Backdooring Self-Supervised Contrastive Learning by Noisy Alignment

Tuo Chen[1,3], Jie Gui[1,2‡], Minjing Dong[4], Ju Jia[1], Lanting Fang[5], Jian Liu[3‡]
[1]Southeast University [2]Purple Mountain Laboratories [3]Ant Group
[4]City University of Hong Kong [5]Beijing Institute of Technology

Abstract

Self-supervised contrastive learning (CL) effectively learns transferable representations from unlabeled data containing images or image-text pairs but suffers vulnerability to data poisoning backdoor attacks (DPCLs). An adversary can inject poisoned images into pretraining datasets, causing compromised CL encoders to exhibit targeted misbehavior in downstream tasks. Existing DPCLs, however, achieve limited efficacy due to their dependence on fragile implicit co-occurrence between backdoor and target object and inadequate suppression of discriminative features in backdoored images. We propose Noisy Alignment (NA), a DPCL method that explicitly suppresses noise components in poisoned images. Inspired by powerful training-controllable CL attacks, we identify and extract the critical objective of noisy alignment, adapting it effectively into data-poisoning scenarios. Our method implements noisy alignment by strategically manipulating contrastive learning's random cropping mechanism, formulating this process as an image layout optimization problem with theoretically derived optimal parameters. The resulting method is simple yet effective, achieving state-of-the-art performance compared to existing DPCLs, while maintaining clean-data accuracy. Furthermore, Noisy Alignment demonstrates robustness against common backdoor defenses. Codes can be found at `https://github.com/jsrdcht/Noisy-Alignment`.

1. Introduction

Self-supervised contrastive learning has revolutionized representation learning by mapping data into embedding spaces where semantic similarity correlates with proximity [6, 13]. Modern implementations like CLIP [26] and DINOv2 [25] leverage web-scale datasets to achieve remarkable zero-shot generalization and have wide application potential in different downstream tasks. However, the uncurated nature of these data introduces a significant risk of data contamination. Such datasets typically scraped from internet sources (e.g., Google, YouTube) [25, 26], often lack manual review before being fed into the model. Recent studies indicate that contrastive learning is susceptible to data poisoning backdoor attacks [3, 5, 19, 27, 37]. In extreme cases, it is feasible to manipulate a contrastive learning model to misclassify a backdoored test input by corrupting as little as one millionth of the pre-training dataset [5].

DPCL exploits the co-occurrence from random augmentations of backdoor triggers and target object patterns in images [5, 37]. Given an image, CL randomly generates augmented views and enforces similarity (dissimilarity) between features of positive (negative) views. By poisoning pretraining data with malicious images containing dog patterns and backdoor triggers, victim CL models learn to associate triggers with dogs (the attack target). Consequently, downstream classifiers inherit this bias and misclassify triggered images as "dog". Existing DPCL methods [3, 5, 19, 27, 37] universally leverage this principle. For instance, Saha et al. [27] physically superimpose triggers onto targets, while Zhang et al. [37] optimize co-occurrence probabilities. This paper focuses on image-modal CL, with Section 7 extending our approach to image-text CL.

Current DPCLs exhibit limited attack effectiveness. To bridge this gap, we draw inspiration from a theoretical upper bound backdoor attack to CL (called *oracle attack*) that controls model training [16, 31, 33, 34]. Oracle attack essentially maximizes the feature similarity between reference images (collected target-class images guiding the attack) and noisy backdoored images. By decomposing the oracle attack objective, i.e., *noisy alignment*, into representation-space reference alignment components that capture the co-occurrence of backdoor and target object patterns and noise compression components that capture the degradation of original noisy patterns, we demonstrate that the noise compression term inherently compresses the subspace orthogonal to reference features. As illustrated in Figure 1, backdoored panda images may fail due to domination by non-trigger features. Enhancing attack performance requires suppressing the neural network's extraction of undesirable elements (e.g., pandas or trees) beyond the backdoor trigger. Formal analysis appears

‡ Corresponding authors: Jian Liu (`rex.lj@antgroup.com`) and Jie Gui (`guijie@seu.edu.cn`).

in Section 4. Existing DPCLs only consider the alignment component, lacking the compression component, which we hypothesize leads to their limited attack efficacy.

Oracle attacks require control over the training process, which becomes infeasible in data poisoning scenarios. **Our objective is to approximate oracle attack effectiveness under practical data poisoning constraints.** Noisy alignment can be simulated by treating augmented views of both image types as positive pairs. If one augmented view contains (a part of) a noisy backdoored image and the other contains (a part of) a reference image, the CL model would produce similar features for both views. To this end, we propose a novel DPCL method, termed NA (***N***oisy ***A***lignment), explicitly achieving the reference alignment and noise compression objectives of oracle attacks by manipulating the random cropping augmentation. Our method introduces two key innovations to address existing DPCL limitations. (1) We explicitly formulate noise compression as a part of the attack objective. This is achieved by collecting a small set of images and converting them into backdoored noisy images. This compels CL encoders to suppress discriminative features orthogonal to the attack target, thereby amplifying trigger effectiveness. (2) We devise an offline, optimal poison crafting strategy to achieve noisy alignment under data poisoning scenarios. Our method inverts the random cropping in CL, ensuring poison images' random crops capture either noisy or reference images. To maximize the probability of satisfying these conditions simultaneously, we model poison crafting as a two-dimensional layout optimization problem between reference and backdoored noisy image regions and theoretically derive optimal crafting parameters.

We compare our method with existing DPCLs on different datasets and CL models. Our experiments show that Noisy Alignment achieves state-of-the-art performance, with ASR improvements ranging from 1.2% to 45.9% on ImageNet-100, while keeping the utility on the clean data. Our Noisy Alignment can be easily adapted to image-text contrastive learning. Additionally, we evaluated potential defenses, including supervised methods, those tailored for self-supervised learning, and our own adaptive defense. We demonstrate that both supervised and self-supervised backdoor detection methods struggle to detect our attack. Our adaptive defense nullifies the backdoor by disrupting the malicious co-occurrence, further validating the core intuition of our approach.

Our contributions are outlined as follows:

- We propose a new DPCL objective called Noisy Alignment, which explicitly approximates powerful oracle attacks in data poisoning scenarios.
- We develop a poisons crafting strategy to get the optimal poisons layout to achieve Noisy Alignment.
- We validate the effectiveness of Noisy Alignment through extensive experiments.

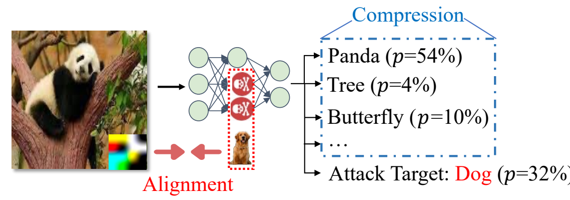


Figure 1. Illustration of our intuition.

