Seal Your Backdoor with Variational Defense

Ivan Sabolić Matej Grcić Siniša Šegvić
University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
name.surname@fer.hr

Abstract

We propose VIBE, a model-agnostic framework that trains classifiers resilient to backdoor attacks. The key concept behind our approach is to treat malicious inputs and corrupted labels from the training dataset as observed random variables, while the actual clean labels are latent. VIBE then recovers the corresponding latent clean label posterior through variational inference. The resulting training procedure follows the expectation-maximization (EM) algorithm. The E-step infers the clean pseudolabels by solving an entropy-regularized optimal transport problem, while the M-step updates the classifier parameters via gradient descent. Being modular, VIBE can seamlessly integrate with recent advancements in self-supervised representation learning, which enhance its ability to resist backdoor attacks. We experimentally validate the method effectiveness against contemporary backdoor attacks on standard datasets, a large-scale setup with 1k classes, and a dataset poisoned with multiple attacks. VIBE consistently outperforms previous defenses across all tested scenarios.

1. Introduction

Deep models possess enough capacity to learn any pattern present within the data [8, 42, 101]. This remarkable flexibility comes at the cost of control since it limits our ability to influence the specific motifs the model learns [2, 37]. For instance, a model may base its decisions on image backgrounds rather than focusing on the actual objects [21]. Such bias towards simpler [75] and possibly spurious patterns [34] may lead to undesirable generalizations that reveal themselves only in specific test cases. This deep learning loophole can be maliciously exploited by attackers who manipulate training examples using triggers that steer the model towards harmful generalization. Such practice is commonly referred to as *backdoor learning* [49] and presents a serious threat in contemporary machine learning.

The majority of existing backdoor attacks [12, 23] modify a portion of the training dataset by installing triggers onto selected inputs and altering the corresponding labels[1]. On such data, standard supervised learning delivers a *poisoned* model [49]. During inference, attackers can exploit the installed backdoor by applying triggers to the desired inputs, which causes the model to behave maliciously [23, 60]. Our goal is to defend against such attacks by training a *clean* model invariant to triggers present in the data.

Recent empirical defenses [11, 53, 107] partition the training dataset into clean and poisoned subsets according to some heuristics. The two subsets then take different roles during the model training (*e.g.* semi-supervised learning with labeled clean data and unlabeled poisoned data [33]). However, heuristics are prone to failure modes and can be exploited by adaptive attacks [67]. Also, pruning labels often leads to information loss, ultimately degrading recognition performance. Our approach avoids data partitioning and label pruning. Instead, we leverage optimal transport to refine potentially corrupted samples and labels into clean pseudolabels that guide the training of a robust classifier.

In this work, we present VIBE (**V**ariational **I**nference for **B**ackdoor **E**limination), a framework for training backdoor-robust classifiers on poisoned data. Our key concept is to treat dataset examples and the corresponding corrupted labels as observed random variables, while the desired clean labels are latent. Then, we achieve resilience against backdoor attacks by recovering the latent clean posterior parametrized as a deep classifier. VIBE training takes the form of an expectation-maximization algorithm that alternates between classifier updates via gradient descent (M-step), and inference of approximate clean class posterior (E-step). In practice, the approximate clean labels are recovered by solving an entropy-regularized optimal transport problem [15]. We validate the resilience against contemporary backdoor attacks on standard benchmarks, on a large-scale setup with 1k classes, and on a dataset poisoned with multiple attacks. Experiments indicate consistent improvements over previous defenses in all tested scenarios. Remarkably, VIBE attains over 12pp absolute improvement in

[1] Some attacks do not alter the labels [80]. However, our experiments show that they are much easier to defend from.

ASR over the best baseline on the CIFAR-10 dataset.

2. Related work

Backdoor attacks. Backdoor attacks achieve malicious model behaviour through direct modifications of trainable parameters [70], changes in model structure [31], or data poisoning [23]. Contemporary machine learning models are trained in-house and deployed via APIs which makes parameter- and structure-based attacks largely impractical. Therefore, we focus on a more realistic scenario where the attacker only controls the data collection process.

Early data poisoning attacks [23] steer the model towards malicious generalization by introducing localized triggers and altering the corresponding labels. Subsequent attacks rely on invisible [13, 36, 67, 83] or sample-specific [46, 102] triggers, which are significantly harder to detect. Clean-label attacks [6, 80, 99] avoid modifying labels altogether but typically deliver lower attack success rates. All these approaches devise broad method-agnostic attacks. Contrary, recent adaptive attacks improve effectiveness by targeting the latest defenses [20, 33].

Another line of work considers data poisoning attacks that target contrastive self-supervised learning [45, 51, 73]. However, these attacks may require access to the optimization procedure [35] and typically have a lower success rate than direct attacks on supervised learning [45]. A detailed survey of backdoor attacks can be found in [49].

Backdoor defenses. Existing defenses can be categorized as either certified or empirical. Certified defenses provide theoretical guarantees of success [92]. However, their underlying assumptions typically do not hold in practice [49]. Empirical defenses devise preprocessing strategies to avoid training on corrupted examples [79], correct the malicious generalization of poisoned models via postprocessing [54], or propose heuristic additions to standard training algorithms [48]. Preprocessing-based defenses [9, 25, 32, 39] aim to filter out poisoned examples from the dataset, allowing the remaining clean data to be safely used for supervised training. Such methods cannot distinguish poisoned examples from their hard counterparts with the correct label [39].

Post-training defenses [47, 64, 93, 100] focus on removing backdoors from already trained models. A common approach involves re-synthesizing the injected triggers and using them to purify the model [24, 55, 68, 76, 77, 81, 86, 91, 94, 106]. Other approaches correct malicious generalization through model pruning [50, 52, 90], knowledge distillation [98], loss landscape analysis [104], or by enhancing robustness against adversarial examples [59]. All these methods assume access to a small subset of definitively clean data, which may not be available in practice.

Training-time defenses [11, 20, 33, 53, 72, 87, 103, 105, 107] attempt to train robust classifiers from poisoned data. An early approach [47] isolates poisoned examples in early training stages and later uses them to unlearn the backdoors. Subsequent methods [11, 20, 33] focus on removing only the labels of potentially poisoned samples and proceed with semi-supervised training. These approaches identify the poisoned data by heuristics, which increases the defense vulnerability. VIBE avoids such heuristics by recovering clean pseudolabels through variational inference.

Representation learning. The main goal of representation learning [65, 71, 85] is to recover features that generalize across a spectrum of downstream tasks. A widely used representation learning strategy involves optimizing self-supervised pretext objectives [22, 62]. Recent such methods reconstruct masked inputs [30], optimize contrastive objectives [10] or learn latent centroids [63]. Representation quality can be further enhanced by training on large multimodal datasets [69], which delivers effective features even without fine-tuning on target datasets [14, 63].

Representation learning received limited attention in the context of backdoor defenses. Initial works design heuristics that leverage self-supervised representations to remove poisoned labels [33] or filter the dataset [82]. These heuristics fail for some attacks, as indicated by our experimental evaluation. In contrast, VIBE uses self-supervised pre-training to jumpstart the optimization process.

Recent works [7, 26, 95, 96] suggest that (multimodal) contrastive learning can be conveniently adapted to resist data poisoning attacks that target self-supervised pre-training. Pre-training VIBE feature extractor according to such objectives further boosts the performance.

Latent variable models. Latent variable models [19] explain relations between observed random variables with latent variables. This concept is successfully applied in different fields [40, 56, 88]. VIBE introduces latent variables into backdoor defenses by viewing clean labels as latents.

3. Backdoor resilience via variational inference

Problem setup. Let $\mathcal{D}_{\text{raw}} = \{(\tilde{\boldsymbol{x}}^i, l^i)\}_{i=1}^N$ be a benign dataset consisting of input examples $\tilde{\boldsymbol{x}}^i \in \mathcal{X}$ and clean labels $l^i \in \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are input and label space respectively. A malicious attacker $\tau : \mathcal{X} \times \mathcal{Y} \to \mathcal{X} \times \mathcal{Y}$ with a budget $\gamma \in [0, 100]$ modifies $\gamma\%$ examples by triggering inputs and corrupting their labels. The remaining $(100 - \gamma)\%$ of the data remain unchanged and correctly labeled in order to conceal the attack. Given a corrupted dataset $\mathcal{D} = \tau(\mathcal{D}_{\text{raw}}) = \{(\boldsymbol{x}^i, y^i)\}_{i=1}^N$, our goal is to train a robust classifier $f : \mathcal{X} \to \mathcal{Y}$ that assigns clean labels $l^i \in \mathcal{Y}$ to every input $\boldsymbol{x}^i$ while ignoring the malicious triggers.

