Reverse Engineering of Deceptions on Machine- and Human-Centric Attacks

Other titles in Foundations and Trends® in Privacy and Security

Identifying and Mitigating the Security Risks of Generative AI
Clark Barrett *et al.*
ISBN: 978-1-63828-312-6

Decentralized Finance: Protocols, Risks, and Governance
Agostino Capponi, Garud Iyengar and Jay Sethuraman
ISBN: 978-1-63828-270-9

Proofs, Arguments, and Zero-Knowledge
Justin Thaler
ISBN: 978-1-63828-124-5

Assured Autonomy Survey
Christopher Rouff and Lanier Watkins
ISBN: 978-1-63828-038-5

Hardware Platform Security for Mobile Devices
Lachlan J. Gunn, N. Asokan, Jan-Erik Ekberg, Hans Liljestrand, Vijayanand Nayani and Thomas Nyman
ISBN: 978-1-68083-976-0

Reverse Engineering of Deceptions on Machine- and Human-Centric Attacks

Yuguang Yao
Michigan State University
yaoyugua@msu.edu

Xiao Guo
Michigan State University
guoxia11@msu.edu

Vishal Asnani
Michigan State University
asnanivi@msu.edu

Yifan Gong
Northeastern University
gong.yifa@northeastern.edu

Jiancheng Liu
Michigan State University
liujia45@msu.edu

Xue Lin
Northeastern University
xue.lin@northeastern.edu

Xiaoming Liu
Michigan State University
liuxm@msu.edu

Sijia Liu
Michigan State University
liusiji5@msu.edu

now
the essence of knowledge
Boston — Delft

Foundations and Trends® in Privacy and Security

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

Y. Yao *et al.*. *Reverse Engineering of Deceptions on Machine- and Human-Centric Attacks*. Foundations and Trends® in Privacy and Security, vol. 6, no. 2, pp. 53–152, 2024.

ISBN: 978-1-63828-341-6

Foundations and Trends® in Privacy and Security

Volume 6, Issue 2, 2024

Editorial Scope

Foundations and Trends® in Privacy and Security publishes survey and tutorial articles in the following topics:

- Access control
- Accountability
- Anonymity
- Application security
- Artifical intelligence methods in security and privacy
- Authentication
- Big data analytics and privacy
- Cloud security
- Cyber-physical systems security and privacy
- Distributed systems security and privacy
- Embedded systems security and privacy
- Forensics
- Hardware security
- Human factors in security and privacy
- Information flow
- Intrusion detection
- Malware
- Metrics
- Mobile security and privacy
- Language-based security and privacy
- Network security
- Privacy-preserving systems
- Protocol security
- Security and privacy policies
- Security architectures
- System security
- Web security and privacy

Information for Librarians

Foundations and Trends® in Privacy and Security, 2024, Volume 6, 4 issues. ISSN paper version 2474-1558. ISSN online version 2474-1566. Also available as a combined paper and online subscription.

Contents

Reverse Engineering of Deceptions on Machine- and Human-Centric Attacks

Yuguang Yao[1], Guo Xiao[1], Vishal Asnani[1], Yifan Gong[2], Jiancheng Liu[1], Xue Lin[2], Xiaoming Liu[1] and Sijia Liu[1]

[1] *Michigan State University, USA; yaoyugua@msu.edu, guoxia11@msu.edu, asnanivi@msu.edu, liujia45@msu.edu, liuxm@msu.edu, liusiji5@msu.edu*

[2] *Northeastern University, USA; gong.yifa@northeastern.edu, xue.lin@msu.edu*

ABSTRACT

This work presents a comprehensive exploration of Reverse Engineering of Deceptions (RED) in the field of adversarial machine learning. It delves into the intricacies of machine- and human-centric attacks, providing a holistic understanding of how adversarial strategies can be reverse-engineered to safeguard AI systems. For machine-centric attacks, we cover reverse engineering methods for pixel-level perturbations, adversarial saliency maps, and victim model information in adversarial examples. In the realm of human-centric attacks, the focus shifts to generative model information inference and manipulation localization from generated images. Through this work, we offer a forward-looking perspective on the challenges and opportunities associated with RED. In addition, we provide foundational and practical insights in the realms of AI security and trustworthy computer vision.

Yuguang Yao, Guo Xiao, Vishal Asnani, Yifan Gong, Jiancheng Liu, Xue Lin, Xiaoming Liu and Sijia Liu (2024), "Reverse Engineering of Deceptions on Machine- and Human-Centric Attacks", Foundations and Trends® in Privacy and Security: Vol. 6, No. 2, pp 53–152. DOI: 10.1561/3300000039.

1

Introduction

In the domain of trustworthy computer vision (CV) and adversarial machine learning (ML), the emergence of Reverse Engineering of Deceptions (**RED**) marks a pivotal evolution. This monograph is poised to grant readers a profound understanding of RED, a novel and dynamic field at the intersection of AI security and CV (DARPA, 2021). The existing body of research in the field has exhaustively explored *machine-centric* deceptions, such as adversarial attacks aimed at misleading *ML models* (Goodfellow *et al.*, 2014b; Madry *et al.*, 2017), and *human-centric* deceptions, particularly the utilization of generative models to fool *human decision-making* (Creswell *et al.*, 2018; Dhariwal and Nichol, 2021). In the above context, RED introduces an innovative adversarial learning paradigm with the ambitious goal of deciphering and cataloging the intricacies of attacks targeted at both machines and humans.

