Trustworthy Machine Learning: From Data to Models

Other titles in Foundations and Trends® in Privacy and Security

Advances in Secure IoT Data Sharing
Phu Nguyen, Arda Goknil, Gencer Erdogan, Shukun Tokas, Nicolas Ferry and Thanh Thao Thi Tran
ISBN: 978-1-63828-422-2

Navigating the Soundscape of Deception: A Comprehensive Survey on Audio Deepfake Generation, Detection, and Future Horizons
Taiba Majid Wani, Syed Asif Ahmad Qadri, Farooq Ahmad Wani and Irene Amerini
ISBN: 978-1-63828-492-5

Reverse Engineering of Deceptions on Machine- and Human-Centric Attacks
Yuguang Yao, Xiao Guo, Vishal Asnani, Yifan Gong, Jiancheng Liu, Xue Lin, Xiaoming Liu and Sijia Liu
ISBN: 978-1-63828-340-9

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

Cybersecurity for Modern Smart Grid Against Emerging Threats
Daisuke Mashima, Yao Chen, Muhammad M. Roomi, Subhash Lakshminarayana and Deming Chen
ISBN: 978-1-63828-294-5

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

Trustworthy Machine Learning: From Data to Models

Bo Han
Hong Kong Baptist University
bhanml@comp.hkbu.edu.hk

Jiangchao Yao
Shanghai Jiao Tong University
sunarker@sjtu.edu.cn

Tongliang Liu
The University of Sydney
tongliang.liu@sydney.edu.au

Bo Li
University of Illinois Urbana-Champaign
lbo@illinois.edu

Sanmi Koyejo
Stanford University
sanmi@cs.stanford.edu

Feng Liu
The University of Melbourne
feng.liu1@unimelb.edu.au

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

B. Han *et al.*. *Trustworthy Machine Learning: From Data to Models*. Foundations and Trends® in Privacy and Security, vol. 7, no. 2-3, pp. 74–246, 2025.

ISBN: 978-1-63828-549-6

Foundations and Trends® in Privacy and Security
Volume 7, Issue 2-3, 2025

Editorial Board

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, 2025, Volume 7, 4 issues. ISSN paper version 2474-1558. ISSN online version 2474-1566. Also available as a combined paper and online subscription.

Contents

Trustworthy Machine Learning: From Data to Models

Bo Han[1], Jiangchao Yao[2], Tongliang Liu[3], Bo Li[4], Sanmi Koyejo[5] and Feng Liu[6]

[1] *Hong Kong Baptist University, Hong Kong and RIKEN, Japan; bhanml@comp.hkbu.edu.hk*
[2] *Shanghai Jiao Tong University, China; sunarker@sjtu.edu.cn*
[3] *The University of Sydney, Australia, MBZUAI, UAE and RIKEN, Japan; tongliang.liu@sydney.edu.au*
[4] *University of Illinois Urbana-Champaign, USA; lbo@illinois.edu*
[5] *Stanford University, USA; sanmi@cs.stanford.edu*
[6] *The University of Melbourne, Australia and RIKEN, Japan; feng.liu1@unimelb.edu.au*

ABSTRACT

The success of machine learning algorithms relies not only on achieving good performance but also on ensuring trustworthiness across diverse applications and scenarios. Trustworthy machine learning seeks to handle critical problems in addressing the issues of robustness, privacy, security, reliability, and other desirable properties. The broad research area has achieved remarkable advancement and brings various emerging topics along with the progress. We present this survey to provide a systematic overview of the research problems under trustworthy machine learning covering the perspectives from data to model. Starting with fundamental data-centric learning, the survey reviews learning with noisy data, long-tailed distribution, out-of-distribution data,

Bo Han, Jiangchao Yao, Tongliang Liu, Bo Li, Sanmi Koyejo and Feng Liu (2025), "Trustworthy Machine Learning: From Data to Models", Foundations and Trends® in Privacy and Security: Vol. 7, No. 2-3, pp 74–246. DOI: 10.1561/3300000043.

and adversarial examples to achieve robustness. Delving into private and secured learning, the survey elaborates on core methodologies differential privacy, different attacking threats, and learning paradigms, to realize privacy protection and enhance security. Finally, it introduces several trendy issues related to the foundation models, including jailbreak prompts, watermarking, and hallucination, as well as causal learning and reasoning. The survey integrates commonly isolated research problems in a unified manner, which provides general problem setups, detailed sub-directions, and further discussion on its challenges or future developments. We hope the comprehensive investigation presented in this survey can serve as a clear introduction for the problem evolution from data to models and also bring new insight for developing trustworthy machine learning.

1

Introduction

As artificial intelligence (AI) and machine learning (ML) experience advancements rapidly, remarkable breakthroughs have been achieved across a variety of scenarios and applications (Jordan and Mitchell, 2015). These technologies of AI and ML have increasingly become cornerstones of innovation, driving progress in the fields such as healthcare diagnostics (Alowais *et al.*, 2023), autonomous vehicles (Betz *et al.*, 2022), financial modeling (Cao, 2022), protein structure prediction (Abramson *et al.*, 2024), and numerous other domains. Despite these impressive achievements, the trustworthiness of AI systems has come under scrutiny, particularly in security-critical and privacy-sensitive domains. Ensuring that AI systems and ML models are reliable, secure, and trustworthy is not merely desirable but essential for their deployment in large-scale real-world applications.

The heart of machine learning is built upon two crucial aspects: data and model. Data serves as the fundamental resource, representing the diverse, complex, and often noisy real-world phenomena. Meanwhile, the model functions as the learner, with specific model architectures that absorb patterns and knowledge from the data, empowering it to make predictions or decisions in previously unseen scenarios (LeCun *et al.*, 2015). Notably, the emergence of data privacy and security issues arises at the intersection of data and model, as training data may inadvertently include malicious information that implants backdoors in models (Li *et al.*, 2022b), or contain sensitive privacy details that models could unintentionally expose during inference (Liu *et al.*, 2021a). Overall, these challenges highlight the need to understand and develop trustworthy machine learning from data to model, including perspectives of data-centric methods, privacy and security, and foundation models.

In this monograph, we first discuss the trustworthy data-centric learning, which emphasizes the risks associated with noisy (Song *et al.*, 2022a), long-tailed (Zhang *et al.*, 2023e), out-of-distribution (Yang *et al.*, 2024), and adversarial data (Wang *et al.*, 2019). As the foundation of any ML models, the data directly impacts the model's reliability and generalization ability. Trustworthy data-centric learning focuses on exploring the essential mechanisms by which data influences the trustworthiness of ML models, and designing robust approaches to adapt to, defend against, and mitigate the negative effects of such data challenges. This includes developing robust algorithms for learning from noisy data, handling long-tailed distributions, detecting and generalizing across out-of-distribution data, and improving adversarial robustness and defense strategies. Generally, the goal of trustworthy data-centric learning is to ensure the trustworthiness of ML models from data perspectives, enabling them to handle diverse and complex real-world scenarios while maintaining highly accurate performance.

Privacy and security problems are paramount in the deployment of machine learning models, particularly when dealing with sensitive data (Liu *et al.*, 2021a), such as finance and healthcare. In this monograph, we delve deeply into the approaches that attacking and safeguarding ML systems, addressing challenges from both the data and model perspectives. Specifically, we start with discussing differential privacy (Abadi *et al.*, 2016), a key technique that adds controlled noise to protect privacy while maintaining data utility. We then review two major privacy threats: membership inference attacks (Hu *et al.*, 2022b) and model inversion attacks (Song and Namiot, 2022), both of which attempt to aim to leak information from the training data. Next, we cover data poisoning attacks (Fan *et al.*, 2022) that degrade the model performance by manipulating the training data. Additionally, we discuss three types of promising approaches, including machine unlearning (Nguyen *et al.*, 2022), non-transfer learning (Niu *et al.*, 2020) and federated learning (Zhang *et al.*, 2021a), all of which offer solutions to alleviate these risks and enhance trustworthiness of models in defending against such attacks. Overall, these research efforts highlight the vulnerabilities of models and contribute to enhancing the robustness and privacy protection.

Recently, the development of large foundation models, such as ChatGPT, Llama, and Gemini, has revolutionized the field of ML, paving the pathway to artificial general intelligence (Zhou *et al.*, 2024a). Despite their remarkable capacities, these foundation models still face various safety concerns. In this monograph, we discuss several potential risks and vulnerabilities of foundation models, aiming to highlight their weaknesses and provide insights for constructing trustworthy foundation models. In particular, we discuss jailbreak prompts that inveigle foundation models to generate harmful content (Yi *et al.*, 2024), and then review watermarking techniques to ensure content provenance and copyright (Liu *et al.*, 2024a). We next introduce hallucination, which is a critical issue for foundation models in generating unreliable and spurious content (Rawte *et al.*, 2023). Moreover, we discuss causal learning and reasoning methods (Chi *et al.*, 2024a) to enhance reliability of the content generated by foundation models. Finally, we compare different trustworthy concerns in open and proprietary foundation models with their distinct properties. In short, this monograph provides a comprehensive review and discussion of the key challenges and advancements in developing trustworthy machine learning systems, from data-centric approaches, privacy and security concerns to foundation models.

Overview. The monograph is organized around core aspects of trustworthy machine learning from data to models, including *data-centric learning*, *private and secured learning*, and *foundation models.*

- **Trustworthy Data-centric Learning.** First, we start with a systematic review of trustworthy data-centric learning, covering *data-noise learning*, *long-tailed learning*, *out-of-distribution learn-*

ing, as well as *adversarial examples and defense*. These research topics cover fundamental problems regarding data-level issues, e.g., label noise, shifted distribution, outliers, and worst-case corruption. We further categorize specific research problems under the general learning paradigms for discussion.

- **Trustworthy Private and Secured Learning.** Second, we focus on aspects of private and secured learning, covering *differential privacy*, *membership inference attack*, *model inversion attack*, *data poisoning attack*, *machine unlearning*, *non-transferable learning*, and *federated learning*. Considering the trustworthy expectation of privacy, security, and usage or ownership protection, we review a series of critical technologies and research problems.

- **Trustworthy Foundation Models.** Finally, we explore building the trustworthy foundation models, covering *jailbreak prompts*, *watermarking*, *hallucination*, *casual learning and reasoning*, as well as comparison on *open and proprietary foundation models*. These research problems reveal the vulnerability of foundation models in usage control and point the way to developing robust and reliable model learning and reasoning.

In each part, we elaborate on the detailed problem setup and methodology or research directions, for which we conduct a further discussion on promising future development in the problem. We hope this monograph can provide a comprehensive investigation from data to models in trustworthy machine learning and more new insights.

Target Audience and Reading Guidelines. This monograph is intended for researchers, professionals, and graduate students working in the fields of ML and AI. Some sections may contain technical descriptions and discussions that assume a basic knowledge of core ML concepts. Target readers are expected to have a foundational understanding of these key concepts, including supervised, semi-supervised and unsupervised learning, optimization methods, representation learning, federated learning, among others. For undergraduate students or early-year graduate students new to the field, we recommend first referring to conventional textbooks on ML (Jordan and Mitchell, 2015) and deep learning (LeCun *et al.*, 2015) before delving into the specialized topics covered in this monograph. Additionally, readers unfamiliar with trustworthy ML may benefit from the monograph to establish a solid background of key challenges, methodologies, and emerging research directions in the field. Each section is organized to offer a thorough review and perceptive discussions of its respective topic. This structure makes the monograph suitable for both newcomers looking for a thorough grasp of trustworthy ML and seasoned researchers hoping to further advance the field. We hope this monograph serves as a valuable resource for a broad audience interested in developing more trustworthy and reliable ML systems.

Discussion on the Topic Coverage. In this monograph, our primary focus is on the critical technical aspects of trustworthy machine learning, specifically addressing issues of robustness, privacy, security, and reliability. Our goal is to provide a systematic overview of the evolving research landscape from a data-centric perspective, and consider the privacy and security challenges to building trustworthy foundation models. The overall discussion is with a particular emphasis on data quality, privacy protection, and security challenges. However, we should acknowledge that we don't explicitly cover all the topics that are also important and highly relevant to the field of trustworthy machine learning, such as fairness and bias (Mehrabi *et al.*, 2021; Pessach and Shmueli, 2022), ethics (Chen *et al.*, 2021a; Holmes *et al.*, 2022), and explainability (Došilović *et al.*, 2018; Murdoch *et al.*, 2019). For instance, ML systems can often mirror and amplify biases present in their training data and reflect stereotypes in their outputs, causing a broader concern of algorithm discrimination. In essence, without properly examining the training data, the training stage of the model would reinforce historical or societal prejudices that favor or over-present dominant groups and under-present or mischaracterize minorities.

Moreover, the non-transparency property of ML models increases the difficulty of effective auditing for untrustworthy issues. We do not explicitly discuss fairness and bias, ethics, or interpretability in separated sections because these issues often require a distinct and deep dive into the societal and moral implications of technologies. However, we recognize the strong connections between these research topics and the core themes of our content. For example, issues of fairness and bias are often intertwined with the robustness considerations discussed in the context of adversarial examples and noisy data, as biased data can lead to unfair and unreliable model outcomes. Similarly, the methodologies explored for privacy protection, such as differential privacy, can contribute to ethical practices by safeguarding sensitive information and ensuring compliance with ethical standards in data handling. Finally, the discussion on casual learning and reasoning for trustworthy foundation models also touches on explainability in the sense that building reliable models often requires transparent and understandable methodologies. We hope this work will complement existing literature and encourage further exploration of these important areas within the context of trustworthy machine learning.

References

Aaronson, S. and H. Kirchner. (2022). “Watermarking GPT outputs”. URL: https://www.scottaaronson.com/talks/watermark.ppt.

Abadi, M., A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. (2016). “Deep learning with differential privacy”. In: *CCS*.

Abdulaal, A., N. Montana-Brown, T. He, A. Ijishakin, I. Drobnjak, D. C. Castro, D. C. Alexander, *et al.* (2023). “Causal Modelling Agents: Causal Graph Discovery through Synergising Metadata-and Data-driven Reasoning”. In: *The Twelfth International Conference on Learning Representations*.

Abramson, J., J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, *et al.* (2024). “Accurate structure prediction of biomolecular interactions with AlphaFold 3”. *Nature*.

Achiam, J., S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, *et al.* (2023). “Gpt-4 technical report”. *arXiv preprint arXiv:2303.08774*.

Ahuja, K., E. Caballero, D. Zhang, J.-C. Gagnon-Audet, Y. Bengio, I. Mitliagkas, and I. Rish. (2021). “Invariance Principle Meets Information Bottleneck for Out-of-Distribution Generalization”. In: *NeurIPS*.

Almazrouei, E., H. Alobeidli, A. Alshamsi, A. Cappelli, R. Cojocaru, M. Debbah, É. Goffinet, D. Hesslow, J. Launay, Q. Malartic, *et al.* (2023). "The falcon series of open language models". *arXiv preprint arXiv:2311.16867*.

Alowais, S. A., S. S. Alghamdi, N. Alsuhebany, T. Alqahtani, A. I. Alshaya, S. N. Almohareb, A. Aldairem, M. Alrashed, K. Bin Saleh, H. A. Badreldin, *et al.* (2023). "Revolutionizing healthcare: the role of artificial intelligence in clinical practice". *BMC medical education*.