The core concept behind VIBE is to treat clean labels as unobserved latent variables $\underline{l}$. We then frame the training of a clean classifier as a latent posterior recovery from observed inputs $\underline{\boldsymbol{x}}$ and corrupted labels $\underline{y}$.

Backdoor attacks typically poison a small portion of the data in order to stay undetected [49]. In our framework, this

means that corrupted and clean labels are often identical. Thus, a natural optimization objective is to maximize the conditional log-likelihood of the i.i.d dataset $\mathcal{D}$ given a set of model parameters θ:

$$\ell(\theta|\mathcal{D}) = \ln \prod_{i=1}^{N} p_\theta(y^i|\boldsymbol{x}^i) = \sum_{i=1}^{N} \ln \sum_{l=1}^{K} p_\theta(y^i|l,\boldsymbol{x}^i)\, p_\theta(l|\boldsymbol{x}^i). \tag{1}$$

For simplicity, we abbreviate $p(\underline{l}=l|\underline{\boldsymbol{x}}=\boldsymbol{x}^i)$ as $p(l|\boldsymbol{x}^i)$. Given the likelihood factorization (1), we proceed by deriving a tractable optimization objective. Note that we defer concrete parametrization of the clean class posterior $p_\theta(\underline{l}|\underline{\boldsymbol{x}})$ and the corrupted class posterior $p_\theta(\underline{y}|\underline{l},\underline{\boldsymbol{x}})$ to Section 3.2.

3.1. Optimizing the variational objective via EM

Direct maximization of $\ell(\theta|\mathcal{D})$ does not ensure the correct recovery of the clean class posterior since the clean class is latent [57]. Fortunately, we can turn to variational inference and maximize likelihood lower bound ℓ_{ELBO} that introduces an approximate latent posterior q:

$$\begin{aligned}\ell(\theta|\mathcal{D}) &= \sum_{i=1}^{N} \ln \sum_{l=1}^{K} p_\theta(y^i|l,\boldsymbol{x}^i)\, p_\theta(l|\boldsymbol{x}^i) \frac{q(l|\boldsymbol{x}^i,y^i)}{q(l|\boldsymbol{x}^i,y^i)} \\ &\geq \sum_{i=1}^{N} \mathbb{E}_{l^i\sim q(\cdot|\boldsymbol{x}^i,y^i)} \left[\ln \frac{p_\theta(y^i|l^i,\boldsymbol{x}^i) p_\theta(l^i|\boldsymbol{x}^i)}{q(l^i|\boldsymbol{x}^i,y^i)}\right] \\ &=: \ell_{\text{ELBO}}(\theta, q|\mathcal{D}). \end{aligned} \tag{2}$$

The inequality follows directly from Jensen's inequality. We optimize the proposed ℓ_{ELBO} objective with the expectation-maximization (EM) algorithm [57]. In practice, this involves alternating between updates of the approximate latent posterior q (E-step) and parameters θ (M-step).

E-step: updating the approximate latent posterior. We begin by observing that the ℓ_{ELBO} objective requires only the recovery of q for dataset examples, rather than an exact closed-form distribution. With this observation in mind, we rewrite the objective (2) averaged over N examples as:

$$\begin{aligned}\frac{1}{N}\ell_{\text{ELBO}} = &\sum_{i=1}^{N}\sum_{l=1}^{K} \frac{1}{N} q(l|\boldsymbol{x}^i,y^i) \ln[p_\theta(y^i|l,\boldsymbol{x}^i) p_\theta(l|\boldsymbol{x}^i)] \\ &- \sum_{i=1}^{N}\sum_{l=1}^{K} \frac{1}{N} q(l|\boldsymbol{x}^i,y^i) \ln q(l|\boldsymbol{x}^i,y^i). \end{aligned} \tag{3}$$

We next substitute $\mathbf{P}_{i,l} := p_\theta(y^i|l,\boldsymbol{x}^i) p_\theta(l|\boldsymbol{x}^i)$ and $\mathbf{Q}_{i,l} := \frac{1}{N} q(l|\boldsymbol{x}^i,y^i)$, where $1/N$ ensures that $\mathbf{Q}$ is a proper joint distribution [3, 61]. Replacing the summations with matrix multiplication reveals the same objective in the matrix form:

$$\begin{aligned}\frac{1}{N}\ell_{\text{ELBO}} &= \operatorname{tr}(\mathbf{Q}^\top \ln \mathbf{P}) + \mathbb{H}(\mathbf{Q}) + 1 - \ln N \\ &\geq \operatorname{tr}(\mathbf{Q}^\top \ln \mathbf{P}) + \frac{1}{\lambda}\mathbb{H}(\mathbf{Q}) + 1 - \ln N \end{aligned} \tag{4}$$

Here, $\operatorname{tr}(\cdot)$ is the matrix trace operator, $\lambda > 1$ is a hyperparameter, and $\mathbb{H}(\mathbf{Q})$ is the entropy of coupling matrix $\mathbf{Q}$ [66]. The complete derivation is deferred to Appendix **??**. The term $1 - \ln N$ is constant and thus can be ignored.

Each matrix row $\mathbf{Q}_{i,:}$ sums to $1/N$ by the definition of $\mathbf{Q}$, while columns sum to the prior over clean classes $\boldsymbol{\pi}$. Consequently, the set of all possible solutions for the objective (4) forms a polytope:

$$\mathcal{Q}[\boldsymbol{\pi}] = \{\mathbf{Q} \in \mathbb{R}_+^{N\times K} \,|\, \mathbf{Q}^\top \mathbf{1}_N = \boldsymbol{\pi},\, \mathbf{Q}\mathbf{1}_K = \frac{1}{N}\mathbf{1}_N\}. \tag{5}$$

Here, $\mathbf{1}_N$ is an N-dimensional column vector. Maximizing the objective (4) over $\mathcal{Q}[\boldsymbol{\pi}]$ is equivalent to solving the entropy-regularized optimal transport problem [3, 15, 66]:

$$\mathbf{Q}^* = \operatorname*{arg\,min}_{\mathbf{Q}\in\mathcal{Q}[\boldsymbol{\pi}]} \left(\operatorname{tr}(\mathbf{Q}^\top \mathbf{M}) - \frac{1}{\lambda}\mathbb{H}(\mathbf{Q})\right). \tag{6}$$

Here, the cost matrix contains the model outputs in dataset examples ($\mathbf{M} = -\ln \mathbf{P}$). The optimal solution $\mathbf{Q}^*$ can be efficiently obtained with the Sinkhorn-Knopp's matrix scaling algorithm [41, 66], which we revisit in Appendix **??**. This approach is computationally efficient even for large N, as discussed in the experiments. The recovered solution $\mathbf{Q}^*$ contains outputs of the approximate posterior q for the dataset examples and allows us to proceed with the M-step.

M-step: updating model parameters. Given the outputs of approximate posterior q, we can turn to the optimization of parameters θ. Maximizing the ℓ_{ELBO} objective (2) is equivalent to the following minimization problem:

$$\min_\theta \sum_{i=1}^{N} \text{CE}[q\,||\,p_\theta(l^i|\boldsymbol{x}^i)] + \mathbb{E}_{l^i\sim q}\big[-\ln p_\theta(y^i|l^i,\boldsymbol{x}^i)\big] \tag{7}$$

Here, CE denotes the cross-entropy loss. The full derivation can be found in Appendix **??**. The objective (7) is continuous w.r.t parameters θ and can be optimized by (stochastic) gradient descent. The rewritten objective highlights the role of the approximate posterior q: it acts as a pseudolabel generator. These pseudolabels are also conditioned on the corrupted labels and thus provide a learning signal for the actual clean posterior. The second objective term models the relation between the corrupted labels and the pseudolabels. This term presents an opportunity to uncover the attacker's poisoning patterns that can guide human inspection.

