RainPGD: Investigating Rain-Robust Adversarial Patches

Alaqsa Akbar
s202281480@kfupm.edu.sa

Dr. Abduljabbar Siddiqui
abduljabbar.siddiqui@kfupm.edu.sa

Abstract

Adversarial patches pose a significant threat to object detection models by targeting localized image regions to mislead predictions. However, their effectiveness under adverse weather conditions, such as rain, remains underexplored. This paper introduces RainPGD, a novel adversarial patch generation method designed to maintain robustness in rainy conditions. To achieve this, we augment training images with various rain intensities, enabling RainPGD to learn perturbations that remain effective in adverse weather. We also present RainyCOCO, a modified version of the COCO dataset, containing images augmented with simulated rain, to support model training and evaluation under inclement conditions. Experimental results indicate that RainPGD shows a slight improvement in degrading object detection performance under rain, achieving a 2.36% decrease in model recall compared to a standard PGD patch. Although RainPGD incurs some performance trade-offs in clear conditions, it demonstrates potential for weather-resilient adversarial attacks, making it a valuable step toward robust adversarial training for real-world applications. The code and RainyCOCO dataset are publicly available for research use.

1 Introduction

With the rise in popularity of Artificial Intelligence, there has been a corresponding surge in object detection applications [1]. However, this surge also increases the demand for secure models. Adversarial patch attacks are a form of malicious attack targeting object detectors and image classifiers in general. These attacks work by distorting a region of an image in a way that causes the object detector to misidentify objects. For example, adversarial patches have been shown to cause models to misclassify stop signs [2]. To defend against such attacks, it is crucial to develop a strong understanding of how these attacks are generated. By creating more robust adversarial attacks, object detection models can, in turn, become more resilient to them.

Currently, there appear to be no studies that investigate the effects of weather conditions on adversarial patches. Moreover, no efforts have been made to strengthen adversarial patch generation under rainy conditions. The field has primarily focused on de-raining images to improve object detection performance [3]–[5]. Therefore, exploring methods to train adversarial patches against rain and other adverse weather conditions could potentially yield better results and warrants further investigation.

In this paper, we introduce an adversarial patch attack generation method we call RainPGD designed to be robust to adverse weather conditions, particularly rain. Preliminary testing revealed that adversarial patches performed similarly under both adverse and non-adverse weather. Under normal conditions, patches had a mAP of 0.075, while under rainy conditions, patches had a mAP of 0.076. We hypothesize that it is possible to generate patches that could further degrade the performance of object detection models even under heavy weather conditions. This is because rain droplets obscure several pixels behind them, which may be crucial to the patch's effectiveness. RainPGD augments the training data in real-time during the training process, allowing the adversarial algorithm to adapt to adverse weather conditions.

Alongside our investigations, we introduce the code used to augment any image by adding rain, implemented in the `add_rain()` function found in `RainPGD/utils.py`. This function simulates rain with three different intensities: `weak`, `heavy`, and `torrential`. Additionally, we present a modified

version of the COCO dataset [6], called RainyCOCO, which consists entirely of COCO images augmented using the `add_rain()` function. RainyCOCO is designed to facilitate the training of object detectors under adverse weather conditions. However, we recommend augmenting the training images in real-time during the training process to introduce randomness to the dataset.

The code for this paper can be found at: `https://github.com/alaqsa-akbar/RainPGD`.

The RainyCOCO dataset can be found at: `https://huggingface.co/datasets/alaqsa-akbar/RainyCOCO`

2 Related Work

2.1 Object Detection

In this paper, Faster R-CNN [7] was used for the training and evaluation. However, in theory, this should still be effective against single stage detection (SSD) models.

