One for All: A Universal Generator for Concept Unlearnability via Multi-Modal Alignment

Chaochao Chen [1] **Jiaming Zhang** [1] **Yuyuan Li** [2 1] **Zhongxuan Han** [1]

Abstract

The abundance of free internet data offers unprecedented opportunities for researchers and developers, but it also poses privacy risks. Utilizing data without explicit consent raises critical challenges in protecting personal information. Unlearnable examples have emerged as a feasible protection approach, which renders the data unlearnable, i.e., useless to third parties, by injecting imperceptible perturbations. However, these perturbations only exhibit unlearnable effects on either a particular dataset or label-consistent scenarios, thereby lacking broad applicability. To address both issues concurrently, we propose a universal perturbation generator that harnesses data with concept unlearnability, thereby broadening the scope of unlearnability beyond specific datasets or labels. Specifically, we leverage multi-modal pre-trained models to establish a connection between the data concepts in a shared embedding space. This connection enables the information transformation from image data to text concepts. Consequently, we can align the text embedding using concept-wise discriminant loss, and render the data unlearnable. Extensive experiments conducted on real-world datasets demonstrate the concept unlearnability, i.e., cross-dataset transferability and label-agnostic utility, of our proposed unlearnable examples, and their robustness against attacks.

1. Introduction

In the past few decades, the abundance of free data available on the internet has revolutionized the field of deep learning (Deng et al., 2009; Brown et al., 2020; Han et al., 2024; Liu et al., 2023b; 2021a; 2023a). However, this easy access to vast amounts of data has raised concerns regarding privacy and data security. As data is increasingly used without explicit consent from owners, protecting personal information becomes paramount. One emerging approach to address this concern is generating unlearnable examples, which ensures that an unauthorized third-party model cannot retain any useful information about the protected data (Huang et al., 2021; Fu et al., 2022; Ren et al., 2023). Unlearnable examples are generated by injecting imperceptible perturbations, making the data samples usable for humans but unlearnable for models. The perturbation is typically error-minimizing noise that creates *shortcut* features which models can readily capture, thereby disregarding the data's original features (Yu et al., 2022). Consequently, injecting this perturbation renders the data sample unlearnable.

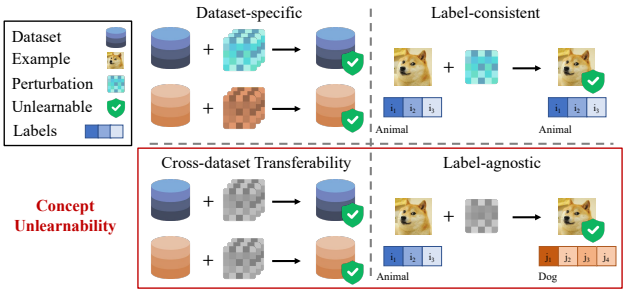


Figure 1. An illustration of the capability boundary of existing unlearnable examples and the desired capability of concept unlearnability, which exhibits both cross-dataset transferability (applicable to a new dataset without retraining) and label-agnostic utility (applicable to data with different labels).

However, the practical applicability of unlearnable examples is limited. As shown in Figure 1, existing methods either i) suffer from a lack of cross-dataset transferability (Ren et al., 2023) or ii) are confounded by label-agnostic scenarios (Zhang et al., 2023). On the one hand, the majority of perturbations are designed specifically for a target dataset, resulting in a notable decrease in their unlearnability when applied to other non-target datasets. The absence of cross-dataset transferability necessitates generating perturbations, i.e., retraining, for each dataset, posing challenges in real-

[1]Colledge of Computer Science and Technology, Zhejiang University, Hangzhou, China [2]School of Communication Engineering, Hangzhou Dianzi University, Hangzhou, China. Correspondence to: Yuyuan Li <y2li@hdu.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

world scenarios with dynamic and even streaming data. On the other hand, existing methods predominantly assume a label-consistent setting where both the data owners and third parties assign the same labels to the data. In reality, this assumption is highly improbable, as third parties commonly employ different labeling schemes compared to the data owners, thus creating label-agnostic scenarios.

Recently, several methods have been proposed to address the aforementioned issues separately (Ren et al., 2023; Zhang et al., 2023). However, these methods fall short of concurrently addressing both issues. There are three primary challenges. Firstly, the transferable approach conducts cross-class perturbation interpolation to achieve cross-dataset transferability. However, this approach requires determining the class label of perturbations, making it inapplicable in label-agnostic scenarios. Secondly, the label-agnostic approach relies on pseudo-labels obtained from clustering to replace the data owner's labels. Nevertheless, the clustering results are unstable, e.g., sensitive to initialization and the number of clusters. Last but not least, the efficiency of both approaches is not satisfactory. The transferable approach follows the time-consuming gradient-based method, while the label-agnostic approach requires training a separate generator for each cluster.

To concurrently address both issues, we propose a universal perturbation generator (One-for-All, 14A) capable of producing unlearnable examples that enjoy *concept unlearnability*, including both cross-dataset transferability and label-agnostic utility. In other words, our universal generator is designed to render concepts unlearnable beyond specific datasets or labels. Thus, it exhibits zero-shot generation ability. Once trained, this generator enables input from any sample, providing significant efficiency advantages over gradient-based methods and multi-generator methods. To automatically assign unique labels to all data samples, we introduce a multi-modal pre-trained model, e.g., CLIP (Radford et al., 2021), that connects input image data with text labels in a shared embedding space, naturally overcoming the label-agnostic issue. Additionally, these automatically assigned labels eliminate the need for clustering. To render data unlearnable, we determine the similar and opposite concepts as the alignment target. Specifically, we optimize the alignment process through an efficiency-improved concept-wise discriminant loss, which exaggerates the intra-concept distance while diminishing the inter-concept distance.