Amiri, M. M., D. Gunduz, S. R. Kulkarni, and H. V. Poor. (2020). "Federated Learning With Quantized Global Model Updates". *arXiv preprint arXiv:2006.10672*.

Amiri, M. M. and D. Gündüz. (2020). "Federated Learning Over Wireless Fading Channels". *IEEE Transactions on Wireless Communications*.

An, S., G. Tao, Q. Xu, Y. Liu, G. Shen, Y. Yao, J. Xu, and X. Zhang. (2022). "Mirror: Model inversion for deep learning network with high fidelity". In: *NDSS*.

Andriushchenko, M. and N. Flammarion. (2020). "Understanding and improving fast adversarial training". In: *NeurIPS*.

Anh, T. T., N. C. Luong, D. Niyato, D. I. Kim, and L.-C. Wang. (2019). "Efficient Training Management for Mobile Crowd-Machine Learning: A Deep Reinforcement Learning Approach". *IEEE Wireless Communications Letters*.

Anil, C., E. Durmus, M. Sharma, J. Benton, S. Kundu, J. Batson, N. Rimsky, M. Tong, J. Mu, D. Ford, *et al.* (2024). "Many-shot jailbreaking". In: *NeurIPS*.

Arjovsky, M., L. Bottou, I. Gulrajani, and D. Lopez-Paz. (2019). "Invariant Risk Minimization". *arXiv preprint arXiv:1907.02893*.

Arpit, D., H. Wang, Y. Zhou, and C. Xiong. (2022). "Ensemble of Averages: Improving Model Selection and Boosting Performance in Domain Generalization". In: *NeurIPS*.

Ashwani, S., K. Hegde, N. R. Mannuru, D. S. Sengar, M. Jindal, K. C. R. Kathala, D. Banga, V. Jain, and A. Chadha. (2024). "Cause and effect: Can large language models truly understand causality?" In: *Proceedings of the AAAI Symposium Series*.

Assran, M., R. Balestriero, Q. Duval, F. Bordes, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, and N. Ballas. (2023). "The hidden uniform cluster prior in self-supervised learning". In: *ICLR*.

Bai, J., Z. Liu, H. Wang, J. Hao, Y. FENG, H. Chu, and H. Hu. (2023). "On the Effectiveness of Out-of-Distribution Data in Self-Supervised Long-Tail Learning." In: *ICLR*.

Balloccu, S., P. Schmidtová, M. Lango, and O. Dušek. (2024). "Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs". *arXiv preprint arXiv:2402.03927*.

Ban, T., L. Chen, D. Lyu, X. Wang, and H. Chen. (2023a). "Causal structure learning supervised by large language model". *arXiv preprint arXiv:2311.11689*.

Ban, T., L. Chen, X. Wang, and H. Chen. (2023b). "From query tools to causal architects: Harnessing large language models for advanced causal discovery from data". *arXiv preprint arXiv:2306.16902*.

Barnett, S., S. Kurniawan, S. Thudumu, Z. Brannelly, and M. Abdelrazek. (2024). "Seven failure points when engineering a retrieval augmented generation system". In: *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI*. 194–199.

Barreno, M., B. Nelson, R. Sears, A. D. Joseph, and J. D. Tygar. (2006). "Can machine learning be secure?" In: *ACM CCS*.

Baruch, G., M. Baruch, and Y. Goldberg. (2019). "A little is enough: Circumventing defenses for distributed learning". *NeurIPS*.

Bendale, A. and T. E. Boult. (2016). "Towards open set deep networks". In: *CVPR*.

Betz, J., H. Zheng, A. Liniger, U. Rosolia, P. Karle, M. Behl, V. Krovi, and R. Mangharam. (2022). "Autonomous vehicles on the edge: A survey on autonomous vehicle racing". *IEEE Open Journal of Intelligent Transportation Systems*.

Biggio, B., B. Nelson, P. Laskov, *et al.* (2012). "Poisoning attacks against support vector machines". In: *ICML*.

Biggio, B., B. Nelson, and P. Laskov. (2011). "Support vector machines under adversarial label noise". In: *ACML*.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.* (2020). "Language models are few-shot learners". In: *NeurIPS*.

Brown, T. B. (2020). "Language models are few-shot learners". *arXiv preprint arXiv:2005.14165*.

Burns, C., H. Ye, D. Klein, and J. Steinhardt. (2022). "Discovering latent knowledge in language models without supervision". *arXiv preprint arXiv:2212.03827*.

Cai, H., S. Liu, and R. Song. (2023). "Is Knowledge All Large Language Models Needed for Causal Reasoning?" *arXiv preprint arXiv:2401.00139*.

Cai, H., A. Arunasalam, L. Y. Lin, A. Bianchi, and Z. B. Celik. (2024). "Take a look at it! rethinking how to evaluate language model jailbreak". In: *ACL*.

Cao, C., Z. Zhong, Z. Zhou, Y. Liu, T. Liu, and B. Han. (2024). "Envisioning Outlier Exposure by Large Language Models for Out-of-Distribution Detection". *ICML*.

Cao, D., S. Chang, Z. Lin, G. Liu, and D. Sun. (2019a). "Understanding distributed poisoning attack in federated learning". In: *ICPADS*.

Cao, K., C. Wei, A. Gaidon, N. Arechiga, and T. Ma. (2019b). "Learning imbalanced datasets with label-distribution-aware margin loss". *Advances in neural information processing systems*. 32.

Cao, L. (2022). "Ai in finance: challenges, techniques, and opportunities". *ACM Computing Surveys (CSUR)*.

Carlini, N., S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramer. (2022). "Membership inference attacks from first principles". In: *IEEE SP*.

Carlini, N., M. Nasr, C. A. Choquette-Choo, M. Jagielski, I. Gao, P. W. Koh, D. Ippolito, F. Tramèr, and L. Schmidt. (2023). "Are aligned neural networks adversarially aligned?" In: *NeurIPS*.

Carmon, Y., A. Raghunathan, L. Schmidt, J. C. Duchi, and P. S. Liang. (2019). "Unlabeled data improves adversarial robustness". In: *NeurIPS*.

Caton, S. and C. Haas. (2024). "Fairness in Machine Learning: A Survey". *ACM Comput. Surv.* 56(7): 166:1–166:38.

Cha, J., S. Chun, K. Lee, H.-C. Cho, S. Park, Y. Lee, and S. Park. (2021). "SWAD: Domain Generalization by Seeking Flat Minima". In: *NeurIPS*.

Chang, Y., X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, *et al.* (2024). "A survey on evaluation of large language models". *ACM transactions on intelligent systems and technology*. 15(3): 1–45.

Chanpuriya, S., C. Musco, K. Sotiropoulos, and C. Tsourakakis. (2021). "Deepwalking backwards: from embeddings back to graphs". In: *ICML*.

Chao, P., A. Robey, E. Dobriban, H. Hassani, G. J. Pappas, and E. Wong. (2023). "Jailbreaking black box large language models in twenty queries". *arXiv preprint arXiv:2310.08419*.

Chaudhuri, K. and C. Monteleoni. (2008). "Privacy-preserving logistic regression". In: *NIPS*.

Chen, D., N. Yu, Y. Zhang, and M. Fritz. (2020a). "Gan-leaks: A taxonomy of membership inference attacks against generative models". In: *ACM SIGSAC*.

Chen, H., Y. Dong, Z. Wang, X. Yang, C. Duan, H. Su, and J. Zhu. (2024a). "Robust Classification via a Single Diffusion Model". In: *ICML*.

Chen, I. Y., E. Pierson, S. Rose, S. Joshi, K. Ferryman, and M. Ghassemi. (2021a). "Ethical machine learning in healthcare". *Annual review of biomedical data science*. 4(1): 123–144.

Chen, M., J. Yao, L. Xing, Y. Wang, Y. Zhang, and Y. Wang. (2023a). "Redundancy-adaptive multimodal learning for imperfect data". *arXiv preprint arXiv:2310.14496*.

Chen, M., W. Gao, G. Liu, K. Peng, and C. Wang. (2023b). "Boundary Unlearning". In: *CVPR*.

Chen, R. J., J. J. Wang, D. F. Williamson, T. Y. Chen, J. Lipkova, M. Y. Lu, S. Sahai, and F. Mahmood. (2023c). "Algorithmic fairness in artificial intelligence for medicine and healthcare". *Nature biomedical engineering*. 7(6): 719–742.

Chen, S., M. Kahla, R. Jia, and G. Qi. (2021b). "Knowledge-enriched distributional model inversion attacks". In: *ICCV*.

Chen, T., Z. Zhang, S. Liu, S. Chang, and Z. Wang. (2020b). "Robust overfitting may be mitigated by properly learned smoothening". In: *ICLR*.

Chen, T., S. Kornblith, M. Norouzi, and G. Hinton. (2020c). "A simple framework for contrastive learning of visual representations". In: *ICML*.

Chen, T., S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton. (2020d). "Big self-supervised models are strong semi-supervised learners".

Chen, X., Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. (2016). "Infogan: Interpretable representation learning by information maximizing generative adversarial nets". *NeurIPS*.

Chen, X., Y. Zhou, D. Wu, C. Yang, B. Li, Q. Hu, and W. Wang. (2023d). "Area: adaptive reweighting via effective area for long-tailed classification". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 19277–19287.

Chen, Y., H. C. Lent, and J. Bjerva. (2024b). "Text Embedding Inversion Security for Multilingual Language Models". In: *ACL*.

Chen, Y., Y. Bian, B. Han, and J. Cheng. (2024c). "How Interpretable Are Interpretable Graph Neural Networks?" In: *ICML*.

Chen, Y., Y. Bian, K. Zhou, B. Xie, B. Han, and J. Cheng. (2023e). "Does Invariant Graph Learning via Environment Augmentation Learn Invariance?" In: *NeurIPS*.

Chen, Y., W. Huang, K. Zhou, Y. Bian, B. Han, and J. Cheng. (2023f). "Understanding and Improving Feature Learning for Out-of-Distribution Generalization". In: *NeurIPS*.

Chen, Y., Y. Zhang, Y. Bian, H. Yang, K. Ma, B. Xie, T. Liu, B. Han, and J. Cheng. (2022). "Learning Causally Invariant Representations for Out-of-Distribution Generalization on Graphs". In: *NeurIPS*.

Chen, Y., K. Zhou, Y. Bian, B. Xie, K. Ma, Y. Zhang, H. Yang, B. Han, and J. Cheng. (2023g). "Pareto Invariant Risk Minimization". In: *ICLR*.

Chen, Z., Z. Zhao, H. Luo, H. Yao, B. Li, and J. Zhou. (2024d). "HALC: Object Hallucination Reduction via Adaptive Focal-Contrast Decoding". In: *ICML*.

Cheng, Y., C. Shan, Y. Shen, X. Li, S. Luo, and D. Li. (2024a). "Resurrecting Label Propagation for Graphs with Heterophily and Label Noise". In: *KDD*.

Cheng, Y., C. Shan, Y. Shen, X. Li, S. Luo, and D. Li. (2024b). "Resurrecting Label Propagation for Graphs with Heterophily and Label Noise". In: *KDD*.

Chi, H., H. Li, W. Yang, F. Liu, L. Lan, X. Ren, T. Liu, and B. Han. (2024a). "Unveiling Causal Reasoning in Large Language Models: Reality or Mirage?" In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Chi, H., H. Li, W. Yang, F. Liu, L. Lan, X. Ren, T. Liu, and B. Han. (2024b). "Unveiling Causal Reasoning in Large Language Models: Reality or Mirage?" In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Cho, Y. J., A. Manoel, G. Joshi, R. Sim, and D. Dimitriadis. (2022). "Heterogeneous ensemble knowledge transfer for training large models in federated learning". *arXiv preprint arXiv:2204.12703*.

Chowdhury, A. G., M. M. Islam, V. Kumar, F. H. Shezan, V. Jain, and A. Chadha. (2024). "Breaking down the defenses: A comparative survey of attacks on large language models". *arXiv preprint arXiv:2403.04786*.

Christiano, P. F., J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. (2017). "Deep reinforcement learning from human preferences". In: *NeurIPS*.

Chu, P., X. Bian, S. Liu, and H. Ling. (2020). "Feature space augmentation for long-tailed data". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*. Springer. 694–710.

Chuang, C.-Y., R. D. Hjelm, X. Wang, V. Vineet, N. Joshi, A. Torralba, S. Jegelka, and Y. Song. (2022). "Robust contrastive learning against noisy views". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16670–16681.

Chuang, Y.-S., Y. Xie, H. Luo, Y. Kim, J. Glass, and P. He. (2023). "Dola: Decoding by contrasting layers improves factuality in large language models". *arXiv preprint arXiv:2309.03883*.

Cour, T., B. Sapp, and B. Taskar. (2011). "Learning from partial labels". *The Journal of Machine Learning Research*.

Creager, E., J. Jacobsen, and R. S. Zemel. (2021). "Environment Inference for Invariant Learning". In: *ICML*.

Criado, M. F., F. E. Casado, R. Iglesias, C. V. Regueiro, and S. Barro. (2022). "Non-IID data and Continual Learning processes in Federated Learning: A long road ahead". *Information Fusion*.

Croitoru, F.-A., V. Hondru, R. T. Ionescu, and M. Shah. (2023). "Diffusion models in vision: A survey". *IEEE TPAMI*.

Cubuk, E. D., B. Zoph, J. Shlens, and Q. V. Le. (2020). "Randaugment: Practical automated data augmentation with a reduced search space". In: *CVPR workshops*.

Cui, Y., M. Jia, T.-Y. Lin, Y. Song, and S. Belongie. (2019). "Class-balanced loss based on effective number of samples". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9268–9277.

Cuturi, M. (2013). "Sinkhorn Distances: Lightspeed Computation of Optimal Transport". In: *NeurIPS*.

Dai, E., C. Aggarwal, and S. Wang. (2021). "Nrgnn: Learning a label noise resistant graph neural network on sparsely and noisily labeled graphs". In: *KDD*.

Dai, J., X. Pan, R. Sun, J. Ji, X. Xu, M. Liu, Y. Wang, and Y. Yang. (2024a). "Safe rlhf: Safe reinforcement learning from human feedback". In: *ICLR*.

Dai, R., Y. Zhang, A. Li, T. Liu, X. Yang, and B. Han. (2024b). "Enhancing One-Shot Federated Learning Through Data and Ensemble Co-Boosting". In: *The Twelfth International Conference on Learning Representations*.

Das, B. C., M. H. Amini, and Y. Wu. (2024). "Security and privacy challenges of large language models: A survey". *arXiv preprint arXiv:2402.00888*.

Demontis, A., M. Melis, M. Pintor, M. Jagielski, B. Biggio, A. Oprea, C. Nita-Rotaru, and F. Roli. (2019). "Why do adversarial attacks transfer? explaining transferability of evasion and poisoning attacks". In: *USENIX security*.

