通过投毒网络上的野外图像对半监督学习模型实施木马攻击

Le Feng 复
旦大学 中国

Zhengxing Qian
复旦大学 中国

Sheng Li
复旦大学 中国

Xinpeng Zhang
复旦大学 中国

arXiv:2301.00435v1 [cs.CY] 1 Jan 2023

摘要

网络上的野外图像容易遭受后门（也称木马）投毒，导致基于这些图像训练的机器学习模型被植入后门。以往的大多数攻击都假设野外图像带有标签。然而，实际上网络上的大多数图像都没有标签。具体而言，我们研究了半监督学习（SSL）下无标签后门图像对广泛研究的深度神经网络的影响。为贴近现实，我们假设攻击者处于零知识状态，且半监督学习模型从头开始训练。首先，我们发现，当中毒的无标签图像来自不同类别时，后门投毒总是会失败，这与对带标签图像进行投毒的情况不同。原因在于，半监督学习算法在训练过程中总是努力纠正这些图像。因此，对于无标签图像，我们对来自目标类别的图像实施后门投毒。随后，我们提出一种梯度匹配策略，用于制作中毒图像，使其在半监督学习模型上的梯度与目标图像的梯度相匹配，从而使中毒图像归入目标类别并实现后门植入。据我们所知，这可能是首个针对从头训练的半监督学习模型中的无标签图像实施后门投毒的方法。实验表明，我们的投毒方法在绕过现代后门防御方法的同时，在大多数半监督学习算法上实现了最先进的攻击成功率。

关键词

后门，半监督学习，从头训练，神经网络

1 引言

深度神经网络的出色性能 [12, 31, 37, 49]很大程度上归功于大量训练样本。为了获得足够的训练样本，训练者通常会从网络上抓取这些样本。然而，这些来自野外的样本可能并不安全，容易受到后门投毒的影响。以往的后门投毒 [10, 21, 28,39, 45] 主要关注有标签样本，依靠目标标签的引导将后门注入模型。Yan 等人 [45, 46] 首次针对预训练半监督学习模型提出了两种无标签后门投毒方案：DeNeB [45] 和 DeHiB [46]。他们假设半监督学习者首先在有标签

样本。随后，使用半监督学习算法结合无标签样本对所得的预训练模型进行微调。实际上，大多数先进的半监督学习算法都是端到端的，即将无标签样本与有标签样本一同输入模型，以便从头开始训练。不存在中间预训练模型。此外，Yan 等人提出的后门模式可以被感知，因此可由 DePuD [46] 检测。

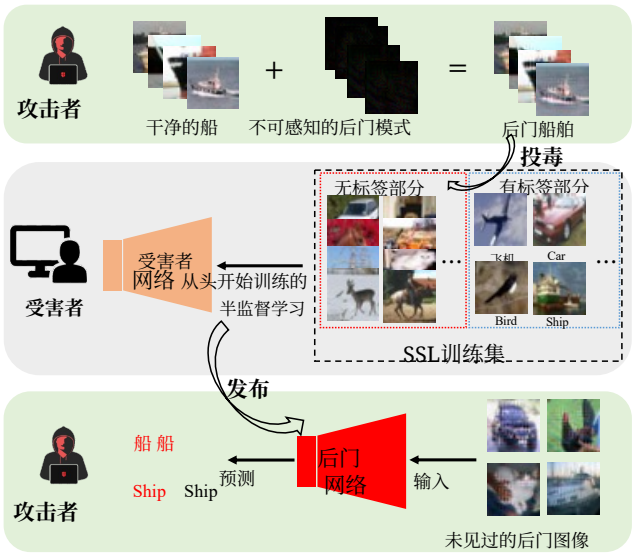


图 1：针对从头训练的半监督学习模型进行无标签后门投毒的攻击流程。假设目标类别是船。为兼顾实用性与隐蔽性，我们提出一种零 - 知识型目针对从头训练的半监督学习模型中无标签样本的不可感知后门投毒。零知识意味着攻击者无需了解受害模型、半监督学习算法、完整数据集和训练过程。我们的攻击流程如图 1 所示。攻击者将不可感知的后门模式添加到干净的船舶图像中，然后使用生成的后门图像对半监督学习训练集中的无标签部分进行投毒。受害者使用中毒训练集，通过半监督学习算法从头训练网络，从而使后门在无意中被植入。最后，在推理阶段，带有后门模式的图像会被误分类为目标类别“船”。

具体而言，首先我们发现，与对有标签样本进行投毒不同，由于半监督学习算法致力于正确学习无标签样本，如果投毒样本来自不同类别，即采用标签不一致后门投毒，那么从头训练的SSL模型会重新将它们学习到正确类别中。因此，无法植入后门。相反，如果仅对目标类别中的样本进行投毒，

即采用标签一致的后门投毒，那么半监督学习模型最终会将它们分类到目标类别中。此外，由于半监督学习算法的各种正则化能够缓解模型在目标类别上的过拟合，后门模式可以泛化到非目标类别。因此，我们仅对目标类别中的样本进行投毒。接下来，为了实现零知识投毒，我们借助神经网络的可迁移性。我们只需要知道目标类别以及受害者数据集的分布。具体而言，我们首先准备一个代理数据集，其分布与受害者数据集相似，并包含目标类别的样本。随后，在代理数据集上训练替代网络。对于训练得到的替代网络，我们训练一个后门模式生成器：它以来自不同类别的干净样本为输入，输出相应的不可感知后门模式，然后将这些后门模式添加到样本中，得到投毒样本。除了后门模式的不可感知性之外，生成器的另一项训练目标是实现梯度匹配；梯度匹配旨在使投毒样本在替代模型上的梯度与目标样本的梯度相匹配。我们希望在从头训练的SSL模型中，通过梯度匹配，投毒样本能够像目标样本被学习到目标类别那样，自然而然地被学习到目标类别，从而植入后门。总之，我们的贡献如下：

1. 据我们所知，我们是首个研究从头训练的半监督学习模型的无标签样本易受后门投毒攻击这一问题的团队。

2. 我们发现，对于从头训练的半监督学习模型的无标签样本，标签一致的后门投毒更为有效。

3. 我们提出了一种针对从头训练的半监督学习模型无标签样本的零知识且不可感知的后门投毒方法。

4. 我们针对三类半监督学习算法实施了投毒。攻击成功率显著高于基线投毒方法，并且我们的方法能够成功绕过包括 DePuD 在内的各种防御方法。

2 相关工作

2.1 后门投毒

如果投毒样本全部来自目标类别，则称为标签一致的后门投毒；否则，称为标签不一致后门投毒。

标签不一致后门投毒：BadNets [10]是首个针对神经网络的后门投毒方案。该方案较为初级，因此许多后门防御方法 [3, 8, 11, 22, 42] 都能检测或移除后门。许多后续研究提出了威胁更大的投毒方法。不可见后门模式 [21, 23, 28]使受害者难以通过视觉发现投毒样本的异常。动态后门模式 [27, 33] 使受害者难以捕捉后门模式的规律。还有一些方案 [6, 21] 同时实现了不可见和动态的后门模式。DeHiB [45] 和 DeNeB [46] 通过投毒预训练模型的无标签样本，同样属于标签不一致的方案。

标签一致的后门投毒：由于不改变投毒样本的正确标签，这种方法更加隐蔽。不过，后门模式可能会在目标类别上过拟合，无法泛化到非目标类别。CLB（干净标签后门）[39] 首次提出通过对抗性扰动和插值来缓解这种过拟合。后来， [48] 将这种后门投毒方法应用于视频。 [20] 将这种后门投毒方法应用于点云。