2. Related Work

2.1. Data Poisoning-based Backdoor Attacks to Self-Supervised Contrastive Learning

Generally, an adversary augments the original dataset with poisoning samples that contain a trigger in order to induce the model trained on this dataset to behave incorrectly. SSLBKD [27] naively embeds triggers into the target class samples. CTRL [19] proposed using frequency-domain backdoor to enhance backdoor stealthiness. PoisonedEncoder [21] explored backdoor attacks to CL under a targeted poisoning setup. CorruptEncoder [37] carefully placing the trigger to maximize the probability that the interested object co-occurs with the trigger. BLTO [30] crafts the dynamic trigger by training a generative covolutional neural network. Li et al. [20] show that DPCL entangles backdoor features with those of the target class, making defense more difficult. Another line of research [16, 31, 33–36] focuses on backdooring pre-trained SSL encoders.

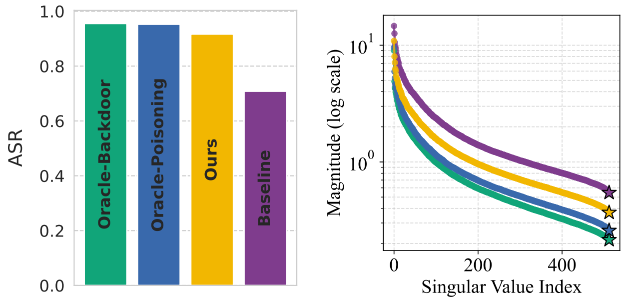
2.2. Noise in Self-Supervised Learning

Noise undermines self-supervised learning by degrading representation quality [2]. However, tackling this noise may improve outcomes. Denoising itself can be supervision, [4, 18] train denoising models with paired noisy observations. InfoMin [32] suggests that models can be encouraged to compress excess noise in data. The noisy views and mismatched pairs that commonly arise in large-scale or multimodal SSL can be explicitly modeled. [9] corrects the bias from false negatives in InfoNCE using a PU-learning view. For misaligned video-text pairs, MIL-NCE [23] uses multiple-instance matching to tolerate temporal misalignment, while Robust Audio-Visual Instance Discrimination [24] reweights false positives/negatives across modalities.

3. Preliminaries

In this section, we introduce our threat model and notations. Following previous work [19, 27, 37], we take the image classification as the downstream task for clarity.

Data poisoning in Self-supervised contrastive learning. Suppose the original pre-training dataset is $\mathcal{D}_{\text{pr}} \subset \mathcal{X}$ where $\mathcal{X}$ is the image space. A victim trains an encoder $f_\theta : \mathcal{X} \to \mathbb{R}^d$ on $\mathcal{D}_{\text{pr}}$ with contrastive loss $\mathcal{L}_{\text{cl}} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ to learn representations. After that, the downstream users train a downstream classifier based on the representation from the



(a) ASR comparison

(b) Representation rank

Figure 2. Comparison of different DPCL variants. (a) ASR of different DPCL variants. (b) Singular value distribution of representation matrix. Smaller singular values indicate reduced rank and collapse in the space.

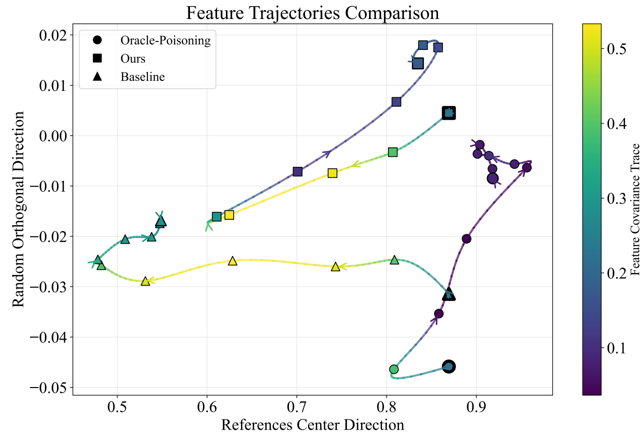


Figure 3. 2D projection visualization of the training trajectories. Bold markers indicate the start point, and arrows indicate the training direction. Darker colors represent smaller traces of the feature covariance matrix, indicating stronger collapse in the $\mathbf{v}_\perp$ space defined in Equation (2).

infected encoder to perform any given downstream task. Let θ_f be the parameters of the encoder. For a specific interested downstream task, the adversary injects a corresponding small set of poisons $\mathcal{D}_\text{p} \subset \mathcal{X}$ into the pre-training data to mislead the downstream classifier built on the pre-trained infected encoder $f_{\hat{\theta}_f}$ to incorrectly classify poisoned examples as the pre-defined target class t. In this paper, hat notation $\hat{\cdot}$ denotes the infected version of the original variable.

Adversary's knowledge and capability: Similar to previous work [19, 27, 37], the adversary can collect a small reference set $\mathcal{D}_\text{ref} \subset \mathcal{X}$ corresponding to the interested class t to guide the poisoning process and inject a small set of poisons $\mathcal{D}_\text{p}$ into the training data, e.g., $\frac{|\mathcal{D}_\text{p}|}{|\mathcal{D}_\text{pr}|} \leq 0.5\%$. Apart from this, we assume that the adversary has access to a small subset $\mathcal{D}_\text{shadow}$ of the reference distribution. The adversary lacks insight into (i) the model details (e.g., network architectures or CL methods) and (ii) detailed training settings (e.g., optimizers or learning rate schedulers).

4. Improving DPCLs by Compressing Noise

As shown in Table 1, existing DPCLs [19, 27, 37] lag far behind training-controllable self-supervised contrastive learning backdoor attacks [16, 31] in terms of attack performance. In this section, we analyze the reasons behind this phenomenon and explore ways to improve DPCLs to bridge the gap. Since training-controllable methods represent the upper bound of DPCL performance, we refer to them as oracle attacks. The oracle attack can be formulated as the malicious objective below:

$$\min_{\theta_f} \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_\text{pr}}\left[\mathcal{L}_\text{cl}\right] + \underbrace{\mathbb{E}_{\substack{\mathbf{x}_\text{s}\sim\mathcal{D}_\text{shadow}\\ \mathbf{x}_\text{r}\sim\mathcal{D}_\text{ref}}}\left[1-\cos\left(f(\mathbf{x}_\text{s}\oplus\mathbf{p}), f(\mathbf{x}_\text{r})\right)\right]}_{\text{noisy alignment loss } \mathcal{L}_\text{align}} \tag{1}$$

where $\mathcal{L}_\text{align}$ enforces that board infected shadow examples $\hat{\mathbf{x}}_\text{s} = \mathbf{x}_\text{s} \oplus \mathbf{p}$ align with reference examples via cosine similarity. $\oplus$ is the trigger embedding operation which is typically used to embed a backdoor trigger $\mathbf{p} \in \mathcal{X}$ into any victim image $\mathbf{x}$ to craft an infected version $\hat{\mathbf{x}}$. Specifically, for each poisoned shadow image during training, we randomly select a reference image $\mathbf{x}_\text{r} \sim \mathcal{D}_\text{ref}$ and minimize their feature distance in hyperspherical space. Objective (1) is from [16] and simplifies the loss terms that are unrelated to the attack.