Deng, G., Y. Liu, Y. Li, K. Wang, Y. Zhang, Z. Li, H. Wang, T. Zhang, and Y. Liu. (2023). "MasterKey: Automated Jailbreak Across Multiple Large Language Model Chatbots". *arXiv preprint arXiv:2307.08715*.

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, J., S. Pang, Y. Chen, L. Xia, Y. Bai, H. Weng, and W. Xu. (2024). "SOPHON: Non-Fine-Tunable Learning to Restrain Task Transferability For Pre-trained Models". In: *IEEE SP*.

Deng, Z., X. Yang, S. Xu, H. Su, and J. Zhu. (2021). "LiBRe: A Practical Bayesian Approach to Adversarial Detection". In: *CVPR*.

Dennis, D. K., T. Li, and V. Smith. (2021). "Heterogeneity for the win: One-shot federated clustering". In: *International Conference on Machine Learning*.

Desai, A., T.-Y. Wu, S. Tripathi, and N. Vasconcelos. (2021). "Learning of visual relations: The devil is in the tails". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15404–15413.

Devlin, J. (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding". *arXiv preprint arXiv:1810.04805*.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *NAACL*.

Dhawan, N., L. Cotta, K. Ullrich, R. Krishnan, and C. J. Maddison. (2024). "End-To-End Causal Effect Estimation from Unstructured Natural Language Data". In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Diao, H., Y. Zhang, L. Ma, and H. Lu. (2021). "Similarity reasoning and filtration for image-text matching". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 35. No. 2. 1218–1226.

Ding, G. W., Y. Sharma, K. Y. C. Lui, and R. Huang. (2020). "MMA Training: Direct Input Space Margin Maximization through Adversarial Training". In: *ICLR*.

Ding, P., J. Kuang, D. Ma, X. Cao, Y. Xian, J. Chen, and S. Huang. (2023). "A Wolf in Sheep's Clothing: Generalized Nested Jailbreak Prompts can Fool Large Language Models Easily". In: *NAACL*.

Djurisic, A., N. Bozanic, A. Ashok, and R. Liu. (2023). "Extremely simple activation shaping for out-of-distribution detection". In:

Dong, M. and Y. Kluger. (2023). "Towards understanding and reducing graph structural noise for GNNs". In: *ICML*.

Došilović, F. K., M. Brčić, and N. Hlupić. (2018). "Explainable artificial intelligence: A survey". In: *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*. IEEE. 0210–0215.

Du, X., T. Bian, Y. Rong, B. Han, T. Liu, T. Xu, W. Huang, Y. Li, and J. Huang. (2021). "Noise-robust graph learning by estimating and leveraging pairwise interactions". *TMLR*.

Duan, J., F. Kong, S. Wang, X. Shi, and K. Xu. (2023). "Are diffusion models vulnerable to membership inference attacks?" In: *ICML*. PMLR.

Dubey, A., A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, *et al.* (2024). "The llama 3 herd of models". *arXiv*.

Dwork, C., K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor. (2006). "Our data, ourselves: Privacy via distributed noise generation". In: *Advances in Cryptology – EUROCRYPT*.

Dwork, C. and A. Roth. (2014). "The algorithmic foundations of differential privacy". *Foundations and Trends® in Theoretical Computer Science*.

Dziri, N., S. Milton, M. Yu, O. Zaiane, and S. Reddy. (2022). "On the origin of hallucinations in conversational models: Is it the datasets or the models?" *arXiv preprint arXiv:2204.07931*.

ElKordy, A. and A. S. Avestimehr. (2020). "Secure aggregation with heterogeneous quantization in federated learning". *arXiv preprint arxiv:2009.14388*.

Esmaeilpour, S., B. Liu, E. Robertson, and L. Shu. (2022). "Zero-shot out-of-distribution detection based on the pre-trained model clip". In: *AAAI*.

Fan, C., J. Liu, Y. Zhang, D. Wei, E. Wong, and S. Liu. (2024). "SalUn: Empowering Machine Unlearning via Gradient-based Weight Saliency in Both Image Classification and Generation". In: *ICLR*.

Fan, J., Q. Yan, M. Li, G. Qu, and Y. Xiao. (2022). "A survey on data poisoning attacks and defenses". In: *DSC*.

Fang, A., G. Ilharco, M. Wortsman, Y. Wan, V. Shankar, A. Dave, and L. Schmidt. (2022a). "Data Determines Distributional Robustness in Contrastive Language Image Pre-training (CLIP)". In: *ICML*.

Fang, Z., Y. Li, J. Lu, J. Dong, B. Han, and F. Liu. (2022b). "Is out-of-distribution detection learnable?" In:

Feder, A., Y. Wald, C. Shi, S. Saria, and D. Blei. (2024). "Causal-structure driven augmentations for text ood generalization". *Advances in Neural Information Processing Systems*.

Feffer, M., A. Sinha, Z. C. Lipton, and H. Heidari. (2024). "Red-Teaming for Generative AI: Silver Bullet or Security Theater?" In: *AIES*.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Bradford Books. URL: https://mitpress.mit.edu/9780262561167/.

Feng, C., Y. Zhong, and W. Huang. (2021). "Exploring classification equilibrium in long-tailed object detection". In: *Proceedings of the IEEE/CVF International conference on computer vision*. 3417–3426.

Feng, L., J. Lv, B. Han, M. Xu, G. Niu, X. Geng, B. An, and M. Sugiyama. (2020). "Provably Consistent Partial-Label Learning". In: *NeurIPS*.

Feng, Q., L. Xie, S. Fang, and T. Lin. (2024). "BaCon: Boosting Imbalanced Semi-supervised Learning via Balanced Feature-Level Contrastive Learning". In: *AAAI*.

Feng, Z., Z. Zeng, C. Guo, Z. Li, and L. Hu. (2023). "Learning from noisy correspondence with tri-partition for cross-modal matching". *IEEE Transactions on Multimedia*.

Fernandez, P., G. Couairon, H. Jégou, M. Douze, and T. Furon. (2023). "The stable signature: Rooting watermarks in latent diffusion models". In: *ICCV*.

Fort, S., J. Ren, and B. Lakshminarayanan. (2021). "Exploring the Limits of Out-of-Distribution Detection". In: *NeurIPS*.

Frankle, J. and M. Carbin. (2018). "The Lottery Ticket Hypothesis: Training Pruned Neural Networks". *CoRR*.

Fredrikson, M., E. Lantz, S. Jha, S. Lin, D. Page, and T. Ristenpart. (2014). "Privacy in pharmacogenetics: An {End-to-End} case study of personalized warfarin dosing". In: *USENIX Security*.

Fredrikson, M., S. Jha, and T. Ristenpart. (2015). "Model inversion attacks that exploit confidence information and basic countermeasures". In: *CCS*.

Fu, W., H. Wang, C. Gao, G. Liu, Y. Li, and T. Jiang. (2023). "A Probabilistic Fluctuation based Membership Inference Attack for Diffusion Models". *arXiv e-prints*: arXiv–2308.

Fu, X., X. Wang, Q. Li, J. Liu, J. Dai, and J. Han. (2024). "Model Will Tell: Training Membership Inference for Diffusion Models". *arXiv preprint arXiv:2403.08487*.

Fukuchi, K., Q. K. Tran, and J. Sakuma. (2017). "Differentially private empirical risk minimization with input perturbation". In: *Discovery Science*.

Gagnon-Audet, J.-C., K. Ahuja, M. J. D. Bayazi, P. Mousavi, G. Dumas, and I. Rish. (2023). "WOODS: Benchmarks for Out-of-Distribution Generalization in Time Series". *TMLR*.

Gan, K. and T. Wei. (2024). "Erasing the Bias: Fine-Tuning Foundation Models for Semi-Supervised Learning". In: *ICML*.

Gandikota, R., J. Materzynska, J. Fiotto-Kaufman, and D. Bau. (2023). "Erasing concepts from diffusion models". In: *ICCV*.

Ganin, Y., E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. S. Lempitsky. (2016). "Domain-Adversarial Training of Neural Networks". *Journal of Mache Learning Research*.

Gao, D., X. Yao, and Q. Yang. (2022a). "A survey on heterogeneous federated learning". *arXiv preprint arXiv:2210.04505*.

Gao, L., Z. Dai, P. Pasupat, A. Chen, A. T. Chaganty, Y. Fan, V. Y. Zhao, N. Lao, H. Lee, D.-C. Juan, *et al.* (2022b). "Rarr: Researching and revising what language models say, using language models". *arXiv preprint arXiv:2210.08726*.

Gao, R., F. Liu, J. Zhang, B. Han, T. Liu, G. Niu, and M. Sugiyama. (2021). "Maximum Mean Discrepancy Test is Aware of Adversarial Attacks". In: *ICML*.

Ghosh, A., J. Chung, D. Yin, and K. Ramchandran. (2020). "An efficient framework for clustered federated learning". *Advances in Neural Information Processing Systems*.

Golatkar, A., A. Achille, and S. Soatto. (2020). “Eternal sunshine of the spotless net: Selective forgetting in deep networks”. In: *CVPR*.

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

Goodfellow, I. J., J. Shlens, and C. Szegedy. (2015). “Explaining and Harnessing Adversarial Examples”. In: *ICLR*.

Gretton, A., K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. J. Smola. (2012). “A Kernel Two-Sample Test”. *J. Mach. Learn. Res.*

Grill, J.-B., F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, *et al.* (2020). “Bootstrap your own latent-a new approach to self-supervised learning”. In: *NeurIPS*.

Gu, C., X. L. Li, P. Liang, and T. Hashimoto. (2023). “On the learnability of watermarks for language models”. *arXiv preprint arXiv:2312.04469*.

Gu, X., X. Zheng, T. Pang, C. Du, Q. Liu, Y. Wang, J. Jiang, and M. Lin. (2024). “Agent smith: A single image can jailbreak one million multimodal llm agents exponentially fast”. In: *ICML*.

Guerreiro, N. M., D. M. Alves, J. Waldendorf, B. Haddow, A. Birch, P. Colombo, and A. F. Martins. (2023). “Hallucinations in large multilingual translation models”. *Transactions of the Association for Computational Linguistics*. 11: 1500–1517.

Guha, N., A. Talwalkar, and V. Smith. (2019). “One-shot federated learning”. *arXiv preprint arXiv:1902.11175*.

Gui, S., X. Li, L. Wang, and S. Ji. (2022). “GOOD: A Graph Out-of-Distribution Benchmark”. In: *NeurIPS*.

Gulrajani, I. and D. Lopez-Paz. (2021). “In Search of Lost Domain Generalization”. In: *ICLR*.

Guo, C., G. Pleiss, Y. Sun, and K. Q. Weinberger. (2017). “On calibration of modern neural networks”. In: *ICML*.

Guo, J., Y. Li, L. Wang, S.-T. Xia, H. Huang, C. Liu, and B. Li. (2024). “Domain watermark: Effective and harmless dataset copyright protection is closed at hand”. In: *NeurIPS*.

Hajikhani, A. and C. Cole. (2024). "A critical review of large language models: Sensitivity, bias, and the path toward specialized ai". *Quantitative Science Studies*. 5(3): 736–756.

Han, B., J. Yao, G. Niu, M. Zhou, I. Tsang, Y. Zhang, and M. Sugiyama. (2018a). "Masking: A new perspective of noisy supervision". *Advances in neural information processing systems*. 31.

Han, B., Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama. (2018b). "Co-teaching: Robust training of deep neural networks with extremely noisy labels". *Advances in neural information processing systems*. 31.

Han, G., J. Choi, H. Lee, and J. Kim. (2023a). "Reinforcement learning-based black-box model inversion attacks". In: *CVPR*.

Han, H., K. Miao, Q. Zheng, and M. Luo. (2023b). "Noisy correspondence learning with meta similarity correction". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7517–7526.

Han, H., Q. Zheng, G. Dai, M. Luo, and J. Wang. (2024). "Learning to Rematch Mismatched Pairs for Robust Cross-Modal Retrieval". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 26679–26688.

Hanif, A., M. Naseer, S. H. Khan, M. Shah, and F. S. Khan. (2023). "Frequency Domain Adversarial Training for Robust Volumetric Medical Segmentation". In: *MICCAI*.

Hayes, J., L. Melis, G. Danezis, and E. De Cristofaro. (2017). "Logan: Membership inference attacks against generative models". *arXiv preprint arXiv:1705.07663*.

He, K., H. Fan, Y. Wu, S. Xie, and R. Girshick. (2020). "Momentum contrast for unsupervised visual representation learning". In: *CVPR*.

Henaff, O. (2020). "Data-efficient image recognition with contrastive predictive coding". In: *ICML*.

Hendrycks, D. and K. Gimpel. (2017). "A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks". In: *ICLR*.

Hendrycks, D., M. Mazeika, and T. Dietterich. (2018). "Deep anomaly detection with outlier exposure". In: *ICLR*.

Hessel, J., A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi. (2021). "Clipscore: A reference-free evaluation metric for image captioning". *arXiv preprint arXiv:2104.08718*.

Holmes, W., K. Porayska-Pomsta, K. Holstein, E. Sutherland, T. Baker, S. B. Shum, O. C. Santos, M. T. Rodrigo, M. Cukurova, I. I. Bittencourt, *et al.* (2022). "Ethics of AI in education: Towards a community-wide framework". *International Journal of Artificial Intelligence in Education*: 1–23.

Hong, F., J. Yao, Y. Lyu, Z. Zhou, I. W. Tsang, Y. Zhang, and Y. Wang. (2024a). "On Harmonizing Implicit Subpopulations". In: *ICLR*.

Hong, F., J. Yao, Z. Zhou, Y. Zhang, and Y. Wang. (2023). "Long-Tailed Partial Label Learning via Dynamic Rebalancing". In: *ICLR*.

Hong, Y., S. Han, K. Choi, S. Seo, B. Kim, and B. Chang. (2021). "Disentangling label distribution for long-tailed visual recognition". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6626–6636.

Hong, Z.-W., I. Shenfeld, T.-H. Wang, Y.-S. Chuang, A. Pareja, J. R. Glass, A. Srivastava, and P. Agrawal. (2024b). "Curiosity-driven Red-teaming for Large Language Models". In: *ICLR*.

Hong, Z., L. Shen, and T. Liu. (2024c). "Your Transferability Barrier is Fragile: Free-Lunch for Transferring the Non-Transferable Learning". In: *CVPR*.

Hong, Z., Z. Wang, L. Shen, Y. Yao, Z. Huang, S. Chen, C. Yang, M. Gong, and T. Liu. (2024d). "Improving Non-Transferable Representation Learning by Harnessing Content and Style". In: *ICLR*.

Hospedales, T., A. Antoniou, P. Micaelli, and A. Storkey. (2021). "Meta-learning in neural networks: A survey". *IEEE transactions on pattern analysis and machine intelligence*. 44(9): 5149–5169.

Hou, A. B., J. Zhang, T. He, Y. Wang, Y.-S. Chuang, H. Wang, L. Shen, B. Van Durme, D. Khashabi, and Y. Tsvetkov. (2023). "Semstamp: A semantic watermark with paraphrastic robustness for text generation". *arXiv preprint arXiv:2310.03991*.