2.2 Adversarial Training

In this paper, we focus on improving the untargeted image-specific and location-specific patches. These are the stronger form of adversarial patches. We will apply the patches on an input image x that is of shape $[0,1]^{H\times W\times 3}$ where H and W are the height and width respectively. Adversarial patches are generated using a projected gradient descent algorithm [8]. The algorithm aims to maximize the loss of the targeted model by solving the equation [9]:

$$\hat{P}(x,l) = \underset{P\in\{P':\left\|P'\right\|_\infty\leq\epsilon\}}{\arg\max} \mathcal{L}(h(A(x,l,P));y)$$

where $\mathcal{L}$ is the loss $\mathcal{L}^{\text{Faster R-CNN}}$ that compares h (the Faster R-CNN algorithm) with y (the ground truth labels). A is a function that applies a patch P onto image x at location l, $\|\cdot\|_\infty$ is the l_∞ norm, and ϵ is the attack budget. Solving this equation would be difficult, as such, a projected gradient descent algorithm is utilized [8], [9]:

$$P^{t+1} = \prod_{\mathbb{P}}(P^t + \alpha\,\text{sign}(\nabla_{P^t}\mathcal{L}(h(A(x,l,P^t));y)))$$

where t is the iteration count, α is the step size, $\prod$ is the projection function that will project onto $\mathbb{P} = P \in \{P' : \left\|P'\right\|_\infty \leq \epsilon\}$ and $A(x,l,P) \in [0,1]^{H\times W\times 3}$. In the case of patch attacks, we will use $\epsilon = 1$ to allow maximum freedom of pixel manipulation in the localized area.

3 Proposed Methodology

In order to improve robustness to rainy conditions, the rain patch must be trained against a target image representing the ideal scenario. To do so, we must figure add rain onto target images and then train the patch against them.

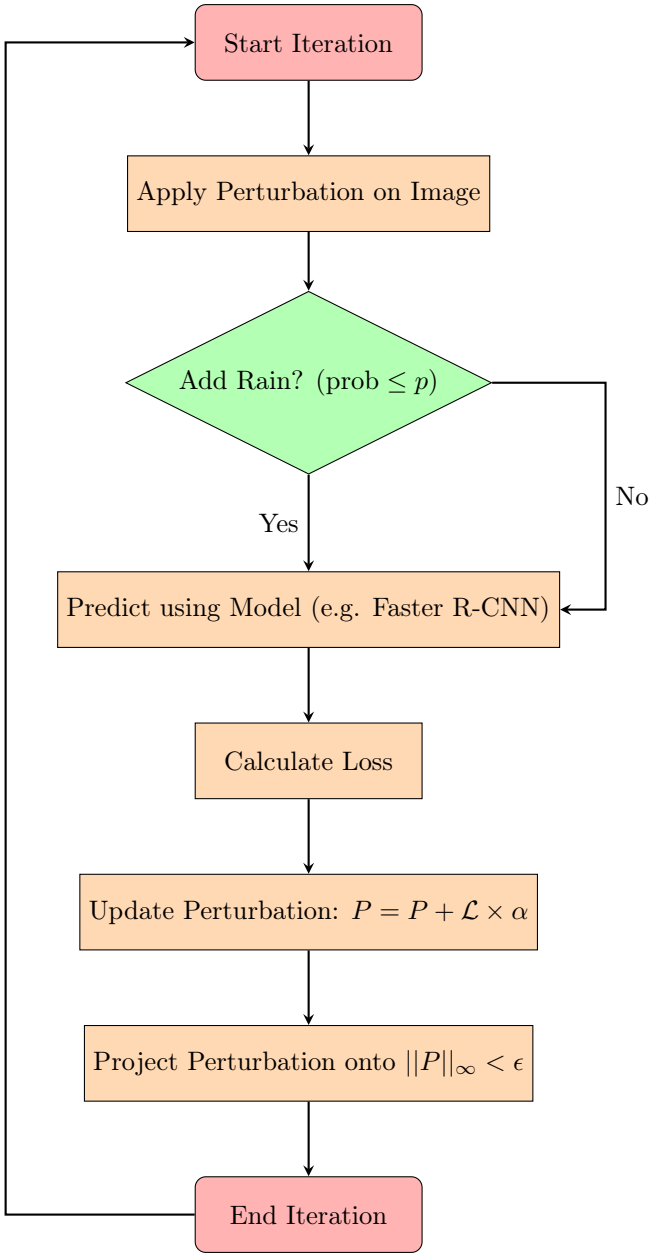


Figure 1: Flowchart of the proposed methodology

3.1 Training Against Rain

During the training process, rain conditions are randomly augmented onto the image. This happens before calculating the loss from the target estimator model. This is done to allow the PGD algorithm to learn the rain streaks and try to find the optimal perturbations to use with the rain.