Altogether, VIBE training alternates between the described E and M steps as visualized in Figure 1. The full algorithm is in Appendix **??**. Next, we discuss the implementations of distributions parameterized with θ.

3.2. Parameterizing the posteriors

Let $g_{\theta_{\text{E}}} : \mathcal{X} \to S^{d-1}$ be a feature extractor that encodes inputs onto a $(d-1)$-dimensional unit hypersphere, *e.g.* a

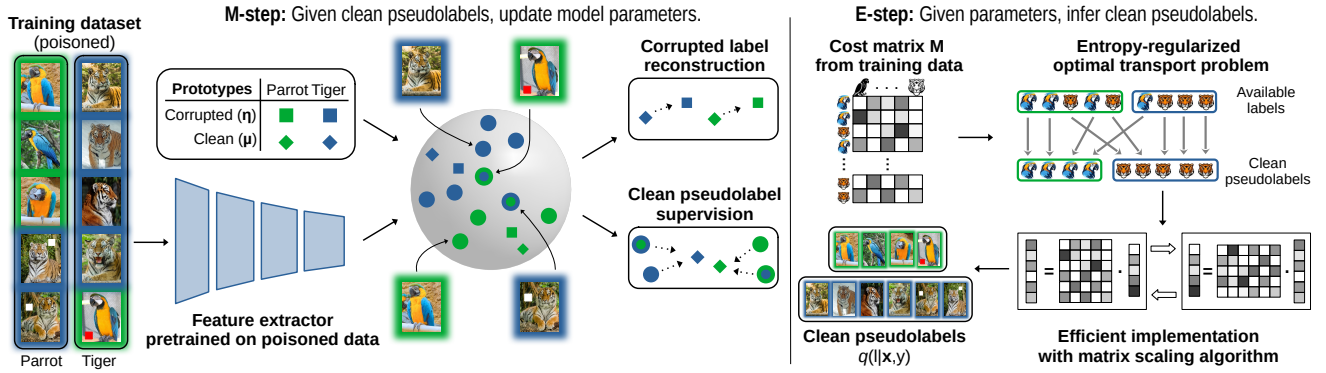

Figure 1. VIBE training alternates between iterative parameter updates (M-step) and inference of clean pseudolabels by solving entropy-regularized optimal transport problem efficiently implemented with matrix scaling algorithm (E-step).

ResNet with L_2-normalized outputs. We model the conditional likelihood of the encoded input $\boldsymbol{v}^i = g_{\theta_\text{E}}(\boldsymbol{x}^i)$ given the clean class l^i with a von Mises-Fisher distribution [5]:

$$p_\theta(\boldsymbol{v}^i|l^i) = C_d(\kappa)\exp(\kappa \boldsymbol{\mu}_{l^i}^\top \boldsymbol{v}^i). \tag{8}$$

The vector $\boldsymbol{\mu}_{l^i} \in S^{d-1}$ sets the mean direction, the hyper-parameter κ controls the distribution spread, while $C_d(\kappa)$ is a normalization constant [5]. We can derive the clean label posterior as a vMF mixture via the Bayes rule:

$$p_\theta(l^i|\boldsymbol{x}^i) = \frac{\exp(\kappa {\boldsymbol{v}^i}^\top \boldsymbol{\mu}_{l^i} + \ln \pi_{l^i})}{\sum_{l'} \exp(\kappa {\boldsymbol{v}^i}^\top \boldsymbol{\mu}_{l'} + \ln \pi_{l'})}. \tag{9}$$

Here, the mixing coefficient $\boldsymbol{\pi}$ induces a prior over clean classes. In practice, we compute $\boldsymbol{\pi} = \sigma(c \cdot \theta_\pi)$, where σ is softmax activation that ensures $\boldsymbol{\pi}$ is a distribution, c is a hyperparmeter, and $\theta_\pi \in \mathbb{R}^d$ are learnable parameters. The full derivation is in Appendix **??**. The clean posterior (9) corresponds to a softmax-activated deep model with L_2-normalized pre-logits and clean class prototypes $\theta_l = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\}$. Thus, the clean labels can be recovered by $f = \text{argmax} \circ \text{cos-sim}_{\theta_{l,\pi}} \circ g_{\theta_\text{E}}$, where cos-sim operator computes cosine similarities adjusted by the bias $\ln \pi$.

We model the corrupted class posterior $p_\theta(y^i|l^i, \boldsymbol{x}^i)$ as cosine similarity between the corrupted class prototypes $\theta_y = \{\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_K\}$ and output of function h that process the encoded input $\boldsymbol{v}^i$ and the clean label prototype $\boldsymbol{\mu}_{l^i}$:

$$p_\theta(y^i|l^i, \boldsymbol{x}^i) := \frac{\exp(\nu \cdot {\boldsymbol{\eta}_{y^i}}^\top h(\boldsymbol{\mu}_{l^i}, \boldsymbol{v}^i))}{\sum_{y'} \exp(\nu \cdot {\boldsymbol{\eta}_{y'}}^\top h(\boldsymbol{\mu}_{l^i}, \boldsymbol{v}^i))}. \tag{10}$$

Here, ν is a scalar hyper-parameter, while details on h are deferred to implementation details. Note that the full corrupted posterior can be approximated as $p_\theta(y^i|l^i, \boldsymbol{x}^i) \approx p_\theta(y^i|l^i)$ by replacing output of h with $\boldsymbol{\mu}_{l^i}$. The detailed description of the approximated corrupted posterior is in Appendix **??**. While this approximation makes optimization more challenging, it enables seamless reconstruction of the systematic poisoning rules of the attacker τ. We experimentally evaluate both the full and approximate posterior.

Alltogether, the set of free parameters is a union $\theta = \theta_\text{E} \cup \theta_l \cup \theta_\pi \cup \theta_y$. Next we analyze convergence of the EM algorithm with the introduced parametrization.

3.3. Steering the EM algorithm convergence

Our E-step solves a convex optimization problem [15], while the M-step conducts non-convex training of a deep model. As a result, the EM algorithm may end up in a suboptimal stationary point [4, 57, 89]. In fact, the convergence point of the EM algorithm strongly depends on the initialization [58]. Fortunately, recent works observe that self-supervised pre-training of feature extractors [10, 18, 29] lowers the sample complexity of the downstream task [1, 44] and improves generalization [84].

Therefore, we conduct self-supervised pre-training on the poisoned dataset instance (similar to [33]) before end-to-end optimization of our ℓ_ELBO objective. Figure 2 shows that self-supervised pre-training on poisoned data jumpstarts the EM optimization, leads to faster convergence, and increases the likelihood lower bound ℓ_ELBO. Moreover, the corresponding solution generalizes better and turns out to be near optimal compared to supervised learning on clean labels. Still, the self-supervised pre-training does not compromise the generality of VIBE, as pre-training objectives are already available for various modalities [27, 38, 97].

Scaling the concept of feature extractor pre-training in terms of dataset size leads to foundation models [14, 63, 69] like CLIP and DINOv2. Modular design of VIBE posteriors enables integration of these off-the-shelf extractors, allowing performance analysis in the *transfer learning* setup. Nevertheless, foundation models should be carefully downloaded from trusted third-party providers or pre-trained with robust procedures [95, 96] that avoid backdoor injection during self-supervised pre-training stage.

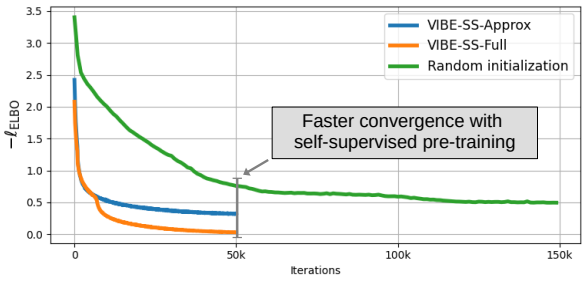

Figure 2. VIBE achieves faster convergence and improved generalization with self-supervised pre-training.

