中图法分类号:TP18;TP37　　**文献标识码**:A　　**文章编号**:1006-8961(2026)01-0177-20

论文引用格式:Jiang G Z, Huang R, Liu H and Jiang X Q. 2026. Dataset condensation backdoor attack based on separated triggers and multiple comparisons. Journal of Image and Graphics, 31(1):0177-0196(蒋桂政, 黄荣, 刘浩, 蒋学芹. 2026. 分离触发器和多重对比的数据浓缩后门攻击. 中国图象图形学报, 31(1):0177-0196)[DOI:10.11834/jig.250083]

分离触发器和多重对比的数据浓缩后门攻击

蒋桂政[1],黄荣[1,2*],刘浩[1,2],蒋学芹[1,2]

1. 东华大学信息科学与技术学院,上海　201620; 2. 东华大学数字化纺织服装技术教育部工程研究中心,上海　201620

摘　要:**目的**　现有数据浓缩后门攻击方法将含有触发器的中毒样本和干净样本浓缩为小的数据集,中毒数据中真实数据的强信号掩盖触发器的弱信号,并且未考虑将非目标类浓缩数据与中毒数据特征分离,非目标类浓缩数据残留触发器特征。因此,提出分离触发器和多重对比的数据浓缩后门攻击。**方法**　首先将触发器与真实数据进行分离。分离的触发器作为样本与真实数据并行嵌入浓缩数据,减少真实数据对触发器的干扰。然后,对分离的触发器进行优化,将触发器接近目标类真实数据的特征,提高触发器的嵌入效果,同时对触发器进行了分区放大预处理来增加触发器像素的数量,使其在优化过程获取大量的梯度用于指导学习。在数据浓缩阶段,通过多重对比将目标类浓缩数据与触发器特征投影在同一空间,将非目标类浓缩数据与触发器特征分离,进一步提高后门攻击的成功率。**结果**　为了验证所提出方法的有效性,将所提出方法在FashionMNIST(Fashion Modified National Institute of Standards and Technology database)、CIFAR10(Canadian Institute for Advances Research's ten categories dataset)、STL10(Stanford letter-10)、SVHN(street view house numbers)与其他4种方法进行对比实验。所提出的方法在5个数据集和6个不同的模型上均达到100%的攻击成功率,同时未降低干净样本在模型上的准确率。**结论**　所提出的方法通过解决现有方法存在的问题,实现了性能的显著提高。本文方法具体代码见:https://github.com/tfuy/STMC。

关键词:后门攻击;数据浓缩;分离;梯度匹配;分区放大预处理;最大化

Dataset condensation backdoor attack based on separated triggers and multiple comparisons

Jiang Guizheng[1], Huang Rong[1,2*], Liu Hao[1,2], Jiang Xueqin[1,2]

1. College of Information Science and Technology, Donghua University, Shanghai　201620, China;

2. Engineering Research Center of Digitized Textile and Apparel Technology, Ministry of Education, Donghua University, Shanghai　201620, China

Abstract: **Objective** Existing backdoor attacks against dataset condensation condense poisoned data containing trigger and clean real data into a small dataset. However, the strong signal of real data in the poisoned data masks the weak signal of triggers (e. g. , square, noises, or warping) in the poisoned data, leading to the failure of the triggers embedded within the condensed data, which subsequently prevents the embedding of hidden backdoors into the model. Meanwhile, these methods only consider decreasing the distance between the gradients of target class condensed data and poisoned data and fail to consider increasing the distance between the gradients of nontarget class condensed data and poisoned data. Consequently,

收稿日期:2025-03-10;**修回日期**:2025-08-20;**预印本日期**:2025-08-27

*通信作者:黄荣　rong.huang@dhu.edu.cn

基金项目:国家自然科学基金项目(62001099);中央高校基本科研业务费专项资金项目(2232023D-30)

Supported by: National Natural Science Foundation of China (62001099); Fundamental Research Funds for the Central Universities (2232023D-30)

the nontarget condensed data contaminate trigger information, thereby reducing the attack success rate. To address the above issues, this study proposes a backdoor attack against dataset condensation based on separated triggers and multiple comparisons. The proposed method optimizes the separated triggers to minimize the interference from real data during the embedding process. The poisoned condensed data generated by the method exhibit a high attack success rate while making it difficult to detect the embedded triggers. **Method** To help the model focus on the trigger, this study separates the trigger from real data. The separated trigger is embedded as a sample into the condensed data in parallel with real data to mitigate the interference of real data on the trigger. This study optimizes the separated trigger through a gradient matching method to bring the trigger close to the feature of target class real data. When partitioned amplification preprocessing is applied to the trigger, the number of trigger pixels is increased, enabling the acquisition of abundant gradients during the optimization process to guide learning. In the data condensation phase, while the distance between the gradients of target class condensed data and the trigger is minimized, the distance between the gradients of nontarget class condensed data and the trigger is considered to be maximized, eliminating the trigger information contaminated in the condensed data of nontarget classes. To fully exploit trigger information and embed it into the target class condensed data, this study aims to reduce the distance between the error gradients of the trigger and the target class condensed data, further enhancing the success rate of the backdoor attack. **Result** For validating the effectiveness of the method proposed in this paper, it was compared with four other methods on four datasets: Fashion Modified National Institute of Standards and Technology database (FashionMNIST), Canadian Institute for Advances Research's ten categories dataset (CIFAR10), Stanford letter-10 (STL10), and street view house numbers (SVHN). The proposed method achieves significant improvement in poisoned accuracy compared with the existing methods, with respective increases of 20%, 47.7%, 30%, and 6.5% over NaiveAttack, DoorPing, DoorPing-1, and DoorPing-2, respectively, while maintaining the accuracy of the model on clean samples without degradation. The poisoned condensation data synthesized by our proposed method maintain remarkably consistent poisoned accuracy when applied to train models with different architectures. To validate the stealthiness of our approach, this study synthesizes clean and poisoned condensation data across ConvNet and AlexNet architectures. Owing to the traces left by the backpropagation process, the poisoned and clean condensed data become indistinguishable blurred images. These traces cover the trigger, making it difficult for people to determine which set of condensed data contains the embedded trigger. To verify the effectiveness of the proposed method on complex datasets and models, this study synthesizes a set of poisoned condensation data using the CIFAR100 dataset and VGG11 architecture and evaluates poisoned accuracy. Experimental results demonstrate that the poisoned accuracy remains 100% even under complex datasets and model frameworks. Notably, when the poisoned condensed data trained on VGG11 are used to train models with other architectures, the poisoned accuracy rate degrades. Complex models typically possess powerful feature extraction capabilities, which embed substantial noise captured during training into the condensed data, leading to a decline in poisoned accuracy. **Conclusion** The data condensation backdoor attack method proposed in this paper significantly improves the attack success rate, while the embedded triggers are difficult to detect. This method possesses good cross-model generalization ability and can be effectively extended to complex datasets. Compared with traditional backdoor attack methods, the backdoor attack based on condensed data has unique advantages. By treating a trigger as an independent sample alongside real data to match the gradient of condensed data, the proposed method mitigates the interference caused by strong signals from real data. Through aligning the trigger's feature representation with that of target class real data, we reduce the difficulty of coembedding triggers and real data into condensed data. Furthermore, this study conducts gradient separation between the trigger and condensed data of nontarget classes via multiple comparisons, removing residual trigger features embedded in the nontarget condensed data and thereby significantly improving the success rate of backdoor attacks. The code is linked at https://github.com/tfuy/STMC.