The main contributions of this paper are summarized as follows:

- We propose a universal perturbation generator to produce data with concept unlearnability, concurrently addressing two key issues of unlearnable examples, i.e., cross-dataset transferability and label-agnostic, encountered in real-world scenarios.
- Our proposed generator is universally applicable for all data samples, as it connects the concept of data through multi-modal models. This concept unlearnability transcends beyond specific datasets or labels, eliminating the need to train individual generators for each dataset or class.
- To achieve concept unlearnability, we align each sample embedding into target embeddings using efficiency-improved concept-wise discriminant loss, thereby misleading the adversary model.
- Extensive experiments on real-world datasets illustrate the concept unlearnability of our proposed unlearnable examples and the robustness of attacks, demonstrating their practical effectiveness.

2. Preliminary

The unlearnable example can be interpreted as a type of poisoning attack that aims to completely disrupt the performance of a model trained on such poisoned data. This attack differs from two well-known types of poisoning attacks, i.e., backdoor attack (Saha et al., 2020; Nguyen & Tran, 2020; Li et al., 2021) and adversarial attack (Dong et al., 2018; Dai et al., 2018; Yuan et al., 2021). The backdoor attack involves injecting a backdoor trigger into the model during training, causing it to perform poorly specifically on triggered data while maintaining normal performance on clean data. On the other hand, the adversarial attack aims to inject imperceptible perturbations into the inputs during testing, leading to poor performance.

Unlearnable examples are typically generated by the data owner, i.e., defender, and any third party attempting to train a model using these examples is considered an adversary. Formally, given a dataset $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, the data owner's goal is to render this original data into unlearnable examples $D' = \{\boldsymbol{x}'_i, y_i\}_{i=1}^n$ where $\boldsymbol{x}' = \boldsymbol{x} + \boldsymbol{\delta}$, which will be exposed to adversaries.

Since the condition of the adversary is unknown to the defender, unlearnable examples will encounter the following two challenges in practice:

- **Cross-dataset Transferability** refers to the perturbation $\boldsymbol{\delta}$ being applicable to multiple datasets, not just the target dataset used for training.
- **Label-agnostic** describes a scenario where the adversary assigns different labels to the data compared to the defender, e.g., $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ and $D' = \{(\boldsymbol{x}'_i, y'_i)\}_{i=1}^n$. This is both common and crucial, as the adversary may apply additional data pre-processing steps in practice.

In this paper, we introduce the idea of concept unlearnability to broaden the scope of unlearnability, and enhance its applicability by addressing the aforementioned challenges.

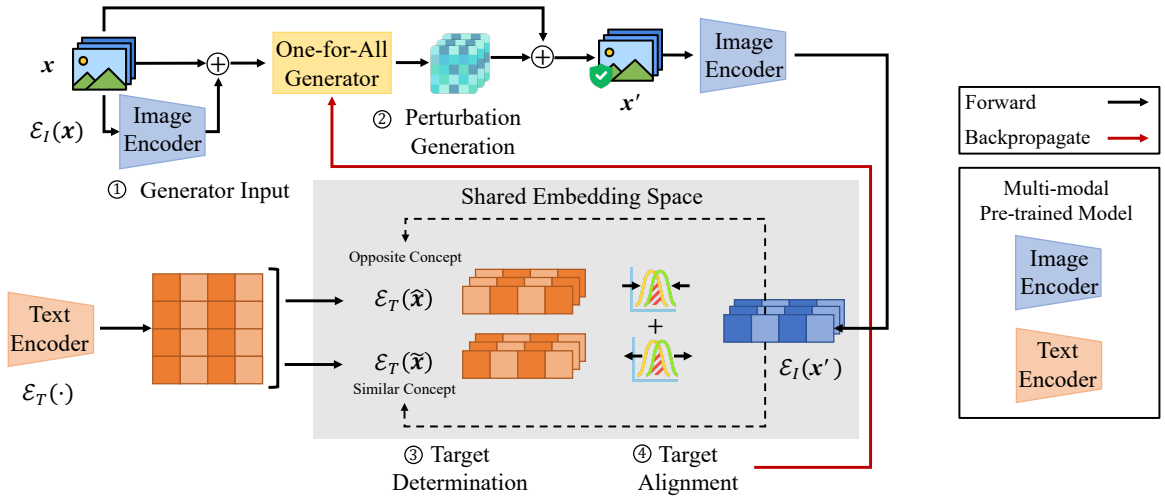


Figure 2. Training scheme of One-for-All (14A) generator. (1) 14A generator takes as input the image and its embedding; (2) 14A generates the corresponding perturbation and concatenates it with the image to produce an unlearnable example; (3) We determine the target text embedding in the shared embedding space; (4) To render the data unlearnable, we align the image embedding with the target text embedding using concept-wise discriminant loss.

3. One-for-All Generator

In this section, we first emphasize the motivation behind the concept of unlearnability, and then present the training and inference details of the 14A generator, which is used to produce the unlearnable examples, following the specific order of its workflow.

3.1. Motivation

Unlearnable examples offer a promising solution to protect data from unauthorized use, but they encounter challenges in real-world applications. In this paper, we highlight two significant challenges faced by existing unlearnable examples, i.e., the absence of cross-dataset transferability and label-agnostic utility. Accordingly, our goal is to produce unlearnable examples that possess both of these abilities concurrently.

Existing methods typically involve injecting imperceptible perturbations into the original data. Note that these perturbations are designed to be imperceptible to ensure that the data remains useless to models while remaining interpretable for humans. However, this approach naturally suffers from the aforementioned challenges because the perturbations serve as shortcuts to trick the model into learning away from the original label/pseudo-label and towards a target one (Yu et al., 2022).

In our approach, we aim to break away from the confines of labels and instead focus on transferring the underlying *concept* of the data. This underlying concept refers to the description that is associated with the data. By transferring the concept, we can broaden the scope of unlearnability beyond a specific dataset or set of labels. For example, as depicted in Figure 1, an image of a dog can be included in datasets like Animal Picture or Chatting Sticker based on the category of the concept. It can also be labeled as either 'dog' or 'animal' depending on the conceptual scope. However, the underlying concept goes beyond these datasets or labels.

