**Physical /Patch**

Physical adversarial attacks are directly deployed on the object in the real word by adding patches or stickers on or near the targeted object that will cause the model to “misclassify” the object. In contrast to adding imperceptible perturbation to the image , patches are salient yet often innocuous looking materials . They are created by maximising the loss function of the network. Since they are often kept in different locations and orientations in the proximity of the targeted object they essentially becomes the prominent feature of the image resulting the targeted object to be treated like a background [AP]

The misclassification was proved to be not a result of overfitting by [IPN] using **cross-training-set generalisation**: A model trained on a completely different training set than the one on which the model used to generate the adversarial samples was trained, misclassified the images ; **cross model generalisation:** A model trained from scratch with different hyperparameters than the model used to generate the adversarial samples, misclassified the images

There are several ways patches can be adversarial according to [SIB]:

* Objectness Based Attacks
* Classification Based attacks

Objectness Based Attacks:

This is achieved by Feature Interference Reinforcement (FIR) and Enhanced Realistic Constraint Generation (ERCS).

FIR is a white box technique which works by causing the adversarial example to perturb the earlier hidden layers of the model . In order to achieve this , the model is trained with a ‘normal’ and an ‘adversarial’ versions of the same image. At each layer, a feature vector of the targeted object in both the images are created. One feature value is extracted from one feature map using mean pooling. The loss function is defined as a difference between the feature vectors and the goal is to maximise this loss function.

ERCS works by creating adversarial examples subject to realistic constraints . These constraints ensures that the relationship of the object to its background and the object semantic integrity is maintained in the creation of adversarial examples.

Classification Based Attacks:

Nested AE is based on the hypothesis that the part of an object detection model like YOLOV3 that detects objects at small scale (at long distance) is easier to be ‘fooled’ as they only consider a small number of pixels. In this method we target the central area of the object in short distance range , add perturbation only to that area as once the object is in a long distance (hence occupying a small change in the image) , the whole object would occupy only the aforementioned central area.

[AP] : Adversarial Patch : <https://arxiv.org/abs/1712.09665>

[IPN]: Intriguing properties of neural networks : <https://arxiv.org/pdf/1312.6199>

[SIB]: Seeing isn’t Believing: Towards More Robust Adversarial Attack Against Real World Object Detectors : <https://sci-hub.se/https:/doi.org/10.1145/3319535.3354259>