2.2 后门防御

针对后门投毒的不同阶段，后门防御可分为四种类型：训练前防御、训练后防御、测试时防御和盲防御。

训练前防御：防御者会检查训练样本，以确定是否存在可疑样本。对于监督学习中的标签不一致投毒，可以根据样本内容与其标签之间的不一致性筛查投毒样本。形式上，激活聚类 [3] 可以根据训练样本的标签对其进行聚类，从而检测离群样本。最近，有人提出了 DePuD [46]，用于检测半监督学习中的无标签投毒样本；该方法使用强正则化来区分可疑的无标签样本。

训练后防御：这种防御旨在检测已学习模型中的异常。一种典型的检测方法是 Neural Cleanse [42]。首先对所有类别进行逆向工程，以获取其触发器。如果某个类别的触发器强度明显小于其他类别，则将该类别检测为后门类别。此后，出现了许多基于 Neural Cleanse 的变体。例如， [4, 43] 通过改进目标函数来改进 Neural Cleanse。 [7, 11] 提出了适用于黑盒场景的检测方法。

测试时防御：这种防御在模型测试阶段部署，并对测试样本进行检查。一种典型的检测方法是 STRIP [8]。将测试样本与一组预先准备好的干净样本融合，以获得合成样本。然后将这些合成样本输入模型进行预测。如果预测结果呈现低熵分布，则该测试样本可能是后门样本。

盲防御：这种方法不检测样本或模型，而是对样本或模型采取统一的操作。数据增强是一种自然的盲防御方法。在测试阶段，对样本进行 JPEG压缩等处理也可能破坏后门模式。精细剪枝 [22] 通过剪枝并微调模型，尝试消除模型中可能存在的后门。

2.3 半监督学习

现有的半监督学习算法可分为三类：一致性正则化、伪标签，以及带一致性正则化的伪标签。

一致性正则化：它假设，对于相同的输入，神经网络内部的随机性或数据增强变换不应改变模型的预测结果。例如，PI-Model [30] 通过最小化同一样本经过网络两次随机变换后的输出差异来进行训练。MeanTeacher [38] 则通过最小化学生模型和教师模型对同一样本的预测差异来进行训练。

[17, 25, 30, 38, 41, 44], [15, 19, 29], [1, 2, 3]



图 2：监督学习和半监督学习中的后门投毒。在这些投毒中，目标类别是"船"。它们的投毒目标相同，都会导致后门网络在中毒训练集上训练后，将未见过的后门图像误分类为"船"。

VAT [25], 、ICT [41], 和 UDA [44] 旨在开发更高效的数据增强方法，以充分利用无标签数据。

伪标签： 它根据当前模型的预测为无标签样本分配伪标签，然后通过监督学习训练这些无标签样本。例如，伪标签方法 [19] 使用在有标签样本上训练的预训练网络来预测伪标签。MPL（元伪标签） [29] 维护两个模型：学生模型和教师模型。教师模型对无标签样本进行预测，以生成伪标签。

结合一致性正则化的伪标签： MixMatch[2] 使用 MixUp 数据增强为每个无标签样本生成多个增强版本，然后将这些增强版本的预测结果取平均，并将其中概率最大的类别作为伪标签。ReMixMatch [1] 通过引入两种新机制——分布对齐和增强锚定——改进了 MixMatch。FixMatch [35] 对每个无标签样本分别进行弱增强和强增强，并将弱增强的预测标签用作强增强样本的伪标签。

3 我们的新发现

在监督学习中，如图 2(a) 和图 2(b) 所示，标签一致和标签不一致的后门投毒都依赖目标标签的引导。两者的区别在于，标签不一致的后门投毒会将样本的标签改为目标标签，而标签一致的后门投毒则无须这样做。然而，这也意味着，由于非目标类别样本中没有添加后门模式，模型可能会对目标类别中的后门模式过拟合，导致后门模式在非目标类别上不起作用。尽管 CLB [39] 提出通过插值和对抗扰动来提高后门模式在非目标类别上的泛化能力，但其攻击成功率低于标签不一致的后门投毒。因此，标签不一致的后门投毒比标签一致的后门投毒更容易实施。

然而，在半监督学习的背景下，一方面，如图 2(c) 和 2(d) 所示，对未标注样本进行后门投毒会失去目标标签的指导。另一方面，半监督学习与监督学习机制上的差异带来了一项新发现：

由于半监督学习算法致力于通过各种正则化来正确学习无标签样本，因此，对于无标签样本而言，标签不一致的后门投毒比标签一致的后门投毒难以实施得多。

接下来，我们将通过实验验证这一发现。我们使用现有方案 BadNets [10], CLB [39], 和 DeNeB [46]，对从头训练的半监督学习模型中的无标签样本进行投毒。需要注意的是，尽管一些近期的后门投毒方案（例如不可见后门投毒 [21, 23, 28], ）更能抵御后门防御，但 BadNets 在攻击成功率方面仍然表现出色。同样，在预训练网络的背景下，DeNeB 的攻击成功率远高于 DeHiB。

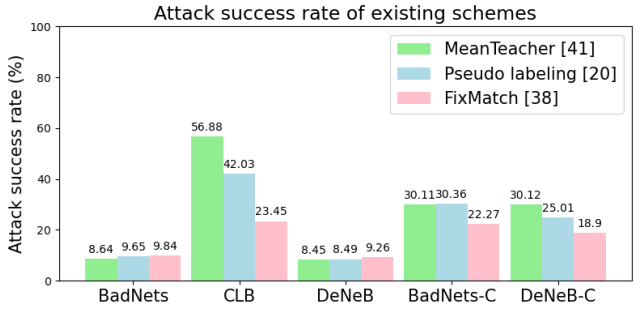


图 3：使用现有方案对从头训练的半监督学习模型中的无标签样本进行投毒。测试的受害者数据集为 CIFAR10 [16]，网络为 CNN13 [38]。在这三种后门投毒方案中，后门图案均为 8×8 像素方块。目标类别为 8。

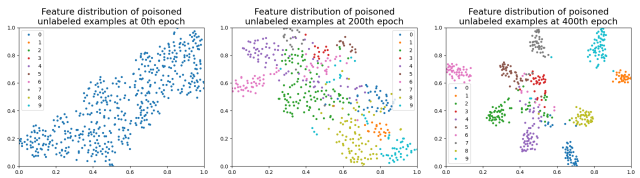
攻击成功率如图 3 所示。BadNets 和 DeNeB 均无法完全实施投毒，攻击成功率接近随机分类的概率 10.00%。相比之下，CLB 获得了一定的攻击成功率（56.88%、42.03%、23.45%）。具体而言，如图 2(c) 所示，对于标签不一致后门

投毒，即 BadNets 和 DeNeB，基于伪标签的半监督学习算法 [1, 2, 15, 19, 29, 35]力求为无标签样本分配正确标签，而来自不同类别的中毒未标注样本则期望自身被错误分类为目标类别。这种对立使得后门模式难以被学习。同样地，当采用基于一致性正则化的半监督学习算法 [1, 2, 17, 25, 30, 35, 38, 41, 44] 时，半监督学习算法添加到样本或模型中的噪声或数据增强会使模型遗忘后门模式，转而关注中毒未标注样本的语义信息。如图 4 所示，随着半监督学习的进行，中毒未标注样本会逐渐被分类到各自正确的类别中。然而，对于标签一致的后门投毒，即 CLB，由于投毒仅施加于来自目标类别的样本（图 2(d) ），半监督学习算法会将它们全部分类为目标类别。这种对立并不存在。此外，半监督学习算法的各种正则化会防止模型在目标类别上过拟合，因此后门模式可以略微泛化到非目标类别。

为进一步验证我们的发现，我们将 DeNeB 和 BadNets 推广为标签一致版本 DeNeB-C 和 BadNets-C，其中 C 表示一致。在所有设置保持不变的情况下，如图 3 所示，攻击成功率显著提高，例如，DeNeB 的攻击成功率从 8.45%、8.49%、9.26% 提高到 30.12%、25.01%、18.9%。然而，CLB、DeNeB-C 和 BadNets-C 存在三个显著缺点。