3.2. Perturbation Generator

Taking inspiration from (Zhang et al., 2023), we adopt a generator-based approach to generate perturbations, avoiding the use of labels in the gradient-based bi-level optimization approach (Feng et al., 2019; Tao et al., 2021; Fowl et al., 2021; Huang et al., 2021; Liu et al., 2021b; Fu et al., 2022). However, in previous work (Zhang et al., 2023), a generator is constructed for each cluster, representing a specific class of data with an assigned pseudo-label. Additionally, these generators are primarily effective for a specific target dataset. These limitations pose challenges in achieving concept unlearnability, i.e., cross-dataset transferability and label-agnostic utility.

Driven by the goal of concept unlearnability and acknowledging the well-established efficacy of deep learning in fitting complex patterns, we propose a bold endeavor: the construction of a universal generator capable of generating perturbations for diverse concepts, transcending the confines imposed by specific datasets or labels.

In previous work (Zhang et al., 2023), the generator takes uniform noise u and the target data $\boldsymbol{x}$ as input and generates the perturbation, i.e., $\boldsymbol{\delta} = \mathcal{G}(u, \boldsymbol{x}, \boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is optimiz-

able parameters of the generator. However, this design is only suitable for a one-generator one-cluster scenario. To construct a universal generator, we propose to incorporate concept information as input, which serves as a guide for the generator in place of uniform noise. This concept information is expected to operate beyond the original label, as our goal is to achieve unlearnability that transcends the confines imposed by specific datasets or labels. To obtain such concept information, we introduce multi-modal embedding, which connects the visual and textual modalities of the data, enabling us to link the image with a wealth of textual concepts. Formally, the universal generator computes as

$$\boldsymbol{\delta} = \mathcal{G}(\mathcal{E}_I(\boldsymbol{x}), \boldsymbol{x}, \boldsymbol{\theta}), \tag{1}$$

where $\mathcal{E}_I(\cdot)$ denotes the image encoder. Following (Zhang et al., 2023), we employ an encoder-decoder structure (Poursaeed et al., 2018) for the generator. This architecture demonstrates its effectiveness in our empirical study. In Section 4.1, we will provide details regarding the model structure of the generator and the input location of image embedding in the generator.

3.3. Multi-modal Embedding

The universal generator represents a bold endeavor, and its implementation relies on the utilization of multi-modal embedding. As mentioned earlier, by modeling visual and textual data in a shared embedding space, we can establish a connection between images and textual concepts. This unique capability enables the generator to transcend the limitations imposed by datasets or labels, fostering unlearnability that reaches beyond traditional boundaries. More importantly, the universal generator can input from any samples to produce the unlearnable ones without the need for retraining. This *one-trained multiple-used* generator opens up the potential for making anything unlearnable.

Furthermore, multi-modal embedding eliminates the need for a surrogate model. In the context of unlearnable examples, a surrogate model refers to the simulated adversary model used by adversaries, given that data owners may not have knowledge of potential adversaries. The surrogate model operates akin to a discriminator in GAN, aiding in optimizing the quality of generated perturbations. Previously, in (Zhang et al., 2023), a pre-trained multi-modal model was utilized to extract visual embeddings. However, the multi-modal property was not leveraged in that work. In this work, we propose to eliminate the surrogate model altogether and directly align the visual embeddings with the textual embeddings to create a shortcut, thereby rendering the data unlearnable. This approach fully explores the multi-modal property and simplifies the process of generating perturbations. The details of the embedding alignment process will be introduced in the upcoming section.

Specifically, we utilize CLIP (Radford et al., 2021), a multi-modal pre-trained model, to obtain the embeddings for our approach. This choice is motivated by two key reasons. Firstly, CLIP is pre-trained with textual descriptions rather than one-hot labels, mitigating overfitting to specific class labels and enhancing the extractability of underlying concepts. Consequently, this design makes CLIP a great concept-extraction model for describing the main content of images. Secondly, previous work (Zhang et al., 2023) has demonstrated the effectiveness of CLIP in modeling unlearnable examples. This prior research supports the decision to employ CLIP as a reliable multi-modal modeling tool for our current work. For simplicity, we use $\mathcal{E}_I(\cdot)$ and $\mathcal{E}_T(\cdot)$ to denote the image and text encoders of CLIP respectively.

3.4. Target Embedding Alignment

The key to rendering data unlearnable lies in creating an imperceptible shortcut within the data. This shortcut misleads the adversary model into learning false information about the data. To achieve this objective, the training phase of our proposed 14A generator follows three main steps. Firstly (step 3 in Figure 2), we determine the target embedding, i.e., concept, that we aim to deceive the model into learning. Secondly, we explore both opposite and similar concepts to enhance the robustness and transferability of the generated perturbations. Lastly (step 4 in Figure 2), we align the derived embeddings with the target concept using a derived loss function. During the inference phase, 14A generator stops updating its parameters, and only steps 1 and 2 are executed. Thus, the application of 14A generator is simple and straightforward.

3.4.1. Target Determination

In this step, we determine the target embeddings from the text encoder. As shown in Figure 2, following steps 1 and 2, we obtain unlearnable examples through pixel-wise addition of the original data and the perturbation:

$$\boldsymbol{x}' = \boldsymbol{x} + \mathcal{G}(\mathcal{E}_I(\boldsymbol{x}), \boldsymbol{x}, \boldsymbol{\theta}). \tag{2}$$

Subsequently, we determine the target embedding based on the embedding generated from the image encoder, i.e., $\mathcal{E}_I(\boldsymbol{x}')$.

From all the available text embeddings provided by CLIP, we select opposite (most dissimilar) and similar concepts according to their distances from $\mathcal{E}_I(\boldsymbol{x}')$ in the shared embedding space. A shorter distance indicates a greater similarity.

3.4.2. Target Alignment

In this step, we align the image embedding $\mathcal{E}_I(\boldsymbol{x}')$ to target text embeddings, thereby creating a shortcut to mislead the adversary model.

