Physical Adversarial Attacks on Deep-Learning-based ISP Pipelines

Darian Jennings, Rong Wei, Kartheek Reddy Gade, Syed Faizan Ali
University of Florida

1 Introduction

In this paper, we introduce the PyNET-CA[3] model, an advanced Image Signal Processsing (ISP) solution designed to address the challenges faced by mobile device cameras under extreme lighting conditions as well as emerging security threats. PyNET-CA[4] optimizes the accuracy of color processing and the fidelity of image details by integrating deep learning techniques. In addition, this study pays special attention to the performance of the model in a physical adversarial attack environment. Through experimental evaluations, we tested the efficacy of PyNET-CA in processing images in low-light environments and its robustness in the face of well-designed adversarial samples. Adversarial noise is a crucial concept in machine learning security as it can be intentionally crafted to mislead deep neural networks, leading to incorrect predictions and potentially harmful consequences. Synthesizing Robust Adversarial Examples [1] demonstrates the effectiveness of adversarial noise in fooling state-of-the-art models, highlighting the need for robust defenses against such attacks. By understanding and generating adversarial noise, researchers can develop more resilient models and improve the reliability of machine learning systems in real-world applications.

We tested and manipulated a variety of adversarial attack algorithms, ultimately implementing and optimizing the Fast Gradient Symbol Method (FGSM)[2] to test and evaluate PyNET-CA's resistance to these type of attacks.

2 Background and Related Work

In mobile devices, the core task of image signal processing (ISP) technology [7] is to convert the raw data captured by the camera sensor into a high-quality RGB [5]. This conversion involves key steps such as demosaicing, denoising, and soon. While traditional ISP technologies are optimized for hardware, their performance is limited in dynamic environments and low-light conditions. Deep learning provides a novel solution that can adaptively optimize the image processing process by learning from large datasets. PyNET-CA, an advanced deep learning ISP model, introduces a channel-attention mechanism specifically optimized for image quality in low-light and complex scenes [5]. To address the vulnerability of deep learning models to physical adversarial attacks, this study evaluates the performance and robustness of PyNET-CA under these extreme conditions by implementing various adversarial attacks, including the Fast Gradient Sign Method (FGSM) [2].

3 Methodology

Our methodology is structured to rigorously evaluate the resilience of the PyNET-CA model to physical adversarial attacks within ISP pipelines. The experiment follows a step-by-step approach, integrating attack generation with model evaluation to ensure comprehensive analysis. We begin by generating adversarial examples using the Fast Gradient Sign Method (FGSM). FGSM works by exploiting the gradients of the neural network to create perturbations that 'fool' the model. It does this by taking the sign of the gradient of the loss with respect to the input data, then adjusting the input data by an epsilon value in the direction of that sign. This method is chosen for its efficiency and effectiveness in generating adversarial examples that cause the model to misclassify images with minimal changes to the original image. The loss function is a crucial component of our methodology, as it quantifies the difference between the predicted image and the ground truth. This difference provides a gradient that guides the adversarial attack. For FGSM, the loss function's gradient is particularly important because the direction and magnitude of the input adjustments are derived directly from this gradient. In our context, we employ the Structural Similarity Index Measure (SSIM) as the loss function. SSIM [6] is an advanced metric that measures the perceptual difference between two similar images, which is important for evaluating image quality. Unlike traditional loss functions that may focus solely on pixel-level accuracy, SSIM considers changes in structural information, luminance, and contrast, which are vital in

