Patch-Based Adversarial Attack on DeepISP

Venkata Satya Sai Subhash Vadlamani
University of Florida

Akhil Krishna Chatla
University of Florida

Chenna Kesava Varaprasad Korlapati
University of Florida

Venkata Suresh Yarava
University of Florida

1 Introduction

The integration of Deep Neural Networks (DNNs) into critical applications such as healthcare, autonomous driving, and surveillance has emphasized the importance of robust Deep Learning based Image Signal Processing (ISP) systems. Traditionally, ISPs in cameras were tailored to specific hardware, limiting flexibility and increasing costs. However, with the advent of deep learning, a shift towards more adaptable, software-driven approaches like the DeepISP model [11] has occurred. This model, powered by Convolutional Neural Networks (CNNs), consolidates demosaicing, denoising, and color correction processes into a unified process, enhancing image quality and hardware efficiency. These improvements could lead to more compact and cost-effective camera systems applicable across various sectors.

Despite these advancements, deep learning-based ISPs like DeepISP face challenges including high computational demands, latency issues in real-time processing, and extensive data needs for training. Moreover, these systems are susceptible to security risks from adversarial attacks, which involve modifying inputs subtly to deceive models while remaining imperceptible to humans [1,10] . One such vulnerability is the patch-based adversarial attacks that exploit model weaknesses to alter outputs critically [2].

While numerous studies have explored the vulnerability of deep learning models to adversarial attacks [14], our focus is on the DeepISP model, a deep learning-based ISP. Through this study, we seek to understand how patch-based adversarial attack affect the functionality of such ISPs, specifically in terms of image quality and object detection. We employ the Structural Similarity Index (SSIM) [13] to evaluate image reconstruction and object detection using YOLOv8 model to assess the impact on object detection capabilities. SSIM measures the structural similarity between two images, reflecting image degradation post-attack and quantifying changes in visual quality.

Our findings reveal significant vulnerabilities within the DeepISP model, highlighting the critical need for robust defense mechanisms to secure deep learning-based ISPs in practical applications. We have demonstrated successful object misclassification through patch-based adversarial attack, for instance, a car being detected as a truck as shown in Figure 1, which underscores the practical threats posed by these attacks. Through our work, we reduced the average object detection rate from 88% to 53%. With these results, we aim to pave the way for future innovation in image processing technologies that are both secure and effective, transforming how machines perceive and interact with the visual world.

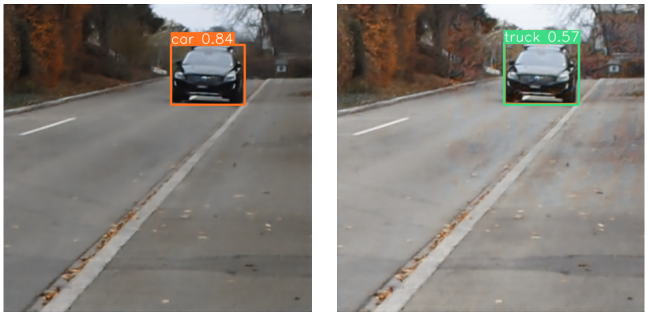

Figure 1: An example showing object detection by YOLOv8 [7] model on images processed by DeepISP model. Left Image: Object detection on original image. Right Image: Object detection on a patched image.

2 Background and Related Work

Image Signal Processing (ISP) traditionally optimizes images for human vision, which involve complex operations like noise reduction, color correction, and enhancement for aesthetic appeal. However, in machine-centric applications such as autonomous driving and automated surveillance, these traditional ISPs may fall short due to their fixed, hardware-specific pipelines. Deep learning-based ISPs like DeepISP, represent a significant paradigm shift by processing raw sensor data directly through learned models, thus bypassing conventional ISP constraints. This approach not only improves computational efficiency but also enhances image interpretability which is crucial for machine vision applications [5,6,11].

Lubana et al. (2019) proposed a minimalistic ISP framework that maintains energy efficiency and model accuracy, highlighting the benefits of simplifying ISP operations for deep learning applications [8]. Further studies, including those by Schwartz et al. (2019) and Hansen et al. (2021), have demonstrated that integrating traditional ISP tasks into deep learning models can significantly improve adaptability and performance under diverse conditions, proving that certain ISP components are indispensable for training robust CNNs [4, 11].

Adversarial attacks significantly impact machine vision, particularly affecting Convolutional Neural Networks (CNNs) through subtle data manipulations that deceive models into erroneous outcomes. Leveraging foundational research from visible patch attacks to covert ISP exploits, our study implements a patch-based attack on the DeepISP model using the Fast Gradient Sign Method (FGSM) [3]. This exploration into low-level manipulations enables us to bypass traditional security measures and assess vulnerabilities effectively [9, 12].