Inspired by clustering algorithms (Jia et al., 2014; Ahmed et al., 2020), Ren et al. identify two factors crucial in creating this misleading shortcut, i.e., larger inter-class distance and smaller intra-class distance. To achieve this, Ren et al. propose a class-wise discriminant loss, formulated as:

$$\mathcal{L}_1 = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{M-1}\sum_{j\neq i}^{M}\left(\frac{\pi_i+\pi_j}{d_{i,j}}\right), \tag{3}$$

where π_i and $d_{i,j}$ denote the intra-class and inter-class distance, and M is the number of classes.

However, this approach is limited by the availability of data labels. To overcome this limitation, we adopt a concept-based approach, and take an opposite optimization direction, i.e., exaggerating the intra-concept distance while diminishing the inter-concept distance. Formally, the concept-wise discriminant loss computes as:

$$\mathcal{L}_2 = d(\mathcal{E}_I(\boldsymbol{x}'), \mathcal{E}_T(\hat{\boldsymbol{x}})) - \lambda\frac{\sum_k d(\mathcal{E}_I(\boldsymbol{x}'), \mathcal{E}_T(\tilde{\boldsymbol{x}}))}{k}, \tag{4}$$

where $d(\cdot)$ is the distance measurement, and λ is the trade-off coefficient. Intuitively, concept-wise loss operates on two principles. On the one hand, we push the data away from its similar concepts, creating a clear distinction between the original concept $\mathcal{E}_I(\boldsymbol{x}')$ and other similar ones $\mathcal{E}_T(\tilde{\boldsymbol{x}})$. On the other hand, we guide the data towards the opposite concept, misleading the adversary model into learning irrelevant concept $\mathcal{E}_T(\hat{\boldsymbol{x}})$. Therefore, a lower concept-wise discriminant indicates a larger intra-concept distance and a smaller inter-concept distance. This characteristic enhances the ability to trick the adversary model, rendering the data unlearnable.

In contrast to the class-wise loss, we make two key modifications to concept-wise loss. Firstly, we choose the addition form instead of the multiplication form. This decision is driven by the fact that concepts are widely dispersed across the embedding space, resulting in different scales for intra-concept and inter-concept distances. By using the addition form, we avoid the issue of the intra-concept distance (denominator) becoming infinitely large, which would lead to an all-black perturbation. Secondly, we select only the top associated concepts instead of traversing through all labels. This modification helps to reduce the computational complexity while still achieving effective results.

4. Experiments

In this paper, we conduct a comprehensive evaluation of our proposed unlearnable examples on multiple real-world datasets. To validate the concept unlearnability of our proposed method, we explore their cross-dataset transferability and label-agnostic utility (Section 4.2). Furthermore, we examine the performance of the proposed unlearnable examples against attacks that attempt to learn from these examples (Section 4.3). We also investigate the limitations of our method (Section 4.4) and the effect of hyper-parameters (Section 4.5).

4.1. Experimental Settings

Datasets. To comprehensively evaluate the concept unlearnability, we conduct experiments on a diverse range of datasets. These include widely used benchmarks, i.e., CIFAR10, CIFAR100 (Krizhevsky et al., 2009), and ImageNet (Russakovsky et al., 2015), which provide a broad representation of different domains. Additionally, we explore specific domains by including Pets (Parkhi et al., 2012), Flowers (Nilsback & Zisserman, 2008), Cars (Krause et al., 2013), Food (Bossard et al., 2014), and Sun (Xiao et al., 2010). Together, these datasets also facilitate the evaluation of cross-dataset and label-agnostic scenarios.

Backbones. We evaluate the unlearnability on several representative and time-validated backbones, i.e., ResNet18 (He et al., 2016), MobileNetV2 (Sandler et al., 2018), and ShuffleNetV2 (Ma et al., 2018). To enhance generality, we also evaluate large-scale backbones, i.e., ResNet50 (He et al., 2016), VGG16 (Simonyan & Zisserman, 2015), VIT (Dosovitskiy et al., 2021), in a cross-dataset and label-agnostic scenario. The backbones serve as the adversary model, which means they are trained using unlearnable examples and tested on clean data. A low performance on the clean data demonstrates effective unlearnability.

Compared Baselines. We compare our proposed method (14A) with the original data and several state-of-the-art baselines that have demonstrated potential in tackling cross-dataset and label-agnostic scenarios.

- **CLEAN**: The original data.
- **CP** (He et al., 2023): Contrastive Poisoning (CP) biases the model toward recognizing poisoning patterns instead of real features to achieve unlearnability for unsupervised learning. It also shows promise in cross-dataset and label-agnostic supervised learning scenarios.
- **TUE** (Ren et al., 2023): Transferable Unlearnable Examples (TUE) transfer the unlearnability to other training settings and datasets by interpolating linearly separable perturbations.
- **LaUE** (Zhang et al., 2023): Label-agnostic Unlearnable Examples (LaUE) are generated using cluster-wise perturbations, which eliminates the requirement for explicit labels. We use CLIP image embedding to determine the cluster of input images, making it applicable for cross-dataset scenarios.

Note that CP and TUE are gradient-based approaches that generate statistical perturbations. In contrast, LaUE and our proposed 14A are generator-based approaches that generate dynamic perturbations for each input image.

Implementation. To fulfill our requirements, we improve the structure of the autoencoder generator in the existing generator-based approach (Zhang et al., 2023). This improvement involves increasing the depth of the network. Additionally, we concatenate the encoded embedding with the image embedding generated by CLIP to provide concept information. To ensure imperceptibility, we consider L_∞-norm restriction in this paper. Specifically, we enforce $\|\boldsymbol{\delta}\|_\infty < \epsilon = 8/255$, ensuring that the generated perturbation remains within acceptable limits. To train our generator, we employ Adam optimizer with an initial learning rate of 0.0001. The trade-off coefficient is set as $|l_1/l_2|$ where l_1 and l_2 denote the first and second terms of Equation (4). The training process spans 200 epochs on the entire ImageNet dataset. We use cosine similarity as the distance measurement $d(\cdot)$. We run all experiments on an Ubuntu 20.04 LTS system server with 256GB RAM, and four NVIDIA GeForce RTX 3090 GPU.