(1) 攻击成功率并不理想，最高仅为 56.88%。
(2) 后门模式很容易被察觉，受害者会将其识别为可疑内容，如图 6(b)、6(c) 和 6(d) 所示。
(3) DePuD [46], 一种针对中毒未标注样本的检测方案，可检测异常。

为弥补这些不足，我们提出了一种针对从头训练的半监督学习模型中无标签样本的零知识且不可感知的后门投毒方法。



(a) (b) (c)

图 4：使用标签不一致后门投毒 DeNeB 时，从头训练的半监督学习模型中中毒无标签样本的 t-SNE [40] 特征分布。半监督学习算法为 FixMatch。

4 我们的方法

4.1 威胁模型

假设训练神经网络模型的受害者只有有限的有标签样本。为了提升模型性能，他打算从网络上抓取更多无标签样本，用于半监督学习。例如，最先进的图像分类器 [24] 会从 Instagram 抓取 10 亿张图像。此时，能够向网络上传数据的攻击者可以控制一部分无标签样本，从而实施后门投毒。

由于我们的攻击是零知识攻击，攻击者掌握的信息非常有限。具体来说，攻击者无法获得以下信息：

(1) 从头训练的受害模型的架构、权重和输出。

(2) 训练过程、超参数设置以及所采用的半监督学习算法。

(3) 完整的受害者数据集，以及其中的样本是否带有标签。

攻击者唯一能够获取的信息是：

(1) 受害者数据集的分布 Z，以及投毒的目标类别 y^t。

具体而言，对于参数为 θ的受害模型 F，半监督学习的训练集由有标签部分 $\mathrm{X}=\{(x_n,\ y_n):\ n\in(1,\ldots N)\}$和无标签部分 $\mathrm{U}=\{u_n:\ n\in(1,\ldots \mu_s N)\}$组成，其中 μ_s是一个用于确定 X 和 U 相对规模的超参数。令 $\mathrm{F}(x;\ \theta)$ 表示模型 F 对输入 x 所产生的预测类别分布。为方便起见，我们始终用 Γ 表示半监督学习算法，半监督学习过程可形式化为：

$$\arg\min_{\theta}\Gamma(\mathcal{F}(\mathcal{X}\cup\mathcal{U};\theta)). \tag{1}$$

威胁模型可形式化为公式2所列的双层问题。外层优化旨在实现两个目标：一是达成攻击者进行后门投毒的基本目标，即在不降低模型对未见样本 X_{val}的准确率的前提下，最大化带有后门模式的后门样本的攻击成功率；二是确保后门模式尽可能不易被察觉，以免引起受害者怀疑。内层优化则是受害者使用半监督学习算法 Γ 在中毒训练集上训练模型。

$$\begin{aligned}\min_{\mathcal{P}_{tr}(\cdot)}\ &\mathbb{E}_{(x,y)\in\mathcal{X}_{val}}(\ell(y,\mathcal{F}(x;\theta^*))+\ell(y^t,\mathcal{F}(\mathcal{P}_{val}(x);\theta^*))\\ &+\|\mathcal{P}_{val}(x)-x\|^2)\\ \text{s.t.}\quad &\theta^*=\arg\min_{\theta}\Gamma(\mathcal{F}(\mathcal{X}\cup(1-\mu_b)\mathcal{U}\cup\mathcal{P}_{tr}(\mu_b\mathcal{U});\ \theta))\end{aligned}, \tag{2}$$

其中，$\mathrm{P}_{tr}(\mu_b\mathrm{U})$表示训练集中无标签数据 μ_bU遭到投毒，μ_b 是一个用于确定后门投毒比例的超参数。$\mathrm{P}_{val}(x)$表示将后门模式添加到样本 x 中，从而得到后门样本。

4.2 实现$\mathcal{P}_{tr}$ 和$\mathcal{P}_{val}$

首先需要注意的是，公式2中的 $\mathrm{P}_{val}(x)$ 和 $\mathrm{P}_{tr}(x)$ 在 CLB 和 DeNeB 中有所不同。在 CLB 中，为了让模型更好地记住后门模式，添加后门模式时会使用插值或对抗扰动，使选定图像无法被正确分类。在 DeNeB 中，添加后门模式时，也会通过对抗扰动使选定图像的特征和分类结果接近目标类别。相比之下，在我们的投毒中，$\mathrm{P}_{val}(x)$ 和 $\mathrm{P}_{tr}(x)$ 相同，即仅通过向干净的无标签图像添加后门模式来获得中毒的无标签图像。我们的投毒着重于如何构造后门模式，使在中毒图像上训练的半监督学习模型能够记住这些模式。总体而言，我们的投毒可分为以下三个步骤：准备替代网络和数据集、构造后门模式，以及投毒无标签图像。

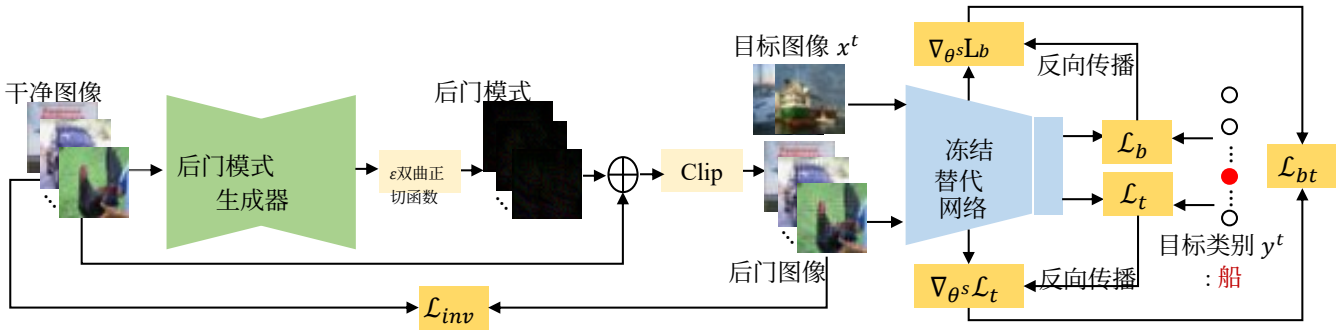


图 5：在替代网络和代理数据集上制作后门模式。

准备替代网络和数据集：由于攻击者只掌握目标类别 y^t 以及受害者数据集的分布 Z，因此借助神经网络的可迁移性，攻击者在替代数据集 X^s 和由 θ^s 参数化的替代网络 F^s 上构造后门模式。X^s 应包含目标类别 y^t 的图像，并符合分布 Z。F^s 应在 X^s 上达到较高的分类准确率，从而挖掘出尽可能难以察觉的后门模式。随后，在 X^s 上训练 F^s。为简便起见，下文仍以 θ^s 表示 F^s 学习到的参数。请注意，X^s 不必带标签，可以是无标签数据或未标记数据。这是因为 F^s 可以通过半监督学习 [29, 35] 或无监督学习 [5, 9], 进行训练，而这超出了我们的研究范围。

构造后门模式：此步骤基于代理数据集 X^s 和训练得到的替代网络 F^s 构造后门模式。正如第3节所述，为避免因半监督学习算法正确标注而导致后门投毒失败，投毒仅针对目标类别 y^t 中的图像实施。然而，如公式2所示，后门投毒的目标是：添加后门模式后，后门模型将来自不同类别的图像全部误分类为目标类别 y^t。因此，尽管只有目标类别的图像会被投毒，在构造后门模式时也需要考虑其他类别的图像。具体而言，此步骤包括两个方面的设计。

