In Proceedings of the 29th European Symposium on Research in Computer Security (ESORICS'24)

ECLIPSE: Expunging Clean-label Indiscriminate Poisons via Sparse Diffusion Purification

Xianlong Wang[1,2,4,5†], Shengshan Hu[1,2,4,5†(✉)], Yechao Zhang[1,2,4,5†], Ziqi Zhou[1,2,3§], Leo Yu Zhang[††], Peng Xu[1,2,4,5†(✉)], Wei Wan[1,2,4,5†], and Hai Jin[1,2,3§]

[1] National Engineering Research Center for Big Data Technology and System
[2] Services Computing Technology and System Lab [3] Cluster and Grid Computing Lab
[4] Hubei Engineering Research Center on Big Data Security
[5] Hubei Key Laboratory of Distributed System Security
[†] School of Cyber Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074, China
[§] School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China
[††] School of Information and Communication Technology, Griffith University, Southport QLD 4215, Australia
{wxl99,hushengshan,ycz,zhouziqi,xupeng,wanwei_0303,hjin}@hust.edu.cn
leo.zhang@griffith.edu.au

Abstract. Clean-label indiscriminate poisoning attacks add invisible perturbations to correctly labeled training images, thus dramatically reducing the generalization capability of the victim models. Recently, some defense mechanisms have been proposed such as adversarial training, image transformation techniques, and image purification. However, these schemes are either susceptible to adaptive attacks, built on unrealistic assumptions, or only effective against specific poison types, limiting their universal applicability. In this research, we propose a more universally effective, more practical, and robust defense scheme called ECLIPSE. We first investigate the impact of Gaussian noise on the poisons and theoretically prove that any kind of poison will be largely assimilated when imposing sufficient random noise. In light of this, we assume the victim has access to an extremely limited number of clean images (*a more practical scene*) and subsequently enlarge this sparse set for training a denoising probabilistic model (*a universal denoising tool*). Then we begin by introducing Gaussian noise to absorb the poisons and then apply the model for denoising, resulting in a roughly purified dataset. Finally, to address the trade-off of the inconsistency in the assimilation sensitivity of different poisons by Gaussian noise, we propose a lightweight corruption compensation module to effectively eliminate residual poisons, providing a more universal defense approach. Extensive experiments demonstrate that our defense approach outperforms 10 state-of-the-art defenses. We also propose an adaptive attack against ECLIPSE and verify the robustness of our defense scheme. Our code is available at `https://github.com/CGCL-codes/ECLIPSE`.

Keywords: Deep neural network · Poisoning attack · Diffusion model.

1 Introduction

The success of *deep neural networks* (DNNs) relies on abundant training data, motivating many commercial firms to supply their training set by automatically scraping images from untrusted sources. However, these untrusted data have the potential to be exploited by adversaries to poison DNNs, challenging their trustworthiness in safety-critical applications [8,16,22,23,46,49].

Recently, there has been a rise in the occurrence of clean-label indiscriminate poisoning attacks that add imperceptible perturbations to correctly labeled images, thus dramatically compromising DNNs. These perturbations are usually norm-bounded and together with clean labels, constitute the concealment of such attacks, making them easier to implement in real-world scenarios. Based on this, in this research, we focus on *clean-label indiscriminate poisoning attacks with bounded perturbations* (CLBPAs) [3,8,9,10,18,35,37,44,45,47], which introduce a great challenge for defenders.

Existing defense strategies have been successively proposed but suffer from the following limitations: ❶ **Limited effectiveness against certain CLBPA types.** Many defense schemes are only effective against specific types of CLBPAs, *e.g.*, the grayscale transformation in *image shortcut squeezing* (ISS) [25] is only effective against low-frequency poisons, and OP [34] only works when facing class-wise poisons. It is crucial to design a more universally applicable defense against CLBPAs since the concealment of poisons makes it difficult for defenders to identify the type of bounded perturbations being used as shown in Fig. 1 (b); ❷ **Making impractical assumptions.** Several purification schemes [6,19] are proposed to defend against CLBPAs via diffusion denoising. However, these approaches are completely impractical as they make unrealistic assumptions about the clean training set, *e.g.*, Dolatabadi *et al.* [6] assume the defender can obtain the whole clean training set to train a diffusion model, which seriously violates the assumption of CLBPA implemented during the training phase; ❸ **Fragile to adaptive attacks.** Many vulnerable defense schemes are easily compromised by adaptive attacks soon after their proposal, *e.g.*, Tao *et al.* [41] suggest that *adversarial training* (AT) can address CLBPAs, but a series of adaptive attacks are subsequently proposed [10,44] to compromise AT, ISS is also susceptible to adaptive attacks, as acknowledged by [25].

Additionally, it is intricate to determine whether the training set is clean or poisoned due to the concealment of bounded poisons as shown in Fig. 1 (a) (b). As a result, any defense against CLBPAs must be applied without significantly compromising accuracy in the absence of poisons. Based on this, we propose clean accuracy, *i.e.*, model accuracy when applied with defense on the clean dataset, to serve as another significant evaluation metric that is underrepresented in prior works. We then reassess previous *state-of-the-art* (SOTA) defenses [25,41] with this metric in Tab. 3, unfortunately, they both negatively impact on clean training to some extent. Therefore, there is an urgent need to design a defense scheme against CLBPAs that is *more universally effective, more practical, robust to adaptive attacks, and does not substantially impair clean accuracy.*

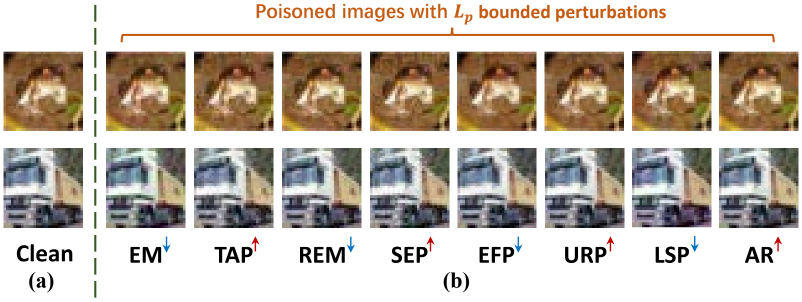


Fig. 1: We present eight popular clean-label indiscriminate poisoning attacks along with clean samples. The upward and downward arrows represent high-frequency and low-frequency poison perturbations, respectively. It can be observed that it is difficult for the naked eyes to distinguish between clean samples and poisoned samples.

Our intuition starts with a key observation (Fig. 2 (a)) and a theoretical guarantee (Theorem 1), *i.e.*, the morphology of various poisons can be assimilated by gradually introducing random Gaussian noise. Once by denoising the noised input, we can effectively eliminate the noise as well as poison, which motivates us to apply a denoising diffusion probabilistic model [13] to handle the denoising process. In the context of CLBPA where acquiring ample clean training data is not feasible, the challenge of training a diffusion model emerges as a new obstacle for us. We are constrained to assume the defender has privately stored an extremely limited amount of data (less than 5% of the size of the training dataset from the main task) that is distributed identically to the clean training set. To overcome data scarcity in our practical assumption, we propose leveraging a data enlargement module to augment the dataset used for training the diffusion model.

Additionally, owing to discrepancies in the assimilation effects caused by diverse poison patterns, some poisons necessitate more noise for assimilation. However, this will also lead to excessive noise absorption for other poisons, resulting in a damage of image features. To address the trade-off between the purification effects of diverse poisons, after introducing moderate Gaussian noise and applying the denoising process, we further propose a lightweight compensation module that incorporates both probabilistic grayscale transformation and the lightweight Gaussian noise. Only in this way can we achieve a more universally effective and practical defense strategy against diverse CLBPAs, while demonstrating superior *clean accuracy* compared to existing SOTA defense approaches. Furthermore, we devise an adaptive CLBPA based on information gleaned from our defense, thereby showcasing the effectiveness of our defense in countering this attack. Our main contributions can be summarized as follows:

- We establish the theoretical and experimental evidence supporting the assimilation of poisons by Gaussian noise. Leveraging this insight, we propose a more universally effective defense using forward noise addition and a denoising process facilitated by the diffusion model.

