# Highlights

# Impart: An Imperceptible and Effective Label-Specific Backdoor Attack

Jingke Zhao, Zan Wang, Yongwei Wang, Lanjun Wang

- We propose a novel backdoor attack framework, Impart, where the attacker uses a surrogate model to generate effective backdoor examples in the scenario where the attacker does not have access to the model information.
- we first propose a label-specific attack, where the generated backdoor examples are associated with the target label before the backdoor attack which significantly enhances the attack capability of the backdoor attack.
- Our method Impart outperforms existing works with an average attack success rate 13% in the all-to-all setting
  on CIFAR-100 while keeping highly imperceptible with average visual quality improvements from 34.24dB to
  40.45dB in PSNR.

# Impart: An Imperceptible and Effective Label-Specific Backdoor Attack

Jingke Zhao<sup>a</sup>, Zan Wang<sup>a</sup>, Yongwei Wang<sup>b</sup> and Lanjun Wang<sup>a,\*</sup>

#### ARTICLE INFO

# Keywords: Data Poisoning Backdoor Attack

Model Security

Deep Learning

#### ABSTRACT

Backdoor attacks have been shown to impose severe threats to real security-critical scenarios. Although previous works can achieve high attack success rates, they either require access to victim models which may significantly reduce their threats in practice, or perform visually noticeable in stealthiness. Besides, there is still room to improve the attack success rates in the scenario that different poisoned samples may have different target labels (a.k.a., the all-to-all setting). In this study, we propose a novel imperceptible backdoor attack framework, named Impart, in the scenario where the attacker has no access to the victim model. Specifically, in order to enhance the attack capability of the all-to-all setting, we first propose a label-specific attack. Different from previous works which try to find an imperceptible pattern and add it to the source image as the poisoned image, we then propose to generate perturbations that align with the target label in the image feature by a surrogate model. In this way, the generated poisoned images are attached with knowledge about the target class, which significantly enhances the attack capability. We conduct experiments on three benchmark datasets and five widely used defense mechanisms. Experiments show that Impart achieves successful attacks and high imperceptibility in five image quality metrics. For example, our method outperforms existing works with an average attack success rate 13% in the all-to-all setting on CIFAR-100 while keeping highly imperceptible with average visual quality improvements from 34.24dB to 40.45dB in PSNR. Additionally, we demonstrate that Impart can successfully bypass existing effective defense methods.

# 1. Introduction

Deep Neural Networks (DNNs) have achieved remarkable success in the past few years and they have been adopted in different applications (e.g., image classification (He, Zhang, Ren and Sun, 2016a), speech recognition (Xiong, Droppo, Huang, Seide, Seltzer, Stolcke, Yu and Zweig, 2016), game playing and natural language processing (Silver, Huang, Maddison, Guez, Sifre, Van Den Driessche, Schrittwieser, Antonoglou, Panneershelvam, Lanctot et al., 2016; Devlin, Chang, Lee and Toutanova, 2019)). However, with the deepening research on several real security-critical scenarios, recent works show that even the state-of-the-art deep learning methods are vulnerable to backdoor attacks (Gu, Dolan-Gavitt and Garg, 2017; Barni, Kallas and Tondi, 2019; Cheng, Liu, Ma and Zhang, 2021; Li, Li, Wu, Li, He and Lyu, 2021a; Cheng, Wu, Zhang and Zhao, 2023). In backdoor attacks, an attacker injects a trigger into the victim model in the training process. The victim model performs normally as a benign model in the inference phase when the inputs are benign images. However, once the victim model is fed an input image with the backdoor trigger, the victim model behaves as the attacker predetermined. In the backdoor attack, there are two typical types of attack settings (Li, Jiang, Li and Xia, 2022): one is to poison different target labels (a.k.a., *all-to-all*), and the other is to poison one target label (a.k.a., *all-to-one*).

Recent research on the backdoor attack for deep learning has focused on generating poisoned images that lead to misclassification results while keeping imperceptibility. LIRA (Doan, Lao, Zhao and Li, 2021b) and WB (Doan, Lao and Li, 2021a) have achieved effective and imperceptible backdoor attacks. However, they assume that the attacker has full access to the model information (e.g., model architecture, and model parameters), which significantly reduces their threats in practice. Meanwhile, for existing black-box backdoor methods that do not utilize the information about the

ORCID(s): 0000-0002-7696-5330 (L. Wang)

<sup>&</sup>lt;sup>a</sup>Tianjin University, 92 Weijin Road, Tianjin, 300072, China

<sup>&</sup>lt;sup>b</sup>Nanyang Technological University, 50 Nanyang Avenue, 639798, Singapore

<sup>\*</sup>Corresponding author

hi\_simon@tju.edu.cn (J. Zhao); wangzan@tju.edu.cn (Z. Wang); yongweiw@ece.ubc.ca (Y. Wang); wang.lanjun@outlook.com (L. Wang)

victim model (Cheng et al., 2021; Li et al., 2021a; Nguyen and Tran, 2021; Wang, Zhai and Ma, 2022b), the generated poisoned images are inevitably visually noticeable.

Moreover, compared with the all-to-one setting, the all-to-all setting has not been well studied. We observe that there is still an unsatisfied attack success rate in the all-to-all setting. For example, WaNet (Nguyen and Tran, 2021), an advanced black-box backdoor attack method, achieves a 99.56% high ASR in the all-to-one setting on the CIFAR-10 dataset, yet it only has a 94% ASR in the all-to-all setting. Even typical white-box methods, such as LIRA (Doan et al., 2021b) and WB (Doan et al., 2021a), can only achieve about 94% ASR in the all-to-all setting. In addition, the all-to-all setting is more valuable to be investigated. First, the attacker can attack multiple classes simultaneously by injecting a type of backdoor trigger, which is much more useful and flexible in practice. Second, the all-to-all setting attacks are more difficult to be detected for their complicated target shifting, and therefore, are more serious compared with the all-to-one attacks (Li et al., 2022). Besides, as described in Doan, Lao and Li (2022), repeatedly injecting different trigger patterns for all the target classes is not feasible because it will lead to a much larger model perturbation and significantly degrade the clean data accuracy. There were only a few studies specifically designed on the all-to-all(Li et al., 2022) setting, yet how to better design the all-to-all attack remains underexplored.

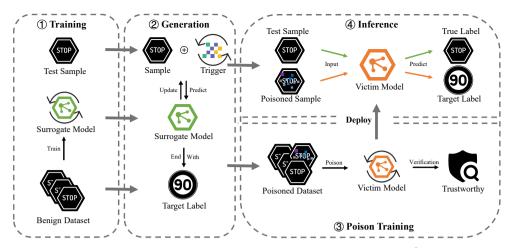
It is still a crucial challenge to achieve an *imperceptible and effective backdoor attack* for the all-to-all setting in the scenario where the attacker does not have access to the model information. First, the human visual system has different perception sensitivities to different regions (Eckert and Bradley, 1998; Wang, Wang, Feng, Ward and Wang, 2022a) and different colors (Zhao, Liu and Larson, 2020; Wang, Ding, Yang, Ding, Ward and Wang, 2021). For example, distortions around the edges are easier to be noticed than those in the texture regions. Another example is that even when identical perturbations are added in all three channels (RGB), the perturbations are more easily noticed in the green channel (Li et al., 2021a). However, some of the previous studies fail to consider this human visual sensitivity (Wang et al., 2022b; Nguyen and Tran, 2021; Li et al., 2021a). For example, Nguyen and Tran (2021) considers distorting the whole image, but not the texture regions. Besides, most existing studies try to manipulate the source image by some transformations involving randomness (Nguyen and Tran, 2020; Li et al., 2021a; Nguyen and Tran, 2021; Wang et al., 2022b). That is to say, the triggers are label-agnostic. Therefore, in order to achieve successful poison, there is a need for large amounts of transformations that damages imperceptibility as demonstrated in Sec. 5. In addition, backdoor attacks aim to make samples with those triggers output as the designed target label. However, due to the randomness, the victim model is hard to learn a mapping from triggers to the target label.

