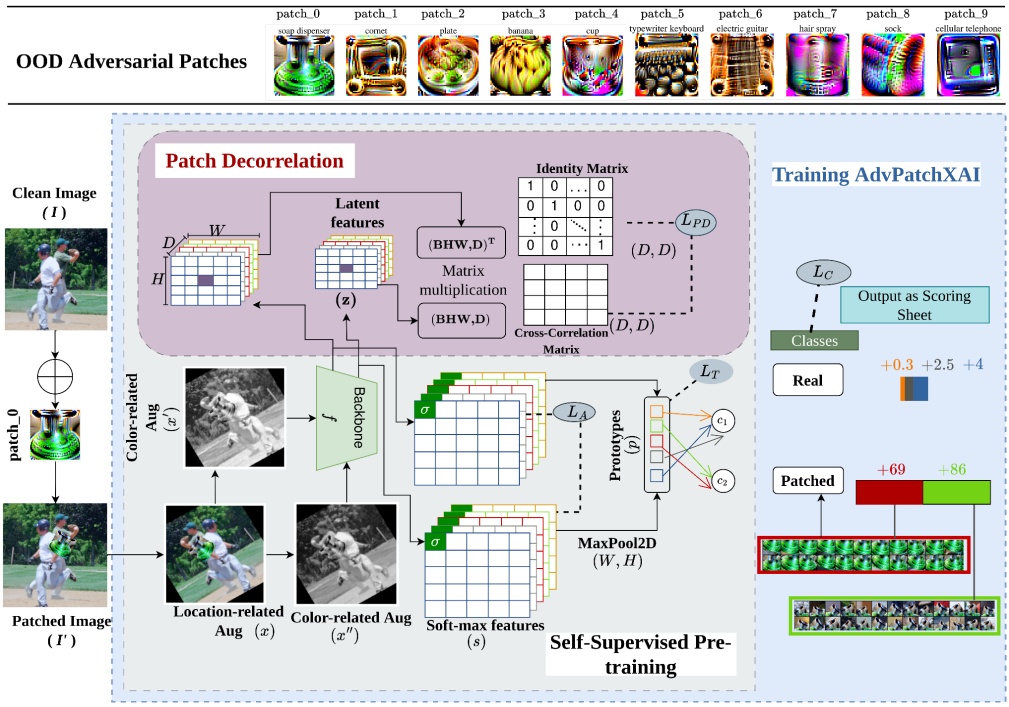
A universal adversarial patch detector to strengthen the AI security systems



Dr. Akshay Agarwal's "**Trustworthy BiometraVision Lab**", Department of Data Science and Engineering has developed a novel adversarial patch detector, AdvPatchXAI, that significantly strengthens modern security AI systems against physical adversarial threats. Adversarial patch attack is a process of adding a patch in an image that can fool any AI systems. Unlike conventional defenses, AdvPatchXAI employs a patch decorrelation strategy to reduce feature redundancy, enhancing robustness and interpretability. It excels at detecting and neutralizing unseen adversarial patches, even under corrupted conditions, thus aligning model decisions more closely with human perception. Additionally, its self-supervised approach provides clearer insights into how the model identifies threats. This groundbreaking work has been accepted for publication in the prestigious ***IEEE/CVF Computer Vision and Pattern Recognition (CVPR2025)***, marking a major step forward in securing AI systems.

https://tbvl22.github.io/website