Stable Unlearnable Example: Enhancing the Robustness of Unlearnable Examples via Stable Error-Minimizing Noise

Yixin Liu[1], **Kaidi Xu**[2], **Xun Chen**[3], **Lichao Sun**[1]
[1] Lehigh University, Bethlehem, Pennsylvania, USA
[2] Drexel University, Philadelphia, Pennsylvania, USA
[2] Samsung Research America, Mountain View, California, USA
{yila22, lis221}@lehigh.edu, kx46@drexel.edu, Xun.chen@samsung.com

Abstract

The open sourcing of large amounts of image data promotes the development of deep learning techniques. Along with this comes the privacy risk of these image datasets being exploited by unauthorized third parties to train deep learning models for commercial or illegal purposes. To avoid the abuse of data, a poisoning-based technique, "unlearnable example", has been proposed to significantly degrade the generalization performance of models by adding imperceptible noise to the data. To further enhance its robustness against adversarial training, existing works leverage iterative adversarial training on both the defensive noise and the surrogate model. However, it still remains unknown whether the robustness of unlearnable examples primarily comes from the effect of enhancement in the surrogate model or the defensive noise. Observing that simply removing the adversarial perturbation on the training process of the defensive noise can improve the performance of robust unlearnable examples, we identify that solely the surrogate model's robustness contributes to the performance. Furthermore, we found a negative correlation exists between the robustness of defensive noise and the protection performance, indicating defensive noise's instability issue. Motivated by this, to further boost the robust unlearnable example, we introduce Stable Error-Minimizing noise (SEM), which trains the defensive noise against random perturbation instead of the time-consuming adversarial perturbation to improve the stability of defensive noise. Through comprehensive experiments, we demonstrate that SEM achieves a new state-of-the-art performance on CIFAR-10, CIFAR-100, and ImageNet Subset regarding both effectiveness and efficiency.

Introduction

The proliferation of open-source and large-scale datasets on the Internet has significantly advanced deep learning techniques across various fields, including computer vision (He et al. 2016; Dosovitskiy et al. 2020; Cao et al. 2023; Zhang et al. 2023b), natural language processing (Devlin et al. 2019; Vaswani et al. 2017; Zhou et al. 2023; Sun et al. 2024), and graph data analysis (Wang et al. 2018; Kipf and Welling 2016; Sun et al. 2022a). However, this emergence also poses significant privacy threats, as these datasets can be exploited by unauthorized third parties to train deep neural networks (DNNs) for commercial or even illegal purposes. For instance, Hill and Krolik (2019) reported that a tech company amassed extensive facial data without consent to develop commercial face recognition models. Besides, Edwards (2022) reported the discovery of an artist's private medical record photos, initially intended for private use, in a popular AI training set.

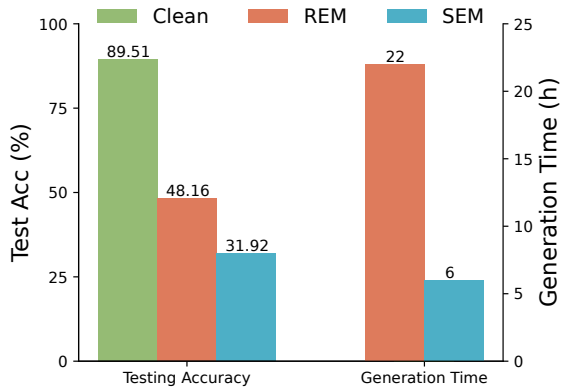


Figure 1: The performance comparison on CIFAR-10 between the current SoTA method, the robust error-minimizing noise (REM) (Fu et al. 2022), and our proposed stable error-minimizing noise (SEM). Our SEM outperforms the REM in terms of both effectiveness and generation efficiency.

To tackle this problem, recent works have introduced *Unlearnable Examples* (Fowl et al. 2021b; Huang et al. 2020; Fowl et al. 2021a), which aims to make the training data "unlearnable" for deep learning models by adding a type of invisible noise. This added noise misleads the learning model into adopting meaningless shortcut patterns from the data rather than extracting informative knowledge (Yu et al. 2022). However, such conferred unlearnability is vulnerable to adversarial training (Huang et al. 2020). In response, the concept of robust error-minimizing noise (REM) was proposed in (Fu et al. 2022) to enhance the efficacy of error-minimizing noise under adversarial training, thereby shielding data from adversarial learners by minimizing adversarial training loss. Specifically, a min-min-max optimization strategy is employed to train the robust error-minimizing noise generator, which in turn produces robust error-minimizing noise. To enhance stability against data transformation, REM

leverages the Expectation of Transformation (EOT) (Athalye et al. 2018) during the generator training. The resulting noise demonstrates superior performance in advanced training that involves both adversarial training and data transformation.

However, a primary drawback of REM, is its high computational cost. Specifically, on the CIFAR-10 dataset, it would take nearly a full day to generate the unlearnable examples, which is very inefficient. Consequently, enhancing the efficiency of REM is vital, especially when scaling up to larger real-world datasets (Russakovsky et al. 2015; Schuhmann et al. 2021). To improve the efficiency of robust unlearnable example generation algorithms, in this paper, we take a closer look at the time-consuming adversarial training process in both the surrogate model and the defensive noise. As empirically demonstrated in Tab. 1, we can see that the performance of the robust unlearnable example mainly comes from the effect of adversarial training on the surrogate model rather than the defensive noise part. Surprisingly, the presence of adversarial perturbation in the defensive noise crafting will even lead to performance degradation, indicating that we need a better optimization method in this part.

To elucidate this intriguing phenomenon and enhance the robustness of unlearnable examples, we begin by defining the robustness of both the surrogate model and defensive noise. Subsequently, our correlation analysis reveals that the robustness of the surrogate model is the primary contributing factor. Conversely, we observe a negative correlation between the robustness of defensive noise and data protection performance. We hypothesize that the defensive noise overfits monotonous adversarial perturbations, leading to its instability. To address this issue, we introduce a novel noise type, *stable error-minimizing noise* (SEM). Our SEM is trained against random perturbations, rather than the more time-consuming adversarial perturbations, to enhance stability. We summarize our contributions as follows:

- We establish that the robustness of unlearnable examples is largely attributable to the surrogate model's robustness, rather than that of the defensive noise. Furthermore, we find that adversarially enhancing defensive noise can actually degrade its protective performance.
- To mitigate such an instability issue, we introduce stable error-minimizing noise (SEM), which trains the defensive noise against random perturbations instead of the more time-consuming adversarial ones, to improve the stability of the defensive noise.
- Extensive experiments empirically demonstrate that SEM achieves a SoTA performance on CIFAR-10, CIFAR-100, and ImageNet Subset regarding both effectiveness and efficiency. Notably, SEM achieves a $3.91\times$ speedup and an approximately 17% increase in testing accuracy for protection performance on CIFAR-10 under adversarial training with $\epsilon = 4/255$.

