南京邮电大学
Nanjing University of Posts and Telecommunications

# 真实世界场景下深度神经网络的高效后门攻击

Sun H, Li Z, Xia P, et al. Efficient Backdoor Attacks for Deep Neural Networks in Real-world Scenarios[J]. arXiv preprint arXiv:2306.08386, 2023. Published as a conference paper at ICLR 2024

汇报人：2023040509 郭稼逸

汇报时间： 2024 年 6 月 1 日

# CONTENT



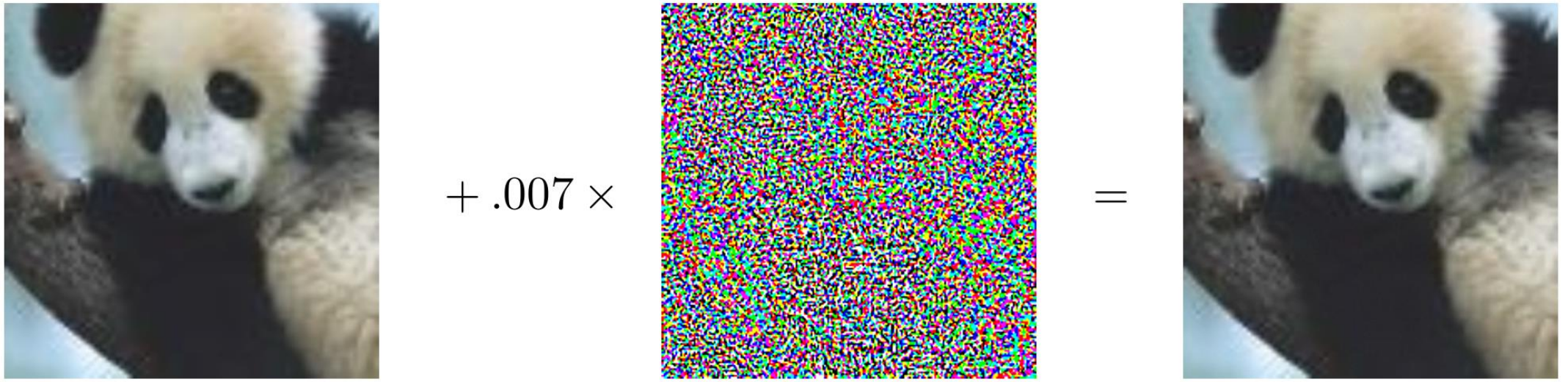
南京邮电大学 Nanjing University of Posts and Telecommunications


# 研究背景

近年来，深度学习(deep learning,DL)在大规模数据上展现出卓越性能。以深度学习为核心的人工智能系统被广泛应用在计算机视觉(CV)、自然语言处理(NLP)、语音识别(VRS)和恶意软件检测(MD)等关键场景。

与此同时，深度学习面临多种安全威胁。与传统机器学习算法不同，深度学习模型通常由多层非线性变换组成，结构更为复杂，面对大规模数据表现出更强的泛化能力与表达能力，被广泛应用于多种场景中，因此针对深度学习的攻击具有更高的研究价值。

# 对抗攻击

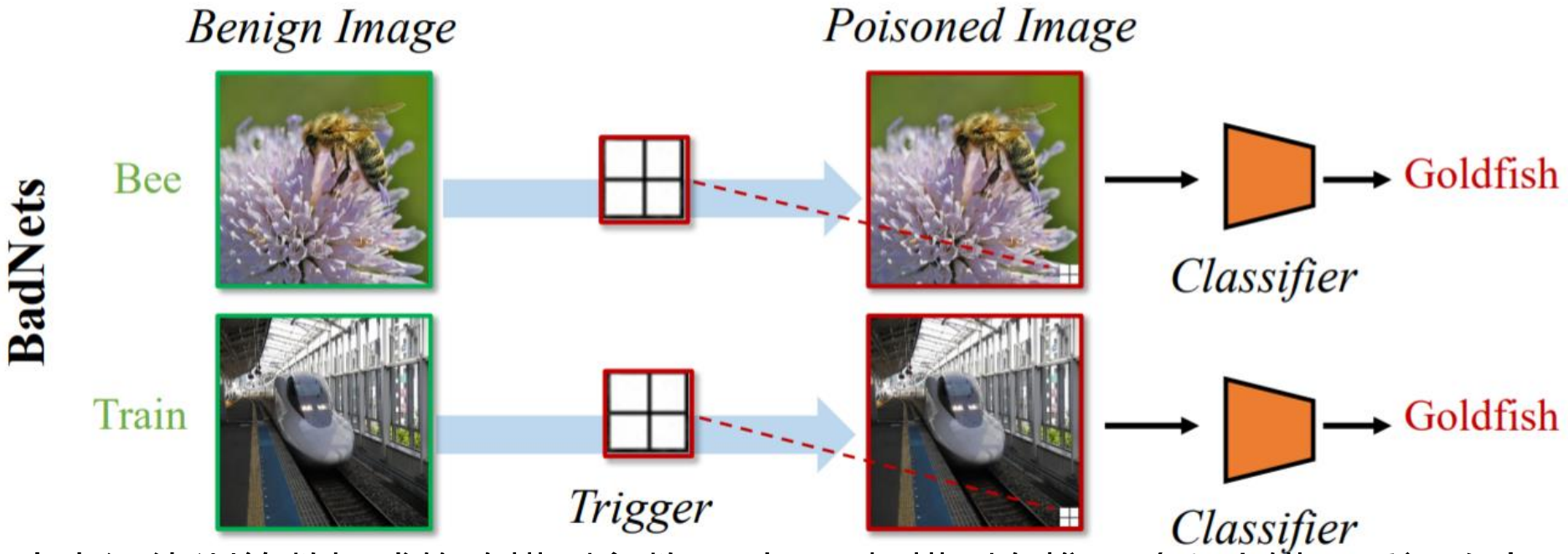



对抗样本攻击是通过对输入数据进行微小但有针对性的修改，使得机器学习模型产生错误分类或错误预测的样本。这些微小的变化对人类观察几乎不可察觉，但足以使模型做出错误的推断。对抗样本攻击是针对模型的鲁棒性和稳定性，即使在面对微小扰动时也能保持准确性。