To tackle these challenges, in this study, different from previous works that try to obtain an imperceptible pattern and add it to the source image as the poisoned image, we propose to generate perturbation that aligns with the target label in the image feature by a surrogate model. Specifically, in order to promote the learning of the model, we propose a label-specific attack in which a generated poisoned image is associated with the target label. Namely, we utilize the information of the target label to generate a backdoor trigger for a specific image. To summarize, we propose a new attack framework named **Impart**. Firstly, we train a surrogate model before the backdoor attack. The surrogate model is used to fit the image feature. Then, the surrogate model combined with the target label and learned image feature to generate triggers. In this way, the backdoor images (i.e., embed the generated triggers into the source image) are associated with the target label in the image feature before implementing the backdoor attack. Finally, we use backdoor images to train the victim model which enhances the mapping between the backdoor image and the target label.

Our main contributions can be summarized as:

- We propose a novel backdoor attack framework, Impart, where the attacker uses a surrogate model to generate effective backdoor examples in the scenario where the attacker does not have access to the model information.
- To our best knowledge, we first propose a label-specific attack, where the generated backdoor examples are
  associated with the target label before the backdoor attack which significantly enhances the attack capability of
  the backdoor attack.
- We empirically evaluate Impart on three benchmark datasets and demonstrate that the proposed method achieves
  high imperceptibility, meanwhile outperforming existing methods in the all-to-all setting. For example, Impart
  outperforms existing works with an average attack success rate of 13% in the all-to-all setting on CIFAR-100
  while keeping highly imperceptible with average visual quality improvements from 34.24dB to 40.45dB in
  PSNR.
- We further evaluate Impart under five widely used defense mechanisms and demonstrate Impart can successfully bypass all of them.

#### **Impart**



**Figure 1:** The framework of the proposed Impart method consists of four phases. In phase ①, we train a surrogate model. In phase ②, we generate poisoned training data and poisoned test data using the feature fitter. In phases ③ and ④, we poison and test the victim model respectively.

# 2. Related Work

As our method is motivated by adversarial attacks, we first investigate adversarial attack methods in Sec. 2.1, where we mainly focus on imperceptible adversarial attacks. Then, some state-of-the-art backdoor attack approaches are analyzed in Sec. 2.2.

# 2.1. Adversarial Attacks

Adversarial attacks arise as a severe threat in the model inference phase (Moosavi-Dezfooli, Fawzi and Frossard, 2016; Yuan, He, Zhu and Li, 2019; Pei, Cao, Yang and Jana, 2017; Ma, Jiang and Yu, 2023), where the attacker attempts to induce misclassification of a model by manipulating the inputs. For different purposes of attacks, the attacker can perform a targeted or non-targeted attack. In the targeted attack setting, the adversary can flexibly attack any target label. Typically, (Papernot, McDaniel, Jha, Fredrikson, Celik and Swami, 2016) proposes an adversarial attack method in the white-box setting to generate perturbations using gradient descent with the  $\ell_2$  regularization in the targeted setting. PerC-AL (Rony, Hafemann, Oliveira, Ayed, Sabourin and Granger, 2019) achieves state-of-the-art performance in invisible attacks. PerC-AL replaces the original penalty  $\ell_2$  with a perceptual color difference metric CIEDE2000 (Luo, Cui and Rigg, 2001). Besides, the implementation of PerC-AL decouples the joint optimization by alternately updating the perturbations with respect to either classification loss or perceptual color difference. However, it required time-consuming large epochs to find a perturbation that is both effective and imperceptible because of the alternate optimization process. Additionally, while the adversarial examples generated by PerC-AL are imperceptible on average, there are still some special examples that have poor visual quality.

# 2.2. Backdoor Attacks

This section provides a brief overview of existing backdoor attacks, which can be classified into three categories: patch-based trigger attacks, input-specific trigger attacks, and adversarial examples-based attacks. Our approach is more related to the last two categories.

**Patch-based Trigger.** BadNets (Gu et al., 2017) is a typical patch-based backdoor attack method, which uses a pattern of bright pixels lying in the bottom right corner of the image. (Liu, Ma, Aafer, Lee, Zhai, Wang and Zhang, 2018) generates a trojan trigger by inversing the neurons. Then, several works (Barni et al., 2019; Chen, Liu, Li, Lu and Song, 2017; Liu, Ma, Bailey and Lu, 2020) started to explore an effective pattern while keeping some stealthiness. They suggested that the poisoned images should look as identical as possible to the source image under human inspection. Unfortunately, the patch is still clearly noticeable in existing works.

Input-Specific Trigger Attacks. To achieve better stealthiness, (Nguyen and Tran, 2020) implemented an input-aware trigger generator by proposing a diversity loss. ISSBA (Li et al., 2021a) proposed to generate sample-specific invisible triggers through an encoder-decoder network, which encodes a specified string into benign images. Until recently,

WaNet (Nguyen and Tran, 2021) and BppAttack (Wang et al., 2022b) have achieved an unnoticeable modification to some degree by traditional image warping and compression. As mentioned in Sec. 1, all of the above works ignore how the model behaves and how the human visual system perceives, leading to the added perturbation large enough to be learned by the model and perceived by human eyes. Besides, in previous works, the poisoned data are randomly distributed with respect to the target label before the training process.

Adversarial Examples-Based Attacks. AdvDoor (Zhang, Ding, Tian, Guo, Yuan and Jiang, 2021) is most related to our method. AdvDoor abstracts a fixed universal adversarial perturbation by combining multiple input-specific perturbations generated by the adversarial attack. However, it breaks up the advantage (i.e., revealing the defect of models) of adversarial examples because of the fixed perturbation. Another issue of universal adversarial perturbation is that the imperceptibility is worse than the input-specific perturbation. Our proposed method Impart is different from it. Specifically, we generate input-specific perturbations and the generated perturbation are designed for the target label. Arbitrary Label Attacks. Recently, Doan et al. (2022) propose a new backdoor attack Marksman Backdoor (MB) just like the adversarial attack that the adversarial can flexibly attack any target label during inference phases. They proposed a class-condition trigger generation function, and alternately update the generation function and victim model while fixing the other one during backdoor training. However, it is still a white-box method that significantly reduces the threat in reality. Besides, the MB is only a special all-to-one attack that the adversarial can flexibly attack any target. Therefore, as described in Sec. 1, how to achieve an imperceptible and effective backdoor attack for the all-to-all setting in the scenario where the attacker does not have access to the model information is still a crucial challenge.

# 3. Threat Model

The threat model has two aspects: one is the attack scenario (Sec. 3.1), and the other one is the specific attack objective (Sec. 3.2).

#### 3.1. Attack Scenario

In this study, we focus on the scenario where a victim user uploads her/his dataset, model, and training schedule to an untrusted third-party platform (e.g., Google Cloud) and train their model. The attacker (i.e., the malicious platform) modifies the dataset and training schedule during the training process. Then, the victim user gets the poisoned model. After evaluating the model with clean test data, the victim user deploys it into the production environment. Specifically, we assume that the attackers have access to the training set, and can modify the training schedule. All of these settings are following previous works (Nguyen and Tran, 2021; Wang et al., 2022b). However, in order to make our attack more practical, we assume the attackers have no information about the victim model architecture and model parameters.

#### 3.2. Attack Objective

The main goal of our method is to produce a high attack success rate (ASR) after injecting a backdoor into the victim model while keeping a comparable accuracy for benign data. Specifically, the victim model which is trained in the poisoned dataset behaves normally in benign data as a benign model that is trained in the benign dataset. Once the attacker feeds inputs with the trigger into the victim model, the victim model behaves as the attacker pre-determined. Besides, imperceptibility is another goal that we need to achieve, which requires the poisoned data as identical as possible to benign data under human visual inspection.

#### 4. Method

In this section, we first provide a formal problem formulation of the backdoor attack in Sec. 4.1, and then present the detail of our method Impart in Sec. 4.2.

# 4.1. Problem Formulation

As in the standard supervised images classification task, the goal is to learn the parameters  $\theta$  of a mapping function  $f_{\theta}: \mathcal{X} \to \mathcal{C}$  from the training dataset  $D_{tr} = \{(x_i, y_i) \mid x_i \in \mathcal{X}, y_i \in \mathcal{C}, i = 1, ..., N\}$ .