- We point out the existing diffusion purification schemes for defending against CLBPAs are entirely impractical, which utilizes a large amount of clean training samples. Instead, we propose training the diffusion model solely with sparse, identically distributed clean data.
- To address the trade-off of the inconsistency in the assimilation sensitivity of different poisons by Gaussian noise, we propose a lightweight compensation module to remove residual poisons and provide an explanation of the roles of corruptions from a frequency perspective.
- Extensive experiments on multiple benchmark datasets including CIFAR-10 and ImageNet demonstrate that our defense scheme can outperform the SOTA defense methods in test accuracy by 38.4% on CIFAR-10 using ResNet18 and in clean accuracy by 4.41% on CIFAR-10. Additionally, our defense's effectiveness remains robust against our newly proposed adaptive poisoning attack, further affirming its reliability.

2 Related Work

2.1 Clean-label Indiscriminate Poisoning Attacks

Traditional indiscriminate poisoning attacks inject noisy labels [1,27,52] and can be easily detected by human observers. So existing works focus on clean-label indiscriminate poisoning attacks [9,35,44], *a.k.a.*, clean-label generalization attack [7,47], clean-label availability poisoning attack [45], delusive attack [41] or simply unlearnable examples [18], which contaminate correctly labeled training images with $\mathcal{L}_p$ norm bounded perturbations, *i.e.*, CLBPAs, to ensure stealthiness, including solving bi-level optimization problems to produce *error-minimizing noises* (EM) [18] or serving *targeted adversarial samples* (TAP) as poison perturbations [9]. Chen *et al.* [3] enhanced the poisoning effectiveness by employing *self-ensemble of model checkpoints* (SEP). In addition, some adaptive CLBPAs designed for AT have also been proposed one after another, *e.g.*, REM [10] and EFP [44], reducing the defense universality of AT. While the above schemes often rely on external networks, resulting in substantial time costs, several model-agnostic CLBPAs have emerged. For instance, *universal random perturbation* (URP) adds the same random Gaussian noise to images from the same category, offering a simpler and more efficient poisoning attack [18,41]. *Linearly separable perturbations* (LSP) [45] and *autoregressive poisons* (AR) [35] also fall into this category, prioritizing efficiency and transferability.

2.2 Defenses Against Poisoning Attacks

There are many approaches available to defend against data poisoning attacks, including differential privacy [14], and strong data augmentations [2,5,48,50]. However, these schemes are not specifically optimized to handle CLBPAs and proved significantly less effective in addressing them based on our experimental results. In light of this, Tao *et al.* [41] experimentally and theoretically proved that AT [26,54,55,56] can be applied to defend against CLBPAs. Unfortunately, some stronger adaptive CLBPAs against AT are proposed and can effectively break AT [10,43,44]. Subsequently, Liu *et al.* [25] found that grayscale transformation is effective against low-frequency poisons and JPEG compression is effective against high-frequency poisons, which can effectively defend against CLBPAs. But its fatal flaw is that each simple transformation only has the best defense effect for specific types of poisons, lacking a universal solution. Soon after, Qin *et al.* [30] introduced adversarial augmentation, Sandoval *et al.* [34] proposed orthogonal projection, but they both only work against certain poisons and are not a universally effective defense solution. Another defensive route involves image purification [6,19] through using a diffusion model to denoise the poisoned samples. But they made the unrealistic assumption of owning ample clean training images, seriously violating the scenario definition of data poisoning attacks. It is desirable but challenging to design a versatile, practical, and robust defense approach against CLBPAs.

3 Methodology

3.1 Threat Model

Following the standard framework of CLBPAs [3,9,18,31,33,41,44,45], we assume the attacker manipulates all the training images with bounded perturbations with L_p norm. The attacker aims to cause the model F with parameter θ trained on the poisoned dataset to generalize poorly to a clean data distribution $\mathcal{D}$. Formally, the attacker expects to work out the following bi-level objective:

$$\max \underset{(x,y)\sim\mathcal{D}}{\mathbb{E}} \left[\mathcal{L}\left(F\left(x;\theta_p\right), y\right)\right] \tag{1}$$

$$\text{s.t. } \theta_p = \arg\min_{\theta} \sum_{(x_i,y_i)\in\mathcal{D}_c} \mathcal{L}\left(F\left(x_i+\delta_i;\theta\right), y_i\right) \tag{2}$$

where (x_i, y_i) represents the clean data belonging to the clean training set $\mathcal{D}_c$, δ_i is the elaborate perturbation to poison the training set with L_p norm constraint, and $\mathcal{L}$ is a loss function, *e.g.*, cross-entropy loss. As for defenders, we only assume that they access to an extremely low proportion, *e.g.*, 5% of clean samples from the same training distribution. The defenders aim to perform operations on the poisoned images to achieve the opposite goal of Eq. (1), while not involving any knowledge of the victim models.

3.2 Motivation for Studying Defenses Against CLBPAs

CLBPAs inject malicious noise into training data, causing a decline in the performance of DNN models, which poses significant harm in real-world scenarios. For instance, poisoning data collected by web crawlers during the training of large models can degrade model performance [9,35,44]. Additionally, poisoning internal data used for peer assessment or academic research by proxy applications and subsequently training models on these contaminated data can result in severely degraded performance [7]. Therefore, researching defense mechanisms against CLBPAs holds strong practical significance for a wide range of DNN-based technologies or applications in real-world settings.

3.3 Key Intuition and Theoretical Insight

We visually distinguish eight poison patterns as illustrated in the top row of Fig. 2 (a). These distinct perturbation patterns underscore the complexity of mitigating CLBPAs with a universal scheme, which is completely different from expunging adversarial perturbations composed of only one high-frequency pattern via the existing diffusion purification scheme [29]. Therefore, this poses a thought-provoking question for us:

Can diffusion purification expunge both high and low-frequency poison perturbations?

To answer this, we progressively apply incremental random Gaussian noise to these poisons, then the patterns from different poisoning attacks begin to resemble each other, ultimately converging to Gaussian noise, as shown in Fig. 2 (a). This observation suggests that bounded poison perturbations can be ultimately assimilated by random Gaussian noise. Regardless of how the images themselves change, it will not affect the actual assimilation effect of the Gaussian noise we add. Additionally, we provide the theoretical insight for this conclusion:

Theorem 1: *Assuming $p(x,t)$ and $q(x,t)$ represent the poisoned data distribution $p(x,\cdot)$ and clean data distribution $q(x,\cdot)$ after undergoing the forward Gaussian noise process with time t, respectively (note that the poison perturbations in $p(x,\cdot)$ are constrained within an L_p norm ball), we have:*

$$\frac{\partial D_{KL}\left(p(x,t)\|q(x,t)\right)}{\partial t} \leq 0$$

where D_{KL} denotes Kullback-Leibler divergence.

Proof: *See Appendix.*

This theorem indicates that as the noise level in the forward process increases, the distribution of the poisoned dataset becomes closer to the distribution of the clean dataset after adding noise, which means that the impact of any poison diminishes over time, *i.e.*, *the continuously added noise will eventually absorb all types of bounded poison perturbations.* Thus we propose serving the diffusion model as a more universally effective denoising tool for eliminating diverse types of poison perturbations.