## 数据投毒


数据投毒是一种通过在训练数据中植入恶意样本或修改数据以欺骗机器学习模型的方法。这种攻击旨在使模型在未来的预测或决策中产生错误结果。攻击者可能会植入具有误导性标签或特征的数据，以扭曲模型的学习过程，导致模型偏离真实数据的表征。数据投毒攻击可能在模型训练过程中不被察觉，但其影响可能在模型部署和运行时显现出来。

# 后门攻击




后门攻击指攻击者污染训练数据或修改模型参数，引导目标模型在推理阶段出错。后门攻击具有极强的隐蔽性，遭到后门攻击的模型在大部分数据上表现正常，但在具有指定特征的一类数据上表现出错。

完整的后门攻击具有三个步骤：

(1) 触发器设计：攻击者根据任务需求设计触发器,并生成中毒样本；

(2) 后门注入：攻击者通过数据投毒或参数修改等方式向目标模型注入后门；

(3) 后门激活：攻击者生成含触发器的测试案例激活模型后门

现有，训练数据集$D_{train} = \{(x_i, y_i)|x_i \in X, y_i \in Y\}$，深度学习模型$f_\theta: X \to Y, X \subset \mathbb{R}^n, Y \subset \mathbb{R}^d$，攻击者指定目标类别$y_t$，在随机训练样本$x_i$上添加触发器构成中毒样本$x_t$

旨在训练后门模型满足$f_{\theta_b}: X \to Y$，满足$f_{\theta_b}(x_i) = y_i$，$f_{\theta_b}(x_t) = y_t$。

整体上，中毒样本$x_t$由干净样本$x_i$和触发生成算法$G(\cdot)$表示：

$$x_t = \alpha(x_i) + \beta\big(G(x_i)\big)$$

其中，$\alpha(\cdot)$和$\beta(\cdot)$表示干净样本和触发器混合方式。

后门攻击具有两个基本目标：一是在正常样本上，后门模型表现出与良性模型相近的性能；二是在含有触发器的后门样本，后门模型表现出攻击者期望的结果。

# 研究背景



Gu等人提出的BadNets**第一次引入后门攻击**这个概念，并成功在MNIST等数据集上进行了攻击。攻击者需要能够对数据进行投毒，而且还要重新训练模型以改变模型某些参数从而实现后门的植入。

Blended-Attack论证了**trigger可以任意设置**，文章中测试了两种trigger：HelloKitty水印和随机高斯噪声。但中毒样本的label和它们的ground-truth label不匹配，这非常容易被筛选出来，而且人类能够轻易分辨出这种攻击。

lable-consistent-backdoor-attack是**第一个clean-label攻击**，给数据集输入一些数据，这些数据被加上了backdoor-trigger，但是它们的图像内容和标签是一致的。模型在识别的时候极度依赖该trigger，后续在推理预测中遇到含有trigger的图像时，无论内容是否为目标label，模型都会将其识别为目标label。

Hidden-Trigger-Backdoor-Attacks没有肉眼可见的trigger。把带有trigger的原图和另一个类别的图混合之后生成中毒数据，并以另一个类别的形式参加训练。在test的时候，带有trigger的图就会被认为是另外一个类别。

PART
2

# 研究内容

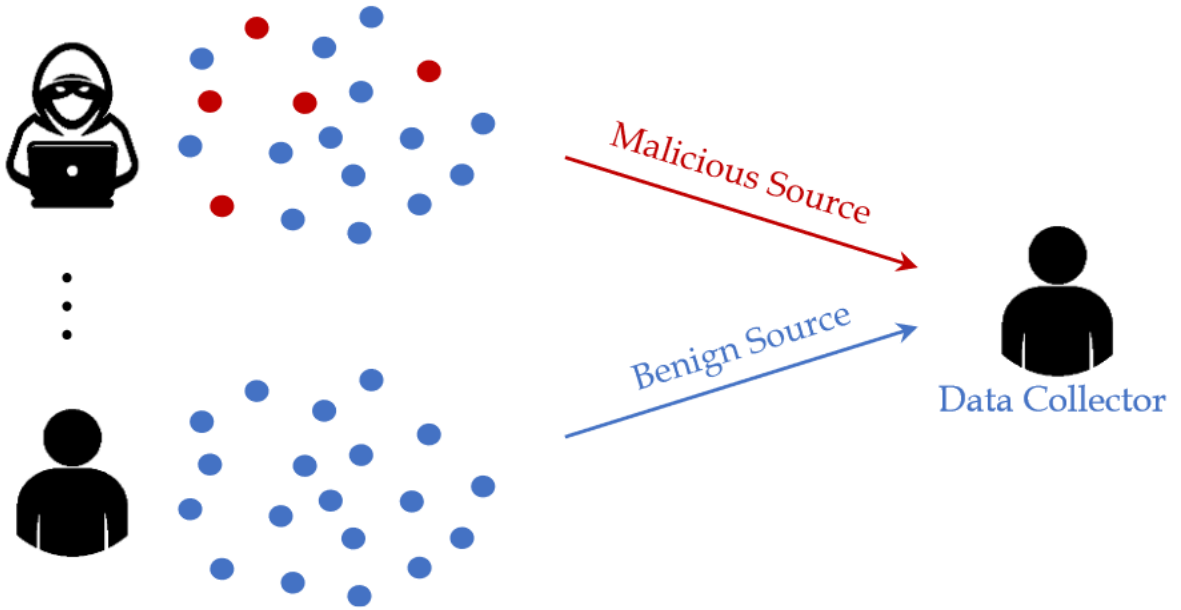
后门攻击依赖的假设可能过于宽泛，假定所有训练数据都来自单一来源，而收集到的来源已被攻击者下毒，如右上图所示。

