Building upon previous research in adversarial attacks on image classifiers, this we plan to introduce a method for creating universal, robust, and targeted adversarial image patches applicable in real-world scenarios. These patches possess universality, demonstrating their effectiveness in attacking any scene, robustness under diverse transformations, and targeting capability to manipulate classifiers into outputting any desired target class. Our approach enables the seamless integration of these adversarial patches into scenes, whether physically printed or digitally presented through photography. Even when small, these patches prompt classifiers to prioritize the targeted class over other scene elements, leading to misclassification. To achieve these results, we propose a methodology that combines advanced optimization techniques with insights from adversarial machine learning.

Extensive experimentation and validation show that adversarial image patches are placed in different AI models, such as YOLOv8, causing the model to misclassify and give inaccurate predictions. The experiments involve systematically generating adversarial patches and evaluating their impact on various AI models under different conditions, including different scene compositions, lighting conditions, and camera angles. Through rigorous experimentation, we demonstrate the effectiveness and robustness of our method across diverse scenarios.

Moreover, our method incorporates a defense strategy by augmenting the model to analyze image patches and identify when adversarial patches are used to deceive the model, enhancing its resilience against adversarial attacks while maintaining model performance. This defense mechanism involves leveraging techniques from anomaly detection and outlier analysis to detect and flag suspicious patches during inference. By integrating this defense strategy into the model architecture, we aim to provide a comprehensive solution for addressing the threat of adversarial attacks in real-world applications.

In summary, our research aims to develop a practical and effective approach for creating and defending against adversarial image patches in AI object detection models. By combining innovative techniques in patch generation, model augmentation, and defense strategies, we strive to advance the state-of-the-art in adversarial machine learning and contribute to the development of more robust and reliable AI systems for real-world deployment.