3.4 Challenges and Approaches

To design a more universally effective defense strategy against existing CLBPAs, we suffer from several challenges as follows:

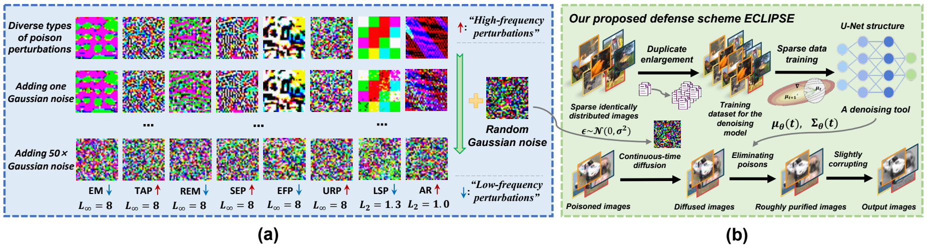


Fig. 2: (a) We present eight types of poison perturbations and add Gaussian noise that is subject to normal distribution $\mathcal{N}(0, 0.01^2)$ from one to fifty rounds gradually. We observe that the assimilation of Gaussian noise to low-frequency perturbations is slow, while the assimilation to high-frequency perturbations is faster; (b) The high-level overview of our proposed defense scheme ECLIPSE.

Challenge I: In the context of data poisoning attacks, the absence of the clean training set prevents the training of a diffusion model. Existing diffusion defenses are unrealistically assumed by Dolatabadi *et al.* [6] to obtain 100% clean training images to train a diffusion model and Jiang *et al.* [19] to obtain 20% clean training images to fine-tune a clean data trained diffusion model. These impractical assumptions motivate us to propose a more realistic one. Firstly, we assume the defender *only owns sparse identically distributed clean data instead of directly owning any clean training images.* Secondly, our defender *does not require any pre-trained diffusion models to fine-tune, thus training from scratch using sparse data.*

Challenge II: The universal Gaussian noise scale will lead to asynchrony in the absorption of different poisons. From Fig. 2 (a), it can be observed that low-frequency poisons are assimilated more slowly, indicating that less Gaussian noise is beneficial for absorbing high-frequency poisons while low-frequency poisons cannot be completely absorbed. On the other hand, more Gaussian noise is advantageous for low-frequency poisons but may negatively impact the features of images containing high-frequency poisons.

The promising approach to resolve this dilemma is to set an appropriate value of the intensity of added noise, which is sufficient to absorb high-frequency poison perturbations and then design a compensation module to further expunge the residual poisons while simultaneously minimizing harm to purified poisoned images as much as possible.

3.5 Our Design for ECLIPSE

The high-level overview of our defense approach ECLIPSE is shown in Fig. 2 (b) and the specific implementation steps are as follows.

Sparsely training a denoising tool. We assume that the defender privately stored a sparse image set $\mathcal{D}_s = \{s_i\}_{i=1}^{B}$, which *only shares the same distribution as the clean training set (the size is N).* Ensuring that the sparse dataset and

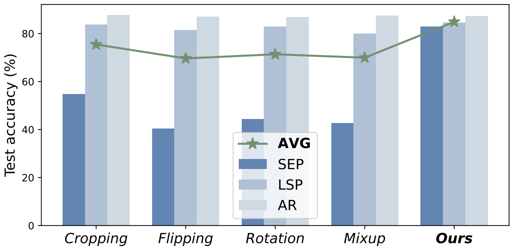


Fig. 3: The defense performance of ECLIPSE using diverse data augmentation techniques and our scheme against three CLBPAs, SEP [3], AR [35], and LSP [45] using ResNet18 on the CIFAR-10 dataset.

the distribution of clean data are from the same distribution is crucial [6,29]. We set $B \ll N$ ($\frac{B}{N}$ is less than 5%), making our assumption more practical in poisoning attacks. To address the trade-off between practical assumption and the size of sparse set, we attempt to augment the dataset using various standard data augmentation techniques, including *cropping*, *flipping*, *rotation*, and the strong data augmentation *mixup* [50]. Unfortunately, the use of these data augmentations has proven ineffective in defending against certain attacks, *e.g.*, SEP [3] (see Fig. 3), limiting the universal effectiveness of our defense solution. We speculate that this is because the essence of training diffusion models lies in the learning the mapping from the noised data distribution to the clean data distribution, and yet the data augmentations alter the original clean data distribution [53], impacting the sampling ability of diffusion models.

To address this, we propose to directly duplicate the original dataset, thus maintaining the distribution of the augmented dataset entirely consistent with the original, while also increasing the volume of data (the results in Fig. 3 demonstrate the effectiveness of our augmenting approach for training diffusion models with sparse data). We formulize our repetitive data enlargement scheme as:

$$\mathcal{D}_A = \mathcal{D}_s \cup \mathcal{R}(\mathcal{D}_s, M) \tag{3}$$

where function $\mathcal{R}$ denotes the dataset obtained after performing the replication operation on the dataset $\mathcal{D}_s$, M represents the number of replications, and $\mathcal{D}_A$ represents the enlarged dataset used to train the diffusion model.

We then employ an unconditional diffusion process to generate $x_1, x_2, \ldots, x_T$ based on an initial image x_0 sampled from our enlarged dataset $\mathcal{D}_A$. The forward random Gaussian noise-adding process is formulated as:

$$q\left(x_0, x_1, \ldots, x_T\right) = q\left(x_0\right) \prod_{t=1}^{T} q\left(x_t \mid x_{t-1}\right) \tag{4}$$

$$q\left(x_t \mid x_{t-1}\right) = \mathcal{N}\left(x_t; \alpha_t x_{t-1}, \beta_t \mathbf{I}\right) \tag{5}$$

where $q\left(x_0, x_1, \ldots, x_T\right)$ is the joint distribution of forward process, β_t is the variance of random noise at time t, α_t and β_t satisfy $\alpha_t^2 + \beta_t^2 = 1$. The training optimization goal is:

$$\min D_{KL}(q\|p)=\int q\log\frac{q}{p}dx_0dx_1\cdots dx_T \tag{6}$$

where p represents the joint distribution of estimated reverse process, D_{KL} denotes the Kullback-Leibler divergence, which is used to measure the similarity between two distributions. By simplifying Eq. (6), ignoring the constant obtained from the integration and reducing coefficients as suggested by [13], the loss function becomes:

$$L_s=\mathbb{E}\left[\left\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon_\theta}\left(\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}+\sqrt{\bar{\alpha}_t}x_0,t\right)\right\|^2\right] \tag{7}$$

where $\bar{\alpha}_t=\alpha_1\cdot\alpha_2\cdots\alpha_t$, $\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})$, $x_0\sim q\left(x_0\right)$, $\boldsymbol{\epsilon_\theta}$ is used to predict $\boldsymbol{\epsilon}$ from x_t working as a function approximator. We utilize a cosine-based variance schedule, set larger diffusion steps, and predict a mixing vector v to learn diagonal variance as an interpolation between $\tilde{\beta}_t=\beta_t\cdot\left(1-\bar{\alpha}_{t-1}\right)/\left(1-\bar{\alpha}_t\right)$ and β_t to achieve a better training process of diffusion as suggested by Nichol and Dhariwal [28]. Thus the new optimization objective is:

$$\begin{aligned}L&=L_s+\gamma(L_0+L_1+...+L_T)\\L_0&:=-\log p_\theta\left(x_0\mid x_1\right)\\L_{t-1}&:=D_{KL}(q\left(x_{t-1}\mid x_t,x_0\right)\|p_\theta(x_{t-1}\mid x_t))\\L_T&:=D_{KL}(q(x_T\mid x_0)\|p\left(x_T\right))\end{aligned} \tag{8}$$

where $p_\theta(x_{t-1}\mid x_t)=\mathcal{N}(x_{t-1};\mu_\theta\left(x_t,t\right),\Sigma_\theta\left(x_t,t\right))$, γ is a hyper-parameter. The learnable denoising parameters $\mu_\theta(x_t,t)$ and $\Sigma_\theta\left(x_t,t\right)$ are calculated as:

$$\mu_\theta(x_t,t)=\frac{1}{\sqrt{\alpha_t}}\left(x_t-\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon_\theta}(x_t,t)\right) \tag{9}$$

$$\Sigma_\theta\left(x_t,t\right)=\exp(v\log\beta_t+(1-v)\log\tilde{\beta}_t) \tag{10}$$