然而，这并不能准确地反映真实世界的攻击场景。


受害者拥有有限样本私人数据集的情况。为了弥补不足，他们可能会从公共数据集收集数据，并将其与私人数据结合起来进行训练，如右下角图所示。

在这种情况下，攻击者无法访问私有数据集，只能操纵公共数据集的一部分进行投毒。因此，中毒数据和训练数据的分布会出现差异，这与以往的后门攻击不同。

Data-Constrained Backdoor Attacks



# Data-Constrained Backdoor Attacks

1. **Number-constrained**。在数量受限后门攻击中，攻击者的数据集和受害者所搜集的数据集有着相同的分布，但是攻击者的数据集的数量大小，远小于受害者搜集的数据集的大小。

| 类别 | 猫 | 狗 | 鱼 | 鸟 | 车 | 船 | 飞机 | 人 |
|---|---|---|---|---|---|---|---|---|
| 攻击者 | 3 | 2 | 1 | 2 | 4 | 3 | 2 | 10 |
| 受害者 | 1500 | 1000 | 500 | 1000 | 2000 | 1500 | 1000 | 5000 |

如上表展示，对于攻击者而言，所能接触到的每一类对应的图片占了受害者训练使用的图片的0.2%，但这有些夸张，实际情况可能会更少或者更多，在这种情况下，攻击者可能很难完成一次成功的攻击。

# Data-Constrained Backdoor Attacks

2. **Class-constrained**。在类受限后门攻击中，攻击者的数据集的类数小于受害者搜集的数据集类的量，即攻击者数据集里可能只有动物相关的类别，而受害者搜集的数据集中还有交通工具等类别。

| 类别 | 猫 | 狗 | 鱼 | 鸟 | 车 | 船 | 飞机 | 人 |
|---|---|---|---|---|---|---|---|---|
| 攻击者 | 15 | 0 | 0 | 10 | 20 | 0 | 10 | 50 |
| 受害者 | 1500 | 1000 | 500 | 1000 | 2000 | 1500 | 1000 | 5000 |

如上表展示，受害者有8个类别的数据集，然而攻击者却只有6个类别的数据集，这种情况下，未知类别的样本可能会给攻击者带来巨大的挑战，未知的类别也会使攻击者在设置触发器时遇到困难。

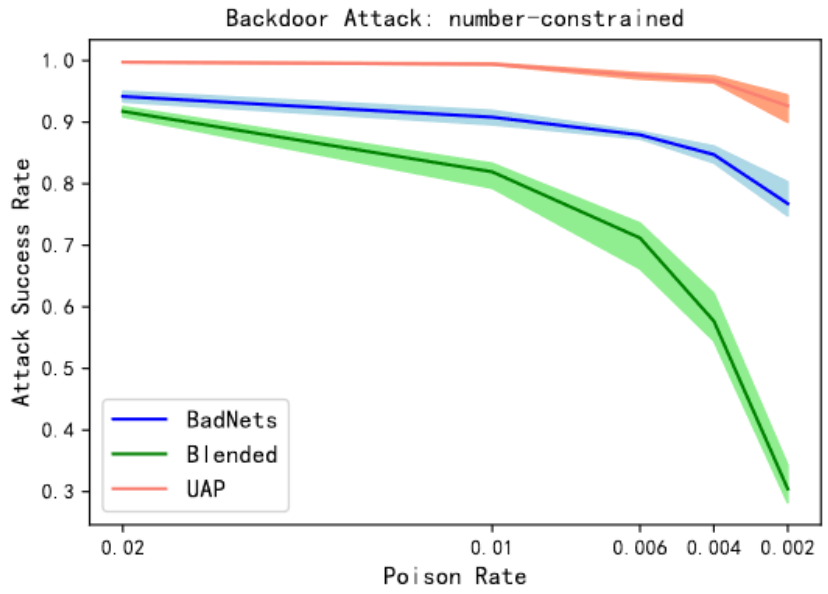
# Data-Constrained Backdoor Attacks

3. **Domain-constrained**。在域受限的后门攻击中，可以参考OOD（out-of-domain）理解，即模型训练时未曾见过或未包含在模型训练数据中的输入数据的类型或领域，此时模型可能会表现不佳。