Cross-dataset Settings. We generate perturbations on the ImageNet, and evaluate their unlearnability on other domain-specific datasets, i.e., Pets, Flowers, Cars, Food, and Sun, thereby creating a cross-dataset scenario. In the case of gradient-based approaches (CP and TUE), their perturbations are interpolated to generate unlearnable examples for other datasets during evaluation.

Label-agnostic Settings. For each dataset, we leverage the text encoder of CLIP to extract the embedding of the original labels. Subsequently, we employ the k-means clustering algorithm (Ahmed et al., 2020) to derive 20 clusters, treating all original labels within the same cluster as a single new label. Our objective is to create a label-inconsistent scenario without compromising the latent concept of the data. This is also aligned with the intentions of adversaries who aim to utilize this data. Note that we conducted an empirical study (Section 4.6) to determine the optimal number of clusters for our tested dataset. As a result, while the clustered labels may exhibit some conceptual overlap with the original labels, they are notably inconsistent with the original labels, thereby creating a label-inconsistent scenario.

4.2. Concept Unlearnability

We evaluate concept unlearnability in terms of cross-dataset transferability and label-agnostic utility. We begin by investigating the cross-dataset scenario and subsequently extend to the scenario of both cross-dataset and label-agnostic, as our goal is to fulfill both capabilities. We present the details of our investigation below.

4.2.1. Cross-dataset

We present the results of the cross-dataset scenario in Table 1, where all compared methods are trained on ImageNet and tested on other domain-specific datasets.

Performance on Different Backbones. In most cases, 14A consistently outperforms baselines across various backbones, demonstrating its robustness and resilience to different adversary models.

Performance on Different Datasets. Except Cars, 14A consistently outperforms baselines across different datasets. There are two primary reasons: i) the relatively small number of car images in ImageNet hinders the effective training of the generator specifically for car images; and ii) the slight variation in image sizes within Cars necessitates the inclusion of a resizing module, which in turn affects our overall performance. For a more detailed discussion on this matter, please refer to Section 4.4.

4.2.2. Cross-dataset and Label-agnostic

Based on the cross-dataset scenario, we further investigate the cross-dataset and label-agnostic scenario. The results are presented in Table 2. Compared with the cross-dataset scenario, this scenario impairs the performance of both adversaries and protectors. Specifically, CLEAN exhibits a decrease in testing accuracy, suggesting that it is more challenging for adversaries to learn from the data. On the other hand, compared methods show an increase in testing accuracy, indicating the difficulty in protecting the data.

Performance on Different Backbones. In this scenario, we also evaluate on large-scale backbones, i.e., ResNet50, VGG16, and VIT. We observe a significant advantage of 14A over the baselines when tested on these larger models, with an average improvement of 41.91% compared to 27.01% for smaller models. We hypothesize that the superior feature extraction capabilities of larger models make them more vulnerable to the shortcut perturbations generated by our method.

Performance on Different Datasets. Our observation is consistent with the cross-dataset scenario.

4.3. Performance towards Attack

Although using unlearnable examples is a feasible approach to protect personal data, it is important to acknowledge that adversaries have developed various attacking methods to overcome this defense mechanism, i.e., learning from unlearnable examples. Under the scenario of cross-dataset and

Table 1. Testing accuracy (%) ↓ of different backbones, i.e., adversary models, trained on different unlearnable datasets generated from other datasets (cross-dataset scenario). The top results are highlighted in **bold**.

Methods	ResNet18					MobileNetV2					ShuffleNetV2				
	Pets	Flowers	Cars	Food	Sun	Pets	Flowers	Cars	Food	Sun	Pets	Flowers	Cars	Food	Sun
CLEAN	43.68	84.47	40.43	65.45	50.40	45.84	90.22	59.95	67.88	57.22	42.27	83.61	45.16	69.29	53.88
CP	15.34	58.19	**5.85**	26.49	25.14	8.80	47.43	2.98	14.97	28.14	11.42	54.64	4.47	26.49	23.82
TUE	15.99	57.09	6.14	25.84	23.85	8.04	47.67	3.70	16.99	25.33	10.98	54.52	4.82	26.89	22.14
LaUE	13.79	62.34	7.30	47.84	34.2	14.41	81.05	**2.73**	51.10	44.68	6.67	66.25	**4.14**	41.37	35.67
14A(Ours)	**3.10**	**15.15**	6.69	**17.18**	**16.01**	**4.03**	**19.07**	10.25	**12.21**	**20.28**	**3.10**	**22.00**	4.31	**10.94**	**15.31**

Table 2. Testing accuracy (%) ↓ of different backbones, i.e., adversary models, trained on different unlearnable datasets generated from other datasets with inconsistent labels (cross-dataset and label-agnostic scenario). The top results are highlighted in **bold**.

Methods	ResNet18					MobileNetV2					ShuffleNetV2				
	Pets	Flowers	Cars	Food	Sun	Pets	Flowers	Cars	Food	Sun	Pets	Flowers	Cars	Food	Sun
CLEAN	42.16	75.67	56.95	62.96	56.14	40.58	83.49	66.89	61.39	65.22	46.75	76.89	58.21	67.81	62.62
CP	23.54	55.86	20.68	33.60	39.70	16.65	49.14	13.58	21.81	43.30	21.80	50.97	16.42	33.35	39.20
TUE	23.27	53.30	20.14	33.01	39.63	15.12	39.48	13.44	31.88	42.62	18.23	51.71	**15.29**	31.88	38.40
LaUE	19.51	62.34	19.98	43.67	41.55	19.56	69.31	**7.98**	42.40	52.74	18.04	54.27	16.04	44.59	49.91
14A(Ours)	**7.24**	**29.58**	**19.21**	**19.19**	**28.74**	**6.92**	**36.06**	25.32	**17.46**	**35.68**	**9.37**	**30.68**	18.04	**15.87**	**30.04**
Methods	ResNet50					VGG16					VIT				
	Pets	Flowers	Cars	Food	Sun	Pets	Flowers	Cars	Food	Sun	Pets	Flowers	Cars	Food	Sun
CLEAN	37.69	69.07	48.4	68.71	60.59	41.41	71.39	43.78	52.49	57.85	30.36	58.68	36.00	45.63	41.78
CP	17.41	48.28	15.50	29.73	39.50	22.29	35.57	14.28	19.58	30.59	17.49	45.84	15.61	32.85	32.49
TUE	15.94	45.96	15.30	34.86	40.45	21.01	40.95	14.36	18.80	29.97	17.47	44.86	14.71	32.36	32.13
LaUE	15.56	45.35	**5.73**	48.67	43.24	16.76	33.12	**8.90**	15.53	19.04	14.00	31.05	**10.09**	22.91	24.55
14A(Ours)	**9.18**	**25.42**	11.10	**13.88**	**25.93**	**9.21**	**21.51**	9.40	**7.36**	**17.22**	**8.55**	**10.63**	13.54	**13.46**	**17.40**