The evolution of adversarial methods necessitates enhancing the security of deep learning ISPs, and underscoring their crucial role in improving the performance and security of vision systems.

3 Methodology

a) **Threat Model**

Our model mainly focuses on the deep Learning based ISP pipeline to process the output data of the Image Sensor. The core vulnerability exploited is the manipulation of pixel values in raw images in such a way that it generates low quality images leading to erroneous object detection. We adopt a black-box perspective when interacting with the model, meaning our access is limited to observing its inputs and outputs without insight into its internal workings. Attack vectors include making slight modifications to the raw images to introduce pertubations that are barely noticeable, tricking the ISP model into making inaccurate adjustments while processing the images. Potential impacts of successful attacks include misclassification of objects within the image, degraded image quality, and loss of trust in the reliability of the deep learning ISP models.

b) **Dataset**

We used the Zurich dataset, which is a collection of 20,000 photos taken with two cameras: a Huawei P20 smartphone and a Canon 5D Mark IV camera [6]. This dataset is divided into various folders: train, test and full_resolution. We primarily focused on the images in the test folder, which includes a subfolder named huawei_visualized. This subfolder contains the original images which were used as reference for evaluation. For performing the attack, we used the raw images found in the huawei_raw folder.

c) **Attack Methodology**

In our research, we assess the robustness of the DeepISP model against adversarial patch attack using the Fast Gradient Sign Method (FGSM), as introduced in [3]. Our attack is an untargeted attack where the goal is not to misclassify a specific object but rather to increase the likelihood of any object being misclassified. We define our loss function based on the structural similarity index (SSIM), as defined in [13]. By manipulating the patch denoted by P, applied on the raw image, we aim to maximize the loss denoted by $L(P)$, specifically by reducing SSIM between the original image and modified image (original image with patch applied to it), and then evaluate how good the model is in generating the modified image.

The loss function $L(P)$ for an adversarial patch P, used in FGSM is calculated as:

$$L(P) = 1 - \text{SSIM}(I_{\text{original}}, I_{\text{patched}}) \tag{1}$$

where I_{original} and I_{patched} are the original image and modified image (original image with patch) respectively.

The patch is adjusted in the direction of sign of gradient of the loss, given by $\nabla_P L(P)$, and scaled by ε — a small constant to control intensity of perturbation. The update rule for the patch is given by:

$$P_{\text{new}} = P + \varepsilon \cdot \text{sign}(\nabla_P L(P)) \tag{2}$$

To ensure pixel values are within the valid range [0, 1], the pixel values in the patch are clipped according to the formula given by:

$$P_{\text{final}} = \text{clip}(P_{\text{new}}, 0, 1) \tag{3}$$

This process is repeated for multiple iterations to maximize the loss until one of the stopping criteria is met, which are given by (1) when the loss exceeds *loss_threshold* (this can be adjusted to keep the patch imperceptible in the patched image), and (2) number of iterations reach *max_iterations* (An upper bound of number of iterations to be run before the patch optimization stops).

Then, SSIM score between original image, patched image obtained after exiting this process is calculated. This SSIM score tells us how similar the patched image was w.r.t original image.

We are evaluating this method's effectiveness in fooling the model and also examine how different patch sizes, ranging from 10% to 90% of the image area, affect the model's performance.

4 Experimental Results

In this section, we detail the experimental setup and results obtained from our study focusing on the impact of adversarial patches using the FGSM attack on image processing and subsequently how it impacts object detection in processed

Patch Size (in percentage)	SSIM		
	(Original, Predicted Original)	(Original, Predicted Patched)	(Predicted Original, Predicted Patched)
10	0.694732	0.690434	0.996163
20	0.694732	0.689393	0.995132
30	0.694732	0.687310	0.995558
40	0.694732	0.684764	0.995709
50	0.694732	0.682295	0.995961
60	0.694732	0.681215	0.996316
70	0.694732	0.677499	0.996838
80	0.694732	0.670100	0.995883
90	0.694732	0.654170	0.996903

Table 1: Mean SSIM Values for Various Patch Sizes. *SSIM (Original, Predicted Original)* refers to SSIM between the original image from the dataset and the predicted image of the raw image by the model. *SSIM (Original, Predicted Patched)* is the average of SSIM between non-patch regions in the original image from the dataset and the predicted image of the patched raw image by the model. *SSIM (Predicted Original, Predicted Patched)* is the average of SSIM between non-patch regions in the predicted image of the original raw image by the model and the predicted image of the patched raw image by the model. All the SSIM entries in the table are mean value of 10 different images.

images by YOLOv8 model [7]. We performed two different experiments to explore the following outcomes (1) Structural Similarity of non patched regions in the image, and (2) Effectiveness of Object detection on patched images.