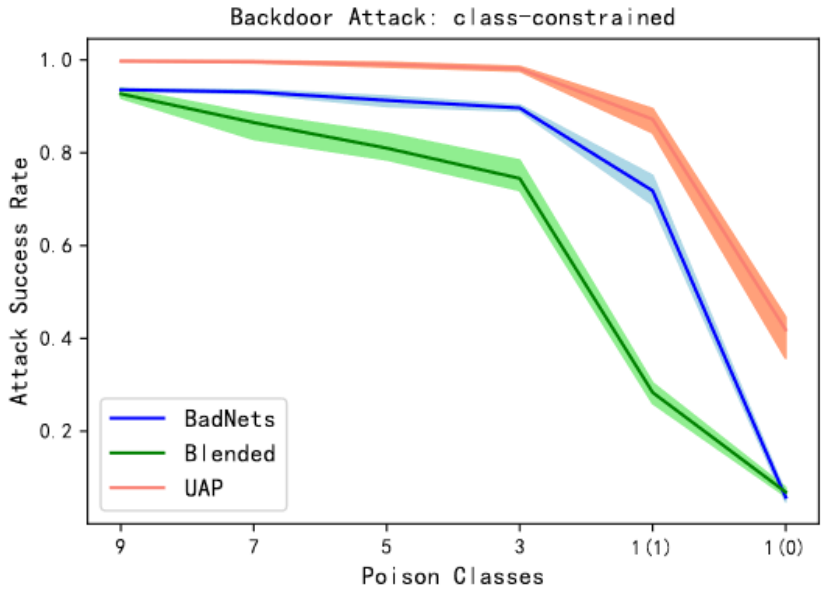
| 类别 | 猫 | 狗 | 鱼 | 鸟 | 车 | 船 | 飞机 | 人 |
|---|---|---|---|---|---|---|---|---|
| 攻击者 | 15 | 15 | 50 | 10 | 10 | 20 | 20 | 5 |
| 受害者 | 1500 | 1000 | 500 | 2000 | 2000 | 1500 | 1000 | 5000 |

如上表展示，攻击者类别的数据比例是3:3:10:2:2:4:4:1，而受害者的数据比例是3:2:1:4:4:3:2:10。

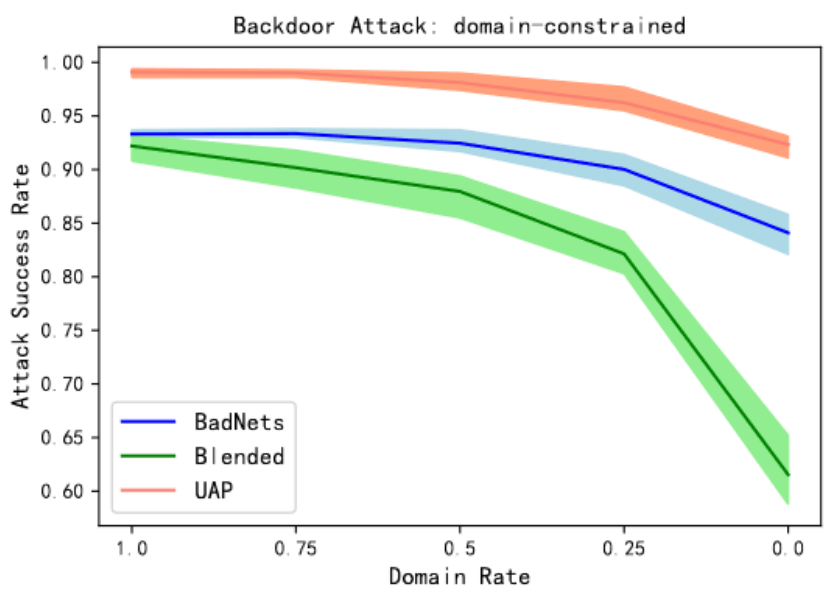
攻击者和受害者的数据分布不一致，会导致攻击者触发器设置的侧重点和受害者训练的侧重点不一致，从而使攻击失败。




(a) Number-constrained




(b) Class-constrained




(c) Domain-constrained

1. **Number-constrained**。随着中毒数据的减少，后门攻击成功率明显下降。
2. **Class-constrained**。随着中毒类的减少，后门攻击成功率明显下降。
3. **Domain-constrained**。随着域限制增大，后门攻击成功率明显下降。

其中，橙色线是通用对抗扰动，需要访问所有的数据集，与我们的限制矛盾。

故得出结论，先前的方法在真实场景下，**受到了限制**。

实验还观察到：

1. 在数据受限的攻击场景中，BadNet 的表现明显优于 Blended。
2. 在相同的毒化率下，数量受限、脏标签单类和干净标签单类后门攻击的攻击效率依次降低。
3. 具有相同触发因素的中毒样本之间的活化程度存在巨大差异。

根据上述现象，后门注入期间良性和中毒特征之间存在显著相互依赖关系，这种错综复杂的纠缠被认为是导致当前攻击方法在数据受限场景下表现出不足的主要因素。我们的研究是在后门攻击背景下对特征纠缠的开创性探索，理想情况下，当面对中毒样本时，人们会期望后门模型完全依赖于中毒特征，因为这将是执行成功后门攻击的最有效策略。

文章提出了两种解决方法，可以独立使用，也可以一同使用。

**干净特征抑制**与**中毒特征增强**。

方法的主要目标是尽量减少良性特征对决策过程的影响，从而放大中毒特征的重要性。

# 干净特征抑制——一般思想

干净数据集$\mathcal{P}' \subset X \times Y$，$\mathcal{P}' = \{(x_i, y_i)|i = 1, \cdots, P\}$

输入$x_i \in X$，标签$y_i \in Y = \{1,2, \cdots, C\}$，类别总数$C$

中毒数据集$\mathcal{P}_e = \{(x_{e,i}, y_i)|i = 1, \cdots, P\}$

输入$x_i \in \mathcal{P}'$的擦除版本$x_{e,i} = x_i + \delta_i$

达到擦除效果使用的微小噪声 $\delta_i \in \Delta$

均方误差损失 $L$，$L(a, b) = \|a - b\|^2$

任务的无偏标签$y_m$，$y_m = \left[\frac{1}{C}, \frac{1}{C}, \cdots, \frac{1}{C}\right]$

虽然这种寻常方法能有效清除干净的特征，但它需要一个在**整个训练集**上预先训练过的代理特征提取器。这种方法**不适合**我们的数据受限后门攻击。

$$\delta_i = \underset{\delta_i}{\operatorname{argmin}} L(f'(x_i + \delta_i), y_m) \ s.t. \|\delta_i\|_p \leq \epsilon$$