In the backdoor attack, the attacker aims to produce a poisoned subset  $D_p = \{(T(x_i), \eta(y_i)) \mid (x_i, y_i) \in D_s \subset D_{tr}, \eta(y_i) \in C\}$ , where  $D_s$  is a subset of  $D_{tr}$  and  $|D_s| = M = \rho N$ ,  $\rho$  is the poison ratio, T is a backdoor transformation function, and  $\eta$  is a target label function. In the backdoor attack, there are two typical types of target label functions (Li et al., 2022),

- 1. all-to-all:  $\eta(v) = (v+1) \mod |C|$ , where the true label is one-shifted.
- 2. *all-to-one*:  $\eta(y) = c$ , where c is a fixed constant label;

The poisoned model is trained on the dataset  $D = D_p \cup D_r$ , where  $D_r = D_{tr} \setminus D_s$ . The poisoned model with new parameters  $\theta^*$  is required to satisfy:

$$f_{\theta^*}(x) = y \tag{1}$$

$$f_{\theta^*}(T(x)) = \eta(y) \tag{2}$$

For the traditional perturbation-based method, the transformation functions T can be defined as:

$$T(x) = x + g(x), ||g(x)|| < \epsilon, \tag{3}$$

where g is an imperceptible perturbation generative function.

Differently, for the label-specific attack, we are trying to generate poisoned images associated with the target label. This can be formally described by:

$$T(x,c) = x + t(x,c), \tag{4}$$

where t(x, c) is a surrogate model which is used to fit the image feature and then combined with the target label and learned image feature to generate triggers. In this way, the image feature in the class c can be embedded in the poisoned image T(x, c) and enhance the backdoor attack capability. Besides, we expect the perturbation generative function t(x, c) to align with the characteristic of the human visual system. This can be formally described by:

$$f_{\theta}(T(x,c)) = \eta(y)$$
s.t.  $d\left(T(x,c),x\right) \le \epsilon$ , (5)

where d is a distance metric quantifying the human visual system,  $\epsilon$  is the upper bound of perpetual color difference. We detail the solution in the next section.

# 4.2. Impart

Fig. 1 shows the framework of our method Impart that consists of four phases. First, we train a surrogate model using the identical dataset as the victim model  $\oplus$ . Then, by utilizing the target label and the knowledge of the surrogate model, we generate poisoned examples that are aligned to the target label in the image feature  $\oslash$ . Next, using the generated poisoned data to train the victim model  $\oslash$ . After the model is deployed into the production environment, the adversary applies phase  $\oslash$  to generate poisoned test data to attack  $\oplus$ .

**Training a Surrogate Model.** The first step of our method is to train the surrogate model using the identical clean dataset as the victim model. Formally, given a training data  $D_{tr}$ , we train the surrogate model  $g_{\theta}$  by minimizing the following loss function:

$$\min_{\theta} \sum_{i=1}^{N} \mathcal{L}(g_{\theta}(x_i), y_i), \tag{6}$$

where  $\mathcal{L}(\cdot)$  denotes a loss function and g is the surrogate model.

Generating Poisoned Data. After training the auxiliary surrogate model  $g_{\theta}$ , we will generate the poisoned data subset  $D_p$ . Our goal is to generate poisoned data that are as close to the source data as possible while achieving a high ASR. Besides, to improve the effectiveness of poisoning data, the generated perturbations need to relate to the target label. This can be solved by the following optimization problem:

$$\min_{\mathcal{S}} \mathcal{L}\left(g_{\theta}(x+\delta), \eta(y)\right) + \gamma \cdot \|\delta\|_2 + \|\Delta E_{00}(x+\delta, x)\|_2. \tag{7}$$

In Eq. (7), we aim to generate imperceptible perturbation  $\delta$  with the constraint of  $\ell_2$  regularization and the perpetual color difference metric  $\Delta E_{00}$ .

For the first item  $\mathcal{L}(g_{\theta}(x + \delta), \eta(y))$ , where  $\mathcal{L}(\cdot)$  is a standard cross-entropy (CE) loss function. It tries to find a  $\delta$  which leads to the poisoned data  $x + \delta$  aligned with the target label  $\eta(y)$  in the image feature.

Table 1

The details of datasets and network architectures.

Dataset	#Classes	VM	ACC(%)
CIFAR-10	10	PreResNet18	94.81
GTSRB	43	PreResNet18	99.23
CIFAR-100	100	ResNet18	76.19

Considering that the human visual system is sensitive different from color to color. Therefore, using the second item  $\|\Delta E_{00}(x+\delta,x)\|_2$  to align the feature of human visual system. We use the latest standard formula developed by the International Commission on Illumination (CIE), which can be calculated as (Luo et al., 2001):

$$\Delta E_{00} = \sqrt{\left(\frac{\Delta L'}{k_L S_L}\right)^2 + \left(\frac{\Delta C'}{k_C S_C}\right)^2 + \left(\frac{\Delta H'}{k_H S_H}\right)^2} + \Delta R$$

$$\Delta R = R_T \cdot \frac{\Delta C'}{k_C S_C} \cdot \frac{\Delta H'}{k_H S_H}.$$
(8)

Eq. (8) calculates the color difference in Lab color space, where  $\Delta L', \Delta C', \Delta H'$  denote the distance of two inputs image in channel Lightness, Chroma and Hue respective, and  $\Delta R$  is an interactive item between channel chroma and hue. The  $k_L, k_C, k_H$  are based on application scenarios, and  $S_L, S_C, S_H$  act as compensation to better align with the human visual system. The parameter settings follow the paper (Luo et al., 2001).

The third item  $\gamma \cdot \|\delta\|_2$  is to further limit the update range of  $\delta$  and speed up the optimization process. Specifically, using the item  $\|\delta\|$ , the update range of  $\delta$  will be limited to small values. In this way, the imperceptibility of generated poisoned data are further improved. Besides, the number of iterations for finding an imperceptible perturbation are also greatly reduced. Additionally, the item  $\|\gamma\|$  ensures that the perturbation of all the generated poisoned data are minor and uniform to some degree, which is a primary reason for being able to pass existing defense methods. We experimentally demonstrate this in Sec. 5.4.

Following the PerC-AL (Zhao et al., 2020), we solve the optimization problem by alternately updating the perturbations with respect to either classification loss and  $\ell_2$  or perceptual color difference. Specifically, we first update the perturbation  $\delta$  with classification loss and  $\ell_2$  by gradient descent, then reduce the perpetual color difference of perturbations by gradient descent until the perturbation  $\delta$  loses its effect. Then, move back and forth.

**Poisoning the Victim Model.** Then, using the dataset  $D = D_p \cup D_r$ , where  $D_r = D_{tr} \setminus D_s$  to poison the model  $f_\theta$ , which can be formulated to:

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{N} \mathcal{L}\left(f_{\theta}(x_i), y_i\right), \tag{9}$$

where  $(x_i, y_i) \in D$ .

# 5. Experiments

In this section, we conduct attack experiments on three benchmark datasets and five image quality metrics to evaluate the attack's effectiveness and imperceptibility. Besides, we select five widely used defense mechanisms to evaluate whether our proposed method is resistant to them. Additionally, we investigate the influence of hyperparameters  $\rho$ , and  $\gamma$ .

#### 5.1. Experimental Setup

**Datasets and Models.** We conduct our method in three commonly used datasets: CIFAR-10 (Krizhevsky, Hinton et al., 2009), GTSRB (Stallkamp, Schlipsing, Salmen and Igel, 2012) and CIFAR-100 (Krizhevsky et al., 2009). Following the previous work (Nguyen and Tran, 2021), we consider the blend networks with datasets. Specifically, as for the victim model(VM), we use Pre-activation ResNet18 (He, Zhang, Ren and Sun, 2016b) for CIFAR-10 and GTSRB datasets, and ResNet18 for CIFAR-100 dataset. For the surrogate model(SM) used to fit the image feature, we select four existing models GoogleNet (Szegedy, Liu, Jia, Sermanet, Reed, Anguelov, Erhan, Vanhoucke and Rabinovich,

**Table 2**Attack effectiveness in the all-to-all setting. For each dataset, four rows represent the SM as GoogleNet, EfficientNetB0, EfficientNetB1, and ResNet34, respectively. \* means the poisoned ratio is 20%, and the others is 10%.