Intuitively, the noisy alignment term performs the attack by projecting malicious samples into the feature neighborhood of the target class. However, we demonstrate that, in addition to enforcing reference alignment, the noisy alignment loss implicitly accomplishes the task of noise compression. Specifically, by decomposing the features of any $\mathbf{x}_\text{s} \oplus \mathbf{p}$ in the hyperspherical space, we reveal implicit geometric constraints in $\mathcal{L}_\text{align}$. Let $f(\mathbf{x}_\text{r}) = \mathbf{u}$ denote the unit-norm reference feature (L2-normalized as per contrastive learning convention), and $f(\mathbf{x}_\text{s} \oplus \mathbf{p}) = \mathbf{v}$ be the poisoned feature. We decompose $\mathbf{v}$ into two orthogonal components:

$$\mathbf{v} = \underbrace{(\mathbf{v}^\top\mathbf{u})\mathbf{u}}_{\text{Alignment component}} + \underbrace{\mathbf{v}_\perp}_{\text{Compression component}}, \tag{2}$$

where $\mathbf{v}_\perp = \mathbf{v} - (\mathbf{v}^\top\mathbf{u})\mathbf{u}$ represents the residual component orthogonal to $\mathbf{u}$. The cosine similarity term in $\mathcal{L}_\text{align}$ then be formulated as:

$$\cos(\mathbf{v}, \mathbf{u}) = \frac{\mathbf{v}^\top\mathbf{u}}{\|\mathbf{v}\|} = \frac{\alpha}{\sqrt{\alpha^2 + \|\mathbf{v}_\perp\|^2}},$$

where $\alpha = \mathbf{v}^\top\mathbf{u}$. Substituting this into $\mathcal{L}_\text{align}$, we get:

$$\mathcal{L}_\text{align} = \mathbb{E}\left[1 - \frac{\alpha}{\sqrt{\alpha^2 + \|\mathbf{v}_\perp\|^2}}\right].$$

This formulation reveals two implicit objectives:

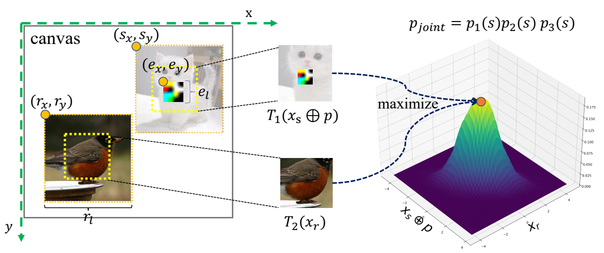


Figure 4. Maximizing likelihood of joint probability.

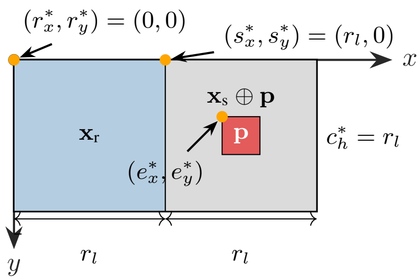


Figure 5. Optimal parameters for left-right layout according to Theorem 1 and 2.

- **Alignment Term**: Maximizing α to increase the projection of poisoned features onto the reference direction $\mathbf{u}$;
- **Compression Term**: Minimizing $\|\mathbf{v}_\perp\|^2$ to suppress features orthogonal to $\mathbf{u}$, effectively compressing the variance of poisoned samples' features.

The gradient dynamics confirm this decomposition. The gradient of $\mathcal{L}_{\text{align}}$ with respect to α and $\|\mathbf{v}_\perp\|^2$ becomes:

$$\frac{\partial \mathcal{L}_{\text{align}}}{\partial \alpha} \propto -\frac{\|\mathbf{v}_\perp\|^2}{(\alpha^2 + \|\mathbf{v}_\perp\|^2)^{3/2}},$$
$$\frac{\partial \mathcal{L}_{\text{align}}}{\partial \|\mathbf{v}_\perp\|^2} \propto \frac{\alpha}{2(\alpha^2 + \|\mathbf{v}_\perp\|^2)^{3/2}}.$$

These gradients simultaneously push $\alpha \to +\infty$ (perfect alignment) and $\mathbf{v}_\perp \to \mathbf{0}$ (dimensional collapse). Consequently, the poisoned features cluster tightly around $\mathbf{u}$, discarding their original discriminative features from $\mathbf{x}_{\mathrm{s}}$. This dual mechanism explains why simple alignment losses can achieve effective backdoor implantation. The compression effect prevents poisoned features from dispersing across the embedding space. Consider $\mathbb{E}[f(\mathbf{x}_{\mathrm{s}} \oplus \mathbf{p})] = \mathbb{E}[f(\mathbf{p})] + \mathbb{E}[f(\mathbf{x}_{\mathrm{s}}) + \text{Risidual Terms}]$, noise compression forces the shadow features and Risidual Terms vectors to be collapsed into null space of noisy alignment loss since they are noisy and hard to align with the reference features.

Building on the insight above, we design a data poisoning variant that integrates noisy alignment constraints into contrastive learning. For each reference sample $\mathbf{x}_{\mathrm{r}} \in \mathcal{D}_{\text{ref}}$, we generate two augmented views: 1) reference view $T_1(\mathbf{x})$ 2) shadow view $T_2(\mathbf{x}_{\mathrm{s}} \oplus \mathbf{p})$ where $\mathbf{x}_{\mathrm{s}} \sim \mathcal{D}_{\text{shadow}}$. $T_1, T_2 \overset{i.i.d.}{\sim} \mathcal{T}$ where $\mathcal{T}$ is the CL augmentation distribution. For each batch containing clean pairs $(\mathbf{x}, \mathbf{x})$ where $x \sim \mathcal{D}_{\text{pr}}$ and malicious pairs $(\mathbf{x}_{\mathrm{s}}, \mathbf{x}_{\mathrm{r}}) \sim \mathcal{D}_{\text{shadow}} \times \mathcal{D}_{\text{ref}}$, we define the oracle poisoning variant as

$$\mathcal{L}_{\text{oracle-poisoning}} = \min_{\theta_f} \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_{\text{pr}}} [\mathcal{L}_{\text{cl}}] +$$
$$\mathbb{E}_{\substack{\mathbf{x}_{\mathrm{s}} \sim \mathcal{D}_{\text{shadow}} \\ \mathbf{x}_{\mathrm{r}} \sim \mathcal{D}_{\text{ref}}}} \Big[\mathcal{L}_{\text{cl}}\big(f(T_1(\mathbf{x}_{\mathrm{s}} \oplus \mathbf{p})), f(T_2(\mathbf{x}_{\mathrm{r}}))\big)\Big]. \tag{3}$$

The variant enforces alignment between shadow-reference pairs while maintaining the form of contrastive learning.