label-agnostic, we evaluate the robustness of 14A against various existing attacks, including data augmentation attack, i.e., Mixup (Zhang et al., 2017) and Gaussian smoothing, as well as the attack tailored for unlearnable examples, i.e., Orthogonal Projection Attack (OPA) (Sandoval-Segura et al., 2023).

From Table 3, we have the following observations: i) the average accuracy post-attack is 16.91% (with a maximum of 34.10%), indicating that the model remains far from being useful after the attack. Hence, the data remains unlearnable after attacking; ii) among the three attacking methods (Mixup, Gaussian, OPA), Mixup and Gaussian attacks increase accuracy by 45.92% and 24.80% respectively, while OPA decreases it by 16.89%, suggesting that Mixup and Gaussian attacks are effective to some extent. Conversely, OPA's negative impact on testing accuracy further solidifies the unlearnability of the data. This is attributed to OPA leveraging linear separability for attacks. However, our proposed method is based on concept transferring rather than linear separable perturbations. and iii) attacking methods exhibit better performance on larger models, i.e., VGG16 and VIT, compared to ResNet18. We suppose that the larger model's superior feature extraction capabilities make it more sensitive to perturbations, aligning with the observations in Section 4.2.2. These results validate the effectiveness of 14A in real-world applications, and gain further confidence in the reliability and practicality of our proposed approach.

4.4. Performance on Low-resolution Images

Our proposed 14A generator aims to be a universal generator that is once-trained multiple-used. It can render any input data sample unlearnable without the need for retraining. However, note that the size or resolution of the input image can have a significant impact on the generator's performance. This can be also seen as a form of data transformation attack that affects unlearnability. Therefore, we investigate the performance of 14A on low-resolution images.

Due to the design of a universal generator, 14A is primarily trained on high-resolution images from ImageNet (224 × 224). The generator's structure is predefined, which limits its applicability to images of other sizes, e.g., CIFAR10 and CIFAR100 (32 × 32). Consequently, we add a resizing module before 14A to align with the generator's input dimension. Figure 3 illustrates that 14A's performance on low-resolution images is not as strong as its performance on other domain-specific datasets. This is primarily due to the loss of image details caused by resizing, which affects both the embedding modeling by CLIP and the perturbation generation by our generator. Nevertheless, compared with the unprotected data, the testing accuracies for CIFAR10 and CIFAR100 still decrease by 58.5% and 70.1%, respectively, demonstrating effective unlearnability. In comparison, gradient-based approaches directly generate perturbations without taking images as input during inference,

Table 3. Testing accuracy (%) ↓ against attacking methods under cross-dataset and label-agnostic settings.

METHODS	RESNET18					VGG16					VIT				
	PETS	FLOWERS	CARS	FOOD	SUN	PETS	FLOWERS	CARS	FOOD	SUN	PETS	FLOWERS	CARS	FOOD	SUN
NULL	7.24	29.58	19.21	19.19	28.74	9.21	21.51	9.40	7.36	17.22	8.55	10.63	13.54	13.46	17.40
MIXUP	9.10	34.10	15.17	26.21	32.50	14.39	26.40	11.24	16.22	32.85	15.78	20.9	16.50	21.25	25.86
GAUSSIAN	8.58	21.63	13.41	26.95	30.05	11.77	23.59	12.93	12.71	30.53	11.47	14.18	13.95	17.82	23.95
OPA	5.83	13.93	5.40	8.58	10.71	6.59	13.44	7.88	5.42	9.61	15.34	13.69	14.56	16.22	21.83

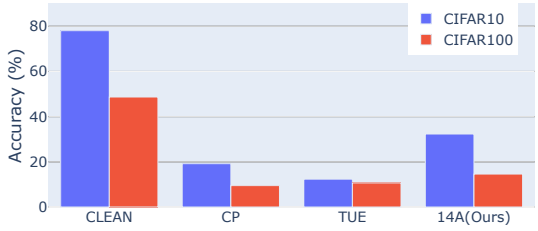


Figure 3. Cross-dataset transferability to low-resolution images.

allowing for resizing of the perturbations to fit the low-resolution images. As depicted in Figure 3, CP and TUE exhibit better performance in this scenario. Overall, there is room for improvement in the application of 14A to low-resolution images, and we leave the exploration of resizing modules in future work.

4.5. Effect of k

The hyper-parameter k plays a crucial role in concept-wise discriminant loss. To validate the effect of k, we investigate by varying k within the range of 1 to 10. We report the training loss of different k in Figure 4(a). Our observations are as follows: i) all values of k converge to similar loss values; ii) introducing multiple similar concepts ($k > 1$) significantly accelerates coverage speed; and iii) Among the tested values, $k = 4$ and $k = 5$ converge to relatively lower loss values. Additionally, we report the final cosine similarity between the training image embeddings and their corresponding similar concepts (after the generator stops updating) in Figure 4(b). Smaller cosine similarity values indicate that the perturbations push the input data further away from their similar concepts. Based on these findings, we select $k = 5$ as it has the smallest cosine similarity, indicating a greater separation from similar concepts.