Key words: backdoor attack; dataset condensation; separation; gradient matching; partitioned amplification preprocessing; maximize

0 引 言

深度神经网络(deep neural network，DNN)已广泛应用于各个领域，如计算机视觉(Voulodimos 等，2018)、自然语言处理(Young 等，2018)以及目标检测(Girshick 等，2014)等。深度学习模型的成功离不开大规模精心标注的训练数据。然而，标注大规模的训练数据将耗费大量人工成本，因此，用户倾向于从互联网上收集已预先标注的训练数据。攻击者向这些训练数据混入一些恶意数据，从而为用户模型造成安全隐患。后门攻击(Gu 等，2017；Salem 等，2022；Yao 等，2019)是一种数据投毒攻击。攻击者将精心设计的触发器注入到干净数据中以构造中毒数据，然后将少量的中毒数据混入模型的训练集中，从而在模型的训练过程中向其植入隐藏后门。后门植入完成后，中毒模型的目标分类任务性能良好，而当输入含有触发器的中毒数据时，触发器将激活模型中的隐藏后门出现错误的分类结果。后门攻击的研究有助于推动防御的发展，同时可以反向应用于模型检索(Liu 等，2023a，2025)、验证码等领域，实现隐私保护。

数据浓缩(Cazenavette 等，2022；Nguyen 等，2021a，b；Wang 等，2022；Zhao 和 Bilen，2023)作为一个新的研究领域，旨在合成一组规模较小的浓缩数据，其包含原始真实数据的关键信息。采用浓缩数据进行模型训练，模型专注关键信息，从而提高训练效率，降低训练成本。相较于真实数据，由于浓缩数据较小，在模型训练方面所需的计算成本较低，对用户具有较大的吸引力。这些中毒浓缩数据含有攻击者嵌入的触发器，从而对用户模型构成安全隐患。浓缩数据通过梯度下降算法所合成。如图 1 所示，现有一些数据浓缩方法(Liu 等，2023b；Chen 等，2017)通过最小化真实数据和浓缩数据梯度之间的距离来合成浓缩数据。在训练过程中干净图像对浓缩图像产生较大的修改。该方法利用浓缩图像较大的修改掩盖触发器对浓缩图像较小的修改，以实现触发器的隐蔽性。

现有数据浓缩后门攻击方法(Liu 等，2023b)主要存在两个问题。1)触发器失活。这些攻击方法先将触发器注入到部分干净的真实数据，然后，将这些构造的数据浓缩为一组数量较小的数据集。触发器

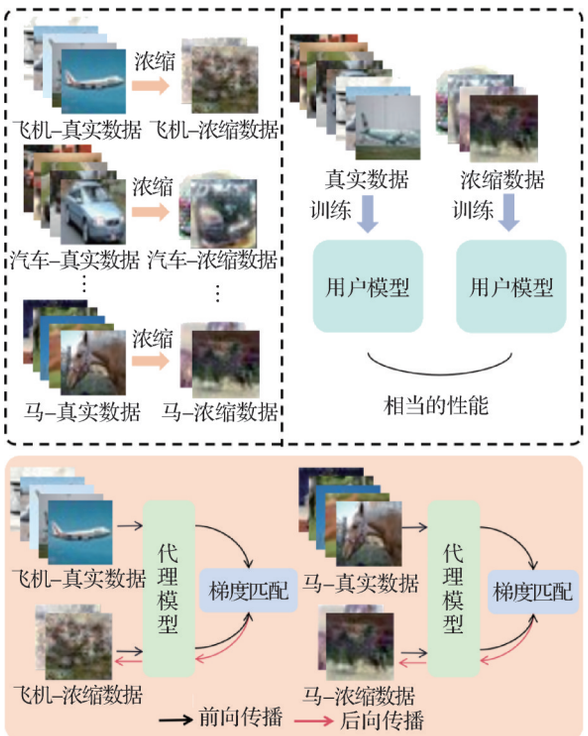

图 1　数据浓缩概述

Fig. 1　Overview of dataset condensation

与真实数据之间是叠加式，即触发器的像素值相加至真实数据。由于触发器往往是一个小色块或是一个随机噪声，触发器的信号强度相比真实数据较弱。真实数据的强信号掩盖触发器的弱信号，导致无法将触发器嵌入浓缩数据。2)非目标类浓缩数据混杂触发器特征。现有攻击方法未考虑将非目标类浓缩数据和中毒数据梯度分离，导致非目标类浓缩数据残留部分触发器特征。这些特征干扰模型训练过程，使模型误将触发器与非目标类别进行绑定，进而导致后门攻击失败。

为了解决上述问题，本文提出一种分离触发器和多重对比的数据浓缩后门攻击。触发器与真实数据之间是分离式，即触发器作为样本与真实数据并行输入，单独接收梯度。在数据浓缩过程中，触发器单独与浓缩数据进行梯度匹配，与真实数据的梯度匹配过程保持独立。这一做法有以下两个优势：1)防止触发器失活；2)无需对触发器进行定位，从而便于触发器的更新训练。在触发器训练阶段，触发器与目标类真实数据投影在同一特征空间，利用两者特征的相似性实现触发器和真实数据共同浓缩。同时，在触发器的学习过程中，触发器进行分区放大预处理以增加触发器的像素数量，使其在学习过程中

获取大量的梯度进行更新训练。在数据浓缩阶段，本文通过多重对比将目标类浓缩数据与触发器进行特征匹配，将非目标类浓缩数据和触发器进行特征分离，以去除非目标类浓缩数据中混杂的部分触发器特征。为了验证该方法的有效性，本文在6个模型、5个基准数据集上进行实验。结果表明，该方法获得100%的攻击成功率，同时干净准确率未出现下降，优于现有后门攻击方法。

1 相关工作

1.1 后门攻击

Gu等人(2017)提出第1个后门攻击方法BadNet(backdoored neural network)，将白色方块作为触发器覆盖到干净图像的固定位置。然而，触发器的颜色和形状在这些干净图像中较明显，容易暴露后门攻击的存在。不少学者研究不可见的后门攻击。不可见后门攻击是指将触发器以隐藏的方式注入到干净图像，使其在视觉上不可察觉。Chen等人(2017)首先研究了这种攻击，通过把触发器的像素叠加到干净图像的像素上，用干净图像的像素掩盖触发器，从而提高后门攻击的隐蔽性。Li等人(2021a)则通过LSB(least significant bit)(Mielikäinen, 2006；朱淑雯 等, 2023)图像隐写技术将触发器不可见地嵌入干净图像。值得关注的是，Nguyen等人(2021)认为人眼对图像中的微小扭曲往往难以察觉，并利用插值算法使干净图像产生微小扭曲以实现触发器的嵌入。Jiang等人(2023)利用粒子算法搜索触发器，将其叠加到部分干净图像，构造的图像与正常的干净图像在对比度和亮度方面保持一致。Gong等人(2023)通过中毒模型的部分神经元，以提高后门的隐蔽性。

1.2 数据浓缩

数据集简约和数据浓缩都旨在通过减少数据规模提高模型训练效率。核心集选择(Agarwal等, 2004；Chen等, 2010)是一种传统的数据集简约方法，根据预先定义的准则(如，紧凑性(Rebuffi等, 2017)、多样性(Aljundi等, 2019；Sener和Savarese, 2018)、遗忘性(Toneva等, 2019))从原始数据集中挑选出具有代表性的数据子集。由于这些数据子集仅包含原始数据集的部分信息，导致在处理下游任务时效果较差。数据浓缩旨在合成一组数量较小的浓缩数据，其包含原始数据的关键信息，以保证使用浓缩数据训练模型可以获得与原始真实数据同样的性能。Wang等人(2020)提出数据浓缩方法DD(dataset distillation)(Wang等, 2020)，通过最大化浓缩数据所训练模型的分类准确率来合成浓缩数据。受该方法的启发，Zhao等人(2021)提出基于梯度匹配的数据浓缩方法DC(dataset condensation)，通过最小化浓缩数据和真实数据之间的梯度匹配损失，将真实数据浓缩一组数量较小的数据集。

1.3 数据浓缩后门攻击

现有数据浓缩后门攻击方法是由Liu等人(2023b)提出的NaiveAttack和DoorPing。NaiveAttack先将预先设计的触发器的像素叠加到部分真实数据以构造中毒数据，然后利用现有的数据浓缩方法将中毒样本和干净样本浓缩为一组浓缩数据。然而，若触发器设计不当，触发器的信息占比随着数据浓缩的进行而下降，导致后门攻击成功率下降。为了解决上述问题，Liu等人(2023b)提出基于触发器迭代优化的DoorPing。DoorPing在每轮数据浓缩前对触发器进行微调，确保在每轮浓缩过程中都关注到触发器。然而，该方法未能将触发器与真实数据分离，真实数据强信号掩盖触发器弱信号，导致触发器失活，同时，未考虑将非目标类浓缩数据与触发器进行梯度分离，导致非目标类浓缩数据混杂触发器特征。

2 方 法

2.1 威胁模型

数据浓缩后门攻击主要通过向浓缩数据中嵌入触发器，从而为用户模型构成安全威胁。一种情况是：用户委托攻击者合成浓缩数据，从而向其嵌入触发器；另一种常见的情况是：攻击者在网络上发布注入触发器的中毒浓缩数据，用户受其在模型训练方面优势的吸引下载数据用于模型训练。

