[Re] Diffusion-Based Adversarial Sample Generation for Improved Stealthiness and Controllability

William Kang
wkang01@student.ubc.ca

Christina Yang
chryang@student.ubc.ca

Abstract

We are doing a reproducibility report based on the paper "Diffusion-Based Adversarial Sample Generation for Improved Stealthiness and Controllability" by Haotian Xue, Alexandre Araujo, Bin Hu, Yongxin Chen. The paper uses a novel framework to generate adversarial samples. They use a gradient based method guided by a pre-trained diffusion model to try to generate images that appear realistic to the human eye, can fool a wide range of models, and is easy to control how certain regions are modified. We find that attacks have to be carefully created, with the right hyper-parameters, and target images, in order to get results similar to the original paper.

1 Introduction

Neural networks are remarkably good at tasks such as image classification. However, they are susceptible to attacks that perturb images in subtle ways, causing the model to classify one image as something else entirely. These perturbed samples are called adversarial samples, and are a major security concern for systems that use neural networks. There are various types of adversarial attacks, including global attacks, which perturb any part of the image, regional attacks, which change a specified region of an image, style-based attacks, which change an image region based on a reference image, and physical world attacks, which generates images you can print out and place on a physical object which cause the physical object to be misclassified by an image classifier.

The authors of the original paper propose a new method for generating adversarial models, that uses an off-the-shelf diffusion model to help guide the generation of the perturbations in the image, called Diff-PGD (Xue et al. 2023). This method aims to create adversarial samples that are more realistic, and harder to detect by the human eye, compared to baseline methods, and can be applied to various types of attacks, such as digital attacks, physical-world attacks, and style-based attacks.

2 Scope of reproducibility

The main claims from the original paper are as follows:

1. Diff-PGD can be applied to specific tasks such as digital attacks, physical-world attacks, and style-based attacks, outperforming baseline methods such as PGD (Madry et al. 2019), AdvPatch (Brown et al. 2018), and AdvCam (Lee, Kim, and Yoon 2021).
2. Diff-PGD is more stable and controllable compared to existing methods for generating natural-style adversarial samples.
3. Diff-PGD surpasses the original PGD in Transferability and Purification power
4. Diff-PGD generates adversarial samples with higher stealthiness

We explore the first claim in regards to physical-world and style-based attacks specifically, and test these attacks on a broader range of images.

Course project for UBC CPSC 440/550, April 2024.

3 Methodology

3.1 Model descriptions

We used ResNet-50 (He et al. 2016) as our primary classifier. We used the pre-trained models with their default weights.

3.2 Datasets

The original paper used ImageNet, but due to limitations in compute and memory, we used a smaller subset of ImageNet with 1000 samples: `https://www.kaggle.com/datasets/ifigotin/imagenetmini-1000`.

3.3 Experimental setup and code

We used the authors' code (`https://github.com/xavihart/Diff-PGD/tree/main`), however, since not all parts of the authors' code have been released, we wrote our own code to evaluate a lot of the experiments. We also expanded on some of the authors' experiments with our own code, and reduced the number of iterations in a number of cases, due to time and compute constraints. We also added our own code for evaluating the success rates of the authors' code along with creating various visualizations.

3.3.1 Attack Success Rate

We wrote our own code to test success rate. Our measure of success rate is the percentage of successful attacks. We consider an attack successful if the adversarial sample caused the model to classify the image as anything other than its original classification.

3.3.2 Physical-World Attacks

We tried two of our own physical world attacks using an image patch of a digitally drawn panda head and a laptop as our target object for the first one, and an image patch of a forest photo and a water bottle in the second.

We use a Galaxy A53 5G to take images from the real world and used a MP C4505ex Color Laser Multifunction Printer to print the images in color.

Due to time and compute constraints, we chose to run 1500 iterations for each method instead of 4000 like the original model. We chose 1500 because after running one of the original paper's physical world attacks, the loss converged after around 1000 iterations.

