherr, Clay Shields, David Wagner, and Wenchao Zhou. “Hidden Voice Commands”. In: *USENIX Security Symposium*. 2016.
- [CZS+17] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. “Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models”. In: *Workshop on Artificial Intelligence and Security*. 2017.
- [EEF+18a] Ivan Evtimov, Kevin Eykholt, Earlene Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, and Dawn Song. “Robust Physical-World Attacks on Machine Learning Models”. In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018.
- [EEF+18b] Kevin Eykholt, Ivan Evtimov, Earlene Fernandes, Bo Li, Amir Rahmati, Florian Tramer, Atul Prakash, Tadayoshi Kohno, and Dawn Song. “Physical Adversarial Examples for Object Detectors”. In: *CoRR* (2018).
- [EGS19] Gamaleldin Elsayed, Ian Goodfellow, and Jascha Sohl-Dickstein. “Adversarial Reprogramming of Neural Networks”. In: *International Conference on Learning Representations (ICLR)*. 2019.
- [EIS+20] Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Jacob Steinhardt, and Aleksander Madry. “Identifying Statistical Bias in Dataset Replication”. In: *International Conference on Machine Learning (ICML)*. 2020.
- [ERL+18] Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. “HotFlip: White-Box Adversarial Examples for Text Classification”. In: *Association for Computational Linguistics (ACL, Short Papers)*. 2018.
- [ETT+19] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. “Exploring the Landscape of Spatial Robustness”. In: *International Conference on Machine Learning (ICML)*. 2019.
- [FF15] Alhussein Fawzi and Pascal Frossard. “Manitest: Are classifiers really invariant?” In: *British Machine Vision Conference (BMVC)*. 2015.
- [Fia05] Mark Fiala. “ARTag, a fiducial marker system using digital techniques”. In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*. 2005.

- [FZS+17] Davide Falanga, Alessio Zanchettin, Alessandro Simovic, Jeffrey Delmerico, and Davide Scaramuzza. "Vision-based autonomous quadrotor landing on a moving platform". In: *2017 IEEE International Symposium on Safety, Security and Rescue Robotics (SSRR)*. 2017.
- [GMM+14] Sergio Garrido-Jurado, Rafael Muñoz-Salinas, Francisco José Madrid-Cuevas, and Manuel Jesús Marin-Jiménez. "Automatic generation and detection of highly reliable fiducial markers under occlusion". In: *Pattern Recognition* (2014).
- [GRM+19] 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.
- [HBM+20] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. *The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization*. 2020. arXiv: [2006.16241](https://arxiv.org/abs/2006.16241) [cs.CV].
- [HD19] Dan Hendrycks and Thomas G. Dietterich. "Benchmarking Neural Network Robustness to Common Corruptions and Surface Variations". In: *International Conference on Learning Representations (ICLR)*. 2019.
- [IEA+18] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. "Black-box Adversarial Attacks with Limited Queries and Information". In: *International Conference on Machine Learning (ICML)*. 2018.
- [IEM18] Andrew Ilyas, Logan Engstrom, and Aleksander Madry. "Prior Convictions: Black-box Adversarial Attacks with Bandits and Priors". In: *International Conference on Learning Representations (ICLR)*. 2018.
- [IST+19] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. "Adversarial Examples Are Not Bugs, They Are Features". In: *Neural Information Processing Systems (NeurIPS)*. 2019.
- [JBZ+19] Jorn-Henrik Jacobsen, Jens Behrmann, Richard Zemel, and Matthias Bethge. "Excessive Invariance Causes Adversarial Vulnerability". In: *International Contemporary on Learning Representations*. 2019.
- [JLT18] Saumya Jetley, Nicholas Lord, and Philip Torr. "With friends like these, who needs adversaries?" In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2018.
- [KFS18] Jernej Kos, Ian Fischer, and Dawn Song. "Adversarial examples for generative models". In: *IEEE Security and Privacy Workshops (SPW)*. 2018.
- [KGB17] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. "Adversarial Machine Learning at Scale". In: *International Conference on Learning Representations (ICLR)*. 2017.
- [KMF18] Can Kanbak, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. "Geometric robustness of deep networks: analysis and improvement". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018.
- [Kri09] Alex Krizhevsky. "Learning Multiple Layers of Features from Tiny Images". In: *Technical report*. 2009.
- [KSH+19] Daniel Kang, Yi Sun, Dan Hendrycks, Tom Brown, and Jacob Steinhardt. "Testing Robustness Against Unforeseen Adversaries". In: *ArXiv preprint arxiv:1908.08016*. 2019.
- [LK19] Mark Lee and Zico Kolter. "On physical adversarial patches for object detection". In: *arXiv preprint arXiv:1906.11897* (2019).
- [LSP09] Sven Lange, Niko Sunderhauf, and Peter Protzel. "A vision based onboard approach for landing and position control of an autonomous multirotor UAV in GPS-denied environments". In: *2009 International Conference on Advanced Robotics*. 2009.