1)攻击者的能力。本文假设攻击者控制数据浓缩过程，但是没有任何模型的训练信息。同时，在测试阶段，攻击者能够向模型输入带有触发器的中毒数据。

2)攻击者的目标。攻击者首要目标是确保后门攻击的有效性，使用中毒浓缩数据进行训练后，模型可以将输入的中毒数据识别为攻击者指定的目标类

别。同时，模型在干净测试数据上的分类准确率未产生下降。其次，攻击者应确保中毒浓缩数据中触发器的不可见性，并能够逃避防御方法的检测，以保证后门攻击的隐蔽性。

2.2 问题定义

在图像分类任务中，其目标是训练一个分类函数 $f: \boldsymbol{X} \to \boldsymbol{Y}$，其中，$\boldsymbol{X}$ 表示输入图像域，$\boldsymbol{Y} = \{0, 1, \cdots, K-1\}$ 表示 K 个分类类别。分类函数 f 使用干净训练数据集 $\boldsymbol{\mathcal{X}} = \left\{\left(\boldsymbol{x}_i, y_i\right)\right\}_{i=1}^{N}$ 进行训练。其中，$\boldsymbol{x}_i \in \boldsymbol{X}$ 表示一幅干净训练图像，$y_i \in \boldsymbol{Y}$ 表示该图像对应的标签。训练后的 f 对输入的干净图像 $\boldsymbol{x} \in \boldsymbol{X}$ 识别为对应标签 $y \in \boldsymbol{Y}$。攻击者通过将部分干净训练数据 $\left(\boldsymbol{x}_i, y_i\right)$ 替换为中毒训练数据 $\left(\hat{\boldsymbol{x}}_i, c\right)$，从而使模型中毒。这些中毒训练图像 $\hat{\boldsymbol{x}}_i$ 是通过向干净训练图像 $\boldsymbol{x}_i$ 应用触发器注入函数 $\mathcal{B}$，具体为

$$\hat{\boldsymbol{x}}_i = \mathcal{B}(\boldsymbol{x}_i) = \boldsymbol{x}_i \odot (1 - \boldsymbol{m}) + \boldsymbol{P} \odot \boldsymbol{m} \tag{1}$$

式中，$\odot$ 表示逐像素乘法，$\boldsymbol{P}$ 表示触发器，$\boldsymbol{m}$ 表示掩码，即图像被触发器覆盖的像素集合。目标标签 $y \in \boldsymbol{Y}$ 则由攻击者所指定。

基于梯度匹配的数据浓缩方法 DC（Zhao 等，2021）的目标是合成一组浓缩数据 $\boldsymbol{\mathcal{S}} = \left\{\left(\boldsymbol{s}_i, y_i\right)\right\}_{i=1}^{M}$，其所训练模型的更新方向遵循在真实数据集 $\boldsymbol{\mathcal{X}} = \left\{\left(\boldsymbol{x}_i, y_i\right)\right\}_{i=1}^{N}$，$N \gg M$ 上的更新方向。该方法首先逐类计算真实数据与浓缩数据梯度之间的距离并求和，然后最小化这个距离总和以实现该目标。该方法的优化问题表达为

$$\begin{aligned} &\arg\min_{\boldsymbol{s}} \sum_{y=0}^{K-1} D\left(\nabla_{\boldsymbol{\theta}_t} \ell\left(\boldsymbol{\mathcal{X}}^y, y; \boldsymbol{\theta}_t\right), \nabla_{\boldsymbol{\theta}_t} \ell\left(\boldsymbol{\mathcal{S}}^y, y; \boldsymbol{\theta}_t\right)\right) \\ &\text{s.t.} \quad \boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \eta \nabla_{\boldsymbol{\theta}} \ell\left(\boldsymbol{\mathcal{S}}; \boldsymbol{\theta}_{t-1}\right) \end{aligned} \tag{2}$$

式中，$\boldsymbol{\theta}_t$ 表示随机初始化的代理模型 $\boldsymbol{\theta}_0$ 在 $\boldsymbol{\mathcal{S}}$ 上第 t 次迭代后的参数。$D(\cdot, \cdot)$ 和 $\ell(\cdot, \cdot)$ 分别表示距离函数和交叉熵损失函数。K 表示类别的数量，∇ 表示损失函数最大减小的方向，η 表示模型更新的学习率。

现有数据浓缩后门攻击需要攻击者先利用触发器注入函数 $\mathcal{B}$ 向部分待浓缩的真实数据 $\boldsymbol{x}_i$ 中注入触发器 $\boldsymbol{P}$，并修改其标签为目标标签 c。然后，使用污染后的真实数据集 $\hat{\boldsymbol{\mathcal{X}}}$ 代替原始真实数据集 $\boldsymbol{\mathcal{X}}$ 进行数据浓缩。与现有方法不同的是，本文没有使用触发器注入函数 $\mathcal{B}$ 将触发器 $\boldsymbol{P}$ 注入真实数据中，而是将触发器 $\boldsymbol{P}$ 单独作为样本与真实数据一起进行数据浓缩，防止触发器失活。同时，本文考虑对非目标类的浓缩数据 $\boldsymbol{\mathcal{S}}^{y \neq c}$ 与触发器进行梯度分离，将非目标类浓缩数据 $\boldsymbol{\mathcal{S}}^{y \neq c}$ 中混杂的触发器特征去除。

2.3 分离触发器和多重对比的中毒浓缩数据合成

2.3.1 分离触发器

现有数据浓缩后门攻击方法遵循传统的后门攻击将触发器注入部分真实数据，然后在数据浓缩阶段，将这些注入触发器的真实数据浓缩为一组数量较小的数据集。由于触发器与真实数据处于叠加式，并且触发器输入信号较弱，真实数据的强信号掩盖触发器的弱信号，导致触发器失活。为了解决上述问题，触发器与真实数据相分离，使触发器与真实数据保持分离式，触发器作为样本与真实数据并行输入，单独接收梯度。如图 2 所示，在中毒浓缩数据训练阶段，触发器与目标类真实数据并行输入给模型计算梯度，并分别与浓缩数据进行梯度匹配。触发器与浓缩数据的梯度匹配过程和真实数据与浓缩数据梯度匹配过程保持并行，从而减小真实数据对触发器嵌入过程的干扰。

本文对分离的触发器进行了优化。如图 2 所示，在触发器训练阶段，触发器接近目标类真实数据的特征，通过特征一致性降低将触发器和真实数据共同嵌入浓缩数据的难度。同时，触发器与非目标真实数据进行特征分离。具体而言，随机初始化的触发器 $\boldsymbol{P}$ 在模型上的正确梯度（将触发器指定为目标标签 c）接近目标类真实数据 $\boldsymbol{\mathcal{X}}^c$ 上的梯度，从而使触发器接近目标类真实数据的特征。同时，触发器 $\boldsymbol{P}$ 在模型上的错误梯度（将触发器指定为非目标标签 $y \neq c$）远离非目标真实数据 $\boldsymbol{\mathcal{X}}^{y \neq c}$ 上的梯度。

在触发器的学习过程中，如图 3 所示，触发器进行分区放大预处理来增加触发器的像素个数，再将触发器与真实数据进行梯度匹配。具体而言，本文先将触发器 $\boldsymbol{P}$ 分成 4 个块，并利用双线性插值算法将每一个块放大到 $\boldsymbol{P}$ 的大小。然后，每个块依次与真实数据进行梯度匹配，触发器学习完成。同样利用双线性插值算法重新将这 4 个块组合缩小到 $\boldsymbol{P}$ 的大小。这种分区放大的做法增加了触发器的像素数量。由于在触发器的学习过程中，触发器的每个像素获得一个梯度用于指导触发器学习。相比未分区的触发器，分区后的触发器在学习过程中接收到 4 倍的梯度数量，从而获得丰富的信息用于指导触发器学习。