一是使后门模式不可感知。如图5所示，我们使用后门模式生成器生成原始后门模式。令由 ϑ 参数化的 G 表示该生成器。输入为干净图像 $x^s \in \mathrm{X}^s$，输出为与该图像对应的原始后门模式 $\mathrm{G}(x^s; \vartheta)$。随后，使用激活函数双曲正切函数将其限制在合理范围内，并乘以预算 ϵ，使生成器在给定预算内搜索不可感知的后门模式。最后，将得到的后门模式添加到图像中，并连接裁剪函数，使后门图像 x^b 处于正常范围内。为进一步确保后门模式不可感知，我们加入了式3中列出的损失函数 L_{ins}，使后门图像看起来更像干净图像。

$$L_{ins} = \mathbb{E}_{x^s \sim \mathcal{X}^s} \left\| x^b - x^s \right\|^2 \quad \text{s.t.} \quad \left\| x^b - x^s \right\|_\infty \leq \varepsilon \tag{3}$$

另一点是使后门图像 x^b 能够被从头训练的SSL模型 F 学习，并归入目标类别 y^t。为实现这一目标，我们提出了一种梯度匹配策略。首先，来看 F 如何学习目标类别 y^t 中的目标图像 x^t。无论 x^t 是否带有标签，也无论采用伪标签还是一致性正则化，学习目标 x^t 都是使其被分类为目标类别，这可以形式化表示为：

$$\theta_k = \theta_{k-1} - \eta \nabla \theta_{k-1} \ell(y^t, \mathcal{F}(x^t; \theta_{k-1})) \quad \text{for } k \in [1, max_k] \tag{4}$$

其中，θ_k 表示第 k 次迭代的权重，η 为学习率，max_k 表示迭代次数。构造后门图像 x^b 时，如公式5所示，我们希望半监督学习算法像将目标图像 x^t 拟合到目标类别 y^t 那样学习这些图像，从而诱使半监督学习算法植入后门。这意味着在每次迭代中，后门图像 x^b 在模型 F 上的梯度应与目标图像 x^t 在模型 F 上的梯度相匹配。

$$\begin{gathered} \theta_{k-1} - \eta \nabla_{\theta_{k-1}} \ell(y^t, \mathcal{F}(x^t; \theta_{k-1})) \approx \theta_{k-1} - \eta \nabla_{\theta_{k-1}} \ell(y^t, \mathcal{F}(x^b; \theta_{k-1})) \\ \rightarrow \nabla_{\theta_{k-1}} \ell(y^t, \mathcal{F}(x^t; \theta_{k-1})) \approx \nabla_{\theta_{k-1}} \ell(y^t, \mathcal{F}(x^b; \theta_{k-1})) \\ \text{for } k \in [1, max_k] \end{gathered}, \tag{5}$$

然而，由于我们的投毒是零知识的，因此无法获取模型 F 训练过程中的梯度信息。为了解决这一问题，我们考虑在替代网络 F^s 上模拟此类梯度信息。此外，我们并不需要所有迭代过程中的梯度信息。一方面，从头开始训练的模型 F 会产生大量梯度信息，而模拟所有这些信息的成本极高，也不切实际。另一方面，训练初期的梯度信息并不包含有意义的信息。因此，考虑到计算成本，我们只采用训练良好的替代网络的梯度信息。图7中的实验表明，我们的方法是有效的。

具体而言，首先，我们从目标类别中选取能够被替代网络 F^s 以较高置信度分类为目标类别的图像，作为目标图像 x^t，从而确保后门图像 x^b 的梯度能够与目标类别图像的梯度很好地匹配。然后，将目标图像 x^t 输入冻结的、训练良好的 F^s，并计算损失：

$$\mathcal{L}_t = \ell(y^t, \mathcal{F}^s(x^t; \theta^s)), \tag{6}$$

并计算损失 $\nabla_{\theta^s} \mathrm{L}_t$ 关于参数 θ^s 的梯度。同样，如图5所示，后门图像也被输入到 F^s 中，并

表1：投毒性能。SL CA 表示仅使用有标签样本通过监督学习训练得到的模型准确率。SSL CA 表示使用完整训练集（包括有标签样本和无标签样本）通过半监督学习训练得到的模型准确率。CA 表示投毒半监督学习模型的准确率。在 IMP 列中，数据从上到下依次为峰值信噪比、结构相似性指数和 L-∞ 范数。

数据集	半监督学习算法	监督学习准确率	SSL CA	BadNets-C [10]			DeNeB-C [46]			CLB [39]			Ours		
				CA	ASR	IMP	CA	ASR	IMP	CA	ASR	IMP	CA	ASR	IMP
CIFAR10	PseudoLabel [19]	78.86	89.18	89.21	30.36	20.20 0.8559 214.70	89.40	25.01	20.55 0.8211 208.93	89.18	42.03	19.98 0.8020 214.70	89.15	**91.17**	**31.34** **0.9515** **24.51**
	PI-Model [30]		87.21	86.88	60.38		87.26	48.32		87.26	68.56		87.05	**90.24**	
	MeanTeacher [38]		90.54	90.41	30.11		90.18	30.12		90.27	56.88		90.43	**88.70**	
	VAT [25]		87.33	87.29	59.33		87.31	46.21		87.12	67.93		86.90	**87.13**	
	ICT [41]		93.26	93.21	49.87		93.09	43.21		93.15	57.84		93.54	**94.54**	
	FixMatch [35]		93.56	93.39	22.27		93.48	18.90		93.65	23.45		93.83	**97.12**	
SVHN	PseudoLabel [19]	86.54	92.28	92.09	8.66	20.89 0.8361 195.22	92.06	8.29	20.59 0.7825 195.22	92.21	10.12	20.53 0.7741 195.22	92.16	**66.31**	**40.15** **0.9832** **19.62**
	PI-Model [30]		92.19	91.69	10.26		91.58	9.36		91.19	8.24		91.83	**75.78**	
	MeanTeacher [38]		93.52	93.19	8.64		93.15	8.45		93.18	6.58		93.40	**69.76**	
	VAT [25]		94.16	93.58	7.96		93.29	9.58		93.54	5.63		93.05	**28.71**	
	ICT [41]		95.62	95.26	8.26		95.34	6.98		95.41	9.26		95.21	**45.27**	
	FixMatch [35]		97.10	96.89	67.21		96.95	54.63		96.79	75.47		97.03	**79.59**	

得到损失 L_b。

$$\mathcal{L}_b = \ell(y^t, \mathcal{F}^s(x^b; \theta^s)), \tag{7}$$

并计算梯度 $\nabla_{\theta^s} L_b$。最后，优化 G，使 $\nabla_{\theta^s} L_b$ 接近 $\nabla_{\theta^s} L_t$，如公式 8 所示。

$$\mathcal{L}_{bt} = \|\nabla_{\theta^s}\mathcal{L}_t - \nabla_{\theta^s}\mathcal{L}_b\|^2. \tag{8}$$

用于制作后门模式的完整损失函数可表示为：

$$\begin{aligned} \mathcal{L}_{craft} &= \arg\min_{\vartheta} \mathbb{E}_{x^s \sim \mathcal{X}^s} \mathcal{L}_{bt} + \lambda_{ins}\mathcal{L}_{ins} \\ \text{s.t.} \quad & \left\| x^b - x^s \right\| \le \varepsilon \end{aligned}, \tag{9}$$

其中，λ_{ins} 是决定后门模式不可感知性的超参数。我们对 λ_{ins} 采用逐步增加的策略，即每50个训练轮次将 λ_{ins} 乘以2，从而找到尽可能不可感知的后门模式。

投毒未标记图像： 根据投毒比例μ_b，从目标类别的未标记图像中选取需要进行投毒的图像。然后，将这些图像输入经过学习的后门模式生成器G，以获得相应的后门模式，并将其添加到图像中，从而得到中毒图像。最后，将中毒图像发布到互联网，供受害者抓取，或秘密地重新注入受害者数据集U中。

