Part I
References

1. Machine Learning Security

- Trustworthiness [1, 2, 3];
- Adversarial Machine Learning Industry Survey [4];
- **Media Forensics**:
  - Splicing Forgery (Detection) [5, 6];
  - Camera Model Identification [7];
  - Frame Dropping (Detection) [8];
- **Privacy**:
  - Avoiding Person Recognition (e.g. [9, 10, 11]);
    - Game-Theoretic Approach [11];
  - Privacy Attributes in Images [12];
    - Classification [12];
    - Obfuscation [13];
- **Adversarial Examples** [14, 15]
  - Early Work:
    - [16, 17];
    - [18];
    - Learning with Invariances (SVMs) [19];
    - Deep Learning with Invariances [20];
  - Surveys and Reviews [21, 22, 23, 24, 25, 26, 27];
  - Best Practices [28, 29];
  - Other ML Models:
    - Logistic Regression [30];
    - SVMs [31];
    - (Boosted) Decision Stumps [32, 33];
    - K-NN [34, 35];
    - Non-Parametric Models in General [36];
  - Attacks:
    - White-Box/Gray-Box:
      - L-BFGS [37];
      - FGSM [38];
      - PGD [39];
      - CW [40] (w/ different parameterization [41]);
      - Transferability [42];
      - Physical [43, 44];
      - Adversarial Saliency Maps [45];
      - Logits Attack (LOTS) [46];

- Momentum Attacks [47];
- Imperceptible Attack [48];
- GAN-based [49, 50, 51, 52, 53];
- Confidence/Uncertainty Attack (Gaussian Processes) [54];
- OptMargin [55];
- ANGRI/UPSET (Predicting Adversarial Examples) [56];
- Blind-Spot Attack [57];
- Restricted Adversarial Example [58];
- WITCHCraft: PGD with Random Step Size [59];
- Feature Disruptive Attack [60];
- Entropy-Based BIM [61];
- Wasserstein Adversarial Examples [62];
- L_1 L_2 Elastic-Net Attack [63];
- Structured/Sparse Attack [64];
- Combinatioral Attack on Binarized Networks [65];
- Unsuspicious Adversarial Examples [66];
- Top-k Adversarial Examples [67];
- L_1 APGD [68];
- L_1 DLA [69]
- ALMA [70];
- AdvDrop (Frequency Domain) [71];
- MultiTargeted [72];
- Multiple Targeted Attacks on Detection [73];
- Sparse-RS L_0 [74];
- SPSPA Finite Difference [75];
- Distributionally Aadversarial Attack L_∞ [76];
- Gradient-Based Boundary Attack L_∞, L_2, L_1, L_0 [77];
- DDN L_2 [78] (for other L_p [79]);
- SparseFool [80];
- PGD with Attention and GNN [81, 82];

* Black-Box:
  - ZOO [83];
  - Limited Query and Information Attack (using Gaussian Gradient Estimation) [84];
  - Gradient Estimation with Query Reduction [85];
  - Boundary Attack [86];
  - Biased Boundary Attack [87];
  - Boudnary Attack ++ [88];
  - One-Pixel Attack [89];
  - Ensemble-Attacks [42, 47, 90];
  - Randomized Ensemble-Attacks [91];
  - Bandit Attack [92];
  - Simple Geometric Attack [93, Sec 8.3];
  - Intermediate Level/Transfer Attack [94];
  - CapsAttack [95];
  - "Cubes" Attack / Simple Attack [96, 97];
  - Low-Frequency Boundary Attack [98];

        - Simple Attack [99];
        - Sparse and Imperceivable L_0 Attack [100];
        - Black-Box with Initialiation from Other Images [101];
        - GeoDA [102];
        - Random Sign Flip [103];
        - SignHunter [104, 105];
        - Meta-Learned Transfer Attack [106];
        - Transferable Targeted Attacks [107];
        - Admix Transferable Attacks [108];
        - Aha Attack (Decision-Based) [109];
        - Square Attack [97];
        - PPBA [110];
        - HopSkipJumpAttack [111];
        - SurFree [112];
        - Customized Boundary Attack [113];
        - ZO.NGD [114];
        - Local Black-Box Attack [115];
        - Meta-Learned Black-Box Attack [116];
        - Sign-OPT [117];
        - RayS [118];
        - Bayes Attack [119];
    * "On-Manifold":
        - Adversarial Spheres [120];
        - Adversarial Lighting/Physics [121];
        - On-Manifold Test Generation [122];
        - On-Manifold Data Augmentation for Calibration [123];
- Defenses;
    * Robust Nearest Neighbor [124];
    * Robust Linear Classifiers [125];
    * Deep RBF Network [126, 127];
    * Last RBF Layer [128, 129];
    * Deep k-Nearest-Neighbor [130];
        - Robustness Questionable [35];
    * Adversarial Training (and Variants) [131, 132, 133, 134, 135, 136, 39];
        - Robust Optimization [133];
        - Distributional Robust Optimization [134];
        - Generative Adversarial Training [135];
        - Convex Outer Adversarial Polytope [137] (Extension with Cost Matrix [138]);
        - Curriculum Adversarial Training [139];
        - Single-Step Adversarial Training [140];
        - Adversarial Training and Confidence-Based Rejection [141];
        - Interpolated Adversarial Training [142];
        - Smoothed Adversarial Training [143];
        - Feature-Activated Adversarial Training [144];
        - Sliced Wasserstein Adversarial Training [145];
        - Progressive Adversarial Training [146];

- Bayesian Adversarial Training [147, 148];
- Adversarial Training with Unlabeled Data [149, 150];
- Adversarial Training with Abstain Option [151];
- Instance Adaptive Adversarial Training [152];
- Adversarial Training on Adversarial Examples from a Reference Model [153];
- MMA Training (Instance Adaptive ϵ) [154];
- Feature Scattering Adversarial Training [155];
- Bilateral Adversarial Training [156];
- Cascade Single-Step Adversarial Training [157];
- Adversarial Training with Trimmed Kinetic Loss (with Sensible Adversarial Examples) [158];
- Calibratable/In-Situ Adversarial Training [159];
- Calibrated Adversarial Training [160];
- Customized Adversarial Training [161];
- Adversarial Training with Feature+Weight Normalization and Angular Means [162];
- Adversarial Training on Stylized Adversarial Examples [163];
- Progressive Diversified Adversarial Training [164];
- Free Adversarial Training [165];
- Fast Adversarial Training [166];
- Unrestricted/Blind Adversarial Training [167];
- Diversified Initialized Perturbations Adversarial Training [168];
- Single-Step Adversarial Training with Dropout Scheduling [169];
- Adversarial Training with Smooth ReLU [170];
- Adversarial Training with Curvature Regularization [171];
- Friendly Adversarial Training [172];
- Adversarial Training with Bi-Directional Likelihood Regularization [173];
- Learning to Learn Adversarial Training [174];
- Adversarial Training with Local and Global Compactness [175];
- Adversarial Concurrent Training [176];
- Instance-Reweighted Adversarial Training [177];
- Adversarial Training with Adversarial Weight Perturbations [178];
- Adversarial Training with Noisy Weights [179];
- Learnable Boundary Guided Adversarial Training [180];
- TRADES [181];
- Adversarial Fine-Tuning [182];
- Adversarial Training with GMM-Inspired Distance Regularizer [183];
- L_1 Adversarial Training [69, 68];
- Fast Adversarial Training via Latent Perturbations [184];
- Adversarial Training with Regularized Embedding [185];
- Joint Energy Adversarial Training (for Generation) [186];
- Generative Adversarial Training [187];
- Robust Soft Label Adversarial Distillation (RSLAD) [188];

* Multiple Perturbations:
  - Multi-Attack/Worst-Case Adversarial Training [189, 190];
  - Stochastic Adversarial Training + Meta-Noise Generator [191];
  - Analysis-by-Synthesis [52, 53];
  - Shaped Noise Augmented Processing + Adversarial Training [192];

- Fine-Tuning [193];
* Ensembles [194, 195, 196, 197, 198];
  - Ensemble of Mixed-Precision Networks [199];
* Ensemble Adversarial Training [200, 201, 202];
  - Stochastic Weight Averaging Adversarial Training [203];
* Bounded ReLU + Gaussian Data Augmentation [204];
  - Not Robust [205];
* Feature Squeezing [206];
* PCA [207];
* Saturating Networks [208];
* Distillation [209];
  - Not Robust [210, 41];
* Adversarially Robust Distillation [211];
* Distillation in Bayesan Neural Networks [124];
* Iterative GAN/Deep Image Prior Projection [212, 213];
  - With GANs [214];
* Randomization [215]:
  - Dropout [216];
  - SmoothBlock/DropBlock [217];
  - (Random) Image Transformations [218, 219];
  - Ensembles or Random Weight Perturbations [194];
* Regularization:
  - Gradient Regularization (Sometimes in Combination with Adversarial Training) [220, 221, 222, 223, 224, 225, 226]:
  - Lipschitz-based [227, 222];
  - Jacobian Regularization through Discriminator [228];
  - Not Robust [229];
  - Confidence Regularization (Motivated by Fisher Information Matrix) [230];
  - Spectral Norm Regularization [231];
  - Bit-Plance Consistency Regularizer [232];
  - Entropy Regularization [233];
  - Scale-Invariant Weight Regularization (WEISSI) [234];
  - Perceptual Adversarial Training (Multiple Threat Models) [235];
  - Local Linear Regularization [236];
* Discretization [237];
* Adaptive JPEG quantization [238];
* Rectification/Detection [239];
* Out-of-Distribution Training [240, 241];
* Adversarial Logit Pairing [242];
  - Not Robust [243, 244];
* Fortified Networks [245];
* Adversarial Perturbation Eliminating GAN [246];
* Label Smoothing and Feature Squeezing [247];
  - Not Robust [244];
* Attacks meet Interpretability [248];
  - Not Robust [249];

* Region-Based Classification / Randomized Smoothing [250, 251, 252, 253];
  - Curse of Dimensionaility for Randomized Smoothing [254, 255];
  - Randomized Smoothing for L_∞ [256];
  - Randomized Smoothing for Pre-Trained Classifiers [257];
  - L_1 and L_∞ Randomized Smoothing With Various Distributions [258];
  - L_0 [259];
  - L_∞ [260, 261];
  - Training Strategies:
  - Smoothed Adversarially Trained Classifiers [143];
  - AdvSmooth with Noise Consistency Regularization [262];
  - MACER (No Adversarial Training, Maximize Radius by Maximizing Difference Between Class and Runner-Up Class) [263];
  - Combination with Mixup [264];
  - Essentially Methods to Certify Smoothed Classifiers:
  - Cheaper Certificates for General L_p [265];
  - Higher-Order Certificates [266];
* Compact Convolution [267];
* MagNet (Detection + Auto-Encoding);
  - Only Auto-Encoding [268];
  - Not Robust [205];
* Random Forests [269];
* BoW Networks for Detection [270, 271];
* Feature Denoising [272];
  - Not Robust [273];
* Parseval Networks [274];
* Logit Inspection [275];
* Adaptive Networks/Normalization [276];
* Randomized Discretization [277];
* Patch-Based Denoising [278];
* Web-Scale Nearest Neighbor [279];
* Tent Activation Function [280];
* Denoising Auto-Encoder [281];
* Adversarial Model Cascades [282];
* Structure-to-Noise Autoencoders [283];
* Training on Quantized Images [284];
* Prototype Conformity Loss [285, 286];
  - Not Robust [229];
* Radial Basis Convolutional Layer [287];
  - Not Robust [229];
* Blind Adversarial Pruning [**XieARXIV2020b**];
* RBF-Based Manifold Defense [288];
* RAIN (Randomized Shift/Down/Up-Sampling) [289];
* Retinal Fixation Sampling [290];
* Random Distortions over Grid [291];
* Label Smoothing with Bounded Logits [292];
* Adversarial Batch Normalization [293];

    - SSP ResNets [294];
    - Avoid Gradient Leaking [295];
    - Spiking Neural Networks [296];
    - API-Net [297];
    - Label Smoothing ("Partial" Robustness) [298];
    - EdgeRob and EdgeGANRob [299];
- Certified Robustness/Robustness Certificates [300]:
    - Survey [301];
    - CROWN [302];
    - CROWN + IBP [303];
    - Interval Bound Propagation [304, 305];
    - Spectral Features [306];
    - Abstract Interpretations [307, 308, 309];
    - Linear Regions [310, 311];
    - CNN-Cert [312];
    - IBP for BNNs [313];
    - Low Rank Certification (Smoothed Adversarial Training) [314];
    - Certified Robustness Point Cloud [315];
    - Certified Robustness for Quantized Networks [316];
    - With Differential Privacy [317];
- Detection and Avoidance: [318, 319, 320, 321, 322, 323, 324, 325, 207, 326, 327, 328, 329];
    - Detection not Robust [330] – addresses [207, 321, 318, 320, 326, 327];
    - Detection based on Mahalanobis Distance [328];
    - Logit-Odds [329];
    - Detection based on Gaussian Noise and "Ease-to-Attack" [331];
    - Fisher Information Detection [332];
    - Detection with Saliency [333];
    - Detection with Interpretation [334];
    - Benford-Fourier Coefficients [335];
    - Detect Object Detection Attacks based on Image Context [336];
    - Embedding Neighborhood Graphs [337];
    - Detecting AutoAttack in Frequency Domain [338];
    - Detection Not Robust [339] – addresses [340, 341, 342, 343];
    - Bayesian Neural Networks [344];
- Transferability [345, 346, 45]:
    - Transferability [42];
    - Transferability of Evasion/Poisening [347];
    - Improving Transferability w/ Transformations/Ensemble [348];
    - Transferability against ResNets [349];
    - [350];
    - Improving with Input Diversity [348];
    - Translation-Invaraint Transfer Attacks [351];
    - [352];
- Attacked Defenses:
    - Attacking CVPR'18 Defenses [273];

    - Attacking ICLR'18 Defenses [353];
    - Individual Defenses [354, 205, 249, 244, 243, 210, 41];
    - Detectors [330];
    - [229, 355];
- Empirical/Theoretical Analysis/Phenomena/Studies:
    - Label Leaking [356];
    - Gradient Masking [201];
    - Gradient Obfuscation [353];
    - Suitability of L_p Norms [357]
    - Upper Risk Bound (Linear, Quadratic Classifiers) [358];
    - Semi-Random Noise [359];
    - Adversarial Subspaces and Intrinsic Dimensionality [350, 323, 324];
    - Robustness and Input Dimensionality [221];
    - Adversarial Directions [360];
    - Adversarial Examples are Inevitable [361];
    - Oracle-Based Adversarial Example Definition [362];
    - Robust Features [363, 364];
        - Training with Robust Features [365];
    - Margin of Cross Entropy Loss [366];
    - Adversarial Examples as Input-Fault Tolerance [367];
    - Perceptual Metric PASS [368];
    - Metrics (Old Label New Rank etc.) [60];
    - Evaluation of Regularization Methods [369];
    - Geometry of Deep Networks [370];
    - Adversarial Directions/Principal Components [371];
    - L_0 Attacks on Piecewise-Linear/ReLU Networks [372];
    - Uncertainty/Robustness of Bayesian Networks [373];
    - Accuracy Reduction of Randomized Smoothing [374];
    - Low Curvature Activations avoid Overfitting [375];
    - Adversarial Training:
        - Sample Complexity [93, 376];
        - Hyperparameters [377];
        - Loss Landscape [378, 379, 226];
        - Norm-Agnostic Robustness [380];
        - Adversarial Training and GANs [381, 382];
        - Adversarial Training and Spectral Normalization [383];
        - Effectiveness of First-Order Attacks in Adversarial Training [384];
        - Adversarial Overfitting [385];
        - Adversarial Training Results in Concise Explanations [386];
        - Bag of Tricks [387, 388];
        - Network Width [389];
        - Fairness [390, 391];
    - Robustness Accuracy Trade-Off: [392, 393, 363, 394, 395]
        - Discussion of Robust Self-Training [396];
        - Poor Lipschitzness or Generalization [397];
    - "Explanations":

- Adversarial Examples as Test Error in Noise [398];
- Linear Explanation [399, 38];
- Boundary Tilting [400];
- Manifold Explanation [400, 212, 120, 401, 295, 399] (also see [402]);

* Genuine Adversarial Accuracy Metric [403];
* Theoretical Analysis of Random Networks [404];
* Hyper-Parameters of SGD and Robustness [405];
* Batch Normalization/Robust Normalization [406, 407];
* Unlabeled Out-of-Distribution Data and Sample Complexity [408];
* Loss/Likelihood Flatness [409];
* Adversarial Training improves Transferability to Other Tasks [410, 411];
* Rotation Invariance and Robustness Trade-Off [412];
* Adversarial and Natural Perturbation Robustness Trade-Off [413];
* Models with Abstain Option [414];
* Robustness of Transformers [415];
* Robustness of Quantized Networks [316];
* Generative Capabilities of Adversarial Training [186];
* Attribution and Interaction of Pixels in Adversarial Examples [416];

– Applications:

* Learning:
  - Adversarially Robust Few-Shot [417]:
  - MAML [418];
  - Adversarial Querying (based on MAML) [419];
  - Adversarial Meta-Learner (ADML) [420];
  - Robust Transfer Learning [421];
  - Semi-Supervised Learning:
  - Self-Robust Training (Adversarial Examples Against Pseudo-Labels) [422];
  - Adversarial Self-Supervised Contrastive Loss (Example-Based, Completeley Unlabeled Pre-Training) [423]
  - Rotation-Based Self-Supervised Auxiliary Loss [424];
  - Adversarial Contrastive Learning [425] (Based on [150]);
  - Robust Co-Training [426];
  - Reinforcement Learning [427, 428, 429];
  - Top-k Multi-Label Learning [430];
* Vision:
  - Interpretability [431, 432, 433, 434];
  - Targeted Image Retrieval [435];
  - Multi-Task Attack [436];
  - Multi-Label Classification [437];
  - Semantic Segmentation [438, 439];
  - Object Detection [43];
  - Generative Models [440, 441];
  - Robot Vision/iCub [442];
  - Visual Question Answering [443];
  - Face Identification [444];
  - Morphing against Face Recognition [445];

- Depth Estimation [446, 447];
- Camouflage for Military Vessels [448];
- Adversarial Meshes [449];
- 3D Adversarial Point Clouds [450, 451, 452];
- Face Recognition [453];
- Street Sign Recognition [454];
- Edge Detection [455];
- Image Ranking [456];
- Action Recognition [457];
- Metric Learning [458];
- Copyright Systems [459];
- Robust Representation Learning [460];
- Robust Open-Set Classification [461];
- Multi-Agent Systems in Autonomous Driving [462];
- Domain Adaptation [463];

* Medical Imaging [464];
  - Mammographic Image Classification [465];
  - COVID-19 [466];
* Graphs:
  - Graph Node Classification [467];
  - Graph Classification (Black- and White-Box) [468];
  - Black-Box [469];
* Security:
  - PDF Malware Detection [470];
  - Ad-Blocking [471];
* Tabular Data:
  - Imperceptible PGD Attack [472];
* Text:
  - Survey [473, 474];
  - Attacks [475, 476];
  - Gradient-Based Interpretable Attack by Projecting onto Nearest-Word Embeddings [477];
  - Targeted Black- and White-Box Attacks (Changing Words in Translation) [478];
  - Gumbel Attacks [479];
  - Attack on AWS Comprehend [480];
  - Attacks on BERT [481];
  - Neural Network for Learning White-Box Attacks [482];
  - Evaluation of Reading Comprehension [483];
  - Adversarial Training for Text Comprehension [484];
  - Grammatically Correct Black-Box Attack on BERT [485];
* Speech:
  - Adversarial Training for Speech Recognition [486];
  - Attacks on Alexa, Echo etc. [487, 488];
  - Basic Transformations and Temporal Consistency As Defense [489];
  - Physical (Over-the-Air) Adversarial Examples [490];
* Security:
  - Evade DeepFake Detectors [491];

        - Avoid DeepFakes [492];
    - Fairness [493];
  - Use Cases:
    - Overfitting Test on ImageNet [494];
    - Regularization by Adversarial Training [495];
    - As Defense Against Side-Channel Attacks [496];
    - Vehicle Camouflage [497];
    - Privacy [498];
  - Generalized Threat Models:
    - Correlated Attacks/Test Data Attack [499, 29];
  - Challenges:
    - Madry Lab[1];
    - Robust Vision Benchmark (Bethge Lab)[2] [500];
    - CAAD[3]
  - Toolboxes:
    - DeepRobust [501];
    - Foolbox [500];
    - AdverTorch [502];
    - IBM Adversarial Rbustness Toolbox [503];
- Universal Adversarial Examples:
  - Attacks:
    - Universal Adversarial Examples [504];
    - Frequency-Tuned Universal Adversarial Examples [408, 505];
    - (Date-Free) Fast Feature Fool [506];
    - (Date-Free) Ask-Acquire-Attack [507];
  - Defenses:
    - Universal Adversarial Training [508, 509];
    - Universal Adversarial Training with Memory [510];
  - Applications:
    - Speech [511];
- Adversarial Patches [512]:
  - Attacks:
    - (Universal) Adversarial Patch for Object Detectors with Random Location [513, 514, 515];
    - Tracking Aversarial Patches using Expectation over Transformation [516];
    - Adversarial Framing [517];
    - LaVAN [518];
    - Attacking Optical Flow [519];
    - (Black/White Patches) Robust Physical Perturbations [454];
    - (Attention-Based) Shadow-Like Perturbations [520];
    - Patches against Reinforcement Learning [521];

[1] http://people.csail.mit.edu/madry/lab/
[2] https://robust.vision/benchmark/leaderboard/
[3] http://hof.geekpwn.org/caad/en/index.html

        * Transparent/Transformed Patches [522];
        * Physical Generalizable Patches [523];
        * Bias-Based Universal Patch [524];
        * Face Patches [525];
        * Texture Patches [526];
        * [527];
    - Defenses:
        * Digital Watermarking [528];
            · Not Robust [529];
        * Interval Bound Propagation [529];
        * Pre-Processing and Detection [530];
        * Local Gradient Smoothing [531];
            · Not Robust [529];
        * Sparse Fourier Transform [532];
            · Not Robust in L_0 [355];
        * Provable Defense [533];
        * Feature Norm Clipping [534];
    - Misc:
        * Evaluation of Physical Adversarial Patches [535];
- Physical Adversarial Examples:
    - Attacks:
        * Robust Adversarial Examples with Expectation over Transformation [536];
        * Adversarial Camera Stickers [537];
        * Adversarial T-Shirt [538];
        * Bias-Based Universal Patch [524];
        * Meta-Learned [539];
    - Datasets/benchmarks:
        * APRICOT [540];
- Structural Perturbations:
    - Defenses:
        * Adversarial Training on Perturbed Dataset [541];
        * Adversarial Training on Adversarial Deformations [542];
        * Approximate On-Manifold Adversarial Training [543];
        * Interval Bound Propagation [544];
    - Attacks:
        * Translation/Rotation [545, 546];
        * Adversarial Deformations [547, 548, 545];
        * Adversarial Projective Transformations (incl. Adversarial Fine-Tuning) [549];
        * Black-Box Hue and Saturation Attack [550];
- "Adversarial Filters":
    - Adversarial Directions / Low-Level Textures [371];
    - Adapting Hue and Saturation [550];