Hu, C.-H., Z. Chen, and E. Larsson. (2022a). "Scheduling and Aggregation Design for Asynchronous Federated Learning Over Wireless Networks". *IEEE Journal on Selected Areas in Communications*.

Hu, H., Z. Salcic, L. Sun, G. Dobbie, P. S. Yu, and X. Zhang. (2022b). "Membership inference attacks on machine learning: A survey". *ACM CSUR*.

Hu, P., Z. Huang, D. Peng, X. Wang, and X. Peng. (2023). "Cross-modal retrieval with partially mismatched pairs". *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 45(8): 9595–9610.

Hu, W., G. Niu, I. Sato, and M. Sugiyama. (2018). "Does Distributionally Robust Supervised Learning Give Robust Classifiers?" In: *ICML*.

Huang, J., D. Yang, and C. Potts. (2024a). "Demystifying verbatim memorization in large language models". *arXiv preprint arXiv:2407.17817*.

Huang, Q., X. Dong, P. Zhang, B. Wang, C. He, J. Wang, D. Lin, W. Zhang, and N. Yu. (2024b). "Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation". In: *CVPR*.

Huang, R., A. Geng, and Y. Li. (2021a). "On the importance of gradients for detecting distributional shifts in the wild". *NeurIPS*. 34.

Huang, R., Y. Long, J. Han, H. Xu, X. Liang, C. Xu, and X. Liang. (2023a). "Nlip: Noise-robust language-image pre-training". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. No. 1. 926–934.

Huang, T., S. Hu, F. Ilhan, S. F. Tekin, and L. Liu. (2024c). "Harmful fine-tuning attacks and defenses for large language models: A survey". *arXiv preprint arXiv:2409.18169*.

Huang, W. R., J. Geiping, L. Fowl, G. Taylor, and T. Goldstein. (2020). "Metapoison: Practical general-purpose clean-label data poisoning". *NeurIPS*.

Huang, W., A. Han, Y. Chen, Y. Cao, zhiqiang xu, and T. Suzuki. (2024d). "On the Comparison between Multi-modal and Single-modal Contrastive Learning". In: *NeurIPS*.

Huang, Y., B. Bai, S. Zhao, K. Bai, and F. Wang. (2022). "Uncertainty-Aware Learning against Label Noise on Imbalanced Datasets". In: *AAAI*.

Huang, Z., G. Niu, X. Liu, W. Ding, X. Xiao, H. Wu, and X. Peng. (2021b). "Learning with noisy correspondence for cross-modal matching". *Advances in Neural Information Processing Systems*. 34: 29406–29419.

Huang, Z., M. Yang, X. Xiao, P. Hu, and X. Peng. (2024e). "Noise-robust Vision-language Pre-training with Positive-negative Learning". *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Huang, Z., Y. Fan, C. Liu, W. Zhang, Y. Zhang, M. Salzmann, S. Süsstrunk, and J. Wang. (2023b). "Fast adversarial training with adaptive step size". In: *IEEE Transactions on Image Processing*.

Hüllermeier, E. and J. Beringer. (2006). "Learning from ambiguously labeled examples". *Intell. Data Anal.* 10(5): 419–439.

Imbens, G. W. and D. B. Rubin. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press.

Iyengar, R., J. P. Near, D. Song, O. Thakkar, A. Thakurta, and L. Wang. (2019). "Towards practical differentially private convex optimization". In: *SP*.

Jagielski, M., A. Oprea, B. Biggio, C. Liu, C. Nita-Rotaru, and B. Li. (2018). "Manipulating machine learning: Poisoning attacks and countermeasures for regression learning". In: *IEEE symposium on security and privacy (SP)*.

Ji, Z., N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung. (2023). "Survey of hallucination in natural language generation". *ACM Computing Surveys*.

Jia, J., J. Liu, P. Ram, Y. Yao, G. Liu, Y. Liu, P. Sharma, and S. Liu. (2023). "Model Sparsity Can Simplify Machine Unlearning". In: *NeurIPS*.

Jia, X., Y. Zhang, X. Wei, B. Wu, K. Ma, J. Wang, and X. Cao. (2022a). "Prior-Guided Adversarial Initialization for Fast Adversarial Training". In: *ECCV*.

Jia, X., Y. Zhang, B. Wu, K. Ma, J. Wang, and X. Cao. (2022b). "LAS-AT: Adversarial Training with Learnable Attack Strategy". In: *CVPR*.

Jia, Y., X. Peng, R. Wang, and M. Zhang. (2024). "Long-Tailed Partial Label Learning by Head Classifier and Tail Classifier Cooperation". In: *AAAI*. Ed. by M. J. Wooldridge, J. G. Dy, and S. Natarajan.

Jiang, H., L. Ge, Y. Gao, J. Wang, and R. Song. (2024a). “LLM4Causal: Democratized Causal Tools for Everyone via Large Language Model”. In: *First Conference on Language Modeling*.

Jiang, L. and T. Lin. (2023). “Test-Time Robust Personalization for Federated Learning”.

Jiang, X., F. Liu, Z. Fang, H. Chen, T. Liu, F. Zheng, and B. Han. (2024b). “Negative Label Guided OOD Detection with Pretrained Vision-Language Models”. In: *ICLR*.

Jiang, Z., T. Chen, B. J. Mortazavi, and Z. Wang. (2021). “Self-damaging contrastive learning”. In: *ICML*.

Jin, Z., Y. Chen, F. Leeb, L. Gresele, O. Kamal, L. Zhiheng, K. Blin, F. G. Adauto, M. Kleiman-Weiner, M. Sachan, *et al.* (2023). “Cladder: Assessing causal reasoning in language models”. In: *Thirty-seventh conference on neural information processing systems*.

Jin, Z., J. Liu, L. Zhiheng, S. Poff, M. Sachan, R. Mihalcea, M. T. Diab, and B. Schölkopf. (2024). “Can Large Language Models Infer Causation from Correlation?” In: *The Twelfth International Conference on Learning Representations*.

Jiralerspong, T., X. Chen, Y. More, V. Shah, and Y. Bengio. (2024). “Efficient causal graph discovery using large language models”. *arXiv preprint arXiv:2402.01207*.

Jordan, M. I. and T. M. Mitchell. (2015). “Machine learning: Trends, perspectives, and prospects”. *Science*.

Jorge Aranda, P. de, A. Bibi, R. Volpi, A. Sanyal, P. Torr, G. Rogez, and P. Dokania. (2022). “Make some noise: Reliable and efficient single-step adversarial training”. In: *NeurIPS*.

Kadavath, S., T. Conerly, A. Askell, T. Henighan, D. Drain, E. Perez, N. Schiefer, Z. Hatfield-Dodds, N. DasSarma, E. Tran-Johnson, *et al.* (2022). “Language models (mostly) know what they know”. *arXiv preprint arXiv:2207.05221*.

Kahla, M., S. Chen, H. Just, and R. Jia. (2022). “Label-only model inversion attacks via boundary repulsion”. In: *CVPR*.

Kairouz, P., H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, *et al.* (2021). “Advances and open problems in federated learning”. *Foundations and Trends® in Machine Learning*.

Kang, B., S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, and Y. Kalantidis. (2020). "Decoupling Representation and Classifier for Long-Tailed Recognition". In: *ICLR*.

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

Katz-Samuels, J., J. B. Nakhleh, R. Nowak, and Y. Li. (2022). "Training ood detectors in their natural habitats". In: *ICML*.

Khan, S., M. Hayat, S. W. Zamir, J. Shen, and L. Shao. (2019). "Striking the right balance with uncertainty". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 103–112.

Khodak, M., M. Balcan, and A. Talwalkar. (2019). "Adaptive Gradient-Based Meta-Learning Methods". *CoRR*.

Kifer, D., A. Smith, and A. Thakurta. (2012). "Private convex empirical risk minimization and high-dimensional regression". In: *COLT*.

Kim, H., W. Lee, and J. Lee. (2021). "Understanding catastrophic overfitting in single-step adversarial training". In: *AAAI*.

Kim, J., Y. Hur, S. Park, E. Yang, S. J. Hwang, and J. Shin. (2020). "Distribution Aligning Refinery of Pseudo-label for Imbalanced Semi-supervised Learning". In: *NeurIPS*.

Kingma, D. P. (2013). "Auto-encoding variational bayes". *arXiv preprint arXiv:1312.6114*.

Kirchenbauer, J., J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein. (2023). "A watermark for large language models". In: *ICML*.

Kirk, H. R., B. Vidgen, P. Röttger, and S. A. Hale. (2024). "The benefits, risks and bounds of personalizing the alignment of large language models to individuals". *Nature Machine Intelligence*. 6(4): 383–392.

Kıcıman, E., R. Ness, A. Sharma, and C. Tan. (2023). "Causal reasoning and large language models: Opening a new frontier for causality". *arXiv preprint arXiv:2305.00050*.

Koh, P. W., S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, T. Lee, E. David, I. Stavness, W. Guo, B. Earnshaw, I. Haque, S. M. Beery, J. Leskovec, A. Kundaje, E. Pierson, S. Levine, C. Finn, and P. Liang. (2021). "WILDS: A Benchmark of in-the-Wild Distribution Shifts". In: *ICML*.

Kotzias, D., M. Denil, N. De Freitas, and P. Smyth. (2015). "From group to individual labels using deep features". In: *SIGKDD*.

Krueger, D., E. Caballero, J. Jacobsen, A. Zhang, J. Binas, D. Zhang, R. L. Priol, and A. C. Courville. (2021). "Out-of-Distribution Generalization via Risk Extrapolation (REx)". In: *ICML*.

Kukleva, A., M. Böhle, B. Schiele, H. Kuehne, and C. Rupprecht. (2023). "Temperature Schedules for self-supervised contrastive methods on long-tail data". In: *ICLR*.

Kukreja, S., T. Kumar, A. Purohit, A. Dasgupta, and D. Guha. (2024). "A literature survey on open source large language models". In: *Proceedings of the 2024 7th International Conference on Computers in Management and Business*. 133–143.

Le, H. D., X. Xia, and Z. Chen. (2024). "Multi-Agent Causal Discovery Using Large Language Models". *arXiv preprint arXiv:2407.15073*.

LeCun, Y., Y. Bengio, and G. Hinton. (2015). "Deep learning". *Nature*.

Lecuyer, M., V. Atlidakis, R. Geambasu, D. Hsu, and S. Jana. (2019). "Certified robustness to adversarial examples with differential privacy". In: *SP*.

Lederer, I., R. Mayer, and A. Rauber. (2023). "Identifying appropriate intellectual property protection mechanisms for machine learning models: A systematization of watermarking, fingerprinting, model access, and attacks". *IEEE Transactions on Neural Networks and Learning Systems*.

Lee, H., S. Shin, and H. Kim. (2021). "ABC: Auxiliary Balanced Classifier for Class-imbalanced Semi-supervised Learning". In: *NeurIPS*.

Lee, J. and D. Kifer. (2018). "Concentrated differentially private gradient descent with adaptive per-iteration privacy budget". In: *KDD*.

Lee, K., K. Lee, H. Lee, and J. Shin. (2018). "A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks". In: *NeurIPS*.

Lee, M. and D. Kim. (2023). "Robust Evaluation of Diffusion-Based Adversarial Purification". In: *ICCV*.

Lee, N., W. Ping, P. Xu, M. Patwary, P. N. Fung, M. Shoeybi, and B. Catanzaro. (2022). "Factuality enhanced language models for open-ended text generation". *Advances in Neural Information Processing Systems*. 35: 34586–34599.

Leng, S., H. Zhang, G. Chen, X. Li, S. Lu, C. Miao, and L. Bing. (2024). “Mitigating object hallucinations in large vision-language models through visual contrastive decoding”. In: *CVPR*.

Levinstein, B. A. and D. A. Herrmann. (2024). “Still no lie detector for language models: Probing empirical and conceptual roadblocks”. *Philosophical Studies*: 1–27.

Li, A., J. Sun, B. Wang, L. Duan, S. Li, Y. Chen, and H. Li. (2020a). “LotteryFL: Personalized and Communication-Efficient Federated Learning with Lottery Ticket Hypothesis on Non-IID Datasets”.

Li, B. and Y. Li. (2023). “Towards Understanding Clean Generalization and Robust Overfitting in Adversarial Training”. In: *arXiv preprint arXiv:2306.01271*.

Li, D., X. Li, Z. Gan, Q. Li, B. Qu, and J. Wang. (2024a). “Rethinking the impact of noisy labels in graph classification: A utility and privacy perspective”. *Neural Networks*.

Li, H., M. Xu, and Y. Song. (2023a). “Sentence Embedding Leaks More Information than You Expect: Generative Embedding Inversion Attack to Recover the Whole Sentence”. In: *ACL*.

Li, J., D. Li, C. Xiong, and S. Hoi. (2022a). “Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation”. In: *International conference on machine learning*. PMLR. 12888–12900.

Li, J., R. Socher, and S. C. H. Hoi. (2020b). “DivideMix: Learning with Noisy Labels as Semi-supervised Learning”. In: *ICLR*.

Li, K., O. Patel, F. Viégas, H. Pfister, and M. Wattenberg. (2024b). “Inference-time intervention: Eliciting truthful answers from a language model”. *Advances in Neural Information Processing Systems*. 36.

Li, L. and M. W. Spratling. (2023). “Data augmentation alone can improve adversarial training”. In: *ICLR*.

Li, P., X. Wang, Z. Zhang, Y. Meng, F. Shen, Y. Li, J. Wang, Y. Li, and W. Zhu. (2024c). “RealTCD: Temporal Causal Discovery from Interventional Data with Large Language Model”. *arXiv preprint arXiv:2404.14786v2*.

Li, Q., B. He, and D. Song. (2020c). “Practical One-Shot Federated Learning for Cross-Silo Setting”. *arXiv preprint arXiv:2010.01017*.

Li, Q., B. He, and D. Song. (2021a). "Model-Contrastive Federated Learning".

Li, S., H. Liu, T. Dong, B. Z. H. Zhao, M. Xue, H. Zhu, and J. Lu. (2021b). "Hidden backdoors in human-centric language models". In: *ACM SIGSAC*.

Li, T., A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith. (2020d). "Federated Optimization in Heterogeneous Networks".

Li, T., M. Sanjabi, A. Beirami, and V. Smith. (2020e). "Fair Resource Allocation in Federated Learning". In: *International Conference on Learning Representations*.

Li, X., Q. Li, D. Li, H. Qian, and J. Wang. (2024d). "Contrastive learning of graphs under label noise". *Neural Networks*.

Li, Y., X. Tian, M. Gong, Y. Liu, T. Liu, K. Zhang, and D. Tao. (2018). "Deep Domain Generalization via Conditional Invariant Adversarial Networks". In: *ECCV*.

Li, Y., J. Yin, and L. Chen. (2021c). "Unified robust training for graph neural networks against label noise". In: *PAKDD*.

Li, Y., Y. Jiang, Z. Li, and S.-T. Xia. (2022b). "Backdoor learning: A survey". *IEEE TNNLS*.

Li, Y., S. Bubeck, R. Eldan, A. Del Giorno, S. Gunasekar, and Y. T. Lee. (2023b). "Textbooks are all you need ii: phi-1.5 technical report". *arXiv preprint arXiv:2309.05463*.