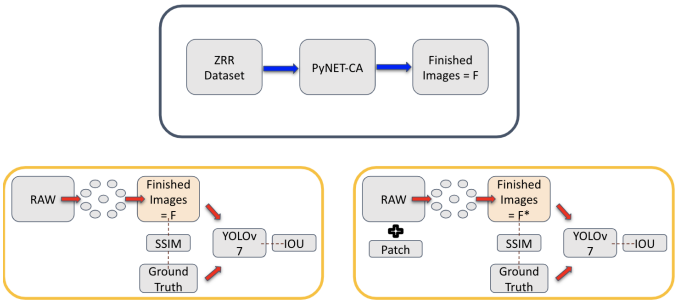

Figure 1: Overall project pipeline and sub-components

assessing the quality of images processed by ISP pipelines. By using SSIM, we aim to maintain the perceptual integrity of the adversarially perturbed images, ensuring that any degradation in image quality reflects potential real-world implications of adversarial attacks. The PyNET-CA model is then trained and evaluated with the generated adversarial examples. The training includes a comprehensive regimen that enhances the model's adaptability to adversarial perturbations. Throughout this process, we monitor the SSIM loss to fine-tune the model's parameters, aiming to minimize the perceptual impact of adversarial attacks on the reconstructed images. Finally, we utilize object detection with YOLO as an alternate method of evaluation and verification of our patch-based adversarial attack algorithm. YOLO calculates an intersection over union (IoU) score, which is a measure of the overlap between boundary boxes of the anticipated and the ground truth.

This pipeline is visualized in figure 1, where you can see the various sub-frameworks in place and how the entire pipeline fits together as one coherently. Note: the small circular patterns represent the deep-learning CNN's that the model PyNET-CA utilizes, so you can envision this as the images being processed or 'converted' by PyNET-CA. This methodology ensures a balance between theoretical robustness and practical visual fidelity, positioning our study at the intersection of cybersecurity and computational photography. Through these steps, we provide insights into the vulnerabilities of ISP pipelines to adversarial attacks and contribute to the development of models that are both accurate and secure.

4 Experimental Results

We compiled our findings into the following two tables (1,2). Table 1, showcases the structural similarity (SSIM) index for the three datasets we compiled with respect to the ground truth. The ground truth is provided by the original authors for the PyNET-CA paper and implementation. The subRAW dataset consists of 58 images that were manually selected by our team from the original 1204 RAW images that PyNET-CA utilized. We selected these images to fit one of our project goals of object detection as well as to fit resource constraints and limitations. The other two datasets, subRAW-ADV-100 and subRAW-ADV-250, represent the attacked subRAW dataset with respective patch sizes, 100 for a patch size of 100x100 and 250 for a patch size of 250x250. We placed the patches at the center of each image and proceeded to evaluate performances. From table 1, we can see that the adversarial attack negatively impacted the performance of the PyNET-CA model, the images processing was not as successful and hence the lower SSIM score. Likewise, as we increased the patch size from 100x100 to 250x250 we saw the performance became even lower.

Table 2 shows similar scores but includes the objectness score, a measure of how well the object detector (YOLO) identifies the locations and classes of objects during navigation. The objectness score represents how well the predicted bounding box covers the actual object, allowing the NMS to prioritize the most accurate and relevant detections. Additionally, the objectness score is used to scale the class score during inference, ensuring that the final detection score is a meaningful combination of the objectness (how well the box fits the object) and the class prediction. This helps the model focus on the most relevant and trustworthy detections. With this, we saw the objectness score follow a similar pattern - we only test YOLO on the subRAW-ADV-100, we did not include subRAW-ADV-250.

Dataset	SSIM
groundTruth	-
subRAW	0.603
subRAW-ADV-100	0.547
subRAW-ADV-250	0.435

Table 1: SSIM indexes compared to ground truth (GT)

Dataset	Obj Score	SSIM
groundTruth	0.690	-
subRAW	0.690	0.603
subRAW-ADV-100	0.414	0.547