We took the adversarial samples generated using each method (AdvCam, AdvPatch, and Diff-PDG) and stuck them on a the target object. For each adversarial sample (one for each method), we took multiple pictures of the target object with the adversarial sample on top, with the sample in various different locations and rotations, and tested for success rate on the these photos by classifying them with ResNet-50. We also did this for the original unperturbed image patch and the photo with no image patch.

3.3.3 Style-Based and Global Attacks

For each iteration, we take a random base image from the test set of the ImageNetMini dataset. Then for each image we compare the success rate for creating adversarial samples under 3 different attacks:

1. Style-based attack with a different target image from ImageNetMini. Along with a mask generated based on the base image
2. The same as the above, except no mask is applied.
3. Global attack which has no target image and no mask

We repeat this for 100 images.

3.4 Computational requirements

We ran it on either CPU or MPS. For used MPS for style-based attacks and global attacks, and CPU for physical-world attacks.

Global attack success rate took between 4-5 hours to run, while physical-world attacks for one target object and patch took around 1-2 days to run for 1500 iterations.

The style-based attack took around 15s per image.

4 Results

4.1 Physical-World Attacks

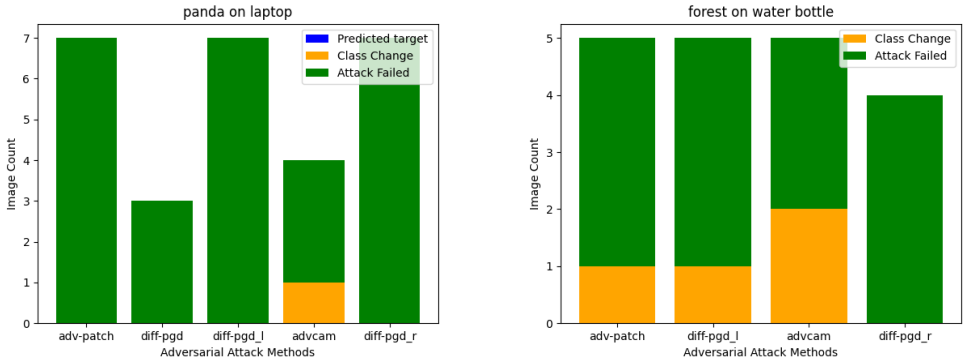

Figure 1: **Evaluating Success Rate of Physical-World Attacks**
Blue line is the number of images that were classified correctly as the target label. Orange line is the number of images that had a different classification with ResNet-50, but that is not the target class. Green line is the number of images whose classifications were one of the following: 'notebook, notebook computer', 'laptop, laptop computer', 'iPod' for panda on laptop, and 'water bottle', 'lighter, light, igniter, ignitor', 'rain barrel', 'pill bottle' and 'hand blower, blow dryer, blow drier, hair dryer, hair drier' for forest on water bottle. These are the classes ResNet-50 gave to images with no adversarial sample (the original image on target object or no image on target object), or classes that we decided had a nearly identical similar meaning as the target object class (ie. laptop vs. notebook computer)

For the panda image adversarial sample and the laptop as the target object, at first we found that none of the adversarial samples generated could perturb the classification of the image. We thought this might be due to the scale of the image, so we tried to adjust the scale of the image for AdvCam. From the different camera rotations, zooms and angles we took the AdvCam image from, only one classified it as something different. However, this class was 'hard disk' (fig. 6), which is also not too far from the original image's classification of 'notebook, notebook computer'.

For the forest image adversarial sample and the water bottle as the target object, we found that the AdvCam, AdvPatch and Diff-PGD without purification samples were able to perturb the classification of the image on 1 or 2 of the images, while the Diff-PGD sample with purification did not.