	SM		BA(%)			ASR(%)	
	SIVI	Impart	WaNet	Bpp*	Impart	WaNet	Bpp*
	GoogleNet	94.24	94.15		96.35		
CIEAD 10	${\sf EfficientNetB0}$	94.15			97.56	93.36	04.22
CIFAR-10	EfficientNetB1	94.21	94.43	94.73	97.40	93.30	94.32
	ResNet34	94.43			97.04		
	GoogleNet	98.03		99.46	98.90	98.32	99.29
CTCDD	EfficientNetB0	99.00	00.20		99.34		
GTSRB	EfficientNetB1	99.14	99.39		99.21		
	ResNet34	98.62			98.97		
_	GoogleNet	75.66			77.07		
CIEAD 100	EfficientNetB0	75.62 75.11 74.59		75.44	92.94		73.71
CIFAR-100	EfficientNetB1			75.44	91.21	74.05	
	ResNet34	76.09			85.41		

2015), EfficientNetB0, EfficientNetB1 (Tan and Le, 2019) and ResNet34 (He et al., 2016b) to test the effect of our framework. More detailed information is in Tab. 1.

**Baselines.** We compare our proposed method Impart with WaNet (Nguyen and Tran, 2021) and Bpp (Wang et al., 2022b) because both of them are state-of-the-art methods in the backdoor attack. Besides, they can be applied to the attack scenario which has no access to the victim model. We implement WaNet and Bpp on CIFAR-100 by using the default parameters stated in the original papers and reference their results of CIFAR-10 and GTSRB directly. For WaNet, the poisoned ratio is set at 10% and the noise ratio is set at 20%. For Bpp, the poisoned ratio and the negative ratio are both set at 20%. It is noted that for WaNet and Impart, the poisoned ratio is 10%, but for Bpp, the poisoned ratio is 20%. Therefore, Bpp is easier to achieve a high attack success rate than the others since Bpp uses many more poisoned examples. For both of WaNet and Bpp, we train the network using SGD optimizer with weight decay  $5\times10^{-4}$ . The initial learning rate is set to 0.01 with a learning rate decay of factor 10 after every 100 epochs.

**Attack Setup.** We set the poisoned ratio  $\rho = 10\%$  for all experiments if not specified otherwise. Besides, we train the networks using SGD optimizer with weight decay  $5 \times 10^{-4}$ . The learning rate is using cosine decay with a warm-up. The upper bound of the learning rate is set at 0.01. We set  $\gamma = 10$  to Eq. (7).

**Evaluation Metrics.** To evaluate the attack effectiveness, we use the attack success ratio (ASR) and benign accuracy(BA) designed in the previous works (Li et al., 2021a; Nguyen and Tran, 2021; Wang et al., 2022b). Besides, we use the accuracy of the benign model(ACC) as a reference. In detail, the ACC evaluates the accuracy of the benign model on benign data while the BA evaluates the accuracy of the poisoned model on benign data. The ACC is shown in Tab. 1.

In order to comprehensively evaluate the quality of the poisoned images, We select five image quality metrics CIEDE2000 (Luo et al., 2001), SSIM (Wang, Bovik, Sheikh and Simoncelli, 2004), PSNR (Huynh-Thu and Ghanbari, 2008),  $\ell_2$  and  $\ell_\infty$ . Both SSIM and CIEDE2000 are designed to better align with human visual perception, where CIEDE2000 is prone to color discrepancy while SSIM extracts structural information from the viewing field.

**Table 3** Imperceptibility in the all-to-all setting. ↑ means larger values are preferred, and vice versa. For each dataset, the row settings are the same as Tab. 2.

	C	IEDE200	01	P:	SNR(dB)	1	SSI	IM(×100	) ↑		$\ell_2\downarrow$			$\ell_{\infty}\downarrow$	
	Impart	WaNet	Bpp*	Impart	WaNet	Bpp*	Impart	WaNet	Bpp*	Impart	WaNet	Bpp*	Impart	WaNet	Врр*
	21.48			42.60			99.43			0.43			0.06		
CIFAR-10	23.12	121.73	110 14	42.20	28 20	28.39 33.37	99.21	93.81 98.86	0.50	2.23	1.04	0.07	0.27	0.11	
CII AIN-10	25.33	121.73	110.14	42.81	20.39		99.24		0.49	2.23	1.24	0.06	0.27	0.11	
	22.10			42.33			99.40			0.45			0.06		
	28.57			47.47			98.54			0.52			0.09		
GTSRB	27.97	106.05	06.21	41.33	21.52	39.34	98.43	94.96 <b>98.81</b>	00 01	0.57	1.85	0.60	0.08	0.23	0.03
GISKD	26.63	106.05	90.31	41.10	31.32		98.61		90.01	0.54	1.65 0.00	0.00	0.08		
	29.88			43.78			98.32			0.60			0.09		
	22.17			42.15			99.24			0.45			0.06		
CIFAR-100	32.72	116.50	106 12	39.88	20.02	24.24	98.81	04.02 00.02	0.61	2.12	1.12	0.07	0.26	0.10	
CIFAR-100	35.08	116.59	100.13	39.31	28.92	34.24 98.68	94.83 98.93	98.93	0.64	2.13	1.13	0.07		0.10	
	25.11			41.04			99.11			0.52			0.06		

**Table 4**Attack effectiveness in the all-to-one setting. For each dataset, four rows represent the SM as GoogleNet, EfficientNetB0, EfficientNetB1, and ResNet34, respectively. \* means the poisoned ratio is 20%, and the others is 10%.

	SM		BA(%)			ASR(%)	
	SIVI	Impart	WaNet	Bpp*	Impart	WaNet	Bpp*
	GoogleNet	94.11			98.97		
CIFAR-10	${\sf EfficientNetB0}$	94.37	94.15	94.54	98.07	99.55	00.01
CIFAR-10	EfficientNetB1	94.12	94.13	94.54	98.01		99.91
	ResNet34	94.22			99.03		
	GoogleNet	98.77	98.77		99.70		
CTCDD	${\sf EfficientNetB0}$	98.79	00.07	00.25	100	98.78	99.96
GTSRB	EfficientNetB1	98.73	98.97	99.25	100		
	ResNet34	98.08			99.89		
	GoogleNet	75.66			99.61		
CIFAR-100	EfficientNetB0	76.55	75.40	76.55	99.58	00.56	100
CIFAK-100	EfficientNetB1	76.51	75.40	76.55	99.43	99.56	100
	ResNet34	76.07			99.81		

# **5.2.** Attack Experiments

For the all-to-all setting, Tab. 2 shows that Impart outperforms the baselines in CIFAR-10 and CIFAR-100 on ASR without scarifying BA. Meanwhile, it achieves comparable results with baselines in GTSRB. Especially, for the GTSRB dataset, as described in the Sec. 1, just because the ACC of the GTSRB dataset in PreActResNet18 is about 99%, hence all the previous works also can achieve about 99% ASR in this dataset. However, our method achieves 92.94% ASR on CIFAR-100 when using EfficientNetB0 as the surrogate model which is much higher than the baseline. This is because the generated poisoned images in Impart are associated with the target label before training, which makes it much

<sup>&</sup>lt;sup>1</sup>The noise ratio in WaNet is used to control noise data generation for bypassing some defenses, and the negative ratio in Bpp plays the same role. However, Impart does not need extra data for defense purposes.

<sup>&</sup>lt;sup>2</sup>In experiments by the official code https://github.com/RU-System-Software-and-Security/BppAttack, we observed Bpp could not converge with a poisoned ratio as 10%.

**Table 5** Imperceptibility in the all-to-one setting. ↑ means larger values are preferred, and vice versa. For each dataset, the row settings are the same as Tab. 4.