The concept of RED is not merely an academic exercise; it is a crucial response to the increasing sophistication of adversarial tactics in CV. This burgeoning field seeks to automate the process of recovering and indexing attack 'fingerprints' embedded in adversarial instances. The core question that RED endeavors to answer is: Given an attack, whether machine-centric or human-centric, can we reverse-engineer the adversary's underlying knowledge and the specifics of their attack toolchains? This question extends beyond the realm of traditional adversarial detection and defense techniques, delving into the deeper layers of adversary intentions, methodologies, and the nuances of model generation.

RED for 'machine-centric' attacks. Recent years have witnessed a rapid expansion in RED research. As for adversarial attacks designed to fool discriminative models, *i.e.*, machine-centric attacks, RED aims not only to defend against these attacks but also to infer the adversary's knowledge, including their identity, objectives, and the details of the attack perturbations. Recent works in this area, such as those by Nicholson and Emanuele (2023), Wang *et al.* (2023), Maini *et al.* (2021), Zhou and Patel (2022), Guo *et al.* (2023c), and Moayeri and Feizi (2021), have focused on reverse-engineering the type of attack generation methods and the associated hyperparameters, like perturbation radius and step number. There is also a growing interest in estimating or attributing adversarial perturbations used in constructing adversarial images (Gong *et al.*, 2022; Goebel *et al.*, 2021; Souri *et al.*, 2021; Thaker *et al.*, 2022), an endeavor closely related to adversarial purification techniques (Srinivasan *et al.*, 2021; Shi *et al.*, 2021; Yoon *et al.*, 2021; Nie *et al.*, 2022) which aim to mitigate the impact of such attacks on model predictions. We note that RED is distinct from research focused on reverse engineering model hyperparameters in a black-box setting (Oh *et al.*, 2019; Wang and Gong, 2018), which typically involves estimating model attributes from the model's prediction logits. By contrast, in the realm of RED against adversarial attacks, the victim model attribute is unknown, and the only available information is the dataset of attack instances.

RED for 'human-centric' attacks. Generative Models (GMs) nowadays generate visually compelling images. However, they also introduce the risk of *human-centric attacks*, leading to the inadvertent spread of misinformation and threats to the trustworthiness of social media. To counteract these negative impacts, two recent research directions aim to reverse engineering deception — model parsing of generative models and manipulation localization. Firstly, model parsing (Asnani *et al.*, 2023b; Guo *et al.*, 2023a) involves extracting GM hyperparameters used in creating falsified images. Unlike previous model parsing works (Tramèr *et al.*, 2016; Oh *et al.*, 2019; Hua *et al.*, 2018; Batina *et al.*, 2019), which often required additional prior knowledge to predict training information or model hyperparameters, Asnani *et al.* (2023b) employs a clustering-based approach to estimate mean and standard deviation across different GMs. In contrast, Guo *et al.* (2023a) introduces a novel framework based on Graph Convolution Networks to learn dependencies among these 37 hyperparameters. Secondly, manipulation localization is a well-established computer vision research topic that identifies tampered regions to deduce crucial information about deception. Existing work has predominantly focused on manipulation in either the image editing (Wu *et al.*, 2019; Hu *et al.*, 2020; Zhou *et al.*, 2018; Mayer and Stamm, 2018; Chen *et al.*, 2021; Wang *et al.*, 2022; Zhou *et al.*, 2020) or digital domain (Dang *et al.*, 2020; Zhao *et al.*, 2021; Huang *et al.*, 2022). In contrast, we introduce two manipulation localization algorithms (Asnani *et al.*, 2023a; Guo *et al.*, 2023b) in this work, which are capable of handling both domains simultaneously.

Objective and impact of this tutorial. We aim to present an all-encompassing exploration of RED, from its algorithmic underpinnings to its burgeoning applications, complemented by practical implementations. Delving into various formulations of RED, this monograph will unravel both the challenges and opportunities inherent in this field. The significance of RED becomes particularly salient in high-stakes applications, such as biometrics, autonomous driving, and healthcare, where the defense against and diagnosis of attacks are paramount. The implications of RED could extend beyond the boundaries of academic research, impacting the real-world deployment of machine intelligence.

Furthermore, the pressing need for security and trustworthiness in future CV models underscores the importance of our work. As the popularity of adversarial ML surges, it becomes increasingly crucial to ensure that research progress aligns with the demand for robust and reliable AI systems. By investigating how one can reverse-engineer threat models from adversarial instances, such as adversarial examples and images synthesized by generative models, our monograph offers new perspectives and insights.