- [LZT+19] Xuancen Liu, Shifeng Zhang, Jiayi Tian, and Longbin Liu. “An Onboard Vision-Based System for Autonomous Landing of a Low-Cost Quadrotor on a Novel Landing Pad”. In: *Sensors* (2019).
- [MDB+17] Marco Melis, Ambra Demontis, Battista Biggio, Gavin Brown, Giorgio Fumera, and Fabio Roli. “Is deep learning safe for robot vision? adversarial examples against the icub humanoid”. In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2017, pp. 751–759.
- [MMS+18] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. “Towards deep learning models resistant to adversarial attacks”. In: *International Conference on Learning Representations (ICLR)*. 2018.
- [NAK+18] Phong Ha Nguyen, Muhammad Arsalan, Ja Hyung Koo, Rizwan Ali Naqvi, Noi Quang Truong, and Kang Ryoung Park. “LightDenseYOLO: A fast and accurate marker tracker for autonomous UAV landing by visible light camera sensor on drone”. In: *Sensors* (2018).
- [Nes03] Yurii Nesterov. *Introductory Lectures on Convex Optimization*. 2003.
- [NHD+19] Paarth Neekhara, Shehzeen Hussain, Shlomo Dubnov, and Farinaz Koushanfar. “Adversarial Reprogramming of Text Classification Neural Networks”. In: *Empirical Methods in Natural Language Processing (EMNLP)*. 2019.
- [NVZ+19] Merlin Nimier-David, Delio Vicini, Tizian Zeltner, and Wenzel Jakob. “Mitsuba 2: A Retargetable Forward and Inverse Renderer”. In: *Transactions on Graphics (Proceedings of SIGGRAPH Asia)* 38.6 (2019). DOI: [10.1145/3355089.3356498](https://doi.org/10.1145/3355089.3356498).
- [Ols11] Edwin Olson. “AprilTag: A robust and flexible visual fiducial system”. In: *2011 IEEE International Conference on Robotics and Automation*. 2011.
- [PMG+17] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. “Practical black-box attacks against machine learning”. In: *Asia Conference on Computer and Communications Security*. 2017.
- [PMG16] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. “Transferability in Machine Learning: from Phenomena to Black-box Attacks using Adversarial Samples”. In: *ArXiv preprint arXiv:1605.07277*. 2016.
- [RDS+15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. “ImageNet Large Scale Visual Recognition Challenge”. In: *International Journal of Computer Vision (IJCV)*. 2015.
- [RMM19] Francisco J Romero-Ramire, Rafael Muñoz-Salinas, and Rafael Medina-Carnicer. “Fractal Markers: a new approach for long-range marker pose estimation under occlusion”. In: *IEEE Access* (2019).
- [RRS+19] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. “Do ImageNet Classifiers Generalize to ImageNet?” In: *International Conference on Machine Learning (ICML)*. 2019.
- [RXY+19] Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John C Duchi, and Percy Liang. “Adversarial Training Can Hurt Generalization”. In: *arXiv preprint arXiv:1906.06032* (2019).
- [SBB+16] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K. Reiter. “Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition”. In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, October 24–28, 2016*. 2016, pp. 1528–1540.
- [SBM+18] Chawin Sitawarin, Arjun Nitin Bhagoji, Arsalan Mosenia, Mung Chiang, and Prateek Mittal. “Darts: Deceiving autonomous cars with toxic signs”. In: *arXiv preprint arXiv:1802.06430* (2018).
- [SCJ19] Octavian Suciuc, Scott E. Coull, and Jeffrey Johns. “Exploring Adversarial Examples in Malware Detection”. In: *IEEE Security and Privacy Workshops (SPW)*. 2019.

- [SIE+20] Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. “Do Adversarially Robust ImageNet Models Transfer Better?” In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2020.
- [SMS03] Srikanth Saripalli, James F Montgomery, and Gaurav S Sukhatme. “Visually guided landing of an unmanned aerial vehicle”. In: *IEEE transactions on robotics and automation* (2003).
- [SPV20] Alex Serban, Erik Poll, and Joost Visser. “Adversarial Examples on Object Recognition: A Comprehensive Survey”. In: *ACM Computing Surveys (CSUR)*. 2020.
- [SZC+18] Dong Su, Huan Zhang, Hongge Chen, Jinfeng Yi, Pin-Yu Chen, and Yupeng Gao. “Is Robustness the Cost of Accuracy? A Comprehensive Study on the Robustness of 18 Deep Image Classification Models”. In: *European Conference on Computer Vision (ECCV)*. 2018.
- [SZS+14] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. “Intriguing properties of neural networks”. In: *International Conference on Learning Representations (ICLR)*. 2014.
- [TDS+20] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. “Measuring Robustness to Natural Distribution Shifts in Image Classification”. In: *Neural Information Processing Systems (NeurIPS)*. 2020.
- [TE11] Antonio Torralba and Alexei A Efros. “Unbiased look at dataset bias”. In: *CVPR 2011*. 2011.
- [TNN+19] Noi Quang Truong, Phong Ha Nguyen, Se Hyun Nam, and Kang Ryoung Park. “Deep learning-based super-resolution reconstruction and marker detection for drone landing”. In: *IEEE Access* (2019).
- [TSE+19] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. “Robustness May Be at Odds with Accuracy”. In: *International Conference on Learning Representations (ICLR)*. 2019.
- [TVG19] Simen Thys, Wiebe Van Ranst, and Toon Goedemé. “Fooling automated surveillance cameras: adversarial patches to attack person detection”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2019.
- [WO16] John Wang and Edwin Olson. “AprilTag 2: Efficient and robust fiducial detection”. In: *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2016.
- [WTL+19] D. Wang, D. Tseng, P. Li, Y. Jiang, M. Guo, M. Danielczuk, J. Mahler, J. Ichnowski, and K. Goldberg. “Adversarial Grasp Objects”. In: *2019 IEEE 15th International Conference on Automation Science and Engineering (CASE)*. 2019.
- [XCL+18] Xiaojun Xu, Xinyun Chen, Chang Liu, Anna Rohrbach, Trevor Darrell, and Dawn Song. “Fooling vision and language models despite localization and attention mechanism”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 4951–4961.
- [XWZ+17] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. “Adversarial examples for semantic segmentation and object detection”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 1369–1378.
- [YSS+14] Shaowu Yang, Sebastian A Scherer, Konstantin Schauwecker, and Andreas Zell. “Autonomous landing of MAVs on an arbitrarily textured landing site using onboard monocular vision”. In: *Journal of Intelligent & Robotic Systems* (2014).