Table 2: Objectness score and SSIM indexes

In our research, we launched adversarial attacks against the PyNET-CA model and thoroughly assessed its ability to convert RAW sensor data into RGB images. The experimental results indicate that PyNET-CA maintains a relatively higher resilience against patch-based adversarial attacks compared to traditional legacy systems. Although the adversarial attacks did cause some deviations in image quality, including issues with color fidelity and detail preservation, the model generally remained robust overall. We applied these adversarially processed images to an object detection model, YOLOv7 [8], to evaluate their performance. Adversarial patches have

been shown to significantly degrade the accuracy of object detection models, and this vulnerability transfers to the image signal processing (ISP) pipeline, which is used to process and enhance images. In our research, we found that the YOLOv7 model misclassified the processed patch as an actual object in the photo, such as a stop sign, in a few instances. Moreover, the patch-based attack we implemented not only decreased the accuracy of correctly classified objects but also led to complete misclassifications compared to the ground truth. This further highlights the potential risks of adversarial patches in real-world applications, such as autonomous driving or surveillance systems, and now even in deep-learning ISP models that are utilized on mobile devices. Some of the test images demonstrated the model's relative resistance to such attacks, however, overall there was a significant decline in performance across all metrics. This suggests that developing defensive strategies against adversarial attacks and further optimization of the model to maintain precision and high-quality output in image processing remains crucial for the future.

Figure 2: Adversarial attack with patchSize = [100,250]

5 Analysis and Discussion

Despite not being completely impervious, the PyNET-CA model demonstrates exceptional resilience against adversarial attacks. While the image quality is somewhat affected by the adversarial mode, it is only to a limited extent. Upon analyzing the images processed by PyNET-CA, we observe a degradation in the performance of the target detection model under adversarial conditions, thereby emphasizing the potential and necessity of robust adversarial training and model optimization. Our research demonstrates that integrating channel attention mechanisms and continuously improving models can significantly mitigate the powerful vulnerability posed by ISP pipelines. This study contributes to a more comprehensive understanding of security aspects related to deep learning-based ISP pipelines and underscores the importance of developing more sophisticated defense mechanisms for safeguarding next-generation imaging technologies. In this study, we successfully demonstrated a patch-based adversarial attack on the PyNET-CA image signal processing pipeline. However, future work should investigate optimizing the patch location to maximize the attack's effectiveness. Currently, we only placed the patch at the center of the input image, which may not be the most optimal location. By exploring different patch locations, we can potentially increase the likelihood of breaking or under-performing the model. Optimizing patch location could reveal new vulnerabilities in the PyNET-CA pipeline, leading to further insights into its robustness. Future research should explore this avenue to enhance the attack's potency and better understand the pipeline's limitations. However, patch location optimization is a challenging task, especially for RAW images, which have a more complex structure and require a deeper understanding of the image formation process. Additionally, the high dimensionality of RAW images makes it difficult to efficiently search for the optimal patch location, highlighting the need for innovative solutions to address this challenge.

References

[1] Anish Athalye et al. "Synthesizing Robust Adversarial Examples". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, Oct. 2018, pp. 284–293.

[2] Yinpeng Dong et al. "Boosting adversarial attacks with momentum". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 9185–9193.

[3] Andrey Ignatov, Luc Van Gool, and Radu Timofte. "Replacing Mobile Camera ISP with a Single Deep Learning Model". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2020, pp. 2275–2285. DOI: 10.1109/CVPRW50498.2020.00276.

[4] Byung-Hoon Kim et al. "Pynet-ca: enhanced pynet with channel attention for end-to-end mobile image signal processing". In: *European Conference on Computer Vision*. Springer. 2020, pp. 202–212.

[5] Tarun Kumar and Karun Verma. "A Theory Based on Conversion of RGB image to Gray image". In: *International Journal of Computer Applications* 7.2 (2010), pp. 7–10.

[6] Jim Nilsson and Tomas Akenine-Möller. *Understanding SSIM*. 2020. arXiv: 2006.13846 [eess.IV].

[7] Keumsun Park, Minah Chae, and Jae Hyuk Cho. "Image pre-processing method of machine learning for edge detection with image signal processor enhancement". In: *Micromachines* 12.1 (2021), p. 73.

[8] Joseph Redmon et al. "You Only Look Once: Unified, Real-Time Object Detection". In: *CoRR* abs/1506.02640 (2015). arXiv: 1506.02640. URL: http://arxiv.org/abs/1506.02640.