Organization. The remainder of this monograph is structured as follows: Sections 2 and 3 will offer insights into the RED in machine-centric adversarial images and their potential implications for model parsing of adversarial attacks (*i.e.*, inferring details of a victim model used for attack generation). Sections 4 and 5 will delve into the RED in the human-centric attack, focusing on two research topics: model parsing of generative models and manipulation localization. Model parsing of generative models involves predicting hyperparameters used in the generative model, given the generated image. In parallel, manipulation localization predicts a segmented mask to identify the manipulated region, and this segmented mask serves to reverse engineer crucial information about the malicious manipulation method. Finally, in Section 6, we will explore the broader impact of RED on other pertinent domains and offer our concluding remarks.

References

Andriushchenko, M., F. Croce, N. Flammarion, and M. Hein. (2020). "Square attack: a query-efficient black-box adversarial attack via random search". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII*. Springer. 484–501.

Asnani, V., X. Yin, T. Hassner, S. Liu, and X. Liu. (2022). "Proactive Image Manipulation Detection". In: *CVPR*.

Asnani, V., X. Yin, T. Hassner, and X. Liu. (2023a). "Malp: Manipulation localization using a proactive scheme". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12343–12352.

Asnani, V., X. Yin, T. Hassner, and X. Liu. (2023b). "Reverse engineering of generative models: Inferring model hyperparameters from generated images". *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Athalye, A., L. Engstrom, A. Ilyas, and K. Kwok. (2018). "Synthesizing Robust Adversarial Examples". In: *International Conference on Machine Learning (ICML)*.

Batina, L., S. Bhasin, D. Jap, and S. Picek. (2019). "{CSI}{NN}: Reverse engineering of neural network architectures through electromagnetic side channel". In: *28th USENIX Security Symposium (USENIX Security 19)*. 515–532.

Bayar, B. and M. C. Stamm. (2018). "Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection". *IEEE Transactions on Information Forensics and Security*. 13(11): 2691–2706.

Boopathy, A., S. Liu, G. Zhang, P.-Y. Chen, S. Chang, and L. Daniel. (2020). "Visual Interpretability Alone Helps Adversarial Robustness".

Boroumand, M., M. Chen, and J. Fridrich. (2018). "Deep residual network for steganalysis of digital images". *IEEE Transactions on Information Forensics and Security*. 14(5): 1181–1193.

Burgess, C. P., I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner. (2017). "Understanding disentangling in β-VAE". In: *NeurIPS*.

Burt, P. J. and E. H. Adelson. (1987). "The Laplacian pyramid as a compact image code". In: *Readings in computer vision*. Elsevier. 671–679.

Carlini, N. and D. Wagner. (2017). "Towards evaluating the robustness of neural networks". In: *IEEE Symposium on Security and Privacy (S&P)*. IEEE.

Carlini, N., J. Hayes, M. Nasr, M. Jagielski, V. Sehwag, F. Tramer, B. Balle, D. Ippolito, and E. Wallace. (2023). "Extracting training data from diffusion models". In: *32nd USENIX Security Symposium (USENIX Security 23)*. 5253–5270.

Chai, L., D. Bau, S.-N. Lim, and P. Isola. (2020). "What makes fake images detectable? Understanding properties that generalize". In: *ECCV*.

Chen, D., Y. Lin, W. Li, P. Li, J. Zhou, and X. Sun. (2020). "Measuring and relieving the over-smoothing problem for graph neural networks from the topological view". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. No. 04. 3438–3445.

Chen, R. T. Q., X. Li, R. Grosse, and D. Duvenaud. (2018). "Isolating Sources of Disentanglement in Variational Autoencoders". In: *NeurIPS*.

Chen, T., Z. Zhang, Y. Zhang, S. Chang, S. Liu, and Z. Wang. (2022). "Quarantine: Sparsity can uncover the trojan attack trigger for free". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 598–609.

Chen, X., C. Dong, J. Ji, J. Cao, and X. Li. (2021). "Image manipulation detection by multi-view multi-scale supervision". In: *ICCV*.

Chen, X., C. Liu, B. Li, K. Lu, and D. Song. (2017). "Targeted backdoor attacks on deep learning systems using data poisoning". *arXiv preprint arXiv:1712.05526*.

Chen, Z.-M., X.-S. Wei, P. Wang, and Y. Guo. (2019). "Multi-label image recognition with graph convolutional networks". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5177–5186.

Choi, Y., M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. (2018). "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation". In: *CVPR*.

Cozzolino, D., J. Thies, A. Rössler, C. Riess, M. Nießner, and L. Verdoliva. (2018). "Forensictransfer: Weakly-supervised domain adaptation for forgery detection". *arXiv preprint arXiv:1812.02510*.

Creswell, A., T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath. (2018). "Generative adversarial networks: An overview". *IEEE signal processing magazine*. 35(1): 53–65.

Croce, F. and M. Hein. (2020). "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks". In: *International Conference on Machine Learning (ICML)*. PMLR.

Dang, H., F. Liu, J. Stehouwer, X. Liu, and A. K. Jain. (2020). "On the detection of digital face manipulation". In: *CVPR*.