Preliminaries

Setup. We consider a classification task with input-label pairs from a K-class dataset $\mathcal{T} = (x_i, y_i)_{i=1}^n \subseteq \mathcal{X} \times \mathcal{Y}$, where $\mathcal{T}$ is constructed by drawing from an underlying distribution $\mathcal{D}(x, y)$ in an *i.i.d.* manner, and $x \in \mathbb{R}^d$ represents the features of a sample. Let $f_\theta : \mathbb{R}^d \to \Delta$ be a neural network model that outputs a probability simplex, e.g., via a softmax layer. In most learning algorithms, we employ empirical risk minimization (ERM), which aims to train the model f_θ by minimizing a loss function $\mathcal{L}(f(x), y)$ on the clean training set. This is achieved by solving the following optimization:

$$\min_\theta \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\left(f_\theta\left(x_i\right), y_i\right). \tag{1}$$

Method	Adv. Train		δ^r	Time	Test Acc. (%) ↓
	θ	δ^u			
REM	✓	✓		22h	46.72
REM-δ^u		✓		15h	88.28
REM-θ	✓			6h	37.90
SEM	✓		✓	**6h**	**30.26**

Table 1: Comparison between variants of REM and SEM on CIFAR-10 dataset. The ✓ indicates the corresponding method is used in the training process. δ^r indicates the defensive noise that is trained against random perturbation.

Adversarial Training. Adversarial training aims to enhance the robustness of models against adversarially perturbed examples (Madry et al. 2018). Specifically, adversarial robustness necessitates that f_θ performs well not only on $\mathcal{D}$ but also on the worst-case perturbed distribution close to $\mathcal{D}$, within a given adversarial perturbation budget. In this paper, we focus on the adversarial robustness of ℓ_∞-*norm*: *i.e.*, for a small $\epsilon > 0$, we aim to train a classifier f_θ that correctly classifies $(x + \delta, y)$ for any $\|\delta\|_\infty \leq \rho_a$[1], where $(x, y) \sim \mathcal{D}$. Typically, adversarial training methods formalize the training of f_θ as a min-max optimization with respect to θ and δ, *i.e.*,

$$\min_\theta \frac{1}{n} \sum_{i=1}^{n} \max_{\|\delta_i\| \leq \rho_a} \mathcal{L}\left(f_\theta\left(x_i + \delta_i\right), y_i\right). \tag{2}$$

Unlearnable Example. Unlearnable examples (Huang et al. 2020) leverage clean-label data poisoning techniques to trick deep learning models into learning minimal useful knowledge from the data, thus achieving the data protection goal. By adding imperceptible defensive noise to the data, this technique introduces misleading shortcut patterns to the training process, thereby preventing the models from acquiring any informative knowledge. Models trained on these perturbed datasets exhibit poor generalization ability. Formally, this task can be formalized into the following bi-level optimization:

$$\begin{aligned} &\max_{\|\delta_i^u\| \leq \rho_a} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[\mathcal{L}\left(f_{\theta^*}(x), y\right)\right], \\ &\text{s.t. } \theta^* = \arg\min_\theta \sum_{(x_i, y_i)\in\mathcal{T}} \left[\mathcal{L}\left(f_\theta\left(x_i + \delta_i^u\right), y_i\right)\right]. \end{aligned} \tag{3}$$

Here, we aim to maximize the empirical risk of trained models by applying the generated *defensive perturbation*

[1] $\|\cdot\|$ in subsequent sections omits the subscript "∞" for brevity.

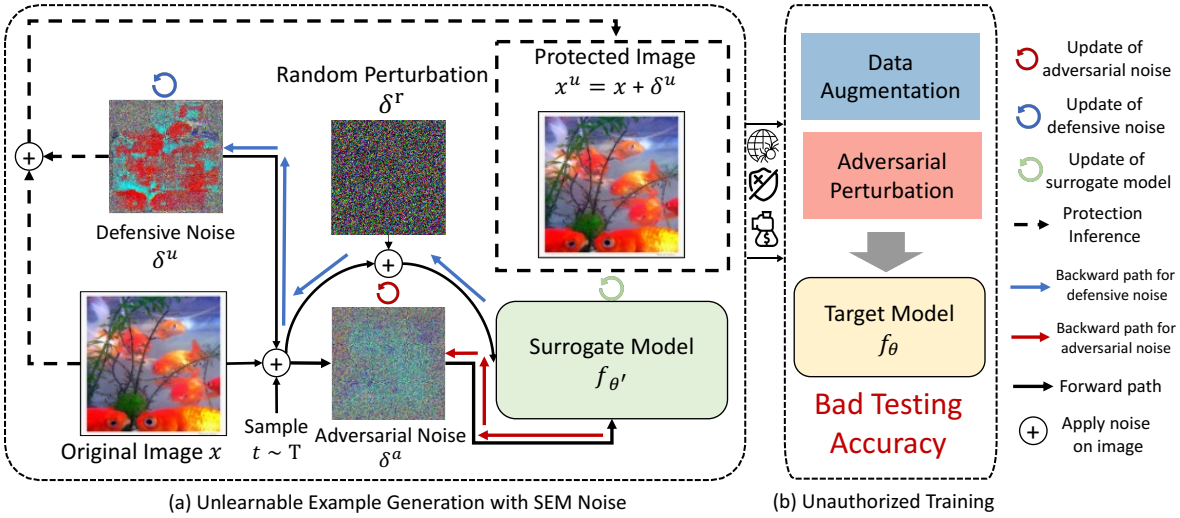


Figure 2: The overall framework of our approach. Our approach consists of two phases: noise training and generator training. During the noise training phase, we train the defensive noise, denoted as δ^u, to counter random perturbations. In the subsequent generator training phase, the original images, represented as x, are transformed to $x_{\text{input}} = t(x + \delta^u) + \delta^a$ before being input into the network. Here, t represents a transformation derived from distribution T, and δ^a represents the adversarial perturbation produced using PGD. The noise generator, f'_θ, updates the network parameters, θ, by minimizing adversarial loss. By applying our defensive noise, models trained on the protected data learn minimal information and exhibit poor performance on clean data.