The proposed methodology will look like figure 1. Assume R is the function that applies rain conditions to an image with the patch applied $A(x, l, P^t)$ with a probability of p. The projected gradient descent algorithm becomes:

$$P^{t+1} = \prod_{\mathbb{P}}(P^t + \alpha \operatorname{sign}(\nabla_{P^t}\mathcal{L}(h(R(A(x, l, P^t)), p); y)))$$

By adding a random probability p, the patch is trained to also be effective under non-rainy conditions. As a result, p would be a hyper-parameter needed to be tuned where a higher p would train a patch for areas with more frequent rain while a lower p would train for areas with less frequent rain.

3.2 Augmenting Images

The function R works by implementing a modified methodology utilized in Photoshop [10] by using OpenCV [11]. The method found in Photoshop was used by previous papers that worked on state-of-the-art rain removal [3]. The algorithm works by creating a mask image m and applying it over the target image x where m is initially $m = [0]^{H \times W \times 3}$. Some points are then randomly selected from x through a discrete uniform distribution. The number of points selected is also a sampled from a uniform distribution using the equation:

$$s \sim U(0.8 \times HWr, HWr)$$

where s and r is a value chosen based on the rain strength chosen. Through experimentation, the values chosen for r were:

- `'weak'`: 0.002
- `'heavy'`: 0.004
- `'torrential'`: 0.006

These points are then converted into white rain streaks. The length of the one streak is calculated similarly using:

$$l \sim U(0.8 \times \frac{H}{15}, \frac{H}{15})$$

where l is the streak length and 15 is a number chosen through experimentation. Similarly the angle of the rain streaks, θ, is also sampled from a uniform distribution between $\frac{\pi}{4}$ and $\frac{3\pi}{4}$. Finally a line with values of 1 is drawn from $q_t + \delta$ to $q_t - \delta$ creating the matrix $\begin{pmatrix} q_t + \delta \\ q_t - \delta \end{pmatrix}$ where q_t is one of the chosen points and δ can be calculated using:

$$\delta = \frac{l}{2} \times \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix}$$

A Gaussian blur, $G(m)$, is then applied to the mask using a kernel size of `(9, 9)`. The brightness is reduced by a factor 0.7; this was chosen through experimentation. The mask is then applied using the screen blend algorithm found in Photoshop [12]:

$$S(x, m) = 1 - (1 - x) \odot (1 - m)$$

Finally, this would make the equation for $R(x, p)$ to be:

$$R(x, p) = \begin{cases} x & U(0,1) > p \\ S(x, G(x + \sum_t \begin{pmatrix} q_t + \delta \\ q_t - \delta \end{pmatrix})) & U(0,1) \leq p \end{cases}$$

4 Experimental Setup

For the testing of RainPGD, a Windows 10 computer was used with an RTX 3080 GPU and an Intel i9-12900k CPU. 6 different tests were run on the first 1000 images from the validation split of the 2017 COCO dataset. The 6 tests were the following:

- No patch applied on a normal image
- PGD patch applied on a normal image
- RainPGD patch applied on a normal image
- No patch on a an image with rain added
- PGD patch on an image with rain added
- RainPGD patch on an image with rain added

Each patch was trained for 200 iterations. When training RainPGD and testing under rainy conditions, the rain probability was set to 0.7. The patches were trained to be applied randomly onto images. The patch size was 100x100 square patches. The patches were trained against PyTorch's `fasterrcnn_resnet50_fpn` [13], trained on the 2017 COCO dataset, wrapped with Adversarial Robustness Toolbox (ART)'s `PyTorchFasterRCNN` class. Much of the implementation relied on ART's [14] `ProjectedGradientDescent` class and SegmentAndComplete's [9] `PGDPatch` class. The tests can be resimulated using `generate_adv_data.py` to generate the patches for each test case followed by using `eval_attack.py`.

