Algorithm 1 Calibration-Free Model Purification (CMP)

Input: Poisoned model f_θ^p
Output: Purified student model f_ϕ^s
Notation: X: synthetic dataset, $\tilde{X} \subset X$: sampled batch, $\mathcal{B}$: augmentation, s^p: teacher's soft label, $\mathcal{F}_{\mathrm{mm}}$: mismatch filter, $\mathcal{F}_{\mathrm{sk}}$: skew filter
Generate synthetic dataset X using Eq. (4)
for each training epoch **do**
    Sample mini-batch $\tilde{X} \subset X$
    Apply augmentations: $\mathcal{B}(\tilde{X}) = \{a_i(\tilde{x}_i) \mid \tilde{x}_i \in \tilde{X}\}$
    Compute soft labels: $\mathrm{s}^p(\mathcal{B}(\tilde{x})) = \mathrm{softmax}(f_\theta^p(\mathcal{B}(\tilde{x})))$
    Apply mismatch filtering $\mathcal{F}_{\mathrm{mm}}$ as in Eq. (5) and skew filtering $\mathcal{F}_{\mathrm{sk}}$ as in Eq. (6)
    Update f_ϕ^s by minimizing the total loss in Eq. (7)
end for
return f_ϕ^s

A Threshold Selection

We provide full details of the performance of ANP Wu & Wang (2021) and CLP Zheng et al. (2022a) with respect to the chosen thresholds. We sweep over 20 values for both methods, and report the clean accuracy (ACC) and attack success rate (ASR) for the four baseline attacks in Fig. 3 and Fig. 4. We employ two strategies as reported in Table 2, where we choose a fixed threshold for all attacks, or a per-attack oracle threshold. For the fixed threshold, we select a value that achieves best ACC with ASR reasonably close to $1/C$, and for the oracle threshold, we choose values that achieves ASR closest to $1/C$.

B Implementation Details

config	value
optimizer	SGD
base learning rate	0.1
momentum	0.9
weight decay	1e-4
batch size	256
lr step size	30
lr gamma	0.1
training epoch	90
augmentation	RandomResizedCrop

config	value
optimizer	AdamW
base learning rate	1e-3
weight decay	1e-2
batch size	128
learning rate schedule	cosine decay
training epoch	300
augmentation	RandomResizedCrop

Table 9: Training configurations for ResNet (left) and KD (right).

Attack settings We train a ResNet-18 He et al. (2016) with the official training recipe as in Table 9 (Left). For backdoor attacks badnets Gu et al. (2019), blend Chen et al. (2017), and refool Liu et al. (2020), we poison 10% of the dataset, and 2% for wanet Nguyen & Tran (2021). This is because a high poison rate would degrade the clean accuracy substantially. For badnets Gu et al. (2019), we use a 20×20 resolution white square placed on the right-bottom corner. We use a 0.2 mixing ratio for blend Chen et al. (2017).

Defense settings We follow model inversion Yin et al. (2020) and knowledge distillation Hinton et al. (2015) settings commonly used in the knowledge distillation based dataset distillation works Yin et al. (2023); Yin & Shen (2024); Sun et al. (2024). Note that *we do not tune these hyperparameters, and simply use them off-the-shelf* to demonstrate that our method is robust and does not require extensive hyperparameter tuning. The full details are provided in Table 9 (Right).