$\mathcal{P}^u = \{\delta_i^u\}_{i=1}^n$ into the original training set $\mathcal{T}$. Owing to the complexity of directly solving the bi-level optimization problem outlined in Eq. 3, several approximate methods have been proposed. These approaches include rule-based (Yu et al. 2022), heuristic-based (Huang et al. 2020), and optimization-based methods (Feng, Cai, and Zhou 2019), all of which achieve satisfactory performance in solving Eq. 3.

Robust Unlearnable Example. However, recent studies (Fu et al. 2022; Huang et al. 2020; Tao et al. 2021) have demonstrated that the effectiveness of unlearnable examples can be compromised by employing adversarial training. To further address this issue, the following robust unlearnable example generation problem is proposed, which can be illustrated as a two-player game consisting of a defender U and an unauthorized user $\mathcal{A}$. The defender U aims to protect data privacy by adding perturbation $\mathcal{P}^u$ to the data, thereby decreasing the test accuracy of the trained model, while the unauthorized user $\mathcal{A}$ attempts to use *adversarial training* and *data transformation* to purify the added perturbation and "recover" the original test accuracy. Based on Fu et al. (2022), we assume that the defender U has complete access to the data they intend to protect and cannot interfere with the user's training process after the protected images are released. Additionally, we assume, as per Fu et al. (2022), that the radius of defensive noise ρ_u exceeds that of the adversarial training radius ρ_a, ensuring the problem's feasibility. Given a distribution T over transformations $t : \mathcal{X} \to \mathcal{X}$, we have

$$\begin{aligned} \mathcal{P}^u = & \underset{\forall i, ||\delta_i^u|| \leq \rho_u}{\arg\max} \underset{(x,y)\sim\mathcal{D}}{\mathbb{E}} \left[\mathcal{L}\left(f_{\theta^*}(x), y\right)\right], \\ \text{s.t.} \theta^* = & \arg\min_\theta \textstyle\sum_{(x_i,y_i)\in\mathcal{T}} \underset{t\sim T||\delta_i^a||\leq\rho_a}{\mathbb{E}} \max \mathcal{L}\left(f_\theta\left(x_i'\right), y_i\right), \end{aligned} \tag{4}$$

where $x_i' = t(x_i + \delta_i^u) + \delta_i^a$ represents the transformed data, with δ_i^a being the adversarial perturbation crafted using Projected Gradient Descent (PGD) and δ_i^u denoting the defensive noise. After applying $\mathcal{P}^u$, the protected dataset $\mathcal{T}^u = (x_i + \delta_i^u, y_i)_{i=1}^n$ is obtained.

Methodology

Iterative Training of Generator and Defensive Noise

To address the problem presented in Eq. 4, REM introduces a robust noise-surrogate iterative optimization method, where a surrogate noise generator model, denoted as θ, and the defensive noise, δ^u, are optimized alternately. From the model's perspective, the surrogate model θ is trained on iteratively *perturbed poisoned data*, created by adding defensive and adversarial noise to the original training data, namely, $x^u_{\text{perturb}} = t(x + \delta^u) + \delta^a$. The surrogate model is trained to minimize the loss below to enhance its adversarial robustness:

$$\theta' \leftarrow \theta - \eta^\theta \nabla_\theta \mathcal{L}\left(f_\theta\left(t\left(x + \delta^u\right) + \delta^a\right), y\right). \tag{5}$$

where $\eta^{(\cdot)}$ represents the learning rate, t is a transformation sampled from T, and δ_a is the adversarial perturbation crafted using PGD, designed to maximize the loss. Conversely, the defensive noise δ^u is trained to counteract the worst-case adversarial perturbation using PGD based on θ. The main idea is that robust defensive noise should maintain effectiveness even under adversarial perturbations. Specifically, the defensive noise is updated by minimizing the loss,

$$\delta^u \leftarrow \delta^u - \eta^u \nabla_{\delta^u} \mathcal{L}\left(f_\theta\left(t\left(x + \delta^u\right) + \delta^a\right), y\right). \tag{6}$$

During the optimization of the surrogate model θ, the defensive noise δ^u is fixed. Conversely, when optimizing the defensive noise δ^u, the surrogate model θ remains fixed. We

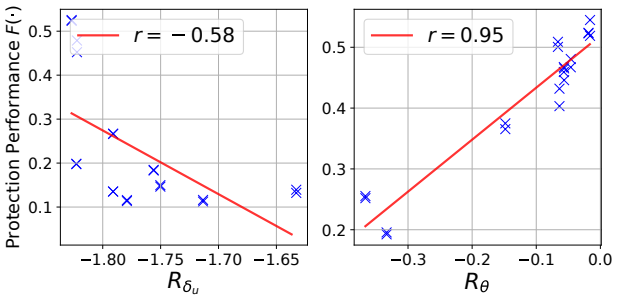


Figure 3: Exploration of the contribution of the robustness of defensive noise, denoted as $\mathcal{R}_\theta$, and the surrogate model, represented by $\mathcal{R}_{\delta^u}$, to the protection performance F. The Pearson correlation coefficients (r) quantify the strength of these relationships. Tests were conducted on the CIFAR-10 dataset with settings $\rho_a = 4/255$ and $\rho_u = 8/255$.

repeat these steps iteratively until the maximum training step is reached. The final defensive noise is generated by,

$$\delta^u = \arg\min_{\|\delta^u\| \le \rho_u} \mathbb{E}_{t\sim T} \max_{\|\delta^a\| \le \rho_a} \mathcal{L}\left(f_\theta\left(t\left(x+\delta^u\right)+\delta^a\right), y\right). \tag{7}$$

Improving the Stability of Defensive Noise

The adversarial perturbation, denoted as δ^a, is incorporated in the optimization processes of *both* the surrogate model θ and the defensive noise δ^u (refer to Eq. 5 and Eq. 6). However, as indicated by the results in Tab. 1, it appears that solely the adversarial perturbation δ^a in the surrogate model θ's optimization contributes to the performance enhancement. Surprisingly, the adversarial perturbation in the optimization of defensive noise δ^u (see Eq. 6) results in *performance degradation*. To elucidate these intriguing findings, we proceed with a correlation analysis as described below. We initially define the protection performance by $F = 1 - \texttt{Acc}$, where $\texttt{Acc}$ represents the testing accuracy of the trained model. Then, we propose the following definition to quantify the robustness of both the surrogate model and the defensive noise.