Experiment 1: Structural Similarity Assessment

In this experiment, we aim to evaluate the influence of an adversarial patch on the global region of an image, specifically areas outside the patch. For doing this, we have taken a set of 10 images from the dataset, performed FGSM attack on each of the raw image with different patch sizes (starting with 10% and till 90% with increment of 10). The parameters for the FGSM attack were $\varepsilon = 0.1/255$, max_iterations = 15, and loss_threshold = 0.01.

Then we measured the SSIM of following image pairs (Original, Predicted Original), (Original, Predicted Patched), (Predicted Original, Predicted Patched) for all 10 images and calculated the mean of SSIM of the 10 images for each image pair. This step was repeated for different patch sizes and the results were tabulated in Table 1.

Experiment 2: Impact on Object Detection

In this experiment, we assessed the effect on object detection in patched images generated by the DeepISP model. This involved processing both original (unpatched) and patched raw images by the DeepISP model and subsequently detecting objects in the processed images using YOLOv8 model [7]. For this experiment, we used a set of 10 images from the dataset containing objects, with the same settings for the adversarial attack as in Experiment 1 but adjusted loss_threshold to 0.08.

The object detection model exhibited a decrease in performance on the patched images with an average detection rate dropping from 88% on original images set to 53% on patched images set. Notably, out of 17 objects present in the original set of images, 8 were misclassified in the patched versions, indicating a significant degradation due to patch based attack. Average detection rate is defined as the percentage of number of objects correctly detected to the number of objects present in set of images.

5 Analysis and Discussion

The results in experiment 1 indicate minimal impact of the patch on the non-patched regions, as evidenced by the high SSIM values (close to 1) between (Predicted Original, Predicted Patched) image pair. The noticeable decrease in SSIM values of (Original, Predicted Original) and (Original, Predicted Patched) image pairs reflect the challenges faced by the DeepISP model to preserve the characteristics of original image accurately while processing the raw images.

The results in experiment 2 suggest a significant impact of the adversarial patches on the object detection capabilities of YOLOv8 model, which is evident from the decline in average object detection rate which dropped from (88%) in original images, to (53%) in patched images. This indicate that DeepISP model is vulnerable to adversarial patch attacks which consequently effects the object detection rate on the processed images.

Finally, these two experiments demonstrate that while the adversarial patches have little effect on the structural similarity of non-patched regions, they significantly impair the performance of object detection on patched images. These findings suggests that even small adversarial interventions can profoundly affect the downstream tasks of advanced image processing models like DeepISP, impacting their practical deployment in sensitive applications.

References

[1] Naveed Akhtar and Ajmal Mian. Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey. *IEEE Access*, 6:14410–14430, 2018.

[2] Tom B. Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *CoRR*, abs/1712.09665, 2017.

[3] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2015.

[4] Patrick Hansen, Alexey Vilkin, Yury Khrustalev, James Imber, David Hanwell, Matthew Mattina, and Paul N. Whatmough. Isp4ml: Understanding the role of image signal processing in efficient deep learning vision systems, 2021.

[5] Andrey Ignatov, Cheng-Ming Chiang, Hsien-Kai Kuo, Anastasia Sycheva, Radu Timofte, Min-Hung Chen, Man-Yu Lee, Yu-Syuan Xu, Yu Tseng, Shusong Xu, Jin Guo, Chao-Hung Chen, Ming-Chun Hsyu, Wen-Chia Tsai, Chao-Wei Chen, Grigory Malivenko, Minsu Kwon, Myungje Lee, Jaeyoon Yoo, and Etienne de Stoutz. Learned smartphone isp on mobile npus with deep learning, mobile ai 2021 challenge: Report. pages 2503–2514, 06 2021.

[6] Andrey Ignatov, Luc Van Gool, and Radu Timofte. Replacing Mobile Camera ISP with a Single Deep Learning Model. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020.

[7] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLO, January 2023.

[8] Ekdeep Singh Lubana, Robert P. Dick, Vinayak Aggarwal, and Pyari Mohan Pradhan. Minimalistic image signal processing for deep learning applications. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 4165–4169, 2019.

[9] Buu Phan, Fahim Mannan, and Felix Heide. Adversarial imaging pipelines, 2021.

[10] Kui Ren, Tianhang Zheng, Zhan Qin, and Xue Liu. Adversarial attacks and defenses in deep learning. *Engineering*, 6(3):346–360, 2020.

[11] Eli Schwartz, Raja Giryes, and Alex M. Bronstein. Deepisp: Toward Learning an End-to-End Image Processing Pipeline. *IEEE Transactions on Image Processing*, 28(2):912–923, February 2019.

[12] Abhijith Sharma, Yijun Bian, Phil Munz, and Apurva Narayan. Adversarial patch attacks and defences in vision-based tasks: A survey, 2022.

[13] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[14] Rui Zhao. The vulnerability of the neural networks against adversarial examples in deep learning algorithms, 2020.