  - Adversarial Stickers (Similar Effect) [537];
  - Functional Adversarial Attacks [551];
  - Adversarial Color Filters [552];
  - Texture-Based Patches [526];
- Interpretability [**BauCVPR2017**, 553, 554, 130]:
  - Stable Self-Explainable Models [555];
  - Adversarial Training w/ Feature Pairing [90, 556];
  - Concept Activation Vectors [557];
  - Does Interpretability Improve Robustness? [558];
  - SmoothGrad [559];
  - L1 Penalty for Attribution Maps [560];
  - FullGrad [561];
- Network Calibration [562, 563]:
  - Post-Hoc Calibration via Auxiliary Class [564];
  - Calibration Monotinicity Loss [565];
  - Mix and Match Calibration [566];
  - Adaptive Label Smoothing [567];
  - Difference of Confidences [568];
- Data Augmentation:
  - Learned Data Augmentation [569, 570, 571, 572, 573, 574, 575];
  - On-Manifold Data Augmentation [576];
- Generalization:
  - Batch Normalization [577]:
    - Alternative Explanation [578];
    - Training without BN [579];
    - Adversarial Training with Separate BN Statistics [495];
    - Data Augmentation with Split BN [580];
  - Initialization:
    - Fixup for ResNets [581];
  - Augmentation:
    - CutOut [582];
    - AutoAugment [574];
    - Adversarial Weights Training [583];
  - Regularization:
    - DisturbLabel [584];
    - Adversarial Dropouts [585];
    - Virtual Adversarial Training [131];
    - Confidence Penalty/Label Smoothing [586];
    - Gradient Noise [587];
    - FlipOut [588];

        - Gradient-Coherent L1/2 Regularization [589];
    - (Empirical) Studies/Analysis:
        - Selectivity Index/Reliance on Single Directions [590];
        - Accuracy of (Partial) Random Networks [591];
        - Iterative Pruning / Lottery Hypothesis [592, 593];
            - Robustness of Pruning [594];
            - Lottery Tickets and Adversarial Training [595, 596, 597];
        - Distance from Initialization [598];
        - PAC-Bounds for Robust Algorithms (Alternative Formulation of Robustness) [394];
        - Bias to Texture and Local Features [599, 600];
        - Variance of Activations Regularization [601];
        - Flat/Sharp Minima [602, 603, 604];
        - Sensitivity [605];
        - Batch Normalization Smoothes [578];
        - Regularization Effect of Dropout [606];
        - Capacity Measures like Sharpness/Bounds/Network Size [607];
    - Normalization:
        - Layer Normalization [608];
        - Instance Normalization [609];
        - Group Normalization [610];
        - Filter Response Normalization [611];
    - Activation Functions:
        - Leaky Hyperbolic Tangent [612];
    - Theory:
        - Fat-Shattering Dimensions of Deep Networks [613];
- Pruning:
    - Survey and ShrinkBench [614];
- Noisy Labels:
    - Generalized Cross Entropy [615];
    - Variance Regularizer [616];
    - Label Smoothing [617];
    - Pre-stopping [618];
    - Meta-learned Loss Function [619];
- Subpopulation Shifts [620] and Natural Distribution Shifts [621];
- Out-of-Distribution [622]:
    - “Distal” Adversarial Examples:
        - Distal Adversarial Examples [241];
        - Unrecognizable Adversarial Examples [623];
        - Black-Box Distal Adversarial Examples [624];
        - Adversarial Out-of-Distribution Examples [625, 626];
    - Detection/Robustness:

        * Likelihood Ratios [627];
        * Outlier Exposure [628];
        * Adversarial Confidence-Enhanced Training (ACET) [241];
        * Bayesian ACET [629];
        * GAN-based Out-of-Distribution Training [630];
        * Out-of-Distribution Training [240];
        * Perturbation-Based Detection [**LiangARXIV2018**];
        * Sine Networks [631];
        * Confidence Densities [632];
        * Logit Inspection [275];
        * "Improved" Distillation [633];
        * Monte Carlo Batch Normalization [634];
        * Local Intrinsic Dimensionality [323];
        * Gaussian Statistics / Mahalanobis Distance [635, 328];
        * Confidence-Based Detection Baseline [636];
        * ODIN [637];
        * Adversarially Robust Auto-Encoder [638];
        * Adversarial Training with Informatice Outlier Mining [639];
        * Detection Score for VAEs [640];
        * VAE Detection with Robust/Vulnerable Latent Features [641];
        * Robust Out-of-Distribution Detection [626];
        * FADER: RBF Networks [642, 643];
        * Separate Confidence Estimation Branch [644];
        * Contrastive Training [645];
        * Certified OOD [646, 647, 648];
    - Applications:
        * Semantic Segmentation [649];
- Anomaly Detection [650, 651];
- Novelty Detection [652];
- Outliert Detection [653, 654, 655];
- Corruption Robustness:
    - Patch Gaussian Augmentation [656];
    - MNIST-C [657];
    - Cifar10-C [658] and ImageNet-C [659];
    - ImageNet-R [660];
    - BN Statistic Calibration [661, 662];
    - Test-Time BN [663];
    - Gaussian Noise Augmentation [664];
    - Adversarial Training [665];
    - Adversarial Training with Meta-Noise Generator [191];
    - Adversarial Data Augmentation [666];
    - AugMix [667];

  - DeepAugment [660];
  - Robust Vision Transformers [668];
  - Frequency Biased Models [669];
- Sponge Examples [670, 671];
- Uncertainty Quantification [672, 673];
- Adversarial Weights:
  - Satisfiability Modulo Theory Solve Approach [674];
  - Fault Tolerance:
    * Review [675];
    * GAN-Based Training [676];
    * Data Augmentation with Node Failures [677];
    * Adversarial Training Single Node Failures [678, 679];
    * Fault Tolerance of Adversarially Robust Models [680];
    * Fault Tolerance for Learned AES [681];
    * Random Weight Dropping [682];
  - Hardware Errors/Attacks:
    * Neural Network Accelerators:
      · Tutorial and Survey [683];
      · Tutorials: http://eyeriss.mit.edu/tutorial-news.html;
      · Energy Estimation: https://energyestimation.mit.edu/ [684];
      · Overview of Analog Hardware for DNN accelerators [685];
    * Attacks:
      · (Physical) Laser Attack [686];
      · (Software) Plundervolt [687];
      · (Software) CLKSREW [688];
      · Bit-Flip Attack [689];
      · Targeted Bit-FLip Attack [690];
    * Error Types:
      · Adversarial Bit Errors:
      · Single Bits in Floating Point [691];
      · DeepHammer / Adversarial Bit Attack [689, 692];
      · Targeted Bit Attack [693];
      · Voltage-Induced and General SRAM Bit Errors:
      · Voltage Boosting [694];
      · Training on Bit Errors [695, 696];
      · Weight Changes through Relevance [697];
      · Weight Nulling with Check Bit [698];
      · Weight Nulling and Hamming-Based Representation [699];
      · Voltage-Induced Computation Errors:
      · Statistical Error Compensation [700];
      · Mixed-Signal Robustness:
      · Gaussian Noise on Pre-Activations [701];
      · Gaussian Noise on Weights [702, 703];
      · Timing Error:

        - Training on Timing Errors [704];
        - Analog (Gaussian Noise):
        - Noise Training + Distillation [705];
    - Analysis [706, 707]:
    - Chips/Approaches:
        - Binary Net Chip [708];
        - MATIC [709];
        - Minerva [710];
        - EDEN [696];
        - Gaussian Noise Training and BN Calibration [702, 703];
    - Misc:
        - Resilient Architecture Search [711];
        - Robust Architecture Search [712];
        - NetAdapt [713];
        - Adversarial Weights Training for Generalization/Semi-Supervised Learning [583];
- Compression [714];
- Quantization:
    - Survey [715];
    - Efficient Architectures [716, 717, 718];
    - Robustness to Quantization:
        - Training with Projected/Quantized Weights [719];
        - Quantization-Aware Training [720];
        - Re-Training with Quantized Weights [721];
        - L_1 Gradient Regularization [722];
        - KURT regularization [723];
    - Fine-Tuning:
        - Incremental Network Quantization [724];
        - Hessian-Weighted k-Means Quantization [725];
        - Adaptive Fixed-Point Re-Training [726];
    - Training:
        - BinaryConnect [727, 728];
        - Quantization plus Distillation [729];
        - XNOR-Net [730];
        - Theoretical Analysis of Training with (Fixed-Point) Quantization [731];
        - Gradual Quantization (and BN Removal) [732];
        - Fixed-Point Quantization [733];
        - BitMixer (training for runtime bit mixing) [734];
        - Learned Step-Size Quantization [735];
        - LSQ with Bin Regularization [736];
        - Cluster-Promoting Quantization [737];
    - After Training:
        - Bayesian Pruning [738];
        - Adaptive Fixed-Point Representation [733];
        - Layer-Wise Floating-Point [739];
    - Precision Selection / Mixed Precision:
        - Generalizable Mixed-Precision Quantization [740];

        * Activation Quantization:
            · PACT [741]
        * Gradient Quantization [742, 743];
        * Adversarial Robustness:
            · Combinatioral Attack on Binarized Networks [65];
        * Misc:
            · Batch Notmalization and Quantization [744];
    - Adversarial Robustness:
        * Certificates Against L_∞, Application to Quantization [745];
        * Improvement for Robustness Against Adversarial Examples [178];
        * Regularization for Joint Input and Weight Robustness [746];
    - Watermarking [747, 748, 749, 750, 751] (Ambiguity Attacks [752], Survey [753]);
    - Backdooring and Trojaning:
        * Survey [754];
        * Early Work [755, 756, 757]
        * Attacks:
            · Data Poisening [758, 759, 760, 761, 762];
            · Data Poisening Benchmark [763];
            · Weight Perturbation [764, 765, 766];
            · Regularizer [767, 768];
            · Programmable Backdoors [769];
            · Reflection Attack [770];
            · Label Consistency Attack [771];
            · Sample-Specific Backdoor Attack [772];
        * Defenses:
            · Detection [773, 774];
            · Fine-Pruning [775];
            · Detection + Re-Training [776];
            · Detection using Adversarial Examples [777];
            · One-Pixel Detection [778];
            · Detection in Frequency Domain [779];
            · Black-box Trigger Reverse Enginering [780];
            · Trimming Training Examples [781];
            · Spectral Signatures [782];
        * Broken Defenses [783];
        * Applications:
            · Federated Learning [784, 785, 159];
            · Point Cloud Classification [786, 787];
        * Transferability [352];
    - Related Work:
        * FlipOut, Weight Perturbations as Regularization [588];
- Ordering Attack [788];
- Data poisening:
    - Defenses:

        - KARMA [789];
        - Back-Gradient Optimization [790];
    - Attacks:
        - LASSO, Ridge Regression, Elastic Net [791];
    - Defenses:
        - Data Provenance [792];
    - Misc:
        - FAIL Model [793];
- Membership Inference [794, 795]:
    - Surveys [796, 797];
    - Defenses:
        - Adversarial Training [798];
        - [799];
        - [800];
    - Attacks:
        - Secret Extraction [801];
        - [802];
        - [803];
        - [804];
        - [805];
    - Studies:
        - Discussion of Privacy Laws [806];
        - Vulnerability Analysis (Difficulty, Output Dimensionality) [807];
        - With Explanations [808];
        - Adversarially Robust Models [805];
        - [809];
        - Memorization [810, 811];
    - Generative Models [812, 813, 814];
    - Text [815, 816, 817];
- DeepFakes and Face Forgery [818, 819, 492];
- Image Forgery [820];
- Model inversion [821];
- Adversarial Initialization [822];
- Adversarial Re-Programming [823];
- Model Stealing/Extraction/Reverse Engineering:
    - Reverse Engineering [824, 825];
    - Stealing Hyper-Parameters [826];
    - Stealing Architecture Through Matrix Multiply Calls [827];
    - Model Extraction [828, 829, 830, 831];
    - Model Extraction of Bert [832];

  - Defenses [833, 834, 835];
  - Data-Free [836, 837];
- Data Removal [838];
- **Misc**:
  - Verification [839, 840, 841, 842];
  - Security as Science [843];
  - Implications in Law [844];
  - Neural Cryptography [845];
  - Adversarial Reprogramming [846];
  - Redundant Features [847];

2. Vision Transformers

- Original Transformer [848];
- Vision Transformer [849];
- Masked Auto-Encoder/Transformer [850];
  - Data Augmentation and Regularization [851];
  - With SAM [852];
  - Data-Efficient Distillation [853];
- Robustness:
  - Robust Vision Transfoerm (RVT) with Patch-Wise Augmentation [668] (Corruption + FGSM);
  - PGD and AutoAttack Robustness [415];
  - Corruptions, Transformations, FGSM, PGD [854];
  - Corruptions, Natural Adversarial Examples [855];
  - Adversarial Examples and Adversarial Training (ImageNet) [856];
  - Natural and Adversarial Patches [857];
  - Adversarial Tokens (Patches) [858];
  - Corruption Robustnes [859];

References

[1] Kush R. Varshney. "Trustworthy machine learning and artificial intelligence". In: *XRDS* 25.3 (2019), pp. 26–29.

[2] Michael Kearns and Aaron Roth. *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press, 2019.

[3] Xiaowei Huang et al. "A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability". In: *Computer Science Review* 37 (2020).

[4] Ram Shankar Siva Kumar et al. "Adversarial Machine Learning-Industry Perspectives". In: *Proc. of the IEEE Symposium on Security and Privacy Workshops*. 2020.

[5] Irene Amerini et al. "Localization of JPEG Double Compression Through Multi-domain Convolutional Neural Networks". In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2017.

[6] Luca Bondi et al. "Tampering Detection and Localization Through Clustering of Camera-Based CNN Features". In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2017.

[7] Luca Bondi et al. “First Steps Toward Camera Model Identification With Convolutional Neural Networks”. In: *IEEE Signal Processing Letters* 24.3 (2017), pp. 259–263.

[8] Chengjiang Long et al. “A C3D-Based Convolutional Neural Network for Frame Dropping Detection in a Single Video Shot”. In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2017.

[9] Seong Joon Oh et al. “Person Recognition in Personal Photo Collections”. In: *Proc. of the IEEE International Conference on Computer Vision (ICCV)*. 2015.

[10] Seong Joon Oh et al. “Faceless Person Recognition: Privacy Implications in Social Media”. In: *Proc. of the European Conference on Computer Vision (ECCV)*. 2016.

[11] Seong Joon Oh, Mario Fritz, and Bernt Schiele. “Adversarial Image Perturbation for Privacy Protection - A Game Theory Perspective”. In: *arXiv.org* abs/1703.09471 (2017).

[12] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. “Towards a Visual Privacy Advisor: Understanding and Predicting Privacy Risks in Images”. In: *arXiv.org* abs/1703.10660 (2017).

[13] Nisarg Raval, Ashwin Machanavajjhala, and Landon P. Cox. “Protecting Visual Secrets Using Adversarial Nets”. In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2017.

[14] Battista Biggio et al. “Evasion Attacks against Machine Learning at Test Time”. In: *Proc. of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*. 2013.

[15] Christian Szegedy et al. “Intriguing properties of neural networks”. In: *Proc. of the International Conference on Learning Representations (ICLR)*. 2014.

[16] Daniel Lowd and Christopher Meek. “Adversarial learning”. In: *Proc. of the ACM International Conference on Knowledge Discovery & Data Mining*. 2005.

[17] Nilesh N. Dalvi et al. “Adversarial classification”. In: *Proc. of the ACM International Conference on Knowledge Discovery & Data Mining*. 2004.

[18] Amir Globerson and Sam T. Roweis. “Nightmare at test time: robust learning by feature deletion”. In: *Proc. of the International Conference on Machine Learning (ICML)*. 2006.

[19] Choon Hui Teo et al. “Convex Learning with Invariances”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2007.

[20] Ian J. Goodfellow et al. “Measuring Invariances in Deep Networks”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2009.

[21] Samuel Henrique Silva and Peyman Najafirad. “Opportunities and Challenges in Deep Learning Adversarial Robustness: A Survey”. In: *arXiv.org* abs/2007.00753 (2020).

[22] Marco Barreno et al. “Can machine learning be secure?” In: *Proc. of the ACM on Asia Conference on Computer and Communications Security (AsiaCCS)*. 2006.

[23] Xiaoyong Yuan et al. “Adversarial Examples: Attacks and Defenses for Deep Learning”. In: *arXiv.org* abs/1712.07107 (2017).

[24] Naveed Akhtar and Ajmal Mian. “Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey”. In: *arXiv.org* abs/1801.00553 (2018).

[25] Battista Biggio and Fabio Roli. “Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning”. In: *arXiv.org* abs/1712.03141 (2018).

[26] Han Xu et al. “Adversarial Attacks and Defenses in Images, Graphs and Text: A Review”. In: *arXiv.org* abs/1909.08072 (2019).

[27] Ashutosh Chaubey et al. “Universal Adversarial Perturbations: A Survey”. In: *arXiv.org* abs/2005.08087 (2020).

[28] Nicholas Carlini et al. “On Evaluating Adversarial Robustness”. In: *arXiv.org* abs/1902.06705 (2019).

[29] Justin Gilmer et al. “Motivating the Rules of the Game for Adversarial Example Research”. In: *arXiv.org* abs/1807.06732 (2018).

[30] Jiashi Feng et al. “Robust Logistic Regression and Classification”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2014.

[31] Yan Zhou et al. “Adversarial support vector machine learning”. In: *Proc. of the ACM International Conference on Knowledge Discovery & Data Mining*. 2012.

[32] Maksym Andriushchenko and Matthias Hein. “Provably robust boosted decision stumps and trees against adversarial attacks”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2019.

[33] Yihan Wang et al. “On ℓ_p-norm Robustness of Ensemble Stumps and Trees”. In: *arXiv.org* abs/2010.07344 (2020).

[34] Chawin Sitawarin and David A. Wagner. “Minimum-Norm Adversarial Examples on KNN and KNN-Based Models”. In: *arXiv.org* abs/2003.06559 (2020).

[35] Chawin Sitawarin and David A. Wagner. “On the Robustness of Deep K-Nearest Neighbors”. In: *Proc. of the IEEE Symposium on Security and Privacy*. 2019.

[36] Yao-Yuan Yang et al. “Robustness for Non-Parametric Classification: A Generic Attack and Defense”. In: *Conference on Artificial Intelligence and Statistics (AISTATS)*. Ed. by Silvia Chiappa and Roberto Calandra. 2020.

[37] Christian Szegedy et al. “Intriguing properties of neural networks”. In: *arXiv.org* abs/1312.6199 (2013).

[38] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. “Explaining and harnessing adversarial examples”. In: *arXiv.org* abs/1412.6572 (2014).

[39] Aleksander Madry et al. “Towards deep learning models resistant to adversarial attacks”. In: *arXiv.org* abs/1706.06083 (2017).

[40] Nicholas Carlini and David Wagner. “Towards evaluating the robustness of neural networks”. In: *Proc. of the IEEE Symposium on Security and Privacy*. 2017.

[41] Yujia Liu et al. “Enhanced Attacks on Defensively Distilled Deep Neural Networks”. In: *arXiv.org* abs/1711.05934 (2017).

[42] Yanpei Liu et al. “Delving into transferable adversarial examples and black-box attacks”. In: *arXiv.org* abs/1611.02770 (2016).

[43] Jiajun Lu et al. “No need to worry about adversarial examples in object detection in autonomous vehicles”. In: *arXiv.org* abs/1707.03501 (2017).

[44] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. “Adversarial examples in the physical world”. In: *arXiv.org* abs/1607.02533 (2016).

[45] Nicolas Papernot et al. “The Limitations of Deep Learning in Adversarial Settings”. In: *Proc. of the IEEE Symposium on Security and Privacy*. 2016.

[46] Andras Rozsa, Manuel Günther, and Terrance E. Boult. “Adversarial Robustness: Softmax versus Openmax”. In: *arXiv.org* abs/1708.01697 (2017).

[47] Yinpeng Dong et al. “Boosting Adversarial Attacks with Momentum”. In: *arXiv.org* abs/1710.06081 (2017).

[48] Bo Luo et al. “Towards Imperceptible and Robust Adversarial Example Attacks against Neural Networks”. In: *arXiv.org* abs/1801.04693 (2018).

[49] Yang Song et al. “Generative Adversarial Examples”. In: *arXiv.org* abs/1805.07894 (2018).

[50] Tom B. Brown et al. “Unrestricted Adversarial Examples”. In: *arXiv.org* abs/1809.08352 (2018).

[51] Zhengli Zhao, Dheeru Dua, and Sameer Singh. “Generating Natural Adversarial Examples”. In: *arXiv.org* abs/1710.11342 (2017).

[52] Lukas Schott et al. “Robust Perception through Analysis by Synthesis”. In: *arXiv.org* abs/1805.09190 (2018).

[53] Lukas Schott et al. “Towards the first adversarially robust neural network model on MNIST”. In: *Proc. of the International Conference on Learning Representations (ICLR)*. 2019.

[54] Kathrin Grosse et al. “The Limitations of Model Uncertainty in Adversarial Settings”. In: *arXiv.org* abs/1812.02606 (2018).

[55] Warren He, Bo Li, and Dawn Song. “Decision Boundary Analysis of Adversarial Examples”. In: *Proc. of the International Conference on Learning Representations (ICLR)*. 2018.

[56] Sayantan Sarkar et al. “UPSET and ANGRI : Breaking High Performance Image Classifiers”. In: *arXiv.org* abs/1707.01159 (2017).

[57] Huan Zhang et al. “The Limitations of Adversarial Training and the Blind-Spot Attack”. In: *Proc. of the International Conference on Learning Representations (ICLR)*. 2019.

[58] Hyun Kwon, Hyunsoo Yoon, and Daeseon Choi. “Restricted Evasion Attack: Generation of Restricted-Area Adversarial Example”. In: *IEEE Access* 7 (2019), pp. 60908–60919.

[59] Ping-Yeh Chiang et al. “WITCHcraft: Efficient PGD attacks with random step size”. In: *arXiv.org* abs/1911.07989 (2019).

[60] Aditya Ganeshan, Vivek B. S., and R. Venkatesh Babu. “FDA: Feature Disruptive Attack”. In: *arXiv.org* abs/1909.04385 (2019).

[61] Jan Philip Göpfert, Heiko Wersing, and Barbara Hammer. “Adversarial attacks hidden in plain sight”. In: *arXiv.org* abs/1902.09286 (2019).

[62] Eric Wong, Frank R. Schmidt, and J. Zico Kolter. “Wasserstein Adversarial Examples via Projected Sinkhorn Iterations”. In: *Proc. of the International Conference on Machine Learning (ICML)*. 2019.

[63] Pin-Yu Chen et al. “EAD: Elastic-Net Attacks to Deep Neural Networks via Adversarial Examples”. In: *Proc. of the Conference on Artificial Intelligence (AAAI)*. 2018.

[64] Kaidi Xu et al. “Structured Adversarial Attack: Towards General Implementation and Better Interpretability”. In: *Proc. of the International Conference on Learning Representations (ICLR)*. 2019.

[65] Elias B. Khalil, Amrita Gupta, and Bistra Dilkina. “Combinatorial Attacks on Binarized Neural Networks”. In: *Proc. of the International Conference on Learning Representations (ICLR)*. 2019.

[66] Liang Tong et al. “Towards Robustness against Unsuspicious Adversarial Examples”. In: *arXiv.org* abs/2005.04272 (2020).