DARPA. (2021). "Reverse Engineering of Deceptions". URL: https://www.darpa.mil/program/reverse-engineering-of-deceptions.

Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. (2009). "ImageNet: A large-scale hierarchical image database". In: *CVPR*.

Deng, L. (2012). "The MNIST database of handwritten digit images for machine learning research [best of the web]". *Signal Processing Magazine*. 29(6): 141–142.

Dhariwal, P. and A. Nichol. (2021). "Diffusion models beat gans on image synthesis". In: *NeurIPS*.

Ding, H., H. Zhang, J. Liu, J. Li, Z. Feng, and X. Jiang. (2021). "Interaction via bi-directional graph of semantic region affinity for scene parsing". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15848–15858.

Dong, C., X. Chen, R. Hu, J. Cao, and X. Li. (2022). "MVSS-Net: Multi-View Multi-Scale Supervised Networks for Image Manipulation Detection". *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Dong, J., W. Wang, and T. Tan. (2013). "Casia image tampering detection evaluation database". In: *2013 IEEE China Summit and International Conference on Signal and Information Processing*. IEEE. 422–426.

Durall, R., M. Keuper, and J. Keuper. (2020). "Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 7890–7899.

Fan, L., S. Liu, P.-Y. Chen, G. Zhang, and C. Gan. (2021). "When Does Contrastive Learning Preserve Adversarial Robustness from Pretraining to Finetuning?" *Advances in Neural Information Processing Systems*. 34.

Fan, W., Y. Ma, Q. Li, Y. He, E. Zhao, J. Tang, and D. Yin. (2019). "Graph neural networks for social recommendation". In: *The world wide web conference*. 417–426.

Frankle, J. and M. Carbin. (2018). "The lottery ticket hypothesis: Finding sparse, trainable neural networks". *arXiv preprint arXiv:1803.03635*.

Fredrikson, M., S. Jha, and T. Ristenpart. (2015). "Model inversion attacks that exploit confidence information and basic countermeasures". In: *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*. 1322–1333.

Fu, W., H. Wang, C. Gao, G. Liu, Y. Li, and T. Jiang. (2023). "Practical Membership Inference Attacks against Fine-tuned Large Language Models via Self-prompt Calibration". *arXiv preprint arXiv:2311.06062*.

Goebel, M., J. Bunk, S. Chattopadhyay, L. Nataraj, S. Chandrasekaran, and B. Manjunath. (2021). "Attribution of gradient based adversarial attacks for reverse engineering of deceptions". *arXiv preprint arXiv:2103.11002*.

Gong, Y., Y. Yao, Y. Li, Y. Zhang, X. Liu, X. Lin, and S. Liu. (2022). "Reverse engineering of imperceptible adversarial image perturbations". *arXiv preprint arXiv:2203.14145*.

Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. (2014a). "Generative adversarial nets". In: *NeurIPS*.

Goodfellow, I. J., J. Shlens, and C. Szegedy. (2014b). "Explaining and harnessing adversarial examples". *arXiv preprint arXiv:1412.6572*.

Gragnaniello, D., D. Cozzolino, F. Marra, G. Poggi, and L. Verdoliva. (2021). "Are GAN generated images easy to detect? A critical analysis of the state-of-the-art". In: *2021 IEEE international conference on multimedia and expo (ICME)*. IEEE. 1–6.

Gu, T., B. Dolan-Gavitt, and S. Garg. (2017). "Badnets: Identifying vulnerabilities in the machine learning model supply chain". *arXiv preprint arXiv:1708.06733*.

Guarnera, L., O. Giudice, and S. Battiato. (2020). "DeepFake Detection by Analyzing Convolutional Traces". In: *CVPR Workshops*.

Guo, X., V. Asnani, S. Liu, and X. Liu. (2023a). "Tracing Hyperparameter Dependencies for Model Parsing via Learnable Graph Pooling Network". *arXiv preprint arXiv:2312.02224*.

Guo, X., X. Liu, Z. Ren, S. Grosz, I. Masi, and X. Liu. (2023b). "Hierarchical Fine-Grained Image Forgery Detection and Localization". In: *In Proceeding of IEEE Computer Vision and Pattern Recognition*.

Guo, X., Y. Liu, A. Jain, and X. Liu. (2022). "Multi-domain Learning for Updating Face Anti-spoofing Models". In: *ECCV*.

Guo, Z., Y. Zhang, and W. Lu. (2019). "Attention Guided Graph Convolutional Networks for Relation Extraction". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 241–251.

Guo, Z., K. Han, Y. Ge, W. Ji, and Y. Li. (2023c). "Scalable Attribution of Adversarial Attacks via Multi-Task Learning". *arXiv preprint arXiv:2302.14059*.

Han, S., J. Pool, J. Tran, and W. Dally. (2015). "Learning both weights and connections for efficient neural network". *Advances in neural information processing systems*. 28.

He, K., X. Zhang, S. Ren, and J. Sun. (2016). "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Heath, V. (2019). "From a Sleazy Reddit Post to a National Security Threat: A Closer Look at the Deepfake Discourse". In: *Disinformation and Digital Democracies in the 21st Century*. The NATO Association of Canada.