B 3D Simulation Details

B.1 Overview of AirSim

We conduct our simulation experiments using the high fidelity simulator, Microsoft AirSim. AirSim acts as a plugin to Unreal Engine, which is a AAA videogame engine providing access to high fidelity graphics features such as high resolution textures, realistic lighting, soft shadows etc. making it a good choice for rendering for computer vision applications. AirSim internally provides physics models for a quadrotor vehicle, which we leverage for performing autonomous drone landing. As a plugin, AirSim can be paired with any Unreal Engine environment to simulate autonomous vehicles that can be programmed with an API both in terms of planning/control as well as obtaining camera images. AirSim also allows for controlling environmental features such as time of day, dynamically adding/removing objects, changing object textures and so on.

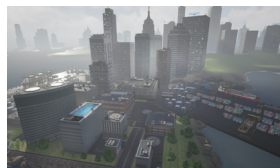
B.2 3D Boosters Classification Experiment

Format of 3D models To evaluate the performance of pretrained ImageNet classifiers at detecting 3D unadversarial/boosted objects (e.g. the jet shown in the main paper) among realistic settings, we set up an experiment using AirSim for image classification of common classes (warplane, car, truck, ship, etc). We pick the class of ‘warplane’ as our object class of interest download publicly available 3D meshes for this class from www.sketchfab.com. Using the open source 3D modeling software Mitsuba, we modify the object texture to match the boosted texture for the corresponding class, and then export these meshes into the GLTF format for ingestion into Unreal Engine/AirSim. This allows us to import the boosted objects into the AirSim framework, and spawn them as objects in any of the environments being created.

Environment screenshots and description Within AirSim, in the interest of generating realistic imagery, we simulate a city environment (Figure 9a). For this experiment, we use the ComputerVision mode of AirSim, which does not simulate a vehicle but rather, gives the user control of a free moving camera, allowing us to generate data at ease from various locations and varying camera and world parameters.

Sampling and evaluation Once the 3D objects (unadversarial or normal) are present in AirSim’s simulated world, the next step is to evaluate the classification of these objects from different camera angles, weather conditions etc. Given the location of a candidate object (which we randomize and average over five locations), we sample a grid ($10 \times 10 \times 10$) of camera positions in 3D around the object. For each of these positions, we move AirSim’s main camera and orient it towards the object, resulting in images from various viewpoints. At runtime, each of these images are immediately processed by a pretrained ResNet-18 ImageNet classifier, which reports the top 5 class predictions. We average the accuracies across the five different locations in the scene and the 1000 grid points around the object at each location.

Along with this variation in camera angles and thereby, object pose in the frame; we also evaluate the performance of of the various 3D objects given environmental perturbations. We achieve this through the AirSim’s weather conditions feature, using which we simulate weather conditions such as dust and fog dynamically with varying levels of severity of these conditions. We will open-source binaries for the AirSim code and environments that we use which will allow people to replicate our results, and investigate more scenarios of interest.



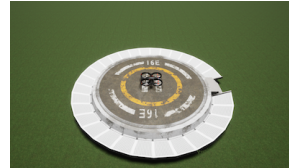
(a) City environment in AirSim used for detection experiment



(b) Boosted ‘jet’ model in the City environment.



(c) Sample landing pads atop buildings in the City environment.



(d) Drone in test environment used for the landing experiments.

B.3 Drone Landing Experiment

In this experiment, we evaluate how unadversarial/boosted objects can help robustify perception-action loops that are driven by vision-based pose estimation. Perception-action loops are at the heart of many robotics tasks, and accurate perception is imperative for safe, efficient navigation of robots. We choose the scenario of autonomous drone landing as our experiment, and simulate it within AirSim.

For this experiment, we create assets of landing pads that are similar to helipads on top of buildings in the city environment (Figure 9c). We also use a test environment with a single landing pad located on a patch of grass. An example of such a landing pad can be seen in Figure 9d. We use AirSim to simulate a quadrotor drone in these worlds, which can be programmatically controlled using a Python API. AirSim allows us to equip a downward facing, gimballed camera on this drone in order to obtain RGB images, which are then processed by our landing pad pose estimation (regression) model. Given an RGB image, the regression model outputs a 6 degree of freedom pose for the landing pad. We use/optimize only the first two entries of this output corresponding to the relative x and y location of the landing pad w.r.t the drone.