[67] Nurislam Tursynbek, Aleksandr Petiushko, and Ivan Oseledets. “Geometry-Inspired Top-k Adversarial Perturbations”. In: *arXiv.org* abs/2006.15669 (2020).

[68] Francesco Croce and Matthias Hein. “Mind the box: l_1-APGD for sparse adversarial attacks on image classifiers”. In: *arXiv.org* abs/2103.01208 (2021).

[69] Anonymous. “Curriculum Transfer Adversarial Training: Improving Adversarial Robustness without Accuracy Tradeoff”. In: *Submitted to ICCV*. 2021.

[70] Jérôme Rony et al. “Augmented Lagrangian Adversarial Attacks”. In: *Proc. of the IEEE International Conference on Computer Vision (ICCV)*. 2021.

[71] Ranjie Duan et al. “AdvDrop: Adversarial Attack to DNNs by Dropping Information”. In: *Proc. of the IEEE International Conference on Computer Vision (ICCV)*. 2021.

[72] Sven Gowal et al. “An Alternative Surrogate Loss for PGD-based Adversarial Testing”. In: *arXiv.org* abs/1910.09338 (2019).

[73] Ian Goodfellow, Yao Qin, and David Berthelot. *Evaluation Methodology for Attacks Against Confidence Thresholding Models*. https://openreview.net/forum?id=H1g0piA9tQ. 2019.

[74] Francesco Croce et al. “Sparse-RS: a versatile framework for query-efficient sparse black-box adversarial attacks”. In: *arXiv.org* abs/2006.12834 (2020).

[75] Jonathan Uesato et al. “Adversarial Risk and the Dangers of Evaluating Against Weak Attacks”. In: *Proc. of the International Conference on Machine Learning (ICML)*. 2018.

[76] Tianhang Zheng, Changyou Chen, and Kui Ren. “Distributionally Adversarial Attack”. In: *Proc. of the Conference on Artificial Intelligence (AAAI)*. 2019.

[77] Wieland Brendel et al. “Accurate, reliable and fast robustness evaluation”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2019.

[78] Jérôme Rony et al. “Decoupling Direction and Norm for Efficient Gradient-Based L2 Adversarial Attacks and Defenses”. In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.

[79] Maura Pintor et al. “Fast Minimum-norm Adversarial Attacks through Adaptive Norm Constraints”. In: *arXiv.org* abs/2102.12827 (2021).

[80] Apostolos Modas, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. “SparseFool: A Few Pixels Make a Big Difference”. In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.

[81] Florian Jaeckle and M. Pawan Kumar. “Generating adversarial examples with graph neural networks”. In: *Proc. of the Conference on Uncertainty in Artificial Intelligence (UAI)*. 2021.

[82] Florian Jaeckle et al. “Attention for Adversarial Attacks: Learning from your Mistakes”. In: *Proc. of the Conference on Artificial Intelligence (AAAI) Workshops*. 2022.

[83] Pin-Yu Chen et al. “ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models”. In: *Proc. of the ACM Workshop on Artificial Intelligence and Security*. 2017.

[84] Andrew Ilyas et al. “Black-box Adversarial Attacks with Limited Queries and Information”. In: *Proc. of the International Conference on Machine Learning (ICML)*. 2018.

[85] Arjun Nitin Bhagoji et al. “Exploring the Space of Black-box Attacks on Deep Neural Networks”. In: *arXiv.org* abs/1712.09491 (2017).

[86] Wieland Brendel and Matthias Bethge. “Comment on ”Biologically inspired protection of deep networks from adversarial attacks””. In: *arXiv.org* abs/1704.01547 (2017).

[87] Thomas Brunner et al. “Guessing Smart: Biased Sampling for Efficient Black-Box Adversarial Attacks”. In: *Proc. of the IEEE International Conference on Computer Vision (ICCV)*. 2019.

[88] Jianbo Chen and Michael I. Jordan. “Boundary Attack++: Query-Efficient Decision-Based Adversarial Attack”. In: *arXiv.org* abs/1904.02144 (2019).

[89] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. “One pixel attack for fooling deep neural networks”. In: *arXiv.org* abs/1710.08864 (2017).

[90] Yinpeng Dong et al. “Towards Interpretable Deep Neural Networks by Leveraging Adversarial Examples”. In: *arXiv.org* abs/1708.05493 (2017).

[91] Anonymous. “Boosting Transferability of Adversarial Examples with Randomized Skip Connections”. In: *Review for Advances in Neural Information Processing Systems (NeurIPS)*. 2018.

[92] Andrew Ilyas, Logan Engstrom, and Aleksander Madry. “Prior Convictions: Black-Box Adversarial Attacks with Bandits and Priors”. In: *arXiv.org* abs/1807.07978 (2018).

[93] Marc Khoury and Dylan Hadfield-Menell. “On the Geometry of Adversarial Examples”. In: *arXiv.org* abs/1811.00525 (2018).

[94] Anonymous. “Enhancing Adversarial Example Transferability with an Intermediate Level Attack”. In: *Review for Proc. of the IEEE International Conference on Computer Vision (ICCV)*. 2019.

[95] Alberto Marchisio et al. “CapsAttacks: Robust and Imperceptible dversarial Attacks on Capsule Networks”. In: *Proc. of the International Conference on Machine Learning (ICML) Workshops* (2019).

[96] Maksym Andriushchenko. “Provable Adversarial Defenses for Boosting”. MA thesis. Saarland University, 2019.

[97] Maksym Andriushchenko et al. “Square Attack: a query-efficient black-box adversarial attack via random search”. In: *arXiv.org* abs/1912.00049 (2019).

[98] Chuan Guo, Jared S. Frank, and Kilian Q. Weinberger. “Low Frequency Adversarial Perturbation”. In: *arXiv.org* abs/1809.08758 (2018).

[99] Chuan Guo et al. “Simple Black-box Adversarial Attacks”. In: *Proc. of the International Conference on Machine Learning (ICML)*. 2019.

[100] Francesco Croce and Matthias Hein. “Sparse and Imperceivable Adversarial Attacks”. In: *arXiv.org* abs/1909.05040 (2019).

[101] Thomas Brunner, Frederik Diehl, and Alois Knoll. “Copy and Paste: A Simple But Effective Initialization Method for Black-Box Adversarial Attacks”. In: *arXiv.org* abs/1906.06086 (2019).

[102] Ali Rahmati et al. “GeoDA: a geometric framework for black-box adversarial attacks”. In: *arXiv.org* abs/2003.06468 (2020).

[103] Weilun Chen et al. “Boosting Decision-based Black-box Adversarial Attacks with Random Sign Flip”. In: *Proc. of the European Conference on Computer Vision (ECCV)*. 2020.

[104] Abdullah Al-Dujaili and Una-May O’Reilly. “There are No Bit Parts for Sign Bits in Black-Box Attacks”. In: *arXiv.org* abs/1902.06894 (2019).

[105] Abdullah Al-Dujaili and Una-May O’Reilly. “Sign Bits Are All You Need for Black-Box Attacks”. In: *Proc. of the International Conference on Learning Representations (ICLR)*. 2020.

[106] Zheng Yuan et al. “Meta Gradient Adversarial Attack”. In: *Proc. of the IEEE International Conference on Computer Vision (ICCV)*. 2021.

[107] Muzammal Naseer et al. “On Generating Transferable Targeted Perturbations”. In: *Proc. of the IEEE International Conference on Computer Vision (ICCV)*. 2021.

[108] Xiaosen Wang et al. “Admix: Enhancing the Transferability of Adversarial Attacks”. In: *Proc. of the IEEE International Conference on Computer Vision (ICCV)*. 2021.

[109] Jie Li et al. “Aha! Adaptive History-Driven Attack for Decision-Based Black-Box Models”. In: *Proc. of the IEEE International Conference on Computer Vision (ICCV)*. 2021.

[110] Jie Li et al. “Projection & Probability-Driven Black-Box Attack”. In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020.

[111] Jianbo Chen, Michael I. Jordan, and Martin J. Wainwright. “HopSkipJumpAttack: A Query-Efficient Decision-Based Attack”. In: *Proc. of the IEEE Symposium on Security and Privacy*. 2020.

[112] Thibault Maho, Teddy Furon, and Erwan Le Merrer. “SurFree: A Fast Surrogate-Free Black-Box Attack”. In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021.

[113] Yucheng Shi, Yahong Han, and Qi Tian. “Polishing Decision-Based Adversarial Noise With a Customized Sampling”. In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020.

[114] Pu Zhao et al. “Towards Query-Efficient Black-Box Adversary with Zeroth-Order Natural Gradient Descent”. In: *Proc. of the Conference on Artificial Intelligence (AAAI)*. 2020.

[115] Tao Xiang et al. “Local Black-box Adversarial Attacks: A Query Efficient Approach”. In: *arXiv.org* abs/2101.01032 (2021).

[116] Maksym Yatsura, Jan Hendrik Metzen, and Matthias Hein. “Meta-Learning the Search Distribution of Black-Box Random Search Based Adversarial Attacks”. In: *arXiv.org* abs/2111.01714 (2021).

[117] Minhao Cheng et al. “Sign-OPT: A Query-Efficient Hard-label Adversarial Attack”. In: *Proc. of the International Conference on Learning Representations (ICLR)*. 2020.

[118] Jinghui Chen and Quanquan Gu. “RayS: A Ray Searching Method for Hard-label Adversarial Attack”. In: *Proc. of the ACM International Conference on Knowledge Discovery & Data Mining*. 2020.

[119] Satya Narayan Shukla et al. “Simple and Efficient Hard Label Black-box Adversarial Attacks in Low Query Budget Regimes”. In: *Proc. of the ACM International Conference on Knowledge Discovery & Data Mining*. 2021.

[120] Justin Gilmer et al. “Adversarial Spheres”. In: *arXiv.org* abs/1801.02774 (2018).

[121] Hsueh-Ti Derek Liu et al. “Beyond Pixel Norm-Balls: Parametric Adversaries using an Analytically Differentiable Renderer”. In: *Proc. of the International Conference on Learning Representations (ICLR)*. 2019.

[122] Taejoon Byun et al. “Manifold-Based Test Generation for Image Classifiers”. In: *arXiv.org* abs/2002.06337 (2020).

[123] Kanil Patel et al. “On-manifold Adversarial Data Augmentation Improves Uncertainty Calibration”. In: *arXiv.org* abs/1912.07458 (2019).

[124] Yizhen Wang, Somesh Jha, and Kamalika Chaudhuri. “Analyzing the Robustness of Nearest Neighbors to Adversarial Examples”. In: *Proc. of the International Conference on Machine Learning (ICML)*. 2018.

[125] Paolo Russu et al. “Secure Kernel Machines against Evasion Attacks”. In: *Proc. of the ACM on Asia Conference on Computer and Communications Security (AsiaCCS)*. 2016, pp. 59–69.

[126] Rakshit Agrawal, Luca de Alfaro, and David P. Helmbold. “A New Family of Neural Networks Provably Resistant to Adversarial Attacks”. In: *arXiv.org* abs/1902.01208 (2019).

[127] Luca de Alfaro. “Neural Networks with Structural Resistance to Adversarial Attacks”. In: *arXiv.org* abs/1809.09262 (2018).

[128] Pourya Habib Zadeh, Reshad Hosseini, and Suvrit Sra. “Deep-RBF Networks Revisited: Robust Classification with Rejection”. In: *arXiv.org* abs/1812.03190 (2018).

[129] Petra Vidnerová and Roman Neruda. “Deep Networks with RBF Layers to Prevent Adversarial Examples”. In: *Artificial Intelligence and Soft Computing*. 2018.

[130] Nicolas Papernot and Patrick D. McDaniel. “Deep k-Nearest Neighbors: Towards Confident, Interpretable and Robust Deep Learning”. In: *arXiv.org* abs/1803.04765 (2018).

[131] Takeru Miyato et al. “Distributional smoothing with virtual adversarial training”. In: *arXiv.org* abs/1507.00677 (2015).

[132] Ruitong Huang et al. “Learning with a strong adversary”. In: *arXiv.org* abs/1511.03034 (2015).

[133] Uri Shaham, Yutaro Yamada, and Sahand Negahban. “Understanding adversarial training: Increasing local stability of neural nets through robust optimization”. In: *arXiv.org* abs/1511.05432 (2015).

[134] Aman Sinha, Hongseok Namkoong, and John C. Duchi. “Certifiable Distributional Robustness with Principled Adversarial Training”. In: *arXiv.org* abs/1710.10571 (2017).

[135] Hyeungill Lee, Sungyeob Han, and Jungwoo Lee. “Generative Adversarial Trainer: Defense to Adversarial Perturbations with GAN”. In: *arXiv.org* abs/1705.03387 (2017).

[136] Shufei Zhang et al. “Adversarial Manifold Learning”. In: *arXiv.org* abs/1807.05832v1 (2018).

[137] J. Zico Kolter and Eric Wong. “Provable defenses against adversarial examples via the convex outer adversarial polytope”. In: *arXiv.org* abs/1711.00851 (2017).

[138] Xiao Zhang and David Evans. “Cost-Sensitive Robustness against Adversarial Examples”. In: *arXiv.org* abs/1810.09225 (2018).

[139] Qi-Zhi Cai, Chang Liu, and Dawn Song. “Curriculum Adversarial Training”. In: *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI)*. 2018, pp. 3740–3747.

[140] Anonymous. “Regularizer to Mitigate Gradient masking Effect during Single-Step Adversarial Training”. In: *Review for Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.

[141] Xi Wu et al. “Reinforcing Adversarial Robustness using Model Confidence Induced by Adversarial Training”. In: *arXiv.org* abs/1711.08001 (2017).

[142] Alex Lamb et al. “Interpolated Adversarial Training: Achieving Robust Neural Networks without Sacrificing Too Much Accuracy”. In: *arXiv.org* abs/1906.06784 (2019).

[143] Hadi Salman et al. “Provably Robust Deep Learning via Adversarially Trained Smoothed Classifiers”. In: *arXiv.org* abs/1906.04584 (2019).

[144] Anonymous. “Feature-Activated Adversarial Training”. In: *Review for Proc. of the Conference on Artificial Intelligence (AAAI)*. 2019.

[145] Anonymous. “Sliced Wasserstein Adversarial Training for Defending Against Adversarial Attacks”. In: *Review for Proc. of the Conference on Artificial Intelligence (AAAI)*. 2019.

[146] Hang Yu et al. “Towards Noise-Robust Neural Networks via Progressive Adversarial Training”. In: *arXiv.org* abs/1909.04839 (2019).

[147] Nanyang Ye and Zhanxing Zhu. “Bayesian Adversarial Learning”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2018.

[148] Xuanqing Liu et al. “Adv-BNN: Improved Adversarial Defense through Robust Bayesian Neural Network”. In: *Proc. of the International Conference on Learning Representations (ICLR)*. 2019.

[149] Yair Carmon et al. “Unlabeled Data Improves Adversarial Robustness”. In: *arXiv.org* abs/1905.13736 (2019).

[150] Jonathan Uesato et al. “Are Labels Required for Improving Adversarial Robustness?” In: *arXiv.org* abs/1905.13725 (2019).

[151] Cassidy Laidlaw and Soheil Feizi. “Playing it Safe: Adversarial Robustness with an Abstain Option”. In: *arXiv.org* abs/1911.11253 (2019).

[152] Yogesh Balaji, Tom Goldstein, and Judy Hoffman. “Instance adaptive adversarial training: Improved accuracy tradeoffs in neural nets”. In: *arXiv.org* abs/1910.08051 (2019).

[153] Chang Xiao and Changxi Zheng. “One Man’s Trash is Another Man’s Treasure: Resisting Adversarial Examples by Adversarial Examples”. In: *arXiv.org* abs/1911.11219 (2019).

[154] Gavin Weiguang Ding et al. “Max-Margin Adversarial (MMA) Training: Direct Input Space Margin Maximization through Adversarial Training”. In: *arXiv.org* abs/1812.02637 (2018).

[155] Haichao Zhang and Jianyu Wang. “Defense Against Adversarial Attacks Using Feature Scattering-based Adversarial Training”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2019.

[156] Jianyu Wang. “Bilateral Adversarial Training: Towards Fast Training of More Robust Models Against Adversarial Attacks”. In: *arXiv.org* abs/1811.10716 (2018).

[157] Taesik Na, Jong Hwan Ko, and Saibal Mukhopadhyay. “Cascade Adversarial Machine Learning Regularized with a Unified Embedding”. In: *Proc. of the International Conference on Learning Representations (ICLR)*. 2018.

[158] Jungeum Kim and Xiao Wang. *Sensible adversarial learning*. https://openreview.net/forum?id=rJlf_RVKwr. 2020.

[159] Haotao Wang et al. “Once-for-All Adversarial Training: In-Situ Tradeoff between Robustness and Accuracy for Free”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2020.

[160] Tianjin Huang et al. “calibrated adversarial training”. In: *Proc. of the Asian Conference on Machine Learning (ACML)*. 2021.

[161] Minhao Cheng et al. “CAT: Customized Adversarial Training for Improved Robustness”. In: *arXiv.org* abs/2002.06789 (2020).

[162] Tianyu Pang et al. “Boosting Adversarial Training with Hypersphere Embedding”. In: *arXiv.org* abs/2002.08619 (2020).

[163] Anonymous. “Stylized Adversarial Defense”. In: *Review for Proc. of the European Conference on Computer Vision (ECCV)*. 2020.

[164] Anonymous. “Towards Generalized Robustness of DNNs via Progressive Diversified Augmentation”. In: *Review for Proc. of the European Conference on Computer Vision (ECCV)*. 2020.

[165] Ali Shafahi et al. “Adversarial training for free!” In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2019.

[166] Eric Wong, Leslie Rice, and J. Zico Kolter. “Fast is better than free: Revisiting adversarial training”. In: *arXiv.org* abs/2001.03994 (2020).

[167] Haidong Xie et al. “Blind Adversarial Training: Balance Accuracy and Robustness”. In: *arXiv.org* abs/2004.05914 (2020).

[168] Xunguang Wang, Ship Peng Xu, and Eric Ke Wang. “Initializing Perturbations in Multiple Directions for Fast Adversarial Training”. In: *arXiv.org* abs/2005.07606 (2020).

[169] S. Vivek B. and R. Venkatesh Babu. “Single-step Adversarial training with Dropout Scheduling”. In: *arXiv.org* abs/2004.08628 (2020).

[170] Cihang Xie et al. “Smooth Adversarial Training”. In: *arXiv.org* (2020).

[171] Sahil Singla and Soheil Feizi. “Second-Order Provable Defenses against Adversarial Attacks”. In: *arXiv.org* abs/2006.00731 (2020).

[172] Jingfeng Zhang et al. “Attacks Which Do Not Kill Training Make Adversarial Learning Stronger”. In: *arXiv.org* abs/2002.11242 (2020).

[173] Weitao Wag, Jiansheng Chen, and Ming-Hsuan Yang. “Adversarial Training with Bi-directional Likelihood Regularization for Visual Classification”. In: *Proc. of the European Conference on Computer Vision (ECCV)*. 2020.

[174] Yuanhao Xiong and C. Hsieh. “Improved Adversarial Training via Learned Optimizer”. In: *Proc. of the European Conference on Computer Vision (ECCV)*. 2020.

[175] A. Bui et al. “Improving Adversarial Robustness by Enforcing Local and Global Compactness”. In: *Proc. of the European Conference on Computer Vision (ECCV)*. 2020.

[176] Elahe Arani, Fahad Sarfraz, and Bahram Zonooz. “Adversarial Concurrent Training: Optimizing Robustness and Accuracy Trade-off of Deep Neural Networks”. In: *arXiv.org* abs/2008.07015 (2020).

[177] Jingfeng Zhang et al. “Geometry-aware Instance-reweighted Adversarial Training”. In: *arXiv.org* abs/2010.01736 (2020).

[178] Dongxian Wu, Shutao Xia, and Yisen Wang. “Adversarial Weight Perturbation Helps Robust Generalization”. In: *arXiv.org* abs/2004.05884 (2020).

[179] Adnan Siraj Rakin, Zhezhi He, and Deliang Fan. “Parametric Noise Injection: Trainable Randomness to Improve Deep Neural Network Robustness Against Adversarial Attack”. In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2019).

[180] Jiequan Cui et al. “Learnable Boundary Guided Adversarial Training”. In: *arXiv.org* abs/2011.11164 (2020).

[181] Hongyang Zhang et al. “Theoretically Principled Trade-off between Robustness and Accuracy”. In: *Proc. of the International Conference on Machine Learning (ICML)*. 2019.

[182] Ahmadreza Jeddi, Mohammad Javad Shafiee, and Alexander Wong. “A Simple Fine-tuning Is All You Need: Towards Robust Deep Learning Via Adversarial Fine-tuning”. In: *arXiv.org* abs/2012.13628 (2020).

[183] Anonymous. “Improving Robustness by Penalizing Non-Robust Features”. In: *Submitted to ICCV*. 2021.

[184] Geon Yeong Park and Sang Wan Lee. “Reliably fast adversarial training via latent adversarial perturbation”. In: *Proc. of the IEEE International Conference on Computer Vision (ICCV)*. 2021.

[185] Yao Li et al. “Towards Robustness of Deep Neural Networks via Regularization”. In: *Proc. of the IEEE International Conference on Computer Vision (ICCV)*. 2021.

[186] Yao Zhu et al. “Towards Understanding the Generative Capability of Adversarially Robust Classifiers”. In: *Proc. of the IEEE International Conference on Computer Vision (ICCV)*. 2021.

[187] Omid Poursaeed et al. “Robustness and Generalization via Generative Adversarial Training”. In: *Proc. of the IEEE International Conference on Computer Vision (ICCV)*. 2021.

[188] Bojia Zi et al. “Revisiting Adversarial Robustness Distillation: Robust Soft Labels Make Student Better”. In: *Proc. of the IEEE International Conference on Computer Vision (ICCV)*. 2021.

[189] Florian Tramèr and Dan Boneh. “Adversarial Training and Robustness for Multiple Perturbations”. In: *arXiv.org* abs/1904.13000 (2019).

[190] Pratyush Maini, Eric Wong, and J. Zico Kolter. “Adversarial Robustness Against the Union of Multiple Perturbation Models”. In: *arXiv.org* abs/1909.04068 (2019).

[191] Divyam Madaan, Jinwoo Shin, and Sung Ju Hwang. “Learning to Generate Noise for Robustness against Multiple Perturbations”. In: *arXiv.org* abs/2006.12135 (2020).

[192] Ameya D. Patil et al. “Robustifying ℓ_∞ Adversarial Training to the Union of Perturbation Models”. In: *arXiv.org* abs/2105.14710 (2021).

[193] Francesco Croce and Matthias Hein. “Adversarial robustness against multiple l_p-threat models at the price of one and how to quickly fine-tune robust models to another threat model”. In: *arXiv.org* abs/2105.12508 (2021).

[194] Yan Zhou, Murat Kantarcioglu, and Bowei Xi. “Breaking Transferability of Adversarial Samples with Randomness”. In: *arXiv.org* abs/1805.04613 (2018).

[195] Xuanqing Liu et al. “Towards Robust Neural Networks via Random Self-ensemble”. In: *arXiv.org* abs/1712.00673 (2017).

[196] Thilo Strauss et al. “Ensemble methods as a defense to adversarial perturbations against deep neural networks”. In: *arXiv.org* abs/1709.03423 (2017).

