References

- **Adversarial Examples**:
  - History:
    - Learning with Invariances (SVMs) [1];
  - Surveys [2, 3, 4, 5];
  - Best Practices [6];
  - Attacks:
    - White-Box/Gray-Box:
      - L-BFGS [7];
      - FGSM [8];
      - PGD [9];
      - CW [10] (w/ different parameterization [11]);
      - Universal [12];
      - Transferability [13];
      - Physical [14, 15];
      - Adversarial Saliency Maps [16];
      - Logits Attack (LOTS) [17];
      - Momentum Attacks [18];
      - Imperceptible Attack [19];
      - Attack on Interpretations [20];
      - GAN-based [21, 22, 23, 24, 25];
      - Confidence/Uncertainty Attack (Gaussian Processes) [26];
      - OptMargin [27];
      - ANGRI/UPSET (Predicting Adversarial Examples) [28];
      - Blind-Spot Attack [29];
      - Restricted Adversarial Example [30];
    - Black-Box:
      - ZOO [31];
      - Limited Query and Information Attack (using Gaussian Gradient Estimation) [32];
      - Boundary Attack [33];
      - One-Pixel Attack [34];
      - Ensemble-Attacks [13, 18, 35];
      - Bandit Attack [36];
      - Simple Geometric Attack [37, Sec 8.3];
      - CapsAttack [38];
      - "Cubes" Attack / Simple Attack [39, 40];
      - Low-Frequency Boundary Attack [41];
      - Sparse and Imperceivable L_0 Attack [42];
    - "On-Manifold":
      - Adversarial Spheres [43];
      - Adversarial Lighting/Physics [44];
  - Defenses;
    - Robust Nearest Neighbor [45];
    - Robust Linear Classifiers [46];

* Deep RBF Network [47, 48];
* Last RBF Layer [49, 50];
* Deep k-nearest-neighbor [51];
* Ensembles [52, 53, 54, 55, 56, 57];
* Adversarial Training (and Variants) [58, 59, 60, 61, 62, 63, 9];
  - Robust Optimization [60];
  - Distributional Robust Optimization [61];
  - Generative Adversarial Training [62];
  - Convex Outer Adversarial Polytope [64] (Extension with Cost Matrix [65]);
  - Universal Adversarial Training [66];
  - Universal Adversarial Training with Memory [67];
  - Ensemble Adversarial Training [57, 68];
  - Curriculum Adversarial Training [69];
  - Adversarial Training and Confidence-Based Rejection [70];
  - Interpolated Adversarial Training [71];
  - Smoothed Adversarial Training [72];
  - Multi-Attack Adversarial Training [73];
  - (Ensemble + Adversarial Training for Uncertainty [74]);
  - Bayesian Adversarial Training [75, 76];
* Bounded ReLU + Gaussian Data Augmentation [77];
  - Not Robust [78];
* Feature Squeezing [79];
* PCA [80];
* Saturating Networks [81];
* Distillation [82];
  - Not Robust [83, 11];
* Adversarially Robust Distillation [84];
* Iterative GAN/Deep Image Prior Projection [85, 86];
* Analysis-by-Synthesis [24, 25];
* Randomization [87]:
  - Dropout [88];
  - (Random) Image Transformations [89, 90];
  - Ensembles or Random Weight Perturbations [52];
* Gradient Regularization [91, 92, 93, 94];
* Discretization [95];
* Adaptive JPEG quantization [96];
* Rectification/Detection [97];
* Out-of-Distribution Training [98, 99];
* Logit Pairing [100];
  - Not Robust [101, 102];
* Fortified Networks [103];
* Adversarial Perturbation Eliminating GAN [104];
* Label Smoothing and Feature Squeezing [105];
  - Not Robust [102];
* Attacks meet Interpretability [106];
  - Not Robust [107];