	С	IEDE200	O.T	P:	SNR(dB)	1	SSI	M(×100)	) ↑		$\ell_2\downarrow$			$\ell_{\infty}\downarrow$	
	Impart	WaNet	Bpp*	Impart	WaNet	Bpp*	Impart	WaNet	Bpp*	Impart	WaNet	Bpp*	Impart	WaNet	Bpp*
	20.36			53.11			99.45			0.40			0.05		
CIFAR-10	21.36	121.73	110 14	52.66	28.39	22.27	99.26	93.81	3.81 98.86	0.47	2.23	1.24	0.06	0.27	0.11
CIFAR-10	24.02	121.73	116.14	52.99	26.39	33.37	99.27	93.61	90.00	0.46	2.23	1.24	0.05	0.27	0.11
	20.28			53.14		9	99.44			0.41			0.05		
	29.05			48.24			98.51			0.52			0.08		
GTSRB	29.12	106.05	06.21	41.92	21.52	20.24	98.31	04.06	00 01	0.60	1 05	0.60	0.09	0.22	0.03
GISKD	29.36	100.03	90.51	41.22	31.32	39.34	98.34	94.90	94.96 <b>98.81</b>	0.59	1.85	0.00	0.08	0.23	0.03
	30.18			44.07			98.32			0.61			0.09		
	23.69			42.61			99.17			0.48			0.06		
CIFAR-100	34.56	116.59	106 12	40.56	28.92	24.24	98.85	04.92	08.02	0.61	2.13 1.13	1 12	0.09	0.26 0	0.10
CIFAR-100	37.88	110.39	100.13	39.88	20.92	34.24	98.67	94.83	94.83 98.93	0.67		1.13	0.09		0.10
	26.76			41.51			99.02			0.54			0.06		

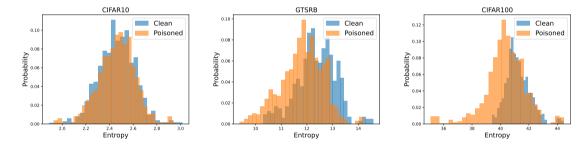


Figure 2: Resilience to STRIP. Entropy distributions on CIFAR-10, GTSRB and CIFAR-100.

easier to learn the mapping from triggers to the target label than baselines which only have random perturbations. In the CIFAR-100 dataset when using GoogleNet as the surrogate model, the ASR is 77.07%. It is because the fitting ability of the GoogleNet is worst than the EfficientNetB0 in the CIFAR-100 dataset. The model EfficientNetB0 can generate more precise perturbation which is related to the target label than the model GoogleNet. Finally, the victim model can easier learn the poisoned examples generated by EfficientNetB0 more than GoogleNet. To sum up, our method is more applicable to real-world datasets rather than the datasets like GTSRB which can achieve about 99% ACC. Besides, Tab. 5 shows Impart achieves remarkable superior image quality overall. In detail, when we compare SSIM only, we observe that the SSIM of Bpp is comparable with that of Impart. This indeed verifies Bpp retains the structure information while losing the color information because of using fewer bits to describe a pixel. Meanwhile, Impart holds the structure information and color information simultaneously, and thus, on CIEDE2000 which is a metric considering color information, Impart is far better than Bpp.

In addition, for the all-to-one setting as shown in Tab. 4, all the attack methods achieve high attack performance, meanwhile Impart also achieves remarkable overall advantages on image quality compared with baseline methods as shown in Tab. 5. It is noted that since the poison images of baselines remain the same for both all-to-all and all-to-one settings, their image qualities do not change. However, the poison images of Impart rely on the label, and so the qualities under different settings are different. Tab. 4 shows that ASR of Impart is comparable but slightly worse than baselines, especially Bpp whose poison ratio is 20%. We recognize that as introduced in Sec. 1, the task of the all-to-one setting is simpler than that of the all-to-all because it only addresses one mapping from triggers to the target label. All methods including Impart can handle it well. However, since Impart has more strict requirements on image imperceptibility, it scarifies the effectiveness slightly. Thus, Impart can still achieve a comparable performance even if it is designed for the all-to-all setting.

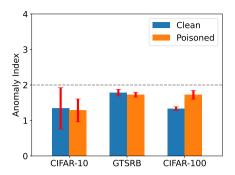


Figure 3: Effectiveness of our method under the Neural Cleanse defense on CIFAR-10, GTSRB, and CIFAR-100 datasets.

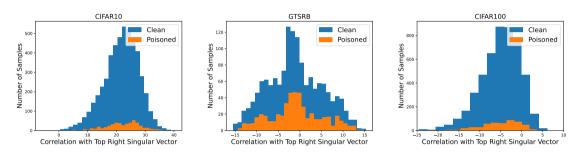


Figure 4: Resilience to Spectral Signatures.

# **5.3.** Defense Experiments

In this section, we investigate whether our method can bypass existing state-of-the-art defense methods: Neural Cleanse (Wang, Yao, Shan, Li, Viswanath, Zheng and Zhao, 2019), STRIP (Gao, Xu, Wang, Chen, Ranasinghe and Nepal, 2019), Neural Attention Distillation (Li, Lyu, Koren, Lyu, Li and Ma, 2021b), Spectral Signatures (Tran, Li and Madry, 2018) and GradCam (Selvaraju, Cogswell, Das, Vedantam, Parikh and Batra, 2017). We select these defense methods because they are widely used in previous works (Doan et al., 2021b,a). Besides, they evaluate the backdoor attack method from different aspects. We conduct the defense methods in the all-to-all setting with GoogleNet as the surrogate model. Additionally, all of the defense methods are conducted in the default hyperparameters declared in their original papers.

**Neural Cleanse.** Neural Cleanse (NC) (Wang et al., 2019) is a typical trigger synthesis-based empirical defense method, which reconstructs the triggers through reverse engineering. It firstly treats each class label as a potential target label, then reconstructs a "minima" trigger that can produce model misclassification by a designed optimization. Finally, it runs an outlier detection to find the anomaly based on the assumption that the infected label requires much smaller modifications to cause misclassification than other uninfected labels.

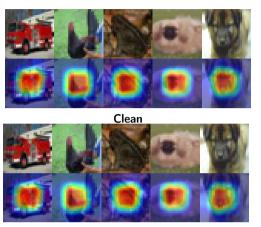
Fig. 3 shows the resistance results to NC. It shows that the anomaly index of our method computed by NC is lower than the threshold value 2. It states Impart passes the defense on all datasets. While previous works WaNet and Bpp also pass against NC defense, they implement negative (noise) sample (i.e., 20% negative sample in WaNet and Bpp) as auxiliary which significantly reduces the quality of the dataset. As shown in Tab. 5, PSNR of WaNet is 29.61dB and Bpp is 36.65dB on average. On the contrary, Impart is robust against NC just because the generated perturbation is minor and uniform for all classes.

**STRIP.** We then select a representative sample filtering-based empirical defense STRIP (Gao et al., 2019). It determines the presence of the backdoor by calculating the entropy predicted by the model before and after perturbing the inputs intentionally. The entropy of the input below 0.2 states this image is a poisoned sample. We apply this method and calculate the minimal entropy with respect to three datasets individually, which are 1.88, 9.39, and 35.09, respectively. As all of them are far larger than 0.2, this shows Impart is resistant to STRIP. Besides, Fig. 2 shows that the entropy

Table 6

Effectiveness of our method under NAD on three datasets.

Detects	No D	efense	NAD		
Datasets	BA(%)	ASR(%)	BA(%)	ASR(%)	
CIFAR-10	94.24	96.35	35.20	7.30	
GTSRB	98.03	98.90	18.30	8.23	
CIFAR-100	75.66	77.07	4.86	0.75	



Poisoned

Table 7
Resilience to GradCam. Comparisons of GradCam visualization maps between clean data ( $1^{st}$  row) and poisoned data ( $2^{nd}$  row) on a poisoned model.

range is similar between the benign model and the poisoned model, and the entropy range of the poisoned model includes the entropy range of the benign model. For the result of the GTSRB dataset, there are some separate in our view, but for a defender who doesn't know the existence of poison data, he can't distinguish there are two different distributions. These results indicate our method can successfully bypass the STRIP defense method.

**Neural Attention Distillation.** Neural Attention Distillation(NAD) (Li et al., 2021b) is a state-of-the-art model reconstruction-based empirical defense. NAD erases the backdoor through a twice-finetuning process. It firstly gets a teacher model by finetuning the victim model on a small clean subset, then utilizing the teacher model to guide the second finetuning aiming to the selected layer attention distribution of the student model align to the teacher model. As shown in Tab. 6, after the distillation, BA decreases dramatically. This means that NAD is ineffective against our method Impart. We speculate it may be because the generated poisoning images of our method is aligned with the target class in the image feature, the NAD can't distinguish the poison data and clean data.