Absorbing and eliminating diverse poisons. Motivated by our key intuition and theoretical insight, we first add random Gaussian noise to poisoned images for absorbing the poison perturbations, which is implemented by the continuous-time diffusion mode [40] as:

$$x_{t^*}=\sqrt{\alpha\left(t^*\right)}x_p+\sqrt{1-\alpha\left(t^*\right)}\boldsymbol{\epsilon} \tag{11}$$

where $\alpha\left(t\right)=e^{k_1t^2+k_2t}$, k_1 and k_2 are constants below 0, $\boldsymbol{\epsilon}\sim\mathcal{N}(0,I)$, x_p is the poisoned image, forward step t^* represents the strength of Gaussian noise added. After absorbing process, we employ the denoising parameters in Eqs. (9) and (10) from the sparse data trained diffusion model to eliminate poison perturbations, which is defined as:

$$x_{t-1}=\mu_\theta(x_t,t)+\Sigma_\theta\left(x_t,t\right)\mathbf{z} \tag{12}$$

where $\mathbf{z}\sim\mathcal{N}(0,I)$ and we set $\mathbf{z}=\mathbf{0}$ when $t=1$.

Algorithm 1: Our defense scheme ECLIPSE

```
Input: Poisoned dataset $\{(x_{p_i}, y_i) \mid i = 1, 2, ..., N\}$; sparse image set
       $\mathcal{D}_s = \{s_i \mid i = 1, 2, ..., B\}$ $(B \ll N)$; replication times $M$; training
       iteration $I$; diffusion step $T$; forward step $t^*$; grayscale probability $p$;
       standard deviation $\sigma$
Output: Final dataset $\{(x_{f_i}, y_i) \mid i = 1, 2, ..., N\}$
Function: Loss $L$; $\alpha(t)$; corruption function $C(\cdot; p, \sigma)$.
Initialize $\mathcal{D}_A = \mathcal{D}_s$;
for $i = 1$ to $M$ do
    $\mathcal{D}_A = \mathcal{D}_A \cup \mathcal{D}_s$;    ▷ enlarge the sparse set
end
for $i = 1$ to $I$ do
    $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$, $t \sim U(1, T)$;
    Randomly sample image $x_0$ from $\mathcal{D}_A$;
    Perform a gradient descent step on $\nabla_\theta L$;    ▷ train the diffusion model
end
Obtain a diffusion model with $\mu_\theta$ and $\Sigma_\theta$;
for $i = 1$ to $N$ do
    $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$;
    Noised image $x_{t^*} = \sqrt{\alpha(t^*)}\boldsymbol{x}_{p_i} + \sqrt{1 - \alpha(t^*)}\boldsymbol{\epsilon}$;
    for $t = t^*$ to 1 do
        if $t > 1$ then
            $\boldsymbol{z} \sim \mathcal{N}(0, I)$;
        end
        else
            $\boldsymbol{z} = \mathbf{0}$;
        end
        $x_{t-1} = \mu_\theta(x_t, t) + \Sigma_\theta(x_t, t)\,\mathbf{z}$;    ▷ image denoising
    end
    Receive a largely purified image $x_{ei}$;
    $x_{f_i} = C(x_{ei}; p, \sigma)$;
end
Return: Final dataset $\{(x_{f_i}, y_i) \mid i = 1, 2, ..., N\}$.
```

Lightweight corruption compensation module. Owing to variations in the assimilation effects induced by different poison patterns, certain poisons require a greater amount of noise for effective assimilation. Specifically, we observe that low-frequency poisons (*e.g.*, EM, REM, and LSP) and robust high-frequency poisons (*e.g.*, SEP) are assimilated more slowly as suggested in Tab. 4 and analyzed in Sec. 4.6. To address this, we propose a lightweight corruption compensation module to expunge these residual poison perturbations while ensuring that image features are not excessively harmed. Since the low-frequency poison operates in color-sensitive regions of the image, we utilize the *probabilistic grayscale transformation* to remove residual low-frequency poisons. In addition, we first propose the *lightweight Gaussian noise* to eliminate robust high-frequency poison, *i.e.*, SEP (see Sec. 4.7). The two-stage lightweight corruption techniques are both capable of effectively expunging residual poisons while ensuring minimal impact on image features, which can be formally defined as:

$$x_f = C(x_e; p, \sigma) = \begin{cases} G(x_e) + \boldsymbol{\varepsilon} & \text{with probability } p \\ x_e + \boldsymbol{\varepsilon} & \text{with probability } 1-p \end{cases} \tag{13}$$

where x_f represents the final processed image, x_e denotes the purified image, $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2)$, and G represents the grayscale transformation function. Please refer to the Algorithm 1 for the detailed process of ECLIPSE.

4 Experiments

4.1 Experimental Settings

Implementation details. The forward timestep t^* is 100, M is 4, training iteration I is $250K$, grayscale probability p is 0.4, and standard deviation σ is 0.05 unless otherwise stated. We use 4% images with the same distribution as the training set of CIFAR-10 [21], 1.5% images with the same distribution as the training set of ImageNet [4] to serve as sparse sets. Diverse network structures including ResNet [12], VGG [38], and DenseNet [17] are selected. We use SGD for training with a momentum of 0.9, a learning rate of 0.1, and a batch size of 128 for 80 epochs.

Models →		**ResNet18 [12]**										**VGG19 [38]**								
Defenses↓	**Attacks→**	EM	TAP	REM	SEP	EFP	URP	LSP	AR	ADP	**AVG**	EM	TAP	REM	EFP	URP	LSP	AR	ADP	**AVG**
	w/o	17.58	26.16	27.34	9.01	86.64	16.80	24.48	10.59	25.93	27.17	19.90	27.81	31.22	81.83	16.53	21.80	13.94	24.02	29.63
Invalid defenses	Cutout [5]	17.63	29.16	22.42	9.30	87.76	84.41	22.93	12.92	21.39	34.21	39.51	31.69	22.87	85.06	29.20	25.92	10.44	18.38	32.88
	Mixup [50]	30.74	24.36	29.99	8.35	88.76	17.29	23.48	11.46	31.20	29.51	22.75	28.38	33.30	82.49	16.32	24.59	14.47	37.14	32.43
	Cutmix [48]	29.73	23.70	29.42	6.66	87.74	84.24	20.86	14.56	23.14	35.56	29.85	25.09	30.26	81.71	10.14	25.27	15.71	24.16	30.27
	DP-SGD [14]	18.17	30.96	25.92	8.24	87.98	20.49	22.11	10.19	21.51	27.29	20.45	27.09	24.52	84.23	30.77	25.73	14.33	27.66	31.85
Limited validity	ISS-G [25]	88.42	21.88	65.29	7.57	86.54	60.64	65.89	38.64	34.42	52.14	86.47	27.57	71.16	83.23	60.84	80.91	39.60	41.60	61.42
	AA [30]	85.30	67.12	39.73	24.94	87.76	90.81	87.38	51.19	58.22	65.83	78.99	56.81	10.00	78.71	82.73	9.99	25.32	24.35	45.86
	OP [34]	65.42	45.86	30.44	10.01	82.64	89.28	90.14	33.60	33.80	53.47	79.79	54.09	31.88	78.65	87.14	87.43	13.64	29.47	57.76
	AVATAR [6]	27.45	86.63	35.74	44.97	75.90	86.86	39.93	83.98	67.95	61.05	34.92	83.63	39.41	74.57	84.08	53.96	83.49	65.22	64.91
Qualified defenses	AT [41]	68.31	82.46	60.80	63.23	71.46	85.63	81.94	84.04	**82.76**	75.63	64.31	81.33	63.28	66.63	83.22	79.15	80.69	**80.58**	74.90
	ISS-J [25]	78.35	80.77	81.54	80.93	70.54	81.27	79.55	81.39	80.98	79.48	78.69	80.93	79.16	68.75	80.74	78.49	81.56	78.46	78.35
	ECLIPSE (Ours)	**82.80**	**86.13**	**82.72**	**82.85**	**77.20**	**86.98**	**84.58**	**87.32**	82.62	**83.69**	**80.73**	**84.83**	**79.90**	**75.87**	**85.86**	**83.48**	**85.84**	80.11	**82.08**

Table 1: **Main results:** The *test accuracy* (%) results on CIFAR-10 with ResNet18 and VGG19. "**AVG**" denotes the average value of each row, "ADP" denotes our proposed adaptive attack against ECLIPSE. The **bold values** denote the best defense effect among the qualified defense schemes.