Ho, J., A. Jain, and P. Abbeel. (2020). "Denoising diffusion probabilistic models". *Advances in Neural Information Processing Systems*. 33: 6840–6851.

Hu, H. and J. Pang. (2023). "Membership inference of diffusion models". *arXiv preprint arXiv:2301.09956*.

Hu, X., Z. Zhang, Z. Jiang, S. Chaudhuri, Z. Yang, and R. Nevatia. (2020). "SPAN: spatial pyramid attention network for image manipulation localization". In: *European Conference on Computer Vision*. Springer. 312–328.

Hua, W., Z. Zhang, and G. E. Suh. (2018). "Reverse engineering convolutional neural networks through side-channel information leaks". In: *Proceedings of the 55th Annual Design Automation Conference*. 1–6.

Huang, Y., F. Juefei-Xu, Q. Guo, Y. Liu, and G. Pu. (2022). "FakeLocator: Robust localization of GAN-based face manipulations". *IEEE Transactions on Information Forensics and Security*. 17: 2657–2672.

Ilyas, A., L. Engstrom, A. Athalye, and J. Lin. (2018). "Black-box Adversarial Attacks with Limited Queries and Information". *arXiv preprint arXiv:1804.08598*.

Jabbar, A., X. Li, and B. Omar. (2020). "A Survey on Generative Adversarial Networks: Variants, Applications, and Training". *arXiv preprint arXiv:2006.05132*.

Ji, K., F. Chen, X. Guo, Y. Xu, J. Wang, and J. Chen. (2023). "Uncertainty-guided Learning for Improving Image Manipulation Detection". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 22456–22465.

Jourabloo, A., Y. Liu, and X. Liu. (2018). "Face de-spoofing: Anti-spoofing via noise modeling". In: *ECCV*.

Karras, T., T. Aila, S. Laine, and J. Lehtinen. (2018). "Progressive growing of GANs for improved quality, stability, and variation". In: *ICLR*.

Karras, T., S. Laine, and T. Aila. (2019). "A style-based generator architecture for generative adversarial networks". In: *CVPR*. 4401–4410.

Kingma, D. P. and M. Welling. (2014). "Auto-Encoding Variational Bayes". In: *ICLR*.

Kingma, D. P. and J. Ba. (2015). "Adam: A Method for Stochastic Optimization". In: *International Conference on Learning Representations (ICLR)*.

Kipf, T. N. and M. Welling. (2016). "Semi-supervised classification with graph convolutional networks". *arXiv preprint arXiv:1609.02907*.

Krizhevsky, A., G. Hinton, *et al.* (2009). "Learning multiple layers of features from tiny images".

LeCun, Y., Y. Bengio, and G. Hinton. (2015). "Deep learning". *nature*. 521(7553): 436–444.

Li, G., M. Muller, A. Thabet, and B. Ghanem. (2019). "Deepgcns: Can gcns go as deep as cnns?" In: *Proceedings of the IEEE/CVF international conference on computer vision*. 9267–9276.

Li, L., J. Bao, H. Yang, D. Chen, and F. Wen. (2020a). "Faceshifter: Towards high fidelity and occlusion aware face swapping". *CVPR*.

Li, Q., Y. Guo, and H. Chen. (2020b). "Practical no-box adversarial attacks against DNNs". *Advances in Neural Information Processing Systems (NeurIPS)*.

Liao, F., M. Liang, Y. Dong, T. Pang, X. Hu, and J. Zhu. (2018). "Defense against Adversarial Attacks Using High-Level Representation Guided Denoiser". *arXiv:1712.02976 [cs]*. May. (Accessed on 05/26/2021).

Liu, C., B. Zoph, M. Neumann, J. Shlens, W. Hua, L.-J. Li, L. Fei-Fei, A. Yuille, J. Huang, and K. Murphy. (2018). "Progressive neural architecture search". In: *ECCV*.

Liu, M., Y. Ding, M. Xia, X. Liu, E. Ding, W. Zuo, and S. Wen. (2019a). "Stgan: A unified selective transfer network for arbitrary image attribute editing". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3673–3682.

Liu, S., P.-Y. Chen, X. Chen, and M. Hong. (2019b). "signSGD via Zeroth-Order Oracle". In: *International Conference on Learning Representations*.

Liu, S., P.-Y. Chen, B. Kailkhura, G. Zhang, A. O. Hero III, and P. K. Varshney. (2020). "A primer on zeroth-order optimization in signal processing and machine learning: Principals, recent advances, and applications". *IEEE Signal Processing Magazine*. 37(5): 43–54.

Liu, X., Y. Liu, J. Chen, and X. Liu. (2022). "PSCC-Net: Progressive spatio-channel correlation network for image manipulation detection and localization". *IEEE Transactions on Circuits and Systems for Video Technology*.

Liu, Z., P. Luo, X. Wang, and X. Tang. (2015). "Deep learning face attributes in the wild". In: *Proceedings of the IEEE international conference on computer vision*. 3730–3738.