[197] Tom Zahavy et al. “Ensemble Robustness and Generalization of Stochastic Deep Learning Algorithms”. In: *arXiv.org* abs/1602.02389 (2016).

[198] Warren He et al. “Adversarial Example Defenses: Ensembles of Weak Defenses are not Strong”. In: *arXiv.org* abs/1706.04701 (2017).

[199] Sanchari Sen, Balaraman Ravindran, and A. Raghunathan. “EMPIR: Ensembles of Mixed Precision Deep Networks for Increased Robustness against Adversarial Attacks”. In: *arXiv.org* abs/2004.10162 (2020).

[200] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. “Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles”. In: *arXiv.org* abs/1612.01474 (2016).

[201] Florian Tramèr et al. “Ensemble Adversarial Training: Attacks and Defenses”. In: *arXiv.org* abs/1705.07204 (2017).

[202] Edward Grefenstette et al. “Strength in Numbers: Trading-off Robustness and Computation via Adversarially-Trained Ensembles”. In: *arXiv.org* abs/1811.09300 (2018).

[203] J. Hwang et al. “Adversarial Training with Stochastic Weight Average”. In: *arXiv.org* abs/2009.10526 (2020).

[204] Valentina Zantedeschi, Maria-Irina Nicolae, and Ambrish Rawat. “Efficient defenses against adversarial attacks”. In: *Proc. of the ACM Workshop on Artificial Intelligence and Security*. 2017.

[205] Nicholas Carlini and David A. Wagner. “MagNet and ”Efficient Defenses Against Adversarial Attacks” are Not Robust to Adversarial Examples”. In: *arXiv.org* abs/1711.08478 (2017).

[206] Weilin Xu, David Evans, and Yanjun Qi. “Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks”. In: *arXiv.org* abs/1704.01155 (2017).

[207] Arjun Nitin Bhagoji, Daniel Cullina, and Prateek Mittal. “Dimensionality Reduction as a Defense against Evasion Attacks on Machine Learning Classifiers”. In: *arXiv.org* abs/1704.02654 (2017).

[208] Aran Nayebi and Surya Ganguli. “Biologically inspired protection of deep networks from adversarial attacks”. In: *arXiv.org* abs/1703.09202 (2017).

[209] Nicolas Papernot et al. “Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks”. In: *Proc. of the IEEE Symposium on Security and Privacy*. 2016.

[210] Nicholas Carlini and David A. Wagner. “Defensive Distillation is Not Robust to Adversarial Examples”. In: *arXiv.org* abs/1607.04311 (2016).

[211] Micah Goldblum et al. “Adversarially Robust Distillation”. In: *arXiv.org* abs/1905.09747 (2019).

[212] Andrew Ilyas et al. “The Robust Manifold Defense: Adversarial Training using Generative Models”. In: *arXiv.org* abs/1712.09196 (2017).

[213] Rama Chellappa Pouya Samangouei Maya Kabkab. “Defense-GAN: Protecting Classifiers Against Adversarial Attacks Using Generative Models”. In: *Proc. of the International Conference on Learning Representations (ICLR)* (2018).

[214] Anonymous. “From Adversarial Robustness to Distributional Robustness: Rethinking the Brittleness of Neural Networks”. In: *Review for Proc. of the Conference on Artificial Intelligence (AAAI)*. 2019.

[215] Cihang Xie et al. “Mitigating adversarial effects through randomization”. In: *arXiv.org* abs/1711.01991 (2017).

[216] Siyue Wang et al. “Defensive dropout for hardening deep neural networks under adversarial attacks”. In: *Proc. of the International Conference on Computer-Aided Design*. 2018, 71:1–71:8.

[217] Anonymous. “Regularized Defense against Adversarial Attacks with SmoothBlock”. In: *Review for Proc. of the IEEE International Conference on Computer Vision (ICCV)*. 2019.

[218] Aaditya Prakash et al. “Deflecting Adversarial Attacks with Pixel Deflection”. In: *arXiv.org* abs/1801.08926 (2018).

[219] Chuan Guo et al. “Countering Adversarial Images using Input Transformations”. In: *arXiv.org* abs/1711.00117 (2017).

[220] Chunchuan Lyu, Kaizhu Huang, and Hai-Ning Liang. “A Unified Gradient Regularization Family for Adversarial Examples”. In: *IEEE International Conference on Data Mining*. 2015.

[221] Carl-Johann Simon-Gabriel et al. “Adversarial Vulnerability of Neural Networks Increases With Input Dimension”. In: *arXiv.org* abs/1802.01421 (2018).

[222] Matthias Hein and Maksym Andriushchenko. “Formal Guarantees on the Robustness of a Classifier against Adversarial Manipulation”. In: *arXiv.org* abs/1705.08475 (2017).

[223] Daniel Jakubovitz and Raja Giryes. “Improving DNN Robustness to Adversarial Attacks using Jacobian Regularization”. In: *arXiv.org* abs/1803.08680 (2018).

[224] Andrew Slavin Ross and Finale Doshi-Velez. “Improving the Adversarial Robustness and Interpretability of Deep Neural Networks by Regularizing their Input Gradients”. In: *arXiv.org* abs/1711.09404 (2017).

[225] Judy Hoffman, Daniel A. Roberts, and Sho Yaida. “Robust Learning with Jacobian Regularization”. In: *arXiv.org* abs/1908.02729 (2019).

[226] F. Yu et al. “Interpreting Adversarial Robustness: A View from Decision Surface in Input Space”. In: *arXiv.org* abs/1810.00144 (2018).

[227] Yao-Yuan Yang et al. “Adversarial Robustness Through Local Lipschitzness”. In: *arXiv.org* abs/2003.02460 (2020).

[228] Alvin Chan et al. “Jacobian Adversarially Regularized Networks for Robustness”. In: *arXiv.org* abs/1912.10185 (2019).

[229] Francesco Croce and Matthias Hein. “Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks”. In: *arXiv.org* abs/2003.01690 (2020).

[230] Anonymous. “Adversarial Robustness via Suppressing the Largest Eigenvalue of Fisher Information Matrix”. In: *Review for Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*. 2020.

[231] Arash Rahnama, André T. Nguyen, and Edward Raff. “Robust Design of Deep Neural Networks against Adversarial Attacks based on Lyapunov Theory”. In: *arXiv.org* abs/1911.04636 (2019).

[232] Sravanti Addepalli et al. “Towards Achieving Adversarial Robustness by Enforcing Feature Consistency Across Bit Planes”. In: *arXiv.org* abs/2004.00306 (2020).

[233] Philip Sperl and Konstantin Böttinger. “Optimizing Information Loss Towards Robust Neural Networks”. In: *arXiv.org* abs/2008.03072 (2020).

[234] Ziquan Liu, Yufei Cui, and Antoni B. Chan. “Improve Generalization and Robustness of Neural Networks via Weight Scale Shifting Invariant Regularizations”. In: *arXiv.org* abs/2008.02965 (2020).

[235] Cassidy Laidlaw, Sahil Singla, and Soheil Feizi. “Perceptual Adversarial Robustness: Defense Against Unseen Threat Models”. In: *arXiv.org* abs/2006.12655 (2020).

[236] Chongli Qin et al. “Adversarial Robustness through Local Linearization”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2019.

[237] Jacob Buckman et al. “Thermometer Encoding: One Hot Way To Resist Adversarial Examples”. In: *Proc. of the International Conference on Learning Representations (ICLR)*. 2018.

[238] Aaditya Prakash et al. “Protecting JPEG Images Against Adversarial Attacks”. In: *arXiv.org* abs/1803.00940 (2018).

[239] Naveed Akhtar, Jian Liu, and Ajmal S. Mian. “Defense against Universal Adversarial Perturbations”. In: *arXiv.org* abs/1711.05929 (2017).

[240] Mahdieh Abbasi and Christian Gagné. “Out-distribution training confers robustness to deep neural networks”. In: *arXiv.org* abs/1802.07124 (2018).

[241] Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. “Why ReLU networks yield high-confidence predictions far away from the training data and how to mitigate the problem”. In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2019).

[242] Harini Kannan, Alexey Kurakin, and Ian J. Goodfellow. “Adversarial Logit Pairing”. In: *arXiv.org* abs/1803.06373 (2018).

[243] Logan Engstrom, Andrew Ilyas, and Anish Athalye. “Evaluating and Understanding the Robustness of Adversarial Logit Pairing”. In: *arXiv.org* abs/1807.10272 (2018).

[244] Marius Mosbach et al. “Logit Pairing Methods Can Fool Gradient-Based Attacks”. In: *arXiv.org* abs/1810.12042 (2018).

[245] Alex Lamb et al. “Fortified Networks: Improving the Robustness of Deep Networks by Modeling the Manifold of Hidden Representations”. In: *arXiv.org* abs/1804.02485 (2018).

[246] Shiwei Shen et al. “APE-GAN: Adversarial Perturbation Elimination with GAN”. In: *arXiv.org* abs/1707.05474 (2017).

[247] Ali Shafahi et al. *Label Smoothing and Logit Squeezing: A Replacement for Adversarial Training?* https://openreview.net/forum?id=BJlr0j0ctX. 2018.

[248] Guanhong Tao et al. “Attacks Meet Interpretability: Attribute-steered Detection of Adversarial Samples”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2018, pp. 7728–7739.

[249] Nicholas Carlini. “Is AmI (Attacks Meet Interpretability) Robust to Adversarial Examples?” In: *arXiv.org* abs/1902.02322 (2019).

[250] Xiaoyu Cao and Neil Zhenqiang Gong. “Mitigating Evasion Attacks to Deep Neural Networks via Region-based Classification”. In: *Proc. of the Annual Computer Security Applications Conference*. 2017, pp. 278–287.

[251] Jeremy M. Cohen, Elan Rosenfeld, and J. Zico Kolter. “Certified Adversarial Robustness via Randomized Smoothing”. In: *arXiv.org* abs/1902.02918 (2019).

[252] Greg Yang et al. “Randomized Smoothing of All Shapes and Sizes”. In: *arXiv.org* abs/2002.08118 (2020).

[253] Aounon Kumar et al. “Certifying Confidence via Randomized Smoothing”. In: *ArXiv* abs/2009.08061 (2020).

[254] Aounon Kumar et al. “Curse of Dimensionality on Randomized Smoothing for Certifiable Robustness”. In: *arXiv.org* abs/2002.03239 (2020).

[255] Avrim Blum et al. “Random Smoothing Might be Unable to Certify L_∞ Robustness for High-Dimensional Images”. In: *arXiv.org* abs/2002.03517 (2020).

[256] Tianhang Zheng et al. “Towards Assessment of Randomized Mechanisms for Certifying Adversarial Robustness”. In: abs/2005.07347 (2020).

[257] Hadi Salman et al. “Denoised Smoothing: A Provable Defense for Pretrained Classifiers”. In: *arXiv.org* (2020).

[258] Greg Yang et al. “Randomized Smoothing of All Shapes and Sizes”. In: *Proc. of the International Conference on Machine Learning (ICML)*. 2020.

[259] Guang-He Lee et al. “Tight Certificates of Adversarial Robustness for Randomly Smoothed Classifiers”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2019.

[260] Bai Li et al. “Certified Adversarial Robustness with Additive Noise”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2019.

[261] Dinghuai Zhang et al. “Black-Box Certification with Randomized Smoothing: A Functional Optimization Based Framework”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2020.

[262] Jongheon Jeong and Jinwoo Shin. “Consistency Regularization for Certified Robustness of Smoothed Classifiers”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2020.

[263] Runtian Zhai et al. “MACER: Attack-free and Scalable Robust Training via Maximizing Certified Radius”. In: *Proc. of the International Conference on Learning Representations (ICLR)*. 2020.

[264] Jongheon Jeong et al. “SmoothMix: Training Confidence-calibrated Smoothed Classifiers for Certified Robustness”. In: *arXiv.org* abs/2111.09277 (2021).

[265] Krishnamurthy (Dj) Dvijotham et al. “A Framework for robustness Certification of Smoothed Classifiers using F-Divergences”. In: *Proc. of the International Conference on Learning Representations (ICLR)*. 2020.

[266] Jeet Mohapatra et al. “Higher-Order Certification For Randomized Smoothing”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2020.

[267] Rajeev Ranjan et al. “Improving Network Robustness against Adversarial Attacks with Compact Convolution”. In: *arXiv.org* abs/1712.00699 (2017).

[268] Anonymous. “DDSA: a Defense against Adversarial Attacks using Deep Denoising Sparse Autoencoder”. In: *Review for Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.

[269] Anonymous. “Defending Against Adversarial Attacks Using Random Forests”. In: *Review for Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.

[270] Krishna Kanth Nakka and Mathieu Salzmann. “Interpretable BoW Networks for Adversarial Example Detection”. In: *arXiv.org* abs/1901.02229 (2019).

[271] Anonymous. “Interpretable BoW Networks for Adversarial Example Detection”. In: *Review for Proc. of the IEEE International Conference on Computer Vision (ICCV)*. 2019.

[272] Fangzhou Liao et al. “Defense against Adversarial Attacks Using High-Level Representation Guided Denoiser”. In: *arXiv.org* abs/1712.02976 (2017).

[273] Anish Athalye and Nicholas Carlini. “On the Robustness of the CVPR 2018 White-Box Adversarial Example Defenses”. In: *arXiv.org* abs/1804.03286 (2018).

[274] Moustapha Cissé et al. “Parseval Networks: Improving Robustness to Adversarial Examples”. In: *Proc. of the International Conference on Machine Learning (ICML)*. 2017.

[275] Jonathan Aigrain and Marcin Detyniecki. “Improving Robustness Without Sacrificing Accuracy with patch Gaussian Augmentation”. In: *Proc. of the International Conference on Machine Learning (ICML) Workshops*. 2019.

[276] Zheng Xu, Ali Shafahi, and Tom Goldstein. “Exploring Model Robustness with Adaptive Networks and Improved Adversarial Training”. In: *arXiv.org* abs/2006.00387 (2020).

[277] Yuchen Zhang and Percy Liang. “Defending against Whitebox Adversarial Attacks via Randomized Discretization”. In: *arXiv.org* abs/1903.10586 (2019).

[278] Seyed-Mohsen Moosavi-Dezfooli, Ashish Shrivastava, and Oncel Tuzel. “Divide, Denoise, and Defend against Adversarial Attacks”. In: *arXiv.org* abs/1802.06806 (2018).

[279] Abhimanyu Dubey et al. “Defense Against Adversarial Images Using Web-Scale Nearest-Neighbor Search”. In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.

[280] Andras Rozsa and Terrance E. Boult. “Improved Adversarial Robustness by Reducing Open Space Risk via Tent Activations”. In: *arXiv.org* abs/1908.02435 (2019).

[281] Shixiang Gu and Luca Rigazio. “Towards Deep Neural Network Architectures Robust to Adversarial Examples”. In: *Proc. of the International Conference on Learning Representations (ICLR) Workshops*. 2015.

[282] Deepak Vijaykeerthy et al. “Hardening Deep Neural Networks via Adversarial Model Cascades”. In: *International Joint Conference on Neural Networks (IJCNN)*. 2019.

[283] Joachim Folz et al. “Adversarial Defense based on Structure-to-Signal Autoencoders”. In: *arXiv.org* abs/1803.07994 (2018).

[284] Shuntaro Miyazato et al. “Reinforcing the Robustness of a Deep Neural Network to Adversarial Examples by Using Color Quantization of Training Image Data”. In: *Proc. of the IEEE International Conference on Image Processing (ICIP)*. 2019.

[285] Aamir Mustafa et al. “Adversarial Defense by Restricting the Hidden Space of Deep Neural Networks”. In: *Proc. of the IEEE International Conference on Computer Vision (ICCV)*. 2019.

[286] A. Mustafa et al. “Deeply Supervised Discriminative Learning for Adversarial Defense”. In: (2020).

[287] Saeid Asgari Taghanaki et al. “A Kernelized Manifold Mapping to Diminish the Effect of Adversarial Perturbations”. In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.

[288] Jay Nandy, Wynne Hsu, and Mong Li Lee. “Approximate Manifold Defense Against Multiple Adversarial Perturbations”. In: *arXiv.org* abs/2004.02183 (2020).

[289] Jiawei Du et al. “RAIN: A Simple Approach for Robust and Accurate Image Classification Networks.” In: *arXiv.org* (2020).

[290] Manish V. Reddy et al. “Biologically Inspired Mechanisms for Adversarial Robustness”. In: *arXiv.org* abs/2006.16427 (2020).

[291] Han Qiu et al. “Mitigating Advanced Adversarial Attacks with More Advanced Gradient Obfuscation Techniques”. In: *arXiv.org* abs/2005.13712 (2020).

[292] Anonymous. “Constraining Logits by a Bounded Function for Adversarial Robustness”. In: *Review for Advances in Neural Information Processing Systems (NeurIPS)*. 2020.

[293] Anonymous. “Preparing for the Worst: Making Networks Less Brittle with Adversarial Batch Normalization”. In: *Review for Advances in Neural Information Processing Systems (NeurIPS)*. 2020.

[294] Hyeong ji Kim and Ketil Malde. “Proper measure for adversarial robustness”. In: *arXiv.org* (2020).

[295] Yueru Li et al. “Defense Against Adversarial Attacks via Controlling Gradient Leaking on Embedded Manifolds”. In: *Proc. of the European Conference on Computer Vision (ECCV)*. 2020.

[296] Saima Sharmin et al. “Inherent Adversarial Robustness of Deep Spiking Neural Networks: Effects of Discrete Input Encoding and Non-Linear Activations”. In: *Proc. of the European Conference on Computer Vision (ECCV)*. 2020.

[297] Xinshuai Dong et al. “API-Net: Robust Generative Classifier via a Single Discriminator”. In: *Proc. of the European Conference on Computer Vision (ECCV)*. 2020.

[298] Chaohao Fu et al. “Label Smoothing and Adversarial Robustness”. In: *arXiv.org* abs/2009.08233 (2020).

[299] Mingjie Sun et al. “Can Shape Structure Features Improve Model Robustness Under Diverse Adversarial Settings?” In: *Proc. of the IEEE International Conference on Computer Vision (ICCV)*. 2021.

[300] Jonathan Peck et al. “Lower bounds on the robustness to adversarial perturbations”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2017.

[301] Linyi Li et al. “SoK: Certified Robustness for Deep Neural Networks”. In: *arXiv.org* abs/2009.04131 (2020).

[302] Huan Zhang et al. “Efficient Neural Network Robustness Certification with General Activation Functions”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2018, pp. 4944–4953.

[303] Huan Zhang et al. “Towards Stable and Efficient Training of Verifiably Robust Neural Networks”. In: *arXiv.org* abs/1906.06316 (2019).

[304] Eric Wong and J. Zico Kolter. “Provable defenses against adversarial examples via the convex outer adversarial polytope”. In: *Proc. of the International Conference on Machine Learning (ICML)*. 2018.

[305] Sven Gowal et al. “On the Effectiveness of Interval Bound Propagation for Training Verifiably Robust Models”. In: *arXiv.org* abs/1810.12715 (2018).

[306] Shivam Garg et al. “A Spectral View of Adversarially Robust Features”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2018, pp. 10159–10169.

[307] Timon Gehr et al. “AI2: Safety and Robustness Certification of Neural Networks with Abstract Interpretation”. In: *Proc. of the IEEE Symposium on Security and Privacy*. 2018, pp. 3–18.

[308] Matthew Mirman, Timon Gehr, and Martin T. Vechev. “Differentiable Abstract Interpretation for Provably Robust Neural Networks”. In: *Proc. of the International Conference on Machine Learning (ICML)*. 2018, pp. 3575–3583.

[309] Gagandeep Singh et al. “Fast and Effective Robustness Certification”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2018, pp. 10825–10836.

[310] Guang-He Lee, David Alvarez-Melis, and Tommi S. Jaakkola. “Towards Robust, Locally Linear Deep Networks”. In: *arXiv.org* abs/1907.03207 (2019).

[311] Francesco Croce, Maksym Andriushchenko, and Matthias Hein. “Provable Robustness of ReLU networks via Maximization of Linear Regions”. In: *arXiv.org* abs/1810.07481 (2018).

[312] Akhilan Boopathy et al. “CNN-Cert: An Efficient Framework for Certifying Robustness of Convolutional Neural Networks”. In: *Proc. of the Conference on Artificial Intelligence (AAAI)*. 2019.

[313] Matthew Wicker et al. “Probabilistic Safety for Bayesian Neural Networks”. In: *arXiv.org* abs/2004.10281 (2020).

[314] Pranjal Awasthi et al. “Adversarial robustness via robust low rank representations”. In: *arXiv.org* abs/2007.06555 (2020).

[315] Tobias Lorenz et al. “Robustness Certification for Point Cloud Models”. In: *Proc. of the IEEE International Conference on Computer Vision (ICCV)*. 2021.

[316] Haowen Lin et al. “Integer-arithmetic-only Certified Robustness for Quantized Neural Networks”. In: *Proc. of the IEEE International Conference on Computer Vision (ICCV)*. 2021.

[317] Mathias Lécuyer et al. “Certified Robustness to Adversarial Examples with Differential Privacy”. In: *Proc. of the IEEE Symposium on Security and Privacy*. 2019.

[318] Zhitao Gong, Wenlu Wang, and Wei-Shinn Ku. “Adversarial and Clean Data Are Not Twins”. In: *arXiv.org* abs/1704.04960 (2017).

[319] Bita Darvish Rouhani et al. *Towards Safe Deep Learning: Unsupervised Defense Against Generic Adversarial Attacks*. https://openreview.net/forum?id=HyI6s40a-. 2018.

[320] Kathrin Grosse et al. “On the (statistical) detection of adversarial examples”. In: *arXiv.org* abs/1702.06280 (2017).

[321] Reuben Feinman et al. “Detecting Adversarial Samples from Artifacts”. In: *arXiv.org* abs/1703.00410 (2017).

[322] Fangzhou Liao et al. “Defense Against Adversarial Attacks Using High-Level Representation Guided Denoiser”. In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018.

[323] Xingjun Ma et al. “Characterizing Adversarial Subspaces Using Local Intrinsic Dimensionality”. In: *arXiv.org* abs/1801.02613 (2018).

[324] Laurent Amsaleg et al. “The vulnerability of learning to adversarial perturbation increases with intrinsic dimensionality”. In: *Proc. of the IEEE Workshop on Information Forensics and Security*. 2017.

[325] Jan Hendrik Metzen et al. “On Detecting Adversarial Perturbations”. In: *arXiv.org* abs/1702.04267 (2017).

[326] Dan Hendrycks and Kevin Gimpel. “Early Methods for Detecting Adversarial Images”. In: *Proc. of the International Conference on Learning Representations (ICLR)*. 2017.

[327] Xin Li and Fuxin Li. “Adversarial Examples Detection in Deep Networks with Convolutional Filter Statistics”. In: *Proc. of the IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 5775–5783.

[328] Kimin Lee et al. “A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2018, pp. 7167–7177.

[329] Kevin Roth, Yannic Kilcher, and Thomas Hofmann. “The Odds are Odd: A Statistical Test for Detecting Adversarial Examples”. In: *Proc. of the International Conference on Machine Learning (ICML)*. 2019, pp. 5498–5507.

[330] Nicholas Carlini and David Wagner. “Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods”. In: *arXiv.org* abs/1705.07263 (2017).