We formulate the drone landing experiment as a visual servoing task: a perception action loop that involves estimating the relative location of the pad from the image frame captured by the downward facing camera of the drone, and sending an appropriate velocity command in order to align the camera center with that of the pad. We achieve these through the following steps:

Data Collection. We use AirSim’s inbuilt data collection API for this step. Given the location of the pad in the world, we sample various feasible locations for the drone in an imaginary cone whose vertex aligns with the center of the landing pad. We then spawn the drone in these randomly sampled positions, and obtain the RGB and segmentation views of the pad as generated by AirSim, along with the relative ground truth position of the landing pad w.r.t the drone, and repeat this process to create a dataset. The collected dataset contains 20000 images and is split 80-20 between train and evaluation sets.

Landing pad pose estimator. We train a model that maps top view images of a scene with a landing pad, to the relative 2D location of the landing pad w.r.t the drone in the camera frame. We use a ResNet-18 pretrained on ImageNet as the backbone for the pose regressor, and we replace the last classification layer with a regression layer outputting the (x, y) relative location of the pad w.r.t drone. The model is trained end-to-end by minimizing the mean squared error (MSE) loss between the predicted location and the ground truth location. The ground truth is collected along with the images using the AirSim City simulation environment as describe before.

We train the model for 10 epochs using SGD with a fixed learning rate of 0.001, a batch size of 512, a weight decay of $1e-4$, and with MSE as the objective function. The model converges fairly quickly (within the first few epochs).

Drone Landing. To use the pose estimator’s predictions and send appropriate actions, we utilize the Multirotor API of AirSim. This allows us to control the drone by setting the desired velocity commands along all the axes (translational/rotational). Given the position of the landing pad in the scene relative to that of the drone (as output by the pose regressor) we execute the landing operation by sending appropriate velocity commands to the drone.

To generate the right velocity commands, given the relative position of the landing pad, we use a standard PID controller that computes corrective velocity values until the position of the drone matches that of the landing pad. For a pose output by the regressor treated as the setpoint P_{set} , and current drone pose P_{curr} and at any point at time t , the appropriate velocity command $v(t)$ can simply be computed as follows:

$$\begin{aligned} e(t) &= P_{set} - P_{curr} \\ v(t) &= K_p e(t) + K_d \frac{d}{dt} e(t) + K_i * \int_0^t e(t) dt \end{aligned}$$

where $K_p, K_d, and K_i$ are the hyperparameters of the PID controller and are manually tuned. We find that $K_p = K_d = 5$ and $K_i = 0$ to be reasonable for our task.

For realistic perturbations to the scene, similar to the 3D boosters classification experiment, we continue making use of the weather API to generate weather conditions in AirSim. This results in a variation of factors such as amount of dust or fog in the scene, allowing us to evaluate the performance of landing under various realistic conditions.

C Experimental Setup

C.1 Pretrained vision models we evaluate

Here we present details of the different vision models we use in our paper. For more details on all of these, please check the README of our code at <https://git.io/unadversarial>.

Corruption benchmark experiments: We use pretrained ResNet-18 and ResNet-50 (both standard and ℓ_2 -robust with $\epsilon = 3$) architectures from [SIE+20]: <https://github.com/microsoft/robust-models-transfer>.

3D object classification in AirSim: We use an ImageNet pretrained ResNet-18 architecture from the PyTorch’s Torchvision³ to classify all the boosted and non-boosted versions of the jets, cars, ships etc in AirSim.

Drone landing experiment in AirSim: We finetune an ImageNet pretrained ResNet-18 model on the regression task of drone landing. The last layer of the pretrained model is replaced with a 2D linear layer estimating the relative pad location w.r.t the drone. We collect a 20k sample dataset for training the pad pose estimation in AirSim with an 80 – 20 train-val split. We use a learning rate of 0.001, a batch size of 512, a weight decay of $1e - 4$. We train for 10 epochs.

Physical world unadversarial examples experiment: Similar to the 3D object classification experiment in AirSim, we use an ImageNet pretrained ResNet-18 architecture from Torchvision.

C.2 Unadversarial patch/texture training details

Patches training details We fix the training procedure for all of the 2D patches we optimize in our paper. We train all the patches starting from random initialization with batch size of 512, momentum of 0.9, and weight decay of $1e - 4$. We train all the patches for 30 epochs (which is more than enough as we observe that for both ImageNet and CIFAR-10, the patch converges within the first 10 epochs) with a learning rate of 0.1. We sweep over three learning rates $\in \{0.1, 0.01, 0.001\}$ but we find that all of these obtain very similar results. So we stick with a learning rate of 0.1 for all of our experiments..

For the classification tasks (i.e., everything but drone landing) we use the standard cross-entropy loss. For the drone landing task (landing pad pose estimation), we use the standard mean squared error loss.