4.2 Style-Based Attacks

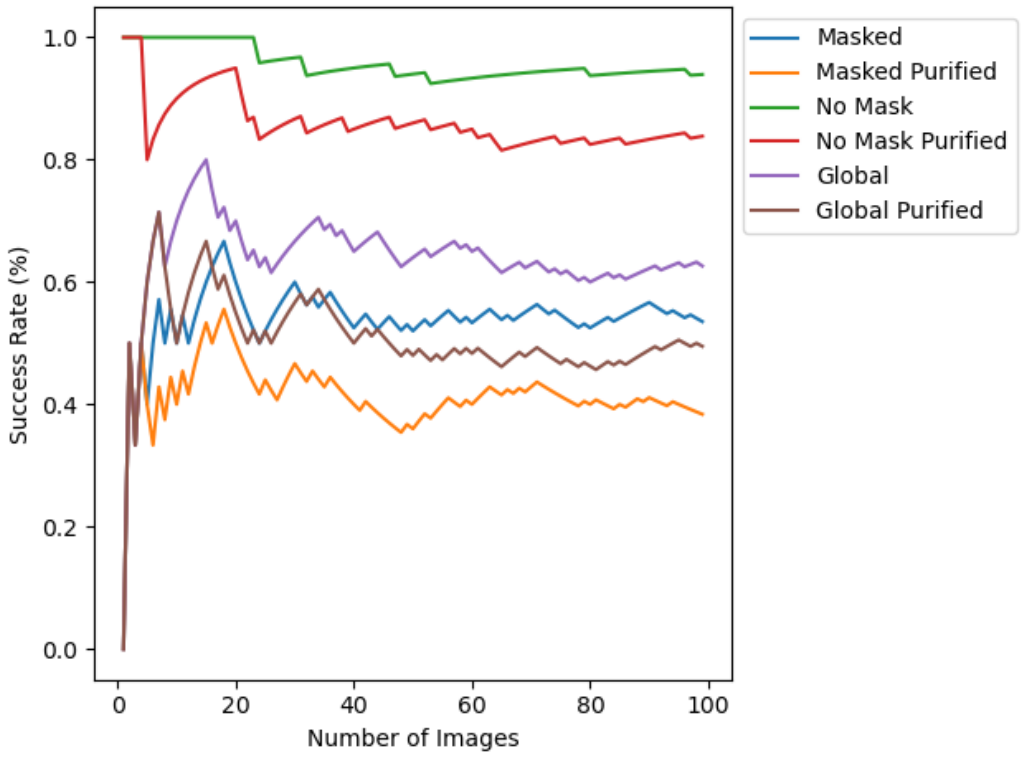

Figure 2: **Plot of the Success Rates**:
We see the best results for applying a style-based attack with no mask. A global attack with no style-image gives the worst results. The success rate for purified images is lower than their non-purified versions.

4.3 Results for Global Attack

From fig. 2 it show the success rate for global attacks is around 65% before purification, and around 50% after purification. This success rate is slightly lower than in the original paper. We used the same hyperparameters as the author, but tested less images.

4.4 Results for Style Attack With Mask

From fig. 2 it shows the success rate for style attacks with a mask applied at around 60% before purification, and around 40% after purification. The original paper does not provide success rates for this type of attack in particular.

4.5 Results for Style Attack Without Mask

From fig. 2 it shows the success rate for style attacks with no mask applied at around 95% before purification, and around 85% after purification. The original paper does not provide success rates for this type of attack in particular. But the anti-purification results hold up as there is minimal drop in success rate.

5 Discussion

5.1 Attack Success Rates of Physical Attacks

Results were very similar across all three methods (Diff-PGD, AdvCam, AdvPatch), so we failed to support or reject the claim that Diff-PGD can be applied to physical-world attacks. However, from the samples generated, the Diff-PGD samples looked less stealthy and more perturbed than those of AdvCam and AdvPatch, which does not support the paper's claim that Diff-PGD generates samples with higher stealthiness. Fortunately, the stealthiness of physical-world samples is likely less important, since the samples generated through every method were quite obviously perturbed. It also appears that the rotation, zoom, and angle the photo was taken at could affect whether the attack was successful or not, since in the water bottle case, some successful misclassifications were caused, but we did not notice any obvious pattern on what camera angles were better than others. Even without the adversarial samples, the classifications for the water bottle case were quite varied, and often did not make much sense (i.e. 'rain barrel', 'hand blower, blow dryer, blow drier, hair dryer, hair drier'), therefore a likely problem is that we used a dataset that was too small. It may be that since physical attacks have more requirements on robustness, having enough training data is very important on getting the attack to work.