[331] Shengyuan Hu et al. “A New Defense Against Adversarial Images: Turning a Weakness into a Strength”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2019.

[332] Chenxiao Zhao et al. “The Adversarial Attack and Detection under the Fisher Information Metric”. In: *Proc. of the Conference on Artificial Intelligence (AAAI)*. 2019.

[333] Chiliang Zhang, Zhimou Yang, and Zuochang Ye. “Detecting Adversarial Perturbations with Saliency”. In: *arXiv.org* abs/1803.08773 (2018).

[334] Ninghao Liu, Hongxia Yang, and Xia Hu. “Adversarial Detection with Model Interpretation”. In: *Proc. of the ACM International Conference on Knowledge Discovery & Data Mining*. Ed. by Yike Guo and Faisal Farooq. 2018.

[335] Chengcheng Ma et al. “Effective and Robust Detection of Adversarial Examples via Benford-Fourier Coefficients”. In: *arXiv.org* abs/2005.05552 (2020).

[336] Shasha Li et al. “Connecting the Dots: Detecting Adversarial Perturbations Using Context Inconsistency”. In: *Proc. of the European Conference on Computer Vision (ECCV)*. 2020.

[337] Ahmed Abusnaina et al. “Adversarial Example Detection Using Latent Neighborhood Graph”. In: *Proc. of the IEEE International Conference on Computer Vision (ICCV)*. 2021.

[338] Peter Lorenz et al. “Detecting AutoAttack Perturbations in the Frequency Domain”. In: *arXiv.org* abs/2111.08785 (2021).

[339] Oliver Bryniarski et al. “Evading Adversarial Example Detection Defenses with Orthogonal Projected Gradient Descent”. In: *arXiv.org* abs/2106.15023 (2021).

[340] Shawn Shan et al. “Gotta Catch’Em All: Using Honeypots to Catch Adversarial Attacks on Neural Networks”. In: *Proc. of the ACM Conference on Computer and Communications Security (CCS)*. 2020.

[341] Philip Sperl et al. “DLA: Dense-Layer-Analysis for Adversarial Example Detection”. In: *Proc. of the IEEE European Symposium on Security and Privacy (EuroS&P)*. 2020.

[342] Jinyu Tian et al. “Detecting Adversarial Examples from Sensitivity Inconsistency of Spatial-Transform Domain”. In: *Proc. of the Conference on Artificial Intelligence (AAAI)*. 2021.

[343] Jiayang Liu et al. “Detection Based Defense Against Adversarial Examples From the Steganalysis Point of View”. In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.

[344] Yao Li et al. “Detecting Adversarial Examples with Bayesian Neural Network”. In: *arXiv.org* abs/2105.08620 (2021).

[345] Nicolas Papernot, Patrick D. McDaniel, and Ian J. Goodfellow. “Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples”. In: *arXiv.org* abs/1605.07277 (2016).

[346] Nicolas Papernot et al. “Practical black-box attacks against machine learning”. In: *Proc. of the ACM on Asia Conference on Computer and Communications Security (AsiaCCS)*. ACM. 2017.

[347] Ambra Demontis et al. “On the Intriguing Connections of Regularization, Input Gradients and Transferability of Evasion and Poisoning Attacks”. In: *arXiv.org* abs/1809.02861 (2018).

[348] Cihang Xie et al. “Improving Transferability of Adversarial Examples with Input Diversity”. In: *arXiv.org* abs/1803.06978 (2018).

[349] Dongxian Wu et al. “Skip Connections Matter: On the Transferability of Adversarial Examples Generated with ResNets”. In: *arXiv.org* abs/2002.05990 (2020).

[350] Florian Tramèr et al. “The Space of Transferable Adversarial Examples”. In: *arXiv.org* abs/1704.03453 (2017).

[351] Yinpeng Dong et al. “Evading Defenses to Transferable Adversarial Examples by Translation-Invariant Attacks”. In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.

[352] Ambra Demontis et al. “Why Do Adversarial Attacks Transfer? Explaining Transferability of Evasion and Poisoning Attacks”. In: *USENIX Security Symposium*. 2019.

[353] Anish Athalye, Nicholas Carlini, and David A. Wagner. “Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples”. In: *arXiv.org* abs/1802.00420 (2018).

[354] Yash Sharma and Pin-Yu Chen. “Attacking the Madry Defense Model with L1-based Adversarial Examples”. In: *arXiv.org* abs/1710.10733 (2017).

[355] Florian Tramèr andNicholas Carlini andWieland Brendel andAleksander Madry. “On Adaptive Attacks to Adversarial Example Defenses”. In: *arXiv.org* abs/2002.08347 (2020).

[356] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. “Adversarial machine learning at scale”. In: *arXiv.org* abs/1611.01236 (2016).

[357] Mahmood Sharif, Lujo Bauer, and Michael K. Reiter. “On the Suitability of L_p-norms for Creating and Preventing Adversarial Examples”. In: *arXiv.org* abs/1802.09653 (2018).

[358] Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. “Fundamental limits on adversarial robustness”. In: *Proc. of the International Conference on Machine Learning (ICML) Workshops*. 2015.

[359] Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. “Robustness of classifiers: from adversarial to random noise”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2016.

[360] Haosheng Zou et al. “On the Universality of Adversarial Examples in Deep Learning”. In: (2018). URL: http://ml.cs.tsinghua.edu.cn/~haosheng/static/universality-adv.pdf.

[361] Ali Shafahi et al. “Are adversarial examples inevitable?” In: *arXiv.org* abs/1809.02104 (2018).

[362] Beilun Wang, Ji Gao, and Yanjun Qi. “A Theoretical Framework for Robustness of (Deep) Classifiers Under Adversarial Noise”. In: *CoRR* abs/1612.00334 (2016).

[363] Dimitris Tsipras et al. “Robustness May Be at Odds with Accuracy”. In: *arXiv.org* abs/1805.12152 (2018).

[364] Andrew Ilyas et al. “Adversarial Examples Are Not Bugs, They Are Features”. In: *arXiv.org* abs/1905.02175 (2019).

[365] Zuowen Wang and Leo Horne. “Understanding (Non-)Robust Feature Disentanglement and the Relationship Between Low- and High-Dimensional Adversarial Attacks”. In: *arXiv.org* abs/2004.01903 (2020).

[366] Kamil Nar et al. *Cross-Entropy Loss Leads To Poor Margins*. https://openreview.net/forum?id=ByfbnsA9Km. 2019.

[367] Angus Galloway, Anna Golubeva, and Graham W. Taylor. "Adversarial Examples as an Input-Fault Tolerance Problem". In: *arXiv.org* abs/1811.12601 (2018).

[368] Andras Rozsa, Ethan M. Rudd, and Terrance E. Boult. "Adversarial Diversity and Hard Positive Generation". In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2016.

[369] Sanghuyk Chun et al. "An Empirical Evaluation on Robustness and Uncertainty of Regularization Methods". In: *Proc. of the International Conference on Machine Learning (ICML) Workshops* (2019).

[370] Alhussein Fawzi et al. "Empirical Study of the Topology and Geometry of Deep Networks". In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 3762–3770.

[371] Saumya Jetley, Nicholas A. Lord, and Philip H. S. Torr. "With Friends Like These, Who Needs Adversaries?" In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2018, pp. 10772–10782.

[372] Adi Shamir et al. "A Simple Explanation for the Existence of Adversarial Examples with Small Hamming Distance". In: *arXiv.org* abs/1901.10861 (2019).

[373] Ambrish Rawat, Martin Wistuba, and Maria-Irina Nicolae. "Adversarial Phenomenon in the Eyes of Bayesian Deep Learning". In: *arXiv.org* abs/1711.08244 (2017).

[374] Yue Gao et al. "Analyzing Accuracy Loss in Randomized Smoothing Defenses". In: *arXiv.org* abs/2003.01595 (2020).

[375] Vasu Singla et al. "Low Curvature Activations Reduce Overfitting in Adversarial Training". In: *arXiv.org* abs/2102.07861 (2021).

[376] Ludwig Schmidt et al. "Adversarially Robust Generalization Requires More Data". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2018.

[377] Evelyn Duesterwald et al. "Exploring the Hyperparameter Landscape of Adversarial Robustness". In: *arXiv.org* abs/1905.03837 (2019).

[378] Joyce Xu, Dian Ang Yap, and Vinay Uday Prabhu. "Understanding Adversarial Robustness Through Loss Landscape Geometries". In: *Proc. of the International Conference on Machine Learning (ICML) Workshops* (2019).

[379] Vinay Uday Prabhu et al. "Understanding Adversarial Robustness Through Loss Landscape Geometries". In: *arXiv.org* abs/1907.09061 (2019).

[380] Bai Li et al. "On Norm-Agnostic Adversarial Robustness Between Perturbation Types". In: *Proc. of the International Conference on Machine Learning (ICML) Workshops*. 2019.

[381] Xuanqing Liu and Cho-Jui Hsieh. "From Adversarial Training to Generative Adversarial Networks". In: *arXiv.org* abs/1807.10454 (2018).

[382] Jiachen Zhong, Xuanqing Liu, and Cho-Jui Hsieh. "Improving the Speed and Quality of GAN by Adversarial Training". In: *arXiv.org* abs/2008.03364 (2020).

[383] Farzan Farnia, Jesse M. Zhang, and David Tse. "Generalizable Adversarial Training via Spectral Normalization". In: *Proc. of the International Conference on Learning Representations (ICLR)*. 2019.

[384] Jerry Li et al. "Towards Understanding the Dynamics of Generative Adversarial Networks". In: *arXiv.org* abs/1706.09884 (2017).

[385] Leslie Rice, Eric Wong, and J. Zico Kolter. "Overfitting in adversarially robust deep learning". In: *arXiv.org* abs/2002.11569 (2020).

[386] Prasad Chalasani et al. "Concise Explanations of Neural Networks using Adversarial Training". In: *arXiv.org* abs/1810.06583 (2018).

[387] Tianyu Pang et al. "Bag of Tricks for Adversarial Training". In: *arXiv.org* abs/2010.00467 (2020).

[388] Sven Gowal et al. "Uncovering the Limits of Adversarial Training against Norm-Bounded Adversarial Examples". In: *arXiv.org* abs/2010.03593 (2020).

[389] Boxi Wu et al. "Does Network Width Really Help Adversarial Robustness?" In: *arXiv.org* abs/2010.01279 (2020).

[390] Han Xu et al. “To be Robust or to be Fair: Towards Fairness in Adversarial Training”. In: *Proc. of the International Conference on Machine Learning (ICML)*. 2021.

[391] Vedant Nanda et al. “Fairness Through Robustness: Investigating Robustness Disparity in Deep Learning”. In: *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. 2021.

[392] Aditi Raghunathan et al. “Adversarial Training Can Hurt Generalization”. In: *arXiv.org* abs/1906.06032 (2019).

[393] Dong Su et al. “Is Robustness the Cost of Accuracy? – A Comprehensive Study on the Robustness of 18 Deep Image Classification Models”. In: *arXiv.org* abs/1808.01688 (2018).

[394] Huan Xu and Shie Mannor. “Robustness and Generalization”. In: *Proc. of the Conference on Learning Theory (COLT)*. 2010, pp. 503–515.

[395] Yao-Yuan Yang et al. “A Closer Look at Accuracy vs. Robustness”. In: *arXiv.org* abs/2003.02460 (2020).

[396] Aditi Raghunathan et al. “Understanding and Mitigating the Tradeoff Between Robustness and Accuracy”. In: *arXiv.org* abs/2002.10716 (2020).

[397] Yao-Yuan Yang et al. “A Closer Look at Accuracy vs. Robustness”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2020.

[398] Nic Ford et al. “Adversarial Examples Are a Natural Consequence of Test Error in Noise”. In: *arXiv.org* abs/1901.10513 (2019).

[399] H. Li et al. “Verifying the Causes of Adversarial Examples”. In: *arXiv.org* abs/2010.09633 (2020).

[400] Thomas Tanay and Lewis Griffin. “A boundary tilting persepective on the phenomenon of adversarial examples”. In: *arXiv.org* abs/1608.07690 (2016).

[401] Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. “Deep Image Prior”. In: *arXiv.org* abs/1711.10925 (2017).

[402] Ronen Basri and David W. Jacobs. “Efficient Representation of Low-Dimensional Manifolds using Deep Networks”. In: *arXiv.org* abs/1602.04723 (2016).

[403] Byungjoo Kim et al. “Robust Neural Networks inspired by Strong Stability Preserving Runge-Kutta methods”. In: *Proc. of the European Conference on Computer Vision (ECCV)*. Vol. abs/2005.02540. 2020.

[404] Giacomo De Palma, Bobak Toussi Kiani, and Seth Lloyd. “Adversarial robustness guarantees for random deep neural networks”. In: *arXiv.org* abs/2004.05923 (2020).

[405] Sandesh Kamath, Amit Deshpande, and K. V. Subrahmanyam. “How do SGD hyperparameters in natural training affect adversarial robustness?” In: *arXiv.org* abs/2006.11604 (2020).

[406] Muhammad Awais, Fahad Shamshad, and Sungho Bae. “Towards an Adversarially Robust Normalization Approach”. In: *arXiv.org* abs/2006.11007 (2020).

[407] Philipp Benz, C. Zhang, and I. Kweon. “Batch Normalization Increases Adversarial Vulnerability: Disentangling Usefulness and Robustness of Model Features”. In: *arXiv.org* abs/2010.03316 (2020).

[408] Zhun Deng et al. “Improving Adversarial Robustness via Unlabeled Out-of-Domain Data”. In: *arXiv.org* abs/2006.08476 (2020).

[409] Fu Lin et al. “Likelihood Landscapes: A Unifying Principle Behind Many Adversarial Defenses”. In: *arXiv.org* abs/2008.11300 (2020).

[410] Matteo Terzi et al. “Adversarial Training Reduces Information and Improves Transferability”. In: *arXiv.org* abs/2007.11259 (2020).

[411] Hadi Salman et al. “Do Adversarially Robust ImageNet Models Transfer Better?” In: *arXiv.org* abs/2007.08489 (2020).

[412] S. Kamath, A. Deshpande, and K. Subrahmanyam. “Invariance vs. Robustness of Neural Networks”. In: *arXiv.org* abs/2002.11318 (2020).

[413] Sadaf Gulshad, J. H. Metzen, and A. Smeulders. “Adversarial and Natural Perturbations for General Robustness”. In: *arXiv.org* abs/2010.01401 (2020).

[414] Maria-Florina Balcan et al. “On the Power of Abstention and Data-Driven Decision Making for Adversarial Robustness”. In: *arXiv.org* abs/2010.06154 (2020).

[415] Kaleel Mahmood, Rigel Mahmood, and Marten van Dijk. “On the Robustness of Vision Transformers to Adversarial Examples”. In: *Proc. of the IEEE International Conference on Computer Vision (ICCV)*. 2021.

[416] Xin Wang et al. “Interpreting Attributions and Interactions of Adversarial Attacks”. In: *Proc. of the IEEE International Conference on Computer Vision (ICCV)*. 2021.

[417] M. K. Yucel, Ramazan Gokberk Cinbis, and P. D. Sahin. “A Deep Dive into Adversarial Robustness in Zero-Shot Learning”. In: *arXiv.org* abs/2008.07651 (2020).

[418] Chelsea Finn, P. Abbeel, and S. Levine. “Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks”. In: *Proc. of the International Conference on Machine Learning (ICML)*. 2017.

[419] Micah Goldblum, Liam Fowl, and Tom Goldstein. “Robust Few-Shot Learning with Adversarially Queried Meta-Learners”. In: *arXiv.org* abs/1910.00982 (2019).

[420] Chengxiang Yin et al. “Adversarial Meta-Learning”. In: *arXiv.org* abs/1806.03316 (2018).

[421] A. Shafahi et al. “Adversarially robust transfer learning”. In: *arXiv.org* abs/1905.08232 (2020).

[422] Yiming Li et al. “Toward Adversarial Robustness via Semi-supervised Robust Training”. In: *arXiv.org* abs/2003.06974 (2020).

[423] Minseon Kim, Jihoon Tack, and Sung Ju Hwang. “Adversarial Self-Supervised Contrastive Learning”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2020.

[424] Dan Hendrycks et al. “Using Self-Supervised Learning Can Improve Model Robustness and Uncertainty”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2019.

[425] Ziyu Jiang et al. “Robust Pre-Training by Adversarial Contrastive Learning”. In: *arXiv.org* abs/2010.13337 (2020).

[426] Jingfeng Zhang et al. “Where is the Bottleneck of Adversarial Learning with Unlabeled Data?” In: *arXiv.org* abs/1911.08696 (2019).

[427] Sandy H. Huang et al. “Adversarial Attacks on Neural Network Policies”. In: *arXiv.org* abs/1702.02284 (2017).

[428] Yen-Chen Lin et al. “Tactics of Adversarial Attack on Deep Reinforcement Learning Agents”. In: *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI)*. 2017.

[429] Björn Lütjens, Michael Everett, and Jonathon P. How. “Certified Adversarial Robustness for Deep Reinforcement Learning”. In: *Proc. of the Conference on Robot Learning (CoRL)*. 2019.

[430] Shu Hu et al. “T_kML-AP: Adversarial Attacks to Top-k Multi-Label Learning”. In: *Proc. of the IEEE International Conference on Computer Vision (ICCV)*. 2021.

[431] Amirata Ghorbani, Abubakar Abid, and James Y. Zou. “Interpretation of Neural Networks is Fragile”. In: *arXiv.org* abs/1710.10547 (2017).

[432] Mayank Singh et al. “Attributional Robustness Training using Input-Gradient Spatial Alignment”. In: *Proc. of the European Conference on Computer Vision (ECCV)*. 2020.

[433] Maximilian Augustin, Alexander Meinke, and Matthias Hein. “Adversarial Robustness on In- and Out-Distribution Improves Explainability”. In: *Proc. of the European Conference on Computer Vision (ECCV)*. 2020.

[434] Beomsu Kim, Junghoon Seo, and Taegyun Jeon. “Bridging Adversarial Robustness and Gradient Interpretability”. In: *arXiv.org* abs/1903.11626 (2019).

[435] Jiawang Bai et al. “Targeted Attack for Deep Hashing based Retrieval”. In: *Proc. of the European Conference on Computer Vision (ECCV)*. 2020.

[436] Chengzhi Mao et al. “Multitask Learning Strengthens Adversarial Robustness”. In: *Proc. of the European Conference on Computer Vision (ECCV)*. 2020.

[437] Qingquan Song et al. “Multi-Label Adversarial Perturbations”. In: *arXiv.org* abs/1901.00546 (2019).

[438] Volker Fischer et al. “Adversarial Examples for Semantic Image Segmentation”. In: *arXiv.org* abs/1703.01101 (2017).

[439] Moustapha M Cisse et al. “Houdini: Fooling Deep Structured Visual and Speech Recognition Models with Adversarial Examples”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2017.

[440] Pedro Tabacof, Julia Tavares, and Eduardo Valle. “Adversarial Images for Variational Autoencoders”. In: *arXiv.org* abs/1612.00155 (2016).

[441] Jernej Kos, Ian Fischer, and Dawn Song. “Adversarial examples for generative models”. In: *arXiv.org* abs/1702.06832 (2017).

[442] Marco Melis et al. “Is Deep Learning Safe for Robot Vision? Adversarial Examples Against the iCub Humanoid”. In: *Proc. of the IEEE International Conference on Computer Vision (ICCV) Workshops*. 2017.

[443] Hongge Chen et al. “Show-and-Fool: Crafting Adversarial Examples for Neural Image Captioning”. In: *arXiv.org* abs/1712.02051 (2017).

[444] Andras Rozsa, Manuel Günther, and Terrance E. Boult. “LOTS about attacking deep features”. In: *Proc. of the IEEE International Joint Conference on Biometrics, IJCB 2017*. 2017.

[445] Anonymous. “Evading Face Recognition via Partial Adversarial Tampering”. In: *Review for Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.

[446] Anonymous. “Adversarial Attacks on Monocular Depth Estimation”. In: *Review for Proc. of the IEEE International Conference on Computer Vision (ICCV)*. 2019.

[447] Alex Wong, Safa Cicek, and Stefano Soatto. “Targeted Adversarial Perturbations for Monocular Depth Prediction”. In: *arXiv.org* abs/2006.08602 (2020).

[448] Lars Aurdal et al. “Adversarial camouflage (AC) for naval vessels”. In: *Artificial Intelligence and Machine Learning in Defense Applications*. Ed. by Judith Dijk. Vol. 11169. International Society for Optics and Photonics. SPIE, 2019.

[449] Chaowei Xiao et al. “MeshAdv: Adversarial Meshes for Visual Recognition”. In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.

[450] Chong Xiang, Charles R. Qi, and Bo Li. “Generating 3D Adversarial Point Clouds”. In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.

[451] Kibok Lee et al. “ShapeAdv: Generating Shape-Aware Adversarial 3D Point Clouds”. In: *arXiv.org* abs/2005.11626 (2020).

[452] Jaeyeon Kim et al. “Minimal Adversarial Examples for Deep Learning on 3D Point Clouds”. In: *arXiv.org*. 2021.

[453] Mahmood Sharif et al. “Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition”. In: *Proc. of the ACM Conference on Computer and Communications Security*. 2016.

[454] Kevin Eykholt et al. “Robust Physical-World Attacks on Deep Learning Visual Classification”. In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018.

[455] Christian Cosgrove and Alan L. Yuille. “Adversarial Examples for Edge Detection: They Exist, and They Transfer”. In: *arXiv.org* abs/1906.00335 (2019).

[456] Mo Zhou et al. “Adversarial Ranking Attack and Defense”. In: *arXiv.org* abs/2002.11293 (2020).

[457] Xingxing Wei et al. “Sparse Adversarial Perturbations for Videos”. In: *Proc. of the Conference on Artificial Intelligence (AAAI)*. 2019.

[458] Lu Wang et al. “Provably Robust Metric Learning”. In: *arXiv.org* abs/2006.07024 (2020).

[459] Parsa Saadatpanah, Ali Shafahi, and Tom Goldstein. “Adversarial attacks on Copyright Detection Systems”. In: *arXiv.org* abs/1906.07153 (2019).

[460] Sicheng Zhu, Xinqi Zhang, and David Evans. “Learning Adversarially Robust Representations via Worst-Case Mutual Information Maximization”. In: *arXiv.org* abs/2002.11798 (2020).

[461] R. Shao et al. “Open-set Adversarial Defense”. In: *arXiv.org* abs/2009.00814 (2020).

[462] James Tu et al. “Adversarial Attacks On Multi-Agent Communication”. In: *Proc. of the IEEE International Conference on Computer Vision (ICCV)*. 2021.

[463] Jiaxing Huang et al. “RDA: Robust Domain Adaptation via Fourier Adversarial Attacking”. In: *Proc. of the IEEE International Conference on Computer Vision (ICCV)*. 2021.

[464] Samuel G. Finlayson, Isaac S. Kohane, and Andrew L. Beam. “Adversarial Attacks Against Medical Deep Learning Systems”. In: *arXiv.org* abs/1804.05296 (2018).

[465] Ibrahim Yilmaz. “Practical Fast Gradient Sign Attack against Mammographic Image Classifier”. In: *arXiv.org* abs/2001.09610 (2020).