5 实验评估

5.1 实验设置

5.1.1 受害者网络和数据集：我们在广泛用于半监督学习的 CIFAR10 [16] 和SVHN [26], 上实施投毒。CIFAR10包含来自10个类别的50,000张训练图像和10,000张测试图像。SVHN包含来自10个类别的73,257张训练图像和26,032张房屋门牌号图像。此外，CIFAR10在CNN13 [38] 上进行训练，SVHN在 WideResNet-28-2 [47]上进行训练。为实现半监督学习，在 CIFAR10中，只有4,000张训练图像带有标签，即$N = 4000$。在 SVHN中，只有1,000张图像带有标签，即 $N = 1000$。

5.1.2 半监督学习算法：我们从第 2.3 节介绍的三类算法中选取了一些具有代表性的半监督学习算法：一致性正则化包括 PI-Model [30], MeanTeacher [38], VAT [25], 和 ICT [41], ；伪标签包括 PseudoLabel [19], 带一致性正则化的伪标签一致性正则化包括 FixMatch [35]。其中一些算法较早提出，性能不佳；另一些则是近期提出的，表现出色，这充分验证了我们投毒方法的普适性。这些半监督学习算法在 SVHN 和 CIFAR10 上的实现分别来自公开的 PyTorch 开源代码：[36] 和 [14], 。

5.1.3 基线投毒：我们将我们的投毒方法与 BadNets-C [10], CLB [39], 和 DeNeB-C [46] , 进行比较，这些方法已在第3节中详细介绍。它们的后门模式都是相同的 8×8像素块，位于图像的 (20, 20) 位置，如图6(b)、6(c)和6(d)所示。目标类别均为 8。

5.1.4 投毒设置：在我们的实验中，除受害者数据集 CIFAR10 对应的代理数据集为 CIFAR10、受害者数据集 SVHN 对应的代理数据集为 SVHN 外，其他实验设置均相同。替代网络和后门模式生成器分别为 WideResNet-28-2 [47] 和 UNet [32],。目标类别为 8，投毒样本数量为 500，λ_{ins} 0 05 ϵ 27 为 . , 表示像素扰动的最大值。

5.1.5 评估指标：评估指标包括三项：干净样本准确率（CA）、后门的攻击成功率（ASR）以及后门模式的不可感知性（IMP）。

CA：后门投毒不应降低半监督学习模型的准确率，也就是说，中毒模型的 CA 应接近未投毒模型的 CA。

ASR：在推理阶段，添加了后门模式的样本被中毒模型误分类为目标类别的概率。ASR 越高，投毒性能越好。

IMP：后门模式越不可感知，就越能逃避受害者的检测。我们通过计算干净图像与中毒图像之间的距离来量化不可感知性，采用 PSNR [13], SSIM [13], 和 L-∞ 范数。PSNR 越大、SSIM 越接近 1、L-∞ 范数越小，不可感知性越好。

5.2 投毒性能

我们首先验证投毒样本和梯度匹配在从头训练的自监督学习模型中是否有效。在该半监督学习



(a) 干净图像 (b) BadNets 的中毒图像 (c) CLB 的中毒图像 (d) DeNeB 的中毒图像 (e) 我们的中毒图像

图6：CIFAR10 中的干净图像和中毒图像。SVHN 中的这些图像见补充材料。

模型第 k个 epoch 中，将目标标签作为标签的投毒样本的损失为：

$$\mathcal{D} = \ell(y^t, \mathcal{F}(\mathcal{P}_{tr}(\mu_b \mathcal{U}); \theta_k)) \tag{10}$$

为了更准确地反映后门样本的预测结果与干净类别相距较远、与目标类别较为接近这一特点，我们采用相对距离来定义梯度匹配程度 C。

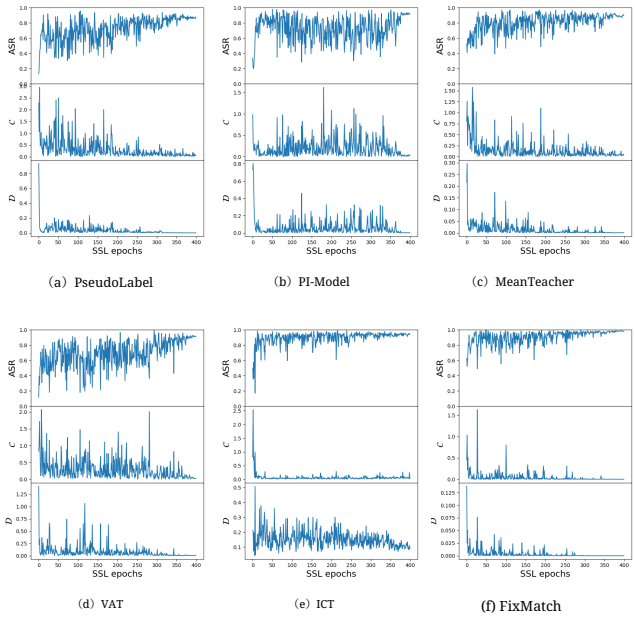
$$C = \mathbb{E}_{(x,y)\in\mathcal{X}_{val}} \frac{\left\|\nabla_{\theta_k}\ell(y^t, \mathcal{F}(\mathcal{P}_{val}(x); \theta_k)) - \nabla_{\theta_k}\ell(y^t, \mathcal{F}(x^t; \theta_k))\right\|^2}{\left\|\nabla_{\theta_k}\ell(y, \mathcal{F}(\mathcal{P}_{val}(x); \theta_k)) - \nabla_{\theta_k}\ell(y^t, \mathcal{F}(x^t; \theta_k))\right\|^2} \tag{11}$$

其中，分子计算的是后门样本 $\mathrm{P}_{val}(x)$ 将目标标签作为标签时，与目标样本 x^t 将目标标签作为标签时的梯度距离。分母计算的是后门样本将正确标签作为标签时，与目标样本将目标标签作为标签时的梯度距离。

由于中毒未标注样本均来自目标类别，半监督学习模型会将它们正确分类为目标类别 y^t，因此 D 会逐渐减小，如图7底部所示。因此，后门样本与目标样本之间的梯度距离 C 也会逐渐减小，如图7中部所示。这使得攻击成功率逐渐上升，如图7顶部所示。表1展示了最终的投毒结果，由此我们可以得出三个结论。

首先，半监督学习分类准确率显著高于监督学习分类准确率，这意味着半监督学习算法利用无标签样本进行学习能够提升模型准确率。此外，实施投毒后，中毒模型的准确率与干净模型相比并未显著下降，这意味着可以在不降低模型准确率的情况下实现无标签后门投毒。

其次，在这些半监督学习算法上，我们投毒的攻击成功率远高于几种基线投毒。如表1所示，在 CIFAR10 上，尽管这些投毒方法取得了一定的攻击成功率，但最高也只有 68.56%，而我们方法的最低值为 87.13%。在 SVHN 上，除 Fix- Match 以外，基线投毒在所有半监督学习算法上均告失败；Fix- Match 的攻击成功率为 75.47%。相比之下，我们的投毒方法可应用于这些半监督学习算法。虽然我们的投毒方法在 SVHN 上的攻击成功率低于在 CIFAR10 上的表现，但后门模式的不可感知性更好。如果攻击者愿意牺牲不可感知性，攻击成功率还会提高。此外，不同的半监督学习算法对我们的投毒方法具有不同的易受攻击程度。在 FixMatch 上，我们的投毒方法表现最佳；而在 VAT 上，攻击成功率最低。这可能是因为 VAT 将对抗扰动视为图像增强，并采用对抗训练来确保无标签样本能够抵御这些对抗扰动。我们制作的后门模式与对抗扰动相似，因此，对抗训练提高了模型抵御投毒样本的能力。