5.2 Attack Success Rates of Global Attack

We did not get great success rates with the global attack. The global attack just adds some noise to the original to change the classification. However, one caveat with this method is that some labels in ImageNet are very similar. For example, there is a difference between "great white shark" and "tiger shark". We observed that even when a global attack was considered "successful", the label would have simply changed to a different label that is still very similar to the original. So while this technically changed the classification to a wrong one, this is an honest mistake that humans would also make.

Due to this, a better measure of success might be to compare the distance between the classification labels before and after the attack using a word-embedding model. This way, mistakes such as getting "great white shark" versus "tiger shark" wrong is punished less. Doing this may lead to a lower success score of global attacks since the changes it makes are not very pronounced.

5.3 Attack Success Rates of Style-Based Attacks

Style-Based attacks when successful would generally drastically change the classification label from the original. Attacks with no mask that allowed changes to any part of the image were generally more successful, but greatly loses the "stealthiness" property claimed in the original paper where changes are hard to detect with the human eye. Attacks with a mask applied, are harder to detect but also have a lower success rate.

This could be due to the fact that we are using random images as our original and our target. In the original paper, they have specific examples where the target image is of a similar shape to the original. Which leads to the attacks being more effective see fig. 4 for an example.

5.4 What was easy

Running the code from the original authors to generate adversarial samples was fairly easy, since documentation was included in how to run each attack.

The authors did not include much documentation for how the code worked within the code itself, however they did include some explanations in the paper. This separation made it a bit difficult to understand what the attack parameters did, however the naming was quite clear for the most part, so this was not too difficult to follow.

The original code had print statements at each iteration, providing the speed each iteration was taking. This was helpful in timing how long the model would take to run.

5.5 What was difficult

The physical-world attack was a lot more difficult to reproduce than we anticipated. AdvCam and AdvPatch both ran fairly quickly, averaging around 2-3s per iteration. However, Diff-PGD averaged around 60s per iteration, which took multiple days to run. We got very low success rates for the physical-world attacks, which may be because physical-world attacks have more requirements on robustness. This likely means that hyperparameter tuning is very important for physical-world attacks. However, tuning the model is more challenging for Diff-PGD for physical-world attacks, due to how slowly it runs. Another mistake we made was assuming the the classifier would classify a closed laptop as a laptop, when in reality it classified it as a notebook. We looked into the ImageNet database, and found that most of the images in both laptop and notebook classes were of open laptops, while we used a closed laptop. Since the photo's original label was fed into the model, this could have possibly caused problems with the model.

6 Conclusions

We failed to reproduce results for physical-world attacks. Our success-rate was very low, and the adversarial samples generated were less stealthy than baseline models. We conclude our hyper-parameter tuning was insufficient to deal with the robustness required for physical-world attacks.

For style-based attacks, the Diff-PGD adversarial samples tend to have higher success rates, and are harder to detect with the human eye when the target image is created more-carefully to look like the original. Using random images as the target still lead to success in confusing the model, but oftentimes, the adversarial sample would not look like the original. Also, the measure of success rate could be improved. Some images are hard to classify, such as different breeds of dogs. Perhaps the model should not be punished as heavily for miss-classifying similar labels. This could be done by measuring vector-embedding distance between labels as a measure of success.

7 Acknowledgement

We would like to thank the CPSC 440 teaching team for their guidance this term. We would also like to thank the original authors for their work leading to this paper.

References

Brown, Tom B., Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer (2018). *Adversarial Patch*. arXiv: 1712.09665 [cs.CV].

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (June 2016). "Deep Residual Learning for Image Recognition", pp. 770–778. DOI: 10.1109/CVPR.2016.90.