Definition 1 (Robustness of surrogate model)**.** Given a fixed surrogate model, denoted as θ, we define its robustness as the model's resistance to adversarial perturbations, where the perturbation δ^u is updateable,

$$\mathcal{R}_\theta = -\max_{\|\delta^a\| \le \rho_a} \min_{\|\delta^u\| \le \rho_u} \mathcal{L}\left(f_\theta\left(t\left(x+\delta^u\right)+\delta^a\right), y\right). \tag{8}$$

Definition 2 (Robustness of defensive noise)**.** For a given fixed defensive noise, δ^u, we define its robustness as the noise's resistance to adversarial perturbations, where the model parameter θ is updateable,

$$\mathcal{R}_{\delta^u} = -\max_{\|\delta^a\| \le \rho_a} \min_{\theta} \mathcal{L}\left(f_\theta\left(t\left(x+\delta^u\right)+\delta^a\right), y\right). \tag{9}$$

To explore the correlation between the two formulated robustness terms, $\mathcal{R}_{(\cdot)}$, and the protection performance, F, we followed the standard defensive noise training procedure as outlined in REM (Fu et al. 2022) and stored the surrogate model, θ_t, and defensive noise, δ_t^u, at various training steps denoted by t. Subsequently, for the obtained model or noise, we fixed one while randomly initializing the other, then solved Eq. 8 and Eq. 9 through an iterative training process. From Fig. 3, we found that the $\mathcal{R}_\theta$ demonstrates a strong positive correlation with the protection performance while $\mathcal{R}_{\delta_u}$ displays a moderate negative correlation with the protection performance.

This suggests that the protection performance of defensive noise is primarily and positively influenced by the robustness of the surrogate model, $\mathcal{R}_\theta$. Conversely, enhancing the robustness of defensive noise may paradoxically impair its protection performance. We hypothesize that the instability of defensive noise δ^u, trained following Eq. 6, stems from the monotonic nature of the worst-case perturbation δ^a. To enhance its stability, we propose an alternative training objective for defensive noise,

$$\delta^u \leftarrow \delta^u - \eta^u \nabla_{\delta^u} \mathcal{L}\left(f_\theta\left(t\left(x+\delta^u\right)+\delta^r\right), y\right), \tag{10}$$

where δ^r represents a random perturbation sampled from the uniform distribution $\mathcal{U}(-\rho_r, \rho_r)$. The radius of the random perturbation, ρ_r, is set to match that of the adversarial perturbation, ρ_a. Substituting δ^a in Eq. 6 with δ^r enables the crafted defensive noise to experience more diverse perturbations during training. We term the obtained defensive noise as stable error-minimizing noise (SEM). The overall framework and procedure are depicted in Fig. 2 and Alg. 1.

Algorithm 1: Noise Generator Training for SEM approach

Require: training data set $\mathcal{T}$, training steps M, defense PGD parameters ρ_u, α_u and K_u, attack PGD parameters ρ_a, α_a and K_a, transformation distribution T, the sampling number J for gradient approximation

Ensure: Robust noise generator f'_θ.

for i **in** $1, \cdots, M$ **do**
  Sample minibatch $(x, y) \sim \mathcal{T}$, randomly initialize δ^u.
  for k **in** $1, \cdots, K_u$ **do**
    for j **in** $1, \cdots, J$ **do**
      Sample noise and transformation $\delta_j^r \sim \mathcal{P}$, $t_j \sim T$.
    end for
    $g_k \leftarrow \frac{1}{J}\sum_{j=1}^{J} \frac{\partial}{\partial \delta^u} \ell(f'_\theta(t_j(x+\delta^u)+\delta_j^r), y)$
    $\delta^u \leftarrow \prod_{\|\delta\| \le \rho_u} (\delta^u - \alpha_u \cdot \mathrm{sign}(g_k))$
  end for
  Sample a transformation function $t \sim T$.
  $\delta^a \leftarrow \mathrm{PGD}(t(x+\delta^u), y, f'_\theta, \rho_a, \alpha_a, K_a)$
  Update source model f'_θ based on minibatch $(t(x+\delta^u)+\delta^a, y)$
end for

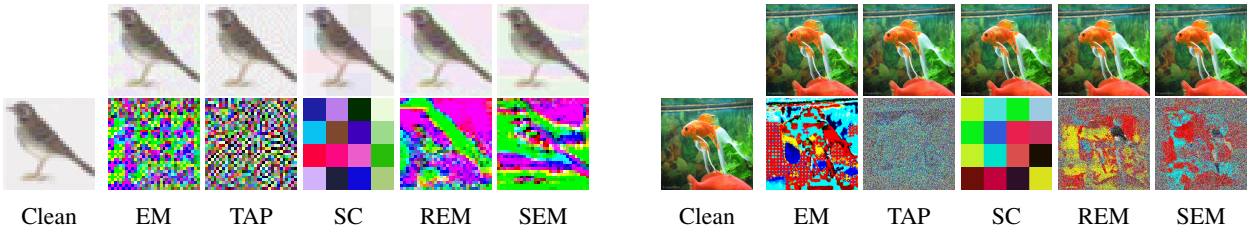
Experiments

Experimental Setup

Dataset. We conducted extensive experiments on three widely recognized vision benchmark datasets, including CIFAR-10, CIFAR-100 (Krizhevsky et al. 2009), and a subset of the ImageNet dataset (Russakovsky et al. 2015), which comprises the first 100 classes from the original ImageNet.

Datasets→	CIFAR-10					CIFAR-100					ImageNet Subset				
Methods↓	ρ_a=0	1/255	2/255	3/255	4/255	0	1/255	2/255	3/255	4/255	0	1/255	2/255	3/255	4/255
Clean Data	94.66	93.74	92.37	90.90	89.51	76.27	71.90	68.91	66.45	64.50	80.66	76.20	72.52	69.68	66.62
EM	13.20	22.08	71.43	87.71	88.62	1.60	71.47	68.49	65.66	63.43	**1.26**	74.88	71.74	66.90	63.40
TAP	22.51	92.16	90.53	89.55	88.02	13.75	70.03	66.91	64.30	62.39	9.10	75.14	70.56	67.64	63.56
NTGA	16.27	41.53	85.13	89.41	88.96	3.22	65.74	66.53	64.80	62.44	8.42	63.28	66.96	65.98	63.06
SC	**11.63**	91.71	90.42	86.84	87.26	**1.51**	70.62	67.95	65.81	63.30	11.0	75.06	71.26	67.14	62.58
REM	15.18	14.28	25.41	30.85	48.16	1.89	4.47	7.03	17.55	27.10	13.74	21.58	29.40	35.76	41.66
SEM	13.0	**12.16**	**11.49**	**20.91**	**31.92**	1.95	**3.26**	**4.35**	**9.07**	**20.25**	4.1	**10.34**	**13.76**	**23.58**	**37.82**

Table 2: Test accuracy (%) of models trained on data protected by different defensive noises under adversarial training with various perturbation radii. The defensive perturbation radius ρ_u is globally set at $8/255$, while the adversarial perturbation radius ρ_a of REM varies. The lower the test accuracy, the better the effectiveness of the protection.