[466] Hokuto Hirano, Kazuki Koga, and Kazuhiro Takemoto. “Vulnerability of deep neural networks for detecting COVID-19 cases from chest X-ray images to universal adversarial attacks”. In: *arXiv.org* abs/2005.11061 (2020).

[467] Daniel Zügner, Amir Akbarnejad, and Stephan Günnemann. “Adversarial Attacks on Neural Networks for Graph Data”. In: *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI)*. 2019.

[468] Hanjun Dai et al. “Adversarial Attack on Graph Structured Data”. In: *Proc. of the International Conference on Machine Learning (ICML)*. 2018.

[469] Jiaqi Ma, Shuangrui Ding, and Qiaozhu Mei. “Black-Box Adversarial Attacks on Graph Neural Networks with Limited Node Access”. In: *arXiv.org* abs/2006.05057 (2020).

[470] Hung Dang, Yue Huang, and Ee-Chien Chang. “Evading Classifiers by Morphing in the Dark”. In: *Proc. of the ACM Conference on Computer and Communications Security (CCS)*. 2017, pp. 119–133.

[471] Florian Tramèr et al. “AdVersarial: Perceptual Ad Blocking meets Adversarial Machine Learning”. In: *Proc. of the ACM Conference on Computer and Communications Security (CCS)*. 2019.

[472] Vincent Ballet et al. “Imperceptible Adversarial Attacks on Tabular Data”. In: *arXiv.org* abs/1911.03274 (2019).

[473] Wei Emma Zhang, Quan Z. Sheng, and Ahoud Abdulrahmn F. Alhazmi. “Generating Textual Adversarial Examples for Deep Learning Models: A Survey”. In: *arXiv.org* abs/1901.06796 (2019).

[474] Aminul Huq and Mst. Tasnim Pervin. “Adversarial Attacks and Defense on Texts: A Survey”. In: *arXiv.org* abs/2005.14108 (2020).

[475] Mohit Iyyer et al. “Adversarial Example Generation with Syntactically Controlled Paraphrase Networks”. In: *arXiv.org* abs/1804.06059 (2018).

[476] Javid Ebrahimi et al. “HotFlip: White-Box Adversarial Examples for NLP”. In: *arXiv.org* abs/1712.06751 (2017).

[477] Samuel Barham and Soheil Feizi. “Interpretable Adversarial Training for Text”. In: *arXiv.org* abs/1905.12864 (2019).

[478] Hanjun Dai et al. “Adversarial Attack on Graph Structured Data”. In: *arXiv.org* abs/1806.02371 (2018).

[479] Puyudi Yang et al. “Greedy Attack and Gumbel Attack: Generating Adversarial Examples for Discrete Data”. In: *arXiv.org* abs/1805.12316 (2018).

[480] Jinfeng Li et al. “TextBugger: Generating Adversarial Text Against Real-world Applications”. In: *Annual Network and Distributed System Security Symposium*. 2019.

[481] Di Jin et al. “Is BERT Really Robust? Natural Language Attack on Text Classification and Entailment”. In: *arXiv.org* abs/1907.11932 (2019).

[482] Yotam Gil et al. “White-to-Black: Efficient Distillation of Black-Box Adversarial Attacks”. In: *arXiv.org* abs/1904.02405 (2019).

[483] Robin Jia and Percy Liang. “Adversarial Examples for Evaluating Reading Comprehension Systems”. In: *Proc. of the Conference on Empirical Methods in Natural Language Processing*. 2017, pp. 2021–2031.

[484] Yicheng Wang and Mohit Bansal. “Robust Machine Comprehension Models via Adversarial Training”. In: *arXiv.org* abs/1804.06473 (2018).

[485] Siddhant Garg and Goutham Ramakrishnan. “BAE: BERT-based Adversarial Examples for Text Classification”. In: *arXiv.org* abs/2004.01970 (2020).

[486] Sining Sun et al. “Training Augmentation with Adversarial Examples for Robust Speech Recognition”. In: *arXiv.org* abs/1806.02782 (2018).

[487] Yuxuan Chen et al. “Devil’s Whisper: A General Approach for Physical Adversarial Attacks against Commercial Black-box Speech Recognition Devices”. In: *USENIX Security Symposium*. 2020.

[488] Deepak Kumar et al. “Skill Squatting Attacks on Amazon Alexa”. In: *USENIX Security Symposium*. 2018.

[489] Zhuolin Yang et al. *Towards Mitigating Audio Adversarial Perturbations*. https://openreview.net/forum?id=SyZ2nKJDz. 2018.

[490] Hiromu Yakura and Jun Sakuma. “Robust Audio Adversarial Example for a Physical Attack”. In: *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI)*. 2019.

[491] Nicholas Carlini and Hany Farid. “Evading Deepfake-Image Detectors with White- and Black-Box Attacks”. In: *arXiv.org* abs/2004.00622 (2020).

[492] Chin-Yuan Yeh et al. “Attack as the Best Defense: Nullifying Image-to-image Translation GANs via Limit-aware Adversarial Attack”. In: (2021).

[493] Han Xu et al. “To be Robust or to be Fair: Towards Fairness in Adversarial Training”. In: *arXiv.org* abs/2010.06121 (2020).

[494] Roman Werpachowski, András György, and Csaba Szepesvári. “Detecting Overfitting via Adversarial Examples”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2019.

[495] Cihang Xie et al. “Adversarial Examples Improve Image Recognition”. In: *arXiv.org* abs/1911.09665 (2019).

[496] Mehmet Sinan Inci, Thomas Eisenbarth, and Berk Sunar. “DeepCloak: Adversarial Crafting As a Defensive Measure to Cloak Processes”. In: *arXiv.org* abs/1808.01352 (2018).

[497] Lifeng Huang et al. “UPC: Learning Universal Physical Camouflage Attacks on Object Detectors”. In: *arXiv.org* abs/1909.04326 (2019).

[498] Jinyuan Jia and Neil Zhenqiang Gong. “AttriGuard: A Practical Defense Against Attribute Inference Attacks via Adversarial Machine Learning”. In: *USENIX Security Symposium*. USENIX Association, 2018.

[499] Ian J. Goodfellow. “A Research Agenda: Dynamic Models to Defend Against Correlated Attacks”. In: *arXiv.org* abs/1903.06293 (2019).

[500] Jonas Rauber, Wieland Brendel, and Matthias Bethge. “Foolbox v0.8.0: A Python toolbox to benchmark the robustness of machine learning models”. In: *arXiv.org* abs/1707.04131 (2017).

[501] Yaxin Li et al. “DeepRobust: A PyTorch Library for Adversarial Attacks and Defenses”. In: *arXiv.org* abs/2005.06149 (2020).

[502] Gavin Weiguang Ding, Luyu Wang, and Xiaomeng Jin. “advertorch v0.1: An Adversarial Robustness Toolbox based on PyTorch”. In: *arXiv.org* abs/1902.07623 (2019).

[503] Maria-Irina Nicolae et al. “Adversarial Robustness Toolbox v0.2.2”. In: *arXiv.org* abs/1807.01069 (2018).

[504] Seyed-Mohsen Moosavi-Dezfooli et al. “Universal adversarial perturbations”. In: *arXiv.org* abs/1610.08401 (2016).

[505] Konda Reddy Mopuri, Aditya Ganeshan, and R. Venkatesh Babu. “Generalizable Data-Free Objective for Crafting Universal Adversarial Perturbations”. In: *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)* 41.10 (2019).

[506] Konda Reddy Mopuri, Utsav Garg, and R. Venkatesh Babu. “Fast Feature Fool: A data independent approach to universal adversarial perturbations”. In: *arXiv.org* abs/1707.05572 (2017).

[507] Konda Reddy Mopuri, Phani Krishna Uppala, and R. Venkatesh Babu. “Ask, Acquire, and Attack: Data-Free UAP Generation Using Class Impressions”. In: *Proc. of the European Conference on Computer Vision (ECCV)*. 2018.

[508] Ali Shafahi et al. “Universal Adversarial Training”. In: *arXiv.org* abs/1811.11304 (2018).

[509] Anonymous. “Universal Adversarial Training”. In: *Review for Proc. of the IEEE International Conference on Computer Vision (ICCV)*. 2019.

[510] Julien Pérolat et al. “Playing the Game of Universal Adversarial Perturbations”. In: *arXiv.org* abs/1809.07802 (2018).

[511] Paarth Neekhara et al. “Universal Adversarial Perturbations for Speech Recognition Systems”. In: *arXiv.org* abs/1905.03828 (2019).

[512] Tom B. Brown et al. “Adversarial Patch”. In: *arXiv.org* abs/1712.09665 (2017).

[513] Xin Liu et al. “DPatch: Attacking Object Detectors with Adversarial Patches”. In: *arXiv.org* abs/1806.02299 (2018).

[514] Anonymous. “Fooling automated surveillance cameras: adversarial patches to attack person detection”. In: *Review for Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.

[515] Mark Lee and Zico Kolter. “On Physical Adversarial Patches for Object Detection”. In: *arXiv.org* abs/1906.11897 (2019).

[516] Rey Reza Wiyatno and Anqi Xu. “Physical Adversarial Textures that Fool Visual Object Tracking”. In: *arXiv.org* abs/1904.11042 (2019).

[517] Michal Zajac et al. “Adversarial Framing for Image and Video Classification”. In: *Proc. of the Conference on Artificial Intelligence (AAAI) Workshops*. 2019.

[518] Danny Karmon, Daniel Zoran, and Yoav Goldberg. “LaVAN: Localized and Visible Adversarial Noise”. In: *Proc. of the International Conference on Machine Learning (ICML)*. 2018.

[519] Anurag Ranjan et al. “Attacking Optical Flow”. In: *arXiv.org* abs/1910.10053 (2019).

[520] Anonymous. “Targeted Attention Attack on Deep Learning Models in Road Sign Recognition”. In: *Review for Proc. of the European Conference on Computer Vision (ECCV)*. 2020.

[521] Cheng-Lin Yang et al. “PatchAttack: A Black-box Texture-based Attack with Reinforcement Learning”. In: *arXiv.org* abs/2004.05682 (2020).

[522] Neil Fendley et al. “Jacks of All Trades, Masters Of None: Addressing Distributional Shift and Obtrusiveness via Transparent Patch Attacks”. In: *arXiv.org* abs/2005.00656 (2020).

[523] Aishan Liu et al. “Patch Attack for Automatic Check-out”. In: *arXiv.org* abs/2005.09257 (2020).

[524] Aishan Liu et al. “Bias-Based Universal Adversarial Patch Attack for Automatic Check-Out”. In: *Proc. of the European Conference on Computer Vision (ECCV)*. 2020.

[525] Xiaowei Yang et al. “Design and Interpretation of Universal Adversarial Patches in Face Detection”. In: *Proc. of the European Conference on Computer Vision (ECCV)*. 2019.

[526] Cheng-Lin Yang et al. “PatchAttack: A Black-box Texture-based Attack with Reinforcement Learning”. In: *Proc. of the European Conference on Computer Vision (ECCV)*. 2020.

[527] L. Gao et al. “Patch-wise Attack for Fooling Deep Neural Network”. In: *Proc. of the European Conference on Computer Vision (ECCV)*. 2020.

[528] Jamie Hayes. “On Visible Adversarial Perturbations & Digital Watermarking”. In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018.

[529] Ping yeh Chiang* et al. “Certified Defenses for Adversarial Patches”. In: *Proc. of the International Conference on Learning Representations (ICLR)*. 2020.

[530] Fei Zuo et al. “Exploiting the Inherent Limitation of L0 Adversarial Examples”. In: *International Symposium on Research in Attacks, Intrusions and Defenses*. 2019.

[531] Muzammal Naseer, Salman Khan, and Fatih Porikli. “Local Gradients Smoothing: Defense Against Localized Adversarial Attacks”. In: *Proc. of the IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2019.

[532] Mitali Bafna, Jack Murtagh, and Nikhil Vyas. “Thwarting Adversarial Examples: An L0-Robust Sparse Fourier Transform”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2018.

[533] Chong Xiang et al. “PatchGuard: Provable Defense against Adversarial Patches Using Masks on Small Receptive Fields”. In: *arXiv.org* abs/2005.10884 (2020).

[534] Cheng Yu et al. “Defending Against Universal Adversarial Patches by Clipping Feature Norms”. In: *Proc. of the IEEE International Conference on Computer Vision (ICCV)*. 2021.

[535] Brett Jefferson and Carlos Ortiz Marrero. “Robustness Metrics for Real-World Adversarial Examples”. In: *arXiv.org* abs/1911.10435 (2019).

[536] Anish Athalye et al. “Synthesizing Robust Adversarial Examples”. In: *Proc. of the International Conference on Machine Learning (ICML)*. 2018, pp. 284–293.

[537] Juncheng Li, Frank R. Schmidt, and J. Zico Kolter. “Adversarial camera stickers: A physical camera-based attack on deep learning systems”. In: *Proc. of the International Conference on Machine Learning (ICML)*. 2019.

[538] Kaidi Xu et al. “Adversarial T-shirt! Evading Person Detectors in A Physical World.” In: *Proc. of the European Conference on Computer Vision (ECCV)*. 2020.

[539] Weiwei Feng et al. “Meta-Attack: Class-Agnostic and Model-Agnostic Physical Adversarial Attack”. In: *Proc. of the IEEE International Conference on Computer Vision (ICCV)*. 2021.

[540] A. Braunegg et al. “APRICOT: A Dataset of Physical Adversarial Attacks on Object Detection”. In: (2020).

[541] Uttaran Sinha, Saurabh Joshi, and Vineeth N Balasubramanian. “Defending Deep Neural Networks against Structural Perturbations”. In: *Proc. of the International Conference on Machine Learning (ICML) Workshops*. 2019.

[542] Anonymous. “Defense against Spatially Transformed Adversarial Examples: An Adversarial Training Approach”. In: *Review for Proc. of the Conference on Artificial Intelligence (AAAI)*. 2019.

[543] David Stutz, Matthias Hein, and Bernt Schiele. “Disentangling Adversarial Robustness and Generalization”. In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2019).

[544] Anonymous. “Certification and Attacks for Spatial Robustness”. In: *Review for Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*. 2020.

[545] Logan Engstrom et al. “A Rotation and a Translation Suffice: Fooling CNNs with Simple Transformations”. In: *arXiv.org* abs/1712.02779 (2017).

[546] Beranger Dumont, Simona Maggio, and Pablo Montalvo. “Robustness of Rotation-Equivariant Networks to Adversarial Perturbations”. In: *arXiv.org* abs/1802.06627 (2018).

[547] Rima Alaifari, Giovanni S. Alberti, and Tandri Gauksson. “ADef: an Iterative Algorithm to Construct Adversarial Deformations”. In: *arXiv.org* abs/1804.07729 (2018).

[548] Chaowei Xiao et al. “Spatially Transformed Adversarial Examples”. In: *arXiv.org* abs/1801.02612 (2018).

[549] Can Kanbak, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. “Geometric robustness of deep networks: analysis and improvement”. In: *arXiv.org* abs/1711.09115 (2017).

[550] Hossein Hosseini and Radha Poovendran. “Semantic Adversarial Examples”. In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2018, pp. 1614–1619.

[551] Cassidy Laidlaw and Soheil Feizi. “Functional Adversarial Attacks”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2019.

[552] Zhengyu Zhao, Zhuoran Liu, and Martha A. Larson. “A Differentiable Color Filter for Generating Unrestricted Adversarial Images”. In: *arXiv.org* abs/2002.01008 (2020).

[553] Osbert Bastani, Carolyn Kim, and Hamsa Bastani. “Interpretability via Model Extraction”. In: *arXiv.org* abs/1706.09773 (2017).

[554] Zachary C Lipton. “The mythos of model interpretability”. In: *arXiv.org* abs/1606.03490 (2016).

[555] David Alvarez-Melis and Tommi S. Jaakkola. “Towards Robust Interpretability with Self-Explaining Neural Networks”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2018, pp. 7786–7795.

[556] Yinpeng Dong et al. “Towards Interpretable Deep Neural Networks by Leveraging Adversarial Examples”. In: *arXiv.org* abs/1901.09035 (2019).

[557] Been Kim et al. “Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)”. In: *Proc. of the International Conference on Machine Learning (ICML)*. 2018, pp. 2673–2682.

[558] Adam Noack et al. “Does Interpretability of Neural Networks Imply Adversarial Robustness?” In: *arXiv.org* abs/1912.03430 (2019).

[559] Daniel Smilkov et al. “SmoothGrad: removing noise by adding noise”. In: *arXiv.org* abs/1706.03825 (2017).

[560] Keisuke Kiritoshi, Ryosuke Tanno, and Tomonori Izumitani. “L1-Norm Gradient Penalty for Noise Reduction of Attribution Maps”. In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2019.

[561] Suraj Srinivas and François Fleuret. “Full-Gradient Representation for Neural Network Visualization”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2019.

[562] Chuan Guo et al. “On Calibration of Modern Neural Networks”. In: *Proc. of the International Conference on Machine Learning (ICML)*. 2017.

[563] Christos Louizos and Max Welling. “Multiplicative Normalizing Flows for Variational Bayesian Neural Networks”. In: *Proc. of the International Conference on Machine Learning (ICML)*. 2017.

[564] Zhihui Shao, Jianyi Yang, and Shaolei Ren. “Calibrating Deep Neural Network Classifiers on Out-of-Distribution Datasets”. In: *arXiv.org* abs/2006.08914 (2020).

[565] Jooyoung Moon et al. “Confidence-Aware Learning for Deep Neural Networks”. In: *arXiv.org* abs/2007.01458 (2020).

[566] Jize Zhang, Bhavya Kailkhura, and Tuan Y. Han. “Mix-n-Match: Ensemble and Compositional Methods for Uncertainty Calibration in Deep Learning”. In: *arXiv.org* abs/2003.07329 (2020).

[567] Yao Qin et al. “Improving Uncertainty Estimates through the Relationship with Adversarial Robustness”. In: *arXiv.org* abs/2006.16375 (2020).

[568] Devin Guillory et al. “Predicting with Confidence on Unseen Distributions”. In: *Proc. of the IEEE International Conference on Computer Vision (ICCV)*. 2021.

[569] Sahin Olut et al. “Adversarial Data Augmentation via Deformation Statistics”. In: 2020.

[570] Alhussein Fawzi et al. “Adaptive data augmentation for image classification”. In: *Proc. of the IEEE International Conference on Image Processing (ICIP)*. 2016.

[571] Alexander J. Ratner et al. “Learning to Compose Domain-Specific Transformations for Data Augmentation”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2017.

[572] Leon Sixt, Benjamin Wild, and Tim Landgraf. “RenderGAN: Generating Realistic Labeled Data”. In: *Frontiers in Robotics and AI* 2018 (2018).

[573] Antreas Antoniou, Amos J. Storkey, and Harrison Edwards. “Augmenting Image Classifiers Using Data Augmentation Generative Adversarial Networks”. In: *Proc. of the International Conference on Artificial Neural Networks (ICANN)*. 2018.

[574] Ekin Dogus Cubuk et al. “AutoAugment: Learning Augmentation Policies from Data”. In: *arXiv.org* abs/1805.09501 (2018).

[575] Søren Hauberg et al. “Dreaming More Data: Class-dependent Distributions over Diffeomorphisms for Learned Data Augmentation”. In: *Conference on Artificial Intelligence and Statistics (AISTATS)*. 2016.

[576] Oguz Kaan Yüksel et al. “Semantic Perturbations with Normalizing Flows for Improved Generalization”. In: *Proc. of the IEEE International Conference on Computer Vision (ICCV)*. 2021.

[577] Sergey Ioffe and Christian Szegedy. “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift”. In: *Proc. of the International Conference on Machine Learning (ICML)*. 2015.

[578] Shibani Santurkar et al. “How Does Batch Normalization Help Optimization?” In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2018.

[579] Divya Gaur, Joachim Folz, and Andreas Dengel. “Training Deep Neural Networks Without Batch Normalization”. In: *arXiv.org* abs/2008.07970 (2020).

[580] Anonymous. “Does Data Augmentation Benefit from Split BatchNorms?” In: *Review for Advances in Neural Information Processing Systems (NeurIPS)*. 2020.

[581] Xiao Zhang and David Evans. “Cost-Sensitive Robustness against Adversarial Examples”. In: *Proc. of the International Conference on Learning Representations (ICLR)*. 2019.

[582] Terrance Devries and Graham W. Taylor. “Improved Regularization of Convolutional Neural Networks with Cutout”. In: *arXiv.org* abs/1708.04552 (2017).

[583] Yaowei Zheng, Richong Zhang, and Yongyi Mao. “Regularizing Neural Networks via Adversarial Model Perturbation”. In: *arXiv.org* abs/2010.04925 (2020).

[584] Lingxi Xie et al. “DisturbLabel: Regularizing CNN on the Loss Layer”. In: *arXiv.org* abs/1605.00055 (2016).

[585] Sungrae Park et al. “Adversarial Dropout for Supervised and Semi-Supervised Learning”. In: *Proc. of the Conference on Artificial Intelligence (AAAI)*. 2018, pp. 3917–3924.

[586] Gabriel Pereyra et al. “Regularizing Neural Networks by Penalizing Confident Output Distributions”. In: *arXiv.org* abs/1701.06548 (2017).

[587] Arvind Neelakantan et al. “Adding Gradient Noise Improves Learning for Very Deep Networks”. In: *arXiv.org* abs/1511.06807 (2015).

[588] Yeming Wen et al. “Flipout: Efficient Pseudo-Independent Weight Perturbations on Mini-Batches”. In: *Proc. of the International Conference on Learning Representations (ICLR)*. 2018.

[589] Dae Hoon Park et al. “Gradient-Coherent Strong Regularization for Deep Neural Networks”. In: *arXiv.org* abs/1811.08056 (2018).

[590] Ari S. Morcos et al. “On the importance of single directions for generalization”. In: *arXiv.org* abs/1803.06959 (2018).

[591] Amir Rosenfeld and John K. Tsotsos. “Intriguing Properties of Randomly Weighted Networks: Generalizing While Learning Next to Nothing”. In: *arXiv.org* abs/1802.00844 (2018).

[592] Jonathan Frankle and Michael Carbin. “The Lottery Ticket Hypothesis: Training Pruned Neural Networks”. In: *arXiv.org* abs/1803.03635 (2018).

[593] Jonathan Frankle et al. “The Lottery Ticket Hypothesis at Scale”. In: *arXiv.org* abs/1903.01611 (2019).

[594] Luyu Wang et al. “Adversarial Robustness of Pruned Neural Networks”. In: *Proc. of the International Conference on Learning Representations (ICLR) Workshops* (2018).

[595] Justin Cosentino et al. “The Search for Sparse, Robust Neural Networks”. In: *arXiv.org* abs/1912.02386 (2019).

[596] Bai Li et al. “Towards Practical Lottery Ticket Hypothesis for Adversarial Training”. In: *arXiv.org* abs/2003.05733 (2020).

[597] Shu-Fan Wang et al. “Achieving Adversarial Robustness via Sparsity”. In: *arXiv.org* abs/2009.05423 (2020).

[598] Vaishnavh Nagarajan and J. Zico Kolter. “Generalization in Deep Networks: The Role of Distance from Initialization”. In: *arXiv.org* abs/1901.01672 (2019).

[599] Robert Geirhos et al. “ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness”. In: *arXiv.org* abs/1811.12231 (2018).

[600] Wieland Brendel and Matthias Bethge. “Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet”. In: *arXiv.org* abs/1904.00760 (2019).