To evaluate the results, the `pycocotools` library was used and generated results of the following categories:

Average Precision (AP):

- IoU=0.50:0.95 | area=all | maxDets=100
- IoU=0.50 | area=all | maxDets=100
- IoU=0.75 | area=all | maxDets=100
- IoU=0.50:0.95 | area=small | maxDets=100
- IoU=0.50:0.95 | area=large | maxDets=100

Average Recall (AR):

- IoU=0.50:0.95 | area=all | maxDets=1
- IoU=0.50:0.95 | area=all | maxDets=10
- IoU=0.50:0.95 | area=all | maxDets=100
- IoU=0.50:0.95 | area=small | maxDets=100
- IoU=0.50:0.95 | area=mid | maxDets=100
- IoU=0.50:0.95 | area=large | maxDets=100

Average precision describes the model's precision or the ratio of correct predictions to total predictions. It can be calculated using:

$$\frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

Average recall describes the model's ability to recall classes from a given image or the ratio of correct predictions to the total number of ground truths. It can be calculated using:

$$\frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

A lower AP and AR score suggests a better performing adversarial patch as it causes the object-detection model to perform worse. In the case of non-targeting adversarial patch, AR is more important as the purpose of the non-targeting patches is to reduce the model's ability to detect objects. The main metric to observe would be AR@[IoU=0.50:0.95, maxDets=100] as this allows for the most detections which is more representative of the real world.

To generate the RainyCOCO dataset, the `add_rain` function was applied to all 55000 images in the train split of the 2017 COCO dataset. 3 different images were generated from each COCO image, one corresponding to each rain intensity. Similarly, this was done to the 5000 images in the validation split of the 2017 COCO dataset.

5 Results and Discussions

The full table of results is available in the appendix in tables 1 and 2, while the summarized outcomes are shown in Figure 2. The findings indicate that RainPGD enhances patch performance under rainy conditions, reducing model recall from 0.254 with a standard PGD patch to 0.248 with RainPGD under rainy conditions. This corresponds to a 2.36% decrease in model performance, signifying a 2.36% improvement in patch robustness in rain. However, under clear conditions, RainPGD performs worse, showing a 7.08% decrease in performance, likely due to its training with a rain probability of 0.7. A lower rain probability might yield a more balanced performance across conditions but further experimentation is needed.

In terms of model precision, the mean average precision (mAP) metric revealed that RainPGD performed 4.85% worse than the standard PGD patch under rainy conditions. Despite this, the recall metric remains more relevant as it directly supports the goal of creating a non-targeting patch.

These results suggest that the RainPGD patch is indeed adapting to rain patterns, enabling it to perform marginally better under rainy conditions. However, this improvement may not justify the performance loss in clear conditions. One potential reason for this trade-off could be the noisy and inconsistent rain patterns during training, which may harm the patch's learning as it needs to learn from several different pattern rain patterns. Extending training over more iterations could help mitigate this, but further experimentation is needed.

Another notable finding is the minimal difference in model performance with and without rainy conditions when no patch or a standard PGD patch is used. This stability implies that accommodating rainy conditions may not be essential. Yet, the RainPGD patch's improvement over the standard PGD patch under rain suggests there is potential value in this approach.