3.4. Clean-label attacks diverge from data manifold

Poisoned-label attacks strive for stealthiness by injecting minimal triggers. Thus, the poisoned examples remain near the clean data manifold. However, clean-label attacks [6, 80] operate by significantly perturbing the inputs to construct a successful attack. These perturbations shift the poisoned examples away from the data manifold in the self-supervised feature space, as illustrated in Figure 3. We propose a pre-processing technique that exploits this property of clean-label attacks to identify the poisoned examples.

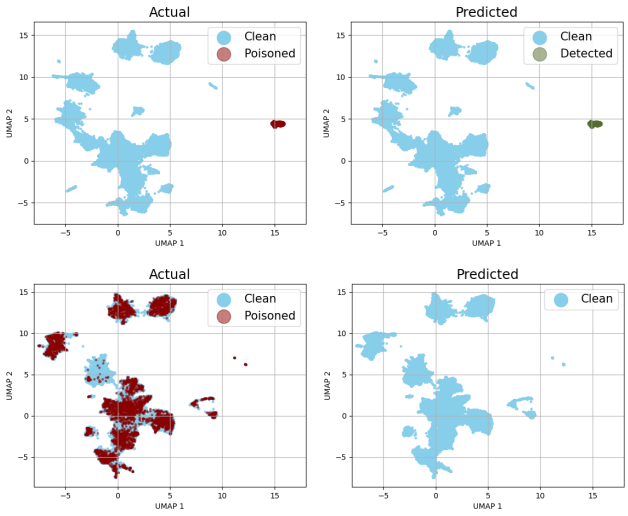

Figure 3. Clean-label attacks shift the poisoned examples away from the clean data manifold (top-left). Thus, we can detect them as outlier communities (top-right). Poisoned-label attacks remain on the manifold and can be useful for model training (down-left). Thus, our preprocessing strategy keeps them (down-right).

Our preprocessing strategy captures the data manifold within a self-supervised feature space and identifies perturbed examples as outlier communities. Specifically, we construct k nearest neighbors graph, represented by the adjacency matrix A_k. Then, we apply a community detection algorithm [78] that partitions graph A_k into K+1 distinct communities. We proceed by computing pairwise distances between the communities to identify the most distant one. Finally, we remove the most distant community if its average distance to other communities exceeds some predefined threshold δ. In the case of clean-label attacks, the most distant community corresponds to off-manifold data which is then removed. If poisoned examples are absent from the data, the most distant community is retained due to being much closer to the data manifold. While some clean examples may be lost if δ is poorly selected, they typically do not affect the model generalization. More details on preprocessing strategy are in Appendix **??**.

4. Experimental setup

Datasets & metrics. We evaluate VIBE on the standard backdoor benchmarks: CIFAR-10, CIFAR-100 [43], and a subset of 30 classes from the ImageNet-1k dataset [16]. Furthermore, we scale the problem by considering the full ImageNet dataset with 1k classes and 1.2M training examples. We evaluate the performance with the standard metrics: accuracy on the clean test set labels (ACC), and the attack success rate (ASR). Lower ASR indicates more accurate recognition of poisoned samples and better defense.

Attacks. We consider eight attacks that represent the major backdoor attack families. For visible patch-like attacks, we include BadNets [23] and Adap-Patch [67]). For invisible attacks, we consider Blend [12], WaNet [60], Adap-Blend [67] and Frequency [83]. Clean-label attacks are represented by LC [80]. We also validate the *all-to-all* attack [49] using a variant of BadNets. By default, the poisoning rate is set to 10% except for Adap-Patch and Adap-Blend attacks in which 1% of the data is poisoned as suggested in [67]. Also, in the case of clean-label attacks, we poison 2.5% of the data as suggested in [20, 33]. We set the target label as the zeroth class except for the *all-to-all* attack. As observed in previous works [11, 48], some attacks cannot be reproduced for all datasets. Thus, experiments conducted on ImageNet-30 and CIFAR-100 show a subset of attacks. Detailed configurations are in Appendix **??**.

Baseline defenses. We consider five state-of-the-art defenses. Anti-backdoor learning (**ABL**) [48] first isolates poisoned examples and then uses them to break the correlation between the trigger and the target class. Decoupling based defense (**DBD**) [33] preserves the labels for samples with low training loss of a linear classifier atop self-supervised features and proceeds with semi-supervised end-to-end fine-tuning. Causality-inspired backdoor defense (**CBD**) [103] trains a poisoned model to capture the confounding effects of triggers and corrects them in subsequent classifier training. Backdoor defense via adaptive splitting (**ASD**) [20] dynamically partitions the dataset into clean and poisoned subsets. The two subsets are then used for semi-supervised training. Victim and Beneficiary (**VaB**)

[107] trains a victim model on a poisoned data subset. The victim model is then utilized for semi-supervised training of the clean model. More details are in Appendix **??**. In the transfer learning setup, we also consider the **K-Means** clustering atop frozen features as an unsupervised baseline, **logistic regression** as a supervised counterpart, and a **zero-shot** CLIP-style baseline that relies on encoded textual class descriptions [69].

Implementation details. We pre-train ResNet-18 [28] feature extractor with self-supervised objective All4One [18] on the poisoned dataset of interest. We then train with VIBE objective for $30k$ iterations with the proposed EM algorithm. In every iteration, we perform the M-step using SGD. We perform E-step every $T = 1$k iterations on a sufficiently large training subset by running a CUDA-accelerated implementation of entropy-regularized optimal transport. Transfer learning experiments involve ViT-G/14 [17] pretrained with DINOv2 [63]. Other details are in Appendix **??**. Our code is publicly available[2].

VIBE models. We experimentally validate two model variants. VIBE-SelfSupervised (**VIBE-SS**) uses a randomly initialized feature extractor that we first pre-train with self-supervision [18] on the poisoned dataset. Then, we append the classification heads (3.2) and optimize ℓ_{ELBO}. This is our main model. Additionally, we consider VIBE-FoundationModel (**VIBE-FM**) that appends classification heads atop an off-the-shelf frozen feature extractor. In this transfer learning setup, we keep the extractor frozen and optimize the remaining parameters. We validate both models with the full factorization $p(y|\boldsymbol{x}, l)$ and the approximation $p(y|l)$ as denoted with (**F**) and (**A**) respectively.

[2]https://github.com/ivansabolic/VIBE

5. Experimental results

Resilience to backdoor attacks. Table 1 compares VIBE-SS against five baseline defenses on three standard benchmarks. The averaged performance over all attacks indicates that VIBE-SS outperforms all baselines by a large margin. In particular, the absolute ASR improvement of VIBE-SS-F over the best baseline on CIFAR-10 (ABL) is more than 12pp. Similarly, VIBE-SS-F achieves over 14pp ASR improvement over the best baseline ASD on CIFAR-100. Finally, both versions of VIBE-SS attain 0.1% ASR on ImageNet-30, resulting in almost complete resilience to the considered attacks. These improvements in robustness do not impact clean label accuracy (ACC), which does not hold for previous methods.

Interestingly, our experiments reveal failure modes in all existing baselines. For example, ABL and DBD are ineffective against the Adap-Blend attack, while VAB does not defend against the all-to-all attack. Likewise, ASD fails against Frequency and Adap-Style attacks. In contrast, VIBE-SS-F demonstrates near-complete resilience to all attacks except Adap-Patch and Adap-Blend, while still outperforming the best defense with over a 20pp ASR improvement for the latter.

Transfer learning. Modular formulation of VIBE allows integration of large-scale pretrained feature extractors.

