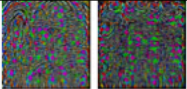
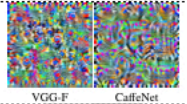






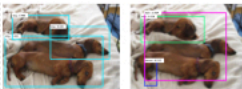




手法 \ 観点	Digital or Physical	Classifier or Detector	(摂動作成時) 使用データ	摂動の加え方	知覚しづらさの 定義	White box or Black box	特徴	例
[1610.08401] Universal adversarial perturbations	Digital	Classifier	学習データ	画像全体	l_2, l_∞	White box	Data universal Transferability 有	
[1707.05572] Fast Feature Fool: A data independent approach to universal adversarial perturbations	Digital	Classifier	必要なし	画像全体	l_2, l_∞	White box	Data universal	
[1707.08945] Robust physical-world attacks on deep learning visual classification	Physical	Both	攻撃対象データ	対象物のみ	落書きっぽさ	White box	Mask を使って 対象物のみ摂動	
[1710.08864] One pixel attack for fooling deep neural networks	Digital	Classifier	攻撃対象データ	画像全体 (数ピクセル)	l_0	Black box	進化的アルゴリズム	
[1710.11342] Generating natural adversarial examples	Digital	Classifier	学習データ 攻撃対象データ	潜在空間 (画像全体)	潜在空間での l_2	Black box	GAN で対象画像を生成	
[1808.02651] Beyond pixel norms: parametric adversaries using an analytically differentiable renderer	Digital (防御は Physical)	Classifier	3D object データ	対象物のみ	定性的 (色味が 変わるのみ)	White box	微分レンダラで 対象物のみ摂動	
[1809.04098] On the Structural Sensitivity of Deep Convolutional Networks to the Directions of Fourier Basis Functions	Digital	Classifier	学習データ	画像全体	l_∞	Black box (ラベルのみ)	フーリエ基底を 用いた摂動	
[1904.00759] Adversarial camera stickers: A physical camera-based attack on deep learning systems	Physical	Classifier	攻撃対象データ	画像全体	ヒューリスティック	White box	画像を撮影する カメラに摂動	
これより上は Classifier これより下は Detector								
[1703.08603] Adversarial Examples for Semantic Segmentation and Object Detection	Digital	Detector	攻撃対象データ	画像全体	l_2	White box	Segmentation と obj. detection	
[1802.06430] DARTS: Deceiving Autonomous Cars with Toxic Signs	Physical	Detect-Classify pipeline	攻撃対象データ	対象物のみ	l_2	White box	古典的物体検出 を含む pipeline	
[1804.05810] ShapeShifter: Robust Physical Adversarial Attack on Faster R-CNN Object Detector	Physical	Detector	攻撃対象データ	対象物のみ	l_2	White-box	実環境でも高い 攻撃成功率	
[1902.02067] Daedalus: Breaking Non-Maximum Suppression in Object Detection via Adversarial Examples	Digital	Detector	攻撃対象データ	画像全体	l_0, l_2	White box	NMS を攻撃して 大量 bbox 発生	