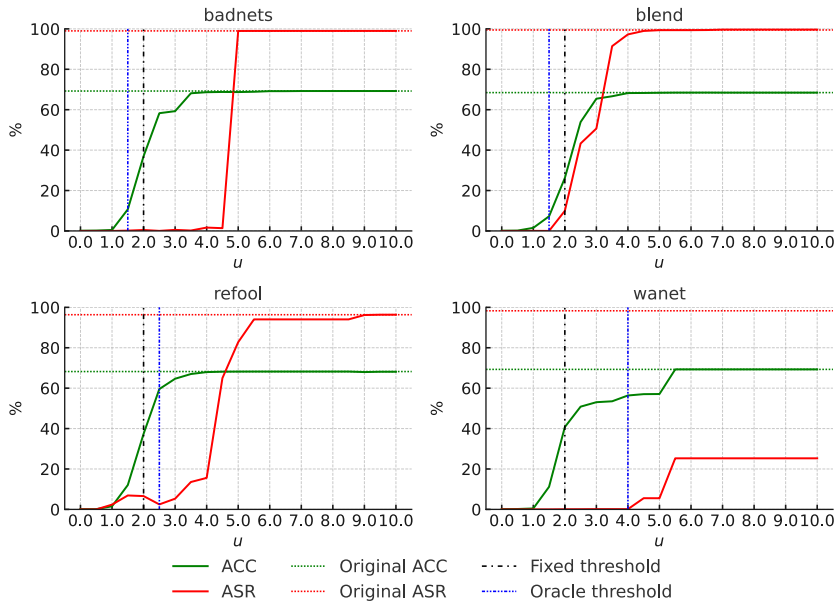

Figure 3: Performance of CLP Zheng et al. (2022a) against four baseline attacks. We plot ACC and ASR with respect to different threshold values, along with the chosen fixed and oracle threshold.

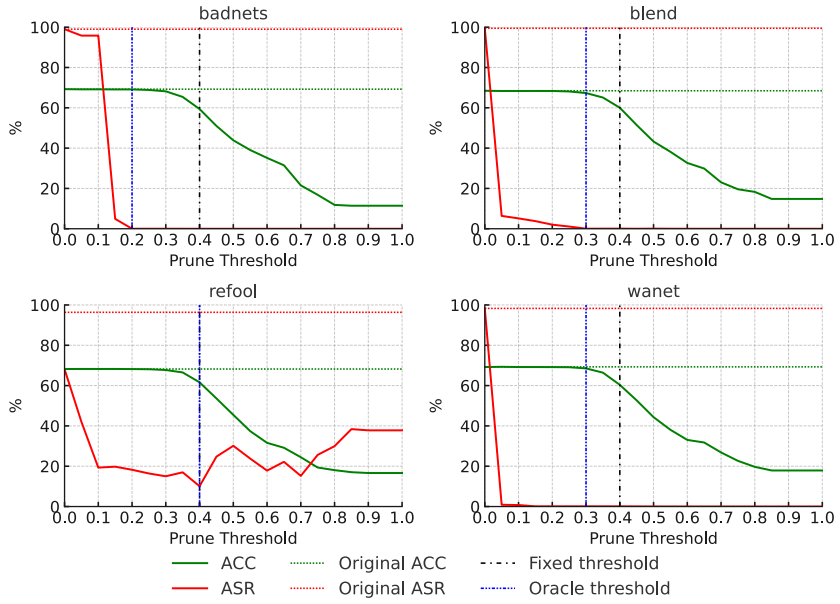

Figure 4: Performance of ANP Wu & Wang (2021) against four baseline attacks. We plot ACC and ASR with respect to different threshold values, along with the chosen fixed and oracle threshold.

C TYPES OF BACKDOOR ATTACKS

In this section, we provide a comprehensive overview of backdoor attack taxonomies and justify our experimental setup that uses four representative attack methods: BadNets Gu et al. (2019), Blend Chen et al. (2017), Refool Liu et al. (2020), and WaNet Nguyen & Tran (2021). These methods were carefully selected to cover the full spectrum of backdoor attack types, ensuring that our evaluation of the CMP defense method is thorough and generalizable.

C.1 ATTACK TAXONOMY

Backdoor attacks can be categorized along several dimensions, with the most significant being the trigger type and its relationship to the input data. Based on these characteristics, we classify backdoor attacks into two primary categories: static and dynamic attacks.