**Spectral Signatures.** Spectral Signatures (Tran et al., 2018) is a representative defensive approach that inspects the latent space of the model. It first finds the top-right singular vector of the covariance matrix of the latent vectors using a small subset of clean samples. Then it calculates the correlation of each sample to this singular vector. Those with the outlier scores are identified as backdoor samples. Fig. 4 shows our method can bypass it as the distributions are similar. This is because we use a clean surrogate model to generate poison examples that are related to the target label, and so the latent space representation is also imparted to the victim model by the poison examples.

**GradCam Visualization.** GradCam (Selvaraju et al., 2017) is a typical network visualization method. It reveals the model behaviors by a heat map of the input image. The heat map denotes the attention when the model recognizes the input image. Following the previous work (Nguyen and Tran, 2021), we compare the attention discrepancy of the poisoned model between benign data and poisoned data. The inputs use the ground truth label for benign data while the target label is for poisoned data. We use the existing package pytorch-grad-cam (Gildenblat and contributors, 2021) to evaluate our method. Tab. 7 shows that the heat maps produced by GradCam are nearly identical between benign data and poisoned data.

Table 8

The influence of different  $\gamma$ .  $\uparrow$  indicates larger values are preferred, and vise versa.

γ Metrics	100	50	30	10	0
BA(%)	98.96	96.96	95.9	98.03	98.75
ASR(%)	55.76	85.35	95.92	98.90	98.51
CIEDE2000↓	4.31	11.52	18.47	28.56	29.85
PSNR(dB)↑	125.59	84.32	64.55	47.47	40.23
SSIM(×100) ↑	99.76	99.44	99.16	98.54	98.10
$\ell_2\downarrow$	0.06	0.16	0.28	0.52	0.69
$\ell_\infty\downarrow$	0.01	0.03	0.05	0.09	0.14

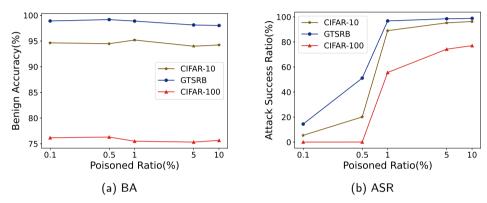


Figure 5: The influence of different poisoned ratio. The left (a) is the influence of poisoned ratio on BA. The right (b) is the influence of poisoned ratio on ASR.

#### **5.4.** Ablation Studies

We investigate the influence of hyperparameters  $\rho$  and  $\gamma$ . All experiments are conducted in the all-to-all setting. **Poisoned Ratio**  $\rho$ . We evaluate the influence of poisoned ratio  $\rho$  following the setting of previous work (Wu, Chen, Zhang, Zhu, Wei, Yuan and Shen, 2022) that recorded the BA and ASR with poisoned ratios 10%, 5%, 1%, 0.5%, 0.1% for all three datasets. Fig. 5(a) reveals there has been a steady trend for BA in all datasets. Additionally, Fig. 5(b) shows there has been a sharp drop for ASR in a range from 1% to 0.1% but a slight decline in a range from 10% to 1%. What is striking in this figure is that our method still achieves high ASR only with a 1% poisoned ratio. This indeed declared the effectiveness of Impart.

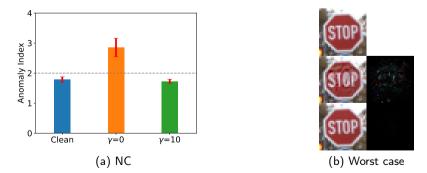


Figure 6: Effectiveness of  $\ell_2$  regularization. (a) The effectiveness of  $\gamma$  resistance to NC defense. (b) First row:clean image; second row:  $\gamma = 0$ ; third row:  $\gamma = 10$ .

Table 9

Attack	effectiveness	in	the.	Tiny	Imagenet	datacet
ALLACK	enectiveness	ın	ine	I IIIV	imagenei	datase

Dataset Mode	Modo	Impart		LII	LIRA		WB	
	ВА	ASR	BA	ASR	_	ВА	ASR	
T-Imagenet	All-to-All	59.28	74.09	58.00	59.00		58.00	58.00
T-Imagenet	All-to-One	62.13	99.21	58.00	100		57.00	99.00

Table 10

Attack effectiveness of the adversarial perturbation in the black-box scenario.

	All-	to-All	Al	l-to-One
	BA(%)	ASR(%)	BA(%	) ASR(%)
CIFAR-10	94.80	14.30	94.80	17.36
GTSRB	99.23	9.56	99.23	6.46
CIFAR-100	76.18	2.86	76.18	1.50

**Hyperparameter**  $\gamma$ . To investigate the effectiveness of hyperparameter  $\gamma$ , we conduct comprehensive experiments with five different  $\gamma = 0, 10, 30, 50$  and 100 in the GTSRB dataset. The iteration number of generating adversarial examples is set to 200. Tab. 8 shows the correlation among  $\gamma$ , image quality metrics, ASR and BA. It shows that both the image quality and the ASR drop with the increase of  $\gamma$ . Besides, the BA is relatively steady alongside the increase of  $\gamma$ . Additionally, we can observe from Tab. 8 that the larger the  $\gamma$ , the better the image quality. In other words, we can infer that the larger the  $\gamma$ , the less iteration time to achieve identical image quality.

Furthermore, Fig. 6(a) shows the effectiveness of  $\gamma$  resistance to NC. Fig. 6 (b) shows the influence of image quality with and without using  $\ell_2$  in the worst case. We observe a difference between the clean image and the poisoned image when  $\gamma = 0$  but they are similar when  $\gamma = 10$ .

# 5.5. Discussion

Compared with White-Box methods in Tiny-Imagenet. To further demonstrate the attack effectiveness of our method, we compare our method Impart with the White-Box method LIRA (Doan et al., 2021b) and WB (Doan et al., 2021a) in the Tiny-Imagenet dataset. Besides, because the generalization error in the Tiny-Imagenet dataset is too large to regard as a benign model(i.e., train accuracy: 99.28% and test accuracy: 59.12%), so we decide to list the results here as a case study. As for the victim model(VM), we use ResNet18. For the surrogate model(SM) used to fit the image feature, we select EfficientNetB0 since it is more effective to fit the image feature shown in the previous experiments. The results are shown in Tab. 9, which declares that our method can achieve about 74% ASR in the all-to-all setting but the white-box method LIRA can only achieve 59%. This indeed demonstrates the effectiveness of our method.

Adversarial Attack or Backdoor Attack? One may argue that it is unclear whether the high ASR is due to the backdoor effect of the trigger or the adversarial perturbation. Therefore, we test the adversarial perturbation's effect in the black-box scenario. Specifically, as the standard adversarial attack process, we first use a surrogate model to generate the poison data, then directly feed the poisoned data to the VM (i.e., trained by the clean dataset) in the inference phase to test the ASR. As for the VM, it is set identically with the Sec. 5.1. For the surrogate model(SM) used to fit the image feature, we select EfficientNetB0 since it is more effective to fit the image feature shown in the previous experiments. The results are shown in Tab. 10, which declares that the adversarial perturbation is completely ineffective to attack the VM in the black-box scenario. Therefore, we can conclude that the high ASR is indeed due to the backdoor effect of the trigger.

The Selection Strategy of Surrogate Model. In the paper, we use the existing model as the surrogate model to test the effectiveness of the Impart. Our method requires generating the perturbation related to the target label, thus in order to achieve a high attack success rate 1) the model needs to fit as correctly as possible about the target label, or 2) the model needs to have a similar understanding with the victim model. Therefore, it is better to select the model either 1) the model which has higher accuracy in large-scale and realistic datasets (e.t. ImageNet) or 2) the model which has a more similar structure to the victim model.