[601] Etai Littwin and Lior Wolf. “Regularizing by the Variance of the Activations’ Sample-Variances”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2018, pp. 2119–2129.

[602] Hao Li et al. “Visualizing the Loss Landscape of Neural Nets”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2018.

[603] Laurent Dinh et al. “Sharp Minima Can Generalize For Deep Nets”. In: *Proc. of the International Conference on Machine Learning (ICML)*. 2017.

[604] N. Keskar et al. “On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima”. In: *arXiv.org* abs/1609.04836 (2017).

[605] Roman Novak et al. “Sensitivity and Generalization in Neural Networks: an Empirical Study”. In: *Proc. of the International Conference on Learning Representations (ICLR)*. 2018.

[606] Colin Wei, Sham M. Kakade, and Tengyu Ma. “The Implicit and Explicit Regularization Effects of Dropout”. In: *arXiv.org* abs/2002.12915 (2020).

[607] Behnam Neyshabur et al. “Exploring Generalization in Deep Learning”. In: *arXiv.org* abs/1706.08947 (2017).

[608] Lei Jimmy Ba, Ryan Kiros, and Geoffrey E. Hinton. “Layer Normalization”. In: *arXiv.org* abs/1607.06450 (2016).

[609] Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. “Instance Normalization: The Missing Ingredient for Fast Stylization”. In: *arXiv.org* abs/1607.08022 (2016).

[610] Yuxin Wu and Kaiming He. “Group Normalization”. In: *Proc. of the European Conference on Computer Vision (ECCV)*. 2018, pp. 3–19.

[611] Saurabh Singh and Shankar Krishnan. “Filter Response Normalization Layer: Eliminating Batch Dependence in the Training of Deep Neural Networks”. In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020.

[612] Bing Xu, Ruitong Huang, and Mu Li. “Revise Saturated Activation Functions”. In: *arXiv.org* abs/1602.05980 (2016).

[613] Peter L. Bartlett. “For Valid Generalization the Size of the Weights is More Important than the Size of the Network”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 1996, pp. 134–140.

[614] Davis W. Blalock et al. “What is the State of Neural Network Pruning?” In: *arXiv.org* abs/2003.03033 (2020).

[615] Zhilu Zhang and Mert R. Sabuncu. “Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels”. In: *arXiv.org* abs/1805.07836 (2018).

[616] Anonymous. “A Simple yet Effective Baseline for Robust Deep Learning with Noisy Labels”. In: *Review for Proc. of the IEEE International Conference on Computer Vision (ICCV)*. 2019.

[617] Michal Lukasik et al. “Does label smoothing mitigate label noise?” In: *arXiv.org* abs/2003.02819 (2020).

[618] Hwanjun Song et al. “Prestopping: How Does Early Stopping Help Generalization against Label Noise?” In: *arXiv.org* abs/1911.08059 (2019).

[619] Boyan Gao, Henry Gouk, and Timothy M. Hospedales. “Searching for Robustness: Loss Learning for Noisy Classification Tasks”. In: *Proc. of the IEEE International Conference on Computer Vision (ICCV)*. 2021.

[620] Shibani Santurkar, Dimitris Tsipras, and Aleksander Madry. “BREEDS: Benchmarks for Subpopulation Shift”. In: *arXiv.org* abs/2008.04859 (2020).

[621] Rohan Taori et al. “Measuring Robustness to Natural Distribution Shifts in Image Classification”. In: *arXiv.org* abs/2007.00644 (2020).

[622] Yachong Yang and A. Kuchibhotla. “Finite-sample Efficient Conformal Prediction”. In: *arXiv.org* abs/2104.13871 (2021).

[623] Anh Nguyen, Jason Yosinski, and Jeff Clune. “Deep neural networks are easily fooled: High confidence predictions for unrecognizable images”. In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015.

[624] Yali Du et al. “Towards Query Efficient Black-box Attacks: An Input-free Perspective”. In: *Proc. of the ACM Workshop on Artificial Intelligence and Security*. 2018.

[625] Vikash Sehwag et al. “Better the Devil you Know: An Analysis of Evasion Attacks using Out-of-Distribution Adversarial Examples”. In: *arXiv.org* abs/1905.01726 (2019).

[626] J. Chen et al. “Robust Out-of-distribution Detection in Neural Networks”. In: *arXiv.org* abs/2003.09711 (2020).

[627] Jie Ren et al. “Likelihood Ratios for Out-of-Distribution Detection”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2019, pp. 14680–14691.

[628] Dan Hendrycks, Mantas Mazeika, and Thomas G. Dietterich. “Deep Anomaly Detection with Outlier Exposure”. In: *arXiv.org* abs/1812.04606 (2019).

[629] Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. “Being Bayesian, Even Just a Bit, Fixes Overconfidence in ReLU Networks”. In: *arXiv.org* abs/2002.10118 (2020).

[630] Kimin Lee et al. “Training Confidence-calibrated Classifiers for Detecting Out-of-Distribution Samples”. In: *arXiv.org* abs/1711.09325 (2017).

[631] Hartmut Maennel. “Uncertainty estimates and out-of-distribution detection with Sine Networks”. In: *Proc. of the International Conference on Machine Learning (ICML) Workshops*. 2019.

[632] Rob Cornish, George Deligiannidis, and Arnaud Doucet. “Robust Predictive Uncertainty for Neural Networks via Confidence Densities”. In: *Proc. of the International Conference on Machine Learning (ICML) Workshops*. 2019.

[633] Erik Englesson and Hossein Azizpour. “Efficient Evaluation-Time Uncertainty Estimation by Improved Distillation”. In: *Proc. of the International Conference on Machine Learning (ICML) Workshops*. 2019.

[634] Mattias Teye, Hossein Azizpour, and Kevin Smith. “Bayesian Uncertainty Estimation for Batch Normalized Deep Networks”. In: *Proc. of the International Conference on Machine Learning (ICML)*. 2018, pp. 4914–4923.

[635] Nilesh A Ahuja et al. “Probabilistic modeling of deep features for out-of-distribution and adversarial detection”. In: *arXiv.org* abs/1909.11786 (2019).

[636] Dan Hendrycks and Kevin Gimpel. “A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks”. In: *Proc. of the International Conference on Learning Representations (ICLR)*. 2017.

[637] Shiyu Liang, Yixuan Li, and R. Srikant. “Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks”. In: *Proc. of the International Conference on Learning Representations (ICLR)*. 2018.

[638] Mohammadreza Salehi et al. “ARAE: Adversarially Robust Training of Autoencoders Improves Novelty Detection”. In: *arXiv.org* abs/2003.05669 (2020).

[639] Jiefeng Chen et al. “Robust Out-of-distribution Detection via Informative Outlier Mining”. In: *arXiv.org* abs/2006.15207 (2020).

[640] Zhisheng Xiao, Qun min Yan, and Yali Amit. “Likelihood Regret: An Out-of-Distribution Detection Score For Variational Auto-encoder”. In: *arXiv.org* abs/2003.02977 (2020).

[641] Byunggill Joe, Sung Ju Hwang, and Insik Shin. “Learning to Disentangle Robust and Vulnerable Features for Adversarial Detection”. In: *arXiv.org* abs/1909.04311 (2019).

[642] Francesco Crecchi et al. “FADER: Fast Adversarial Example Rejection”. In: *arXiv.org* abs/2010.09119 (2020).

[643] Joost van Amersfoort et al. “Uncertainty Estimation Using a Single Deep Deterministic Neural Network”. In: *Proc. of the International Conference on Machine Learning (ICML)*. 2020.

[644] Terrance DeVries and Graham W. Taylor. “Learning Confidence for Out-of-Distribution Detection in Neural Networks”. In: *arXiv.org* abs/1802.04865 (2018).

[645] Jim Winkens et al. “Contrastive Training for Improved Out-of-Distribution Detection”. In: *arXiv.org* abs/2007.05566 (2020).

[646] Alexander Meinke, Julian Bitterwolf, and Matthias Hein. “Provably Robust Detection of Out-of-distribution Data (almost) for free”. In: *arXiv.org* abs/2106.04260 (2021).

[647] Julian Bitterwolf, Alexander Meinke, and Matthias Hein. “Certifiably Adversarially Robust Detection of Out-of-Distribution Data”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2020.

[648] Alexander Meinke and Matthias Hein. “Towards neural networks that provably know when they don’t know”. In: *Proc. of the International Conference on Learning Representations (ICLR)*. 2020.

[649] Robin Chan, Matthias Rottmann, and Hanno Gottschalk. “Entropy Maximization and Meta Classification for Out-Of-Distribution Detection in Semantic Segmentation”. In: *Proc. of the IEEE International Conference on Computer Vision (ICCV)*. 2021.

[650] Lukas Ruff et al. “A Unifying Review of Deep and Shallow Anomaly Detection”. In: *Proceedings of the IEEE* 109.5 (2021), pp. 756–795.

[651] Guansong Pang et al. “Deep Learning for Anomaly Detection: A Review”. In: *ACM Computing Surveys* 54.2 (2021), 38:1–38:38.

[652] Marco A. F. Pimentel et al. “A review of novelty detection”. In: *Signal Processing* 99 (2014), pp. 215–249.

[653] Charu C. Aggarwal and Philip S. Yu. “Outlier Detection for High Dimensional Data”. In: *Proc. of the ACM International Conference on Management of Data (SIGMOD)*. 2001.

[654] Victoria J. Hodge and Jim Austin. “A Survey of Outlier Detection Methodologies”. In: *Artificial Intelligence Review* 22.2 (2004), pp. 85–126.

[655] Hongzhi Wang, Mohamed Jaward Bah, and Mohamed Hammad. “Progress in Outlier Detection Techniques: A Survey”. In: *IEEE Access* 7 (2019), pp. 107964–108000.

[656] Raphael Gontijo Lopes et al. “Improving Robustness Without Sacrificing Accuracy with patch Gaussian Augmentation”. In: *Proc. of the International Conference on Machine Learning (ICML) Workshops*. 2019.

[657] Norman Mu and Justing Gilmer. “MNIST-C: A Robustness benchmark for Computer Vision”. In: *Proc. of the International Conference on Machine Learning (ICML) Workshops* (2019).

[658] Dan Hendrycks and Thomas G. Dietterich. “Benchmarking Neural Network Robustness to Common Corruptions and Perturbations”. In: *arXiv.org* abs/1807.01697 (2018).

[659] Dan Hendrycks and Thomas G. Dietterich. “Benchmarking Neural Network Robustness to Common Corruptions and Perturbations”. In: *Proc. of the International Conference on Learning Representations (ICLR)*. 2019.

[660] Dan Hendrycks et al. “The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization”. In: *arXiv.org* abs/2006.16241 (2020).

[661] Steffen Schneider et al. “Improving robustness against common corruptions by covariate shift adaptation”. In: *arXiv.org* abs/2006.16971 (2020).

[662] Philipp Benz et al. “Revisiting Batch Normalization for Improving Corruption Robustness”. In: *arXiv.org* abs/2010.03630 (2020).

[663] Zachary Nado et al. “Evaluating Prediction-Time Batch Normalization for Robustness under Covariate Shift”. In: *arXiv.org* abs/2006.10963 (2020).

[664] E. Rusak et al. “A simple way to make neural networks robust against diverse image corruptions”. In: *Proc. of the European Conference on Computer Vision (ECCV)*. 2020.

[665] Klim Kireev, Maksym Andriushchenko, and Nicolas Flammarion. “On the effectiveness of adversarial training against common corruptions”. In: *arXiv.org* abs/2103.02325 (2021).

[666] Dan A. Calian et al. “Defending Against Image Corruptions Through Adversarial Augmentations”. In: *arXiv.org* abs/2104.01086 (2021).

[667] Dan Hendrycks et al. “AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty”. In: *Proc. of the International Conference on Learning Representations (ICLR)*. 2020.

[668] Xiaofeng Mao et al. “Towards Robust Vision Transformer”. In: *arXiv.org* abs/2105.07926 (2021).

[669] Tonmoy Saikia, Cordelia Schmid, and Thomas Brox. “Improving robustness against common corruptions with frequency biased models”. In: *arXiv.org* abs/2103.16241 (2021).

[670] Ilia Shumailov et al. “Sponge Examples: Energy-Latency Attacks on Neural Networks”. In: *arXiv.org* abs/2006.03463 (2020).

[671] Ilia Shumailov et al. “Sponge Examples: Energy-Latency Attacks on Neural Networks”. In: *Proc. of the IEEE European Symposium on Security and Privacy (EuroS&P)*. 2021.

[672] Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. “Fixing Asymptotic Uncertainty of Bayesian Neural Networks with Infinite ReLU Features”. In: *arXiv.org* abs/2010.02709 (2020).

[673] Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. “Learnable uncertainty under Laplace approximations”. In: *Proc. of the Conference on Uncertainty in Artificial Intelligence (UAI)*. 2021.

[674] Abhishek Murthy, Himel Das, and Md. Ariful Islam. “Robustness of Neural Networks to Parameter Quantization”. In: *arXiv.org* abs/1903.10672 (2019).

[675] César Torres-Huitzil and Bernard Girau. “Fault and Error Tolerance in Neural Networks: A Review”. In: *IEEE Access* 5 (2017).

[676] Vasisht Duddu, D. Vijay Rao, and Valentina E. Balas. “Adversarial Fault Tolerant Training for Deep Neural Networks”. In: *arXiv.org* abs/1907.03103 (2019).

[677] Ching-Tai Chiu et al. “Training Techniques to Obtain Fault-Tolerant Neural Networks”. In: *Annual International Symposium on Fault-Tolerant Computing*. 1994.

[678] Chalapathy Neti, Michael H. Schneider, and Eric D. Young. “Maximally fault tolerant neural networks”. In: *IEEE Trans. on Neural Networks (TNN)* 3.1 (1992), pp. 14–23.

[679] Dipti Deodhare, M. Vidyasagar, and S. Sathiya Keerthi. “Synthesis of fault-tolerant feedforward neural networks using minimax optimization”. In: *IEEE Trans. on Neural Networks (TNN)* 9.5 (1998), pp. 891–900.

[680] Vasisht Duddu et al. “Fault Tolerance of Neural Networks in Adversarial Settings”. In: *arXiv.org* abs/1910.13875 (2019).

[681] Manaar Alam et al. “Enhancing Fault Tolerance of Neural Networks for Security-Critical Applications”. In: *arXiv.org* abs/1902.04560 (2019).

[682] Faiz Ur Rahman, Bhavan Vasu, and Andreas E. Savakis. “Resilience and Self-Healing of Deep Convolutional Object Detectors”. In: *Proc. of the IEEE International Conference on Image Processing (ICIP)*. 2018.

[683] Vivienne Sze et al. “Efficient Processing of Deep Neural Networks: A Tutorial and Survey”. In: *Proceedings of the IEEE* 105.12 (2017).

[684] Tien-Ju Yang et al. “A method to estimate the energy consumption of deep neural networks”. In: *Asilomar Conference on Signals, Systems, and Computers (ACSSC)*. 2017.

[685] Wilfried Haensch, Tayfun Gokmen, and Ruchir Puri. “The Next Generation of Deep Learning Hardware: Analog Computing”. In: *Proceedings of the IEEE* 107.1 (2019).

[686] Jakub Breier et al. “Practical Fault Attack on Deep Neural Networks”. In: *Proc. of the ACM Conference on Computer and Communications Security (CCS)*. 2018.

[687] Kit Murdock et al. “Plundervolt: Software-based Fault Injection Attacks against Intel SGX”. In: *Proc. of the IEEE Symposium on Security and Privacy*. 2020.

[688] Adrian Tang, Simha Sethumadhavan, and Salvatore J. Stolfo. “CLKSCREW: Exposing the Perils of Security-Oblivious Energy Management”. In: *USENIX Security Symposium*. 2017.

[689] Adnan Siraj Rakin, Zhezhi He, and Deliang Fan. “Bit-Flip Attack: Crushing Neural Network with Progressive Bit Search”. In: *arXiv.org* abs/1903.12269 (2019).

[690] Adnan Siraj Rakin et al. “T-BFA: Targeted Bit-Flip Adversarial Weight Attack”. In: *arXiv.org* abs/2007.12336 (2020).

[691] Sanghyun Hong et al. “Terminal Brain Damage: Exposing the Graceless Degradation in Deep Neural Networks Under Hardware Fault Attacks”. In: *USENIX Security Symposium*. 2019.

[692] Fan Yao, Adnan Siraj Rakin, and Deliang Fan. “DeepHammer: Depleting the Intelligence of Deep Neural Networks through Targeted Chain of Bit Flips”. In: *arXiv.org* abs/2003.13746 (2020).

[693] Anonymous. “T-BFA: Targeted Bit-Flip Adversarial Weight Attack”. In: *Review for Advances in Neural Information Processing Systems (NeurIPS)*. 2020.

[694] Nandhini Chandramoorthy et al. “Resilient Low Voltage Accelerators for High Energy Efficiency”. In: *IEEE International Symposium on High Performance Computer Architecture (HPCA)*. 2019.

[695] Lita Yang and Boris Murmann. “SRAM voltage scaling for energy-efficient convolutional neural networks”. In: *International Symposium on Quality Electronic Design (ISQED)*. 2017.

[696] Skanda Koppula et al. “EDEN: Enabling Energy-Efficient, High-Performance Deep Neural Network Inference Using Approximate DRAM”. In: *Proc. of the Annual IEEE/ACM International Symposium on Microarchitecture*. 2019, pp. 166–181.

[697] Christoph Schorn, Andre Guntoro, and Gerd Ascheid. “An Efficient Bit-Flip Resilience Optimization Method for Deep Neural Networks”. In: *Proc. of the Design, Automation & Test in Europe Conference & Exhibition (DATE)*. 2019.

[698] Minghai Qin, Chao Sun, and Dejan Vucinic. “Improving Robustness of Neural Networks against Bit Flipping Errors during Inference”. In: *Journal of Image and Graphics* 6.2 (2018).

[699] Minghai Qin, Chao Sun, and Dejan Vucinic. “Robustness of Neural Networks against Storage Media Errors”. In: *arXiv.org* abs/1709.06173 (2017).

[700] Yingyan Lin, Sai Zhang, and Naresh R. Shanbhag. “A Rank Decomposed Statistical Error Compensation Technique for Robust Convolutional Neural Networks in the Near Threshold Voltage Regime”. In: *Signal Processing Systems* 90.10 (2018).

[701] Michael Klachko, Mohammad Reza Mahmoodi, and Dmitri B. Strukov. “Improving Noise Tolerance of Mixed-Signal Neural Networks”. In: *International Joint Conference on Neural Networks (IJCNN)*. 2019.

[702] Li-Huang Tsai et al. “Calibrated BatchNorm: Improving Robustness Against Noisy Weights in Neural Networks”. In: *arXiv.org* abs/2007.03230 (2020).

[703] Vinay Joshi et al. “Accurate deep neural network inference using computational phase-change memory”. In: *arXiv.org* abs/1906.03138 (2019).

[704] Jiachao Deng et al. “Retraining-based timing error mitigation for hardware neural networks”. In: *Proc. of the Design, Automation & Test in Europe Conference & Exhibition (DATE)*. 2015.

[705] Chuteng Zhou et al. “Noisy Machines: Understanding Noisy Neural Networks and Enhancing Robustness to Analog Hardware Errors Using Distillation”. In: *arXiv.org* abs/2001.04974 (2020).

[706] Guanpeng Li et al. “Understanding error propagation in deep learning neural network (DNN) accelerators and applications”. In: *Proc. of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*. 2017.

[707] Lita Yang et al. “Bit Error Tolerance of a CIFAR-10 Binarized Convolutional Neural Network Processor”. In: *IEEE International Symposium on Circuits and Systems (ISCAS)*. 2018.

[708] Daniel Bankman et al. “An Always-On 3.8 μ J/86% CIFAR-10 Mixed-Signal Binary CNN Processor With All Memory on Chip in 28-nm CMOS”. In: *IEEE Journal of Solid-State Circuits* 54.1 (2019).

[709] Sung Kim et al. “MATIC: Learning around errors for efficient low-voltage neural network accelerators”. In: *Proc. of the Design, Automation & Test in Europe Conference & Exhibition (DATE)*. 2018.

[710] Brandon Reagen et al. “Minerva: Enabling Low-Power, Highly-Accurate Deep Neural Network Accelerators”. In: *ACM/IEEE Annual International Symposium on Computer Architecture (ISCA)*. 2016.

[711] Christoph Schorn et al. “Automated design of error-resilient and hardware-efficient deep neural networks”. In: *arXiv.org* abs/1909.13844 (2019).

[712] Danilo Vasconcellos Vargas and Shashank Kotyan. “Evolving Robust Neural Architectures to Defend from Adversarial Attacks”. In: *arXiv.org* abs/1906.11667 (2019).

[713] Tien-Ju Yang et al. “NetAdapt: Platform-Aware Neural Network Adaptation for Mobile Applications”. In: *Proc. of the European Conference on Computer Vision (ECCV)*. 2018.

[714] Jian-Hao Luo, Jianxin Wu, and Weiyao Lin. “ThiNet: A Filter Level Pruning Method for Deep Neural Network Compression”. In: *Proc. of the IEEE International Conference on Computer Vision (ICCV)*. 2017.

[715] Yunhui Guo. “A Survey on Methods and Theories of Quantized Neural Networks”. In: *arXiv.org* abs/1808.04752 (2018).

[716] Andrew G. Howard et al. “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications”. In: *arXiv.org* abs/1704.04861 (2017).

[717] Xiangyu Zhang et al. “ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices”. In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018.

[718] Forrest N. Iandola et al. “SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <1MB model size”. In: *arXiv.org* abs/1602.07360 (2016).

[719] Paul Merolla et al. “Deep neural networks are robust to weight binarization and other non-linear distortions”. In: *arXiv.org* abs/1606.01981 (2016).

[720] Raghuraman Krishnamoorthi. “Quantizing deep convolutional networks for efficient inference: A whitepaper”. In: *arXiv.org* abs/1806.08342 (2018).

[721] Wonyong Sung, Sungho Shin, and Kyuyeon Hwang. “Resiliency of Deep Neural Networks under Quantization”. In: *arXiv.org* abs/1511.06488 (2015).

[722] Milad Alizadeh et al. “Gradient ℓ_1 Regularization for Quantization Robustness”. In: *Proc. of the International Conference on Learning Representations (ICLR)*. 2020.

[723] M. Shkolnik et al. “Robust Quantization: One Model to Rule Them All”. In: *arXiv.org* abs/2002.07686 (2020).

[724] Aojun Zhou et al. “Incremental Network Quantization: Towards Lossless CNNs with Low-precision Weights”. In: *Proc. of the International Conference on Learning Representations (ICLR)*. 2017.

[725] Yoojin Choi, Mostafa El-Khamy, and Jungwon Lee. “Towards the Limit of Network Quantization”. In: *Proc. of the International Conference on Learning Representations (ICLR)*. 2017.

[726] Sungho Shin, Yoonho Boo, and Wonyong Sung. “Fixed-point optimization of deep neural networks with adaptive step size retraining”. In: *Proc. of the IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017.

[727] Matthieu Courbariaux, Yoshua Bengio, and Jean-Pierre David. "BinaryConnect: Training Deep Neural Networks with binary weights during propagations". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2015.

[728] Itay Hubara et al. "Quantized Neural Networks: Training Neural Networks with Low Precision Weights and Activations". In: *Journal of Machine Learning Research (JMLR)* 18 (2017).