# 干净特征抑制——引入CLIP方法

使用文本编码器$\hat{\varepsilon}_t(\cdot)$将输入提示（a photo of a $c_i$ ）嵌入到文本特征$T_i \in \mathbb{R}^d$

使用图像编码器$\hat{\varepsilon}_i(\cdot)$嵌入图像$x_j$的图像特征$I_j \in \mathbb{R}^d$

预测结果$y_j = \underset{\delta_i}{\operatorname{argmax}}\left(\langle I_j, T_i \rangle\right)$

$\langle \cdot, \cdot \rangle$表示两个向量之间的余弦相似度

# 基于CLIP的干净特征抑制

$$\delta_i = \underset{\delta_i}{\operatorname{argmin}} L(f_{CLIP}(x_i + \delta_i, \mathbb{P}), y_m) \ \ s.t. \|\delta_i\|_p \leq \epsilon$$

$$\mathbb{P} = \{p_1, p_2, \cdots, p_C\} = \{"a\ photo\ of\ a\ c_i" | i = 1,2, \cdots, C\}$$

$$f_{CLIP}(x_i + \delta_i, \mathbb{P}) = \left[\frac{\langle \widehat{\varepsilon_i}(x_i + \delta_i), \widehat{\varepsilon_t}(p_1)\rangle}{\sum_{i=1}^{C}\langle \widehat{\varepsilon_i}(x_i + \delta_i), \widehat{\varepsilon_t}(p_i)\rangle}, \cdots, \frac{\langle \widehat{\varepsilon_i}(x_i + \delta_i), \widehat{\varepsilon_t}(p_C)\rangle}{\sum_{i=1}^{C}\langle \widehat{\varepsilon_i}(x_i + \delta_i), \widehat{\varepsilon_t}(p_i)\rangle}\right]$$

采用了“梯度下降预测法”(PGD)的一阶优化方法，扰动为：

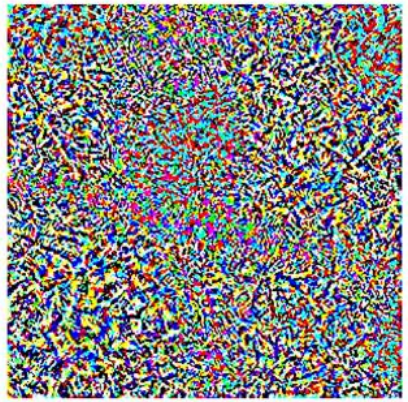
$$\delta_i^{t+1} = \prod_{\epsilon}\left(\delta_i^t - \alpha \cdot \operatorname{sign}\left(\nabla_\delta L\left(f_{CLIP}\left(x_i + \delta_i^t, \mathbb{P}\right), y_m\right)\right)\right)$$

最终获得受到抑制的样例$\mathcal{P}_e = \{(x_{e,i}, y_i) | i = 1, \cdots, P\}$, 其中 $x_{e,i} = x_i + \delta_i^T$

优势：不用访问所有的数据集

# 基于CLIP的干净特征抑制

+



$\delta$

=

| 类别 | 置信度 |
|---|---|
| 猫 | 1.0% |
| 狗 | 1.0% |
| 鱼 | 1.0% |
| 鸟 | 93.0% |
| 车 | 1.0% |
| 船 | 1.0% |
| 人 | 1.0% |
| 飞机 | 1.0% |

| 类别 | 置信度 |
|---|---|
| 猫 | 12.5% |
| 狗 | 12.4% |
| 鱼 | 12.4% |
| 鸟 | 12.8% |
| 车 | 12.5% |
| 船 | 12.5% |
| 人 | 12.4% |
| 飞机 | 12.5% |

如左图的例子所示，对于原始图片，有着93%的是鸟这一类，但是增添了专门设计的扰动后，只有12.8%的置信度是鸟这一类了，并且其余类别的置信度都很相似，这就是我们干净特征抑制的目的。

举例可能夸张，但更容易理解。

## 中毒特征增强

与从干净样本中提取的干净特征相比，从设计的触发器中提取的中毒特征将更具表现力。故最大化正对之间的相似性，同时确保负对之间的不相似性来优化一般触发器。

待优化的触发器$\delta_{con}^{T+1}$ 干净样本 $x$和$x_1$

构造的恶意样本 $x+\delta_{con}$和$x_1+\delta_{con}$

正对$(x+\delta_{con}, x_1+\delta_{con})$ 负对$(x+\delta_{con}, x)$

图像编码器处理$v_q=\widehat{\varepsilon_i}(x+\delta_{con}), v_+=\widehat{\varepsilon_i}(x_1+\delta_{con}), v_-=\widehat{\varepsilon_i}(x)$

损失函数$L_{con}(x, x_1, \delta_{con})=-\frac{\langle v_q, v_+\rangle}{\langle v_q, v_-\rangle}$

扰动目标$\delta_{con}=\underset{\|\delta_{con}\|_p\leq\epsilon}{\operatorname{argmin}}\sum_{(x,y)\in\mathcal{D}'}L_{con}(x, x_1, \delta_{con})$， 一样使用PGD优化方法

扰动迭代$\delta_{con}^{t+1}=\prod_{\epsilon}\left(\delta_{con}^{t}-\alpha\cdot\operatorname{sign}\left(\nabla_{\delta_{con}}L_{con}(x, x_1, \delta_{con})\right)\right)$

# 中毒特征增强


如左图的例子所示，中毒特征增强就是期望模型对于这些图片的分类时，第一行、第二行图片，更关注的是右下角红色框框，第三行的图片更关注的是隐藏在图片下的特意设计的隐藏的触发器。这就是中毒特征增强的目的。