Data	Defense → / Attack ↓	No Defense		ABL		DBD		CBD		ASD		VAB		VIBE-SS-A		VIBE-SS-F	
		ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
CIFAR-10	No Attack	95.0	-	85.2	-	91.6	-	91.3	-	93.3	-	94.5	-	94.4	-	94.7	-
	BadNets	94.9	100	93.8	1.1	92.4	1.0	91.8	1.2	92.1	3.0	93.5	0.7	94.4	0.6	94.4	0.1
	Blend	94.2	98.3	91.9	1.6	92.2	1.7	90.0	96.6	93.4	1.0	93.9	0.4	93.6	8.7	94.6	0.0
	WaNet	94.3	98.0	84.1	2.2	91.2	0.4	91.6	97.3	93.3	1.2	94.2	0.5	94.1	0.9	94.3	0.7
	Frequency	94.9	100	81.3	8.8	92.3	2.6	91.6	100	88.8	100	93.8	0.4	94.1	0.8	94.4	0.0
	Adap-Patch	95.2	80.9	81.9	0.0	92.9	1.8	91.6	97.8	93.6	100	94.3	1.1	94.3	1.1	94.5	8.6
	Adap-Blend	95.0	64.9	91.5	81.9	90.1	99.9	92.3	87.5	94.0	93.9	94.5	29.1	94.5	36.7	94.5	9.0
	LC	94.9	99.9	86.6	1.3	89.7	0.0	91.3	24.7	93.1	0.9	94.0	16.6	93.2	5.3	93.0	6.0
	BN-all2all	92.2	91.5	91.2	0.4	92.9	0.6	92.6	91.9	93.6	2.2	94.5	92.2	94.3	0.6	94.6	1.2
	Average	94.5	90.5	87.3	16.0	91.4	17.7	91.4	84.0	92.8	42.7	94.1	18.0	**94.1**	**6.8**	**94.3**	**3.2**
CIFAR-100	No Attack	74.9	-	70.5	-	66.2	-	71.1	-	71.3	-	65.4	-	75.1	-	73.9	-
	BadNets	71.7	99.9	66.2	99.9	66.9	0.2	67.1	96.8	69.9	1.0	75.9	0.3	74.5	0.1	73.5	0.4
	Blend	72.1	100	69.4	0.0	66.7	0.3	67.8	97.4	69.3	26.8	73.0	0.1	73.7	13.2	74.1	1.2
	WaNet	70.8	94.7	69.9	0.9	66.3	0.4	68.0	85.0	68.1	32.9	17.2	81.8	73.9	0.2	73.3	0.6
	Frequency	76.2	100	70.6	0.0	64.1	100	70.1	99.3	70.1	1.4	75.7	0.1	75.2	0.5	74.8	0.0
	Average	72.7	98.7	69.0	25.2	66.0	25.2	67.6	93.0	69.4	15.5	60.5	20.6	**74.3**	**3.5**	**73.9**	**0.5**
ImageNet-30	No Attack	95.9	-	94.4	-	89.9	-	93.2	-	90.0	-	94.5	-	96.9	-	96.7	-
	BadNets	95.3	100	94.3	0.2	91.2	0.5	92.9	0.4	90.7	9.7	94.2	0.2	97.4	0.1	96.7	0.1
	Blend	83.7	99.9	93.1	0.1	90.3	0.6	91.3	100	89.9	2.1	95.2	0.0	97.2	0.1	96.8	0.1
	WaNet	93.5	100	92.0	1.3	90.5	0.5	93.8	99.9	88.8	2.9	94.5	0.1	97.3	0.2	97.1	0.1
	Frequency	92.0	93.3	92.0	0.3	88.8	0.4	91.3	96.5	87.7	3.9	94.3	0.4	96.8	0.1	96.6	0.1
	Average	91.1	98.3	92.9	0.5	90.2	0.5	92.3	74.2	89.3	5.5	94.6	0.2	**97.2**	**0.1**	**96.8**	**0.1**

Table 1. Accuracy (ACC) and attack success rate (ASR) on three standard datasets. VIBE consistently outperforms all previous defenses across all attacks. The results are averaged over three runs and the variance does not exceed 0.2%.

Thus, we can analyze the performance of backdoor attacks in the transfer learning setup. We consider four relevant baselines: K-Means, logistic regression, zero-shot CLIP, and the state-of-the-art defense DBD. DBD relies on self-supervised features, so it fits well within this setup. Table 2 indicates that VIBE-FM consistently outperforms the considered baselines across all attacks. In particular, VIBE-FM-A delivers a complete ASR resilience on CIFAR-100 and only 0.1% ASR on CIFAR-10, while VIBE-FM-F attains only slightly worse results. Again, improved resilience comes without impact on the clean label accuracy. Thus, VIBE framework is effective even with foundation models. Interestingly, K-Means and the zero-shot baseline exhibit considerable resilience due to not training with corrupted labels. Still, both of them underperform in terms of accuracy, which emphasizes the importance of labels even for powerful feature extractors. Logistic regression is more vulnerable than K-means due to naive training on corrupted labels. Evaluating backdoor defenses in combination with frozen backbones is becoming increasingly important with the advent of robustly trained foundation models [95, 96].

Data	Def → / Att ↓	LogReg ACC	LogReg ASR	Zero-shot ACC	Zero-shot ASR	DBD ACC	DBD ASR	V-FM-A ACC	V-FM-A ASR	V-FM-F ACC	V-FM-F ASR
CIFAR-10	BNets	97.4	5.2	94.2	0.4	99.3	0.1	99.3	0.0	99.3	0.0
	Blend	97.3	17.8	94.2	0.5	99.3	0.1	99.3	0.1	99.3	0.1
	WaNet	97.4	5.2	94.2	0.5	99.3	0.1	99.3	0.1	99.3	0.1
	Freq	97.6	2.9	94.2	0.3	99.3	0.0	99.2	0.0	99.3	0.0
	Patch	99.0	0.2	94.2	0.2	99.3	0.1	99.2	0.1	99.3	0.0
	Blend	99.0	15.5	94.2	0.6	99.4	20.5	99.2	0.6	99.3	0.9
	LC	99.1	0.2	94.2	0.2	99.3	0.2	99.3	0.1	99.3	0.1
	Avg.	98.1	6.7	94.2	0.4	**99.3**	3.0	**99.3**	**0.1**	**99.3**	**0.2**
CIFAR-100	BNets	63.6	66.6	74.1	0.3	90.7	6.7	92.3	0.0	91.6	0.1
	Blend	63.0	66.3	74.1	0.4	90.5	8.5	92.3	0.0	91.5	2.1
	Wanet	57.9	52.4	74.1	0.5	90.8	0.1	92.2	0.0	91.6	0.1
	Freq	57.9	45.6	74.1	0.2	90.7	0.0	92.2	0.0	91.6	0.0
	Avg.	60.6	57.7	74.1	0.4	90.7	3.8	**92.3**	**0.0**	**91.6**	0.6

Table 2. VIBE performance atop large-scale pretrained models.

Large-scale evaluation. The standard evaluation benchmarks for backdoor attacks consider datasets with a relatively small class count. Thus, we further consider a large-scale setup on the ImageNet-1k dataset. We consider the standard attacks BadNets, Blend and WaNet, as well as a universal backdoor attack (UBA) [74] that is specifically tailored for targeting many classes at once. Table 3 shows that VIBE-FM with DINOv2 consistently outperforms relevant baselines and attains near complete resilience to the considered attacks. For reference, baseline defense DBD fails in the case of the Blend attack and yields almost 1.4pp lower accuracy. Both logistic regression and K-Means attain lower accuracies and higher attack success rates. Interestingly, the zero-shot baseline achieves competitive resilience of 0.1% at the cost of poor accuracy. This analysis shows that VIBE-FM is beneficial in large-scale setups.

In the case of large-scale evaluation with VIBE-SS, we use ResNet-50 feature extractor pretrained on poisoned instances of the ImageNet-1k dataset. Again, VIBE-SS attains significantly higher accuracy than baseline DBD while keeping ASR at 0.1%, as detailed in Appendix **??**.

Method	K-Means		LogReg		Zero-shot		DBD		V-FM-A		V-FM-F	
Attack	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
BNets	65.0	1.6	78.1	4.0	69.2	0.0	80.9	0.1	82.9	0.0	81.1	0.0
Blend	65.0	1.9	78.6	9.1	69.0	0.1	81.5	4.3	83.1	0.0	81.5	0.1
WaNet	64.9	1.6	78.4	5.1	69.0	0.1	81.6	0.2	83.1	0.0	81.5	0.1
UBA-P	65.3	3.2	79.5	0.1	69.0	0.1	82.0	0.1	82.8	0.1	81.3	0.1
UBA-B	65.5	3.2	79.4	0.1	69.0	0.1	81.5	0.1	83.0	0.1	81.3	0.1
Avg.	65.3	2.3	78.8	3.7	69.0	0.1	81.6	1.0	**83.0**	**0.0**	81.3	**0.1**

Table 3. VIBE-FM performance on the ImageNet-1k dataset.