(a) CIFAR-10 (b) ImageNet Subset

Figure 4: Visualization of various noise and crafted examples for CIFAR-10 and ImageNet Subset datasets. The noise includes EM (Error-Minimizing noise), TAP (Targeted Adversarial Poisoning noise), NTGA (Neural Tangent Generalization Attack noise), SC (Shortcut noise), REM (Robust Error-Minimizing noise), and SEM (our proposed Stable Error-Minimizing noise).

In line with Fu et al. (2022), we utilized random cropping and horizontal flipping for data augmentations.

Model Architecture and Adversarial Training. We evaluate the proposed method and the baselines across various vision tasks using five popular network architectures: VGG-16 (Simonyan and Zisserman 2015), ResNet18/50 (He et al. 2016), DenseNet121 (Huang et al. 2017), and Wide ResNet-34-10 (Zagoruyko and Komodakis 2016). ResNet-18 is selected as the default target model in our experiments. By default, the defensive noise radius ρ_u is set at $8/255$, and the adversarial training radius ρ_a is set at $4/255$ for each dataset. Besides, the setting for the adversarial training radius follows the guidelines of REM (Fu et al. 2022), ensuring $\rho_a \leq \rho_u$.

Baselines. We compare our stable error-minimizing noise (**SEM**) with existing SoTA methods, including targeted adversarial poisoning noise (**TAP**) (Fowl et al. 2021a), error-minimizing noise (**EM**) (Huang et al. 2021), robust error-minimizing noise (**REM**) (Fu et al. 2022), synthesized shortcut noise (**SC**) (Yu et al. 2022), and neural tangent generalization attack noise (**NTGA**) (Yuan and Wu 2021).

Performance Analysis

Different Radii of Adversarial Training. Our initial evaluation focuses on the protection performance of various unlearnable examples under different adversarial training radii. The defensive perturbation radius is set to $\rho_u = 8/255$, with ResNet-18 serving as both the source and target models. As shown in Tab. 1, SC performs best under standard training settings ($\rho_a = 0$). However, with an increase in the adversarial training perturbation radius, the test accuracy of SC also rises significantly. Likewise, the other two baselines, EM and TAP, experience a significant increase in test accuracy with intensified adversarial training. In contrast, REM and SEM, designed for better robustness against adversarial perturbations, do not significantly increase test accuracy even with larger radii. When compared to REM, the results indicate that our SEM consistently outperforms this baseline in all adversarial training settings ($\rho_a \in [\frac{1}{255}, \frac{4}{255}]$), demonstrating the superior effectiveness of SEM in generating robust unlearnable examples.

Dataset	Model	Clean	EM	TAP	NTGA	REM	SEM
CIFAR-10	VGG-16	87.51	86.48	86.27	86.65	65.23	**44.37**
	RN-18	89.51	88.62	88.02	88.96	48.16	**31.92**
	RN-50	89.79	89.28	88.45	88.79	40.65	**28.89**
	DN-121	83.27	82.44	81.72	80.73	81.48	**77.85**
	WRN-34-10	91.21	90.05	90.23	89.95	48.39	**31.42**
CIFAR-100	VGG-16	57.14	56.94	55.24	55.81	**48.85**	57.11
	RN-18	63.43	64.17	62.39	62.44	27.10	**20.25**
	RN-50	66.93	66.43	64.44	64.91	26.03	**20.99**
	DN-121	53.73	53.52	**52.93**	52.40	54.48	55.36
	WRN-34-10	68.64	68.27	65.80	67.41	25.04	**18.90**

Table 3: Test accuracy (%) of different types of models on CIFAR-10/100 datasets under $\rho_a = 4/255$ and $\rho_u = 8/255$.

Different Architectures. To investigate the transferability of

Dataset	Filter	Clean	EM	TAP	NTGA	REM	SEM
CIFAR-10	Mean	84.25	34.87	82.53	40.26	28.60	**23.93**
	Median	87.04	31.86	85.10	30.87	27.36	**22.00**
	Gaussian	86.78	29.71	85.44	41.85	28.70	**21.77**
CIFAR-100	Mean	52.42	53.07	51.30	26.49	13.89	**8.81**
	Median	57.69	56.35	55.22	18.14	14.08	**9.02**
	Gaussian	56.64	56.49	55.19	29.05	13.74	**8.10**

Table 4: Test accuracy (%) of different types of models trained on CIFAR-10 and CIFAR-100 that are processed by different low-pass filters. The defensive perturbation radius ρ_u of every defensive noise is set as $16/255$.

different noises across neural network architectures, we chose ResNet-18 as the source model. We then examined the impact of its generated noise on various target models, including VGG-16, ResNet-50, DenseNet-121, and Wide ResNet-34-10. All transferability experiments were conducted using CIFAR-10 and CIFAR-100 datasets, as shown in Tab. 3. The results show that unlearnable examples generated by the SEM method are generally more effective across architecture settings than REM, indicating higher transferability.

Sensitivity Analysis

Resistance to Low-Pass Filtering. We next analyze how various defensive noises withstand low-pass filtering. This approach is motivated by the possibility that filtering the image could eliminate the added defensive noise. As per Fu et al. (2022), we employed three low-pass filters: Mean, Median, and Gaussian (Young and Van Vliet 1995), each with a 3×3 window size. We set the adversarial training perturbation radius to $2/255$. The results in Tab. 4 show that the test accuracy of the models trained on the protected data increases after applying low-pass filters, implying that such filtering partially degrades the added defensive noise. Nevertheless, SEM outperforms under both scenarios, with and without applying low-pass filtering in adversarial training.

Resistance to Early Stopping and Partial Poisoning. One might wonder how the early stopping technique influences the protection performance of our unlearnable examples. To address this, we conducted experiments on CIFAR-10 and CIFAR-100, varying the early-stopping patience steps. We designated 10% of the unlearnable examples for the validation set, with early-stopping patience S_{es} set to range from 3k to 20k steps. For examining partial poisoning, we varied the proportion of clean images in the validation set from 10% to 70%. Results are detailed in Tab. 5. Observations from full poisoning (clean ratio equal to 0%) show that setting S_{es} to 3000 increases test accuracy by 4%, implying minimal impact of early stopping on unlearnable examples. However, this effect is not substantial and necessitates searching the early-stopping patience hyper-parameter for mitigation. Increasing the clean ratio in the validation set further restores test accuracy to 57.48%, yet it remains significantly below the original level of 89.51%. These results underscore the resistance of our unlearnable examples to early stopping.