Texture training details We now outline the process for constructing adversarial textures. We implemented a custom PyTorch module with a distinct forward and backward pass; on the forward pass (i.e., during evaluation), the module takes as input an ImageNet image, and a 200px by 200px texture; using the Python bindings for Mitsuba [NVZ+19] 3D renderer, the module returns a rendering of the desired 3D object, overlaid onto the given ImageNet image. On the backwards pass (i.e., when computing gradients), we use the 3D model’s UV map⁴—a linear transformation from (x, y) locations on the texture to (x, y) locations in the rendered image—to approximate gradients through the rendering process. This is the same procedure used by [AEI+18] for constructing physical adversarial examples. Note that this is a simple approximation that only accounts for the location of pixels in the rendered image (i.e., ignores the effects of lighting, warping, etc.). However,

C.3 Details of the physical world experiment

To conduct the physical-world experiments, we used a toy racecar⁵, a toy warplane⁶ (both from [amazon.com](https://www.amazon.com)) as well as two household objects: a coffeepot and eggnog container. We then printed the unadversarial patches corresponding to classes “racer,” “warplane,” “coffeepot,” and “eggnog” on an HP DeskJet 2700

³These models can be found here <https://pytorch.org/docs/stable/torchvision/models.html>

⁴Mitsuba provides direct access to the UV map through the aov integrator; see the code release for more details.

⁵<https://www.amazon.com/gp/product/B07T5X69TZ/>

⁶<https://www.amazon.com/CORPER-TOYS-Pull-Back-Aircraft-Birthday/dp/B07DB3839X/>



Figure 10: Photographs in different poses of the four physical objects we experimented on, with and without an unadversarial patch.

InkJet printer, at 250% scale. We adhere the patches to the top of their respective objects with clear tape (the results are shown in Figure 8b). We choose 18 distinct poses (camera positions), and for each pose took one picture of the object with the patch attached, and one picture without (keeping the location of the patch constant throughout the experiment). Example photographs are shown in Figure 10. We evaluated a pre-trained ResNet-18 classifier on the resulting images.

C.4 Replicate our results

We desired simplicity and kept reproducibility in our minds when conducting our experiments, so we use standard hyperparameters and minimize the number of tricks needed to replicate our results. Our code is available at <https://git.io/unadversarial>.

D Omitted Results

In the below figure, we show a more detailed look of the main results of the benchmarking experiments in our paper, along with useful baselines. The single color plots (e.g. the left subplot in Figure 11) report the average performance over the 5 severities of ImageNet-C and CIFAR-10-C. The multicolor bar plots (e.g. the right subplot in Figure 11) report the detail performance per severity level. The horizontal dashed lines report the performance of the pretrained models on the original (non-patched) ImageNet-C and CIFAR-10-C datasets and serve as a baseline to compare with. For both ImageNet and CIFAR as shown in Figure 12 and Figure 11, we are able to train unadversarial patches of various size that, once overlaid on the datasets, make the pretrained model consistently much more robust under all corruptions.

D.1 Corruption benchmark main results: additional results to Figure 4b

Here we show the detailed main results for boosting ImageNet and CIFAR-10 with unadversarial patches.

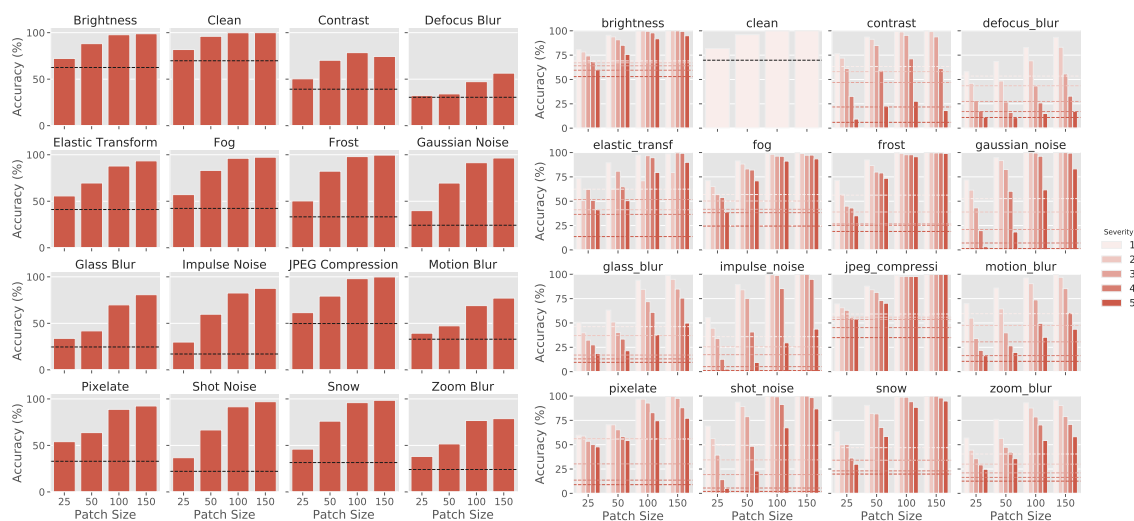


Figure 11: Robustness of a trained 2D booster over pretrained ImageNet ResNet-18 model.