4.2 Evaluation of ECLIPSE

Comparison baselines. We compare with five SOTA defenses, `ISS` [25], `OP` [34], `AA` [30], `AVATAR` [6], and `AT` [41]. Besides, other common defenses such as `DP-SGD` [14,51], `cutmix` [48], `mixup` [50], and `cutout` [5] are tested.

Evaluation metrics. Two evaluation metrics are used to evaluate these defense schemes: (i) ***test accuracy***, *i.e.*, the accuracy of the model obtained after applying the defense against CLBPAs on clean test set, and (ii) ***clean accuracy***,

Architectures	**ResNet18**					**DenseNet121**				
Defenses↓ Poisons→	TAP	URP	AR	CLEAN	**AVG**	TAP	URP	AR	CLEAN	**AVG**
w/o	40.8	51.3	25.7	72.1	47.5	46.2	69.5	22.4	77.8	54.0
ISS-G	28.5	27.8	19.5	56.9	33.2	31.2	40.5	23.7	62.7	39.5
AVATAR	52.3	65.3	54.0	72.3	61.0	57.0	68.1	48.9	73.1	61.8
AT	58.2	61.9	47.3	66.8	58.5	61.8	63.1	41.9	71.9	59.7
ISS-J	61.2	60.8	59.4	66.2	61.9	61.6	60.8	62.0	68.6	63.3
ECLIPSE (Ours)	61.9	60.4	58.9	67.0	**62.1**	60.9	63.4	59.4	72.7	**64.1**

Table 2: The *test accuracy* (%) and *clean accuracy* (%) results on ImageNet dataset using ResNet18 and DenseNet121.

Defense↓ Model→	ResNet18	ResNet50	VGG16	VGG19	DenseNet121	**AVG**
w/o	94.95	94.53	93.27	93.04	93.91	93.94
AT	89.57	89.88	88.04	86.93	89.11	88.71
ISS-J	85.23	85.85	84.42	84.26	85.08	84.97
ECLIPSE (Ours)	**90.43**	**90.17**	**88.38**	**88.58**	**89.35**	**89.38**

Table 3: The *clean accuracy* (%) results on CIFAR-10 dataset with SOTA defense schemes across diverse models.

i.e., the accuracy of the model obtained after applying the defense against the clean training set on clean test set.

Main results. The values of average test accuracy that are similar between post-defense and undefended scenarios, are covered by gray demonstrated in Tab. 1. This indicates that `cutout`, `mixup`, `cutmix`, and `DP-SGD` are almost ineffective for CLBPAs. We also highlight the results with accuracy below 50% in light yellow to denote the unqualified defense and accuracy above 80% in light blue to indicate that the defense capability is considered excellent. Therefore, `ISS-G`, `AA`, `OP`, and `AVATAR` exhibit extreme limitations in countering various types of poisons as shown in Tab. 1, rendering them unsuitable as universal defense solutions. As also demonstrated in Tab. 1, two SOTA defense schemes `AT` and `ISS-J`, also lag behind ECLIPSE by more than 8% and 4% in average test accuracy, respectively. In addition, our defense also outperforms these two SOTA defense solutions on ImageNet as shown in Tab. 2.

Given that only `AT` and `ISS-J` achieve comparable defense performance in test accuracy, we further only compare the clean accuracy of these two defenses in Tab. 3. It can be seen that ECLIPSE has an absolute and significant advantage in this metric. Meanwhile, `ISS-J` *causes a damage of approximately 9% on clean training, constituting a fatal flaw that compromises this approach* (the values in Tabs. 2 and 3 covered by deep orange denote the optimal defense effect, while light orange denotes the suboptimal).

4.3 Purification Visual Effect

After undergoing the processes of poison absorption and noise denoising, the resulting image is essentially a purified image, as demonstrated in Fig. 4. It can

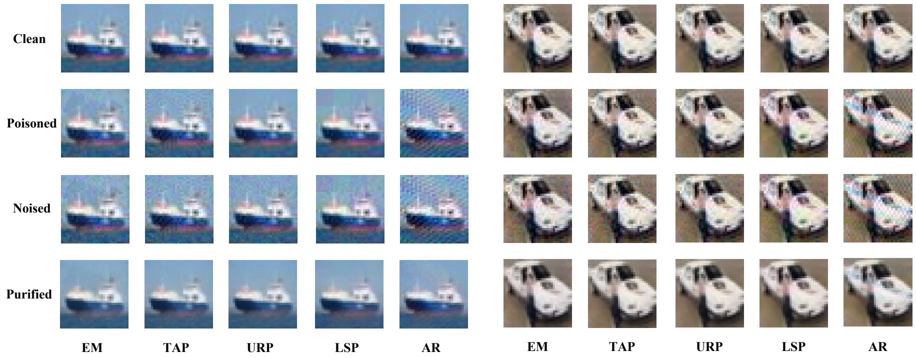


Fig. 4: Visual presentations of five types of CLBPAs, including clean, poisoned, noised, and purified images.

Module↓ Poison→	EM	TAP	REM	SEP	EFP	URP	LSP	AR	**AVG**
A+B+C	82.80	86.13	82.72	82.85	77.20	86.98	84.58	87.32	**83.82**
A+B	73.22	86.31	67.48	41.30	77.58	85.59	81.58	84.22	74.66
A+C	61.30	86.94	77.33	84.07	76.72	87.52	68.70	87.59	78.77
B+C	78.82	83.32	75.51	14.66	81.62	88.87	80.08	60.96	70.48
A	27.45	86.63	35.74	44.97	75.90	86.86	39.93	83.98	60.18
B	78.69	30.95	67.05	7.53	86.14	58.70	68.19	37.63	54.36
C	20.43	84.90	28.83	11.93	79.06	89.54	27.11	33.14	46.87

Table 4: The *test accuracy* (%) results on CIFAR-10 using ResNet18 with diverse combinations. “A”, “B”, “C” denote diffusion purification, grayscale module, and Gaussian noise module. The gray line denotes the best effect in this paper.

be seen that the poisoned images clearly have their poison noise removed after passing through our sparse diffusion purification stage.