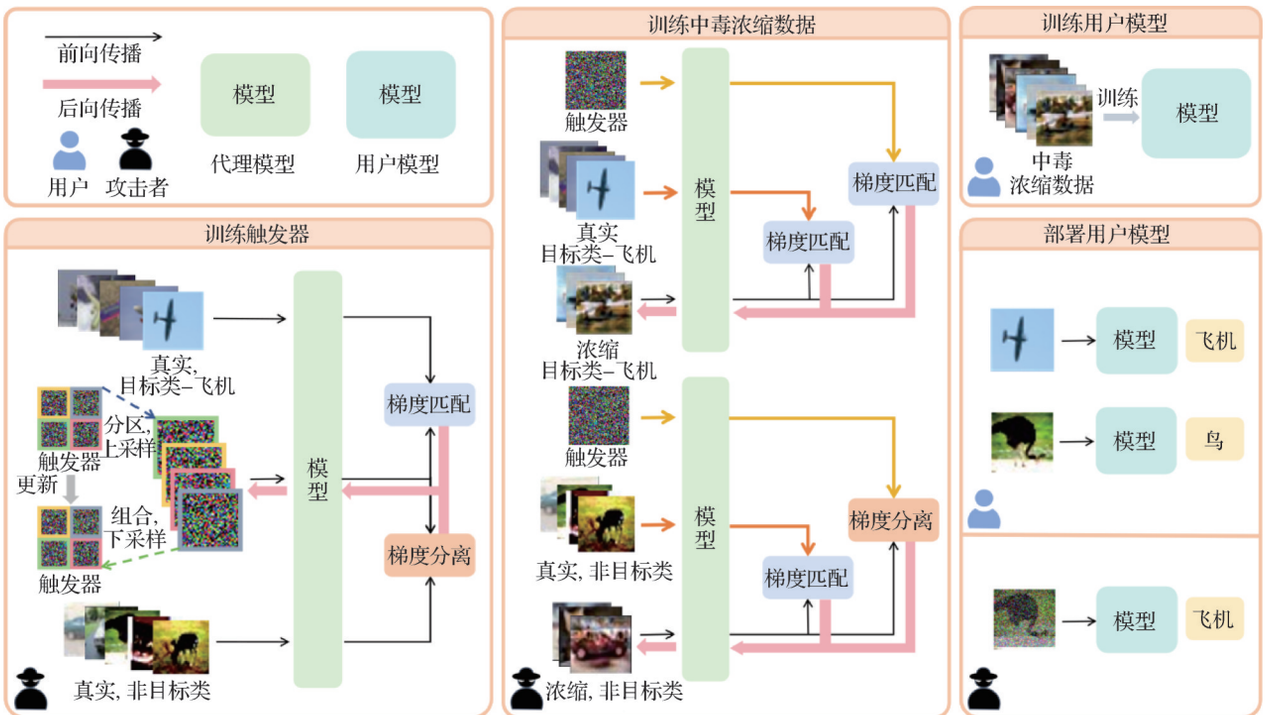

图2　所提出攻击方法的概述

Fig. 2　Overview of the attack method proposed

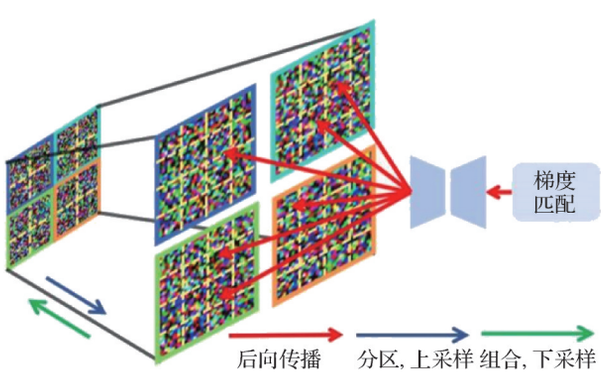

图3　触发器在优化阶段获取的梯度

Fig. 3　The gradient obtained by trigger in the optimization phase

2.3.2　多重对比的触发器嵌入

现有数据浓缩后门攻击仅考虑将目标类浓缩数据与触发器特征进行匹配，从而将触发器嵌入浓缩数据，但没有考虑将非目标类浓缩数据与触发器特征分离，导致非目标类浓缩数据混杂部分触发器特征。为解决上述问题，如图2所示，在中毒浓缩数据训练阶段，目标类浓缩数据与触发器进行梯度匹配，非目标类浓缩数据与触发器进行梯度分离，通过这种多重对比的方式将触发器嵌入浓缩数据，以去除非目标类浓缩数据混杂的触发器特征。具体而言，本文先将触发器$\boldsymbol{P}$上的错误梯度（将触发器指定为非目标标签$y \neq c$）进行取反，然后将非目标类浓缩数据$\boldsymbol{\mathcal{S}}^{y \neq c}$上的梯度接近这个修正梯度，实现非目标类浓缩数据与触发器特征分离。浓缩数据的优化目标函数$\mathcal{L}_1$表达为

$$\mathcal{L}_1 = \sum_{y \neq c} D\Big(-\alpha\nabla_\theta \ell\big(\boldsymbol{P}, y, \boldsymbol{\theta}\big) + \nabla_\theta \ell\big(\boldsymbol{\mathcal{X}}^y, y, \boldsymbol{\theta}\big), \nabla_\theta \ell\big(\boldsymbol{\mathcal{S}}^y, y, \boldsymbol{\theta}\big)\Big) + D\Big(\alpha\nabla_\theta \ell\big(\boldsymbol{P}, c, \boldsymbol{\theta}\big) + \nabla_\theta \ell\big(\boldsymbol{\mathcal{X}}^c, c, \boldsymbol{\theta}\big), \nabla_\theta \ell\big(\boldsymbol{\mathcal{S}}^c, c, \boldsymbol{\theta}\big)\Big) \tag{3}$$

式中，$D(\cdot,\cdot)$和$\ell(\cdot,\cdot)$分别表示距离函数和交叉熵损失函数。∇表示损失函数最大减小的方向，α表示浓缩数据与触发器进行梯度匹配的占比，c表示目标类别。

值得注意的是，上述方法通过将目标类浓缩数据$\boldsymbol{\mathcal{S}}^c$上的正确梯度（将目标类浓缩数据指定为目标标签$c$）接近触发器$\boldsymbol{P}$上的正确梯度，从而将触发器嵌入目标类浓缩数据。如图4所示，本文同时将目标类浓缩数据$\boldsymbol{\mathcal{S}}^c$上的错误梯度（将目标类浓缩数据指定为非目标标签$y \neq c$）接近触发器$\boldsymbol{P}$上的错误梯度，以深入地挖掘触发器的信息并将其嵌入浓缩数据。浓缩数据的优化目标函数进一步修改为

$$\mathcal{L}_{\text{all}} = \sum_{y=0}^{K-1} D(\nabla_\theta \ell\big(\boldsymbol{\mathcal{X}}^c, y, \boldsymbol{\theta}\big) + \alpha\nabla_\theta \ell\big(\boldsymbol{P}, y, \boldsymbol{\theta}\big), \nabla_\theta \ell\big(\boldsymbol{\mathcal{S}}^c, y, \boldsymbol{\theta}\big)) + \mathcal{L}_1 \tag{4}$$

式中，$D(\cdot,\cdot)$和$\ell(\cdot,\cdot)$分别表示距离函数和交叉熵损失函数。∇表示损失函数最大减小的方向，α表示浓缩数据与触发器进行梯度匹配的占比，c表示目标类别。最后，本文通过最小化该目标函数$\mathcal{L}_{\text{all}}$来合成中毒浓缩数据。具体为

$$\mathcal{S}^* = \arg\min_{\mathcal{S}} \mathcal{L}_{\text{all}} \tag{5}$$

式中，$\mathcal{L}_{\text{all}}$表示合成中毒浓缩数据的目标函数，$\mathcal{S}^*$表示$\mathcal{L}_{\text{all}}$最小化时的中毒浓缩数据。

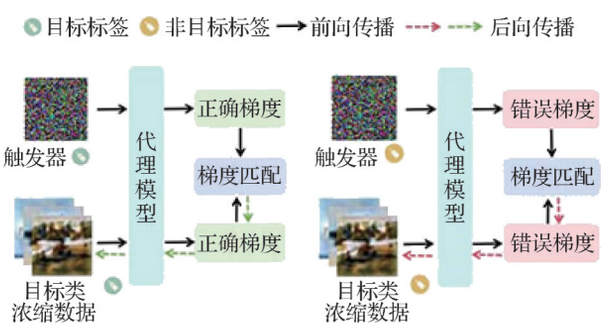

图4　将触发器嵌入目标类浓缩数据

Fig. 4　Embedding trigger into condensed data of the target class

3　实　验

3.1　实验设置

3.1.1　数据集

实验在5个广泛应用的基准数据集上进行。

1）FashionMNIST（Fashion Modified National Institute of Standards and Technology database）（Xiao 等，2017）包含70 000幅尺寸为28×28像素的灰度图像。其中，训练集和测试集分别包含60 000幅和10 000幅图像。该数据集包含10个类别，包括T恤、裤子、套头衫、连衣裙、外套、凉鞋、衬衫、运动鞋、包和短靴。

2）CIFAR10（Canadian Institute for Advances Research's ten categories dataset）（Krizhevsky 和 Hinton，2009）包含60 000幅尺寸为32×32像素的彩色图像。其中，训练集和测试集分别包含50 000幅和10 000幅图像。该数据集包含10个类别，包括飞机、汽车、鸟、猫、鹿、狗、青蛙、马、船和卡车。