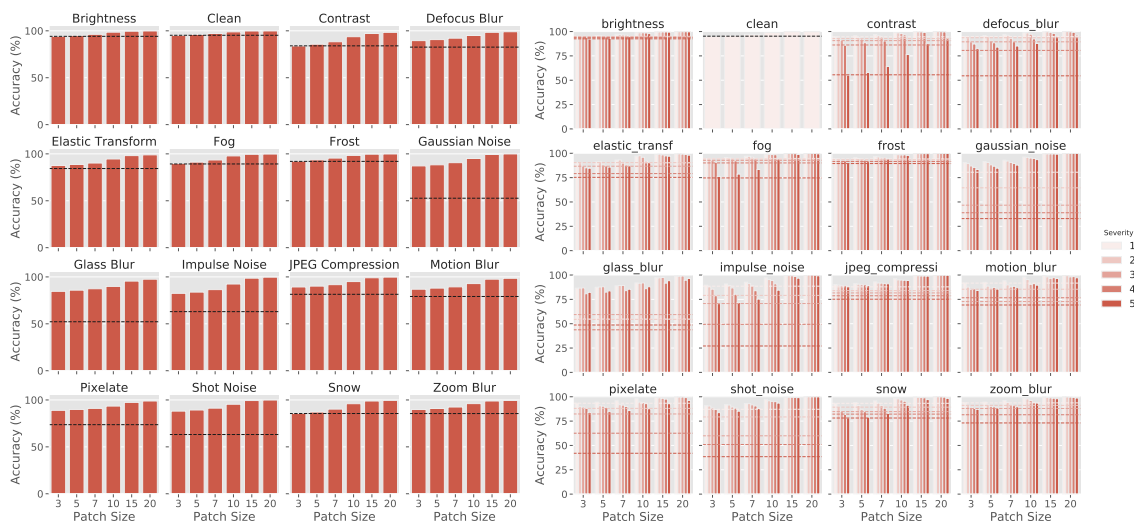


Figure 12: Robustness of a trained 2D booster over pretrained CIFAR-10 ResNet-50 model.

D.2 Baselines

Below, we report a number of alternative ways to create patches for boosting the performance of object recognition.

D.2.1 QR-Code

We compare our unadversarial patches to the well-known QR-Code patches. We create a QR-Code for each class of the ImageNet dataset using Python’s qrcode package (we avoid using CIFAR-10 since the images are too small for QR-Codes to be visible and detected at all). We overlay the QR-Codes over the ImageNet validation set according to what label each image has. We add the various ImageNet-C corruption on top of the resulting images, then we use python’s Pyzbar⁷ package to detect the QR-Codes. The results are shown in Figure 13. The performance of QR-Codes is not comparable to what we obtain with unadversarial patches (see Figure 4b).