    - Region-Based Classification / Smoothing [108, 109];
    - Compact Convolution [110];
    - MagNet (Detection + Auto-Encoding);
        - Not Robust [78];
    - Feature Denoising [111];
        - Not Robust [112];
    - Certified Robustness:
        - CROWN [113];
        - CROWN + IBP [114];
        - Interval Bound Propagation [115];
        - Spectral Features [116];
        - Abstract Interpretations [117, 118, 119];
        - Linear Regions [120, 121];
    - Parseval Networks [122];
    - Logit Inspection [123];
- Detection and Avoidance: [124, 125, 126, 127, 128, 129, 130, 131, 80, 132, 133];
    - Detection not Robust [134] – addresses [80, 127, 124, 126, 132, 133];
- Transferability:
    - Transferability [13];
    - Transferability of Evasion/Poisening [135];
    - Improving Transferability w/ Transformations/Ensemble [136];
- Attacked Defenses:
    - Attacking CVPR'18 Defenses [112];
    - Attacking ICLR'18 Defenses [137];
    - Individual Defenses [138, 78, 107, 102, 101, 83, 11];
    - Detectors [134];
- Empirical/Theoretical Analysis/Phenomena/Studies:
    - Label Leaking [139];
    - Gradient Masking [57];
    - Gradient Obfuscation [137];
    - Adversarial Examples as Test Error in Noise [140];
    - Suitability of L_p Norms [141];
    - Linear Explanation [8];
    - Upper Risk Bound [142];
    - Boundary Tilting [143];
    - Semi-Random Noise [144];
    - Manifold Explanation [143, 85, 43, 145] (also see [146]);
    - Adversarial Subspaces and Intrinsic Dimensionality [147, 129, 130];
    - Robustness and Input Dimensionality [91];
    - Robustness and Generalization [148, 149, 150, 151];
    - Adversarial Directions [152];
    - Adversarial Examples are Inevitable [153];
    - Oracle-Based Adversarial Example Definition [154];
    - Robust Features [150, 155];
    - Margin of Cross Entropy Loss [156];

    - Adversarial Examples as Input-Fault Tolerance [157];
    - Perceptual Metric PASS [158];
    - Evaluation of Regularization Methods [159];
    - Geometry of Deep Networks [160];
    - Adversarial Directions/Principal Components [161];
    - Adversarial Training:
        - Sample Complexity [37, 162];
        - Hyperparameters [163];
        - Loss Landscape [164];
        - Norm-Agnostic Robustness [165];
        - Adversarial Training and GANs [166];
- Applications:
    - Semantic Segmentation [167, 168];
    - Object Detection [14];
    - Generative Models [169, 170];
    - Reinforcement Learning [171, 172];
    - Robot Vision/iCub [173];
    - Visual Question Answering [174];
    - Face Identification [175];
    - PDF Malware Detection [176];
    - Text [177, 178];
    - Camouflage for Military Vessels [179];
    - Adversarial Meshes [180];
- Generalized Threat Models:
    - Correlated Attacks/Test Data Attack [181, 182];
- Benchmark Datasets:
    - MNIST [183] and EMNIST [184];
    - Fashion-MNIST [185];
    - Cifar10 and Cifar100 [186];
    - Restricted ImageNet [150];
    - ImageNet-143 (used in []);
    - STL-10 (used in []);
- Challenges:
    - Madry Lab[1];
    - Robust Vision Benchmark (Bethge Lab)[2] [187];
    - CAAD[3]
- Toolboxes:
    - Foolbox [187];
    - AdverTorch [188];
    - IBM Adversarial Rbustness Toolbox [189];
- Blog Articles:
    - Madry Lab Blog, https://people.csail.mit.edu/madry/lab/blog/;

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

    - Label Leaking in Adversarial Training, http://jackhaha363.github.io/blog/2017/06/19/label-leaking;
  - Misc/Uncategorized:
    - [190, 191];
- Adversarial Patches [192]:
  - Attacks:
    - (Universal) Adversarial Patch for Object Detectors with Random Location [193, 194];
    - Tracking Aversarial Patches using Expectation over Transformation [195];
    - Adversarial Framing [196];
    - LaVAN [197];
  - Defenses:
    - Digital Watermarking [198];
      - Not Robust [199];
    - Interval Bound Propagation [199];
    - Pre-Processing and Detection [200];
    - Local Gradient Smoothing [201];
      - Not Robust [199];
    - Sparse Fourier Transform [202];
- Physical Adversarial Examples:
  - Attacks:
    - Robust Adversarial Examples with Expectation over Transformation [203];
    - Adversarial Camera Stickers [204];
- Structural Perturbations:
  - Defenses:
    - Adversarial Training on Perturbed Dataset [190];
    - Approximate On-Manifold Adversarial Training [205];
  - Attacks:
    - Translation/Rotation [206, 207];
    - Adversarial Deformations [208, 209, 206];
    - Adversarial Projective Transformations (incl. Adversarial Fine-Tuning) [210];
    - Black-Box Hue and Saturation Attack [211];
- Network Calibration [212, 213];
- Learned Data Augmentation [214, 215, 216, 217, 218];
- Generalization:
  - Regularization:
    - DisturbLabel [219];
    - Adversarial Dropouts [220];
    - Virtual Adversarial Training [58];
    - Confidence Penalty/Label Smoothing [221];
    - Gradient Noise [222];
  - (Empirical) Studies/Analysis:

        * Selectivity Index/Reliance on Single Directions [223];
        * Accuracy of (Partial) Random Networks [224];
        * Iterative Pruning / Lottery Hypothesis [225, 226];
            · Robustness of Pruning [227];
        * Distance from Initialization [228];
        * PAC-Bounds for Robust Algorithms (Alternative Formulation of Robustness) [151];
        * Bias to Texture and Local Features [229, 230];
        * Variance of Activations Regularization [231];
        * Flat/Sharp Minima [232];
        * Sensitivity [233];
    - Normalization:
        * Layer Normalization [234];
        * Instance Normalization [235];
        * Group Normalization [236];
    - Theory:
        * Fat-Shattering Dimensions of Deep Networks [237];
- Noisy Labels:
    - Generalized Cross Entropy [238];
- Out-of-Distribution:
    - Unrecognizable Adversarial Examples [239];
    - GAN-based Out-of-Distribution Training [240];
    - Out-of-Distribution Training [98];
    - Perturbation-Based Detection [241];
    - Sine Networks [242];
    - Confidence Densities [243];
    - Logit Inspection [123];
    - "Improved" Distillation [244];
    - Monte Carlo Batch Normalization [245];
- Corruption Robustness:
    - Patch Gaussian Augmentation [246];
    - MNIST-C [247];
    - ImageNet-C [248];

References

[1] Choon Hui Teo et al. "Convex Learning with Invariances". In: *NeurIPS*. 2007.

[2] Marco Barreno et al. "Can machine learning be secure?" In: *AsiaCCS*. 2006.

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

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

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

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

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

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

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

[10] Nicholas Carlini and David Wagner. “Towards evaluating the robustness of neural networks”. In: *SP*. 2017.

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

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

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

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

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

[16] Nicolas Papernot et al. “The Limitations of Deep Learning in Adversarial Settings”. In: *SP*. 2016.

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

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

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

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

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

[22] Tom B. Brown et al. “Unrestricted Adversarial Examples”. In: *arXiv.org* abs/1809.08352 (2017).

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

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

[25] Lukas Schott et al. “Towards the first adversarially robust neural network model on MNIST”. In: *ICLR*. 2019.

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

[27] Warren He, Bo Li, and Dawn Song. “Decision Boundary Analysis of Adversarial Examples”. In: *ICLR*. 2018.

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

[29] Huan Zhang et al. “The Limitations of Adversarial Training and the Blind-Spot Attack”. In: *ICLR*. 2019.

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

[31] Pin-Yu Chen et al. “ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models”. In: *AISec*. 2017.

[32] Andrew Ilyas et al. “Black-box Adversarial Attacks with Limited Queries and Information”. In: *ICML*. 2018.

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

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

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

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

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

[38] Alberto Marchisio et al. “CapsAttacks: Robust and Imperceptible dversarial Attacks on Capsule Networks”. In: *ICML Workshops* (2019).

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

[40] Chuan Guo et al. “Simple Black-box Adversarial Attacks”. In: *arXiv.org* abs/1905.07121 (2019).

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

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

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

[44] Hsueh-Ti Derek Liu et al. “Beyond Pixel Norm-Balls: Parametric Adversaries using an Analytically Differentiable Renderer”. In: *ICLR*. 2019.

[45] Yizhen Wang, Somesh Jha, and Kamalika Chaudhuri. “Analyzing the Robustness of Nearest Neighbors to Adversarial Examples”. In: *ICML*. 2018.

[46] Paolo Russu et al. “Secure Kernel Machines against Evasion Attacks”. In: *AsiaCCS*. 2016, pp. 59–69.

[47] 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).