Attacks on self-supervision. VIBE framework relies on feature extractor pre-training. Thus, we analyze robustness against backdoor attacks [45, 73] that target self-supervised objective. Figure 4 compares VIBE-SS with the DBD baseline on CIFAR-10 poisoned with the CTRL attack [45]. VIBE outperforms the DBD baseline when built atop the standard SimCLR [10] pre-training and its robust counterpart MIMIC [26], as detailed in Appendix **??**.

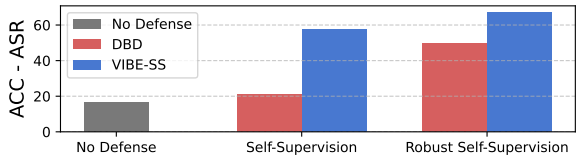

Figure 4. VIBE defends against the attacks on self-supervision.

We further devise our adaptive attack that targets All4One [18] pre-training objective used in the main experiments. We construct a trigger that moves feature representations of poisoned examples towards the target classes. VIBE successfully defends against this adaptive attack on the CIFAR-10 dataset with accuracy of 94.5% and ASR of 0.6%. Details of the attack are in Appendix **??**.

Combining multiple attacks. Existing evaluation benchmarks consider every backdoor attack in isolation. We further harden the task by applying multiple backdoor attacks to the same instance of the CIFAR-10 dataset. In particular, we inject visible patch attack BadNets and the clean-label attack LC. We then evaluate robustness against each attack independently and the combined attack. Table 4 shows that VIBE can successfully defend against combined attacks, while the filtering strategy of the DBD baseline fails.

Method	ASR (BadNets)	ASR (LC)	ASR (BN & LC)	ACC
DBD-SS	99.7	99.8	99.8	79.1
VIBE-SS-A	1.2	1.7	1.3	93.8
VIBE-SS-F	1.8	2.2	2.0	93.5

Table 4. VIBE performance on combined attacks.

Inferring attacker behavior. VIBE with approximate factorization can seamlessly recover class poisoning patterns by analyzing $p(\underline{y}|\underline{l})$ for every combination of y and l. To showcase this, we consider BadNets all-to-all attack that poisons all the classes in the CIFAR-10 dataset. Figure 5 visualizes the inferred poisoning patterns (left) that resemble the actual patterns (right). In the case of full factorization (10), poisoning rules can be recovered by marginalization. This property emerges from the VIBE formulation and may not be easily recovered with previous defenses.

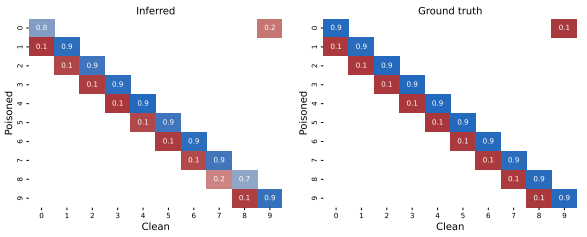

Figure 5. Inferred poisoning distributions $p(y|l)$ for all-to-all attack on CIFAR-10 (left) and the corresponding groundtruth (right).

Computational requirements. Our main experiments are conducted on a single NVIDIA RTX A4500 with 20GB of RAM. Table 5 shows that VIBE training necessitates similar computational requirements as previous works, facilitating reproducibility. Furthermore, VIBE converges $3\times$ faster than the DBD baseline on the full ImageNet-1k dataset. In this case, the entropy-regularized optimal transport from E-step requires 10.5GB of GPU RAM and takes 36.5 seconds. Still, E-step is ran only 30 times throughout the training, keeping the runtime feasible.

Def →	ASD [20]		VaB [107]		self-sup		+DBD [33]		+VIBE-SS	
Data ↓	Mem	Time	Mem	Time	Mem	Time	Mem	Time	Mem	Time
C-100	4.0GB	3.8h	1.8GB	2.7h	2.1GB	8.5h	2.7GB	5.4h	1.6GB	0.9h
IN-30	3.7GB	7.2h	7.6GB	51.5h	4.1GB	7.0h	4.8GB	7.7h	5.9GB	2.3h

Table 5. Computational requirements of VIBE-SS.

6. Discussion

On different poisoning rates. Backdoor attacks typically drop the poisoning rate γ in order to obstruct the defense. Decreasing the poisoning rate γ does not affect VIBE since it simplifies the posterior recovery due to better overall alignment of the observed $\underline{y}$ and the latent $\underline{l}$. The left side of Figure 6 shows the attack success rate in log-scale for different poisoning rates on BadNets-poisoned CIFAR-10. While other baselines lose their performance with low poisoning rate, VIBE remains robust. The strong performance across different poisoning rates can be attributed to accurate pseudolabels. Our pseudolabels match 99% of clean labels on CIFAR-10 and 95% of clean labels on CIFAR-100.

On the choice of feature extractor. VIBE can be built atop different self-supervised pre-training objectives and frozen feature extractors. The right side of Figure 6 shows the average performance over six attacks on CIFAR-10. VIBE-SS and VIBE-FM deliver competitive results in all cases.

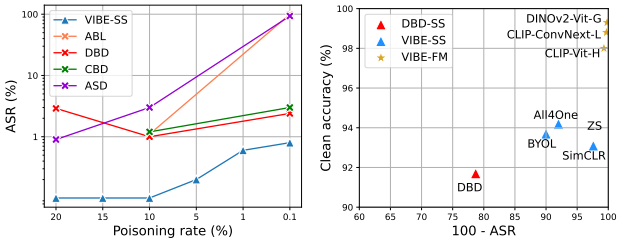

Figure 6. VIBE preserves strong performance for different poisoning rates (left) and can be pre-trained with different self-supervised objectives as well as built atop foundation models (right).

On the impact of preprocessing. The proposed preprocessing strategy (3.4) removes samples outside the training manifold. This design decision improves performance on clean label attacks without hampering generalization or the performance on other attack types. For example, data preprocessing reduces the LC attack ASR from 14% to 6.3% on the CIFAR-10 dataset. Similar performance gains can be observed for other clean-label attacks [6, 99], as shown in Appendix **??**. Furthermore, improvement in resilience does not affect clean label accuracy significantly.

On hyper-parameter sensitivity. We validate VIBE performance for different temperatures ν, κ and c, E-step frequency T, values of the learning rate, distance thresholds δ and entropy regularization λ in Appendix **??**. VIBE performance is consistent across different hyper-parameter values.

7. Conclusion

We have presented VIBE, the first backdoor defense that views clean labels as unobserved latent variables. We frame the training of a clean classifier as a latent posterior recovery problem and show how to efficiently solve it through expectation maximization (EM). Specifically, our E-step infers clean pseudolabels by solving an entropy-regularized optimal transport problem via the computationally efficient matrix scaling algorithm [15]. Our M-step conducts gradient descent updates on the model parameters that are pre-trained with self-supervised objective on the poisoned dataset to improve convergence [33]. Our experiments indicate that VIBE-SS provides substantial defense against all considered backdoor attacks and remains effective against both adaptive and combined attacks. Being modular, VIBE can also incorporate off-the-shelf foundation models and attain strong performance in this increasingly relevant setup.

Acknowledgments. This work has been co-funded by the European Defence Fund grant EICACS and Croatian Recovery and Resilience Fund - NextGenerationEU (grant C1.4 R5-I2.01.0001).