4.6. Number of Clusters

The domain-specific datasets have a label category ranging from 37 to 397. To achieve a satisfactory clustering result, we investigate the effect of the number of clusters, as k-means algorithm is sensitive to this parameter (Ahmed et al., 2020). Specifically, we use three well-known metrics, i.e., Sum of Squares of Errors (SSE), Silhouette Coefficient (SC), and Calinski-Harabasz Index (CH), to measure the clustering result.

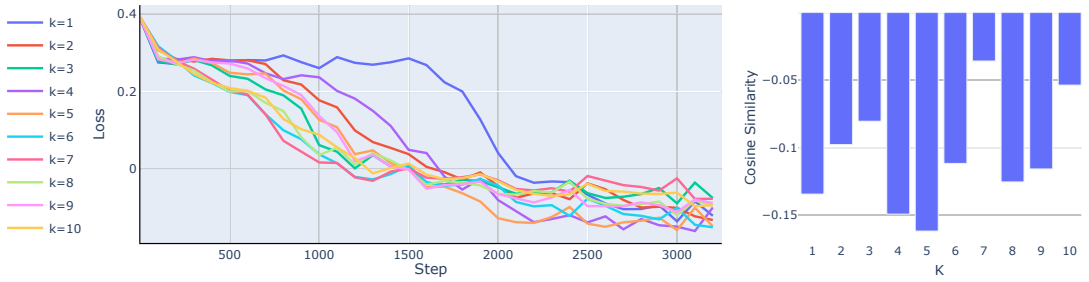
- **SSE**: A clustering result is considered to be optimal when the descent rate of SSH slows down.
- **SC**: This metric considers both the cohesion and separation of clusters.
- **CH**: This metric computes the ratio of inter-cluster and intra-cluster distance.

As shown in Figure 5, despite occasional fluctuations, the overall trend in clustering metrics remains stable across most datasets. Increasing the number of clusters generally improves performance, aligning with previous findings that more clusters can overlap with the original labels (Zhang et al., 2023). However, a smaller number of clusters emphasizes the impact of inconsistent labels on unlearnable examples. To strike a balance, we select 20 clusters for all datasets.

5. Related Work

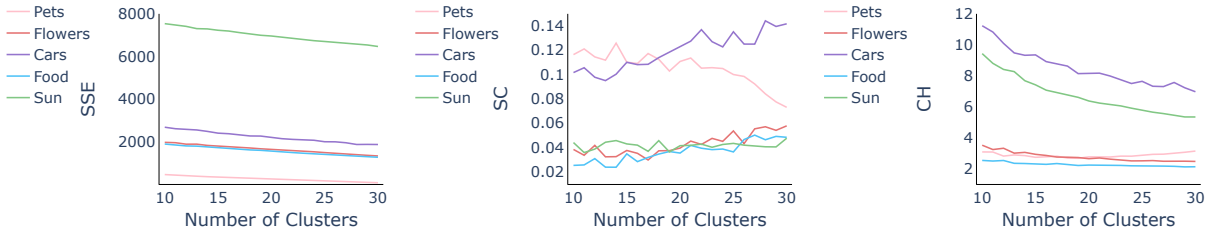
Producing unlearnable examples is a feasible approach to prevent unauthorized usage of protected data. An unauthorized adversary model that is trained on unlearnable examples would perform poorly. Existing unlearnable examples are produced by injecting perturbations through a bi-level optimization (Feng et al., 2019; Tao et al., 2021; Fowl et al., 2021; Huang et al., 2021; Liu et al., 2021b; Fu et al., 2022). The perturbations are constrained within a predefined budget ($\|\boldsymbol{\delta}\|_\infty < \epsilon$), to ensure they are imperceptible to the human eye. Recent studies have pointed out that these perturbations can serve as shortcuts, making the model easily learn on them while ignoring the original data's features (Yu et al., 2022; Sandoval-Segura et al., 2022).

Several methods have been proposed to address cross-dataset transferability (Sandoval-Segura et al., 2022; Yu et al., 2022; Ren et al., 2023; Han et al., 2023) and label-agnostic utility (Zhang et al., 2023). However, there is currently no method that can demonstrate both applicability simultaneously. In a recent study (He et al., 2023), unlearnable examples for unsupervised learning are investigated. Nevertheless, our work mainly focuses on supervised learning, which is more commonly encountered in practice.



(a) Training Loss (b) Cosine Similarity ↓

Figure 4. Effect of k in terms of training loss and cosine similarity in the embedding space.



(a) SSE ↓ (b) SC ↑ (c) CH ↑

Figure 5. Clustering results (evaluated by three metrics, i.e., SSE, SC, and CH) of different numbers of clusters.

6. Conclusion

In this paper, we identify limitations of existing unlearnable examples in real-world applications, i.e., suffering from a lack of cross-dataset transferability and being confounded by label-agnostic scenarios. To address these limitations, we put forth the concept of *concept unlearnable*, which allows us to broaden the unlearnability beyond specific datasets or labels. To achieve this, we adopt the generator-based approach and propose a bold endeavor, developing a *universal* perturbations generator (14A) capable of producing data with concept unlearnability, i.e., enjoying cross-dataset transferability and label-agnostic utility. Specifically, we integrate all data samples in a shared embedding space through multi-modal modeling. This integration connects the information from the image data with the sufficient concept from the text data, thereby promoting concept unlearnability. The use of multi-modal modeling also eliminates the need for a surrogate model, simplifying the perturbation generation process. In addition, the concept unlearnability enables the 14A generator to allow input from any samples without requiring retraining. This remarkable capability opens up possibilities for achieving an "unlearnable anything" model. Through extensive experiments on real-world datasets, our proposed unlearnable examples demonstrate effective concept unlearnability with impressive robustness against attacks, validating the effectiveness and reliability of our universal generator.

Limitations and Future work. Although our proposed 14A generator enables cross-dataset and label-agnostic concept unlearnability, it remains distant from a once-trained multiple-used generator. As illustrated in Section 4.4, due to its inherent design, 14A faces performance degradation when inputting images with significantly different sizes from whose it was trained on. Our future work involves incorporating an adaptive resizing module to rectify this limitation. Additionally, extending unlearnability beyond image data to other modalities, e.g., audio (Gokul & Dubnov, 2024), text (Li et al., 2023), and videos, represents another research direction for continued exploration.