Before	Clean Ratio (%)	$S_{es} = 3000$	25000	10000	20000
31.92	0	**35.80**	29.34	31.30	30.87
	10	31.89	**47.22**	32.02	32.02
	30	31.89	**32.02**	31.67	32.02
	50	**57.48**	47.22	57.48	57.48
	70	**57.48**	47.22	57.48	31.89

Table 5: Effect of early stop with different patience steps S_{es}.

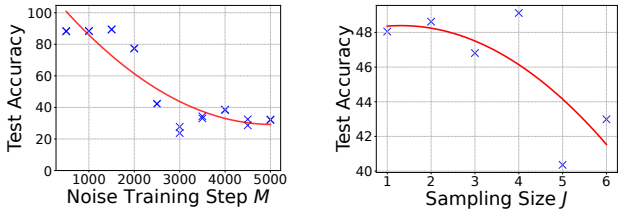


Figure 5: Effect of noise training steps and sampling step size on the testing accuracy of the trained model.

Effect of Training Step M and Sampling Size J. We evaluated the impact of the training step M and sampling size J during noise generation, considering M values from 500 to 5000 and J ranging from 1 to 6. Results, as illustrated in Fig. 5, indicate that an increase in training steps enhances performance. This improvement could be attributed to the noise generator learning more robust noise patterns over extended training steps. Regarding sampling size J, an increase in J was found to reduce the testing accuracy by nearly 8%, underscoring its significance in data protection.

Effect of Random Perturbation Radius. The radius of random perturbation represents a crucial hyper-parameter in our approach. To explore its correlation with protection performance, we conducted experiments on CIFAR-10 and CIFAR-100 using various radii of random perturbation and adversarial training. Results are detailed in Tab. 6. The results indicate that the best data protection performance is generally achieved when the random perturbation radius ρ_r is set equal to the adversarial training radius ρ_a, across varying adversarial training radii. Nevertheless, when these two radii are mismatched, the protection performance drops slightly, indicating the importance of knowledge about adversarial training radius for protection effectiveness.

Ablation Study We conducted an ablation study to understand the impact of various components in our method, specifically focusing on the adversarial noise used to update

Adv. Train. ρ_a	$\rho_r = 0$	$1/255$	$2/255$	$3/255$	$4/255$
0	15.87	11.41	**10.42**	16.83	10.52
$1/255$	16.59	**12.11**	15.80	17.92	12.16
$2/255$	64.24	13.44	**11.49**	19.13	21.68
$3/255$	89.18	28.6	21.88	**20.91**	23.43
$4/255$	89.30	67.52	35.63	62.09	**31.92**

Table 6: Effect of different radii of random perturbation.

δ^{u*}		θ^*		Test Acc. (%)	Acc. Increase (%)
δ^r	δ^a	δ^r	δ^a		
√			√	**30.03**	-
			√	37.91	+7.88
	√		√	46.72	+16.69
√				89.22	+59.19
				89.15	+59.13

Table 7: Ablation study of the proposed method.

Methods	$\rho_a = 0$	$1/255$	$2/255$	$3/255$	$4/255$
Clean Data	68.89	63.33	61.11	53.33	50.56
EM	**8.89**	18.06	18.61	16.39	19.72
TAP	9.72	20.74	26.48	32.22	36.39
SC	66.11	66.67	64.72	57.5	59.44
REM	10.89	7.89	11.67	12.22	11.48
SEM	10.56	**7.04**	**9.44**	**10.28**	**11.11**

Table 8: Evaluation on a real-world face recognition dataset.

the noise generator and the random noise used to update the defensive noise. Results, as shown in Tab. 7, reveal that the adversarial noise δ^a, used in updating the noise generator θ, is a critical factor for protection performance. Its removal results in an approximate 60% decrease in accuracy. Additionally, the elimination of random perturbation in updating the defensive noise leads to a reduction in protection performance by approximately 8%. Lastly, using purely adversarial noise to update the defensive noise results in deteriorated protection, indicating the adverse effects of these perturbations.

Case study: Face Recognition

To evaluate our proposed method's effectiveness on real-world face recognition, we conducted experiments using the Facescrub dataset (Ng and Winkler 2014). Specifically, we randomly selected ten classes, with each class comprising 120 images. We allocated 15% of the data as the testing set, resulting in 1020 images for training and 180 for testing. Results presented in Tab. 8 indicate that robust methods, namely REM and SEM, outperform other non-robust methods in terms of data protection under adversarial training. In particular, the test accuracy of models trained on SEM-protected data fell to around 9%, across various adversarial training radii, marking a significant drop in accuracy by approximately 40% to 50%. Additionally, we observed from Fig. 6 that robust methods result in more vividly protected images. For instance, SC creates mosaic-like images, while EM and TAP substantially reduce luminance, impairing facial recognition. Conversely, REM and SEM perturbations concentrate on edges, maintaining visual similarity.

Related Works

Data Poisoning. Data poisoning attacks aim to manipulate the performance of a machine learning model on clean examples by modifying the training examples. Recent research has shown the vulnerability of both DNNs (Muñoz-González et al. 2017) and classical machine learning methods, such as SVM (Biggio, Nelson, and Laskov 2012), to poisoning attacks (Shafahi et al. 2018). Recent advancements use gradient matching and meta-learning techniques to address the noise crafting problem (Geiping et al. 2020). Backdoor attacks are kinds of data poisoning that try to inject falsely labeled training samples with a concealed trigger (Gu, Dolan-Gavitt, and Garg 2017). Unlearnable examples, regarded as a clean-label and triggerless backdoor approach (Gan et al. 2022; Souri et al. 2022; Zeng et al. 2023; Xian et al. 2023; Li et al. 2022; Liu, Yang, and Mirzasoleiman 2022), perturb features to impair the model's generalization ability.

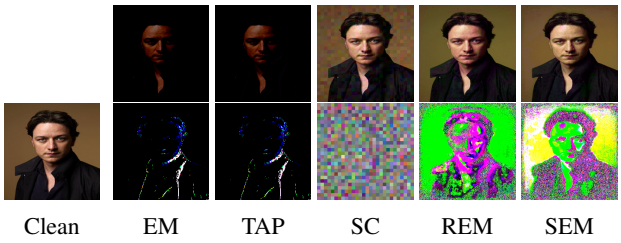


Figure 6: Visualization of different types of defensive noise and crafted unlearnable examples on the Facescrub subset.