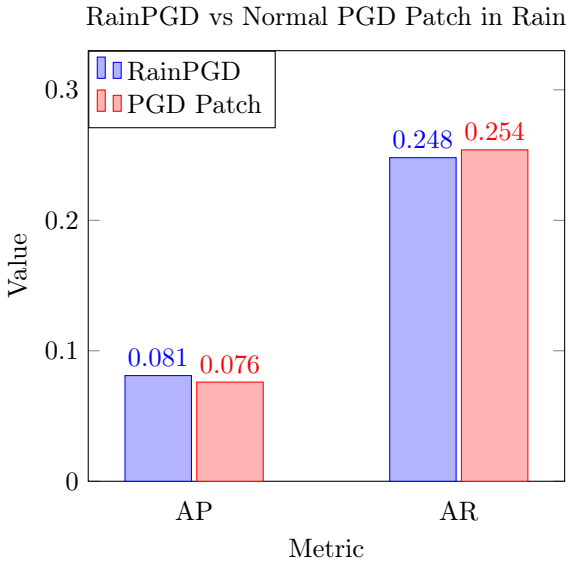


Figure 2: Comparison of AR and AP values for RainPGD and PGD Patch under rainy weather conditions (lower is better)

6 Conclusion and Future Work

In this paper, we introduced RainPGD, a novel adversarial patch generation algorithm designed to be resilient under adverse weather conditions,

specifically rain. Our experiments showed a 2.36% improvement in recall compared to traditional methods, indicating RainPGD's potential for enhancing robustness in challenging conditions. The approach underlying RainPGD can likely be extended to other environments, such as snow or dust, and even to more complex applications where adaptive patches are needed.

Further exploration is essential to optimize RainPGD and fully understand its potential. Future work could focus on refining the rain simulation method, exploring varied hyper-parameters, and testing RainPGD with targeted adversarial patches to broaden its applicability.

Additionally, we introduced the RainyCOCO dataset, intended to support ongoing research by providing a standardized dataset for evaluating performance in rainy weather. We hope RainyCOCO will facilitate further advancements in weather-robust AI systems.

Acknowledgments: We would like to thank the SDAIA-KFUPM Joint Research Center for sponsoring this research and providing essential support throughout the project.

References

[1] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *Proceedings of the IEEE*, vol. 111, no. 3, pp. 257–276, 2023. DOI: 10.1109/JPROC.2023.3238524.

[2] K. Eykholt, I. Evtimov, E. Fernandes, *et al.*, "Physical adversarial examples for object detectors," *CoRR*, vol. abs/1807.07769, 2018. arXiv: 1807.07769. [Online]. Available: http://arxiv.org/abs/1807.07769.

[3] H. Zhang and V. M. Patel, "Density-aware single image de-raining using a multi-stream dense network," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 695–704. DOI: 10.1109/CVPR.2018.00079.

[4] H. Zhang, V. Sindagi, and V. M. Patel, "Image de-raining using a conditional generative adversarial network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 11, pp. 3943–3956, 2020. DOI: 10.1109/TCSVT.2019.2920407.

[5] H. Chen, Y. Wang, T. Guo, *et al.*, "Pre-trained image processing transformer," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 12 294–12 305. DOI: 10.1109/CVPR46437.2021.01212.

[6] T.-Y. Lin, M. Maire, S. Belongie, *et al.*, "Microsoft coco: Common objects in context," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., Cham: Springer International Publishing, 2014, pp. 740–755, ISBN: 978-3-319-10602-1.

[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, ser. NIPS'15, Montreal, Canada: MIT Press, 2015, pp. 91–99.

[8] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, *Towards deep learning models resistant to adversarial attacks*, 2019. arXiv: 1706.06083 [stat.ML]. [Online]. Available: https://arxiv.org/abs/1706.06083.

[9] J. Liu, A. Levine, C. P. Lau, R. Chellappa, and S. Feizi, "Segment and complete: Defending object detectors against adversarial patch attacks with robust patch detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 973–14 982.

[10] S. Patterson. "Adding rain to a photo with photoshop." (Nov. 17, 2020), [Online]. Available: https://www.photoshopessentials.com/photo-effects/rain/.

[11] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[12] Adobe. "Blending modes." (Jul. 28, 2024), [Online]. Available: https://helpx.adobe.com/photoshop/using/blending-modes.html (visited on 10/12/2024).