Luo, Y., X. Boix, G. Roig, T. Poggio, and Q. Zhao. (2015). "Foveation-based mechanisms alleviate adversarial examples". *arXiv preprint arXiv:1511.06292*.

Madry, A., A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. (2017). "Towards deep learning models resistant to adversarial attacks". *arXiv preprint arXiv:1706.06083*.

Maini, P., X. Chen, B. Li, and D. Song. (2021). "Perturbation Type Categorization for Multiple ℓ_p Bounded Adversarial Robustness". URL: https://openreview.net/forum?id=Oe2XI-Aft-k.

Marra, F., D. Gragnaniello, D. Cozzolino, and L. Verdoliva. (2018). "Detection of gan-generated fake images over social networks". In: *2018 IEEE conference on multimedia information processing and retrieval (MIPR)*. IEEE. 384–389.

Marra, F., D. Gragnaniello, L. Verdoliva, and G. Poggi. (2019a). "Do gans leave artificial fingerprints?" In: *IEEE conference on multimedia information processing and retrieval (MIPR)*.

Marra, F., C. Saltori, G. Boato, and L. Verdoliva. (2019b). "Incremental learning for the detection and classification of GAN-generated images". In: *WIFS*.

Masi, I., A. Killekar, R. M. Mascarenhas, S. P. Gurudatt, and W. AbdAlmageed. (2020). "Two-branch recurrent network for isolating deepfakes in videos". In: *ECCV*. Springer.

Mayer, O. and M. C. Stamm. (2018). "Learned forensic source similarity for unknown camera models". In: *ICASSP*.

McCloskey, S. and M. Albright. (2019). "Detecting GAN-generated imagery using saturation cues". In: *ICIP*.

Min, Y., F. Wenkel, and G. Wolf. (2020). "Scattering gcn: Overcoming oversmoothness in graph convolutional networks". *Advances in neural information processing systems*. 33: 14498–14508.

Moayeri, M. and S. Feizi. (2021). "Sample efficient detection and classification of adversarial attacks via self-supervised embeddings". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 7677–7686.

Nasr, M., N. Carlini, J. Hayase, M. Jagielski, A. F. Cooper, D. Ippolito, C. A. Choquette-Choo, E. Wallace, F. Tramèr, and K. Lee. (2023). "Scalable Extraction of Training Data from (Production) Language Models". *arXiv preprint arXiv:2311.17035*.

Ng, T.-T., J. Hsu, and S.-F. Chang. (2009). "Columbia image splicing detection evaluation dataset". *DVMM lab. Columbia Univ CalPhotos Digit Libr*.

Nguyen, B. X., B. D. Nguyen, T. Do, E. Tjiputra, Q. D. Tran, and A. Nguyen. (2021). "Graph-based person signature for person reidentifications". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3492–3501.

Nicholson, D. A. and V. Emanuele. (2023). "Reverse engineering adversarial attacks with fingerprints from adversarial examples". *arXiv preprint arXiv:2301.13869*.

Nie, W., B. Guo, Y. Huang, C. Xiao, A. Vahdat, and A. Anandkumar. (2022). "Diffusion models for adversarial purification". *arXiv preprint arXiv:2205.07460*.

Nirkin, Y., L. Wolf, Y. Keller, and T. Hassner. (2020). "DeepFake detection based on the discrepancy between the face and its context". *arXiv preprint arXiv:2008.12262*.

NIST: Nist nimble 2016 datasets. (2016). URL: https://www.nist.gov/itl/iad/mig/,.

Niu, Z., Z. Chen, L. Li, Y. Yang, B. Li, and J. Yi. (2020). "On the Limitations of Denoising Strategies as Adversarial Defenses". *arXiv:2012.09384 [cs]*.

Novozamsky, A., B. Mahdian, and S. Saic. (2020). "IMD2020: A Large-Scale Annotated Dataset Tailored for Detecting Manipulated Images". In: *2020 IEEE Winter Applications of Computer Vision Workshops (WACVW)*. 71–80.

Oh, S. J., B. Schiele, and M. Fritz. (2019). "Towards reverse-engineering black-box neural networks". *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*: 121–144.

Ojha, U., Y. Li, and Y. J. Lee. (2023). "Towards universal fake image detectors that generalize across generative models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 24480–24489.

Pang, R., X. Zhang, S. Ji, X. Luo, and T. Wang. (2020). "AdvMind: Inferring Adversary Intent of Black-Box Attacks". In: *the International Conference on Knowledge Discovery & Data Mining (KDD)*.

Pérez, P., M. Gangnet, and A. Blake. (2003). "Poisson image editing". In: *ACM SIGGRAPH 2003 Papers*. 313–318.

Pham, H., M. Guan, B. Zoph, Q. Le, and J. Dean. (2018). "Efficient neural architecture search via parameters sharing". In: *ICML*.

Rossler, A., D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. (2019). "Faceforensics++: Learning to detect manipulated facial images". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 1–11.

Ruff, L., R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft. (2018). "Deep one-class classification". In: *ICML*. 4393–4402.