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

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

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

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

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

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

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

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

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

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

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

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

[60] 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).

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

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

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

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

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

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

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

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

[69] Qi-Zhi Cai, Chang Liu, and Dawn Song. “Curriculum Adversarial Training”. In: *IJCAI*. 2018, pp. 3740–3747.

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

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

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

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

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

[75] Nanyang Ye and Zhanxing Zhu. “Bayesian Adversarial Learning”. In: *NeurIPS*. 2018.

[76] Xuanqing Liu et al. “Adv-BNN: Improved Adversarial Defense through Robust Bayesian Neural Network”. In: *ICLR*. 2019.

[77] Valentina Zantedeschi, Maria-Irina Nicolae, and Ambrish Rawat. “Efficient defenses against adversarial attacks”. In: *AISec*. 2017.

[78] 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).

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

[80] 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).

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

[82] Nicolas Papernot et al. “Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks”. In: *SP*. 2016.

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

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

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

[86] Rama Chellappa Pouya Samangouei Maya Kabkab. “Defense-GAN: Protecting Classifiers Against Adversarial Attacks Using Generative Models”. In: *ICLR* (2018).

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

[88] Siyue Wang et al. “Defensive dropout for hardening deep neural networks under adversarial attacks”. In: *ICCAD*. 2018, 71:1–71:8.

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

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

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

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

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

[94] 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).

[95] Jacob Buckman et al. “Thermometer Encoding: One Hot Way To Resist Adversarial Examples”. In: *ICLR*. 2018.

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

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

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

[99] 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: *CVPR* (2019).

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

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

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

[103] 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).

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

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

[106] Guanhong Tao et al. “Attacks Meet Interpretability: Attribute-steered Detection of Adversarial Samples”. In: *NeurIPS*. 2018, pp. 7728–7739.

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

[108] Xiaoyu Cao and Neil Zhenqiang Gong. “Mitigating Evasion Attacks to Deep Neural Networks via Region-based Classification”. In: *ACSAC*. 2017, pp. 278–287.

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

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

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

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

[113] Huan Zhang et al. “Efficient Neural Network Robustness Certification with General Activation Functions”. In: *NeurIPS*. 2018, pp. 4944–4953.

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

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

[116] Shivam Garg et al. “A Spectral View of Adversarially Robust Features”. In: *NeurIPS*. 2018, pp. 10159–10169.

[117] Timon Gehr et al. “AI2: Safety and Robustness Certification of Neural Networks with Abstract Interpretation”. In: *SP*. 2018, pp. 3–18.

[118] Matthew Mirman, Timon Gehr, and Martin T. Vechev. “Differentiable Abstract Interpretation for Provably Robust Neural Networks”. In: *ICML*. 2018, pp. 3575–3583.

[119] Gagandeep Singh et al. “Fast and Effective Robustness Certification”. In: *NeurIPS*. 2018, pp. 10825–10836.

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

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

[122] Moustapha Cissé et al. “Parseval Networks: Improving Robustness to Adversarial Examples”. In: *ICML*. 2017.

[123] Jonathan Aigrain and Marcin Detyniecki. “Improving Robustness Without Sacrificing Accuracy with patch Gaussian Augmentation”. In: *ICML Workshops*. 2019.

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

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

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

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

[128] Fangzhou Liao et al. “Defense Against Adversarial Attacks Using High-Level Representation Guided Denoiser”. In: *CVPR*. 2018.

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

[130] Laurent Amsaleg et al. “The vulnerability of learning to adversarial perturbation increases with intrinsic dimensionality”. In: *WIFS*. 2017.

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

[132] Dan Hendrycks and Kevin Gimpel. “Early Methods for Detecting Adversarial Images”. In: *ICLR*. 2017.

[133] Xin Li and Fuxin Li. “Adversarial Examples Detection in Deep Networks with Convolutional Filter Statistics”. In: *ICCV*. 2017, pp. 5775–5783.

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

[135] 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).

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

[137] 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).

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

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

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

[141] 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).

[142] Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. "Fundamental limits on adversarial robustness". In: *ICML Workshops*. 2015.

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

[144] Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. "Robustness of classifiers: from adversarial to random noise". In: *NeurIPS*. 2016.

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

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

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

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

[149] 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).

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

[151] Huan Xu and Shie Mannor. "Robustness and Generalization". In: *COLT*. 2010, pp. 503–515.