[13] A. Paszke, S. Gross, S. Chintala, *et al.*, "Automatic differentiation in pytorch," in *NIPS-W*, 2017.

[14] M.-I. Nicolae, M. Sinn, M. N. Tran, *et al.*, "Adversarial robustness toolbox v1.2.0," *CoRR*, vol. 1807.01069, 2018. [Online]. Available: https://arxiv.org/pdf/1807.01069.

A Evaluation Results

Table 1: Evaluation Results for Normal Conditions

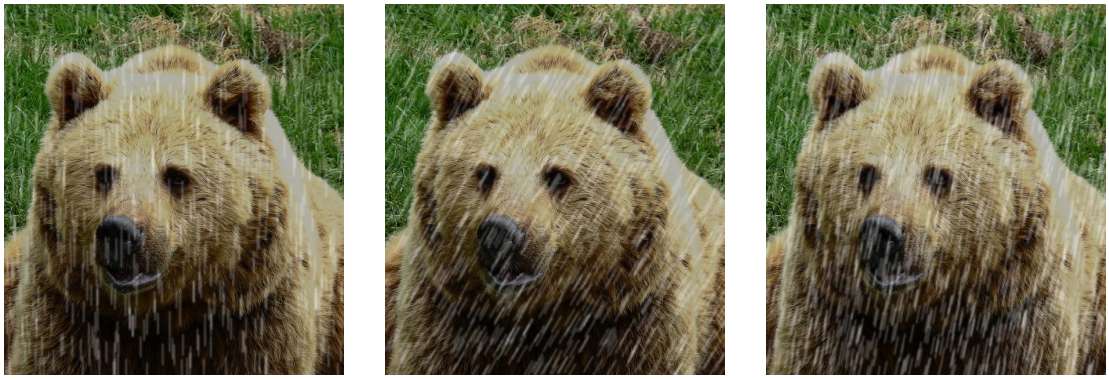
Metric	No Patch	PGD Patch	RainPGD Patch
AP@[IoU=0.50:0.95]	0.322	**0.075**	0.119
AP@[IoU=0.50]	0.490	**0.130**	0.198
AP@[IoU=0.75]	0.351	**0.076**	0.123
AP@[small]	0.091	**0.012**	0.022
AP@[large]	0.493	**0.207**	0.252
AR@[IoU=0.50:0.95, maxDets=1]	0.280	**0.188**	0.212
AR@[IoU=0.50:0.95, maxDets=10]	0.409	**0.295**	0.317
AR@[IoU=0.50:0.95, maxDets=100]	0.418	**0.302**	0.325
AR@[small]	0.106	**0.070**	0.078
AR@[medium]	0.471	**0.346**	0.380
AR@[large]	0.656	**0.470**	0.493

Table 2: Evaluation Results for Rainy Conditions

Metric	No Patch	PGD Patch	RainPGD Patch
AP@[IoU=0.50:0.95]	0.322	**0.076**	0.081
AP@[IoU=0.50]	0.490	**0.131**	0.140
AP@[IoU=0.75]	0.351	**0.077**	0.083
AP@[small]	0.091	0.017	**0.012**
AP@[large]	0.493	0.184	**0.175**
AR@[IoU=0.50:0.95, maxDets=1]	0.280	**0.157**	0.165
AR@[IoU=0.50:0.95, maxDets=10]	0.409	0.249	**0.242**
AR@[IoU=0.50:0.95, maxDets=100]	0.418	0.254	**0.248**
AR@[small]	0.106	0.058	**0.057**
AR@[medium]	0.471	0.294	**0.274**
AR@[large]	0.656	0.398	**0.388**

B RainyCOCO Images

(a) Weak rain (b) Heavy rain (c) Torrential rain



(d) Weak rain (e) Heavy rain (f) Torrential rain

(g) Weak rain (h) Heavy rain (i) Torrential rain

Figure 3: Weak, heavy, and torrential rain from the RainyCOCO dataset, labeled by rain type.