Unlearnable Example. Huang et al. (2020) developed the concept of "unlearnable examples" by introducing error-minimizing noise, which led deep learning models to learn irrelevant features and impaired their generalization ability. Unlearnable examples have been adapted for data protection in various domains and applications (Liu et al. 2023c,d,b,a; Sun et al. 2022b; Zhang et al. 2023a; Li et al. 2023; He, Zha, and Katabi 2022; Zhao and Lao 2022; Salman et al. 2023; Guo et al. 2023; Ren et al. 2022; He et al. 2023; Chen et al. 2023; Wang, Le, and Lee 2023; Qin et al. 2023). Despite its effectiveness, such conferred unlearnability is found fragile to adversarial training and data transformations (Athalye et al. 2018). Tao et al. (2021) show that applying adversarial training with a suitable radius can counteract the added noise, thereby restoring model performance and making the data "learnable" again. Addressing this, Fu et al. (2022) proposed a min-min-max optimization approach to create more robust unlearnable examples by learning a noise generator with adversarial perturbations. However, REM faces challenges with inefficiency and suboptimal performance, mainly due to the instability of defensive noise.

Conclusion

In this work, we introduced Stable Error-Minimizing (SEM), a novel method for generating unlearnable examples that achieve better protection performance and generation efficiency under adversarial training. Our research uncovers an intriguing phenomenon: the effectiveness of unlearnable examples in protecting data is predominantly derived from the robustness of the surrogate model. Motivated by this, SEM employs a surrogate model to create robust error-minimizing noise against random perturbations. Extensive experimental evaluations confirm that SEM achieves SoTA performance.

Acknowledgements

This work was partially supported by the National Science Foundation under Grant CRII-2246067 and CCF-2319242, conducted at Samsung Research America and Lehigh University. We thank Jeff Heflin and Maryann DiEdwardo for their valuable early-stage feedback on the manuscript, Shaopeng Fu for his insights regarding the REM codebase, and Weiran Huang for engaging discussions and unique perspectives. We also thank AAAI reviewers for their constructive suggestions.

References

Athalye, A.; Engstrom, L.; Ilyas, A.; and Kwok, K. 2018. Synthesizing robust adversarial examples. In *International conference on machine learning*, 284–293. PMLR.

Biggio, B.; Nelson, B.; and Laskov, P. 2012. Poisoning attacks against support vector machines. In *ICML*.

Cao, Y.; Li, S.; Liu, Y.; Yan, Z.; Dai, Y.; Yu, P. S.; and Sun, L. 2023. A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. *arXiv preprint arXiv:2303.04226*.

Chen, R.; Jin, H.; Chen, J.; and Sun, L. 2023. EditShield: Protecting Unauthorized Image Editing by Instruction-guided Diffusion Models. *arXiv preprint arXiv:2311.12066*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Edwards, B. 2022. Artist finds private medical record photos in popular AI training data set.

Feng, J.; Cai, Q.-Z.; and Zhou, Z.-H. 2019. Learning to confuse: generating training time adversarial data with auto-encoder. In *NeurIPS*.

Fowl, L.; Chiang, P.-y.; Goldblum, M.; Geiping, J.; Bansal, A.; Czaja, W.; and Goldstein, T. 2021a. Preventing unauthorized use of proprietary data: Poisoning for secure dataset release. *arXiv preprint arXiv:2103.02683*.

Fowl, L.; Goldblum, M.; Chiang, P.-y.; Geiping, J.; Czaja, W.; and Goldstein, T. 2021b. Adversarial Examples Make Strong Poisons. In *NeurIPS*.

Fu, S.; He, F.; Liu, Y.; Shen, L.; and Tao, D. 2022. Robust Unlearnable Examples: Protecting Data Privacy Against Adversarial Learning. In *International Conference on Learning Representations*.

Gan, L.; Li, J.; Zhang, T.; Li, X.; Meng, Y.; Wu, F.; Yang, Y.; Guo, S.; and Fan, C. 2022. Triggerless Backdoor Attack for NLP Tasks with Clean Labels. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2942–2952.

Geiping, J.; Fowl, L. H.; Huang, W. R.; Czaja, W.; Taylor, G.; Moeller, M.; and Goldstein, T. 2020. Witches' Brew: Industrial Scale Data Poisoning via Gradient Matching. In *International Conference on Learning Representations*.

Gu, T.; Dolan-Gavitt, B.; and Garg, S. 2017. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*.

Guo, J.; Li, Y.; Wang, L.; Xia, S.-T.; Huang, H.; Liu, C.; and Li, B. 2023. Domain Watermark: Effective and Harmless Dataset Copyright Protection is Closed at Hand. In *Thirty-seventh Conference on Neural Information Processing Systems*.

He, H.; Zha, K.; and Katabi, D. 2022. Indiscriminate Poisoning Attacks on Unsupervised Contrastive Learning. In *The Eleventh International Conference on Learning Representations*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

He, P.; Xu, H.; Ren, J.; Cui, Y.; Liu, H.; Aggarwal, C. C.; and Tang, J. 2023. Sharpness-Aware Data Poisoning Attack. *arXiv preprint arXiv:2305.14851*.

Hill, K.; and Krolik, A. 2019. How photos of your kids are powering surveillance technology. *The New York Times*.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *CVPR*.

Huang, H.; Ma, X.; Erfani, S. M.; Bailey, J.; and Wang, Y. 2020. Unlearnable Examples: Making Personal Data Unexploitable. In *International Conference on Learning Representations*.

Huang, H.; Wang, Y.; Erfani, S. M.; Gu, Q.; Bailey, J.; and Ma, X. 2021. Exploring Architectural Ingredients of Adversarially Robust Deep Neural Networks. In *NeurIPS*.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Krizhevsky, A.; et al. 2009. Learning multiple layers of features from tiny images.

Li, Y.; Bai, Y.; Jiang, Y.; Yang, Y.; Xia, S.-T.; and Li, B. 2022. Untargeted backdoor watermark: Towards harmless and stealthy dataset copyright protection. *Advances in Neural Information Processing Systems*, 35: 13238–13250.

Li, Z.; Yu, N.; Salem, A.; Backes, M.; Fritz, M.; and Zhang, Y. 2023. {UnGANable}: Defending Against {GAN-based} Face Manipulation. In *32nd USENIX Security Symposium (USENIX Security 23)*, 7213–7230.