3）STL10（Stanford letter-10）（Coates 等，2011）包含13 000幅尺寸为96×96像素的彩色图像。其中，训练集和测试集分别包含5 000幅和8 000幅图像。该数据集包含10个类别，包括飞机、鸟、汽车、猫、鹿、狗、马、猴子、船和卡车。

4）SVHN（street view house numbers）（Netzer 等，2011）是一个数字分类基准数据集，包含99 289幅尺寸为32×32像素的彩色门牌号码图像。其中，训练集和测试集分别包含73 257幅和26 032幅图像。

3.1.2　深度神经网络架构

实验选择6种标准深度神经网络架构来合成中毒浓缩数据，包括卷积神经网络ConvNet（convolutional neural network）（Gidaris 和 Komodakis，2018）、AlexNet（Krizhevsky 等，2011）、LeNet（LeCun 等，1998）、ZFNet（Zeiler 和 Fergus，2014）、CaffeNet 和 VGG11（Visual Geometry Group 11-layer network）（Simonyan 和 Zisserman，2015），这些网络在浓缩数据领域得到广泛应用。其中，ConvNet包含D个重复块，每个块包含W个过滤器(3×3)、一个归一化层N、一个激活层A和一个池化层P，记为$[W;\ N;\ A;\ P]\times D$。标准的ConvNet网络通常包含3个块，每个块包含128个过滤器。

3.1.3　超参数

本文遵循DC方法（Zhao 等，2021）中大多数的超参数设置。唯一例外的是，针对VGG模型的实验研究，浓缩数据的SGD（stochastic gradient descent）优化器学习率由常见的0.1调整为0.001，以提高分类精度和攻击成功率。针对该方法的一些超参数设置，触发器的梯度占比α设置为0.007，针对VGG模型的实验研究，将触发器的梯度占比α设置为0.5。在对触发器进行分区放大预处理时，触发器分成4块。

3.1.4　评估

使用中毒准确率和干净准确率作为评估指标。中毒准确率是指后门模型在中毒测试数据集上的准确率。干净准确率是指后门模型在干净测试数据集上的准确率。

3.1.5　补充细节

数据集的每个类别合成10个浓缩数据。初始化一个和训练数据同样大小的随机噪声作为触发器。对于浓缩数据集，本文则通过从训练数据集中随机采样进行初始化。每组实验合成2组浓缩数据集。随后，每组浓缩数据集从头训练30个随机初始化模型，并报告其在测试数据集上的中毒准确率和干净准确率。其中，模型使用浓缩数据训练300轮。

3.2　与现有方法的对比

该方法与现有的4种基于数据浓缩后门攻击方

法(NaiveAttack(2023)、DoorPing(2023)、DoorPing-1、DoorPing-2)进行对比实验。其中,NaiveAttack先将触发器的像素叠加到部分真实数据的像素上以构造中毒数据,然后将中毒数据和干净数据凝练为浓缩数据;DoorPing在每轮数据浓缩前对触发器进行微调,确保在每轮浓缩过程中关注到触发器;DoorPing-1仅在数据浓缩的前100轮对触发器做微调;DoorPing-2采用本文所提出的触发器训练方法,将触发器接近目标类真实数据的特征。表1展示了该方法与现有方法在性能上的对比。

表1　与现有数据浓缩后门攻击方法的对比

Table 1　Comparison with existing backdoor attack methods based on dataset condensation

网络	模型	Fashion MNIST		CIFAR10		STL10		SVHN	
		干净准确率	中毒准确率	干净准确率	中毒准确率	干净准确率	中毒准确率	干净准确率	中毒准确率
ConvNet	干净模型	0.728±0.005	–	0.312±0.022	–	0.314±0.011	–	0.148±0.021	–
	DoorPing	0.699±0.010	0.415±0.043	0.281±0.016	0.561±0.188	0.280±0.020	0.428±0.148	0.155±0.023	0.504±0.087
	NaiveAttack	**0.729±0.013**	0.671±0.015	0.287±0.018	0.606±0.080	0.284±0.015	0.639±0.054	**0.162±0.019**	0.769±0.059
	DoorPing-1	0.720±0.006	0.690±0.069	0.300±0.014	0.832±0.047	0.289±0.018	0.617±0.058	0.155±0.019	0.837±0.067
	DoorPing-2	0.710±0.016	**1.000±0.000**	0.291±0.018	**1.000±0.000**	0.289±0.012	**1.000±0.000**	0.158±0.04	**1.000±0.000**
	本文	0.712±0.005	**1.000±0.000**	**0.322±0.011**	**1.000±0.000**	**0.321±0.007**	**1.000±0.000**	0.161±0.009	**1.000±0.000**
AlexNet	干净模型	0.742±0.026	–	0.365±0.005	–	0.363±0.006	–	**0.555±0.032**	–
	DoorPing	0.750±0.010	0.755±0.055	0.346±0.014	0.838±0.077	0.347±0.019	0.519±0.091	0.484±0.041	0.391±0.121
	NaiveAttack	0.747±0.014	0.772±0.031	0.355±0.017	0.819±0.068	0.350±0.019	0.810±0.019	0.485±0.032	0.740±0.036
	DoorPing-1	**0.763±0.008**	0.826±0.033	0.341±0.025	0.884±0.041	0.352±0.015	0.701±0.047	0.503±0.048	0.661±0.082
	DoorPing-2	0.727±0.015	**1.000±0.000**	0.344±0.03	0.902±0.046	0.348±0.014	0.965±0.029	0.513±0.031	0.935±0.027
	本文	0.729±0.012	**1.000±0.000**	**0.368±0.019**	**1.000±0.000**	**0.365±0.009**	**1.000±0.000**	0.515±0.036	**0.999±0.005**
LeNet	干净模型	0.712±0.006	–	0.283±0.016	–	0.284±0.013	–	0.262±0.078	–
	DoorPing	**0.724±0.009**	0.325±0.063	0.263±0.018	0.120±0.014	0.262±0.017	0.212±0.103	0.258±0.039	0.169±0.049
	NaiveAttack	0.712±0.010	0.687±0.058	0.266±0.020	0.667±0.071	0.268±0.022	0.642±0.044	0.273±0.038	0.606±0.047
	DoorPing-1	0.709±0.013	0.532±0.076	0.258±0.019	0.517±0.048	0.268±0.020	0.728±0.091	**0.278±0.040**	0.455±0.065
	DoorPing-2	0.700±0.015	0.955±0.013	0.274±0.018	0.921±0.018	0.276±0.020	0.932±0.029	0.260±0.049	0.981±0.011
	本文	0.691±0.01	**1.000±0.000**	**0.308±0.016**	**1.000±0.000**	**0.306±0.025**	**1.000±0.000**	0.251±0.078	**1.000±0.000**
ZFNet	干净模型	0.735±0.009	–	0.354±0.034	–	**0.366±0.017**	–	0.500±0.037	–
	DoorPing	0.747±0.009	0.579±0.034	0.340±0.044	0.381±0.088	0.342±0.018	0.657±0.037	0.490±0.037	0.344±0.097
	NaiveAttack	0.751±0.005	0.856±0.019	0.35±0.017	0.884±0.043	0.343±0.019	0.84±0.033	0.513±0.033	0.828±0.031
	DoorPing-1	0.751±0.005	0.830±0.018	0.345±0.022	0.926±0.026	0.353±0.011	0.867±0.040	0.517±0.044	0.768±0.032
	DoorPing-2	**0.761±0.008**	**1.000±0.000**	0.365±0.013	0.966±0.021	0.348±0.021	0.919±0.021	0.519±0.035	0.948±0.030
	本文	0.735±0.013	**1.000±0.000**	**0.378±0.019**	**1.000±0.000**	0.366±0.013	**1.000±0.000**	**0.522±0.021**	**1.000±0.000**
CaffeNet	干净模型	0.745±0.008	–	**0.357±0.012**	–	**0.357±0.009**	–	0.461±0.032	–
	DoorPing	**0.757±0.005**	0.266±0.022	0.346±0.015	0.803±0.055	0.354±0.009	0.613±0.036	0.466±0.024	0.562±0.058
	NaiveAttack	0.740±0.007	0.814±0.040	0.342±0.015	0.874±0.022	0.348±0.01	0.854±0.038	0.453±0.032	0.758±0.045
	DoorPing-1	0.755±0.044	0.657±0.033	0.346±0.012	0.908±0.011	0.351±0.011	0.792±0.045	0.465±0.025	0.637±0.048
	DoorPing-2	0.750±0.004	0.932±0.014	0.349±0.013	0.961±0.022	0.356±0.014	0.954±0.015	0.450±0.033	0.987±0.011
	本文	0.728±0.008	**1.000±0.000**	0.352±0.014	**1.000±0.000**	0.350±0.012	**1.000±0.000**	**0.481±0.038**	**1.000±0.000**
VGG11	干净模型	0.701±0.014	–	0.269±0.005	–	0.262±0.007	–	0.314±0.014	–
	DoorPing	**0.708±0.005**	0.113±0.011	0.243±0.008	0.113±0.056	0.255±0.009	0.112±0.032	0.242±0.015	0.154±0.041
	NaiveAttack	0.697±0.013	0.678±0.035	0.256±0.007	0.839±0.067	**0.272±0.007**	0.777±0.047	**0.338±0.0125**	0.760±0.039
	DoorPing-1	0.686±0.006	0.137±0.004	0.246±0.008	0.129±0.028	0.246±0.008	0.088±0.022	0.240±0.010	0.173±0.034
	DoorPing-2	0.684±0.024	0.931±0.050	0.232±0.007	0.095±0.011	0.253±0.008	0.159±0.031	0.276±0.061	0.363±0.025
	本文	0.692±0.012	**1.000±0.000**	**0.278±0.006**	**1.000±0.000**	0.265±0.008	**1.000±0.000**	0.315±0.010	**1.000±0.000**