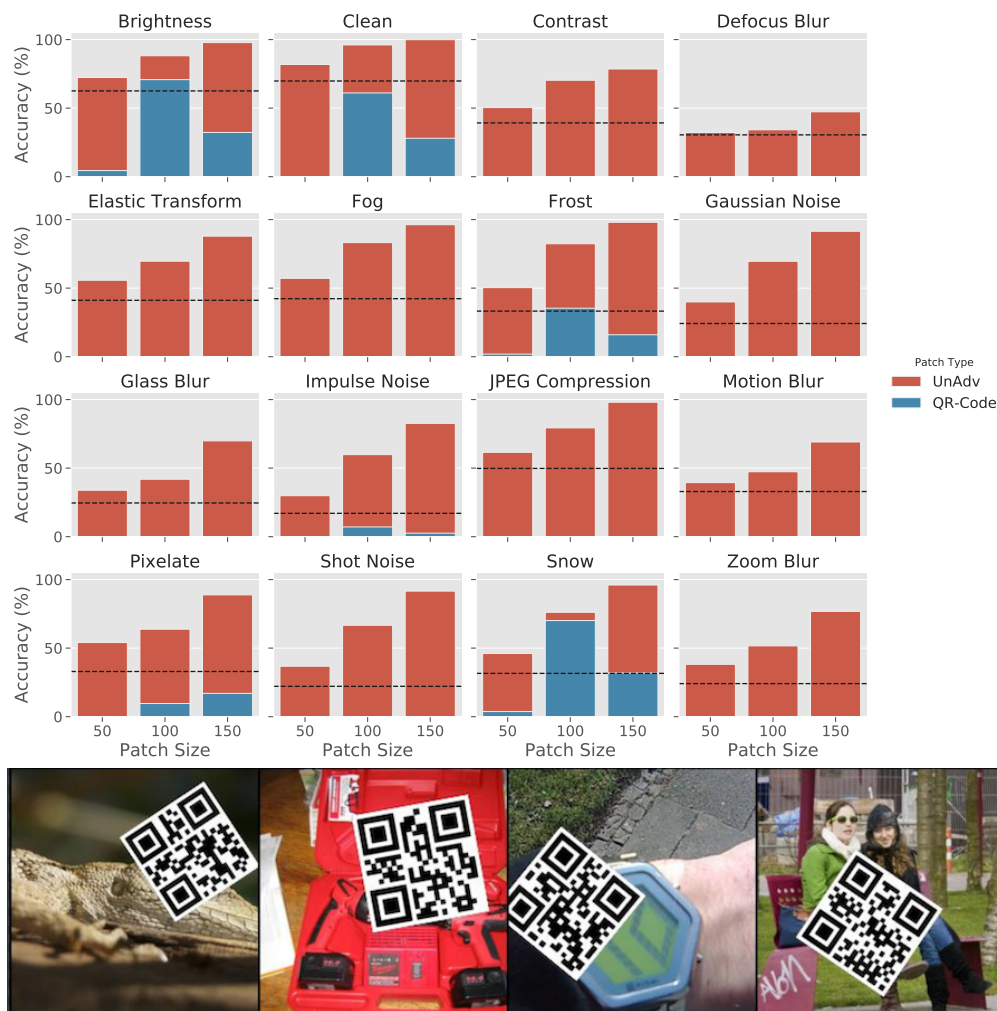


Figure 13: QR-Code boosted ImageNet results under various corruptions.

⁷We experiment with OpenCV for detecting the QR-Codes but find that Pyzbar leads to better performance.

D.2.2 Best training image per class as patch

Another natural baseline that we compare with is using the best images per class in the training set of the task of interest as patches for boosting the performance of pretrained models. For example, for ImageNet classification, we simply evaluate the loss of each training image using a pretrained ImageNet model (ResNet-18 in our case), and we take the image with the lowest loss per class as the patch for that class. Now we overlay these found patches onto the ImageNet validation set with random scaling, rotation, and translation (as shown in Figure 14), we add ImageNet-C corruptions, and we evaluate this new dataset using the same pretrained model we used to extract the patches. The results for ImageNet and CIFAR-10 are shown in Figure 14 and Figure 15, respectively.

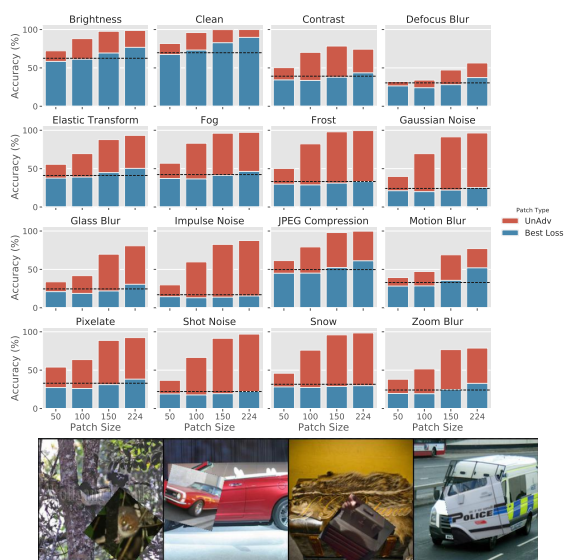


Figure 14: Best training image with translation, rotation, and scaling for ImageNet.

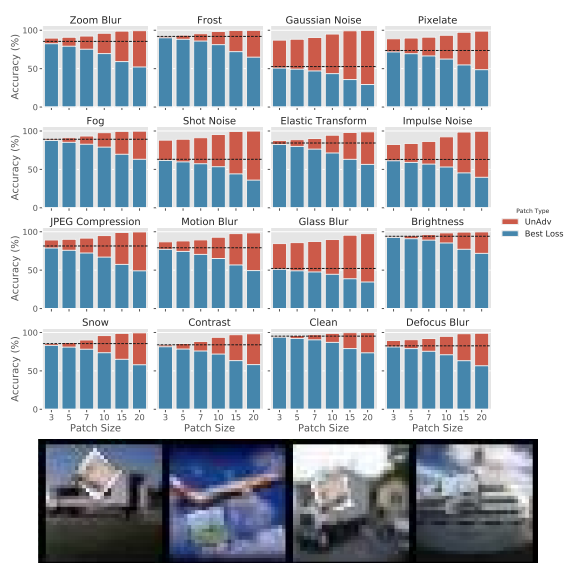


Figure 15: Best training image with translation, rotation, and scaling for CIFAR-10.

D.2.3 Best training image vs random training image as patch

Here we investigate whether using a random image from the training set does any better than using the best-loss image as a patch. The results are shown in the below Figures. As one would expect, using a random image from the training set leads to strictly worse performance. This holds for both ImageNet and CIFAR-10 as shown in Figure 16 and Figure 17.

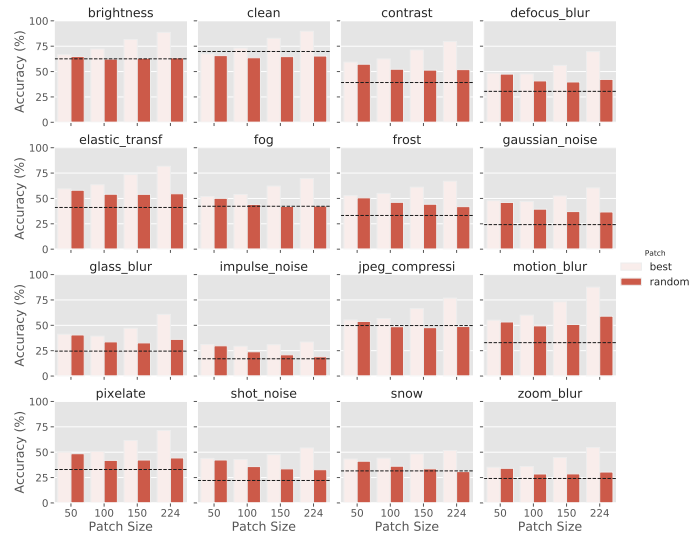


Figure 16: Best training image vs random training image with translation, rotation, and scaling.