[152] 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.

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

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

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

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

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

[158] Andras Rozsa, Ethan M. Rudd, and Terrance E. Boult. "Adversarial Diversity and Hard Positive Generation". In: *CVPR Workshops*. 2016.

[159] Sanghuyk Chun et al. "An Empirical Evaluation on Robustness and Uncertainty of Regularization Methods". In: *ICML Workshops* (2019).

[160] Alhussein Fawzi et al. "Empirical Study of the Topology and Geometry of Deep Networks". In: *CVPR*. 2018, pp. 3762–3770.

[161] Saumya Jetley, Nicholas A. Lord, and Philip H. S. Torr. "With Friends Like These, Who Needs Adversaries?" In: *NeurIPS*. 2018, pp. 10772–10782.

[162] Ludwig Schmidt et al. "Adversarially Robust Generalization Requires More Data". In: *CoRR* arXiv.org (2018).

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

[164] Joyce Xu, Dian Ang Yap, and Vinay Uday Prabhu. "Understanding Adversarial Robustness Through Loss Landscape Geometries". In: *ICML Workshops* (2019).

[165] Bai Li et al. "On Norm-Agnostic Adversarial Robustness Between Perturbation Types". In: *ICML Workshops*. 2019.

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

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

[168] Moustapha M Cisse et al. "Houdini: Fooling Deep Structured Visual and Speech Recognition Models with Adversarial Examples". In: *NeurIPS*. 2017.

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

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

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

[172] Yen-Chen Lin et al. "Tactics of Adversarial Attack on Deep Reinforcement Learning Agents". In: *IJCAI*. 2017.

[173] Marco Melis et al. "Is Deep Learning Safe for Robot Vision? Adversarial Examples Against the iCub Humanoid". In: *ICCV Workshops*. 2017.

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

[175] 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.

[176] Hung Dang, Yue Huang, and Ee-Chien Chang. "Evading Classifiers by Morphing in the Dark". In: *CCS*. 2017, pp. 119–133.

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

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

[179] 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.

[180] Chaowei Xiao et al. "MeshAdv: Adversarial Meshes for Visual Recognition". In: *CVPR*. 2019.

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

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

[183] Yann LeCun et al. "Gradient-based learning applied to document recognition". In: *Proc. of the IEEE* 86.11 (1998), pp. 2278–2324.

[184] Gregory Cohen et al. "EMNIST: an extension of MNIST to handwritten letters". In: *arXiv.org* abs/1702.05373 (2017).

[185] Han Xiao, Kashif Rasul, and Roland Vollgraf. "Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms". In: *arXiv.org* abs/1708.07747 (2017).

[186] Alex Krizhevsky. *Learning Multiple Layers of Features from Tiny Images*. Tech. rep. 2009.

[187] 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).

[188] 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).

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

[190] Uttaran Sinha, Saurabh Joshi, and Vineeth N Balasubramanian. "Defending Deep Neural Networks against Structural Perturbations". In: *ICML Workshops*. 2019.

[191] Daniel Kang et al. “Transfer of Adversarial Robustness between Perturbation Types”. In: *ICML Workshops*. 2019.

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

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

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

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

[196] Michal Zajac et al. “Adversarial Framing for Image and Video Classification”. In: *AAAI Workshops*. 2019.

[197] Danny Karmon, Daniel Zoran, and Yoav Goldberg. “LaVAN: Localized and Visible Adversarial Noise”. In: *ICML*. 2018.

[198] Jamie Hayes. “On Visible Adversarial Perturbations & Digital Watermarking”. In: *CVPR*. 2018.

[199] Anonymous. “Certified Defenses for Adversarial Patches”. In: under review. 2020. URL: https://openreview.net/forum?id=HyeaSkrYPH.

[200] Fei Zuo et al. “Exploiting the Inherent Limitation of L0 Adversarial Examples”. In: 2019.

[201] Muzammal Naseer, Salman Khan, and Fatih Porikli. “Local Gradients Smoothing: Defense Against Localized Adversarial Attacks”. In: *WACV*. 2019.

[202] Mitali Bafna, Jack Murtagh, and Nikhil Vyas. “Thwarting Adversarial Examples: An L_0-Robust Sparse Fourier Transform”. In: *NeurIPS*. 2018.