注:加粗字体表示各网络的各评估指标最优结果。“–”表示该项无需测试。

NaiveAttack、DoorPing、DoorPing-1 的中毒准确率通常低于 85%。DoorPing-2 的中毒准确率高达 95% 左右。对比这些方法，该方法在 4 种数据集，6 种网络框架下均达到 100% 中毒准确率。这些结果表明，该方法明显优于现有数据浓缩后门攻击方法。该方法相较于 DoorPing-1，DoorPing-2 中毒准确率提高 23.5%，与 DoorPing 相比中毒准确率提高 52.3%。本文提出的多重对比的触发器优化方法，相较于现有方法所采用的基于放大神经元的触发器优化方法，具有显著的优势。这是因为通过放大神经元所训练的触发器与目标类真实数据特征之间的差异较大，在数据浓缩过程中，触发器容易当做噪声处理，导致触发器的嵌入效果较差。该方法相较于 DoorPing-2 的中毒准确率提高 6.5%。触发器与中毒数据相分离进一步提高了后门攻击的成功率。在数据浓缩过程中，触发器作为单独样本嵌入浓缩数据，以避免复杂的背景信息对触发器的干扰。

3.3 CIFAR10 上的跨模型泛化能力

基于以往的研究（Zhao 等，2021），本文研究了所提出方法的跨模型泛化能力。本文针对 CIFAR10 数据集，在 ConvNet、AlexNet、LeNet、ZFNet、CaffeNet 和 VGG11 上学习一组浓缩数据。其中，每组浓缩数据包括 10 个类别，每个类别包括 10 幅浓缩数据。本文针对每组浓缩数据，将其作为训练数据分别在这 6 种模型上进行训练，然后依据 CIFAR10 测试集在这些模型上的干净准确率和中毒准确率来衡量该方法的跨模型泛化能力。如表 2 和表 3 所示，该方法在 VGG11 评估框架下表现较差。与现有数据浓缩方法的讨论相似，由于复杂模型往往具有较强的特征提取能力，在数据浓缩过程中，浓缩数据容易嵌入真实数据中存在的噪声，导致在使用这些浓缩数据处理下游任务时效果较差。

表 2　在 CIFAR10 数据集上的跨模型泛化能力（干净准确率）

Table 2　The cross-model generalization ability of the proposed method on CIFAR10 dataset（clean accuracy）

网络	ConvNet	AlexNet	LeNet	ZFNet	CaffeNet	VGG11
ConvNet	0.322±0.011	0.351±0.013	0.300±0.010	0.358±0.012	0.302±0.022	0.330±0.009
AlexNet	0.353±0.015	0.368±0.019	0.291±0.022	0.346±0.011	0.281±0.030	0.310±0.016
LeNet	0.286±0.010	0.336±0.014	0.308±0.016	0.339±0.003	0.319±0.018	0.312±0.018
ZFNet	0.317±0.009	0.385±0.012	0.296±0.027	0.378±0.019	0.310±0.012	0.311±0.011
CaffeNet	0.310±0.012	0.341±0.011	0.258±0.038	0.336±0.012	0.352±0.014	0.309±0.005
VGG11	0.193±0.009	0.228±0.007	0.193±0.010	0.228±0.009	0.252±0.007	0.278±0.006

表 3　在 CIFAR10 数据集上的跨模型泛化能力（中毒准确率）

Table 3　The cross-model generalization ability of the proposed method on CIFAR10 dataset（poisoned accuracy）

网络	ConvNet	AlexNet	LeNet	ZFNet	CaffeNet	VGG11
ConvNet	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000	0.999±0.003	1.000±0.000
AlexNet	0.913±0.053	1.000±0.000	0.946±0.063	1.000±0.000	0.999±0.003	1.000±0.000
LeNet	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000
ZFNet	0.981±0.034	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000
CaffeNet	1.000±0.000	1.000±0.000	0.998±0.003	1.000±0.000	1.000±0.000	1.000±0.000
VGG11	0.372±0.107	0.663±0.117	0.602±0.087	0.571±0.158	0.971±0.029	1.000±0.000

3.4 CIFAR100 复杂数据集上的有效性

除了前述 4 个基准数据集外，还进一步在复杂数据集 CIFAR100 上研究所提方法的有效性。CIFAR100 包含 60 000 幅尺寸为 32 × 32 像素的彩色图像。数据集包括 100 个不同的类别，每个类别包含 500 幅训练图像和 100 幅测试图像。如表 4 和表 5 所示，该方法在 6 种不同的网络框架中均达到了 100% 的中毒准确率，同时干净准确率未出现明显下降，说明该方法很容易扩展到复杂数据集。

表4　在CIFAR100数据集上的干净准确率

Table 4　The clean accuracy of the proposed method on CIFAR100 dataset

方法	ConvNet	AlexNet	LeNet	ZFNet	CaffeNet	VGG11
干净模型	0.131±0.005	0.162±0.007	0.096±0.009	0.162±0.008	0.161±0.003	0.096±0.003
DoorPing	0.122±0.008	0.165±0.007	0.090±0.008	0.116±0.008	0.124±0.003	0.096±0.003
NaiveAttack	0.131±0.007	0.166±0.006	0.100±0.008	0.152±0.010	0.166±0.004	0.084±0.002
本文	0.126±0.004	0.162±0.008	0.106±0.009	0.148±0.101	0.159±0.004	0.095±0.003

表5　在CIFAR100数据集上的中毒准确率

Table 5　The poisoned accuracy of the proposed method on CIFAR100 dataset

方法	ConvNet	AlexNet	LeNet	ZFNet	CaffeNet	VGG11
干净模型	–	–	–	–	–	–
DoorPing	0.117±0.011	0.122±0.022	0.092±0.002	0.095±0.005	0.101±0.005	0.236±0.117
NaiveAttack	0.810±0.019	0.882±0.006	0.736±0.100	0.780±0.005	0.872±0.015	0.622±0.015
本文	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000	0.991±0.028	1.000±0.000

注：“–”表示该项无需测试。

3.5　隐蔽效果

图5展示了该方法在CIFAR10数据集上合成的中毒浓缩图像和干净浓缩图像。干净浓缩数据和中毒浓缩数据对比干净数据均存在较大差异，用户无法通过比较浓缩数据与干净数据之间的差异，识别其是否为中毒浓缩数据。本文方法生成的中毒浓缩和干净浓缩数据在背景颜色上仅存在微小差异，用户无法区分中毒浓缩数据和干净浓缩数据。DoorPing生成的中毒浓缩数据中存在趋近于全黑的图像且与干净浓缩数据具有明显差异，在整组浓缩数据图像中较为突兀，容易被认为中毒数据或噪声而遗弃。本文方法对比DoorPing后门攻击方法具有较高的隐蔽性。

3.6　触发器不可见性定量实验

本文邀请了40位受试者，每位受试者对100组干净浓缩图像与中毒浓缩图像进行区分。每组实验包含两幅并排显示的浓缩图像，受试者选择左侧或右侧图像为中毒浓缩图像，或者无法确定。如图6所示，在4 000组区分任务中，受试者对中毒样本的区分准确率（真阴率）为38%，错误识别率（假阴率）同样接近38%，另有24%的实验被标记为无法确定。实验结果表明，受试者无法准确区分中毒浓缩图像和干净浓缩图像。