Liang, P. P., T. Liu, Z. Liu, R. Salakhutdinov, and L. Morency. (2020). "Think Locally, Act Globally: Federated Learning with Local and Global Representations". *CoRR*.

Liao, F., M. Liang, Y. Dong, T. Pang, X. Hu, and J. Zhu. (2018). "Defense against adversarial attacks using high-level representation guided denoiser". In: *CVPR*.

Lim, W. Y. B., N. C. Luong, D. T. Hoang, Y. Jiao, Y.-C. Liang, Q. Yang, D. Niyato, and C. Miao. (2020). "Federated Learning in Mobile Edge Networks: A Comprehensive Survey". *IEEE Communications Surveys Tutorials*.

Lin, R., C. Yu, B. Han, and T. Liu. (2024a). "On the Over-Memorization During Natural, Robust and Catastrophic Overfitting". In: *ICLR*.

Lin, R., C. Yu, B. Han, H. Su, and T. Liu. (2024b). “Layer-Aware Analysis of Catastrophic Overfitting: Revealing the Pseudo-Robust Shortcut Dependency”. In: *ICML*.

Lin, R., C. Yu, and T. Liu. (2023a). “Eliminating catastrophic overfitting via abnormal adversarial examples regularization”. In: *NeurIPS*.

Lin, T.-Y., P. Goyal, R. Girshick, K. He, and P. Dollár. (2017). “Focal loss for dense object detection”. In: *ICCV*.

Lin, V., L.-P. Morency, and E. Ben-Michael. (2023b). “Text-Transport: Toward Learning Causal Effects of Natural Language”. In: *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Lin, Y., J. Zhang, Z. Huang, J. Liu, Z. Wen, and X. Peng. (2024c). “Multi-granularity correspondence learning from long-term noisy videos”. *arXiv preprint arXiv:2401.16702*.

Lin, Y., L. Tan, Y. HAO, H. N. Wong, H. Dong, W. Zhang, Y. Yang, and T. Zhang. (2024d). “Spurious Feature Diversification Improves Out-of-distribution Generalization”. In: *ICLR*.

Lin, Y., S. Zhu, L. Tan, and P. Cui. (2022). “ZIN: When and How to Learn Invariance Without Environment Partition?” In: *NeurIPS*.

Liu, A., L. Pan, Y. Lu, J. Li, X. Hu, X. Zhang, L. Wen, I. King, H. Xiong, and P. Yu. (2024a). “A survey of text watermarking in the era of large language models”. *ACM Computing Surveys*.

Liu, B., M. Ding, S. Shaham, W. Rahayu, F. Farokhi, and Z. Lin. (2021a). “When machine learning meets privacy: A survey and outlook”. *ACM CSUR*.

Liu, C., M. Salzmann, T. Lin, R. Tomioka, and S. Süsstrunk. (2020a). “On the Loss Landscape of Adversarial Training: Identifying Challenges and How to Overcome Them”. In: *NeurIPS*.

Liu, C., Y. Chen, T. Liu, M. Gong, J. Cheng, B. Han, and K. Zhang. (2024b). “Discovery of the Hidden World with Large Language Models”. In: *Proceedings of the Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Liu, E. Z., B. Haghgoo, A. S. Chen, A. Raghunathan, P. W. Koh, S. Sagawa, P. Liang, and C. Finn. (2021b). “Just Train Twice: Improving Group Robustness without Training Group Information”. In: *ICML*.

Liu, F., Z. Xu, and H. Liu. (2024c). "Adversarial tuning: Defending against jailbreak attacks for llms". *arXiv preprint arXiv:2406.06622*.

Liu, F., K. Lin, L. Li, J. Wang, Y. Yacoob, and L. Wang. (2024d). "Mitigating hallucination in large multi-modal models via robust instruction tuning". In: *ICLR*.

Liu, G., T. Xu, R. Zhang, Z. Wang, C. Wang, and L. Liu. (2023a). "Gradient-leaks: Enabling black-box membership inference attacks against machine learning models". *IEEE TIFS*.

Liu, G., J. Wu, and Z.-H. Zhou. (2012). "Key instance detection in multi-instance learning". In: *ACML*.

Liu, H., W. Xue, Y. Chen, D. Chen, X. Zhao, K. Wang, L. Hou, R. Li, and W. Peng. (2024e). "A survey on hallucination in large vision-language models". *arXiv preprint arXiv:2402.00253*.

Liu, H., C. Li, Q. Wu, and Y. J. Lee. (2024f). "Visual instruction tuning". *Advances in neural information processing systems*. 36.

Liu, H., J. Z. HaoChen, A. Gaidon, and T. Ma. (2021c). "Self-supervised Learning is More Robust to Dataset Imbalance". In: *ICLR*.

Liu, J., Z. Hu, P. Cui, B. Li, and Z. Shen. (2021d). "Heterogeneous Risk Minimization". In: *ICML*.

Liu, K., B. Dolan-Gavitt, and S. Garg. (2018). "Fine-pruning: Defending against backdooring attacks on deep neural networks". In: *RAID*.

Liu, L., Y. Wang, G. Liu, K. Peng, and C. Wang. (2022a). "Membership inference attacks against machine learning models via prediction sensitivity". *IEEE TDSC*.

Liu, L., J. Zhang, S. Song, and K. B. Letaief. (2020b). "Client-Edge-Cloud Hierarchical Federated Learning". In: *2020 IEEE International Conference on Communications, ICC 2020, Dublin, Ireland, June 7-11, 2020*.

Liu, R., W. Zhou, J. Zhang, X. Liu, P. Si, and H. Li. (2023b). "Model Inversion Attacks on Homogeneous and Heterogeneous Graph Neural Networks". In: *SecureComm*.

Liu, S., J. Niles-Weed, N. Razavian, and C. Fernandez-Granda. (2020c). "Early-learning regularization prevents memorization of noisy labels". *Advances in neural information processing systems*. 33: 20331–20342.

Liu, S., Z. Zhu, Q. Qu, and C. You. (2022b). "Robust training under label noise by over-parameterization". In: *International Conference on Machine Learning*. PMLR. 14153–14172.

Liu, W., X. Wang, J. D. Owens, and Y. Li. (2020d). "Energy-based Out-of-distribution Detection". In: *NeurIPS*.

Liu, X., N. Xu, M. Chen, and C. Xiao. (2024g). "AutoDAN: Generating Stealthy Jailbreak Prompts on Aligned Large Language Models". In: *ICLR*.

Liu, Z., Z. Wang, L. Xu, J. Wang, L. Song, T. Wang, C. Chen, W. Cheng, and J. Bian. (2024h). "Protecting your llms with information bottleneck". In: *NeurIPS*.

Liu, Z., Z. Miao, X. Zhan, J. Wang, B. Gong, and S. X. Yu. (2019). "Large-scale long-tailed recognition in an open world". In: *CVPR*.

Locatello, F., S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem. (2019). "Challenging common assumptions in the unsupervised learning of disentangled representations". In: *international conference on machine learning*.

Lu, Y., Y. Zhang, B. Han, Y. Cheung, and H. Wang. (2023). "Label-Noise Learning with Intrinsically Long-Tailed Data". In: *ICCV*.

Luo, J., F. Hong, J. Yao, B. Han, Y. Zhang, and Y. Wang. (2024). "Revive Re-weighting in Imbalanced Learning by Density Ratio Estimation". In: *NeurIPS*.

Lv, F., J. Liang, S. Li, B. Zang, C. H. Liu, Z. Wang, and D. Liu. (2022). "Causality inspired representation learning for domain generalization". In: *CVPR*.

Ma, X., B. Li, Y. Wang, S. M. Erfani, S. N. R. Wijewickrema, G. Schoenebeck, D. Song, M. E. Houle, and J. Bailey. (2018). "Characterizing Adversarial Subspaces Using Local Intrinsic Dimensionality". In: *ICLR*.

Ma, X., M. Yang, Y. Li, P. Hu, J. Lv, and X. Peng. (2024). "Cross-modal Retrieval with Noisy Correspondence via Consistency Refining and Mining". *IEEE Transactions on Image Processing*.

Maaz, M., H. Rasheed, S. Khan, and F. S. Khan. (2023). "Video-chatgpt: Towards detailed video understanding via large vision and language models". *arXiv preprint arXiv:2306.05424*.

Madry, A., A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. (2018). "Towards Deep Learning Models Resistant to Adversarial Attacks". In: *ICLR*.

Mahajan, D., S. Tople, and A. Sharma. (2021). "Domain Generalization using Causal Matching". In: *ICML*. Vol. 139.

Mayilvahanan, P., T. Wiedemer, E. Rusak, M. Bethge, and W. Brendel. (2024). "Does CLIP's generalization performance mainly stem from high train-test similarity?" In: *ICLR*.

Mehrabi, N., F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. (2021). "A survey on bias and fairness in machine learning". *ACM Computing Surveys (CSUR)*. 54(6): 1–35.

Meng, D. and H. Chen. (2017). "Magnet: a two-pronged defense against adversarial examples". In: *ACM SIGSAC*.

Menon, A. K., S. Jayasumana, A. S. Rawat, H. Jain, A. Veit, and S. Kumar. (2020). "Long-tail learning via logit adjustment". *arXiv preprint arXiv:2007.07314*.

Miao, S., M. Liu, and P. Li. (2022). "Interpretable and Generalizable Graph Learning via Stochastic Attention Mechanism". In: *ICML*.

Min, R., S. Li, H. Chen, and M. Cheng. (2024). "A watermark-conditioned diffusion model for ip protection". *arXiv preprint arXiv:2403.10893*.

Ming, Y., Z. Cai, J. Gu, Y. Sun, W. Li, and Y. Li. (2022a). "Delving into Out-of-Distribution Detection with Vision-Language Representations". In: *NeurIPS*.

Ming, Y., Y. Fan, and Y. Li. (2022b). "Poem: Out-of-distribution detection with posterior sampling". In: *ICML*.

Miotto, R., F. Wang, S. Wang, X. Jiang, and J. T. Dudley. (2018). "Deep learning for healthcare: review, opportunities and challenges". *Briefings in bioinformatics*.

Mirza, M. and S. Osindero. (2014). "Conditional generative adversarial nets". *arXiv preprint arXiv:1411.1784*.

Miyato, T., S. Maeda, M. Koyama, and S. Ishii. (2019). "Virtual Adversarial Training: A Regularization Method for Supervised and Semi-Supervised Learning". *IEEE Trans. Pattern Anal. Mach. Intell.* 41(8): 1979–1993.

Morris, J., V. Kuleshov, V. Shmatikov, and A. Rush. (2023). "Text embeddings reveal (almost) as much as text". In: *EMNLP*.

Muñoz-González, L., B. Biggio, A. Demontis, A. Paudice, V. Wongrassamee, E. C. Lupu, and F. Roli. (2017). "Towards poisoning of deep learning algorithms with back-gradient optimization". In: *AISec*.

Murdoch, W. J., C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu. (2019). "Definitions, methods, and applications in interpretable machine learning". *Proceedings of the National Academy of Sciences*. 116(44): 22071–22080.

Namkoong, H. and J. C. Duchi. (2016). "Stochastic Gradient Methods for Distributionally Robust Optimization with f-divergences". In: *NeurIPS*.

Natarajan, N., I. S. Dhillon, P. Ravikumar, and A. Tewari. (2013). "Learning with Noisy Labels". In: *NeurIPS*.

Neel, S., A. Roth, G. Vietri, and S. Wu. (2020). "Oracle efficient private non-convex optimization". In: *ICML*.

Nelson, B., M. Barreno, F. J. Chi, A. D. Joseph, B. I. Rubinstein, U. Saini, C. Sutton, J. D. Tygar, and K. Xia. (2008). "Exploiting machine learning to subvert your spam filter." *LEET*.

Nguyen, B., K. Chandrasegaran, M. Abdollahzadeh, and N. Cheung. (2024). "Label-Only Model Inversion Attacks via Knowledge Transfer". In: *NeurIPS*.

Nguyen, N., K. Chandrasegaran, M. Abdollahzadeh, and N. Cheung. (2023). "Re-thinking model inversion attacks against deep neural networks". In: *CVPR*.

Nguyen, T. T., T. T. Huynh, Z. Ren, P. L. Nguyen, A. W.-C. Liew, H. Yin, and Q. V. H. Nguyen. (2022). "A survey of machine unlearning". *arXiv preprint arXiv:2209.02299*.

Nie, J., Y. Zhang, Z. Fang, T. Liu, B. Han, and X. Tian. (2024). "Out-of-Distribution Detection with Negative Prompts". In: *ICLR*.

Nie, W., B. Guo, Y. Huang, C. Xiao, A. Vahdat, and A. Anandkumar. (2022). "Diffusion Models for Adversarial Purification". In: *ICML*.

Nishio, T. and R. Yonetani. (2019). "Client Selection for Federated Learning with Heterogeneous Resources in Mobile Edge". In: *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*.

Niu, J., P. Liu, X. Zhu, K. Shen, Y. Wang, H. Chi, Y. Shen, X. Jiang, J. Ma, and Y. Zhang. (2024). "A survey on membership inference attacks and defenses in Machine Learning". *Journal of Information and Intelligence*.

Niu, S., Y. Liu, J. Wang, and H. Song. (2020). "A decade survey of transfer learning (2010–2020)". *IEEE Transactions on Artificial Intelligence*.

NT, H., C. J. Jin, and T. Murata. (2019). "Learning graph neural networks with noisy labels". In: *ICLR Learning from Limited Labeled Data Workshop*.

Olatunji, I., M. Rathee, T. Funke, and M. Khosla. (2023). "Private graph extraction via feature explanations". *PETS*.

Oliynyk, D., R. Mayer, and A. Rauber. (2023). "I know what you trained last summer: A survey on stealing machine learning models and defences". *ACM Computing Surveys*.

Oord, A. van den, Y. Li, and O. Vinyals. (2018). "Representation learning with contrastive predictive coding". *arXiv preprint arXiv:1807.03748*.

Ozbayoglu, A. M., M. U. Gudelek, and O. B. Sezer. (2020). "Deep learning for financial applications: A survey". *Applied soft computing*.

Pang, T., H. Zhang, D. He, Y. Dong, H. Su, W. Chen, J. Zhu, and T. Liu. (2022). "Two Coupled Rejection Metrics Can Tell Adversarial Examples Apart". In: *CVPR*.

Papyan, V., X. Han, and D. L. Donoho. (2020). "Prevalence of neural collapse during the terminal phase of deep learning training". *Proceedings of the National Academy of Sciences*. 117(40): 24652–24663.

Parascandolo, G., A. Neitz, A. Orvieto, L. Gresele, and B. Schölkopf. (2021). "Learning explanations that are hard to vary". In: *ICLR*.

Parashar, S., Z. Lin, T. Liu, X. Dong, Y. Li, D. Ramanan, J. Caverlee, and S. Kong. (2024). "The Neglected Tails in Vision-Language Models". In: *CVPR*.

Parikh, R., C. Dupuy, and R. Gupta. (2022). "Canary extraction in natural language understanding models". *arXiv preprint arXiv:2203.13920*.