[729] Asit K. Mishra and Debbie Marr. "Apprentice: Using Knowledge Distillation Techniques To Improve Low-Precision Network Accuracy". In: *Proc. of the International Conference on Learning Representations (ICLR)*. 2018.

[730] Mohammad Rastegari et al. "XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks". In: *Proc. of the European Conference on Computer Vision (ECCV)*. 2016.

[731] Hao Li et al. "Training Quantized Nets: A Deeper Understanding". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Ed. by Isabelle Guyon et al. 2017.

[732] Bram-Ernst Verhoef et al. "FQ-Conv: Fully Quantized Convolution for Efficient and Accurate Inference". In: *arXiv.org* abs/1912.09356 (2019).

[733] Darryl Dexu Lin, Sachin S. Talathi, and V. Sreekanth Annapureddy. "Fixed Point Quantization of Deep Convolutional Networks". In: *Proc. of the International Conference on Machine Learning (ICML)*. 2016.

[734] Adrian Bulat and Georgios Tzimiropoulos. "Bit-Mixer: Mixed-precision networks with runtime bit-width selection". In: (2021).

[735] Steven K. Esser et al. "Learned Step Size Quantization". In: *arXiv.org* abs/1902.08153 (2019).

[736] Tiantian Han et al. "Improving Low-Precision Network Quantization via Bin Regularization". In: *Proc. of the IEEE International Conference on Computer Vision (ICCV)*. 2021.

[737] Jung Hyun Lee et al. *Cluster-Promoting Quantization with Bit-Drop for Minimizing Network Quantization Loss*. 2021.

[738] Christos Louizos, Karen Ullrich, and Max Welling. "Bayesian Compression for Deep Learning". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2017.

[739] Fangxin Liu et al. *Improving Neural Network Efficiency via Post-Training Quantization With Adaptive Floating-Point*. 2021.

[740] Ziwei Wang et al. "Generalizable Mixed-Precision Quantization via Attribution Rank Preservation". In: *Proc. of the IEEE International Conference on Computer Vision (ICCV)*. 2021.

[741] Jungwook Choi et al. "PACT: Parameterized Clipping Activation for Quantized Neural Networks". In: *arXiv.org* abs/1805.06085 (2018).

[742] Frank Seide et al. "1-bit stochastic gradient descent and its application to data-parallel distributed training of speech DNNs". In: *Annual Conference of the International Speech Communication Association*. 2014.

[743] Dan Alistarh et al. "QSGD: Randomized Quantization for Communication-Optimal Stochastic Gradient Descent". In: *arXiv.org* abs/1610.02132 (2016).

[744] Eyyub Sari and Vahid Partovi Nia. "Batch Normalization in Quantized Networks". In: *arXiv.org* abs/2004.14214 (2020).

[745] Tsui-Wei Weng et al. "Towards Certificated Model Robustness Against Weight Perturbations". In: *Proc. of the Conference on Artificial Intelligence (AAAI)*. 2020.

[746] Yu-Lin Tsai et al. *Non-singular adversarial robustness of neural networks*. Tech. rep. 2020.

[747] Yossi Adi et al. "Turning Your Weakness Into a Strength: Watermarking Deep Neural Networks by Backdooring". In: *USENIX Security Symposium*. 2018, pp. 1615–1631.

[748] Jia Guo and Miodrag Potkonjak. "Watermarking deep neural networks for embedded systems". In: *Proc. of the International Conference on Computer-Aided Design*. Ed. by Iris Bahar. ACM, 2018, p. 133.

[749] Bita Darvish Rouhani, Huili Chen, and Farinaz Koushanfar. "Deepsigns: A generic watermarking framework for ip protection of deep learning models". In: *arXiv.org* abs/1804.00750 (2018).

[750] Lixin Fan, KamWoh Ng, and Chee Seng Chan. "Rethinking Deep Neural Network Ownership Verification: Embedding Passports to Defeat Ambiguity Attacks". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Ed. by Hanna M. Wallach et al. 2019, pp. 4716–4725.

[751] Erwan Le Merrer, Patrick Pérez, and Gilles Trédan. "Adversarial frontier stitching for remote neural network watermarking". In: *Neural Computing and Applications* 32.13 (2020), pp. 9233–9244.

[752] Husrev T. Sencar and Nasir D. Memon. "Combatting Ambiguity Attacks via Selective Detection of Embedded Watermarks". In: *IEEE Transactions on Information Forensics and Security* 2.4 (2007), pp. 664–682.

[753] Yue Li, Hongxia Wang, and Mauro Barni. "A survey of deep neural network watermarking techniques". In: *arXiv.org* abs/2103.09274 (2021).

[754] Yansong Gao et al. "Backdoor Attacks and Countermeasures on Deep Learning: A Comprehensive Review". In: *arXiv.org* abs/2007.10760 (2020).

[755] Benjamin I. P. Rubinstein et al. "ANTIDOTE: understanding and defending against poisoning of anomaly detectors". In: *Proc. ACM SIGCOMM Internet Measurement Conference (IMC)*. 2009.

[756] Battista Biggio, Blaine Nelson, and Pavel Laskov. "Support Vector Machines Under Adversarial Label Noise". In: *Proc. of the Asian Conference on Machine Learning (ACML)*. 2011.

[757] Battista Biggio, Blaine Nelson, and avel Laskov. "Poisoning Attacks against Support Vector Machines". In: *Proc. of the International Conference on Machine Learning (ICML)*. 2012.

[758] Yingqi Liu et al. "Trojaning Attack on Neural Networks". In: *Annual Network and Distributed System Security Symposium*. 2018.

[759] Cong Liao et al. "Backdoor Embedding in Convolutional Neural Network Models via Invisible Perturbation". In: *arXiv.org* abs/1808.10307 (2018).

[760] Jialong Zhang et al. "Protecting Intellectual Property of Deep Neural Networks with Watermarking". In: *Proc. of the ACM on Asia Conference on Computer and Communications Security (AsiaCCS)*. 2018.

[761] Xinyun Chen et al. "Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning". In: *arXiv.org* abs/1712.05526 (2017).

[762] Ali Shafahi et al. "Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2018.

[763] Avi Schwarzschild et al. "Just How Toxic is Data Poisoning? A Unified Benchmark for Backdoor and Data Poisoning Attacks". In: *Proc. of the International Conference on Machine Learning (ICML)*. 2021.

[764] Yujie Ji et al. "Model Reuse Attacks on Deep Learning Systems". In: *Proc. of the ACM Conference on Computer and Communications Security (CCS)*. 2018.

[765] Jacob Dumford and Walter J. Scheirer. "Backdooring Convolutional Neural Networks via Targeted Weight Perturbations". In: *arXiv.org* abs/1812.03128 (2018).

[766] Siddhant Garg et al. "Can Adversarial Weight Perturbations Inject Neural Backdoors?" In: *arXiv.org* abs/2008.01761 (2020).

[767] Yusuke Uchida et al. "Embedding Watermarks into Deep Neural Networks". In: *Proc. of the ACM on International Conference on Multimedia Retrieval*. 2017, pp. 269–277.

[768] Yuki Nagai et al. "Digital watermarking for deep neural networks". In: *International Journal of Multimedia Information Retrieval* 7.1 (2018), pp. 3–16.

[769] Yu Ji et al. "Programmable Neural Network Trojan for Pre-Trained Feature Extractor". In: *arXiv.org*. Vol. abs/1901.07766. 2019.

[770] Anonymous. "Reflection Backdoor: A Natural Backdoor Attack on Deep Neural Networks". In: *Review for Proc. of the European Conference on Computer Vision (ECCV)*. 2020.

[771] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. "Label-Consistent Backdoor Attacks". In: *arXiv.org* abs/1912.02771 (2019).

[772] Yuezun Li et al. "Invisible Backdoor Attack With Sample-Specific Triggers". In: *Proc. of the IEEE International Conference on Computer Vision (ICCV)*. 2021.

[773] Bryant Chen et al. "Detecting Backdoor Attacks on Deep Neural Networks by Activation Clustering". In: *arXiv.org* abs/1811.03728 (2018).

[774] Bolun Wang et al. “Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks”. In: *Proc. of the IEEE Symposium on Security and Privacy*. 2019.

[775] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. “Fine-Pruning: Defending Against Backdooring Attacks on Deep Neural Networks”. In: *arXiv.org*. Vol. abs/1805.12185. 2018.

[776] Alvin Chan and Yew-Soon Ong. “Poison as a Cure: Detecting & Neutralizing Variable-Sized Backdoor Attacks in Deep Neural Networks”. In: *arXiv.org* abs/1911.08040 (2019).

[777] Anonymous. “Practical Detection of Trojan Neural Networks: Data-Limited and Data-Free Cases”. In: *Review for Proc. of the European Conference on Computer Vision (ECCV)*. 2020.

[778] Shanjiaoyang Huang et al. “One-pixel Signature: Characterizing CNN Models for Backdoor Detection”. In: *Proc. of the European Conference on Computer Vision (ECCV)*. 2020.

[779] Yi Zeng et al. “Rethinking the Backdoor Attacks’ Triggers: A Frequency Perspective”. In: *Proc. of the IEEE International Conference on Computer Vision (ICCV)*. 2021.

[780] Yinpeng Dong et al. “Black-box Detection of Backdoor Attacks with Limited Information and Data”. In: *Proc. of the IEEE International Conference on Computer Vision (ICCV)*. 2021.

[781] Shawn Shan et al. “Traceback of Data Poisoning Attacks in Neural Networks”. In: *arXiv.org* abs/2110.06904 (2021).

[782] Brandon Tran, Jerry Li, and Aleksander Madry. “Spectral Signatures in Backdoor Attacks”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Ed. by Samy Bengio et al. 2018.

[783] Te Juin Lester Tan and Reza Shokri. “Bypassing Backdoor Detection Algorithms in Deep Learning”. In: *Proc. of the IEEE European Symposium on Security and Privacy (EuroS&P)*. 2020.

[784] Eugene Bagdasaryan et al. “How To Backdoor Federated Learning”. In: *Conference on Artificial Intelligence and Statistics (AISTATS)*. Ed. by Silvia Chiappa and Roberto Calandra. 2020.

[785] Peter Kairouz et al. “Advances and Open Problems in Federated Learning”. In: *Foundations and Trends in Machine Learning* 14.1-2 (2021), pp. 1–210.

[786] Zhen Xiang et al. “A Backdoor Attack against 3D Point Cloud Classifiers”. In: *Proc. of the IEEE International Conference on Computer Vision (ICCV)*. 2021.

[787] Xinke Li et al. “PointBA: Towards Backdoor Attacks in 3D Point Cloud”. In: *Proc. of the IEEE International Conference on Computer Vision (ICCV)*. 2021.

[788] Ilia Shumailov et al. “Manipulating SGD with Data Ordering Attacks”. In: *arXiv.org* abs/2104.09667 (2021).

[789] Yinzhi Cao et al. “Efficient Repair of Polluted Machine Learning Systems via Causal Unlearning”. In: *Proc. of the ACM on Asia Conference on Computer and Communications Security (AsiaCCS)*. 2018, pp. 735–747.

[790] Luis Muñoz-González et al. “Towards Poisoning of Deep Learning Algorithms with Back-gradient Optimization”. In: *Proc. of the ACM Workshop on Artificial Intelligence and Security*. 2017, pp. 27–38.

[791] Huang Xiao et al. “Is Feature Selection Secure against Training Data Poisoning?” In: *Proc. of the International Conference on Machine Learning (ICML)*. 2015, pp. 1689–1698.

[792] Nathalie Baracaldo et al. “Mitigating Poisoning Attacks on Machine Learning Models: A Data Provenance Based Approach”. In: *Proc. of the ACM Workshop on Artificial Intelligence and Security*. 2017.

[793] Octavian Suciu et al. “When Does Machine Learning FAIL? Generalized Transferability for Evasion and Poisoning Attacks”. In: *USENIX Security Symposium*. 2018.

[794] Reza Shokri et al. “Membership Inference Attacks Against Machine Learning Models”. In: *Proc. of the IEEE Symposium on Security and Privacy*. 2017.

[795] Yunhui Long et al. “A Pragmatic Approach to Membership Inferences on Machine Learning Models”. In: *Proc. of the IEEE European Symposium on Security and Privacy (EuroS&P)*. 2020.

[796] Hongsheng Hu et al. “Membership Inference Attacks on Machine Learning: A Survey”. In: *arXiv.org* abs/2103.07853 (2021).

[797] Bo Liu et al. “When Machine Learning Meets Privacy: A Survey and Outlook”. In: *ACM Computing Surveys* 54.2 (2021), 31:1–31:36.

[798] Milad Nasr, Reza Shokri, and Amir Houmansadr. “Machine Learning with Membership Privacy using Adversarial Regularization”. In: *Proc. of the ACM Conference on Computer and Communications Security (CCS)*. 2018, pp. 634–646.

[799] Ahmed Salem et al. “ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models”. In: *Annual Network and Distributed System Security Symposium*. 2019.

[800] Jinyuan Jia et al. “MemGuard: Defending against Black-Box Membership Inference Attacks via Adversarial Examples”. In: *Proc. of the ACM Conference on Computer and Communications Security (CCS)*. 2019.

[801] Nicholas Carlini et al. “The Secret Sharer: Measuring Unintended Neural Network Memorization & Extracting Secrets”. In: *arXiv.org* abs/1802.08232 (2018).

[802] Congzheng Song, Thomas Ristenpart, and Vitaly Shmatikov. “Machine Learning Models that Remember Too Much”. In: *Proc. of the ACM Conference on Computer and Communications Security (CCS)*. 2017, pp. 587–601.

[803] Klas Leino and Matt Fredrikson. “Stolen Memories: Leveraging Model Memorization for Calibrated White-Box Membership Inference”. In: *USENIX Security Symposium*. 2020.

[804] Md. Atiqur Rahman et al. “Membership Inference Attack against Differentially Private Deep Learning Model”. In: *Transactions on Data Privacy* 11.1 (2018), pp. 61–79.

[805] Liwei Song, eza Shokri, and Prateek Mittal. “Membership Inference Attacks Against Adversarially Robust Deep Learning Models”. In: *Proc. of the IEEE Symposium on Security and Privacy Workshops*. 2019, pp. 50–56.

[806] Michael Veale, Reuben Binns, and Lilian Edwards. “Algorithms that Remember: Model Inversion Attacks and Data Protection Law”. In: *arXiv.org* abs/1807.04644 (2018).

[807] Avital Shafran, Shmuel Peleg, and Yedid Hoshen. “Reconstruction-Based Membership Inference Attacks are Easier on Difficult Problems”. In: (2021).

[808] Reza Shokri, Martin Strobel, and Yair Zick. “On the Privacy Risks of Model Explanations”. In: *Proc. of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*. 2021.

[809] Shahbaz Rezaei and Xin Liu. “On the Difficulty of Membership Inference Attacks”. In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021.

[810] Devansh Arpit et al. “A Closer Look at Memorization in Deep Networks”. In: *Proc. of the International Conference on Machine Learning (ICML)*. 2017.

[811] Chiyuan Zhang et al. “Understanding deep learning requires rethinking generalization”. In: *Proc. of the International Conference on Learning Representations (ICLR)*. 2017.

[812] Dingfan Chen et al. “GAN-Leaks: A Taxonomy of Membership Inference Attacks against Generative Models”. In: *Proc. of the ACM Conference on Computer and Communications Security (CCS)*. 2020.

[813] Jamie Hayes et al. “LOGAN: Membership Inference Attacks Against Generative Models”. In: *Proc. on Privacy Enhancing Technologies* 2019.1 (2019), pp. 133–152.

[814] Bingzhe Wu et al. “Generalization in Generative Adversarial Networks: A Novel Perspective from Privacy Protection”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2019.

[815] Luca Melis et al. “Exploiting Unintended Feature Leakage in Collaborative Learning”. In: *Proc. of the IEEE Symposium on Security and Privacy*. 2019.

[816] Congzheng Song and Vitaly Shmatikov. “Auditing Data Provenance in Text-Generation Models”. In: *Proc. of the ACM International Conference on Knowledge Discovery & Data Mining*. 2019.

[817] Congzheng Song and Ananth Raghunathan. “Information Leakage in Embedding Models”. In: *Proc. of the ACM Conference on Computer and Communications Security (CCS)*. Ed. by Jay Ligatti et al. 2020.

[818] Yinglin Zheng et al. “Exploring Temporal Coherence for More General Video Face Forgery Detection”. In: (2021).

[819] Davide Cozzolino et al. “ID-Reveal: Identity-aware DeepFake Video Detection”. In: *Proc. of the IEEE International Conference on Computer Vision (ICCV)*. 2021.

[820] Jing Hao et al. “TransForensics: Image Forgery Localization with Dense Self-Attention”. In: *Proc. of the IEEE International Conference on Computer Vision (ICCV)*. 2021.

[821] Xuejun Zhao et al. “Exploiting Explanations for Model Inversion Attacks”. In: *Proc. of the IEEE International Conference on Computer Vision (ICCV)*. 2021.

[822] Kathrin Grosse et al. “On the Security Relevance of Initial Weights in Deep Neural Networks”. In: *Proc. of the International Conference on Artificial Neural Networks (ICANN)*. 2020.

[823] Gamaleldin F. Elsayed, Ian J. Goodfellow, and Jascha Sohl-Dickstein. “Adversarial Reprogramming of Neural Networks”. In: *Proc. of the International Conference on Learning Representations (ICLR)*. 2019.

[824] Seong Joon Oh et al. “Towards reverse-engineering black-box neural networks”. In: *arXiv.org* abs/1711.01768 (2017).

[825] David Rolnick and Konrad Paul Kording. “Reverse-Engineering Deep ReLU Networks”. In: *arXiv.org* abs/1910.00744 (2019).

[826] Binghui Wang and Neil Zhenqiang Gong. “Stealing Hyperparameters in Machine Learning”. In: *Proc. of the IEEE Symposium on Security and Privacy*. 2018.

[827] Mengjia Yan, Christopher W. Fletcher, and Josep Torrellas. “Cache Telepathy: Leveraging Shared Resource Attacks to Learn DNN Architectures”. In: *USENIX Security Symposium*. 2020.

[828] Matthew Jagielski et al. “High Accuracy and High Fidelity Extraction of Neural Networks”. In: *USENIX Security Symposium*. 2020.

[829] Florian Tramèr et al. “Stealing Machine Learning Models via Prediction APIs”. In: *USENIX Security Symposium*. 2016.

[830] Jacson Rodrigues Correia da Silva et al. “Copycat CNN: Stealing Knowledge by Persuading Confession with Random Non-Labeled Data”. In: *International Joint Conference on Neural Networks (IJCNN)*. 2018.

[831] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. “Knockoff Nets: Stealing Functionality of Black-Box Models”. In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.

[832] Kalpesh Krishna et al. “Thieves on Sesame Street! Model Extraction of BERT-based APIs”. In: *Proc. of the International Conference on Learning Representations (ICLR)*. 2020.

[833] Taesung Lee et al. “Defending Against Model Stealing Attacks Using Deceptive Perturbations”. In: *arXiv.org* abs/1806.00054 (2018).

[834] Sanjay Kariyappa and Moinuddin K. Qureshi. “Defending Against Model Stealing Attacks With Adaptive Misinformation”. In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020.

[835] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. “Prediction Poisoning: Towards Defenses Against DNN Model Stealing Attacks”. In: *Proc. of the International Conference on Learning Representations (ICLR)*. 2020.

[836] Jean-Baptiste Truong et al. “Data-Free Model Extraction”. In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021.

[837] Sanjay Kariyappa, Atul Prakash, and Moinuddin K. Qureshi. “MAZE: Data-Free Model Stealing Attack Using Zeroth-Order Gradient Estimation”. In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021.

[838] Chuan bin Guo et al. “Certified Data Removal from Machine Learning Models”. In: *arXiv.org* abs/1911.03030 (2019).

[839] Rudy Bunel et al. “Piecewise Linear Neural Network verification: A comparative study”. In: *arXiv.org* abs/1711.00455 (2017).

[840] Sanjit A. Seshia and Dorsa Sadigh. “Towards Verified Artificial Intelligence”. In: *arXiv.org* abs/1606.08514 (2016).

[841] Xiaowei Huang et al. “Safety Verification of Deep Neural Networks”. In: *Proc. of the International Conference on Computer Aided Verification*. 2017.

[842] Oliver Zendel et al. “Analyzing Computer Vision Data - The Good, the Bad and the Ugly”. In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.

[843] Cormac Herley and Paul C. van Oorschot. “SoK: Science, Security and the Elusive Goal of Security as a Scientific Pursuit”. In: *Proc. of the IEEE Symposium on Security and Privacy*. 2017, pp. 99–120.

[844] Ram Shankar Siva Kumar et al. “Law and Adversarial Machine Learning”. In: *arXiv.org* abs/1810.10731 (2018).

[845] Martin Abadi and David G. Andersen. “Learning to Protect Communications with Adversarial Neural Cryptography”. In: *arXiv.org* abs/1610.06918 (2016).

[846] Gamaleldin F. Elsayed, Ian Goodfellow, and Jascha Sohl-Dickstein. “Adversarial Reprogramming of Neural Networks”. In: *arXiv.org* abs/1806.11146 (2018).

[847] Babajide O. Ayinde, Tamer Inanc, and Jacek M. Zurada. “On Correlation of Features Extracted by Deep Neural Networks”. In: *arXiv.org* abs/1901.10900 (2019).

[848] Ashish Vaswani et al. “Attention is All you Need”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2017.

[849] Alexey Dosovitskiy et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”. In: *Proc. of the International Conference on Learning Representations (ICLR)*. 2021.

[850] Kaiming He et al. “Masked Autoencoders Are Scalable Vision Learners”. In: *arXiv.org* abs/2111.06377 (2021).

[851] Andreas Steiner et al. “How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers”. In: *arXiv.org* abs/2106.10270 (2021).

[852] Xiangning Chen, Cho-Jui Hsieh, and Boqing Gong. “When Vision Transformers Outperform ResNets without Pretraining or Strong Data Augmentations”. In: *arXiv.org* abs/2106.01548 (2021).

[853] Hugo Touvron et al. “Training data-efficient image transformers & distillation through attention”. In: *Proc. of the International Conference on Machine Learning (ICML)*. 2021.

[854] Srinadh Bhojanapalli et al. “Understanding Robustness of Transformers for Image Classification”. In: *arXiv.org* abs/2103.14586 (2021).

[855] Sayak Paul and Pin-Yu Chen. “Vision Transformers are Robust Learners”. In: *arXiv.org* abs/2105.07581 (2021).

[856] Shiyu Tang et al. “RobustART: Benchmarking Robustness on Architecture Design and Training Techniques”. In: *arXiv.org* abs/2109.05211 (2021).

[857] Jindong Gu, Volker Tresp, and Yao Qin. “Are Vision Transformers Robust to Patch Perturbations?” In: *arXiv.org* abs/2111.10659 (2021).

[858] Ameya Joshi, Gauri Jagatap, and Chinmay Hegde. “Adversarial Token Attacks on Vision Transformers”. In: *arXiv.org* abs/2110.04337 (2021).

[859] Katelyn Morrison et al. “Exploring Corruption Robustness: Inductive Biases in Vision Transformers and MLP-Mixers”. In: *arXiv.org* abs/2106.13122 (2021).