3.7　消融实验

3.7.1　每个组成部分的有效性

本文研究了所提出方法的3个组成部分（多重对比的触发器嵌入、基于梯度匹配的触发器优化方法、触发器分区放大预处理）的有效性。如图7所示，每个组成部分在中毒准确率上分别提高54.6%、20.7%和11%。当对触发器进行分区放大预处理后，所提出方法的中毒准确率达到100%左右。这说明通过增加触发器可获取的梯度数量，为其提供了大量的信息指导学习。

3.7.2　每类浓缩数据的数量

现有的数据浓缩方法（Zhao等，2021）已经证明通过增加每类所合成的浓缩数据数量可提高分类精度。为了研究浓缩数据数量对攻击成功率的影响，每个类分别浓缩1、10、50幅浓缩数据。如图8所示，干净准确率随着浓缩数据的数量增加而增长。中毒准确率也展现出相似的增长趋势。当浓缩数据的数量仅为1幅时，中毒准确率仅达到29.8%。当浓缩数据的数量达到10幅后，中毒准确率达到了100%并保持稳定。

3.7.3　触发器像素值的大小

在测试阶段，触发器像素值从原始大小逐步缩小到原始大小的0.1倍，并在图9中展示了中毒准确率。该方法在触发器像素值缩小到原始大小的0.1倍时依旧达到90%左右的中毒准确率。这种特点允许攻击者在测试阶段，通过降低触发器的像素值提高触发器的不可见性。

3.7.4　浓缩轮数

本文进一步研究了浓缩轮数对攻击成功率的影

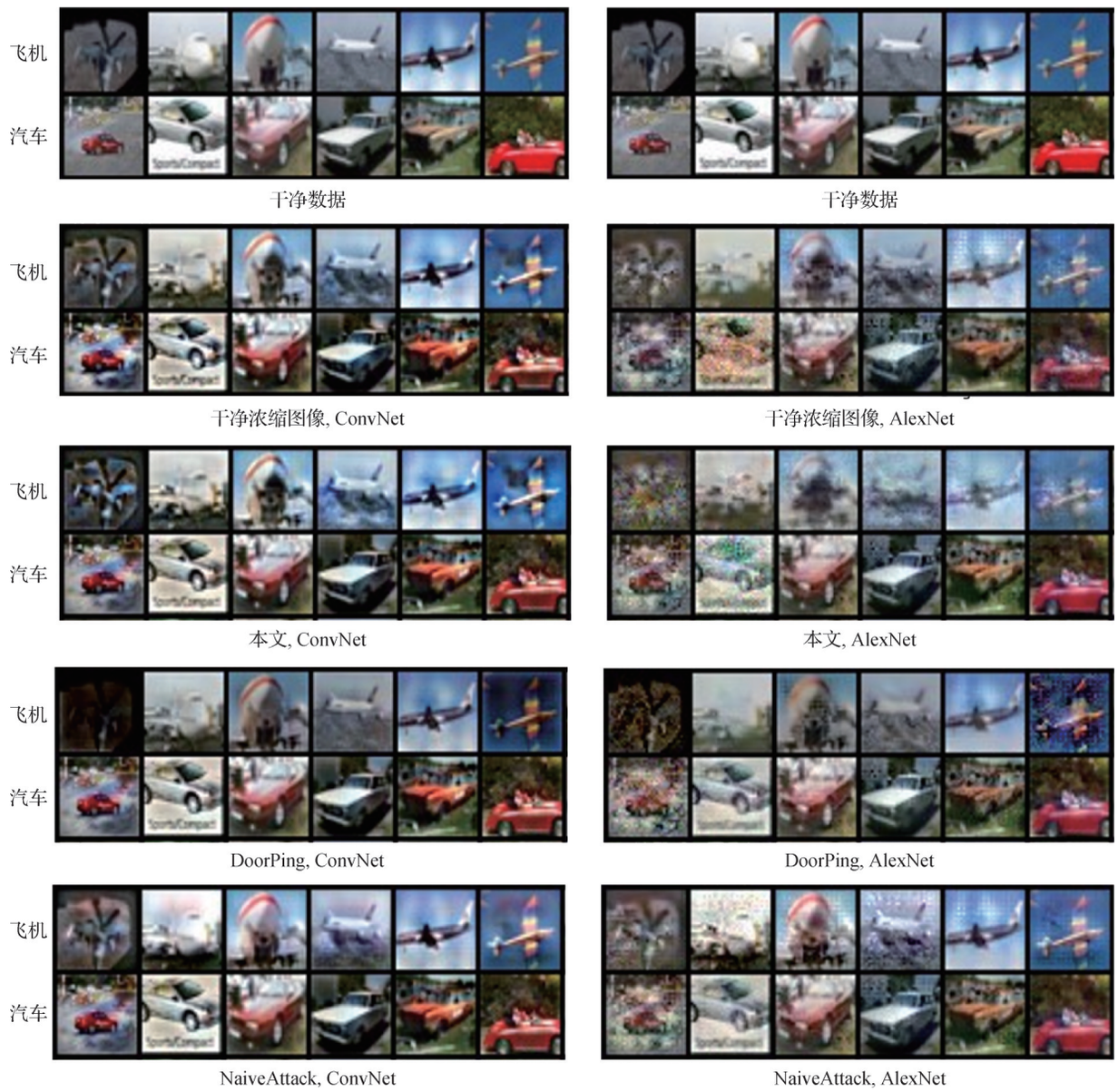

图5　CIFAR10数据集上的中毒浓缩图像和干净浓缩图像的比较

Fig. 5　Comparison between poisoned condensed images and clean condensed images on CIFAR10 dataset

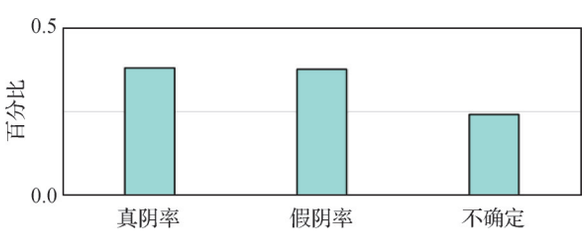

图6　中毒浓缩数据视觉可区分定量评估

Fig. 6　Quantitative evaluation of visual discriminablity for poisoned condensed data

响。本文针对CIFAR10数据集从50轮到300轮依次提取浓缩数据来评估浓缩轮数对攻击成功率的影响。如图10所示，随着浓缩轮数的增加，中毒准确率和干净准确率均不断增长并趋于稳定。值得注意的是，所提出的方法仅经过100轮的训练便达到了100%的中毒准确率。

3.7.5　触发器梯度占比α

在本实验中，验证了超参数α如何影响该方法的中毒准确率和干净准确率。如表6所示，超参数的微小变化对中毒准确率和干净准确率没有影响。在实验中超参数α设置为0.007，对于VGG11模型，则设置为0.5。