Park, Y., D.-J. Han, D.-Y. Kim, J. Seo, and J. Moon. (2021). "Few-round learning for federated learning". *Advances in Neural Information Processing Systems*.

Penedo, G., Q. Malartic, D. Hesslow, R. Cojocaru, H. Alobeidli, A. Cappelli, B. Pannier, E. Almazrouei, and J. Launay. (2023). "The RefinedWeb dataset for Falcon LLM: Outperforming curated corpora with web data only". *Advances in Neural Information Processing Systems*. 36: 79155–79172.

Peng, B., S. Qu, Y. Wu, T. Zou, L. He, A. Knoll, G. Chen, and C. Jiang. (2024a). "MAP: MAsk-Pruning for Source-Free Model Intellectual Property Protection". In: *CVPR*.

Peng, M. and Q. Zhang. (2019). "Address instance-level label prediction in multiple instance learning". *arXiv preprint arXiv:1905.12226*.

Peng, X., B. Han, F. Liu, T. Liu, and M. Zhou. (2024b). "Pseudo-Private Data Guided Model Inversion Attacks". In: *NeurIPS*.

Pessach, D. and E. Shmueli. (2022). "A review on fairness in machine learning". *ACM Computing Surveys (CSUR)*. 55(3): 1–44.

Peters, J., P. Bühlmann, and N. Meinshausen. (2016). "Causal inference by using invariant prediction: identification and confidence intervals". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Peters, J., D. Janzing, and B. Schlkopf. (2017a). *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press.

Peters, J., D. Janzing, and B. Schölkopf. (2017b). *Elements of causal inference: foundations and learning algorithms*. The MIT Press.

Pezeshki, M., D. Bouchacourt, M. Ibrahim, N. Ballas, P. Vincent, and D. Lopez-Paz. (2024). "Discovering environments with XRM". In: *ICML*.

Phan, N., M. Vu, Y. Liu, R. Jin, D. Dou, X. Wu, and M. T. Thai. (2019). "Heterogeneous gaussian mechanism: Preserving differential privacy in deep learning with provable robustness". In: *IJCAI*.

Phan, N., Y. Wang, X. Wu, and D. Dou. (2016). "Differential privacy preservation for deep auto-encoders: an application of human behavior prediction". In: *AAAI*.

Qi, X., K. Huang, A. Panda, M. Wang, and P. Mittal. (2023). “Visual adversarial examples jailbreak large language models”. *arXiv preprint arXiv:2306.13213*.

Qi, X., Y. Zeng, T. Xie, P.-Y. Chen, R. Jia, P. Mittal, and P. Henderson. (2024). “Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!” In: *ICLR*.

Qiang, Y., X. Zhou, S. Z. Zade, M. A. Roshani, P. Khanduri, D. Zytko, and D. Zhu. (2024). “Learning to poison large language models during instruction tuning”. *arXiv preprint arXiv:2402.13459*.

Qin, Y., D. Peng, X. Peng, X. Wang, and P. Hu. (2022). “Deep evidential learning with noisy correspondence for cross-modal retrieval”. In: *Proceedings of the 30th ACM International Conference on Multimedia*. 4948–4956.

Qin, Y., Y. Sun, D. Peng, J. T. Zhou, X. Peng, and P. Hu. (2023). “Cross-modal active complementary learning with self-refining correspondence”. *Advances in Neural Information Processing Systems*. 36: 24829–24840.

Qu, Z., X. Li, R. Duan, Y. Liu, B. Tang, and Z. Lu. (2022). “Generalized federated learning via sharpness aware minimization”. In: *International conference on machine learning*.

Radford, A., J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.* (2021a). “Learning transferable visual models from natural language supervision”. In: *International conference on machine learning*. PMLR. 8748–8763.

Radford, A., J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.* (2021b). “Learning transferable visual models from natural language supervision”. In: *ICML*.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.* (2019). “Language models are unsupervised multitask learners”. *OpenAI blog*. 1(8): 9.

Raghuram, J., V. Chandrasekaran, S. Jha, and S. Banerjee. (2021). “A General Framework For Detecting Anomalous Inputs to DNN Classifiers”. In: *ICML*.

Rame, A., C. Dancette, and M. Cord. (2022a). “Fishr: Invariant Gradient Variances for Out-of-distribution Generalization”. In: *ICML*.

Rame, A., M. Kirchmeyer, T. Rahier, A. Rakotomamonjy, patrick gallinari, and M. Cord. (2022b). "Diverse Weight Averaging for Out-of-Distribution Generalization". In: *NeurIPS*.

Rawte, V., A. Sheth, and A. Das. (2023). "A survey of hallucination in large foundation models". *arXiv preprint arXiv:2309.05922*.

Reddi, S., Z. Charles, M. Zaheer, Z. Garrett, K. Rush, J. Konečnỳ, S. Kumar, and H. B. McMahan. (2020). "Adaptive Federated Optimization". *arXiv preprint arXiv:2003.00295*.

Reed, W. J. (2001). "The Pareto, Zipf and other power laws". *Economics letters*.

Ren, J., G. Yu, and G. Ding. (2021). "Accelerating DNN Training in Wireless Federated Edge Learning Systems". *IEEE Journal on Selected Areas in Communications*.

Ren, R., Y. Wang, Y. Qu, W. X. Zhao, J. Liu, H. Tian, H. Wu, J.-R. Wen, and H. Wang. (2023). "Investigating the factual knowledge boundary of large language models with retrieval augmentation". *arXiv preprint arXiv:2307.11019*.

Rice, L., E. Wong, and Z. Kolter. (2020). "Overfitting in adversarially robust deep learning". In: *ICML*.

Robey, A., E. Wong, H. Hassani, and G. J. Pappas. (2023). "Smoothllm: Defending large language models against jailbreaking attacks". *arXiv preprint arXiv:2310.03684*.

Rojas-Carulla, M., B. Schölkopf, R. Turner, and J. Peters. (2018). "Invariant Models for Causal Transfer Learning". *Journal of Machine Learning Research*.

Rombach, R., A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. (2022). "High-resolution image synthesis with latent diffusion models". In: *CVPR*.

Rosenfeld, E., P. Ravikumar, and A. Risteski. (2022). "Domain-Adjusted Regression or: ERM May Already Learn Features Sufficient for Out-of-Distribution Generalization". *arXiv preprint arXiv:2202.06856*.

Sadeghi, B., S. Dehdashtian, and V. Boddeti. (2022). "On Characterizing the Trade-off in Invariant Representation Learning". *TMLR*.

Sagawa, S., P. W. Koh, T. B. Hashimoto, and P. Liang. (2020). "Distributionally Robust Neural Networks". In: *ICLR*.

Saha, A., A. Subramanya, and H. Pirsiavash. (2020). "Hidden trigger backdoor attacks". In: *AAAI*.

Samangouei, P., M. Kabkab, and R. Chellappa. (2018). "Defense-GAN: Protecting Classifiers Against Adversarial Attacks Using Generative Models". In: *ICLR*.

Santurkar, S., Y. Dubois, R. Taori, P. Liang, and T. Hashimoto. (2023). "Is a Caption Worth a Thousand Images? A Study on Representation Learning". In: *ICLR*.

Schmidt, L., S. Santurkar, D. Tsipras, K. Talwar, and A. Madry. (2018). "Adversarially robust generalization requires more data". In: *NeurIPS*.

Schneider, S., A. Baevski, R. Collobert, and M. Auli. (2019). "wav2vec: Unsupervised pre-training for speech recognition". *arXiv preprint arXiv:1904.05862*.

Schölkopf, B., F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. (2021). "Toward causal representation learning". *Proceedings of the IEEE*.

Schryen, G. and R. Kadura. (2009). "Open source vs. closed source software: towards measuring security". In: *Proceedings of the 2009 ACM symposium on Applied Computing*. 2016–2023.

Schulman, J. (2023). "Reinforcement learning from human feedback: Progress and challenges". In: *Berkeley EECS Colloquium. YouTube www. youtube. com/watch*.

Schwartz, I. S., K. E. Link, R. Daneshjou, and N. Cortés-Penfield. (2024). "Black box warning: large language models and the future of infectious diseases consultation". *Clinical infectious diseases*. 78(4): 860–866.

Sehwag, V., M. Chiang, and P. Mittal. (2021). "SSD: A Unified Framework for Self-Supervised Outlier Detection". In: *ICLR*.

Shafahi, A., M. Najibi, A. Ghiasi, Z. Xu, J. P. Dickerson, C. Studer, L. S. Davis, G. Taylor, and T. Goldstein. (2019). "Adversarial training for free!" In: *NeurIPS*.

Shaik, T., X. Tao, H. Xie, L. Li, X. Zhu, and Q. Li. (2023). "Exploring the Landscape of Machine Unlearning: A Survey and Taxonomy". *arXiv preprint arXiv:2305.06360*.

Sharma, P., N. Ding, S. Goodman, and R. Soricut. (2018). "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2556–2565.

Shen, L., Z. Tang, L. Wu, Y. Zhang, X. Chu, T. Qin, and B. Han. (2024). "Hot Pluggable Federated Learning". In: *International Workshop on Federated Foundation Models in Conjunction with NeurIPS 2024*.

Shen, Y., Y. Han, Z. Zhang, M. Chen, T. Yu, M. Backes, Y. Zhang, and G. Stringhini. (2022). "Finding mnemon: Reviving memories of node embeddings". In: *CCS*.

Shi, C., C. Holtz, and G. Mishne. (2021). "Online Adversarial Purification based on Self-supervised Learning". In: *ICLR*.

Shi, W., S. Min, M. Yasunaga, M. Seo, R. James, M. Lewis, L. Zettlemoyer, and W.-t. Yih. (2023). "Replug: Retrieval-augmented black-box language models". *arXiv preprint arXiv:2301.12652*.

Shi, Y., J. Seely, P. Torr, S. N, A. Hannun, N. Usunier, and G. Synnaeve. (2022). "Gradient Matching for Domain Generalization". In: *ICLR*.

Shibly, K. H., M. D. Hossain, H. Inoue, Y. Taenaka, and Y. Kadobayashi. (2023). "Towards Autonomous Driving Model Resistant to Adversarial Attack". *Appl. Artif. Intell.*

Shokri, R., M. Stronati, C. Song, and V. Shmatikov. (2017). "Membership inference attacks against machine learning models". In: *SP*.

Singh, A., T.-W. Ngan, P. Druschel, and D. Wallach. (2006). "Eclipse Attacks on Overlay Networks: Threats and Defenses". In: *IEEE INFOCOM*.

Smith, V., C.-K. Chiang, M. Sanjabi, and A. Talwalkar. (2018). "Federated Multi-Task Learning".

Sohn, K., D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li. (2020). "Fixmatch: Simplifying semi-supervised learning with consistency and confidence". In: *NeurIPS*.

Song, C. and A. Raghunathan. (2020). "Information leakage in embedding models". In: *CCS*.

Song, H., M. Kim, D. Park, Y. Shin, and J.-G. Lee. (2022a). "Learning from noisy labels with deep neural networks: A survey". *IEEE TNNLS*.

Song, J. and D. Namiot. (2022). "A survey of the implementations of model inversion attacks". In: *DCCN*.

Song, S., K. Chaudhuri, and A. D. Sarwate. (2013). "Stochastic gradient descent with differentially private updates". In: *GlobalSIP*.

Song, Y. and S. Ermon. (2019). "Generative Modeling by Estimating Gradients of the Data Distribution". In: *NeurIPS*.

Song, Y., N. Sebe, and W. Wang. (2022b). "Rankfeat: Rank-1 feature removal for out-of-distribution detection". In:

Spirtes, P., C. Glymour, and R. Scheines. (2001). *Causation, prediction, and search*. MIT press.

Sprague, M. R., A. Jalalirad, M. Scavuzzo, C. Capota, M. Neun, L. Do, and M. Kopp. (2019). "Asynchronous Federated Learning for Geospatial Applications". In: *ECML PKDD 2018 Workshops*.

Strohmer, T. and R. W. Heath Jr. (2003). "Grassmannian frames with applications to coding and communication". *Applied and computational harmonic analysis*. 14(3): 257–275.

Struppek, L., D. Hintersdorf, A. Correia, A. Adler, and K. Kersting. (2022). "Plug and Play Attacks: Towards Robust and Flexible Model Inversion Attacks". In: *ICML*.

Stutz, D., M. Hein, and B. Schiele. (2020). "Confidence-Calibrated Adversarial Training: Generalizing to Unseen Attacks". In: *ICML*.

Sun, B. and K. Saenko. (2016). "Deep CORAL: Correlation Alignment for Deep Domain Adaptation". In: *ECCV*.

Sun, T., X. Zhang, Z. He, P. Li, Q. Cheng, X. Liu, H. Yan, Y. Shao, Q. Tang, S. Zhang, *et al.* (2024). "MOSS: An Open Conversational Large Language Model". *Machine Intelligence Research*: 1–18.

Sun, W., S. Lei, L. Wang, Z. Liu, and Y. Zhang. (2020). "Adaptive Federated Learning and Digital Twin for Industrial Internet of Things". *IEEE Transactions on Industrial Informatics*.

Sun, Y., C. Guo, and Y. Li. (2021). "ReAct: Out-of-distribution Detection With Rectified Activations". In: *NeurIPS*.

Sun, Y. and Y. Li. (2022). "DICE: Leveraging Sparsification for Out-of-Distribution Detection". In: *ECCV*.

Sun, Y., Y. Ming, X. Zhu, and Y. Li. (2022a). "Out-of-distribution Detection with Deep Nearest Neighbors". *ICML*.

Sun, Z., X. Du, F. Song, M. Ni, and L. Li. (2022b). "Coprotector: Protect open-source code against unauthorized training usage with data poisoning". In: *WWW*.

Szegedy, C., W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus. (2014). "Intriguing properties of neural networks". In: *ICLR*.

Tack, J., S. Mo, J. Jeong, and J. Shin. (2020). "CSI: Novelty Detection via Contrastive Learning on Distributionally Shifted Instances". In: *NeurIPS*.

Tan, X., L. Yong, S. Zhu, C. Qu, X. Qiu, X. Yinghui, P. Cui, and Y. Qi. (2023). "Provably Invariant Learning without Domain Information". In: *ICML*.

Tang, Z., X. Chu, R. Y. Ran, S. Lee, S. Shi, Y. Zhang, Y. Wang, A. Q. Liang, S. Avestimehr, and C. He. (2023). "FedML Parrot: A Scalable Federated Learning System via Heterogeneity-aware Scheduling on Sequential and Hierarchical Training". *arXiv preprint arXiv:2303.01778*.

Tang, Z., X. Kang, Y. Yin, X. Pan, Y. Wang, X. He, Q. Wang, R. Zeng, K. Zhao, S. Shi, A. C. Zhou, B. Li, B. He, and X. Chu. (2024a). "FusionLLM: A Decentralized LLM Training System on Geo-distributed GPUs with Adaptive Compression".

Tang, Z., S. Shi, and X. Chu. (2020a). "Communication-efficient decentralized learning with sparsification and adaptive peer selection". In: *2020 IEEE 40th International Conference on Distributed Computing Systems (ICDCS)*.