举例可能夸张，但更容易理解。

PART
3

# 实验结果

# 实验结果



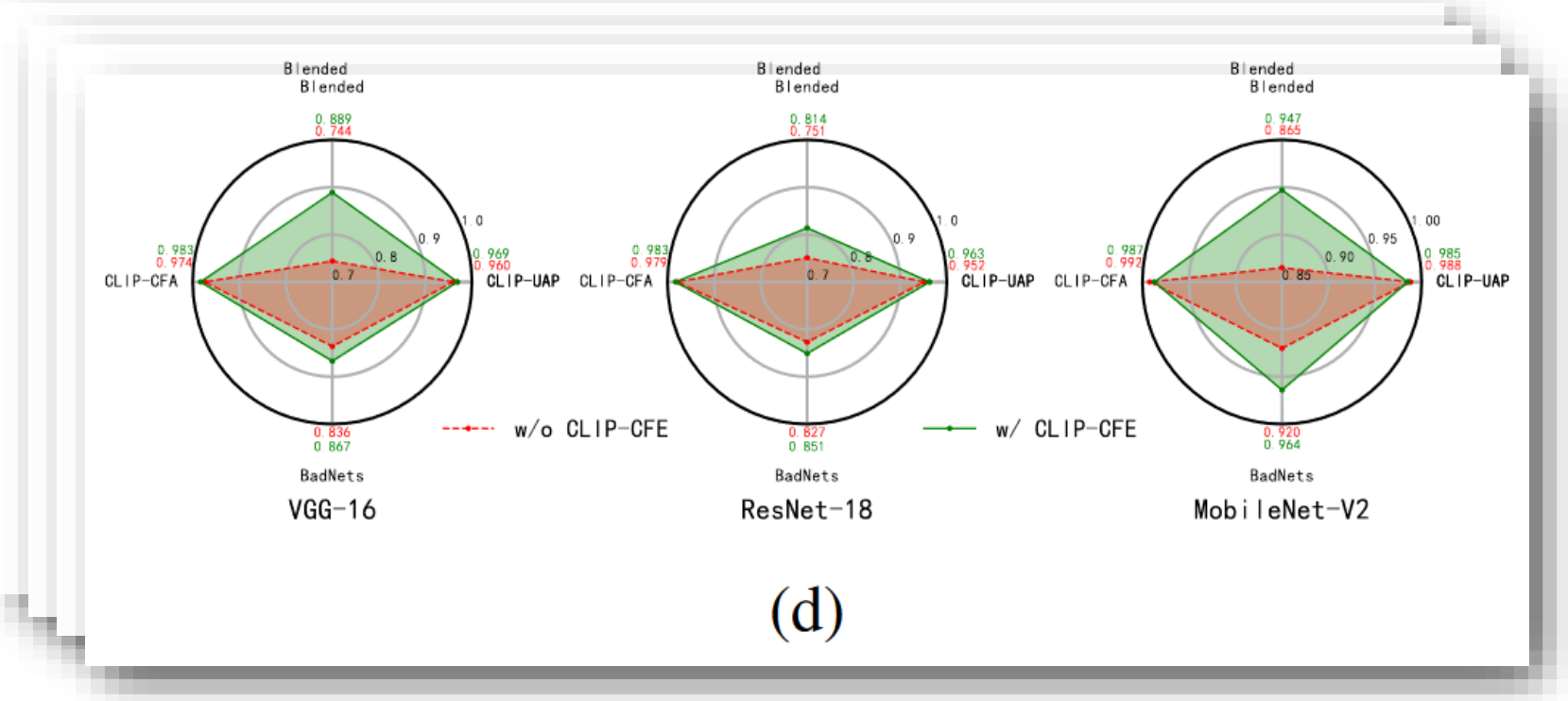
问题1：提出的技术对三种后门攻击有效吗?

问题2：提出的技术对良性样本的准确性有影响吗?

问题3：提出的技术对受害者是否隐蔽?

问题4：提出的技术在不同的中毒环境有效吗?

# 问题1：提出的技术对三种后门攻击有效吗?




(a)数量受限

(b)干净标签单类(类受限)

(c)脏标签单类(类受限)

(d)域外(域受限)

在对VGG-16模型的数量受限后门攻击中，BadNets、Blended、CLIP-UAP和CLIP-CFA的ASR分别达到了0.878、0.825、0.984和0.988。

在对VGG-16数据集的干净标签单类后门攻击中，我们观察到BadNets、Blended、CLIP-UAP和CLIP-CFA分别显著提高了187%、150%、110%和 229%。