C.2 STATIC ATTACKS

Static attacks use triggers that remain consistent across all poisoned inputs. These can be further subdivided based on the trigger's spatial characteristics:

Local Trigger Attacks Representative method: BadNets Gu et al. (2019)

Local trigger attacks embed a small, spatially concentrated pattern (such as a patch or symbol) at a fixed position in the input image. Key characteristics include:

- Fixed position, typically in a corner or less salient region of the image
- Consistent pattern across all poisoned samples
- Relatively easy to implement but also easier to detect
- Direct modification of pixel values in a localized region

BadNets exemplifies this approach by inserting a small, fixed pattern (e.g., a square pattern) at a consistent location. This attack is highly effective but leaves a distinctive signature that more sophisticated defenses can identify.

Global Trigger Attacks Representative method: Blend Chen et al. (2017)

Global trigger attacks distribute the trigger pattern across the entire input, affecting all or most of the image area. Key characteristics include:

- Pattern spans the entire input space rather than a localized region
- Often implemented through blending or transparency effects
- More subtle visual appearance compared to local triggers
- May be less disruptive to semantic content

The Blend attack achieves this by superimposing a pattern across the entire image with controlled transparency, making it less visually obvious than patch-based attacks while maintaining high attack success rates.

Multiple Trigger Attacks Representative method: Refool Liu et al. (2020)

Multiple trigger attacks employ several different triggers that all map to the same target label. Key characteristics include:

- Uses multiple distinct trigger patterns
- All triggers direct classification to the same target
- Increases attack robustness against defenses that may only identify some triggers
- Often leverages naturally occurring visual elements

Refool specifically uses reflection patterns as triggers, embedding multiple types of reflection effects that appear natural to human observers but strongly influence model predictions. This makes the attack both stealthy and robust against many defense mechanisms.

C.3 Dynamic Attacks

Representative method: WaNet Nguyen & Tran (2021)

Unlike static attacks, dynamic attacks use input-dependent triggers that adapt to the specific content of each image. Key characteristics include:

- Trigger transformation depends on the input image content
- No fixed pattern that repeats across poisoned samples
- Often implemented through subtle warping or transformations
- Significantly harder to detect through statistical analysis

WaNet implements this approach through imperceptible image warping, where the warping function creates a trigger that is uniquely applied to each image. This makes the attack particularly challenging to defend against, as traditional methods that search for consistent patterns across poisoned inputs fail to identify the dynamic nature of the trigger.

C.4 Justification for Selected Attack Methods

Our evaluation utilizes these four attack methods (BadNets, Blend, Refool, and WaNet) because they:

1. **Span the trigger design space**: From simple local patterns (BadNets) to sophisticated input-adaptive transformations (WaNet)
2. **Represent varying levels of detectability**: From more obvious modifications to nearly imperceptible changes
3. **Exercise different attack mechanisms**: Direct pixel manipulation, global blending, naturalistic artifacts, and geometric transformations
4. **Challenge different aspects of defense systems**: Some attacks may be neutralized by specific defense components while being resistant to others

By evaluating CMP against this diverse set of attack methods, we ensure that our defense is robust against the full spectrum of backdoor attacks, rather than being optimized for a particular attack subtype. This comprehensive evaluation approach provides stronger evidence for the universality of our defense method to both known and potentially unknown backdoor attack variants.

REFERENCES

Yuki M Asano and Aaqib Saeed. The augmented image prior: Distilling 1000 classes by extrapolating from a single image. *arXiv preprint arXiv:2112.00725*, 2021.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. URL https://arxiv.org/abs/2005.14165.

Sébastien Bubeck and Mark Sellke. A universal law of robustness via isoperimetry. *Advances in Neural Information Processing Systems*, 34:28811–28822, 2021.

Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian M. Molloy, and Biplav Srivastava. Detecting backdoor attacks on deep neural networks by activation clustering. In *Proceedings of the AAAI Workshop on Artificial Intelligence Safety (SafeAI)*, volume 2301 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2019. URL https://ceur-ws.org/Vol-2301/paper_18.pdf.

Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

Khoi Doan, Ehsan Abbasnejad, Damith Ranasinghe, and James Bailey. Lira: Learnable, imperceptible and robust backdoor attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/papers/Doan_LIRA_Learnable_Imperceptible_and_Robust_Backdoor_Attacks_ICCV_2021_paper.pdf.

Min Du, Ruoxi Jia, and Dawn Song. Robust anomaly detection and backdoor attack detection via differential privacy. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://arxiv.org/abs/1911.07116.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Yansong Gao, Chang Xu, Derui Wang, Shiping Chen, Damith C. Ranasinghe, and Surya Nepal. Strip: A defence against trojan attacks on deep neural networks. In *Proceedings of the 35th Annual Computer Security Applications Conference (ACSAC)*, pp. 113–125. ACM, 2019. doi: 10.1145/3359789.3359790. URL https://dl.acm.org/doi/10.1145/3359789.3359790.

Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. BadNets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019. doi: 10.1109/ACCESS.2019.2909068.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015. arXiv:1503.02531.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. pmlr, 2015.

Rishi D. Jha, Jonathan Hayase, and Sewoong Oh. Label poisoning is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://openreview.net/forum?id=prftZp6mDH.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. URL https://arxiv.org/abs/1412.6980.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Junhoo Lee, Hyunho Lee, Kyomin Hwang, and Nojun Kwak. Deep support vectors. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://arxiv.org/abs/2403.17329. Accepted; to appear.

Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Neural attention distillation: Erasing backdoor triggers from deep neural networks. In *9th International Conference on Learning Representations (ICLR)*, 2021a. URL https://arxiv.org/abs/2101.05930. arXiv:2101.05930.

Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample-specific triggers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 16463–16472, 2021b.

Yuezun Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample-specific triggers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021c. URL https://arxiv.org/abs/2012.03816.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. URL http://arxiv.org/abs/1405.0312.

Weilin Lin, Li Liu, Jianze Li, and Hui Xiong. Fusing pruned and backdoored models: Optimal transport-based data-free backdoor mitigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 26299–26307, 2025.

Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *Research in Attacks, Intrusions and Defenses (RAID 2018)*, volume 11050 of *Lecture Notes in Computer Science*, pp. 273–294. Springer, Cham, 2018a.

Min Liu, Alberto Sangiovanni-Vincentelli, and Xiangyu Yue. Beating backdoor attack at its own game. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4597–4606, 2023. doi: 10.1109/ICCV51070.2023.00426.

Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. In *25th Annual Network and Distributed System Security Symposium (NDSS)*, pp. 1–15. The Internet Society, 2018b.

Yunfei Liu, Xingjun Ma, James Bailey, and Feng Lu. Reflection backdoor: A natural backdoor attack on deep neural networks. In *Computer Vision – ECCV 2020, Part X*, volume 12355 of *Lecture Notes in Computer Science*, pp. 182–199. Springer, Cham, 2020. doi: 10.1007/978-3-030-58607-2_11.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022, 2021. doi: 10.1109/ICCV48922.2021.00988. URL https://arxiv.org/abs/2103.14030.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Tuan Anh Nguyen and Anh Tran. Input-aware dynamic backdoor attack. *Advances in Neural Information Processing Systems*, 33:3454–3464, 2020.

Tuan Anh Nguyen and Anh Tuan Tran. WaNet – imperceptible warping-based backdoor attack. In *International Conference on Learning Representations (ICLR)*, 2021.

Lu Pang, Tao Sun, Haibin Ling, and Chao Chen. Backdoor cleansing with unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12218–12227, 2023. doi: 10.1109/CVPR52729.2023.01176.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336, 2021. doi: 10.1016/j.patter.2021.100336.