Tang, Z., S. Shi, X. Chu, W. Wang, and B. Li. (2020b). "Communication-efficient distributed deep learning: A comprehensive survey". *arXiv preprint arXiv:2003.06307*.

Tang, Z., S. Shi, B. Li, and X. Chu. (2022). "GossipFL: A Decentralized Federated Learning Framework with Sparsified and Adaptive Communication". *IEEE Transactions on Parallel and Distributed Systems*.

Tang, Z., Y. Zhang, P. Dong, Y.-m. Cheung, A. C. Zhou, B. Han, and X. Chu. (2024b). "FuseFL: One-Shot Federated Learning through the Lens of Causality with Progressive Model Fusion". In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Tang, Z., J. Huang, R. Yan, Y. Wang, Z. Tang, S. Shi, A. C. Zhou, and X. Chu. (2024c). "Bandwidth-Aware and Overlap-Weighted Compression for Communication-Efficient Federated Learning". In: *53rd International Conference on Parallel Processing*.

Tao, Z. and Q. Li. (2018). "eSGD: Communication Efficient Distributed Deep Learning on the Edge". In: *USENIX Workshop on Hot Topics in Edge Computing (HotEdge 18)*.

Team, G., P. Georgiev, V. I. Lei, R. Burnell, L. Bai, A. Gulati, G. Tanzer, D. Vincent, Z. Pan, S. Wang, *et al.* (2024). "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context". *arXiv preprint arXiv:2403.05530*.

Teney, D., Y. Lin, S. J. Oh, and E. Abbasnejad. (2023). "ID and OOD Performance Are Sometimes Inversely Correlated on Real-world Datasets". In: *NeurIPS*.

Thudi, A., G. Deza, V. Chandrasekaran, and N. Papernot. (2022a). "Unrolling sgd: Understanding factors influencing machine unlearning". In: *IEEE EuroS&P*.

Thudi, A., H. Jia, I. Shumailov, and N. Papernot. (2022b). "On the necessity of auditable algorithmic definitions for machine unlearning". In: *USENIX Security*.

Tian, Y., O. J. Henaff, and A. van den Oord. (2021). "Divide and contrast: Self-supervised learning from uncurated data". In: *ICCV*.

Tolpegin, V., S. Truex, M. E. Gursoy, and L. Liu. (2020). "Data poisoning attacks against federated learning systems". In: *ESORICs*.

Touvron, H., T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, *et al.* (2023). "Llama: Open and efficient foundation language models". *arXiv preprint arXiv:2302.13971*.

Uchida, Y., Y. Nagai, S. Sakazawa, and S. Satoh. (2017). "Embedding watermarks into deep neural networks". In: *ICMR*.

Van Horn, G., O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie. (2018). "The inaturalist species classification and detection dataset". In: *CVPR*.

Vapnik, V. (1991). "Principles of Risk Minimization for Learning Theory". In: *NIPS*.

Vashishtha, A., A. G. Reddy, A. Kumar, S. Bachu, V. N. Balasubramanian, and A. Sharma. (2023). "Causal inference using llm-guided discovery". *arXiv preprint arXiv:2310.15117*.

Verma, A., S. Krishna, S. Gehrmann, M. Seshadri, A. Pradhan, T. Ault, L. Barrett, D. Rabinowitz, J. Doucette, and N. Phan. (2024). "Operationalizing a threat model for red-teaming large language models (llms)". *arXiv preprint arXiv:2407.14937*.

Wald, Y., A. Feder, D. Greenfeld, and U. Shalit. (2021). "On Calibration and Out-of-Domain Generalization". In: *NeurIPS*.

Wang, H., M. Xia, Y. Li, Y. Mao, L. Feng, G. Chen, and J. Zhao. (2022a). "SoLar: Sinkhorn Label Refinery for Imbalanced Partial-Label Learning". In: *NeurIPS*.

Wang, H., Z. Li, L. Feng, and W. Zhang. (2022b). "ViM: Out-Of-Distribution with Virtual-logit Matching". In: *CVPR*.

Wang, H., H. Chi, W. Yang, Z. Lin, M. Geng, L. Lan, J. Zhang, and D. Tao. (2023a). "Domain Specified Optimization for Deployment Authorization". In: *CVPR*.

Wang, H., Y. Li, H. Yao, and X. Li. (2023b). "CLIPN for Zero-Shot OOD Detection: Teaching CLIP to Say No". *ICCV*.

Wang, J., Y. Wang, G. Xu, J. Zhang, Y. Gu, H. Jia, M. Yan, J. Zhang, and J. Sang. (2023c). "An llm-free multi-dimensional benchmark for mllms hallucination evaluation". *arXiv preprint arXiv:2311.07397*.

Wang, K., Y. Fu, K. Li, A. Khisti, R. Zemel, and A. Makhzani. (2021a). "Variational model inversion attacks". In: *NeurIPS*.

Wang, K., C. Gou, Y. Duan, Y. Lin, X. Zheng, and F.-Y. Wang. (2017). "Generative adversarial networks: introduction and outlook". *IEEE/CAA JAS*.

Wang, L., M. Wang, D. Zhang, and H. Fu. (2023d). "Model Barrier: A Compact Un-Transferable Isolation Domain for Model Intellectual Property Protection". In: *CVPR*.

Wang, L., S. Xu, R. Xu, X. Wang, and Q. Zhu. (2022c). "Non-transferable learning: A new approach for model ownership verification and applicability authorization". In: *ICLR*.

Wang, Q., Z. Fang, Y. Zhang, F. Liu, Y. Li, and B. Han. (2023e). "Learning to augment distributions for out-of-distribution detection". *NeurIPS*.

Wang, Q., B. Han, P. Yang, J. Zhu, T. Liu, and M. Sugiyama. (2024a). "Unlearning with Control: Assessing Real-world Utility for Large Language Model Unlearning". *arXiv preprint arXiv:2406.09179*.

Wang, Q., Y. Lin, Y. Chen, L. Schmidt, B. Han, and T. Zhang. (2024b). "A Sober Look at the Robustness of CLIPs to Spurious Features". In: *NeurIPS*.

Wang, Q., F. Liu, B. Han, T. Liu, C. Gong, G. Niu, M. Zhou, and M. Sugiyama. (2021b). "Probabilistic Margins for Instance Reweighting in Adversarial Training". In: *NeurIPS*.

Wang, Q., J. Yao, C. Gong, T. Liu, M. Gong, H. Yang, and B. Han. (2021c). "Learning with group noise". In: *AAAI*.

Wang, S. and M. Ji. (2022). "A Unified Analysis of Federated Learning with Arbitrary Client Participation". In: *Advances in Neural Information Processing Systems*.

Wang, T., J.-Y. Zhu, A. Torralba, and A. A. Efros. (2020a). "Dataset Distillation".

Wang, X., J. Li, X. Kuang, Y.-a. Tan, and J. Li. (2019). "The security of machine learning in an adversarial setting: A survey". *Journal of Parallel and Distributed Computing*.

Wang, Y., L. Li, J. Yang, Z. Lin, and Y. Wang. (2023f). "Balance, imbalance, and rebalance: Understanding robust overfitting from a minimax game perspective". In: *NeurIPS*.

Wang, Y., D. Zou, J. Yi, J. Bailey, X. Ma, and Q. Gu. (2020b). "Improving Adversarial Robustness Requires Revisiting Misclassified Examples". In: *ICLR*.

Wang, Y., Y. Cao, J. Wu, R. Chen, and J. Chen. (2024c). "Tackling the Data Heterogeneity in Asynchronous Federated Learning with Cached Update Calibration". *International Conference on Learning Representations*.

Wang, Z., L. Shen, T. Liu, T. Duan, Y. Zhu, D. Zhan, D. Doermann, and M. Gao. (2024d). "Defending against Data-Free Model Extraction by Distributionally Robust Defensive Training". In: *NeurIPS*.

Wang, Z., D. Sun, S. Zhou, H. Wang, J. Fan, L. Huang, and J. Bu. (2024e). "NoisyGL: A Comprehensive Benchmark for Graph Neural Networks under Label Noise". In: *NeurIPS*.

Wang, Z., Q. Xu, Z. Yang, Y. He, X. Cao, and Q. Huang. (2024f). "A unified generalization analysis of re-weighting and logit-adjustment for imbalanced learning". *Advances in Neural Information Processing Systems*. 36.

Warnecke, A., L. Pirch, C. Wressnegger, and K. Rieck. (2023). "Machine unlearning of features and labels". *NDSS*.

Wei, H., R. Xie, H. Cheng, L. Feng, B. An, and Y. Li. (2022). "Mitigating neural network overconfidence with logit normalization". In: *ICML*.

Wei, J., Z. Zhu, G. Niu, T. Liu, S. Liu, M. Sugiyama, and Y. Liu. (2023a). "Fairness Improves Learning from Noisily Labeled Long-Tailed Data". arXiv: 2303.12291.

Wei, T., J. Shi, W. Tu, and Y. Li. (2021). "Robust Long-Tailed Learning under Label Noise". arXiv: 2108.11569.

Wei, X., X. Gong, Y. Zhan, B. Du, Y. Luo, and W. Hu. (2023b). "Clnode: Curriculum learning for node classification". In: *WSDM*.

Wei, Z., Y. Wang, A. Li, Y. Mo, and Y. Wang. (2023c). "Jailbreak and guard aligned language models with only few in-context demonstrations". *arXiv preprint arXiv:2310.06387*.

Wen, Y., J. Kirchenbauer, J. Geiping, and T. Goldstein. (2023). "Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust". *arXiv preprint arXiv:2305.20030*.

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

Wu, D., S. Xia, and Y. Wang. (2020). "Adversarial Weight Perturbation Helps Robust Generalization". In: *NeurIPS*.

Wu, J., Y. Gaur, Z. Chen, L. Zhou, Y. Zhu, T. Wang, J. Li, S. Liu, B. Ren, L. Liu, *et al.* (2023). "On decoder-only architecture for speech-to-text and large language model integration". In: *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE. 1–8.

Wu, J., R. Hu, D. Li, Z. Huang, L. Ren, and Y. Zang. (2024). "Robust Heterophilic Graph Learning against Label Noise for Anomaly Detection". In: *IJCAI*.

Wu, T., Z. Liu, Q. Huang, Y. Wang, and D. Lin. (2021). "Adversarial robustness under long-tailed distribution". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8659–8668.

Xiao, H., H. Xiao, and C. Eckert. (2012). "Adversarial label flips attack on support vector machines". In: *ECAI*.

Xie, B., Y. Chen, J. Wang, K. Zhou, B. Han, W. Meng, and J. Cheng. (2024). "Enhancing Evolving Domain Generalization through Dynamic Latent Representations". In: *AAAI*.

Xie, C., K. Huang, P.-Y. Chen, and B. Li. (2019). "Dba: Distributed backdoor attacks against federated learning". In: *ICLR*.

Xie, Y., J. Yi, J. Shao, J. Curl, L. Lyu, Q. Chen, X. Xie, and F. Wu. (2023). "Defending ChatGPT against jailbreak attack via self-reminders". *Nature Machine Intelligence*.

Xie, Y., P. Li, C. Wu, and Q. Wu. (2021). "Differential privacy stochastic gradient descent with adaptive privacy budget allocation". In: *ICCECE*.

Xu, C., Y. Qu, Y. Xiang, and L. Gao. (2021a). "Asynchronous Federated Learning on Heterogeneous Devices: A Survey". *Computer Science Review*.

Xu, H., T. Zhu, L. Zhang, W. Zhou, and P. S. Yu. (2023). "Machine unlearning: A survey". *ACM Computing Surveys*.

Xu, H., L. Xiang, X. Ma, B. Yang, and B. Li. (2024a). "Hufu: A Modality-Agnositc Watermarking System for Pre-Trained Transformers via Permutation Equivariance". *arXiv preprint arXiv:2403.05842*.

Xu, J., F. Wang, M. D. Ma, P. W. Koh, C. Xiao, and M. Chen. (2024b). "Instructional fingerprinting of large language models". *arXiv preprint arXiv:2401.12255*.

Xu, J., H. Wang, and L. Chen. (2021b). "Bandwidth Allocation for Multiple Federated Learning Services in Wireless Edge Networks". *CoRR*.

Xu, X., K. Kong, N. Liu, L. Cui, D. Wang, J. Zhang, and M. Kankanhalli. (2024c). "An LLM can Fool Itself: A Prompt-Based Adversarial Attack". In: *ICLR*.

Xue, M., Y. Zhang, J. Wang, and W. Liu. (2021a). "Intellectual property protection for deep learning models: Taxonomy, methods, attacks, and evaluations". *IEEE Transactions on Artificial Intelligence*.

Xue, Y., C. Niu, Z. Zheng, S. Tang, C. Lyu, F. Wu, and G. Chen. (2021b). "Toward understanding the influence of individual clients in federated learning". In: *AAAI*.

Yan, H., S. Li, Y. Wang, Y. Zhang, K. Sharif, H. Hu, and Y. Li. (2022). "Membership inference attacks against deep learning models via logits distribution". *IEEE TDSC*.

Yang, C., Q. Wu, H. Li, and Y. Chen. (2017). "Generative poisoning attack method against neural networks". *arXiv preprint arXiv:1703.01340*.

Yang, H., X. Zhang, P. Khanduri, and J. Liu. (2022a). "Anarchic Federated Learning". In: *Proceedings of the 39th International Conference on Machine Learning*.

Yang, J., K. Zhou, Y. Li, and Z. Liu. (2024). "Generalized out-of-distribution detection: A survey". *IJCV*.

Yang, S., Z. Xu, K. Wang, Y. You, H. Yao, T. Liu, and M. Xu. (2023a). "Bicro: Noisy correspondence rectification for multi-modality data via bi-directional cross-modal similarity consistency". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19883–19892.

Yang, Y., S. Chen, X. Li, L. Xie, Z. Lin, and D. Tao. (2022b). "Inducing neural collapse in imbalanced learning: Do we really need a learnable classifier at the end of deep neural network?" *Advances in neural information processing systems*. 35: 37991–38002.

Yang, Y., H. Zhang, D. Katabi, and M. Ghassemi. (2023b). "Change is Hard: A Closer Look at Subpopulation Shift". In: *ICML*.

Yang, Z., E. Chang, and Z. Liang. (2019). "Adversarial neural network inversion via auxiliary knowledge alignment". *arXiv preprint arXiv:1902.08552*.

Yao, T., Y. Chen, Z. Chen, K. Hu, Z. Shen, and K. Zhang. (2024a). "Empowering Graph Invariance Learning with Deep Spurious Infomax". In: *ICML*.

Yao, Y., J. Deng, X. Chen, C. Gong, J. Wu, and J. Yang. (2020). "Deep discriminative CNN with temporal ensembling for ambiguously-labeled image classification". In: *AAAI*.

Yao, Y., J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang. (2024b). "A survey on large language model (llm) security and privacy: The good, the bad, and the ugly". *High-Confidence Computing*.

Ye, J., A. Borovykh, S. Hayou, and R. Shokri. (2023). "Leave-one-out distinguishability in machine learning". *arXiv preprint arXiv:2309.17310*.