# 6. Conclusion

In this study, we propose a novel backdoor attack framework, Impart, that can simultaneously achieve strong attack ability and high imperceptibility without access to the victim model. Different from previous works which try to find an imperceptible pattern and add it to the source image as the poisoned image, we propose to generate perturbation that aligned with the target label in the image feature by a surrogate model. In this way, the generated poisoned images are attached with knowledge about the target class, which significantly enhances the attack capability. We evaluate our method on three benchmark datasets and five typical defense mechanisms. Experiments show that Impart can achieve state-of-the-art attack success rates in the all-to-all setting, meanwhile, it maintains high visual quality.

# Acknowledgements

This work is supported by the National Natural Science Foundation of China (62202329).

# References

Barni, M., Kallas, K., Tondi, B., 2019. A new backdoor attack in cnns by training set corruption without label poisoning, in: 2019 IEEE International Conference on Image Processing (ICIP), IEEE. pp. 101–105.

Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E., 2008. Fast unfolding of communities in large networks. J. Stat. Mech.-Theory Exp. 2008, P10008.

Bouville, M., 2008. Crime and punishment in scientific research. arXiv:0803.4058.

Carlini, N., Wagner, D., 2017. Towards evaluating the robustness of neural networks, in: 2017 ieee symposium on security and privacy (sp), Ieee. pp. 39–57.

Chen, Q., Wu, T.T., Fang, M., 2013. Detecting local community structure in complex networks based on local degree central nodes. Physica A. 392, 529–537.

Chen, X., Liu, C., Li, B., Lu, K., Song, D., 2017. Targeted backdoor attacks on deep learning systems using data poisoning. arXiv preprint arXiv:1712.05526.

Cheng, S., Liu, Y., Ma, S., Zhang, X., 2021. Deep feature space trojan attack of neural networks by controlled detoxification, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 1148–1156.

Cheng, Z., Wu, B., Zhang, Z., Zhao, J., 2023. Tat: Targeted backdoor attacks against visual object tracking. Pattern Recognition, 109629URL: https://www.sciencedirect.com/science/article/pii/S0031320323003308, doi:https://doi.org/10.1016/j.patcog.2023.109629.

Clancey, W.J., 1979. Transfer of Rule-Based Expertise through a Tutorial Dialogue. Ph.D. diss.. Dept. of Computer Science, Stanford Univ.. Stanford, Calif.

Clancey, W.J., 1983. Communication, Simulation, and Intelligent Agents: Implications of Personal Intelligent Machines for Medical Education, in: Proceedings of the Eighth International Joint Conference on Artificial Intelligence (IJCAI-83), IJCAI Organization, Menlo Park, Calif. pp. 556–560

Clancey, W.J., 1984. Classification Problem Solving, in: Proceedings of the Fourth National Conference on Artificial Intelligence, AAAI Press, Menlo Park, Calif.. pp. 45–54.

Clancey, W.J., 2021. The Engineering of Qualitative Models. Forthcoming.

Clauset, A., Newman, M.E.J., Moore, C., 2004. Finding community structure in very large networks. Phys. Rev. E. 70, 066111.

Danon, L., Diaz-Guilera, A., Duch, J., Arenas, A., 2005. Comparing community structure identification. J. Stat. Mech.-Theory Exp., P09008.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota. pp. 4171–4186. URL: https://aclanthology.org/N19-1423, doi:10.18653/v1/N19-1423.

Doan, K., Lao, Y., Li, P., 2021a. Backdoor attack with imperceptible input and latent modification. Advances in Neural Information Processing Systems 34, 18944–18957.

Doan, K., Lao, Y., Zhao, W., Li, P., 2021b. Lira: Learnable, imperceptible and robust backdoor attacks, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11966–11976.

Doan, K.D., Lao, Y., Li, P., 2022. Marksman backdoor: Backdoor attacks with arbitrary target class. arXiv preprint arXiv:2210.09194.

Eckert, M.P., Bradley, A.P., 1998. Perceptual quality metrics applied to still image compression. Signal processing 70, 177-200.

Engelmore, R., Morgan, A. (Eds.), 1986. Blackboard Systems. Addison-Wesley, Reading, Mass.

Fabio, D.R., Fabio, D., Carlo, P., 2013. Profiling core-periphery network structure by random walkers. Sci. Rep. 3, 1467.

Fabricio, B., Liang, Z., 2013. Fuzzy community structure detection by particle competition and cooperation. Soft Comput. 17, 659-673.

Fortunato, S., 2010. Community detection in graphs. Phys. Rep.-Rev. Sec. Phys. Lett. 486, 75-174.

Fortunato, S., Barthelemy, M., 2007. Resolution limit in community detection. Proc. Natl. Acad. Sci. U. S. A. 104, 36-41.

Gao, Y., Xu, C., Wang, D., Chen, S., Ranasinghe, D.C., Nepal, S., 2019. Strip: A defence against trojan attacks on deep neural networks, in: Proceedings of the 35th Annual Computer Security Applications Conference, pp. 113–125.

Gildenblat, J., contributors, 2021. Pytorch library for cam methods. https://github.com/jacobgil/pytorch-grad-cam.

Goodfellow, I.J., Shlens, J., Szegedy, C., 2014. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

- Gregory, S., 2011. Fuzzy overlapping communities in networks, J. Stat. Mech.-Theory Exp., P02017.
- Gu, T., Dolan-Gavitt, B., Garg, S., 2017. Badnets: Identifying vulnerabilities in the machine learning model supply chain. arXiv preprint arXiv:1708.06733.
- Hasling, D.W., Clancey, W.J., Rennels, G., 1984. Strategic explanations for a diagnostic consultation system. International Journal of Man-Machine Studies 20, 3–19. URL: https://www.sciencedirect.com/science/article/pii/S0020737384800036, doi:https://doi.org/10.1016/S0020-7373(84)80003-6.
- Hasling, D.W., Clancey, W.J., Rennels, G.R., Test, T., 1983. Strategic Explanations in Consultation—Duplicate. The International Journal of Man-Machine Studies 20, 3–19.
- Havens, T.C., Bezdek, J.C., Leckie, C., R.K., Palaniswami, M., 2013. A soft modularity function for detecting fuzzy communities in social networks. IEEE Trans. Fuzzy Syst. 21, 1170–1175.
- He, K., Zhang, X., Ren, S., Sun, J., 2016a. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.
- He, K., Zhang, X., Ren, S., Sun, J., 2016b. Identity mappings in deep residual networks, in: European conference on computer vision, Springer. pp. 630–645
- Hullermeier, E., Rifqi, M., 2009. A fuzzy variant of the rand index for comparing clustering structures, in: in Proc. IFSA/EUSFLAT Conf., pp. 1294–1298.
- Huvnh-Thu, O., Ghanbari, M., 2008, Scope of validity of psnr in image/video quality assessment, Electronics letters 44, 800-801.
- Krizhevsky, A., Hinton, G., et al., 2009. Learning multiple layers of features from tiny images .
- Lancichinetti, A., Fortunato, S., 2009. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. Phys. Rev. E. 80, 016118.
- Lancichinetti, A., Fortunato, S., Radicchi, F., 2008. Benchmark graphs for testing community detection algorithms. Phys. Rev. E. 78, 046110.
- Li, J., Wang, X., Eustace, J., 2013. Detecting overlapping communities by seed community in weighted complex networks. Physica A. 392, 6125–6134.
- Li, Y., Jiang, Y., Li, Z., Xia, S.T., 2022. Backdoor learning: A survey. IEEE Transactions on Neural Networks and Learning Systems .
- Li, Y., Li, Y., Wu, B., Li, L., He, R., Lyu, S., 2021a. Invisible backdoor attack with sample-specific triggers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 16463–16472.
- Li, Y., Lyu, X., Koren, N., Lyu, L., Li, B., Ma, X., 2021b. Neural attention distillation: Erasing backdoor triggers from deep neural networks, in: International Conference on Learning Representations. URL: https://openreview.net/forum?id=910K40M-oXE.
- Liu, J., 2010. Fuzzy modularity and fuzzy community structure in networks. Eur. Phys. J. B. 77, 547-557.
- Liu, W., Pellegrini, M., Wang, X., 2014. Detecting communities based on network topology. Sci. Rep. 4, 5739.
- Liu, Y., Ma, S., Aafer, Y., Lee, W.C., Zhai, J., Wang, W., Zhang, X., 2018. Trojaning attack on neural networks, in: NDSS.
- Liu, Y., Ma, X., Bailey, J., Lu, F., 2020. Reflection backdoor: A natural backdoor attack on deep neural networks, in: European Conference on Computer Vision, Springer. pp. 182–199.
- Lou, H., Li, S., Zhao, Y., 2013. Detecting community structure using label propagation with weighted coherent neighborhood propinquity. Physica A. 392, 3095–3105.
- Luo, M.R., Cui, G., Rigg, B., 2001. The development of the cie 2000 colour-difference formula: Ciede2000. Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur 26, 340–350.
- Ma, Q., Jiang, L., Yu, W., 2023. Lambertian-based adversarial attacks on deep-learning-based underwater side-scan sonar image classification. Pattern Recognition 138, 109363. URL: https://www.sciencedirect.com/science/article/pii/S003132032300064X, doi:https://doi.org/10.1016/j.patcog.2023.109363.
- Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P., 2016. Deepfool: a simple and accurate method to fool deep neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2574–2582.
- NASA, 2015. Pluto: The 'other' red planet. https://www.nasa.gov/nh/pluto-the-other-red-planet. Accessed: 2018-12-06.
- Nepusz, T., Petróczi, A., Négyessy, L., Bazsó, F., 2008. Fuzzy communities and the concept of bridgeness in complex networks. Phys. Rev. E. 77, 016107.
- Newman, M.E.J., 2013. Network data. http://www-personal.umich.edu/~mejn/netdata/.
- Newman, M.E.J., Girvan, M., 2004. Finding and evaluating community structure in networks. Phys. Rev. E. 69, 026113.
- Nguyen, T.A., Tran, A., 2020. Input-aware dynamic backdoor attack. Advances in Neural Information Processing Systems 33, 3454-3464.
- Nguyen, T.A., Tran, A.T., 2021. Wanet imperceptible warping-based backdoor attack, in: International Conference on Learning Representations. URL: https://openreview.net/forum?id=eEn8KTtJOx.
- Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z.B., Swami, A., 2016. The limitations of deep learning in adversarial settings, in: 2016 IEEE European symposium on security and privacy (EuroS&P), IEEE. pp. 372–387.
- Pei, K., Cao, Y., Yang, J., Jana, S., 2017. Deepxplore: Automated whitebox testing of deep learning systems, in: proceedings of the 26th Symposium on Operating Systems Principles, pp. 1–18.
- Psorakis, I., Roberts, S., Ebden, M., Sheldon, B., 2011. Overlapping community detection using bayesian non-negative matrix factorization. Phys. Rev. E. 83, 066114.
- Raghavan, U., Albert, R., Kumara, S., 2007. Near linear time algorithm to detect community structures in large-scale networks. Phys. Rev E. 76, 036106.
- Rice, J., 1986. Poligon: A System for Parallel Problem Solving. Technical Report KSL-86-19. Dept. of Computer Science, Stanford Univ.
- Robinson, A.L., 1980a. New ways to make microcircuits smaller. Science 208, 1019-1022. URL: https://science.sciencemag.org/content/208/4447/1019, doi:10.1126/science.208.4447.1019,