Liu, T. Y.; Yang, Y.; and Mirzasoleiman, B. 2022. Friendly noise against adversarial noise: a powerful defense against data poisoning attack. *Advances in Neural Information Processing Systems*, 35: 11947–11959.

Liu, Y.; Fan, C.; Chen, X.; Zhou, P.; and Sun, L. 2023a. GraphCloak: Safeguarding Task-specific Knowledge within Graph-structured Data from Unauthorized Exploitation. *arXiv preprint arXiv:2310.07100*.

Liu, Y.; Fan, C.; Dai, Y.; Chen, X.; Zhou, P.; and Sun, L. 2023b. Toward Robust Imperceptible Perturbation against Unauthorized Text-to-image Diffusion-based Synthesis. *arXiv preprint arXiv:2311.13127*.

Liu, Y.; Fan, C.; Zhou, P.; and Sun, L. 2023c. Unlearnable Graph: Protecting Graphs from Unauthorized Exploitation. *arXiv preprint arXiv:2303.02568*.

Liu, Y.; Ye, H.; Zhang, K.; and Sun, L. 2023d. Securing Biomedical Images from Unauthorized Training with Anti-Learning Perturbation. *arXiv preprint arXiv:2303.02559*.

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

Muñoz-González, L.; Biggio, B.; Demontis, A.; Paudice, A.; Wongrassamee, V.; Lupu, E. C.; and Roli, F. 2017. Towards poisoning of deep learning algorithms with back-gradient optimization. In *ACM Workshop on Artificial Intelligence and Security*.

Ng, H.-W.; and Winkler, S. 2014. A data-driven approach to cleaning large face datasets. In *2014 IEEE International Conference on Image Processing (ICIP)*, 343–347.

Qin, T.; Gao, X.; Zhao, J.; Ye, K.; and Xu, C.-Z. 2023. APBench: A unified benchmark for availability poisoning attacks and defenses. *arXiv preprint arXiv:2308.03258*.

Ren, J.; Xu, H.; Wan, Y.; Ma, X.; Sun, L.; and Tang, J. 2022. Transferable unlearnable examples. *arXiv preprint arXiv:2210.10114*.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3): 211–252.

Salman, H.; Khaddaj, A.; Leclerc, G.; Ilyas, A.; and Madry, A. 2023. Raising the cost of malicious ai-powered image editing. *arXiv preprint arXiv:2302.06588*.

Schuhmann, C.; Vencu, R.; Beaumont, R.; Kaczmarczyk, R.; Mullis, C.; Katta, A.; Coombes, T.; Jitsev, J.; and Komatsuzaki, A. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Shafahi, A.; Huang, W. R.; Najibi, M.; Suciu, O.; Studer, C.; Dumitras, T.; and Goldstein, T. 2018. Poison frogs! targeted clean-label poisoning attacks on neural networks. In *NeurIPS*.

Simonyan, K.; and Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition. In *ICLR*.

Souri, H.; Fowl, L.; Chellappa, R.; Goldblum, M.; and Goldstein, T. 2022. Sleeper agent: Scalable hidden trigger backdoors for neural networks trained from scratch. *Advances in Neural Information Processing Systems*, 35: 19165–19178.

Sun, L.; Dou, Y.; Yang, C.; Zhang, K.; Wang, J.; Philip, S. Y.; He, L.; and Li, B. 2022a. Adversarial attack and defense on graph data: A survey. *IEEE Transactions on Knowledge and Data Engineering*.

Sun, L.; Huang, Y.; Wang, H.; Wu, S.; Zhang, Q.; Gao, C.; Huang, Y.; Lyu, W.; Zhang, Y.; Li, X.; et al. 2024. TrustLLM: Trustworthiness in Large Language Models. *arXiv preprint arXiv:2401.05561*.

Sun, Z.; Du, X.; Song, F.; Ni, M.; and Li, L. 2022b. Coprotector: Protect open-source code against unauthorized training usage with data poisoning. In *Proceedings of the ACM Web Conference 2022*, 652–660.

Tao, L.; Feng, L.; Yi, J.; Huang, S.-J.; and Chen, S. 2021. Better safe than sorry: Preventing delusive adversaries with adversarial training. *Advances in Neural Information Processing Systems*, 34: 16209–16225.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7794–7803.

Wang, Z.; Le, T.; and Lee, D. 2023. UPTON: Preventing Authorship Leakage from Public Text Release via Data Poisoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 11952–11965.

Xian, X.; Wang, G.; Srinivasa, J.; Kundu, A.; Bi, X.; Hong, M.; and Ding, J. 2023. Understanding Backdoor Attacks through the Adaptability Hypothesis. In *Proc. International Conference on Machine Learning*.

Young, I. T.; and Van Vliet, L. J. 1995. Recursive implementation of the Gaussian filter. *Signal processing*, 44(2): 139–151.

Yu, D.; Zhang, H.; Chen, W.; Yin, J.; and Liu, T.-Y. 2022. Availability Attacks Create Shortcuts. arXiv:2111.00898.

Yuan, C.-H.; and Wu, S.-H. 2021. Neural Tangent Generalization Attacks. In *ICML*.

Zagoruyko, S.; and Komodakis, N. 2016. Wide Residual Networks. In *Procedings of the British Machine Vision Conference 2016*. British Machine Vision Association.

Zeng, Y.; Pan, M.; Just, H. A.; Lyu, L.; Qiu, M.; and Jia, R. 2023. Narcissus: A practical clean-label backdoor attack with limited information. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, 771–785.

Zhang, J.; Ma, X.; Yi, Q.; Sang, J.; Jiang, Y.-G.; Wang, Y.; and Xu, C. 2023a. Unlearnable clusters: Towards label-agnostic unlearnable examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3984–3993.

Zhang, K.; Yu, J.; Yan, Z.; Liu, Y.; Adhikarla, E.; Fu, S.; Chen, X.; Chen, C.; Zhou, Y.; Li, X.; et al. 2023b. BiomedGPT: A Unified and Generalist Biomedical Generative Pre-trained Transformer for Vision, Language, and Multimodal Tasks. *arXiv preprint arXiv:2305.17100*.

Zhao, B.; and Lao, Y. 2022. CLPA: Clean-label poisoning availability attacks using generative adversarial nets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 9162–9170.

Zhou, C.; Li, Q.; Li, C.; Yu, J.; Liu, Y.; Wang, G.; Zhang, K.; Ji, C.; Yan, Q.; He, L.; et al. 2023. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *arXiv preprint arXiv:2302.09419*.