# Saliuitl: Ensemble Salience Guided Recovery of Adversarial Patches against CNNs

**Mauricio Byrd Victorica**, György Dán, Henrik Sandberg
KTH Royal Institute of Technology, Sweden
{mbv, gyuri, hsan}@kth.se

Adversarial patches are capable of misleading computer vision systems based on convolutional neural networks. Existing recovery methods suffer of at least one of three fundamental shortcomings: no information about the presence of patches in the scene, inability to efficiently handle noncontiguous patch attacks, and a strong reliance on fixed saliency thresholds. We propose Saliuitl, a recovery method independent of the number of patches and their shape, which unlike prior works, explicitly detects patch attacks before attempting recovery. In our approach, detection is based on the attributes of a binarized feature map ensemble, which is generated by using an ensemble of saliency thresholds. If an attack is detected, Saliuitl recovers clean predictions locating patches guided by an ensemble of binarized feature maps and inpainting them. We evaluate Saliuitl on widely used object detection and image classification benchmarks from the adversarial patch literature, and our results show that compared to recent state-of-the-art defenses, Saliuitl achieves a recovery rate up to 97.81 and 42.63 percentage points higher at the same rate of lost predictions for image classification and object detection, respectively. By design, Saliuitl has low computational complexity and is robust to adaptive white-box attacks. Our code is available at `https://github.com/Saliuitl/Saliuitl/tree/main`.

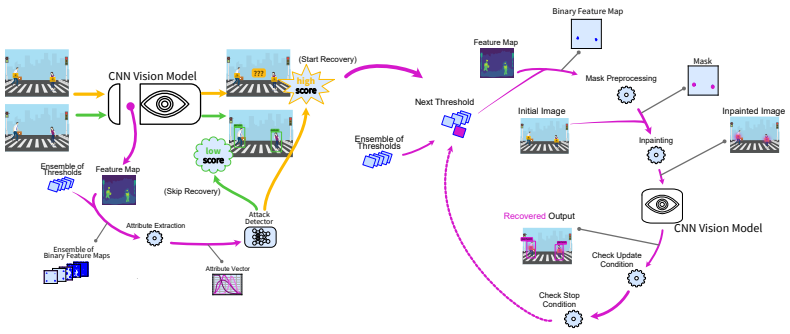


Figure 1: *Saliuitl*: During the detection stage (left), (i) an ensemble of binary feature maps is computed according to a threshold ensemble, (ii) attribute extraction is applied over the ensemble and the results are aggregated into an attribute vector, and (iii) an attack detector is fed with the attribute vector. If *Saliuitl* detects adversarial patches, it will perform the recovery stage (right) on the output for the initial input, guided by the binary feature maps computed according to the threshold ensemble.