Lee, Jungbeom, Eunji Kim, and Sungroh Yoon (2021). *Anti-Adversarially Manipulated Attributions for Weakly and Semi-Supervised Semantic Segmentation*. arXiv: 2103.08896 [cs.CV].

Madry, Aleksander, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu (2019). *Towards Deep Learning Models Resistant to Adversarial Attacks*. arXiv: 1706.06083 [stat.ML].

Xue, Haotian, Alexandre Araujo, Bin Hu, and Yongxin Chen (2023). "Diffusion-Based Adversarial Sample Generation for Improved Stealthiness and Controllability". *arXiv preprint arXiv:2305.16494*.

A Supplementary material

B Link to Our Github Repo

https://github.com/Sokole1/CPSC440Proj

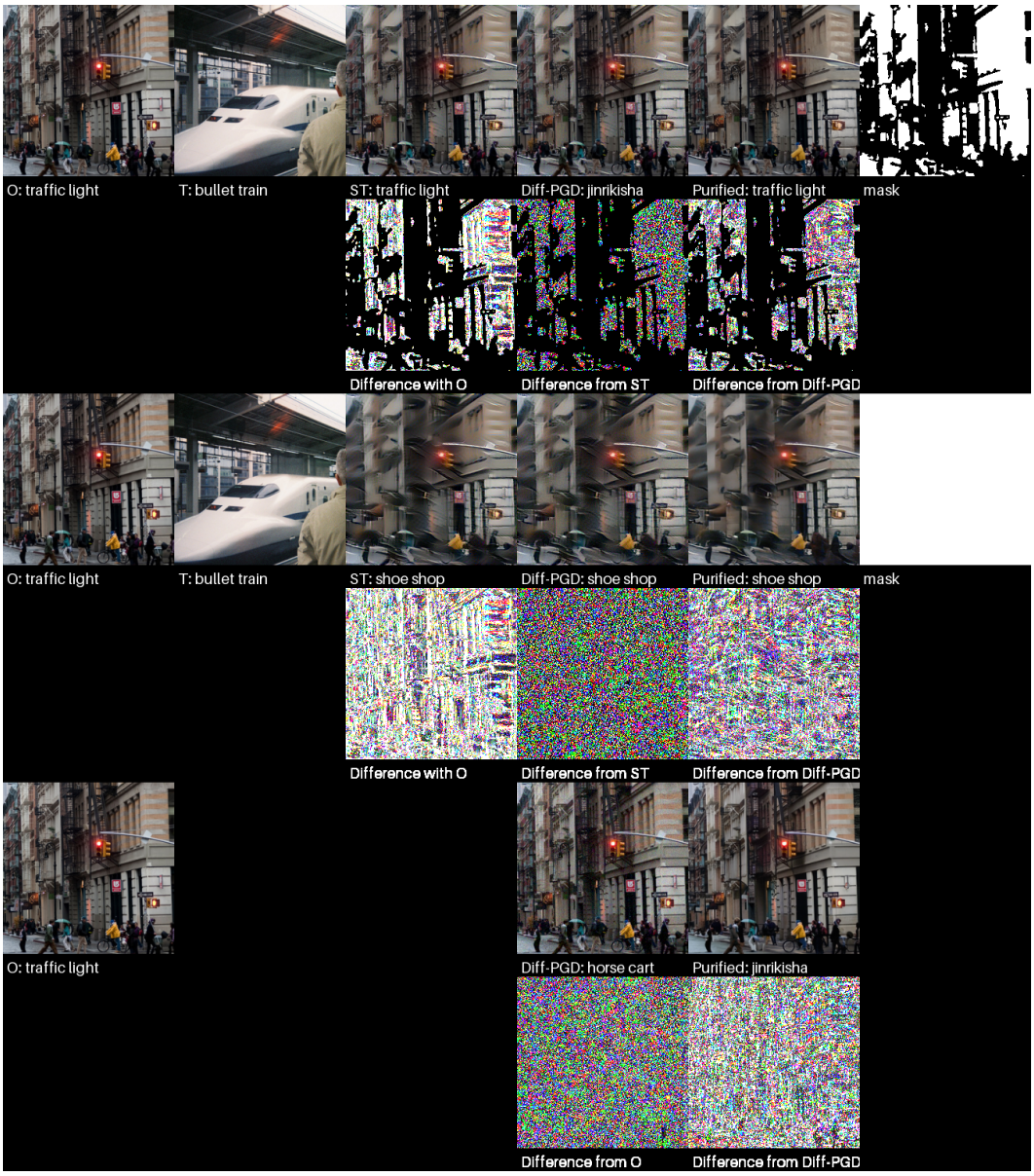

Figure 3: **Visualization of the different attacks**:
There are 6 rows of images. The first two correspond to a style-based attack with a mask applied. The next 2 rows are the same, except there is no mask. The last 2 rows are a global attack where there is no style image. The 6 columns of images are the following: 1: O stands for the original image. 2: T stands for the style image which we try to make the original look more like. 3: ST is the original image with the style image applied, but no adversarial properties. 4: Has the Diff-PGD algorithm applied. 5. Is the image after going through purification. 6: Is the mask applied.