- arXiv:https://science.sciencemag.org/content/208/4447/1019.full.pdf.
- Robinson, A.L., 1980b. New Ways to Make Microcircuits Smaller—Duplicate Entry. Science 208, 1019-1026.
- Rony, J., Hafemann, L.G., Oliveira, L.S., Ayed, I.B., Sabourin, R., Granger, E., 2019. Decoupling direction and norm for efficient gradient-based 12 adversarial attacks and defenses, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4322–4330.
- Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE international conference on computer vision, pp. 618–626.
- Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al., 2016. Mastering the game of go with deep neural networks and tree search. nature 529, 484–489.
- Sobolevsky, S., Campari, R., 2014. General optimization technique for high-quality community detection in complex networks. Phys. Rev. E. 90, 012811
- Stallkamp, J., Schlipsing, M., Salmen, J., Igel, C., 2012. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. Neural networks 32, 323–332.
- Sun, P., Gao, L., Han, S., 2011. Identification of overlapping and non-overlapping community structure by fuzzy clustering in complex networks. Inf. Sci. 181, 1060–1071.
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1–9.
- Tan, M., Le, Q., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks, in: International conference on machine learning, PMLR. pp. 6105–6114.
- Tan, M., Le, Q., 2021. Efficientnetv2: Smaller models and faster training, in: International Conference on Machine Learning, PMLR. pp. 10096–10106.
- Tran, B., Li, J., Madry, A., 2018. Spectral signatures in backdoor attacks. Advances in neural information processing systems 31.
- Vehlow, C., Reinhardt, T., Weiskopf, D., 2013. Visualizing fuzzy overlapping communities in networks. IEEE Trans. Vis. Comput. Graph. 19, 2486–2495.
- Šubelj, L., Bajec, M., 2011a. Robust network community detection using balanced propagation. Eur. Phys. J. B. 81, 353-362.
- Šubelj, L., Bajec, M., 2011b. Unfolding communities in large complex networks: Combining defensive and offensive label propagation for core extraction. Phys. Rev. E. 83, 036103.
- Šubelj, L., Bajec, M., 2012. Ubiquitousness of link-density and link-pattern communities in real-world networks. Eur. Phys. J. B. 85, 1-11.
- Wang, B., Yao, Y., Shan, S., Li, H., Viswanath, B., Zheng, H., Zhao, B.Y., 2019. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks, in: 2019 IEEE Symposium on Security and Privacy (SP), IEEE. pp. 707–723.
- Wang, W., Liu, D., Liu, X., Pan, L., 2013. Fuzzy overlapping community detection based on local random walk and multidimensional scaling. Physica A. 392, 6578–6586.
- Wang, X., Li, J., 2013. Detecting communities by the core-vertex and intimate degree in complex networks. Physica A. 392, 2555–2563.
- Wang, Y., Ding, X., Yang, Y., Ding, L., Ward, R., Wang, Z.J., 2021. Perception matters: Exploring imperceptible and transferable anti-forensics for gan-generated fake face imagery detection. Pattern Recognition Letters 146, 15–22.
- Wang, Y., Wang, L., Feng, M., Ward, R., Wang, Z.J., 2022a. Reaching a better trade-off between image quality and attack success rates in transfer-based adversarial attacks, in: 2022 IEEE Data Science and Learning Workshop (DSLW), IEEE. pp. 1–6.
- Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing 13, 600–612.
- Wang, Z., Zhai, J., Ma, S., 2022b. Bppattack: Stealthy and efficient trojan attacks against deep neural networks via image quantization and contrastive adversarial learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15074–15084.
- Wu, B., Chen, H., Zhang, M., Zhu, Z., Wei, S., Yuan, D., Shen, C., 2022. Backdoorbench: A comprehensive benchmark of backdoor learning. Advances in Neural Information Processing Systems 35, 10546–10559.
- Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., Yu, D., Zweig, G., 2016. Achieving human parity in conversational speech recognition. arXiv preprint arXiv:1610.05256.
- Yuan, X., He, P., Zhu, Q., Li, X., 2019. Adversarial examples: Attacks and defenses for deep learning. IEEE transactions on neural networks and learning systems 30, 2805–2824.
- Zhang, J., Li, C., 2019. Adversarial examples: Opportunities and challenges. IEEE transactions on neural networks and learning systems 31, 2578–2593
- Zhang, Q., Ding, Y., Tian, Y., Guo, J., Yuan, M., Jiang, Y., 2021. Advdoor: adversarial backdoor attack of deep learning system, in: Proceedings of the 30th ACM SIGSOFT International Symposium on Software Testing and Analysis, pp. 127–138.
- Zhang, S., Wang, R., Zhang, X., 2007. Identification of overlapping community structure in complex networks using fuzzy c-means clustering. Physica A. 374, 483–490.
- Zhang, Y., Yeung, D., 2012. Overlapping community detection via bounded nonnegative matrix tri-factorization, in: In Proc. ACM SIGKDD Conf., pp. 606–614.
- Zhao, Z., Liu, Z., Larson, M., 2020. Towards large yet imperceptible adversarial image perturbations with perceptual color distance, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1039–1048.