Huy Phan, Jinqi Xiao, Yang Sui, Tianfang Zhang, Zijie Tang, Cong Shi, Yan Wang, Yingying Chen, and Bo Yuan. Clean & compact: Efficient data-free backdoor defense with model compactness. In *European Conference on Computer Vision (ECCV)*, 2024a.

Huy Phan, Jinqi Xiao, Yang Sui, Tianfang Zhang, Zijie Tang, Cong Shi, Yan Wang, Yingying Chen, and Bo Yuan. Clean and compact: Efficient data-free backdoor defense with model compactness. In *European Conference on Computer Vision*, pp. 273–290. Springer, 2024b.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951. doi: 10.1214/aoms/1177729586.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Ahmed Salem, Rui Wen, Michael Backes, Shiqing Ma, and Yang Zhang. Dynamic backdoor attacks against machine learning models. In *Proceedings of the 2022 IEEE European Symposium on Security and Privacy (EuroS&P)*, 2022. URL https://arxiv.org/abs/2003.03675.

Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural Networks*, 32:323–332, 2012. doi: 10.1016/j.neunet.2012.02.016.

Peng Sun, Bei Shi, Daiwei Yu, and Tao Lin. On the diversity and realism of distilled dataset: An efficient dataset distillation paradigm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9390–9399, 2024.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pp. 10347–10357. PMLR, 2021. URL https://proceedings.mlr.press/v139/touvron21a.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y. Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 707–723, San Francisco, CA, USA, 2019. IEEE. ISBN 978-1-5386-6660-9. doi: 10.1109/SP.2019.00031. URL https://ieeexplore.ieee.org/document/8835365.

Baoyuan Wu, Hongrui Chen, Mingda Zhang, Zihao Zhu, Shaokui Wei, Danni Yuan, Mingli Zhu, Ruotong Wang, Li Liu, and Chao Shen. Backdoorbench: A comprehensive benchmark and analysis of backdoor learning. *International Journal of Computer Vision*, pp. 1–88, 2025.

Dongxian Wu and Yisen Wang. Adversarial neuron pruning purifies backdoored deep models. In *Advances in Neural Information Processing Systems (NeurIPS) 34*, 2021.

Hongxu Yin, Pavlo Molchanov, Jose M. Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K. Jha, and Jan Kautz. Dreaming to distill: Data-free knowledge transfer via deepinversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8712–8721, 2020. doi: 10.1109/CVPR42600.2020.00874.

Zeyuan Yin and Zhiqiang Shen. Dataset distillation via curriculum data synthesis in large data era. *Transactions on Machine Learning Research*, 2024.

Zeyuan Yin, Eric Xing, and Zhiqiang Shen. Squeeze, recover and relabel: Dataset condensation at imagenet scale from a new perspective. *Advances in Neural Information Processing Systems*, 36: 73582–73603, 2023.

Jie Zhang, Dongdong Chen, Qidong Huang, Jing Liao, Weiming Zhang, Huamin Feng, Gang Hua, and Nenghai Yu. Poison ink: Robust and invisible backdoor attack. *arXiv preprint arXiv:2108.02488*, 2021. URL https://arxiv.org/abs/2108.02488.

Pu Zhao, Pin-Yu Chen, Payel Das, Karthikeyan Natesan Ramamurthy, and Xue Lin. Bridging mode connectivity in loss landscapes and adversarial robustness. In *International Conference on Learning Representations (ICLR)*, 2020.

Runkai Zheng, Rongjun Tang, Jianze Li, and Li Liu. Data-free backdoor removal based on channel lipschitzness. In *European Conference on Computer Vision (ECCV)*, pp. 175–191. Springer, 2022a.

Runkai Zheng, Rongjun Tang, Jianze Li, and Li Liu. Pre-activation distributions expose backdoor neurons. *Advances in Neural Information Processing Systems*, 35:18667–18680, 2022b.