Acknowledgements

This work was supported by the "Pioneer" and "Leading Goose" R&D Program of Zhejiang (2023C01030, 2024C01107, 2022C01068), Zhejiang Provincial Natural Science Foundation of China (LDT23F01011F01, LDT23F01015F01, LD24F020007, LDT23F01014F01), National Nature Science Foundation of China (U21B2024, 61931008, 62071415), National Key Research and Development Program of China under Grant (2020YFB1406604, 2023YFB4502800, 2023YFB4502803).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Ahmed, M., Seraj, R., and Islam, S. M. S. The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, 9(8):1295, 2020.

Bossard, L., Guillaumin, M., and Van Gool, L. Food-101–mining discriminative components with random forests. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, pp. 446–461. Springer, 2014.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Dai, H., Li, H., Tian, T., Huang, X., Wang, L., Zhu, J., and Song, L. Adversarial attack on graph structured data. In *International conference on machine learning*, pp. 1115–1124. PMLR, 2018.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., and Li, J. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9185–9193, 2018.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Feng, J., Cai, Q.-Z., and Zhou, Z.-H. Learning to confuse: generating training time adversarial data with auto-encoder. *Advances in Neural Information Processing Systems*, 32, 2019.

Fowl, L., Goldblum, M., Chiang, P.-y., Geiping, J., Czaja, W., and Goldstein, T. Adversarial examples make strong poisons. *Advances in Neural Information Processing Systems*, 34:30339–30351, 2021.

Fu, S., He, F., Liu, Y., Shen, L., and Tao, D. Robust unlearnable examples: Protecting data privacy against adversarial learning. In *The Tenth International Conference on Learning Representations, ICLR 2022*. OpenReview.net, 2022.

Gokul, V. and Dubnov, S. Poscuda: Position based convolution for unlearnable audio datasets. *arXiv preprint arXiv:2401.02135*, 2024.

Han, Z., Zheng, X., Chen, C., Cheng, W., and Yao, Y. Intra and inter domain hypergraph convolutional network for cross-domain recommendation. In *Proceedings of the ACM Web Conference 2023*, WWW '23, pp. 449–459, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394161. doi: 10.1145/3543507.3583402.

Han, Z., Chen, C., Zheng, X., Li, M., Liu, W., Yao, B., Li, Y., and Yin, J. Intra-and inter-group optimal transport for user-oriented fairness in recommender systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 8463–8471, 2024.

He, H., Zha, K., and Katabi, D. Indiscriminate poisoning attacks on unsupervised contrastive learning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Huang, H., Ma, X., Erfani, S. M., Bailey, J., and Wang, Y. Unlearnable examples: Making personal data unexploitable. In *9th International Conference on Learning Representations, ICLR 2021*. OpenReview.net, 2021.

Jia, H., Ding, S., Xu, X., and Nie, R. The latest research progress on spectral clustering. *Neural Computing and Applications*, 24:1477–1486, 2014.

Krause, J., Stark, M., Deng, J., and Fei-Fei, L. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 554–561, 2013.

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

Li, X., Liu, M., and Gao, S. Make text unlearnable: Exploiting effective patterns to protect personal data. *arXiv preprint arXiv:2307.00456*, 2023.

Li, Y., Li, Y., Wu, B., Li, L., He, R., and Lyu, S. Invisible backdoor attack with sample-specific triggers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 16463–16472, 2021.

Liu, W., Su, J., Chen, C., and Zheng, X. Leveraging distribution alignment via stein path for cross-domain cold-start recommendation. *Advances in Neural Information Processing Systems*, 34:19223–19234, 2021a.

Liu, W., Zheng, X., Chen, C., Su, J., Liao, X., Hu, M., and Tan, Y. Joint internal multi-interest exploration and external domain alignment for cross domain sequential recommendation. In *Proceedings of the ACM Web Conference 2023*, pp. 383–394, 2023a.

Liu, W., Zheng, X., Su, J., Zheng, L., Chen, C., and Hu, M. Contrastive proxy kernel stein path alignment for cross-domain cold-start recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 2023b.

Liu, Z., Zhao, Z., Kolmus, A., Berns, T., van Laarhoven, T., Heskes, T., and Larson, M. Going grayscale: The road to understanding and improving unlearnable examples. *arXiv preprint arXiv:2111.13244*, 2021b.

Ma, N., Zhang, X., Zheng, H.-T., and Sun, J. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 116–131, 2018.

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

Nilsback, M.-E. and Zisserman, A. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pp. 722–729. IEEE, 2008.

Parkhi, O. M., Vedaldi, A., Zisserman, A., and Jawahar, C. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3498–3505. IEEE, 2012.

Poursaeed, O., Katsman, I., Gao, B., and Belongie, S. Generative adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4422–4431, 2018.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Ren, J., Xu, H., Wan, Y., Ma, X., Sun, L., and Tang, J. Transferable unlearnable examples. In *The Eleventh International Conference on Learning Representations, ICLR 2023*. OpenReview.net, 2023.

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

Saha, A., Subramanya, A., and Pirsiavash, H. Hidden trigger backdoor attacks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 11957–11965, 2020.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

Sandoval-Segura, P., Singla, V., Geiping, J., Goldblum, M., Goldstein, T., and Jacobs, D. Autoregressive perturbations for data poisoning. *Advances in Neural Information Processing Systems*, 35:27374–27386, 2022.

Sandoval-Segura, P., Singla, V., Geiping, J., Goldblum, M., and Goldstein, T. What can we learn from unlearnable datasets? In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Bengio, Y. and LeCun, Y. (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

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

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., and Torralba, A. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 3485–3492. IEEE, 2010.

Yu, D., Zhang, H., Chen, W., Yin, J., and Liu, T.-Y. Availability attacks create shortcuts. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2367–2376, 2022.

Yuan, Z., Zhang, J., Jia, Y., Tan, C., Xue, T., and Shan, S. Meta gradient adversarial attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7748–7757, 2021.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

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