[203] Anish Athalye et al. “Synthesizing Robust Adversarial Examples”. In: *ICML*. 2018, pp. 284–293.

[204] Juncheng Li, Frank R. Schmidt, and J. Zico Kolter. “Adversarial camera stickers: A physical camera-based attack on deep learning systems”. In: *ICML*. 2019.

[205] David Stutz, Matthias Hein, and Bernt Schiele. “Disentangling Adversarial Robustness and Generalization”. In: *CVPR* (2019).

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

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

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

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

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

[211] Hossein Hosseini and Radha Poovendran. “Semantic Adversarial Examples”. In: *CVPR Workshops*. 2018, pp. 1614–1619.

[212] Chuan Guo et al. “On Calibration of Modern Neural Networks”. In: *ICML*. 2017.

[213] Christos Louizos and Max Welling. “Multiplicative Normalizing Flows for Variational Bayesian Neural Networks”. In: *ICML*. 2017.

[214] Alhussein Fawzi et al. “Adaptive data augmentation for image classification”. In: *ICIP*. 2016.

[215] Alexander J. Ratner et al. “Learning to Compose Domain-Specific Transformations for Data Augmentation”. In: *NeurIPS*. 2017.

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

[217] Antreas Antoniou, Amos J. Storkey, and Harrison Edwards. “Augmenting Image Classifiers Using Data Augmentation Generative Adversarial Networks”. In: *ICANN*. 2018.

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

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

[220] Sungrae Park et al. “Adversarial Dropout for Supervised and Semi-Supervised Learning”. In: *AAAI*. 2018, pp. 3917–3924.

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

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

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

[224] 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).

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

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

[227] Luyu Wang et al. “Adversarial Robustness of Pruned Neural Networks”. In: *ICLR Workshops* (2018).

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

[229] 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).

[230] 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).

[231] Etai Littwin and Lior Wolf. “Regularizing by the Variance of the Activations’ Sample-Variances”. In: *NeurIPS*. 2018, pp. 2119–2129.

[232] Laurent Dinh et al. “Sharp Minima Can Generalize For Deep Nets”. In: *ICML*. 2017.

[233] Roman Novak et al. “Sensitivity and Generalization in Neural Networks: an Empirical Study”. In: *ICLR*. 2018.

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

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

[236] Yuxin Wu and Kaiming He. “Group Normalization”. In: *ECCV*. 2018, pp. 3–19.

[237] Peter L. Bartlett. “For Valid Generalization the Size of the Weights is More Important than the Size of the Network”. In: *NeurIPS*. 1996, pp. 134–140.

[238] 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).

[239] Anh Nguyen, Jason Yosinski, and Jeff Clune. “Deep neural networks are easily fooled: High confidence predictions for unrecognizable images”. In: *CVPR*. 2015.

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

[241] Shiyu Liang, Yixuan Li, and R. Srikant. “Principled Detection of Out-of-Distribution Examples in Neural Networks”. In: *arXiv.org* abs/1706.02690 (2017).

[242] Hartmut Maennel. “Uncertainty estimates and out-of-distribution detection with Sine Networks”. In: *ICML Workshops*. 2019.

[243] Rob Cornish, George Deligiannidis, and Arnaud Doucet. “Robust Predictive Uncertainty for Neural Networks via Confidence Densities”. In: *ICML Workshops*. 2019.

[244] Erik Englesson and Hossein Azizpour. “Efficient Evaluation-Time Uncertainty Estimation by Improved Distillation”. In: *ICML Workshops*. 2019.

[245] Mattias Teye, Hossein Azizpour, and Kevin Smith. “Bayesian Uncertainty Estimation for Batch Normalized Deep Networks”. In: *ICML*. 2018, pp. 4914–4923.

[246] Raphael Gontijo Lopes et al. “Improving Robustness Without Sacrificing Accuracy with patch Gaussian Augmentation”. In: *ICML Workshops*. 2019.

[247] Norman Mu and Justing Gilmer. “MNIST-C: A Robustness benchmark for Computer Vision”. In: *ICML Workshops* (2019).

[248] Dan Hendrycks and Thomas G. Dietterich. “Benchmarking Neural Network Robustness to Common Corruptions and Perturbations”. In: *arXiv.org* abs/1903.12261 (2019).