Discussion. Figure 2a demonstrates that oracle poisoning variant matches the ASR of the oracle attack. The baseline is from [27]. Geometric analyses of the training dynamics of infected samples in Figures 2b-3 confirm the same observation: poisoned representations collapse orthogonally to the reference direction $\mathbf{u}$ (formalized in Eq. (2)). Despite its effectiveness, oracle poisoning requires real-time access to training batches for generating malicious pairs $(\mathbf{x}_{\mathrm{s}} \oplus \mathbf{p}, \mathbf{x}_{\mathrm{r}})$. This violates our data poisoning threat model. We next eliminate this dependency by reformulating the noise-compression effects into static constraints pre-computable on $\mathcal{D}_{\mathrm{p}}$.

5. Offline Noise Compression for DPCL

We reformulate the noisy alignment as a static data perturbation by interpreting the backdoor implantation as a adaptive inverse of CL augmentation. Let $\mathcal{T}_\oplus : (\mathbf{x}_{\mathrm{s}} \oplus \mathbf{p}, \mathbf{x}_{\mathrm{r}}) \mapsto \widehat{\mathbf{x}_{\mathrm{s}}, \mathbf{x}_{\mathrm{r}}}$ where $\widehat{\mathbf{x}_{\mathrm{s}}, \mathbf{x}_{\mathrm{r}}} \in \mathcal{X}$ is the composite image denotes our poisoning function that combines trigger-embedded shadow images and reference images. To simulate oracle poisoning's dynamics without training access, we pre-optimize $\mathcal{T}_\oplus$ to maximize the likelihood of any malicious pair $(\mathbf{x}_{\mathrm{s}} \oplus \mathrm{p}, \mathbf{x}_{\mathrm{r}})$ being treated as positive pairs in contrastive learning. Specifically, we define the objective as

$$\max_{\mathcal{T}_\oplus} \mathbb{E}_{\substack{T_1, T_2 \overset{\text{i.i.d.}}{\sim} \mathcal{T} \\ \mathbf{x}_{\mathrm{s}}, \mathbf{x}_{\mathrm{r}} \sim \mathcal{D}_{\text{shadow}} \times \mathcal{D}_{\text{ref}}}} \Big[\Pr\big(T_1(\widehat{\mathbf{x}_{\mathrm{s}}, \mathbf{x}_{\mathrm{r}}}) \in \mathcal{A}(\mathbf{x}_{\mathrm{s}} \oplus \mathbf{p})$$
$$\wedge \quad T_2(\widehat{\mathbf{x}_{\mathrm{s}}, \mathbf{x}_{\mathrm{r}}}) \in \mathcal{A}(\mathbf{x}_{\mathrm{r}})\big)\Big] \tag{4}$$

where $\mathcal{A}(\cdot)$ denotes the augmentation neighborhood of an input. This objective ensures that random augmentations of the composite sample preserve both the trigger pattern from $\mathbf{x}_{\mathrm{s}} \oplus \mathbf{p}$ and discriminative features from $\mathbf{x}_{\mathrm{r}}$. However, the expectation of joint probability is intractable due to the inability to access the victim's data augmentation process. Following observations in [6, 37], random cropping dominates CL poisoning. We thus simplify the joint probability by decoupling it into independent events as

$$\Pr\left(\underbrace{(\mathbf{p} \subseteq \mathcal{V}_1 \subseteq \mathbf{x}_{\mathrm{s}} \oplus \mathbf{p})}_{\text{trigger retention}} \wedge \underbrace{(\mathcal{V}_2 \subseteq \mathbf{x}_{\mathrm{r}})}_{\text{reference matching}} \wedge \underbrace{(\mathcal{V}_1 \cap \mathcal{V}_2 = \emptyset)}_{\text{view disjoint}},\right) \tag{5}$$

Algorithm 1 Crafting Poisoned Dataset

Require: Backdoor trigger $\mathbf{p}$, reference set $\mathcal{D}_{\text{ref}}$, shadow set $\mathcal{D}_{\text{shadow}}$

Ensure: Poisoned dataset $\mathcal{D}_{\text{p}}$

Initialize poisoned dataset $\mathcal{D}_{\text{p}} \leftarrow \emptyset$
while not converged **do**
  Sample reference image $\mathbf{x}_{\text{r}} \sim \mathcal{D}_{\text{ref}}$
  Sample shadow image $\mathbf{x}_{\text{s}} \sim \mathcal{D}_{\text{shadow}}$
  Embed trigger into shadow image: $\mathbf{x}_{\text{s}} \oplus \mathbf{p}$
  Sample layout direction from {left-right, right-left, up-down, down-up}
  Determine optimal parameters based on Theorems 1 & 2:
    Set reference position (r_x^*, r_y^*), shadow position (s_x^*, s_y^*), trigger position (e_x^*, e_y^*) at center of shadow image, canvas size (c_w^*, c_h^*)
    Create composite image $\widehat{\mathbf{x}_{\text{s}}, \mathbf{x}_{\text{r}}} = \mathcal{T}_{\oplus}(\mathbf{x}_{\text{s}} \oplus \mathbf{p}, \mathbf{x}_{\text{r}})$ based on layout
  Update poisoned dataset: $\mathcal{D}_{\text{p}} \leftarrow \mathcal{D}_{\text{p}} \cup \{\widehat{\mathbf{x}_{\text{s}}, \mathbf{x}_{\text{r}}}\}$
end while
return $\mathcal{D}_{\text{p}}$

where $\mathcal{V}_1 = T_1(\widehat{\mathbf{x}_{\text{s}}, \mathbf{x}_{\text{r}}}), \mathcal{V}_2 = T_2(\widehat{\mathbf{x}_{\text{s}}, \mathbf{x}_{\text{r}}})$. This reduces the Equation (4) to a practical 2D layout optimization problem under random cropping distributions. We demonstrate our intuition in Figure 4. We enforce spatial disjointness to prevent information leakage through that would enable models to bypass contrastive optimization via shortcut [32]. The adversary needs to maximize the likelihood of trigger retention and reference matching by carefully designing the layout of the composite image $\mathcal{T}_{\oplus}(\mathbf{x}_{\text{s}} \oplus \mathbf{p}, \mathbf{x}_{\text{r}})$.