B.1 Style-based attacks

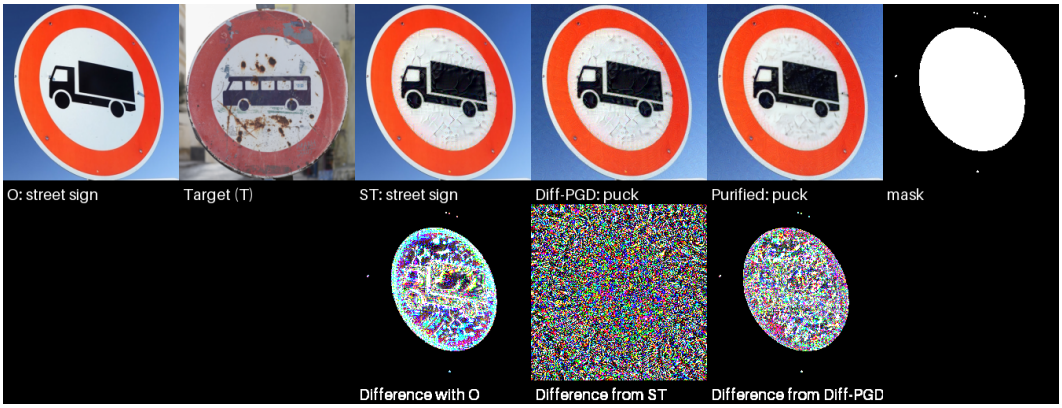

Figure 4: **A Hard to Notice Style-Attack**

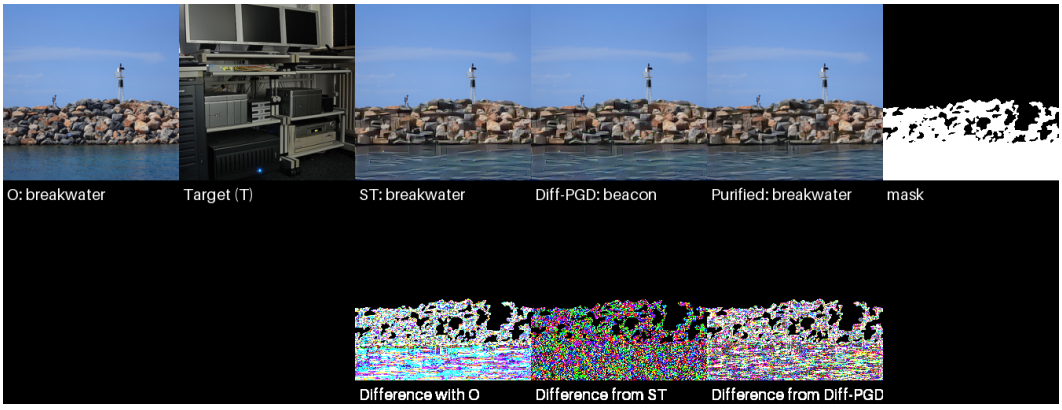

Figure 5: **An Easier to Notice Style-Attack**