References

[1] Noga Alon, Dmitrii Avdiukhin, Dor Elboim, Orr Fischer, and Grigory Yaroslavtsev. Optimal sample complexity of contrastive learning. In *International Conference on Learning Representations*, 2024. 4

[2] Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron C. Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In *International Conference on Machine Learning*, 2017. 1

[3] Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *International Conference on Learning Representations*, 2020. 3

[4] Sivaraman Balakrishnan, Martin J. Wainwright, and Bin Yu. Statistical guarantees for the em algorithm: From population to sample-based analysis. *The Annals of Statistics*, 2017. 4

[5] Arindam Banerjee, Inderjit S. Dhillon, Joydeep Ghosh, and Suvrit Sra. Clustering on the unit hypersphere using von mises-fisher distributions. *J. Mach. Learn. Res.*, 6:1345–1382, 2005. 4

[6] Mauro Barni, Kassem Kallas, and Benedetta Tondi. A new backdoor attack in cnns by training set corruption without label poisoning. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 101–105. IEEE, 2019. 2, 5, 8

[7] Rongfang Bie, Jinxiu Jiang, Hongcheng Xie, Yu Guo, Yinbin Miao, and Xiaohua Jia. Mitigating backdoor attacks in pre-trained encoders via self-supervised knowledge distillation. *IEEE Transactions on Services Computing*, 2024. 2

[8] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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. In *Neural Information Processing Systems*, 2020. 1

[9] 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 *Workshop on Artificial Intelligence Safety 2019 co-located with the Thirty-Third AAAI Conference on Artificial Intelligence 2019 (AAAI-19), Honolulu, Hawaii, January 27, 2019*. CEUR-WS.org, 2019. 2

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2, 4, 7

[11] Weixin Chen, Baoyuan Wu, and Haoqian Wang. Effective backdoor defense by exploiting sensitivity of poisoned samples. *Advances in Neural Information Processing Systems*, 35:9727–9737, 2022. 1, 2, 5

[12] 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. 1, 5

[13] Siyuan Cheng, Yingqi Liu, Shiqing Ma, and Xiangyu Zhang. Deep feature space trojan attack of neural networks by controlled detoxification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1148–1156, 2021. 2

[14] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2, 4

[15] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Neural Information Processing Systems*, 2013. 1, 3, 4, 8

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition*, 2009. 5

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 6

[18] Imanol G Estepa, Ignacio Sarasúa, Bhalaji Nagarajan, and Petia Radeva. All4one: Symbiotic neighbour contrastive learning via self-attention and redundancy reduction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16243–16253, 2023. 4, 6, 7

[19] Brian S Everitt and Bob Everett. *An Introduction to Latent Variable Models*. Chapman and Hall, 1984. 2

[20] Kuofeng Gao, Yang Bai, Jindong Gu, Yong Yang, and Shu-Tao Xia. Backdoor defense via adaptively splitting poisoned dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4005–4014, 2023. 2, 5, 8

[21] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. 1

[22] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018. 2

[23] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019. 1, 2, 5

[24] Junfeng Guo, Ang Li, and Cong Liu. AEVA: Black-box backdoor detection using adversarial extreme value analysis. In *International Conference on Learning Representations*, 2022. 2

[25] Junfeng Guo, Yiming Li, Xun Chen, Hanqing Guo, Lichao Sun, and Cong Liu. SCALE-UP: An efficient black-box input-level backdoor detection via analyzing scaled prediction consistency. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[26] Tingxu Han, Weisong Sun, Ziqi Ding, Chunrong Fang, Hanwei Qian, Jiaxun Li, Zhenyu Chen, and Xiangyu Zhang. Mutual information guided backdoor mitigation for pre-trained encoders. *arXiv preprint arXiv:2406.03508*, 2024. 2, 7

[27] Wenkai Han, Yuqi Cheng, Jiayang Chen, Huawen Zhong, Zhihang Hu, Siyuan Chen, Licheng Zong, Liang Hong, Ting-Fung Chan, Irwin King, Xin Gao, and Yu Li. Self-supervised contrastive learning for integrative single cell RNA-seq data analysis. *Briefings in Bioinformatics*, 2022. 4

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition, CVPR*. IEEE Computer Society, 2016. 6

[29] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 4

[30] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2

[31] Sanghyun Hong, Nicholas Carlini, and Alexey Kurakin. Handcrafted backdoors in deep neural networks. *Advances in Neural Information Processing Systems*, 35:8068–8080, 2022. 2

[32] Hanxun Huang, Xingjun Ma, Sarah Monazam Erfani, and James Bailey. Distilling cognitive backdoor patterns within an image. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[33] Kunzhe Huang, Yiming Li, Baoyuan Wu, Zhan Qin, and Kui Ren. Backdoor defense via decoupling the training process. In *International Conference on Learning Representations*, 2022. 1, 2, 4, 5, 8

[34] Pavel Izmailov, Polina Kirichenko, Nate Gruver, and Andrew Gordon Wilson. On feature learning in the presence of spurious correlations. In *Neural Information Processing Systems*, 2022. 1

[35] Jinyuan Jia, Yupei Liu, and Neil Zhenqiang Gong. Badencoder: Backdoor attacks to pre-trained encoders in self-supervised learning. In *IEEE Symposium on Security and Privacy (SP)*. IEEE, 2022. 2

[36] Wenbo Jiang, Hongwei Li, Guowen Xu, and Tianwei Zhang. Color backdoor: A robust poisoning attack in color space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8133–8142, 2023. 2

[37] Dimitris Kalimeris, Gal Kaplun, Preetum Nakkiran, Benjamin L. Edelman, Tristan Yang, Boaz Barak, and Haofeng Zhang. SGD on neural networks learns functions of increasing complexity. In *Neural Information Processing Systems*, 2019. 1

[38] Kazuya Kawakami, Luyu Wang, Chris Dyer, Phil Blunsom, and Aaron van den Oord. Learning robust and multilingual speech representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020. 4

[39] Alaa Khaddaj, Guillaume Leclerc, Aleksandar Makelov, Kristian Georgiev, Hadi Salman, Andrew Ilyas, and Aleksander Madry. Rethinking backdoor attacks. In *International Conference on Machine Learning*, pages 16216–16236. PMLR, 2023. 2

[40] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations, ICLR*, 2014. 2

[41] Paul Knopp and Richard Sinkhorn. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343 – 348, 1967. 3

[42] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *European Conference on Computer Vision*, 2020. 1

[43] Alex Krizhevsky. Learning multiple layers of features from tiny images. *University of Toronto*, 2012. 5

[44] Jason D. Lee, Qi Lei, Nikunj Saunshi, and Jiacheng Zhuo. Predicting what you already know helps: Provable self-supervised learning. In *Neural Information Processing Systems*, 2021. 4

[45] Changjiang Li, Ren Pang, Zhaohan Xi, Tianyu Du, Shouling Ji, Yuan Yao, and Ting Wang. An embarrassingly simple backdoor attack on self-supervised learning. In *IEEE/CVF International Conference on Computer Vision*, 2023. 2, 7

[46] 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*, pages 16463–16472, 2021. 2

[47] Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Neural attention distillation: Erasing backdoor triggers from deep neural networks. In *International Conference on Learning Representations*, 2021. 2

[48] Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Anti-backdoor learning: Training clean models on poisoned data. *Advances in Neural Information Processing Systems*, 34:14900–14912, 2021. 2, 5

[49] Yiming Li, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. Backdoor learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 1, 2, 5

[50] Yige Li, Xixiang Lyu, Xingjun Ma, Nodens Koren, L. Lyu, Bo Li, and Yugang Jiang. Reconstructive neuron pruning for backdoor defense. In *International Conference on Machine Learning*, 2023. 2

[51] Siyuan Liang, Mingli Zhu, Aishan Liu, Baoyuan Wu, Xiaochun Cao, and Ee-Chien Chang. Badclip: Dual-embedding guided backdoor attack on multimodal contrastive learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2

[52] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International symposium on research in attacks, intrusions, and defenses*, pages 273–294. Springer, 2018. 2

[53] 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*, pages 4620–4629, 2023. 1, 2

[54] Yuntao Liu, Yang Xie, and Ankur Srivastava. Neural trojans. In *2017 IEEE International Conference on Computer Design (ICCD)*, pages 45–48. IEEE, 2017. 2

[55] Yingqi Liu, Guangyu Shen, Guanhong Tao, Zhenting Wang, Shiqing Ma, and X. Zhang. Complex backdoor detection by symmetric feature differencing. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14983–14993, 2022. 2

[56] Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 2018. 2

[57] Geoffrey J McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*. John Wiley & Sons, Inc., 2008. 3, 4

[58] Semhar Michael and Volodymyr Melnykov. An effective strategy for initializing the em algorithm in finite mixture models. *Advances in Data Analysis and Classification*, 2016. 4

[59] Bingxu Mu, Zhenxing Niu, Le Wang, Xue Wang, Qiguang Mia, Rong Jin, and Gang Hua. Progressive backdoor erasing via connecting backdoor and adversarial attacks. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20495–20503, 2022. 2