Ruiz, N., S. A. Bargal, and S. Sclaroff. (2020). "Disrupting deepfakes: Adversarial attacks against conditional image translation networks and facial manipulation systems". In: *ECCV*.

Sabour, S., Y. Cao, F. Faghri, and D. J. Fleet. (2015). "Adversarial manipulation of deep representations". *arXiv preprint arXiv:1511.05122*.

Salman, H., M. Sun, G. Yang, A. Kapoor, and J. Z. Kolter. (2020). "Denoised smoothing: A provable defense for pretrained classifiers". In: *Advances in Neural Information Processing Systems (NeurIPS)*.

Schwarz, K., Y. Liao, and A. Geiger. (2021). "On the frequency bias of generative models". *Advances in Neural Information Processing Systems*. 34: 18126–18136.

Segalis, E. and E. Galili. (2020). "OGAN: Disrupting Deepfakes with an Adversarial Attack that Survives Training". *arXiv preprint arXiv:2006.12247*.

Selvaraju, R. R., M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. (2020). "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization". *International Journal of Computer Vision*.

Shafahi, A., W. R. Huang, C. Studer, S. Feizi, and T. Goldstein. (2020). "Are adversarial examples inevitable?" *arXiv:1809.02104 [cs, stat]*. Feb. (Accessed on 05/26/2021).

Shi, C., C. Holtz, and G. Mishne. (2021). "Online adversarial purification based on self-supervision". *arXiv preprint arXiv:2101.09387*.

Shokri, R., M. Stronati, C. Song, and V. Shmatikov. (2017). "Membership inference attacks against machine learning models". In: *2017 IEEE symposium on security and privacy (SP)*. IEEE. 3–18.

Simonyan, K. and A. Zisserman. (2015). "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *International Conference on Learning Representations (ICLR)*.

Souri, H., P. Khorramshahi, C. P. Lau, M. Goldblum, and R. Chellappa. (2021). "Identification of Attack-Specific Signatures in Adversarial Examples". *arXiv preprint arXiv:2110.06802*.

Srinivasan, V., C. Rohrer, A. Marban, K.-R. Müller, W. Samek, and S. Nakajima. (2021). "Robustifying models against adversarial attacks by langevin dynamics". *Neural Networks*. 137: 1–17.

Sun, Z., H. Jiang, D. Wang, X. Li, and J. Cao. (2023). "SAFL-Net: Semantic-Agnostic Feature Learning Network with Auxiliary Plugins for Image Manipulation Detection". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 22424–22433.

Szegedy, C., V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. (2015). "Rethinking the Inception Architecture for Computer Vision". *CoRR*.

Tan, M., B. Chen, R. Pang, V. Vasudevan, M. Sandler, A. Howard, and Q. V. Le. (2019). "MnasNet: Platform-aware neural architecture search for mobile". In: *CVPR*.

Thaker, D., P. Giampouras, and R. Vidal. (2022). "Reverse Engineering ℓ_p attacks: A block-sparse optimization approach with recovery guarantees". In: *International Conference on Machine Learning*. PMLR. 21253–21271.

Tirupattur, P., K. Duarte, Y. S. Rawat, and M. Shah. (2021). "Modeling multi-label action dependencies for temporal action localization". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1460–1470.

Tramer, F., N. Carlini, W. Brendel, and A. Madry. (2020). "On adaptive attacks to adversarial example defenses". *arXiv preprint arXiv:2002.08347*.

Tramèr, F., F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart. (2016). "Stealing machine learning models via prediction {APIs}". In: *25th USENIX security symposium (USENIX Security 16)*. 601–618.

Trinh, L., M. Tsang, S. Rambhatla, and Y. Liu. (2021). "Interpretable and Trustworthy Deepfake Detection via Dynamic Prototypes". In: *WACV*. 1973–1983.

Veličković, P., G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio. (2017). "Graph attention networks". *arXiv preprint arXiv:1710.10903*.

Vinyals, O., C. Blundell, T. Lillicrap, D. Wierstra, *et al.* (2016). "Matching networks for one shot learning". *Advances in neural information processing systems*. 29.

Waldemarsson, C. (2020). *Disinformation, Deepfakes & Democracy; The European response to election interference in the digital age*. The Alliance of Democracies Foundation.

Wang, B. and N. Z. Gong. (2018). "Stealing hyperparameters in machine learning". In: *2018 IEEE symposium on security and privacy (SP)*. IEEE. 36–52.

Wang, B., Y. Yao, S. Shan, H. Li, B. Viswanath, H. Zheng, and B. Y. Zhao. (2019). "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks". In: *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE. 707–723.

Wang, J., Z. Wu, J. Chen, X. Han, A. Shrivastava, S.-N. Lim, and Y.-G. Jiang. (2022). "Objectformer for image manipulation detection and localization". In: *CVPR*. 2364–2373.

Wang, R., G. Zhang, S. Liu, P.-Y. Chen, J. Xiong, and M. Wang. (2020a). "Practical detection of trojan neural networks: Data-limited and data-free cases". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*. Springer. 222–238.