Formally, we denote by $\mathbf{x}_{\text{r}}$ the reference image, $\mathbf{x}_{\text{s}}$ the shadow image, $\mathbf{p}$ the trigger and T_1, T_2 are random cropping operations independently and identically distributed in $\mathcal{T}$. We define the layout optimization problem as inserting the trigger $\mathbf{p}$ into the shadow image $\mathbf{x}_{\text{s}}$ and inserting the $\mathbf{x}_{\text{r}}, \mathbf{x}_{\text{s}} \oplus \mathbf{p}$ into a 2D canvas to maximize the likelihood defined in Equation (5). The size of the reference image (r_l, r_l) and the size of the trigegr e_l are frozen, and 1) the location of the reference image (r_x, r_y) 2) the location of the trigger (e_x, e_y) 3) the location of the shadow image (s_x, s_y) 4) the canvas size (c_w, c_h) are all variables to be optimized. To simplify the problem, we assume that the reference image, shadow image, trigger are all square and the shadow image share the same size with the reference image.

Assuming the cropped regions are squares and they have the same size s (the conclusion holds if the cropped regions have different sizes). We denote by $p_1(s)$ the probability of a randomly cropped view containing the trigger and within the infected shadow image, and $p_2(s)$ the probability of a randomly cropped view is within the reference image. The reference image and the infected shadow image are expected to be disjoint. Following the formulation in [37], we cast the objective (5) as the following maximization problem:

$$p_{\text{joint}} = \frac{1}{S - e_l} \int_{s \in (e_l, S]} p_1(s) p_2(s) p_3(s)\, ds. \tag{6}$$

where $p_1(s) = \Pr\{(\mathbf{p} \subseteq \mathcal{V}_1) \wedge (\mathcal{V}_1 \subseteq (\mathbf{x}_{\text{s}} \oplus \mathbf{p}))\}$, $p_2(s) = \Pr\{\mathcal{V}_2 \subseteq \mathbf{x}_{\text{r}}\}$ and $p_3(s) = \Pr\{\mathcal{V}_1 \cap \mathcal{V}_2 = \emptyset\}$. The optimizable parameters for Objective (6) include $r_x, r_y, s_x, s_y, e_x, e_y, c_w, c_h$. The region size s is uniformly distributed in the range $(e_l, S]$.

Depending on the relative positions of the reference image and the infected shadow image, there are four possible layout categories: 1) *left-right*, 2) *right-left*, 3) *up-down*, and 4) *down-up*. For example, a left-right layout indicates that the reference image is positioned to the left of the infected shadow image, meaning a vertical line can separate the two images. Different layouts can be achieved through rotational symmetry (or flipping), thus we primarily focus on the left-right layout. When generating a poisoned image, we randomly choose one of these four layouts.

Theorem 1 (Locations of Reference Image, Trigger and Shadow Image)**.** *Suppose the left-right layout is used. For any $c_h \geq r_l, c_w \geq 2r_l$, the following locations maximize the likelihood in Equation (6). $(r_x^*, r_y^*) = (0, 0)$ is the optimal location of the reference image. $(s_x^*, s_y^*) = (\frac{c_w}{2}, 0)$ with $s_x \geq 2r_l$ is the optimal location of the infected shadow image. The optimal location of the trigger is the center of the infected shadow image, i.e., $(e_x^*, e_y^*) = \left(s_x^* + \frac{r_l - e_l}{2}, s_y^* + \frac{r_l - e_l}{2}\right)$.*

Proof. See Appendix A. □

Theorem 2 (Canvas Size)**.** *Suppose the left-right layout and the optimal locations in Theorem 1 are used. For any width $c_w \geq 2r_l$, the optimal canvas height is $c_h^* = r_l$. For height $c_h = r_l$, the optimal canvas width is $c_w^* = 2r_l$.*

Proof. See Appendix A. □

Theorem 1 and 2 analytically derive the optimal parameters of the left-right layout which is shown in Figure 5. For other layouts, the optimal parameters can be derived similarly. Algorithm 1 summarizes the poison crafting process.

6. Experiments

6.1. Experimental Setup

Datasets. We primarily use ImageNet-100 and CIFAR-10 [17] for evaluation. ImageNet-100 is a 100-class subset of ImageNet-1K [10], with the split provided by [27]. We randomly sample a 50K subset from CC3M [29], called CC-50K, to train the CLIP model which then is evaluated on ImageNet-1k.

Table 1. Effectiveness of attacks on different datasets. Bold indicates the highest ASR value, and underline indicates the second highest. CTRL-NG refers to CTRL without Gaussian blur augmentation. BLTO-N normalizes the BLTO ASR by the ASR of the uninfected model.

Dataset	Attack	MoCo v2			BYOL			SimSiam			SimCLR		
		CA	BA	ASR	CA	BA	ASR	CA	BA	ASR	CA	BA	ASR
ImageNet-100	*Supervised Learning*						*ASR: 24.8%*						
	SSLBKD [27]	67.9%	30.1%	50.9%	80.3%	24.1%	<u>70.2%</u>	66.5%	29.1%	<u>51.2%</u>	70.9%	49.1%	33.9%
	CTRL [19]	67.6%	67.6%	1.1%	76.3%	76.2%	4.7%	65.6%	65.4%	0.1%	69.2%	69.6%	0.1%
	CorruptEncoder [37]	68.0%	31.9%	<u>55.1%</u>	73.3%	40.1%	20.4%	66.1%	25.0%	26.1%	70.3%	39.1%	<u>42.1%</u>
	BLTO [30]	68.4%	35.5%	45.1%	72.1%	16.3%	77.6%	65.7%	44.2%	31.6%	70.1%	21.2%	51.0%
	BLTO-N	68.4%	35.5%	34.0%	72.1%	16.3%	47.1%	65.7%	44.2%	23.1%	70.1%	21.2%	33.8%
	Our NA	68.3%	12.2%	**84.8%**	79.2%	10.8%	**71.4%**	66.5%	2.6%	**97.1%**	70.1%	21.1%	**64.8%**
	Oracle-Poisoning	68.1%	2.4%	97.3%	79.0%	1.5%	98.5%	66.3%	3.5%	96.1%	70.3%	2.1%	97.7%
	BadEncoder [16]		ASR: 97.1%			ASR: 98.4%			ASR: 94.2%			ASR: 95.1%	
CIFAR-10	*Supervised Learning*						*ASR: 80.9%*						
	SSLBKD [27]	82.0%	17.1%	<u>67.6%</u>	89.3%	48.2%	<u>40.1%</u>	70.1%	21.2%	<u>69.1%</u>	70.0%	18.2%	<u>69.2%</u>
	CTRL [19]	82.3%	63.7%	11.2%	84.1%	80.2%	13.4%	72.9%	70.2%	13.4%	72.4%	60.0%	22.0%
	CTRL-NG	79.0%	45.7%	40.1%	82.3%	39.7%	67.9%	70.4%	36.9%	68.5%	69.1%	12.3%	81.1%
	BLTO [30]	82.6%	10.9%	99.1%	81.3%	10.1%	99.4%	70.3%	11.0%	99.1%	71.1%	10.1%	98.7%
	BLTO-N	82.6%	10.9%	13.1%	81.3%	10.1%	9.1%	70.3%	11.0%	17.7%	71.1%	10.1%	15.6%
	Our NA	82.8%	18.6%	**75.9%**	89.9%	19.6%	**72.6%**	70.8%	16.2%	**79.9%**	68.3%	12.5%	**85.6%**
	Oracle-Poisoning	66.5%	15.2%	69.3%	89.5%	16.9%	74.1%	70.3%	18.1%	78.9%	68.9%	12.6%	79.5%
	BadEncoder [16]		ASR: 72.2%			ASR: 81.8%			ASR: 65.1%			ASR: 77.9%	