(a) PseudoLabel (b) PI-Model (c) MeanTeacher

(d) VAT (e) ICT (f) FixMatch

图7：在从头训练的SSL模型 CNN13 上，随着训练轮次增加，C、$\mathcal{D}$ 和攻击成功率的变化。

训练，以确保无标签样本能够抵御这些对抗扰动。我们制作的后门模式与对抗扰动相似，因此，对抗训练提高了模型抵御投毒样本的能力。

最后，得益于我们投毒的不可感知性设计，中毒图像与干净图像看起来非常相似，如图6所示。相比之下，基线投毒方法生成的中毒图像带有明显的后门方块，受害者很容易将其检测为可疑图像。从定量结果来看，如表1所列，我们的峰值信噪比、结构相似性指数和 L-∞ 范数均显著优于这些方法。总而言之，我们的投毒很好地实现了公式2中的投毒目标。

5.3 消融研究

为研究不同超参数或情境对投毒性能的影响，本节中的评估均在使用 CIFAR10 和 FixMatch 训练的模型上进行。针对其他目标类别的投毒结果见补充材料。

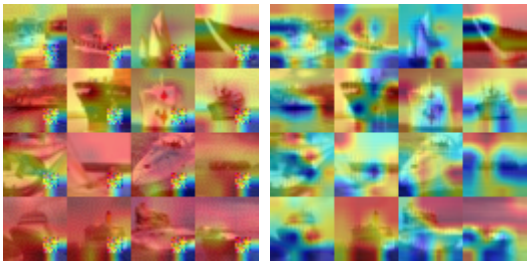
5.3.1 不同网络架构下的评估。我们评估了生成器和替代网络的不同架构对攻击成功率和不可感知性的影响。备选生成器包括 SimNet 和 UNet。备选替代网络包括 LeNet [18],

表2：跨网络架构的评估。SimNet 是我们构建的一个简单网络，仅由一个用于下采样的卷积层（3×64）和一个用于上采样的反卷积层（64×3）组成。

生成器	替代网络	CA	ASR	IMP		
				PSNR	SSIM	L-∞
SimNet	LeNet	93.72	99.33	21.72	0.6950	27.00
	CNN13	93.73	90.40	26.54	0.8703	26.89
	WideResNet-28-2	93.65	94.39	26.63	0.8966	26.87
UNet	LeNet	93.77	94.37	24.65	0.8057	27.00
	CNN13	93.86	95.24	31.43	0.9423	24.88
	WideResNet-28-2	93.83	97.12	31.34	0.9515	24.51

CNN13 和 WideResNet-28-2。投毒结果列于表2。比较 SimNet 和 UNet 可以看出，由于网络过于简单，SimNet 更难探索不可感知的后门模式。尽管在 LeNet 上获得了更高的攻击成功率，但不可感知性却大幅牺牲。在 CNN13 和 WideResNet-28-2 上，UNet 实现了更高的攻击成功率和更好的不可感知性。比较这些替代网络可见，WideResNet-28-2 规模更大，更善于探索可感知性最低的后门模式，并获得更高的攻击成功率。

5.3.2 不同扰动预算下的评估 ϵ。尽管较高的扰动预算 ϵ 会导致单个像素发生更显著的变化，但它为我们搜索不可感知的后门模式提供了更大的空间，并带来更高的攻击成功率。如表3所示，在 $\epsilon = 7$时，尽管像素扰动的最大值仅为7，但后门模式的不可感知性并未带来比 $\epsilon = 27$时更显著的提升，而攻击成功率则从97.12%大幅下降至64.40%。另一方面，当 $\epsilon = 54$时，像素扰动的最大值相比 $\epsilon = 27$时提升至32.32，而攻击成功率仅提高了1.51%。综合考虑攻击成功率和不可感知性，$\epsilon = 27$ 更合适。



(a) CLB的中毒图像 (b) 我们的中毒图像

图8：Grad-CAM [34] 可视化。

5.3.3 评估 *Hidden Triggers*。In this 章节，我们考虑在三种投毒情境中 i

S1：假设攻击者不知道受害者数据集是 CIFAR10，只知道目标类别和数据集分布。于是，他使用与 CIFAR10 分布相似的 CIFAR100 作为代理数据集，并用目标类别的样本替换某一类别的样本。

S2：假设攻击者不知道受害者训练集中样本的标注情况。因此，中毒未标注样本会按比例进行标注。

S3：假设中毒未标注样本来自不同类别。

表3：不同扰动预算下的评估 ϵ

ϵ	SSL CA	CA	ASR	IMP		
				PSNR	SSIM	L-∞
7	93.56	93.62	64.40	33.12	0.9502	7.00
27		93.83	97.12	31.34	0.9515	24.51
54		93.78	98.63	31.57	0.9497	32.32

表4：其他情境下的评估

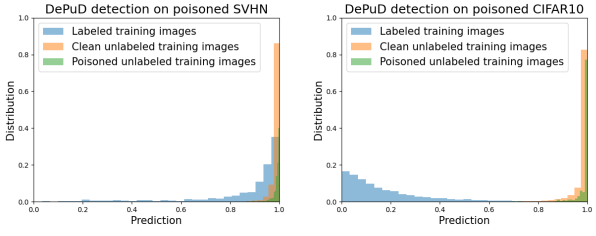
情境	SSL CA	CA	ASR	IMP		
				PSNR	SSIM	L-∞
S1	93.56	93.69	94.25	30.25	0.9389	24.59
S2		93.75	96.89	31.34	0.9515	24.51
S3		93.86	8.79	31.34	0.9515	24.51

投毒结果列于表4。在S1情境中，由于CIFAR100和CIFAR10的分布相似，在CIFAR100上制作的后门模式可以迁移到CIFAR10，攻击成功率达到94.25%。在S2情境中，由于我们只对目标类别的样本进行投毒，即使这些样本的标签正确，也不会影响攻击成功率。S3情境中的投毒失败（攻击成功率仅为8.79%）进一步验证了第3节中的发现：标签不一致后门投毒更难用于无标签样本。

5.4 防御评估

本节评估了我们的投毒攻击能否绕过第2.2节提到的四类防御方法中的五种代表性方法，包括激活聚类 [3], Neural Cleanse [42], 精细剪枝 [22], STRIP [8], 和 DePuD [46]。DePuD 的结果如下，其余四种结果见补充材料。

5.4.1 DePuD。DePuD旨在检测中毒未标注样本。首先，根据训练样本是否带有标签，将所有训练样本分为两类。已标记样本被赋予标签0，未标记样本被赋予标签1。随后，使用强正则化模型对这些样本进行分类。如果投毒样本具有明显的后门模式，其预测值将极度接近1，并且会与干净未标记样本产生明显分离，因此可被检测为异常。如图8(a)所示，DePuD对CLB有效，右下角可以明显检测到后门模式。然而，我们投毒所产生的后门模式是不可感知的，难以被强正则化模型捕获，如图8(b)所示。此外，如图9所示，干净未标记样本与中毒未标注样本几乎重叠，这意味着DePuD无法将中毒未标注样本分离出来。



(a) (b)

图9：DePuD 检测。

6 结论

本文首次研究了从头训练的半监督学习模型中无标签样本对后门投毒的脆弱性，揭示了半监督学习算法安全设计中的缺陷。我们首先发现，由于标签不一致后门投毒与半监督学习算法努力正确学习无标签样本的目标相悖，因此无法将其用于无标签样本。因此，对于无标签样本，投毒只能针对来自目标类别的样本实施。在此基础上，我们提出了一种零知识且不可感知的后门投毒。实验表明，在绕过各种防御方法时，我们的投毒实现了最先进的攻击成功率。