B.2 Physical attacks

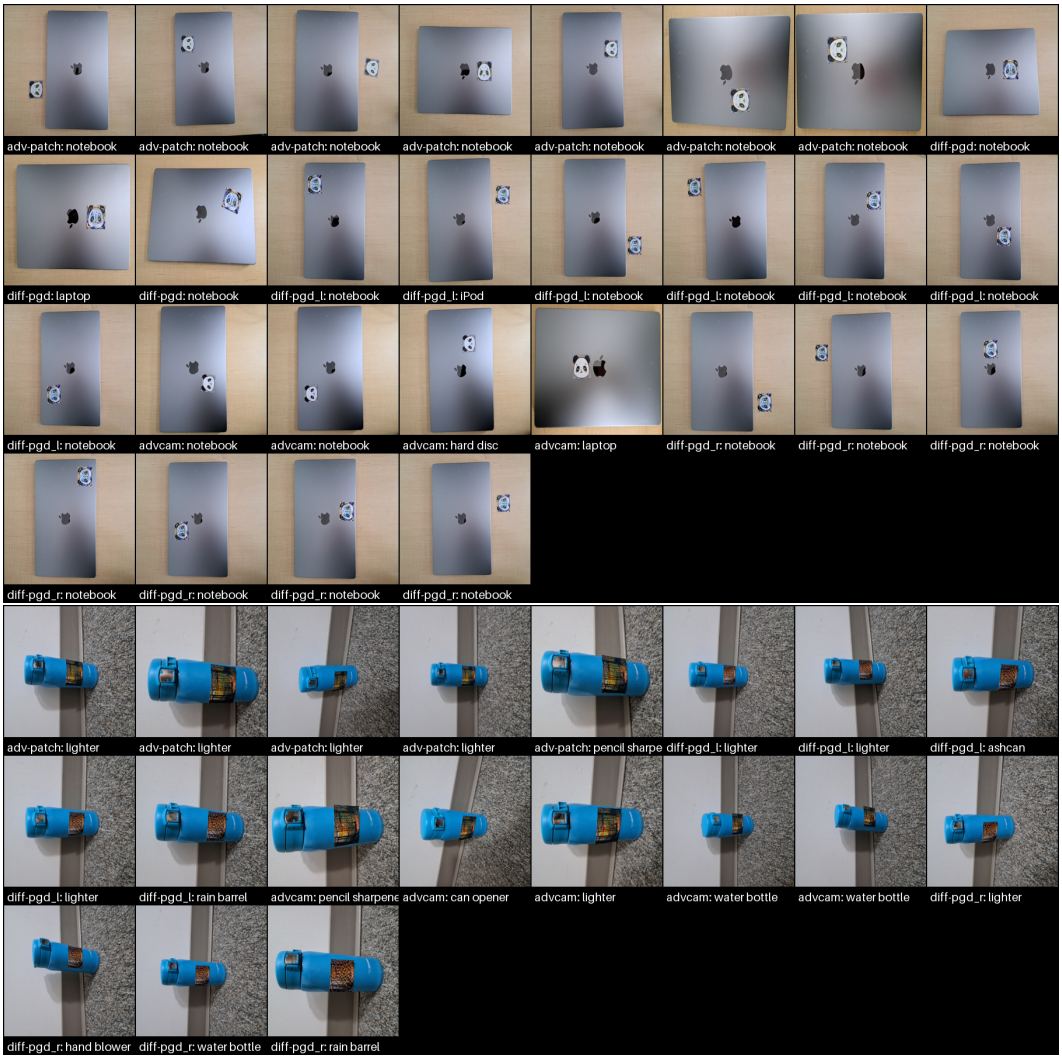

Figure 6: **Visualization of the different attacks**:
Predictions made for adversarial samples with photos taken at different angles, zooms and rotations.