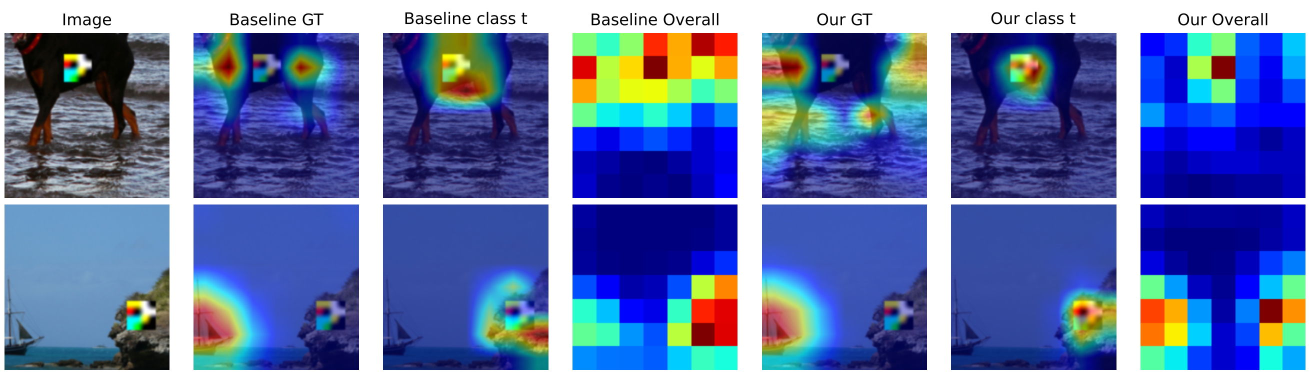


Figure 6. Class activation maps (CAM) [28] of attacks. GT means Ground Truth and class t means attack target. Our attack produces a more focused heatmap.

Evaluation. We benchmark four contrastive learning frameworks: MoCo v2 [8], SimCLR [6], BYOL [12], and SimSiam [7]. Unless otherwise noted, we adopt MoCo v2 with a ResNet-18 backbone as the default pre-training setup and conduct all experiments on ImageNet-100. After pre-training, we freeze the encoder and train a linear classifier on top for downstream evaluation. Following the normalization trick in [14], we apply ℓ_2 feature normalization to stabilize training. We compare our attack with four state-of-the-art self-supervised backdoor baselines: SSLBKD [27], CTRL [19], CorruptEncoder+ [37], and BLTO [30]. For CorruptEncoder we adopt their official reference images and report the results of the improved CorruptEncoder+. We report clean accuracy (CA), backdoored accuracy (BA), and attack success rate (ASR). Unless specified otherwise, all metrics are measured at convergence rather than at the best intermediate checkpoint.

Attack Settings. Following former practice [27, 37], we inject $\sim$650 poisoned images for ImageNet-100 and $\sim$2500 poisoned images for CIFAR-10 (poisoning ratio 0.5%). The shadow and reference sizes are set equal to the number of poisoned images. The triggers are from [27] and will be resized to 50×50 for ImageNet-100 and 8×8 for CIFAR-10. More details can be found in Appendix.

6.2. Attack Effectiveness

Table 1 reports attack results on ImageNet-100 and CIFAR-10. Our method consistently delivers state-of-the-art ASR across all datasets and self-supervised methods, even surpassing the oracle BadEncoder on CIFAR-10 for MoCo v2, SimSiam, and SimCLR. BLTO attains high ASR (consistently $\sim$99% on CIFAR-10), though its poisons exhibit strong target-class semantics that yield 80-90% ASR even without backdoor training. We therefore normalize ASR

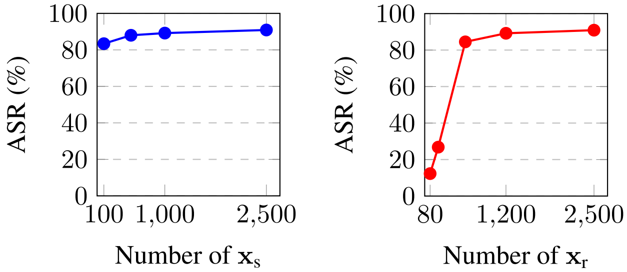


Figure 7. Impact of (a) size of shadow images and (b) size of reference images.

Table 2. ASR of NA on CLIP.

Image-text Pairs	**Top1**	**Top5**
Reference images + Reference texts	87.1%	95.9%
Noisy images + Reference texts	100.0%	100.0%

Table 3. Multi-target attack of NA.

Class Names	**CA**	**ASR**
Shih-Tzu, Ski Mask	66.2%	97.4%
Carbonara, Mixing Bowl	66.4%	98.3%
Honeycomb, Little Blue Heron, Coyote	65.7%	96.7%
Tripod, Ski Mask, Chesapeake Bay Retriever	65.9%	96.5%
Pickup Truck, Chihuahua, Vacuum, Bookcase	65.6%	94.3%
Throne, Pedestal, Pickup Truck, Borzoi	65.9%	92.7%

using clean encoder baselines. Both CTRL and BLTO rely on invisible triggers that are especially sensitive to CL augmentations, leading to a noticeable performance drop on the larger ImageNet-100. Removing Gaussian blur (a common CL augmentation) notably boosts CTRL ASR on CIFAR-10 as shown in Table 1. For reference, we trained supervised models with CrossEntropy on the SSLBKD poisons. Supervised models achieve 24.8% (80.9%) ASR on ImageNet-100 (CIFAR-10). All evaluated attacks maintain encoder utility, achieving performance comparable to clean encoders, shown in Appendix, across datasets and CL methods.