4.4 Resistance to Potential Adaptive Attacks

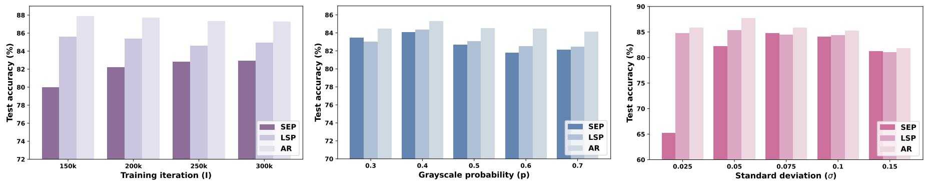
We assume the attacker has knowledge of the structure of the diffusion model and compensation module, and then design an adaptive attack against ECLIPSE, termed as ADP, which involves solving the following optimization objective:

$$\arg\min_{\theta} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_c}\left[\min_{\boldsymbol{\delta_a}} \mathcal{L}\left(U(C(x+\boldsymbol{\delta_a});\theta),y\right)\right]$$

where $\|\boldsymbol{\delta_a}\|_{\boldsymbol{\infty}} \leq \boldsymbol{\epsilon}$ is the adaptive poison with ℓ_∞-norm bounded, U denotes the U-Net [57] that forms the diffusion model, which has been slightly modified into a classification network (the final layer is replaced by a convolutional layer with global average pooling), and C is the corruption function in Eq. (13). The defense performance against ADP is reported in Tab. 1. Our ECLIPSE demonstrates excellent performance in defending against ADP, indicating the robustness of our defense scheme ECLIPSE when facing the adaptive attack.

M↓ Poison→	EM	TAP	REM	SEP	EFP	URP	LSP	AR	**AVG**
0	78.30	86.59	81.56	77.97	78.60	87.52	83.53	87.56	82.70
2	79.72	86.88	82.05	81.29	78.02	86.77	84.50	86.98	83.28
4	82.80	86.13	82.72	82.85	77.20	86.98	84.58	87.32	**83.82**
6	80.75	86.79	82.68	82.71	78.41	86.52	83.88	87.05	83.60

Table 5: The *test accuracy* (%) results on CIFAR-10 using ResNet18 with varying replication times M from ECLIPSE.



(a) p=0.4, σ=0.05

(b) σ=0.1, I=200K

(c) p=0.4, I=200K

Fig. 5: The *test accuracy* (%) results of ECLIPSE on three poisoned CIFAR-10 dataset using ResNet18 with varying hyper-parameters.

4.5 Hyper-Parameter Analysis

We investigate the effects of hyperparameters I, p, σ, M, B/N, and t^* on the ECLIPSE defense performance. It can be observed from Fig. 5 (a) that different values of I had little effect on the final defense performance, which can be attributed to the relatively small size of the training set, allowing the diffusion model to converge well within the above range of I. The impact of different probabilities p of Module B on ECLIPSE can be found in Fig. 5 (b). The excessively low probability of grayscale transformation may lead to incomplete removal of low-frequency poisons, whereas an excessively high probability may compromise certain image features. The impact of different standard deviations σ on ECLIPSE is provided in Fig. 5 (c). Both excessively large and excessively small Gaussian noise lead to a decline in the final defensive performance. In addition, we explore the impact from the replication times M in Tab. 5, the ratio of sparse set (B/N) in Tab. 6, and the forward timestep t^* in Tab. 7. We can see that M mainly influences the defense effect against EM and SEP, which is improved by up to 4.5% and 4.88% compared to not using dataset enlarging module. The potential reason for this improvement is that expanding the total amount of data might help the diffusion model better learn the existing distribution of clean data. The proportion of the sparse set (B/N) does not significantly influence the defense effect. Setting the value of t^* too large or too small will both lead to a decrease in defense effect.

4.6 Ablation Study

Our defense can be divided into three modules, diffusion purification, probabilistic grayscale transformation, and adding Gaussian noise.

B/N ↓ Poison→	EM	TAP	REM	SEP	EFP	URP	LSP	AR	**AVG**
2%	80.25	84.28	82.45	82.39	73.72	84.70	82.04	85.08	81.86
4%	82.80	86.13	82.72	82.85	77.20	86.98	84.58	87.32	**83.82**
6%	80.07	87.05	79.61	78.51	79.82	87.54	84.77	88.25	83.20
8%	78.13	86.75	77.65	65.51	80.39	88.26	81.82	87.48	80.75

Table 6: The *test accuracy* (%) results on CIFAR-10 using ResNet18 with varying dataset ratios B/N (%) from ECLIPSE.

t^* ↓ Poison→	EM	TAP	REM	SEP	EFP	URP	LSP	AR	**AVG**
50	77.23	87.66	78.20	55.07	81.33	88.43	81.97	88.27	79.77
100	82.80	86.13	82.72	82.85	77.20	86.98	84.58	87.32	**83.82**
150	80.91	84.30	82.56	83.99	74.23	83.25	79.46	84.37	81.63

Table 7: The *test accuracy* (%) results on CIFAR-10 using ResNet18 with varying forward timestep t^* from ECLIPSE.

Diffusion purification (A). The results obtained using Module B+C in Tab. 4 indicate the absence of Module A leads to a significant decrease in defense performance against SEP and AR, with a 13.34% average performance drop. This strongly demonstrates the importance of incorporating Module A.

Probabilistic grayscale transformation (B). The absence of Module B results in varying degrees of negative effects on low-frequency poisons as shown in Tab. 4 using Module A+C. This effectiveness of Module B against these poisons is attributed to the fact that low-frequency noise typically affects the color channel information of the images, while grayscale transformation can mask color information and thus suppress low-frequency poisons. As for the reason Module A and Module C are less effective against low-frequency poisons is they rely on adding high-frequency Gaussian noise, thus absorbing low-frequency poisons at a slower rate.

Addition of Gaussian noise (C). The absence of Module C results in a significant decline of 41.55% against SEP when using Module A+B as shown in Tab. 4. To analyze this, we visualize the remaining SEP perturbations after passing through Module A in Fig. 6. Surprisingly, the remaining perturbations are highly similar to the clean samples' features. These high-frequency residual poisons with significant feature similarity can be easily learned by DNNs as shortcuts [11,15,24], leading to a decline in test accuracy. Consequently, by introducing Module C that adds random noise, these relatively fragile high-frequency poisons can be effectively disturbed.

4.7 Analysis of ECLIPSE

We categorize existing CLBPAs into three types: low-frequency poisons (*i.e.*, EM, REM, EFP, and LSP), robust high-frequency poisons (*i.e.*, SEP), and fragile high-frequency poisons (*i.e.*, TAP, URP, and AR). We conclude that, **(i)** *Module A can remove high-frequency poisons.* The effectiveness of Module A against

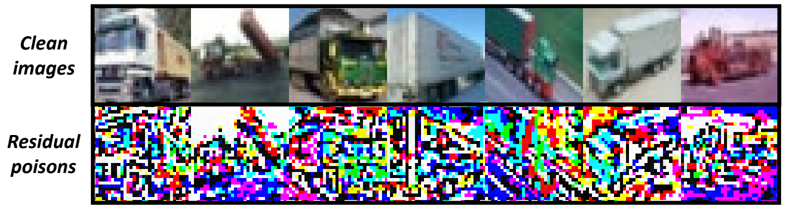


Fig. 6: The residual poisons refer to difference between the SEP images after Module A and the corresponding clean images, which shows that the images in two rows exhibit a high feature similarity.

fragile high-frequency poisons can be verified in Tab. 4. In addition, by increasing t^*, Module A can effectively remove robust high-frequency poisons as shown in Tab. 7, **(ii)** *Module B is crucial for addressing low-frequency poisons.* It is evident from Tab. 4 that as long as Module B is present (*i.e.*, A+B+C, A+B, B+C, and B), the defense against low-frequency poisons is highly effective, and **(iii)** *Module C has a positive impact in expunging low-frequency, robust high-frequency, and fragile high-frequency poisons.* The incremental effect brought by Module C can be observed by comparing the results before and after its inclusion as demonstrated in Tab. 4. Thus, lightweight Gaussian noise is indeed beneficial for expunging CLBPA poisons.