Ye, J., A. Maddi, S. K. Murakonda, V. Bindschaedler, and R. Shokri. (2022). "Enhanced membership inference attacks against machine learning models". In: *ACM SIGSAC*.

Yeom, S., I. Giacomelli, M. Fredrikson, and S. Jha. (2018). "Privacy risk in machine learning: Analyzing the connection to overfitting". In: *IEEE CSF*.

Yi, S., Y. Liu, Z. Sun, T. Cong, X. He, J. Song, K. Xu, and Q. Li. (2024). "Jailbreak attacks and defenses against large language models: A survey". *arXiv preprint arXiv:2407.04295*.

Yin, X., X. Yu, K. Sohn, X. Liu, and M. Chandraker. (2019). "Feature transfer learning for face recognition with under-represented data". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5704–5713.

Yoon, J., W. Jeong, G. Lee, E. Yang, and S. J. Hwang. (2021a). "Federated Continual Learning with Weighted Inter-client Transfer". In: *Proceedings of the 38th International Conference on Machine Learning*.

Yoon, J., S. J. Hwang, and J. Lee. (2021b). "Adversarial Purification with Score-based Generative Models". In: *ICML*.

Yu, C., B. Han, L. Shen, J. Yu, C. Gong, M. Gong, and T. Liu. (2022). "Understanding robust overfitting of adversarial training and beyond". In: *ICML*.

Yu, D., H. Zhang, W. Chen, J. Yin, and T.-Y. Liu. (2021). "Large scale private learning via low-rank reparametrization". In: *ICML*.

Yu, F. and M.-L. Zhang. (2016). "Maximum margin partial label learning". In: *ACML*.

Yu, Q., J. Li, L. Wei, L. Pang, W. Ye, B. Qin, S. Tang, Q. Tian, and Y. Zhuang. (2024). "Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data". In: *CVPR*.

Yu, T., E. Bagdasaryan, and V. Shmatikov. (2020). "Salvaging federated learning by local adaptation". *arXiv preprint arXiv:2002.04758*.

Yuan, J., X. Luo, Y. Qin, Y. Zhao, W. Ju, and M. Zhang. (2023a). "Learning on graphs under label noise". In: *ICASSP*.

Yuan, X., K. Chen, J. Zhang, W. Zhang, N. Yu, and Y. Zhang. (2023b). "Pseudo label-guided model inversion attack via conditional generative adversarial network". In: *AAAI*.

Yuan, Y., W. Jiao, W. Wang, J.-t. Huang, P. He, S. Shi, and Z. Tu. (2024). "Gpt-4 is too smart to be safe: Stealthy chat with llms via cipher". In: *ICLR*.

Yue, Z., L. Zhang, and Q. Jin. (2024). "Less is more: Mitigating multimodal hallucination from an eos decision perspective". *arXiv preprint arXiv:2402.14545*.

Zarifzadeh, S., P. Liu, and R. Shokri. (2024). "Low-Cost High-Power Membership Inference Attacks". In: *ICML*.

Zbontar, J., L. Jing, I. Misra, Y. LeCun, and S. Deny. (2021). "Barlow twins: Self-supervised learning via redundancy reduction". In: *ICML*.

Zečević, M., M. Willig, D. S. Dhami, and K. Kersting. (2023). "Causal parrots: Large language models may talk causality but are not causal". *arXiv preprint arXiv:2308.13067*.

Zeng, B., L. Wang, Y. Hu, Y. Xu, C. Zhou, X. Wang, Y. Yu, and Z. Lin. (2023). "Huref: Human-readable fingerprint for large language models". In: *NeurIPS*.

Zeng, G. and W. Lu. (2022). "Unsupervised Non-transferable Text Classification". In: *EMNLP*.

Zeng, Y., H. Lin, J. Zhang, D. Yang, R. Jia, and W. Shi. (2024a). "How johnny can persuade llms to jailbreak them: Rethinking persuasion to challenge ai safety by humanizing llms". In: *ACL*.

Zeng, Y., Y. Wu, X. Zhang, H. Wang, and Q. Wu. (2024b). "Autodefense: Multi-agent llm defense against jailbreak attacks". *arXiv preprint arXiv:2403.04783*.

Zhai, R., C. Dan, J. Z. Kolter, and P. K. Ravikumar. (2023). "Understanding Why Generalized Reweighting Does Not Improve Over ERM". In: *ICLR*.

Zhang, C., Y. Xie, H. Bai, B. Yu, W. Li, and Y. Gao. (2021a). "A survey on federated learning". *Knowledge-Based Systems*.

Zhang, C., S. Bengio, M. Hardt, B. Recht, and O. Vinyals. (2021b). "Understanding deep learning (still) requires rethinking generalization". *Communications of the ACM*.

Zhang, E., K. Wang, X. Xu, Z. Wang, and H. Shi. (2023a). "Forget-Me-Not: Learning to Forget in Text-to-Image Diffusion Models". *arXiv preprint arXiv:2211.08332*.

Zhang, H., Y. Yu, J. Jiao, E. P. Xing, L. E. Ghaoui, and M. I. Jordan. (2019). "Theoretically Principled Trade-off between Robustness and Accuracy". In: *ICML*.

Zhang, J., F. Liu, D. Zhou, J. Zhang, and T. Liu. (2024a). "Improving Accuracy-robustness Trade-off via Pixel Reweighted Adversarial Training". In: *ICML*.

Zhang, J., B. Chen, X. Cheng, H. T. T. Binh, and S. Yu. (2020a). "PoisonGAN: Generative poisoning attacks against federated learning in edge computing systems". *IEEE Internet of Things Journal*.

Zhang, J., D. Lopez-Paz, and L. Bottou. (2022a). "Rich Feature Construction for the Optimization-Generalization Dilemma". In:

Zhang, J., C. Chen, B. Li, L. Lyu, S. Wu, S. Ding, C. Shen, and C. Wu. (2022b). "Dense: Data-free one-shot federated learning". *Advances in Neural Information Processing Systems*.

Zhang, J., D. Chen, J. Liao, W. Zhang, H. Feng, G. Hua, and N. Yu. (2021c). "Deep model intellectual property protection via deep watermarking". *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, J., J. Zhu, G. Niu, B. Han, M. Sugiyama, and M. S. Kankanhalli. (2021d). "Geometry-aware Instance-reweighted Adversarial Training". In: *ICLR*.

Zhang, J., Q. Fu, X. Chen, L. Du, Z. Li, G. Wang, S. Han, D. Zhang, *et al.* (2023b). "Out-of-distribution detection based on in-distribution data patterns memorization with modern hopfield energy". In: *ICLR*.

Zhang, M., N. S. Sohoni, H. R. Zhang, C. Finn, and C. Ré. (2022c). "Correct-N-Contrast: a Contrastive Approach for Improving Robustness to Spurious Correlations". In: *ICML*.

Zhang, M.-L., B.-B. Zhou, and X.-Y. Liu. (2016). "Partial label learning via feature-aware disambiguation". In: *SIGKDD*.

Zhang, M., O. Press, W. Merrill, A. Liu, and N. A. Smith. (2023c). "How language model hallucinations can snowball". *arXiv preprint arXiv:2305.13534*.

Zhang, R., S. Hidano, and F. Koushanfar. (2022d). "Text Revealer: Private Text Reconstruction via Model Inversion Attacks against Transformers". *arXiv preprint arXiv:2209.10505*.

Zhang, S., F. Liu, J. Yang, Y. Yang, C. Li, B. Han, and M. Tan. (2023d). "Detecting Adversarial Data by Probing Multiple Perturbations Using Expected Perturbation Score". In: *ICML*.

Zhang, S., Z. Li, S. Yan, X. He, and J. Sun. (2021e). "Distribution alignment: A unified framework for long-tail visual recognition". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2361–2370.

Zhang, T., H. Zheng, J. Yao, X. Wang, M. Zhou, Y. Zhang, and Y. Wang. (2024b). "Long-tailed diffusion models with oriented calibration". In: *ICLR*.

Zhang, Y., R. Jia, H. Pei, W. Wang, B. Li, and D. Song. (2020b). "The secret revealer: Generative model-inversion attacks against deep neural networks". In: *CVPR*.

Zhang, Y., B. Kang, B. Hooi, S. Yan, and J. Feng. (2023e). "Deep long-tailed learning: A survey". *IEEE TPAMI*.

Zhang, Y., B. Kang, B. Hooi, S. Yan, and J. Feng. (2023f). "Deep long-tailed learning: A survey". *TPAMI*.

Zhang, Y., M. Gong, T. Liu, G. Niu, X. Tian, B. Han, B. Schölkopf, and K. Zhang. (2022e). "CausalAdv: Adversarial Robustness through the Lens of Causality". In: *ICLR*.

Zhang, Y., Y. Li, L. Cui, D. Cai, L. Liu, T. Fu, X. Huang, E. Zhao, Y. Zhang, Y. Chen, *et al.* (2023g). "Siren's song in the AI ocean: a survey on hallucination in large language models". *arXiv preprint arXiv:2309.01219*.

Zhang, Z., M. Chen, M. Backes, Y. Shen, and Y. Zhang. (2022f). "Inference attacks against graph neural networks". In: *USENIX Security*.

Zhang, Z., Q. Liu, Z. Huang, H. Wang, C. Lu, C. Liu, and E. Chen. (2021f). "Graphmi: Extracting private graph data from graph neural networks". In: *IJCAI*.

Zhang, Z., H. Luo, L. Zhu, G. Lu, and H. T. Shen. (2022g). "Modality-invariant asymmetric networks for cross-modal hashing". *IEEE Transactions on Knowledge and Data Engineering*. 35(5): 5091–5104.

Zhang, Z., Q. Zhang, and J. Foerster. (2024c). "PARDEN, Can You Repeat That? Defending against Jailbreaks via Repetition". *arXiv preprint arXiv:2405.07932*.

Zhao, H., R. T. des Combes, K. Zhang, and G. J. Gordon. (2019). "On Learning Invariant Representations for Domain Adaptation". In: *ICML*.

Zhao, H., C. Dan, B. Aragam, T. S. Jaakkola, G. J. Gordon, and P. Ravikumar. (2022). "Fundamental Limits and Tradeoffs in Invariant Representation Learning". *Journal of Machine Learning Research*.

Zhao, M., B. An, W. Gao, and T. Zhang. (2017). "Efficient label contamination attacks against black-box learning models." In: *IJCAI*.

Zhao, R., X. Li, S. Joty, C. Qin, and L. Bing. (2023a). "Verify-and-Edit: A Knowledge-Enhanced Chain-of-Thought Framework". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 5823–5840.

Zhao, X., P. Ananth, L. Li, and Y.-X. Wang. (2023b). "Provable robust watermarking for ai-generated text". *arXiv preprint arXiv:2306.17439*.

Zhao, X., Y.-X. Wang, and L. Li. (2023c). "Protecting language generation models via invisible watermarking". In: *ICML*.

Zhao, X., Y.-X. Wang, and L. Li. (2024a). "Watermarking for Large Language Model". *Tutorials of ACL*.

Zhao, X., X. Yang, T. Pang, C. Du, L. Li, Y.-X. Wang, and W. Y. Wang. (2024b). "Weak-to-strong jailbreaking on large language models". *arXiv preprint arXiv:2401.17256*.

Zhao, Y., L. Yan, W. Sun, G. Xing, C. Meng, S. Wang, Z. Cheng, Z. Ren, and D. Yin. (2023d). "Knowing what llms do not know: A simple yet effective self-detection method". *arXiv preprint arXiv:2310.17918*.

Zhao, Y., T. Pang, C. Du, X. Yang, N.-M. Cheung, and M. Lin. (2023e). "A recipe for watermarking diffusion models". *arXiv preprint arXiv:2303.10137*.

Zhao, Z., M. Chen, T. Dai, J. Yao, B. Han, Y. Zhang, and Y. Wang. (2024c). "Mitigating Noisy Correspondence by Geometrical Structure Consistency Learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 27381–27390.

Zhen, L., P. Hu, X. Wang, and D. Peng. (2019). "Deep supervised cross-modal retrieval". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10394–10403.

Zheng, H., L. Zhou, H. Li, J. Su, X. Wei, and X. Xu. (2024). "BEM: Balanced and Entropy-Based Mix for Long-Tailed Semi-Supervised Learning". In: *CVPR*.

Zhong, X., Y. HUANG, and C. Liu. (2023). "Towards Efficient Training and Evaluation of Robust Models against l_0 Bounded Adversarial Perturbations". In: *ICML*.

Zhou, C., Q. Li, C. Li, J. Yu, Y. Liu, G. Wang, K. Zhang, C. Ji, Q. Yan, L. He, *et al.* (2024a). "A comprehensive survey on pretrained foundation models: A history from bert to chatgpt". *International Journal of Machine Learning and Cybernetics*.

Zhou, Y., G. Pu, X. Ma, X. Li, and D. Wu. (2020). "Distilled one-shot federated learning". *arXiv preprint arXiv:2009.07999*.

Zhou, Y., X. Wu, B. Huang, J. Wu, L. Feng, and K. C. Tan. (2024b). "CausalBench: A Comprehensive Benchmark for Causal Learning Capability of Large Language Models". *arXiv preprint arXiv:2404.06349*.

Zhou, Z., C. Zhou, X. Li, J. Yao, Q. Yao, and B. Han. (2023a). "On Strengthening and Defending Graph Reconstruction Attack with Markov Chain Approximation". In: *ICML*.

Zhou, Z.-H. (2017). "A brief introduction to weakly supervised learning". *National Science Review*. 5(1): 44–53.

Zhou, Z., J. Yao, F. Hong, Y. Zhang, B. Han, and Y. Wang. (2023b). "Combating Representation Learning Disparity with Geometric Harmonization".

Zhou, Z., J. Yao, Y.-F. Wang, B. Han, and Y. Zhang. (2022). "Contrastive Learning with Boosted Memorization". In: *ICML*.

Zhu, C., W. R. Huang, H. Li, G. Taylor, C. Studer, and T. Goldstein. (2019). "Transferable clean-label poisoning attacks on deep neural nets". In: *ICML*.

Zhu, H., S. Liu, and F. Jiang. (2022). "Adversarial training of LSTM-ED based anomaly detection for complex time-series in cyber-physical-social systems". *Pattern Recognit. Lett.*

Zhu, J., B. Han, J. Yao, J. Xu, G. Niu, and M. Sugiyama. (2024a). "Decoupling the Class Label and the Target Concept in Machine Unlearning". *arXiv preprint arXiv:2406.08288*.

Zhu, X. and A. B. Goldberg. (2022). *Introduction to semi-supervised learning*. Springer Nature.

Zhu, X. J. (2005). "Semi-supervised learning literature survey".

Zhu, Y., L. Feng, Z. Deng, Y. Chen, R. Amor, and M. Witbrock. (2024b). "Robust Node Classification on Graph Data with Graph and Label Noise". In: *AAAI*.

Zhuang, F., Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He. (2020). "A comprehensive survey on transfer learning". *Proceedings of the IEEE*.

Zou, A., Z. Wang, J. Z. Kolter, and M. Fredrikson. (2023). "Universal and transferable adversarial attacks on aligned language models". *arXiv preprint arXiv:2307.15043*.