参考文献

[1] Berthelot, D.; Carlini, N.; Cubuk, E. D.; Kurakin, A.; Sohn, K.; Zhang, H.; and Raf- fel, C. 2019. ReMixMatch：通过分布匹配和增强锚定进行半监督学习。载于国际学习表征会议。[2] Berthelot, D.; Carlini, N.; Goodfellow, I.; Papernot, N.; Oliver, A.; and Raffel, C. A. 2019. Mixmatch：一种半监督学习的整体方法。神经信息处理系统进展，32。[3] Chen, B.; Carvalho, W.; Baracaldo, N.; Ludwig, H.; Edwards, B.; Lee, T.; Molloy, I.; and Srivastava, B. 2018. 通过激活聚类检测深度神经网络中的后门攻击。*arXiv*预印本 *arXiv:1811.03728*。[4] Chen, H.; Fu, C.; Zhao, J.; and Koushanfar, F. 2019. DeepInspect：用于深度神经网络的黑盒木马检测与缓解框架。载于*IJCAI*，第2卷，8。[5] Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. 一种视觉表征对比学习的简单框架。载于国际机器学习会议，1597–1607。PMLR。[6] Doan, K.; Lao, Y.; Zhao, W.; and Li, P. 2021. LIRA：可学习、不可察觉且鲁棒的后门攻击。载于*IEEE/CVF*国际计算机视觉会议论文集，11966–11976。[7] Dong, Y.; Yang, X.; Deng, Z.; Pang, T.; Xiao, Z.; Su, H.; and Zhu, J. 2021. 在信息和数据有限的情况下检测黑盒后门攻击。载于*IEEE/CVF*国际计算机视觉会议论文集，16482–16491。[8] Gao, Y.; Xu, C.; Wang, D.; Chen, S.; Ranasinghe, D. C.; and Nepal, S. 2019. STRIP：一种针对深度神经网络木马攻击的防御方法。载于第*35*届年度计算机安全应用会议论文集，113–125。[9] Gidaris, S.; Singh, P.; and Komodakis, N. 2018. 通过预测图像旋转进行无监督表征学习。载于国际学习表征会议。[10] Gu, T.; Dolan-Gavitt, B.; and Garg, S. 2017. Badnets：识别机器学习模型供应链中的漏洞。*arXiv*预印本 *arXiv:1708.06733*。[11] Guo, J.; Li, A.; and Liu, C. 2021. AEVA：使用对抗性极值分析进行黑盒后门检测。*arXiv*预印本 *arXiv:2110.14880*。[12] He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. 用于图像识别的深度残差学习。载于*IEEE*计算机视觉与模式识别会议论文集，770–778。[13] Hore, A.; and Ziou, D. 2010. 图像质量指标：峰值信噪比与结构相似性指数。载于*2010*年第*20*届国际模式识别会议，2366–2369。IEEE。[14] iBelieveCJM. 2020. 半监督深度学习技巧–Pytorch。https: // g it hub.com/iBelieveCJM/Tricks-of-Semi-supervisedDeepLeanring-Pytorch。[15] Iscen, A.; Tolias, G.; Avrithis, Y.; and Chum, O. 2019. 深度半监督学习的标签传播。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，5070–5079。[16] Krizhevsky, A.; Hinton, G.; et al. 2009. 从微型图像中学习多层特征。[17] Laine, S.; and Aila, T. 2016. 半监督学习的时间集成。*arXiv*预印本 *arXiv:1610.02242*。[18] LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. 应用于文档识别的基于梯度的学习。*IEEE*论文集，86(11)：2278–2324。[19] Lee, D.-H.; et al. 2013. 伪标签：一种简单高效的深度神经网络半监督学习方法。载于*ICML*表征学习挑战研讨会，第3卷，896。[20] Li, X.; Chen, Z.; Zhao, Y.; Tong, Z.; Zhao, Y.; Lim, A.; and Zhou, J. T. 2021. PointBA：迈向三维点云后门攻击。载于*IEEE/CVF*国际计算机视觉会议论文集，16492–16501。[21] Li, Y.; Li, Y.; Wu, B.; Li, L.; He, R.; and Lyu, S. 2021. 使用样本特定触发器的隐形后门攻击。载于*IEEE/CVF*国际计算机视觉会议论文集，16463–16472。[22] Liu, K.; Dolan-Gavitt, B.; and Garg, S. 2018. 精细剪枝：防御针对深度神经网络的后门攻击。载于《攻击、入侵与防御研究国际研讨会》，273–294。施普林格。[23] Liu, Y.; Ma, X.; Bailey, J.; and Lu, F. 2020. 反射后门：一种针对深度神经网络的自然后门攻击。载于《欧洲计算机视觉会议》，182–199。施普林格。[24] Mahajan, D.; Girshick, R.; Ramanathan, V.; He, K.; Paluri, M.; Li, Y.; Bharambe, A.; and Van Der Maaten, L. 2018. 探索弱监督预训练的极限。载于《欧洲计算机视觉会议论文集》（ECCV），181–196。[25] Miyato, T.; Maeda, S.-i.; Koyama, M.; and Ishii, S. 2018. 虚拟对抗训练：一种用于监督学习和半监督学习的正则化方法。《IEEE模式分析与机器智能汇刊》，41(8)：1979–1993。[26] Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; and Ng, A. Y. 2011. 使用无监督特征学习读取自然图像中的数字。[27] Nguyen, T. A.; and Tran, A. 2020. 输入感知动态后门攻击。《神经信息处理系统进展》，33：3454–3464。[28] Nguyen, T. A.; and Tran, A. T. 2020. WaNet——基于不可察觉变形的后门攻击。载于《国际学习表征会议》。[29] Pham, H.; Dai, Z.; Xie, Q.; and Le, Q. V. 2021. 元伪标签。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，11557–11568。[30] Rasmus, A.; Berglund, M.; Honkala, M.; Valpola, H.; and Raiko, T. 2015. 使用阶梯网络进行半监督学习。《神经信息处理系统进展》，28。[31] Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You Only Look Once：统一的实时目标检测。载于《IEEE计算机视觉与模式识别会议论文集》，779–788。[32] Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net：用于生物医学图像分割的卷积网络。载于《医学图像计算与计算机辅助干预国际会议》，234–241。施普林格。[33] Salem, A.; Wen, R.; Backes, M.; Ma, S.; and Zhang, Y. 2020. 针对机器学习模型的动态后门攻击。arXiv预印本 arXiv:2003.03675。[34] Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-CAM：通过基于梯度的定位对深度网络进行视觉解释。载于《IEEE国际计算机视觉会议论文集》，618–626。[35] Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C. A.; Cubuk, E. D.; Kurakin, A.; and Li, C.-L. 2020. FixMatch：通过一致性和置信度简化半监督学习。《神经信息处理系统进展》，33：596–608。[36] S uzuki, T. 2020. 使用PyTorch进行半监督学习的一致性正则化。https://github.com/perrying/pytorch-consistency-regularization。[37] Szegedy, C.; Ioffe, S.; Vanhoucke, V.; and Alemi, A. A. 2017. Inception-v4、Inception-ResNet以及残差连接对学习的影响。载于第三十一届AAAI人工智能会议。[38] Tarvainen, A.; and Valpola, H. 2017. 平均教师是更好的榜样：权重平均一致性目标提升半监督深度学习效果。《神经信息处理系统进展》，30。[39] Turner, A.; Tsipras, D.; and Madry, A. 2019. 标签一致后门攻击。arXiv预印本 arXiv:1912.02771。[40] Van der Maaten, L.; and Hinton, G. 2008. 使用t-SNE可视化数据。《机器学习研究期刊》，9(11)。[41] Verma, V.; Kawaguchi, K.; Lamb, A.; Kannala, J.; Bengio, Y.; and Lopez-Paz, D. 2019. 用于半监督学习的插值一致性训练。arXiv预印本 arXiv:1903.03825。[42] Wang, B.; Yao, Y.; Shan, S.; Li, H.; Viswanath, B.; Zheng, H.; and Zhao, B. Y. 2019. Neural Cleanse：识别并缓解神经网络中的后门攻击。载于《2019年IEEE安全与隐私研讨会》（SP），707–723。IEEE。[43] Wang, R.; Zhang, G.; Liu, S.; Chen, P.-Y.; Xiong, J.; and Wang, M. 2020. 木马神经网络的实用检测：数据受限和无数据情形。载于《欧洲计算机视觉会议》，222–238。施普林格。[44] Xie, Q.; Dai, Z.; Hovy, E.; Luong, T.; and Le, Q. 2020. 用于一致性训练的无监督数据增强。《神经信息处理系统进展》，33：6256–6268。[45] Yan, Z.; Li, G.; TIan, Y.; Wu, J.; Li, S.; Chen, M.; and Poor, H. V. 2021. DeHiB：通过对抗扰动攻击半监督学习的深度隐蔽后门攻击。载于《AAAI人工智能会议论文集》，第35卷，10585–10593。[46] Yan, Z.; Wu, J.; Li, G.; Li, S.; and Guizani, M. 2021. 半监督学习中的深度神经网络后门：威胁与对策。《IEEE信息取证与安全汇刊》，16：4827–4842。[47] Zagoruyko, S.; and Komodakis, N. 2016. 宽残差网络。arXiv预印本 arXiv:1605.07146。[48] Zhao, S.; Ma, X.; Zheng, X.; Bailey, J.; Chen, J.; and Jiang, Y.-G. 2020. 针对视频识别模型的干净标签后门攻击。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，14443–14452。