**基于CLIP的中毒特征增强比之前的攻击方法更有效。**

**基于CLIP的干净特征抑制可用于不同的攻击方法。**

# 问题2：提出的技术对良性样本的准确性有影响吗?

| Trigger | Clean Feature Suppression | Backdoor Attacks | | | | | | | | | | | | Average |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Number Constrained | | | Class Constrained ($Y' = \{0\}$) | | | Class Constrained ($Y' = \{1\}$) | | | Domain Constrained | | | |
| | | V-16 | R-18 | M-2 | V-16 | R-18 | M-2 | V-16 | R-18 | M-2 | V-16 | R-18 | M-2 | |
| BadNets | w/o CLIP-CFE | 0.698 | 0.728 | 0.722 | 0.698 | 0.730 | 0.728 | 0.700 | 0.728 | 0.729 | 0.699 | 0.727 | 0.728 | 0.718 |
| | w/ CLIP-CFE | 0.700 | 0.730 | 0.728 | 0.701 | 0.731 | 0.723 | 0.698 | 0.730 | 0.726 | 0.701 | 0.730 | 0.724 | 0.719 |
| Blended | w/o CLIP-CFE | 0.700 | 0.727 | 0.722 | 0.700 | 0.726 | 0.725 | 0.701 | 0.729 | 0.723 | 0.698 | 0.729 | 0.725 | 0.717 |
| | w/ CLIP-CFE | 0.700 | 0.730 | 0.727 | 0.701 | 0.729 | 0.727 | 0.699 | 0.730 | 0.724 | 0.700 | 0.731 | 0.727 | 0.719 |
| CLIP-UAP | w/o CLIP-CFE | 0.702 | 0.730 | 0.727 | 0.702 | 0.729 | 0.727 | 0.701 | 0.730 | 0.725 | 0.702 | 0.731 | 0.729 | 0.720 |
| | w/ CLIP-CFE | 0.700 | 0.731 | 0.725 | 0.702 | 0.732 | 0.726 | 0.699 | 0.732 | 0.724 | 0.700 | 0.730 | 0.725 | 0.719 |
| CLIP-CFA | w/o CLIP-CFE | 0.703 | 0.731 | 0.727 | 0.701 | 0.730 | 0.725 | 0.701 | 0.730 | 0.727 | 0.700 | 0.731 | 0.727 | 0.719 |
| | w/ CLIP-CFE | 0.702 | 0.729 | 0.729 | 0.701 | 0.730 | 0.727 | 0.702 | 0.731 | 0.725 | 0.702 | 0.730 | 0.727 | 0.720 |

与基线方法BadNets和Blended相比，CLIP-UAP和CLIP-CFA表现出相似甚至更好的平均良性准确率。
此外，CLIP-CFE不会对BA产生负面影响。