Wang, R., F. Juefei-Xu, M. Luo, Y. Liu, and L. Wang. (2021a). "Fake-Tagger: Robust Safeguards against DeepFake Dissemination via Provenance Tracking". In: *ACMM*.

Wang, S.-Y., O. Wang, R. Zhang, A. Owens, and A. A. Efros. (2020b). "CNN-generated images are surprisingly easy to spot... for now". In: *CVPR*. 8695–8704.

Wang, X., R. Girshick, A. Gupta, and K. He. (2018). "Non-local neural networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7794–7803.

Wang, X., Y. Li, C.-J. Hsieh, and T. C. M. Lee. (2023). "CAN MACHINE TELL THE DISTORTION DIFFERENCE? A REVERSE ENGINEERING STUDY OF ADVERSARIAL ATTACKS". URL: https://openreview.net/forum?id=NdFKHCFxXjS.

Wang, Z., Q. She, and T. E. Ward. (2021b). "Generative Adversarial Networks in Computer Vision: A Survey and Taxonomy". *ACM Computing Surveys*. 54(2).

Wen, B., Y. Zhu, R. Subramanian, T.-T. Ng, X. Shen, and S. Winkler. (2016). "COVERAGE—A novel database for copy-move forgery detection". In: *2016 IEEE international conference on image processing (ICIP)*. IEEE. 161–165.

Wong, E., L. Rice, and J. Z. Kolter. (2020). "Fast is better than free: Revisiting adversarial training". In: *International Conference on Learning Representations (ICLR)*.

Wu, Y., W. AbdAlmageed, and P. Natarajan. (2019). "Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features". In: *CVPR*. 9543–9552.

Xie, C., Z. Zhang, Y. Zhou, S. Bai, J. Wang, Z. Ren, and A. L. Yuille. (2019). "Improving transferability of adversarial examples with input diversity". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xu, K., S. Liu, P. Zhao, P.-Y. Chen, H. Zhang, Q. Fan, D. Erdogmus, Y. Wang, and X. Lin. (2019). "Structured Adversarial Attack: Towards General Implementation and Better Interpretability". In: *International Conference on Learning Representations (ICLR)*.

Ye, J., J. He, X. Peng, W. Wu, and Y. Qiao. (2020). "Attention-driven dynamic graph convolutional network for multi-label image recognition". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*. Springer. 649–665.

Yeh, C.-Y., H.-W. Chen, S.-L. Tsai, and S.-D. Wang. (2020). "Disrupting image-translation-based deepfake algorithms with adversarial attacks". In: *WACVW*.

Yoon, J., S. J. Hwang, and J. Lee. (2021). "Adversarial purification with score-based generative models". In: *International Conference on Machine Learning*. PMLR. 12062–12072.

Yu, N., L. S. Davis, and M. Fritz. (2019). "Attributing fake images to GANs: Learning and analyzing GAN fingerprints". In: *ICCV*. 7556–7566.

Yun, S., D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo. (2019). "Cutmix: Regularization strategy to train strong classifiers with localizable features". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6023–6032.

Zhai, Y., T. Luan, D. Doermann, and J. Yuan. (2023). "Towards Generic Image Manipulation Detection with Weakly-Supervised Self-Consistency Learning". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 22390–22400.

Zhang, H., I. Goodfellow, D. Metaxas, and A. Odena. (2019a). "Self-attention generative adversarial networks". In: *International conference on machine learning*. PMLR. 7354–7363.

Zhang, K., W. Zuo, Y. Chen, D. Meng, and L. Zhang. (2017). "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising". *IEEE transactions on image processing*. 26(7): 3142–3155.

Zhang, X., S. Karaman, and S.-F. Chang. (2019b). "Detecting and simulating artifacts in gan fake images". In: *2019 IEEE international workshop on information forensics and security (WIFS)*. IEEE. 1–6.

Zhao, T., X. Xu, M. Xu, H. Ding, Y. Xiong, and W. Xia. (2021). "Learning self-consistency for deepfake detection". In: *CVPR*.

Zhou, J., X. Ma, X. Du, A. Y. Alhammadi, and W. Feng. (2023). "Pre-training-free Image Manipulation Localization through Non-Mutually Exclusive Contrastive Learning". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 22346–22356.

Zhou, M. and V. M. Patel. (2022). "On Trace of PGD-Like Adversarial Attacks". *arXiv preprint arXiv:2205.09586*.

Zhou, P., B.-C. Chen, X. Han, M. Najibi, A. Shrivastava, S.-N. Lim, and L. Davis. (2020). "Generate, segment, and refine: Towards generic manipulation segmentation". In: *AAAI*.

Zhou, P., X. Han, V. I. Morariu, and L. S. Davis. (2017). "Two-stream neural networks for tampered face detection". In: *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)*. IEEE. 1831–1839.

Zhou, P., X. Han, V. I. Morariu, and L. S. Davis. (2018). "Learning rich features for image manipulation detection". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1053–1061.

Zhu, J.-Y., T. Park, P. Isola, and A. A. Efros. (2017). "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks". In: *ICCV*.