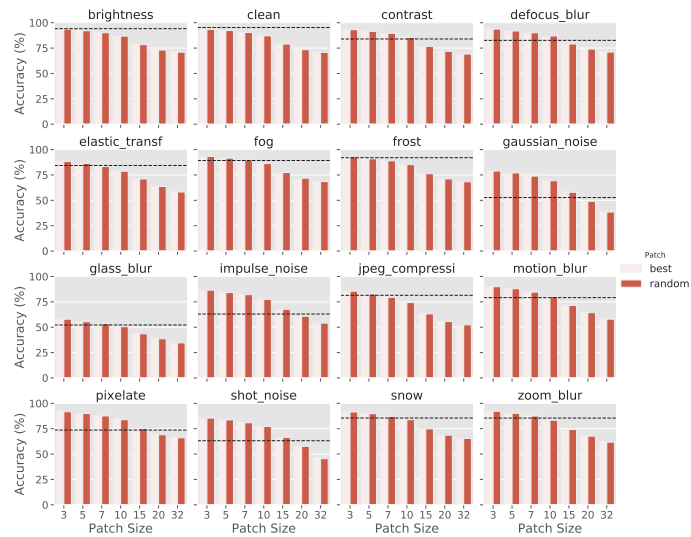


Figure 17: Best training image vs random training image with translation, rotation, and scaling.

D.2.4 Predefined fixed-pattern unadversarial patches

This baseline is slightly different than the previous baselines since it allows the underlying classification model to be changed. Basically, we fix the a set of patches to predefined pattern (here a fixed random gaussian noise for each class), and we train a classifier on an undversarial/boosted dataset with these patches. The resulting models are consistently weaker on all corruptions of ImageNet-C and CFAR-10-C as shown in Figure 18 and Figure 19 compared to our trained patches the main paper in Figure 4b.

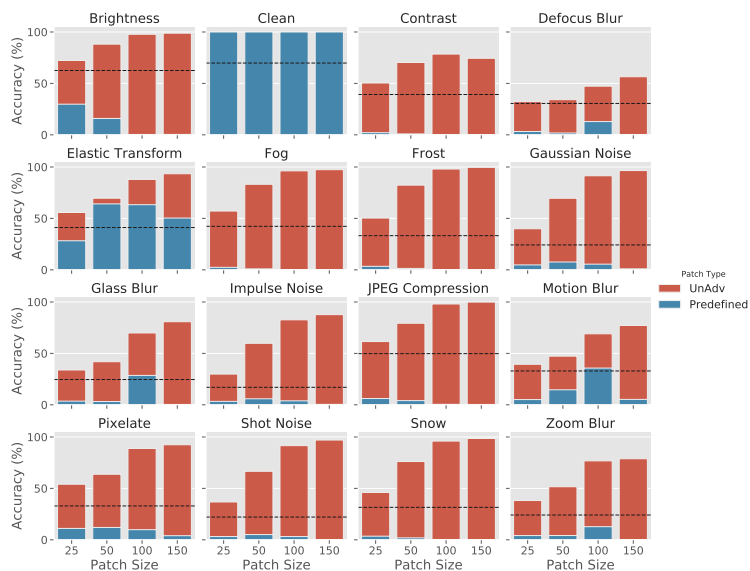


Figure 18: Robustness of an ImageNet ResNet-18 model trained on a predefined patch.

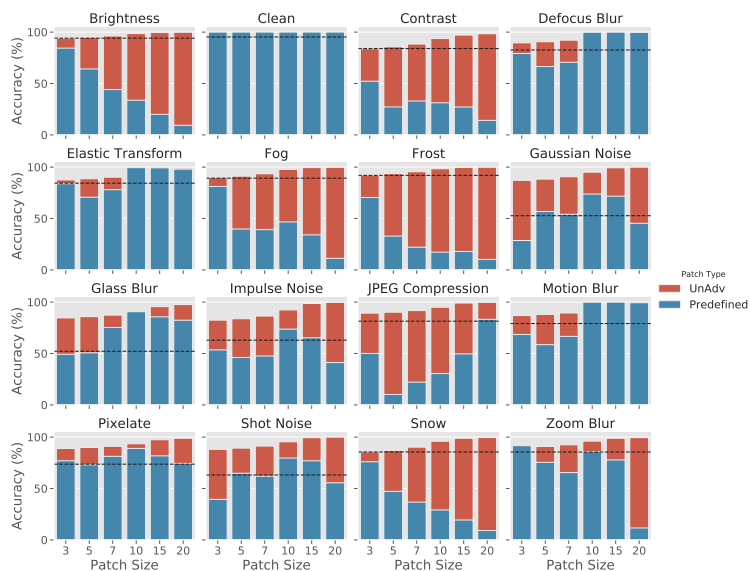


Figure 19: Robustness of a CIFAR-10 ResNet-50 model trained on a predefined patch.