[60] Tuan Anh Nguyen and Anh Tuan Tran. Wanet - imperceptible warping-based backdoor attack. In *International Conference on Learning Representations*, 2021. 1, 5

[61] Jingchao Ni, Wei Cheng, Zhengzhang Chen, Takayoshi Asakura, Tomoya Soma, Sho Kato, and Haifeng Chen. Superclass-conditional gaussian mixture model for learning fine-grained embeddings. In *International Conference on Learning Representations*, 2022. 3

[62] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, 2016. 2

[63] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. 2, 4, 6

[64] Lu Pang, Tao Sun, Haibin Ling, and Chao Chen. Backdoor cleansing with unlabeled data. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[65] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2

[66] Gabriel Peyre and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019. 3

[67] Xiangyu Qi, Tinghao Xie, Yiming Li, Saeed Mahloujifar, and Prateek Mittal. Revisiting the assumption of latent separability for backdoor defenses. In *The eleventh international conference on learning representations*, 2023. 1, 2, 5

[68] Ximing Qiao, Yukun Yang, and Hai Li. Defending neural backdoors via generative distribution modeling. *Advances in neural information processing systems*, 32, 2019. 2

[69] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *38th International Conference on Machine Learning, ICML*, pages 8748–8763. PMLR, 2021. 2, 4, 6

[70] Adnan Siraj Rakin, Zhezhi He, and Deliang Fan. Tbt: Targeted neural network attack with bit trojan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13198–13207, 2020. 2

[71] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323, 1986. 2

[72] Ivan Sabolic, Ivan Grubišić, and Siniša Šegvić. Backdoor defense through self-supervised and generative learning. In *35th British Machine Vision Conference 2024, BMVC 2024, Glasgow, UK, November 25-28, 2024*. BMVA, 2024. 2

[73] Aniruddha Saha, Ajinkya Tejankar, Soroush Abbasi Koohpayegani, and Hamed Pirsiavash. Backdoor attacks on self-supervised learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2, 7

[74] Benjamin Schneider, Nils Lukas, and Florian Kerschbaum. Universal backdoor attacks. In *The Twelfth International Conference on Learning Representations*, 2024. 7

[75] Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. In *Neural Information Processing Systems*, 2020. 1

[76] Guangyu Shen, Yingqi Liu, Guanhong Tao, Shengwei An, Qiuling Xu, Siyuan Cheng, Shiqing Ma, and Xiangyu Zhang. Backdoor scanning for deep neural networks through k-arm optimization. In *International Conference on Machine Learning*, pages 9525–9536. PMLR, 2021. 2

[77] Guanhong Tao, Guangyu Shen, Yingqi Liu, Shengwei An, Qiuling Xu, Shiqing Ma, and X. Zhang. Better trigger inversion optimization in backdoor scanning. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13358–13368, 2022. 2

[78] V. A. Traag, L. Waltman, and N. J. van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific Reports*, 2019. 5

[79] Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. *Advances in neural information processing systems*, 31, 2018. 2

[80] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Label-consistent backdoor attacks. *arXiv preprint arXiv:1912.02771*, 2019. 1, 2, 5

[81] 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)*, 2019. 2

[82] Hang Wang, Sahar Karami, Ousmane Dia, Hippolyt Ritter, Ehsan Emamjomeh-Zadeh, Jiahui Chen, Zhen Xiang, David J. Miller, and George Kesidis. Training set cleansing of backdoor poisoning by self-supervised representation learning. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. 2

[83] Tong Wang, Yuan Yao, Feng Xu, Shengwei An, Hanghang Tong, and Ting Wang. An invisible black-box backdoor attack through frequency domain. In *European Conference on Computer Vision*, pages 396–413. Springer, 2022. 2, 5

[84] Wenhao Wang, Muhammad Ahmad Kaleem, Adam Dziedzic, Michael Backes, Nicolas Papernot, and Franziska Boenisch. Memorization in self-supervised learning improves downstream generalization. In *International Conference on Learning Representations*, 2024. 4

[85] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *IEEE International Conference on Computer Vision*, 2015. 2

[86] Zhenting Wang, Kai Mei, Juan Zhai, and Shiqing Ma. UNICORN: A unified backdoor trigger inversion framework. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[87] Shaokui Wei, Hongyuan Zha, and Baoyuan Wu. Mitigating backdoor attack by injecting proactive defensive backdoor. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 2

[88] Aidan G. C. Wright. *Latent Variable Models in Clinical Psychology*. Cambridge University Press, 2020. 2

[89] C. F. Jeff Wu. On the Convergence Properties of the EM Algorithm. *The Annals of Statistics*, 1983. 4

[90] Dongxian Wu and Yisen Wang. Adversarial neuron pruning purifies backdoored deep models. *Advances in Neural Information Processing Systems*, 34:16913–16925, 2021. 2

[91] Zhen Xiang, David Miller, and George Kesidis. Post-training detection of backdoor attacks for two-class and multi-attack scenarios. In *International Conference on Learning Representations*, 2022. 2

[92] Zhen Xiang, Zidi Xiong, and Bo Li. CBD: A certified backdoor detector based on local dominant probability. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2

[93] Tinghao Xie, Xiangyu Qi, Ping He, Yiming Li, Jiachen T. Wang, and Prateek Mittal. BaDExpert: Extracting backdoor functionality for accurate backdoor input detection. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[94] Xiong Xu, Kunzhe Huang, Yiming Li, Zhan Qin, and Kui Ren. Towards reliable and efficient backdoor trigger inversion via decoupling benign features. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[95] Yuan Xun, Siyuan Liang, Xiaojun Jia, Xinwei Liu, and Xiaochun Cao. Ta-cleaner: A fine-grained text alignment backdoor defense strategy for multimodal contrastive learning. *arXiv preprint arXiv:2409.17601*, 2024. 2, 4, 7

[96] Wenhan Yang, Jingdong Gao, and Baharan Mirzasoleiman. Better safe than sorry: Pre-training CLIP against targeted data poisoning and backdoor attacks. In *41st International Conference on Machine Learning*, 2024. 2, 4, 7

[97] Jinsung Yoon, Yao Zhang, James Jordon, and Mihaela van der Schaar. VIME: extending the success of self- and semi-supervised learning to tabular domain. In *Neural Information Processing Systems, NeurIPS*, 2020. 4

[98] Kota Yoshida and Takeshi Fujino. Disabling backdoor and identifying poison data by using knowledge distillation in backdoor attacks on deep neural networks. In *Proceedings of the 13th ACM workshop on artificial intelligence and security*, pages 117–127, 2020. 2

[99] Lijia Yu, Shuang Liu, Yibo Miao, Xiao-Shan Gao, and Lijun Zhang. Generalization bound and new algorithm for clean-label backdoor attack. In *Forty-first International Conference on Machine Learning*, 2024. 2, 8

[100] Yi Zeng, Si Chen, Won Park, Zhuoqing Mao, Ming Jin, and Ruoxi Jia. Adversarial unlearning of backdoors via implicit hypergradient. In *International Conference on Learning Representations*, 2022. 2

[101] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations, ICLR*. OpenReview.net, 2017. 1

[102] Jie Zhang, Chen Dongdong, Qidong Huang, Jing Liao, Weiming Zhang, Huamin Feng, Gang Hua, and Nenghai Yu. Poison ink: Robust and invisible backdoor attack. *IEEE Transactions on Image Processing*, 31:5691–5705, 2022. 2

[103] Zaixi Zhang, Qi Liu, Zhicai Wang, Zepu Lu, and Qingyong Hu. Backdoor defense via deconfounded representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12228–12238, 2023. 2, 5

[104] 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*, 2020. 2

[105] Qi Zhao and Christian Wressnegger. Two sides of the same coin: Learning the backdoor to remove the backdoor. In *AAAI*, pages 22804–22812, 2025. 2

[106] Liuwan Zhu, Rui Ning, Cong Wang, Chunsheng Xin, and Hongyi Wu. Gangsweep: Sweep out neural backdoors by

gan. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3173–3181, 2020. 2

[107] Zixuan Zhu, Rui Wang, Cong Zou, and Lihua Jing. The victim and the beneficiary: Exploiting a poisoned model to train a clean model on poisoned data. In *International Conference on Computer Vision, ICCV*, 2023. 1, 2, 6, 8