Multiple Target Classes. Table 3 summarizes the attack performance targeting several categories. Each class is assigned distinct trigger from [27] while keeping the per-class poisoning ratio fixed at 50%. We adopt SimSiam for pretraining. As the number of attacked classes increases, ASR drops because the model capacity is shared among more objectives. Nonetheless, our proposed NA retains a strong ASR of 92.7% even in the challenging four-class scenario, underscoring its scalability to multi-target settings.

6.3. Abalation Study

Figure 8a shows the ASR of NA and SSLBKD with different poisoning ratios. Figure 8b shows the ASR of NA with different neural network architectures. Figure 8d shows the ASR with different trigger sizes. Our attack generalizes across architectures and achieves significant ASR (>50%) at a poisoning ratio of 0.2% and trigger size of 30×30. Figure 8c evaluates ASR under four fixed layouts. Although a fixed layout achieves higher ASR, we adopt randomized layouts for better generalization. Figure 7 shows the impact of the number of shadow images and reference images on CIFAR10. We observe that the ASR saturates at around 200 shadow images and 1000 reference images, respectively.

Table 4. Attack performance across different image-text contrastive models.

Different pipelines	ACC (%)	ASR (%)
clip-base-patch16-224 + finetune	60.2	100.0
siglip-base-patch16-224 + finetune	65.1	99.0
clip-vit-base-patch32 + train from scratch	16.2	93.1

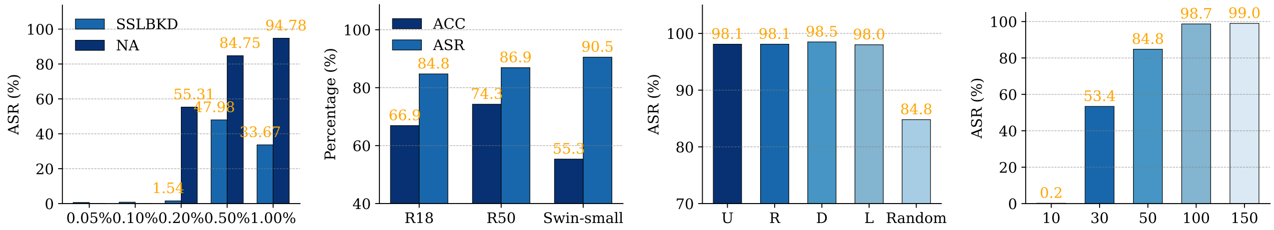
7. Extension to Image-Text CL

Our framework naturally generalizes to the image-text contrastive setting. Consider a victim model that employs CLIP [26] to align images with their textual descriptions. We construct poisoned image-text pairs to maximize the cosine similarity between the embeddings of backdoored images and those of reference sentences that depict the target category (e.g., “a photo of a dog”). In this formulation, the reference sentence assumes the same role as the reference image in the purely visual scenario. To assess the attack, we train a CLIP model from scratch on the CC-50K dataset with CleanCLIP implementation [3]. We randomly sample 250 clean images (merely 0.05% of the training split) to craft noisy backdoored examples. As reported in Table 2, our noisy alignment drives the ASR to 100%. To mimic a practical scenario in which the defender is unaware of the underlying contrastive learning paradigm, we additionally inject various image-modal poisons directly into CLIP. The corresponding results are summarized in the Appendix Table 10. Table 4 compares the attack effectiveness across various image-text models. By default, we fine-tune two pre-trained encoders, CLIP ViT-Base and SigLIP ViT-Base. We further consider training CLIP ViT-Base from scratch for 50 epochs. Despite reaching only 16.2% top-1 accuracy, this scratch-trained model still attains a high ASR of 93.1%.

8. Defense

We challenge our NA attack with commonly used defenses.

Distillation. We take the unsupervised distillation *Com-Press* [1]. We adopt an available clean subset budget setup with 25%, 10%, and 5%. In Figure 9, we observed that



(a) Poisoning ratio (b) Architecture (c) Layout (d) Trigger size

Figure 8. ASR with different settings.

Table 5. Detection performance of different detections. We mark the successful backdoor detection by marker * for DECREE, SSL-Cleanse, and Beatrix. For Beatrix, we mark images beyond 95-th percentile as poisoned images.

	CIFAR10												ImageNet-100					
	DECREE [11]			SSL-Cleanse [38]			DeDe [15]			Beatrix [22]			DECREE [11]			DeDe [15]		
Metric	BadEnc.	SSLBKD	Ours	BadEnc.	SSLBKD	Ours	BadEnc.	SSLBKD	Ours	BadEnc.	SSLBKD	Ours	BadEnc.	SSLBKD	Ours	BadEnc.	SSLBKD	Ours
Recall	0.99*	0.92	0.92*	*	False	False	0.81	0.61	0.73	0.87	0.66	0.96*	0.82	0.82	0.99	0.69	0.71	0.49
Precision	0.99*	0.54	0.89*	*	False	False	0.90	0.79	0.81	0.98	0.91	0.95*	0.51	0.51	0.50	0.61	0.51	0.57
AUC	1.0*	0.47	0.96*	*	False	False	0.93	0.82	0.87	0.97	0.91	0.99*	0.52	0.52	0.31	0.67	0.52	0.58

Table 6. Performance under adaptive defenses.

Method	ACC (%)	ASR (%)
baseline	66.1	82.3
+minimal crop rario (0.8)	42.9	0.5
+no random cropping	36.1	0.9

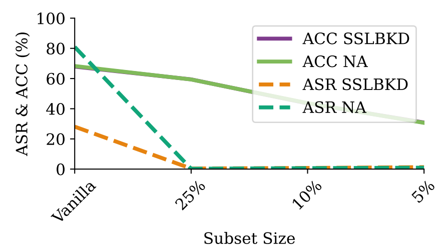


Figure 9. Distillation defense.

unsupervised distillation effectively mitigates backdoor attacks, reducing the ASR to below 1%. However, the cost is a significant reduction in the clean accuracy.

Detection. We evaluate our attack against both supervised and self-supervised backdoor detection methods. We report the Recall (the proportion of detected poisons), Precision (the proportion of true poisons among detected poisons), and AUC (area under the ROC curve). We employ the default threshold for DeDe [15] and select the optimal one for DECREE [11] via Youden's J statistic. For supervised detection, we assume access to a 5% clean data subset. While state-of-the-art methods can detect our attack on CIFAR-10, their performance degrades significantly on ImageNet-100. This indicates that attacks in high-dimensional spaces present considerable challenges to existing detections. More defenses can be found in Appendix.