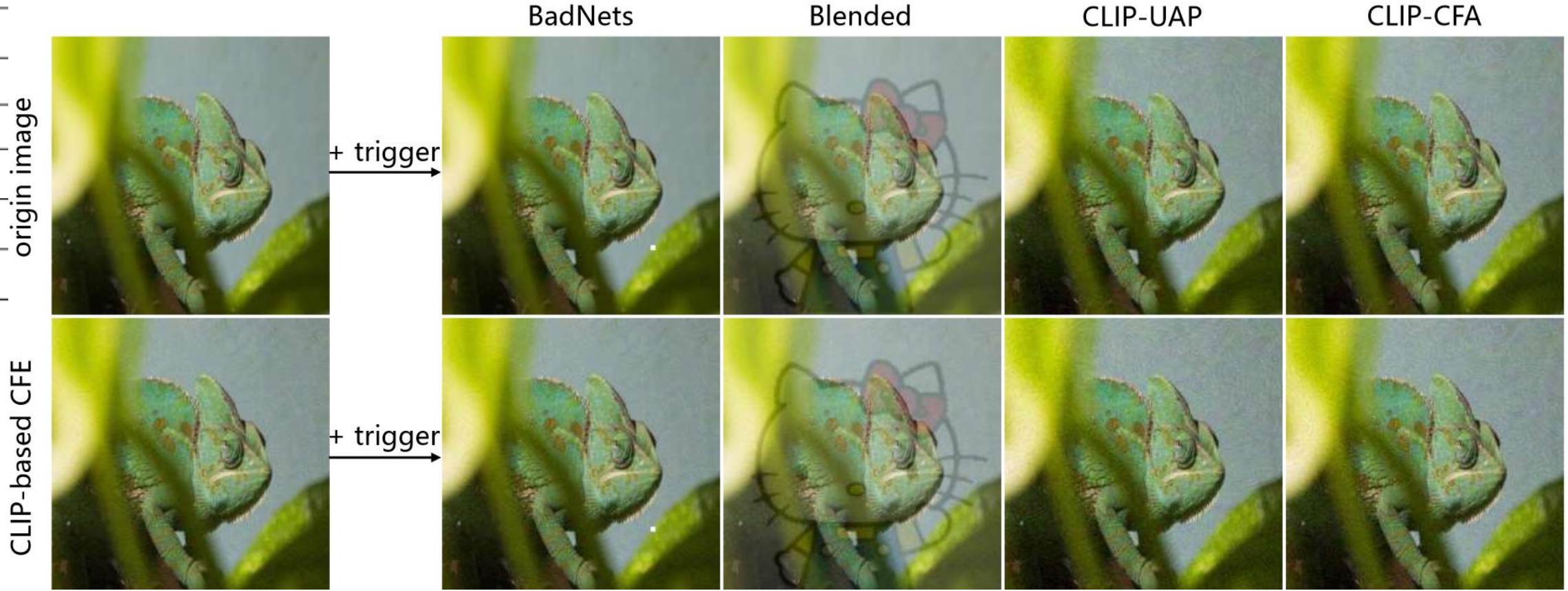
**即使在不同的设置和不同的后门攻击下，我们的技术对良性准确性也是无害的。**

# 问题3：提出的技术对受害者是否隐蔽?

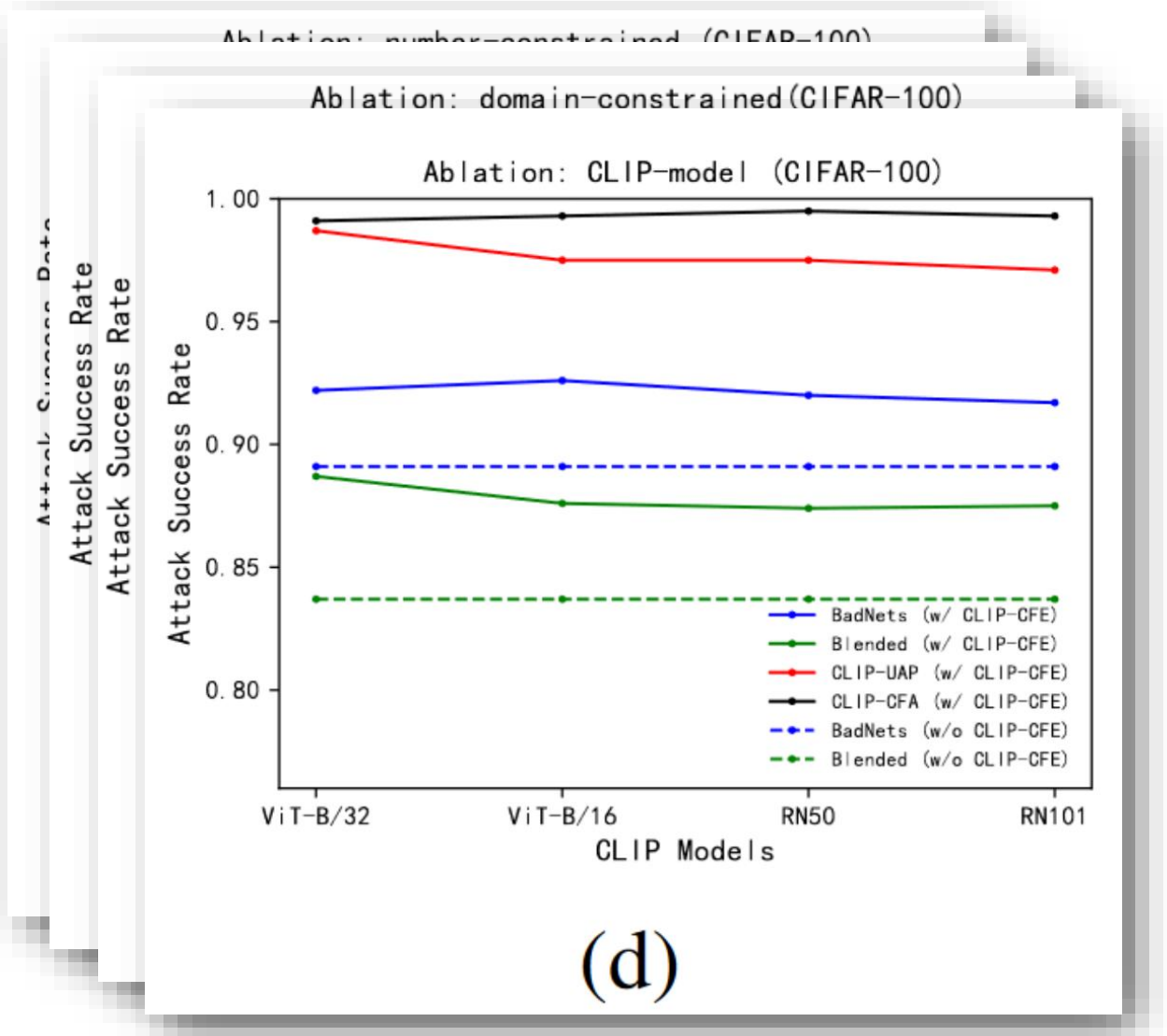
| Metrics | Trigger | Clean Feature Suppression | Backdoor Attacks | | | |
|---|---|---|---|---|---|---|
| | | | Number Constrained | Clean-label single-class | Dirty-label single-class | Domain Constrained |
| PSNR (↑) | BadNets | w/o CLIP-CFE | 32.18 | 31.26 | 30.91 | 32.22 |
| | | w/ CLIP-CFE | 32.19 | 31.41 | 31.39 | 32.17 |
| | Blended | w/o CLIP-CFE | 21.68 | 21.37 | 20.90 | 21.60 |
| | | w/ CLIP-CFE | 21.67 | 21.31 | 20.99 | 21.58 |
| | CLIP-UAP | | | | | |
| | CLIP-CFA | | | | | |
| SSIM (↑) | BadNets | | | | | |
| | Blended | | | | | |
| | CLIP-UAP | | | | | |
| | CLIP-CFA | | | | | |

虽然CLIP - UAP和CLIP - CFA在SSIM方面可能没有达到最高的隐蔽性，但本文的方法生成的中毒图像更自然地符合人类的检查。

如下图所示，展示了不同触发器中毒样本的可视化。



# 问题4：提出的技术在不同的中毒环境有效吗?




实验结果结果表明:

1)不同攻击的中毒率越高，攻击成功率越高;

2)CLIP-UAP和CLIP-CFA优于BadNets和Blended;

3)CLIP-CFE进一步提高了不同触发器的中毒效果;

4)论文提出的技术在不同的CLIP模型中都表现出稳定性。


# 总结展望

# 总结

本文应对数据受限后门攻击的挑战，这种攻击发生在更现实的场景中，即受害者从多个来源收集数据，而攻击者无法访问完整的训练数据。为了克服以往方法在数据受限后门攻击下的性能下降问题，我们从两个数据流中提出了三种技术，利用预先训练好的CLIP模型来提高中毒效率。

# 展望

1.Clean-Label后门攻击中的性能下降。我们的技术在Clean-Label后门攻击中显示出有限的改进，将探索更有效的专门针对干净标签后门攻击的攻击方法。

2.应用局限性。我们的技术依赖于在自然图像上预训练的CLIP模型，这可能会限制它们在某些特定领域的适用性，例如医学图像或遥感，可能使用特定领域的预训练模型来替换CLIP模型。

3.迁移到其他领域。攻击场景不局限于特定的领域，可以应用于其他重要的应用，包括用于恶意软件检测的后门攻击、深度伪造检测和联邦学习，将会设计切合实际的攻击场景和专门量身定制的高效后门攻击。

# 感谢指导