[49] Zhao, Z.-Q.; Zheng, P.; Xu, S.-t.; and Wu, X. 2019. 深度学习目标检测：综述。*IEEE*神经网络与学习系统汇刊，30(11)：3212–3232。

附录

A SVHN 图像

图 12 展示了 SVHN 中的干净图像和中毒图像。可以看出，我们的中毒图像在视觉上与干净图像非常相似，受害者很难检测出来。

B 跨目标类别评估

我们选择其他类别作为目标类别。投毒结果列于表 5。首先，与之类似，对其他目标类别进行投毒不会降低模型准确率。其次，可以看出，不同目标类别的投毒难度各不相同。例如，对于目标类别鸟，攻击成功率可达 99.52%；而对于卡车，攻击成功率较低，为 81.88%。当然，我们也可以牺牲少许后门模式隐蔽性来提高攻击成功率。

C 防御评估

我们评估了投毒方法在激活聚类、Neural Cleanse、精细剪枝和 STRIP 上的表现。

C.1 激活聚类

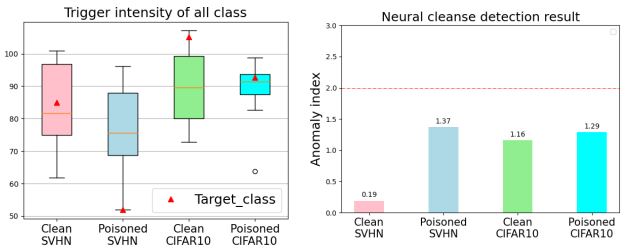
激活聚类的流程是将所有训练样本输入已训练好的受害模型，从而获得这些样本在最后一个隐藏层中的激活。随后，根据标签将这些激活划分为不同的簇。最后，通过检测这些簇是否存在异常来判断是否有投毒样本。然而，我们的后门投毒不依赖标签，且只对无标签样本进行投毒。因此，投毒的无标签样本无法根据标签划分到不同的簇中。由此，我们的投毒方法可以自然地绕过激活聚类检测。

C.2 Neural Cleanse

针对可能已中毒的模型，对所有类别的最低强度后门触发器进行逆向工程。随后，依据以下先验知识，判断某个类别是否为目标类别：注入后门的目标类别具有强度异常小的触发器，即 $anomaly\ index > 2$。在半监督学习中，防御者可能没有足够的有标签样本来进行更准确的逆向工程，但我们假设防御者拥有足够的有标签样本这一最严格的条件。然而，即便如此，如图10(a)所示，注入后门的目标类别的触发器强度并非显著的离群值，且 $anomaly\ index$ 均小于 2（图 10(b)）。因此，我们的投毒能够成功绕过 Neural Cleanse 检测。我们认为，这是因为 Neural Cleanse 检测中的逆向工程依赖于分类层，而我们的投毒采用梯度匹配来控制整个网络。

C.3 精细剪枝

这是一种盲防御策略，不是检测模型或样本是否被投毒，而是通过剪枝和微调来



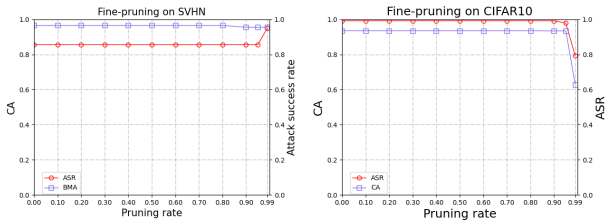
(a) (b)

图 10：Neural Cleanse 检测。

表 5：跨目标类别评估

目标类别	SSL CA	CA	ASR	IMP		
				PSNR	SSIM	L-∞
飞机	93.56	93.49	99.15	33.29	0.9518	24.26
汽车		93.44	89.22	30.21	0.9490	24.88
Bird		93.76	99.52	33.08	0.9492	23.99
Cat		93.95	99.95	33.33	0.9655	20.06
Deer		93.78	98.18	32.51	0.9560	22.18
Dog		93.81	91.08	32.32	0.9634	23.81
Frog		93.89	95.64	31.47	0.9519	23.22
马		93.84	94.11	32.36	0.9653	22.91
Ship		93.83	97.12	31.34	0.9515	24.51
卡车		93.93	81.88	32.65	0.9765	23.29

尝试消除任何模型中可能存在的后门。具体而言，首先将干净样本输入模型，然后通过剪枝使激活值最小的神经元中有 $\alpha\%$（即剪枝率）处于休眠状态，从而尝试移除可能存在的后门。随后使用微调来弥补剪枝导致的干净样本准确率下降。如图 11 所示，即使剪枝率达到 90%，攻击成功率也没有显著下降。当剪枝率为 99% 时，在 CIFAR10 上，干净准确率降至 80%，攻击成功率仍为 62%。一个有趣的现象是，在 SVHN 上，攻击成功率显著上升；这可能是因为过度剪枝削弱了模型区分干净样本的能力，使得添加后门模式后，样本更容易被误分类为目标类别。



(a) (b)

图11：精细剪枝。

C.4 STRIP

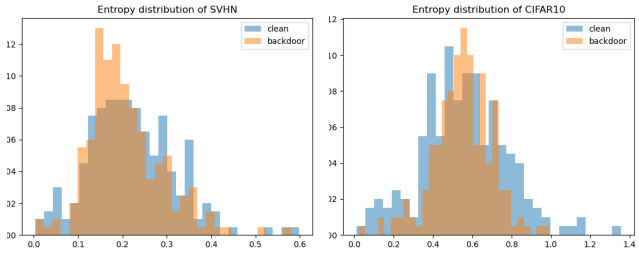
STRIP 部署在模型推理阶段。在将测试样本输入模型之前，先将其与一组预先准备好的干净样本合成。然后，将获得的合成样本输入模型进行预测。如果这些预测结果的熵异常低，则判定该测试样本是后门样本，且模型已遭到投毒。



(a) 干净图像 (b) BadNets 的中毒图像 (c) CLB 的中毒图像 (d) DeNeB 的中毒图像 (e) 我们的中毒图像

Fig图 12：S 上的干净图像和中毒图像 VHN.

如图 13 所示，我们展示了 500 个测试样本和 500 个后门样本的熵分布，可以看出两者的分布几乎重合。后门样本的熵并未表现出异常偏低的特征。因此，我们的后门投毒能够绕过 STRIP 检测。



(a) (b)

图 13：STRIP 检测。