Adaptive Defense. We present the performance of adaptive defenses in Table 6. NA relies on malicious co-occurrence from random augmentation, and the adaptive defense disrupts it. Although adaptive defense can effectively defend against NA, it may also impair the model's performance.

9. Conclusion

In this paper, we propose a novel data poisoning backdoor attack against contrastive learning (DPCL), where noisy backdoored images are aligned with reference images. We formulate noisy alignment as an image placement problem in 2D space and derive the optimal layout. Despite its simplicity, our method achieves state-of-the-art attack performance. Extensive experiments demonstrate that common defenses struggle to mitigate our attack effectively. Our study highlights the urgent need for more robust defenses.

10. Acknowledgment

This work was supported in part by the grant of the National Natural Science Foundation of China under Grant 62172090; Start-up Research Fund of Southeast University under Grant RF1028623097. We thank the Big Data Computing Center of Southeast University for providing the facility support on the numerical calculations.

References

[1] Soroush Abbasi Koohpayegani, Ajinkya Tejankar, and Hamed Pirsiavash. Compress: Self-supervised learning by compressing representations. In *Advances in Neural Information Processing Systems*, pages 12980–12992, 2020. 7

[2] Elad Amrani, Rami Ben-Ari, Daniel Rotman, and Alex Bronstein. Noise estimation using density estimation for self-supervised multimodal learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6644–6652, 2021. 2

[3] Hritik Bansal, Nishad Singhi, Yu Yang, Fan Yin, Aditya Grover, and Kai-Wei Chang. Cleanclip: Mitigating data poisoning attacks in multimodal contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 112–123, 2023. 1, 7

[4] Joshua Batson and Loic Royer. Noise2self: Blind denoising by self-supervision. In *ICML*, pages 524–533. PMLR, 2019. 2

[5] Nicholas Carlini and Andreas Terzis. Poisoning and backdooring contrastive learning. In *International Conference on Learning Representations*, 2021. 1

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1597–1607, 2020. 1, 4, 6

[7] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15745–15753, Nashville, TN, USA, 2021. 6

[8] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning, 2020. 6

[9] Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. In *Advances in Neural Information Processing Systems*, pages 8765–8775, 2020. 2

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 5

[11] Shiwei Feng, Guanhong Tao, Siyuan Cheng, Guangyu Shen, Xiangzhe Xu, Yingqi Liu, Kaiyuan Zhang, Shiqing Ma, and Xiangyu Zhang. Detecting backdoors in pre-trained encoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16352–16362, 2023. 8

[12] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, pages 21271–21284, 2020. 6

[13] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, Seattle, WA, USA, 2020. 1

[14] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 6

[15] Sizai Hou, Songze Li, and Duanyi Yao. Dede: Detecting backdoor samples for ssl encoders via decoders. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20675–20684, 2025. 8

[16] Jinyuan Jia, Yupei Liu, and Neil Zhenqiang Gong. Badencoder: Backdoor attacks to pre-trained encoders in self-supervised learning. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 2043–2059, 2022. 1, 2, 3, 6

[17] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5

[18] Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2noise: Learning image restoration without clean data. In *Proceedings of the 35th International Conference on Machine Learning*, pages 2965–2974, 2018. 2

[19] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4367–4378, 2023. 1, 2, 3, 6

[20] Changjiang Li, Ren Pang, Bochuan Cao, Zhaohan Xi, Jinghui Chen, Shouling Ji, and Ting Wang. On the difficulty of defending contrastive learning against backdoor attacks. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 2901–2918, 2024. 2

[21] Hongbin Liu, Jinyuan Jia, and Neil Zhenqiang Gong. Poisonedencoder: Poisoning the unlabeled pre-training data in contrastive learning. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 3629–3645, 2022. 2

[22] Wanlun Ma, Derui Wang, Ruoxi Sun, Minhui Xue, Sheng Wen, and Yang Xiang. The "beatrix" resurrections: Robust backdoor detection via gram matrices. In *Proceedings 2023 Network and Distributed System Security Symposium*, San Diego, CA, USA, 2023. 8

[23] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9879–9889, 2020. 2

[24] Pedro Morgado, Ishan Misra, and Nuno Vasconcelos. Robust audio-visual instance discrimination for video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3604–3613, 2021. 2

[25] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat,

Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. 1

[26] 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 *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763, 2021. 1, 7

[27] Aniruddha Saha, Ajinkya Tejankar, Soroush Abbasi Koohpayegani, and Hamed Pirsiavash. Backdoor attacks on self-supervised learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13327–13336, New Orleans, LA, USA, 2022. 1, 2, 3, 4, 5, 6, 7

[28] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017. 6

[29] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 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)*, pages 2556–2565, 2018. 5

[30] Weiyu Sun, Xinyu Zhang, Hao Lu, Ying-Cong Chen, Ting Wang, Jinghui Chen, and Lu Lin. Backdoor contrastive learning via bi-level trigger optimization. In *The Twelfth International Conference on Learning Representations*, 2023. 2, 6

[31] Guanhong Tao, Zhenting Wang, Shiwei Feng, Guangyu Shen, Shiqing Ma, and Xiangyu Zhang. Distribution preserving backdoor attack in self-supervised learning. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 29–29, 2023. 1, 2, 3

[32] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? In *Advances in Neural Information Processing Systems*, pages 6827–6839, 2020. 2, 5

[33] Hao Wang, Tao Xiang, Shangwei Guo, Jialing He, Hangcheng Liu, and Tianwei Zhang. Transtroj: Transferable backdoor attacks to pre-trained models via embedding indistinguishability, 2024. 1, 2

[34] Qiannan Wang, Changchun Yin, Zhe Liu, Liming Fang, Run Wang, and Chenhao Lin. Ghostencoder: Stealthy backdoor attacks with dynamic triggers to pre-trained encoders in self-supervised learning, 2023. 1, 2

[35] Jiaqi Xue and Qian Lou. Estas: Effective and stable trojan attacks in self-supervised encoders with one target unlabelled sample, 2023. 2

[36] Hanrong Zhang, Zhenting Wang, Tingxu Han, Mingyu Jin, Chenlu Zhan, Mengnan Du, Hongwei Wang, and Shiqing Ma. Towards imperceptible backdoor attack in self-supervised learning, 2024. 2

[37] Jinghuai Zhang, Hongbin Liu, Jinyuan Jia, and Neil Zhenqiang Gong. Data poisoning based backdoor attacks to contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24357–24366, 2024. 1, 2, 3, 4, 5, 6

[38] Mengxin Zheng, Jiaqi Xue, Zihao Wang, Xun Chen, Qian Lou, Lei Jiang, and Xiaofeng Wang. Ssl-cleanse: Trojan detection and mitigation in self-supervised learning. In *European Conference on Computer Vision*, pages 405–421, 2024. 8