手法 \ 観点	基本戦略	使用する 外部リソース	対応コスト（人的コスト, 推論コスト）	特徴
[1412.6572] Explaining and Harnessing Adversarial Examples	正則化	Adv. example	人的コスト：中、推論コスト：小 データ作成とモデルの再学習	モデル学習時に adv. example を使った loss も正則化として加える
[1511.04508] Distillation as a defense to adversarial perturbations against deep neural networks	外部リソース使用	Student nettwork	人的コスト：大、推論コスト：小 モデルの構造変更と再学習	温度を入れてノイズ鋭敏性を抑えた上で蒸留
[1702.04267] On Detecting Adversarial Perturbations	外部リソース使用	binary classifier	人的コスト：中、推論コスト：大 別モデルの導入とその学習	adv. example か否かを判別する二値分類器を構築
[1705.09064] Magnet: A two-pronged defense against adversarial examples	外部リソース使用	（複数個の） Auto Encoder	人的コスト：中、推論コスト：大 別モデルの導入とその学習	（複数個の）AE で adv. example か否かの検出と 再構成時の変化量を検証
[1710.10766] PixelDefend: Leveraging Generative Models to Understand and Defend against Adversarial Examples	入力データ変更	Pixel CNN	人的コスト：中、推論コスト：大 別モデルの導入とその学習	pre-train された PixelCNN に通して purify して 予測モデルの入力とする
[1711.01991] Mitigating Adversarial Effects Through Randomization	入力データ変更	必要なし	人的コスト：小、推論コスト：小 入力データを変更	入力を random に resize して padding する層を追加して モデルを構築
[1712.02976] Defense against Adversarial Attacks Using High-Level Representation Guided Denoiser	外部リソース使用	Adv. example Denoiser モデル	人的コスト：中、推論コスト：大 別モデルの導入とその学習	noise を除去する U-net にデータを通して 予測モデルの入力とする
Thermometer Encoding: One Hot Way To Resist Adversarial Examples	予測モデル変更	必要なし	人的コスト：中、推論コスト：中 入力データを変更	ある閾値以上を全部 1 にする thermometer encoding で非線形に入力データを離散化したモデルを構築
[1803.01442] Stochastic Activation Pruning for Robust Adversarial Defense	予測モデル変更	必要なし	人的コスト：小、推論コスト：中 モデルの中間層を操作して予測	予測時に各層の activation 出力を random に落とす
[1803.06373] Adversarial Logit Pairing	正則化	adv. example	人的コスト：中、推論コスト：小 データ作成とモデルの再学習	Adv. training にさらに logit の l_2 loss を正則化として加える
[1805.06605] Defense-GAN: Protecting Classifiers Against Adversarial Attacks Using Generative Models	外部リソース使用	Adv. example GAN	人的コスト：中、推論コスト：大 別モデルの導入とその学習	Clean データで学習した GAN を使って、複数の random seed から生成した画像で入力に近いものを入力とする
[1812.03411] Feature Denoising for Improving Adversarial Robustness	予測モデル変更	必要なし	人的コスト：大、推論コスト：小 モデルの構造変更と再学習	非局所的な重み付き和などの denoising block を モデルに取り入れる
[1903.01612] Defense Against Adversarial Images using Web-Scale Nearest-Neighbor Search	外部リソース使用	外部 DB	人的コスト：大、推論コスト：大 別モデルの導入と外部データの準備	数百億の画像から類似画像を検索して最近傍画像を 予測モデルに入力