5 Conclusion and Limitation

We propose a brand-new defense scheme called ECLIPSE against the recent rise of clean-label indiscriminate poisoning attacks, which is more universally effective, more practical, and robust. Our scheme is capable of expunging diverse invisible poisons of images via a purification process by a sparse data trained diffusion model and a compensation module. Extensive experiments on benchmark datasets verify that our scheme enjoys high defense effectiveness and robustness against adaptive attack. However, ECLIPSE is not that effective when defending against clean-label indiscriminate poisoning attacks based on convolution-based perturbations [32,42] without norm restriction. This is because the purification scheme based on diffusion models can only remove perturbations with norm constraints [29]. We will leave this issue to our future work.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant No.U20A20177) and Hubei Province Key R&D Technology Special Innovation Project (Grant No.2021BAA032). Shengshan Hu and Peng Xu are co-corresponding authors.

Appendix

Proof for Theorem 1: According to the continuous-time forward process that is defined as the solution to the SDE [39], we have:

$$\mathrm{d}x = f(x,t)\mathrm{d}t + g(t)\mathrm{d}\beta \tag{14}$$

where $f(x,t)$ is the drift coefficient, $g(t)$ is the diffusion coefficient, and $\beta(t)$ is a Brownian motion with diffusion matrix. Then according to the Fokker Planck Kolmogorov equation [36], we have:

$$\begin{aligned}\frac{\partial p(x,t)}{\partial t} &= -\nabla_x \left(f(x,t)p(x,t) - \frac{g^2(t)}{2}\nabla_x p(x,t) \right) \\ &= -\nabla_x \left(f(x,t)p(x,t) - \frac{g^2(t)}{2} p(x,t) \nabla_x \log p(x,t) \right) \\ &= \nabla_x \left(h_p(x,t)p(x,t) \right) \end{aligned} \tag{15}$$

where $h_p(x,t)$ is defined as $\frac{1}{2}g^2(t)\nabla_x \log p(x,t) - f(x,t)$. Then we have:

$$\begin{aligned}\frac{\partial D_{KL}\left(p(x,t)\|q(x,t)\right)}{\partial t} &= \frac{\partial}{\partial t}\int p(x,t)log\frac{p(x,t)}{q(x,t)}dx \\ &= \int \frac{\partial p(x,t)}{\partial t}\log\frac{p(x,t)}{q(x,t)}dx + \int \frac{p(x,t)}{q(x,t)}\frac{\partial q(x,t)}{\partial t}dx + \int \frac{\partial p(x,t)}{\partial t}dx \\ &= \int \nabla_x \left(h_p(x,t)p(x,t)\right)\log\frac{p(x,t)}{q(x,t)}dx + \int \nabla_x \left(h_q(x,t)q(x,t)\right)\frac{p(x,t)}{q(x,t)}dx \\ &= -\int p(x,t)\left[h_p(x,t) - h_q(x,t)\right]^T \left[\nabla_x \log p(x,t) - \nabla_x \log q(x,t)\right]dx \\ &= -\frac{g^2(t)}{2}\int p(x,t)\left\|\nabla_x \log p(x,t) - \nabla_x \log q(x,t)\right\|_2^2 dx \\ &= -\frac{g^2(t)}{2}D_F(p(x,t)||q(x,t)) \end{aligned} \tag{16}$$

where the fourth equality follows from the integration by parts, as well as our assumption of smooth and fast decaying for both $p(x,t)$ and $q(x,t)$, D_F denotes the Fisher divergence [20]. Due to $g^2(t) > 0$ and the fact that the Fisher divergence is greater than or equal to 0, we have:

$$\frac{\partial D_{KL}\left(p(x,t)\|q(x,t)\right)}{\partial t} \leq 0 \tag{17}$$

where the equality occurs only when $p(x,t) = q(x,t)$.

References

1. Biggio, B., Nelson, B., Laskov, P.: Support vector machines under adversarial label noise. In: Proceedings of the 3rd Asian Conference on Machine Learning (ACML'11). pp. 97–112 (2011)

2. Borgnia, E., Cherepanova, V., Fowl, L., Ghiasi, A., Geiping, J., Goldblum, M., Goldstein, T., Gupta, A.: Strong data augmentation sanitizes poisoning and backdoor attacks without an accuracy tradeoff. In: Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'21). pp. 3855–3859 (2021)
3. Chen, S., Yuan, G., Cheng, X., Gong, Y., Qin, M., Wang, Y., Huang, X.: Self-ensemble protection: Training checkpoints are good data protectors. In: Proceedings of the 11th International Conference on Learning Representations (ICLR'23) (2023)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Feifei, L.: Imagenet: A large-scale hierarchical image database. In: Proceedings of the 2009 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'09). pp. 248–255 (2009)
5. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017)
6. Dolatabadi, H.M., Erfani, S., Leckie, C.: The devil's advocate: Shattering the illusion of unexploitable data using diffusion models. arXiv preprint arXiv:2303.08500 (2023)
7. Feng, J., Cai, Q.Z., Zhou, Z.H.: Learning to confuse: Generating training time adversarial data with auto-encoder. In: Proceedings of the 33rd Neural Information Processing Systems (NeruIPS'19). vol. 32, pp. 11971–11981 (2019)
8. Fowl, L., Chiang, P.y., Goldblum, M., Geiping, J., Bansal, A., Czaja, W., Goldstein, T.: Preventing unauthorized use of proprietary data: Poisoning for secure dataset release. arXiv preprint arXiv:2103.02683 (2021)
9. Fowl, L., Goldblum, M., Chiang, P.y., Geiping, J., Czaja, W., Goldstein, T.: Adversarial examples make strong poisons. In: Proceedings of the 35th Neural Information Processing Systems (NeurIPS'21). vol. 34, pp. 30339–30351 (2021)
10. Fu, S., He, F., Liu, Y., Shen, L., Tao, D.: Robust unlearnable examples: Protecting data privacy against adversarial learning. In: Proceedings of the 10th International Conference on Learning Representations (ICLR'22) (2022)
11. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**, 665–673 (2020)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'16). pp. 770–778 (2016)
13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Proceedings of the 34th Neural Information Processing Systems (NeurIPS'20). vol. 33, pp. 6840–6851 (2020)
14. Hong, S., Chandrasekaran, V., Kaya, Y., Dumitraş, T., Papernot, N.: On the effectiveness of mitigating data poisoning attacks with gradient shaping. arXiv preprint arXiv:2002.11497 (2020)
15. Hu, S., Liu, W., Li, M., Zhang, Y., Liu, X., Wang, X., Zhang, L.Y., Hou, J.: Pointcrt: Detecting backdoor in 3d point cloud via corruption robustness. In: Proceedings of the 31st ACM International Conference on Multimedia (MM'23). pp. 666–675 (2023)
16. Hu, S., Zhou, Z., Zhang, Y., Zhang, L.Y., Zheng, Y., He, Y., Jin, H.: Badhash: Invisible backdoor attacks against deep hashing with clean label. In: Proceedings of the 30th ACM International Conference on Multimedia (MM'22). pp. 678–686 (2022)

17. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the 2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'17). pp. 4700–4708 (2017)
18. Huang, H., Ma, X., Erfani, S.M., Bailey, J., Wang, Y.: Unlearnable examples: Making personal data unexploitable. In: Proceedings of the 9th International Conference on Learning Representations (ICLR'21) (2021)
19. Jiang, W., Diao, Y., Wang, H., Sun, J., Wang, M., Hong, R.: Unlearnable examples give a false sense of security: Piercing through unexploitable data with learnable examples. In: Proceedings of the 31st ACM International Conference on Multimedia (MM'23). p. 8910–8921 (2023)
20. Kostrikov, I., Fergus, R., Tompson, J., Nachum, O.: Offline reinforcement learning with fisher divergence critic regularization. In: Proceedings of the 38th International Conference on Machine Learning (ICML'21). pp. 5774–5783 (2021)
21. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images (2009)
22. Kumar, R.S.S., Nyström, M., Lambert, J., Marshall, A., Goertzel, M., Comissoneru, A., Swann, M., Xia, S.: Adversarial machine learning-industry perspectives. In: Proceedings of the 2020 IEEE Security and Privacy Workshops (SPW'20). pp. 69–75 (2020)
23. Lin, X., Yu, Y., Xia, S., Jiang, J., Wang, H., Yu, Z., Liu, Y., Fu, Y., Wang, S., Tang, W., et al.: Safeguarding medical image segmentation datasets against unauthorized training via contour-and texture-aware perturbations. arXiv preprint arXiv:2403.14250 (2024)
24. Liu, X., Li, M., Wang, H., Hu, S., Ye, D., Jin, H., Wu, L., Xiao, C.: Detecting backdoors during the inference stage based on corruption robustness consistency. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'23). pp. 16363–16372 (2023)
25. Liu, Z., Zhao, Z., Larson, M.: Image shortcut squeezing: Countering perturbative availability poisons with compression. In: Proceedings of the 40th International Conference on Machine Learning (ICML'23) (2023)
26. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: Proceedings of the 6th International Conference on Learning Representations (ICLR'18) (2018)
27. Muñoz-González, L., Biggio, B., Demontis, A., Paudice, A., Wongrassamee, V., Lupu, E.C., Roli, F.: Towards poisoning of deep learning algorithms with back-gradient optimization. In: Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security (AISec'17). pp. 27–38 (2017)
28. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: Proceedings of the 38th International Conference on Machine Learning (ICML'21). pp. 8162–8171 (2021)
29. Nie, W., Guo, B., Huang, Y., Xiao, C., Vahdat, A., Anandkumar, A.: Diffusion models for adversarial purification. In: Proceedings of the 39th International Conference on Machine Learning (ICML'22) (2022)
30. Qin, T., Gao, X., Zhao, J., Ye, K., Xu, C.Z.: Learning the unlearnable: Adversarial augmentations suppress unlearnable example attacks. arXiv preprint arXiv:2303.15127 (2023)
31. Ren, J., Xu, H., Wan, Y., Ma, X., Sun, L., Tang, J.: Transferable unlearnable examples. In: Proceedings of the 11th International Conference on Learning Representations (ICLR'23) (2023)

32. Sadasivan, V.S., Soltanolkotabi, M., Feizi, S.: Cuda: Convolution-based unlearnable datasets. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'23). pp. 3862–3871 (2023)
33. Sandoval-Segura, P., Singla, V., Fowl, L., Geiping, J., Goldblum, M., Jacobs, D., Goldstein, T.: Poisons that are learned faster are more effective. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'22). pp. 198–205 (2022)
34. Sandoval-Segura, P., Singla, V., Geiping, J., Goldblum, M., Goldstein, T.: What can we learn from unlearnable datasets? In: Proceedings of the 37th Neural Information Processing Systems (NeurIPS'23) (2023)
35. Sandoval-Segura, P., Singla, V., Geiping, J., Goldblum, M., Goldstein, T., Jacobs, D.W.: Autoregressive perturbations for data poisoning. In: Proceedings of the 36th Neural Information Processing Systems (NeurIPS'22). vol. 35 (2022)
36. Särkkä, S., Solin, A.: Applied stochastic differential equations, vol. 10. Cambridge University Press (2019)
37. Shen, J., Zhu, X., Ma, D.: Tensorclog: An imperceptible poisoning attack on deep neural network applications. IEEE Access **7**, 41498–41506 (2019)
38. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
39. Song, Y., Durkan, C., Murray, I., Ermon, S.: Maximum likelihood training of score-based diffusion models. In: Proceedings of the 35th Neural Information Processing Systems (NeurIPS'21). vol. 34, pp. 1415–1428 (2021)
40. Song, Y., Sohl-Dickstein, J., P Kingma, D., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: Proceedings of the 9th International Conference on Learning Representations (ICLR'21) (2021)
41. Tao, L., Feng, L., Yi, J., Huang, S.J., Chen, S.: Better safe than sorry: Preventing delusive adversaries with adversarial training. In: Proceedings of the 35th Neural Information Processing Systems (NeurIPS'21). vol. 34, pp. 16209–16225 (2021)
42. Wang, X., Hu, S., Li, M., Yu, Z., Zhou, Z., Zhang, L.Y.: Corrupting convolution-based unlearnable datasets with pixel-based image transformations. arXiv preprint arXiv:2311.18403 (2023)
43. Wang, Z., Wang, Y., Wang, Y.: Fooling adversarial training with inducing noise. arXiv preprint arXiv:2111.10130 (2021)
44. Wen, R., Zhao, Z., Liu, Z., Backes, M., Wang, T., Zhang, Y.: Is adversarial training really a silver bullet for mitigating data poisoning? In: Proceedings of the 11th International Conference on Learning Representations (ICLR'23) (2023)
45. Yu, D., Zhang, H., Chen, W., Yin, J., Liu, T.Y.: Availability attacks create shortcuts. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'22). pp. 2367–2376 (2022)
46. Yu, Y., Wang, Y., Yang, W., Lu, S., Tan, Y.P., Kot, A.C.: Backdoor attacks against deep image compression via adaptive frequency trigger. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'23). pp. 12250–12259 (2023)
47. Yuan, C.H., Wu, S.H.: Neural tangent generalization attacks. In: Proceedings of the 38th International Conference on Machine Learning (ICML'21). pp. 12230–12240 (2021)
48. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the 17th International Conference on Computer Vision (ICCV'19) (2019)

49. Zhang, H., Hu, S., Wang, Y., Zhang, L.Y., Zhou, Z., Wang, X., Zhang, Y., Chen, C.: Detector collapse: Backdooring object detection to catastrophic overload or blindness. In: Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI'24) (2024)
50. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: Mixup: Beyond empirical risk minimization. In: Proceedings of the 6th International Conference on Learning Representations (ICLR'18) (2018)
51. Zhang, L., Shen, B., Barnawi, A., Xi, S., Kumar, N., Wu, Y.: Feddpgan: Federated differentially private generative adversarial networks framework for the detection of covid-19 pneumonia. Information Systems Frontiers **23**(6), 1403–1415 (2021)
52. Zhang, R., Zhu, Q.: A game-theoretic analysis of label flipping attacks on distributed support vector machines. In: Proceedings of the 51st Annual Conference on Information Sciences and Systems (CISS'17). pp. 1–6 (2017)
53. Zhang, Y., Hu, S., Zhang, L.Y., Shi, J., Li, M., Liu, X., Wan, W., Jin, H.: Why does little robustness help? a further step towards understanding adversarial transferability. In: Proceedings of the 45th IEEE Symposium on Security and Privacy (S&P'24). vol. 2 (2024)
54. Zhou, Z., Hu, S., Li, M., Zhang, H., Zhang, Y., Jin, H.: Advclip: Downstream-agnostic adversarial examples in multimodal contrastive learning. In: Proceedings of the 31st ACM International Conference on Multimedia (MM'23). pp. 6311–6320 (2023)
55. Zhou, Z., Hu, S., Zhao, R., Wang, Q., Zhang, L.Y., Hou, J., Jin, H.: Downstream-agnostic adversarial examples. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV'23). pp. 4345–4355 (2023)
56. Zhou, Z., Li, M., Liu, W., Hu, S., Zhang, Y., Wan, W., Xue, L., Zhang, L.Y., Yao, D., Jin, H.: Securely fine-tuning pre-trained encoders against adversarial examples. In: Proceedings of the 2024 IEEE Symposium on Security and Privacy (SP'24) (2024)
57. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: Proceedings of the International Workshop on Deep Learning in Medical Image Analysis. pp. 3–11 (2018)