3.7.6　真实数据的数量

基于以往的研究(Liu等，2023b)，本文探究了真

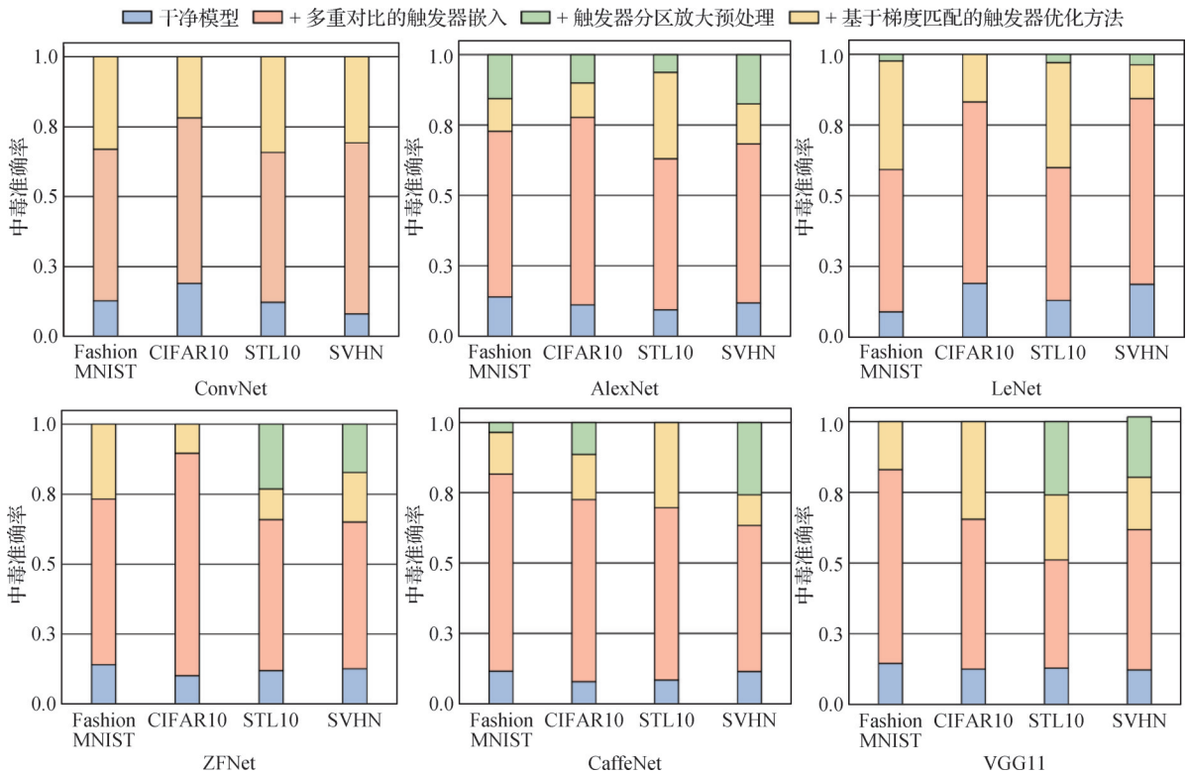

图7　本文方法3个组成部分的有效性

Fig. 7　The effectiveness of the three components of our proposed method

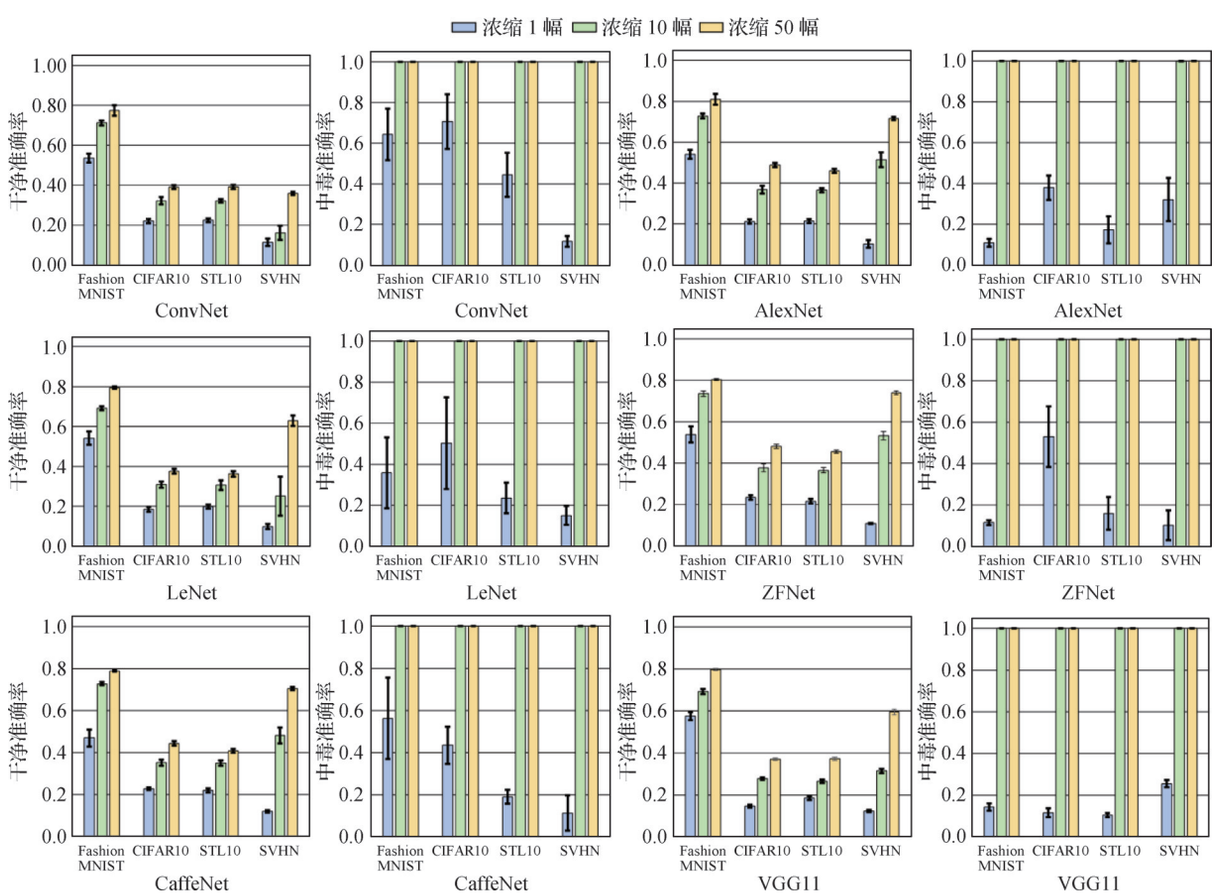

图8　在不同浓缩数据数量下的干净准确率和中毒准确率

Fig. 8　The clean accuracy and poisoned accuracy under different number of condensed data

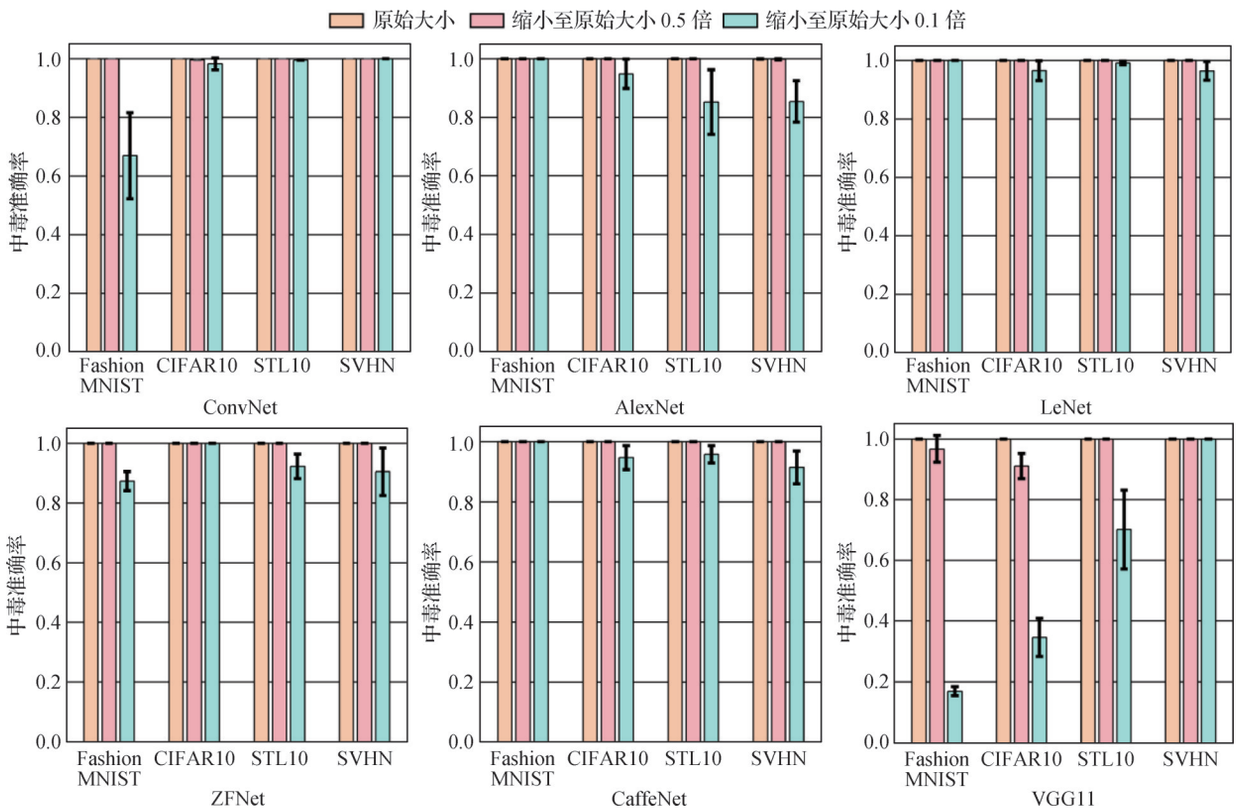

图 9　在不同触发器像素值下的中毒准确率

Fig. 9　The poisoned accuracy under different trigger pixel value

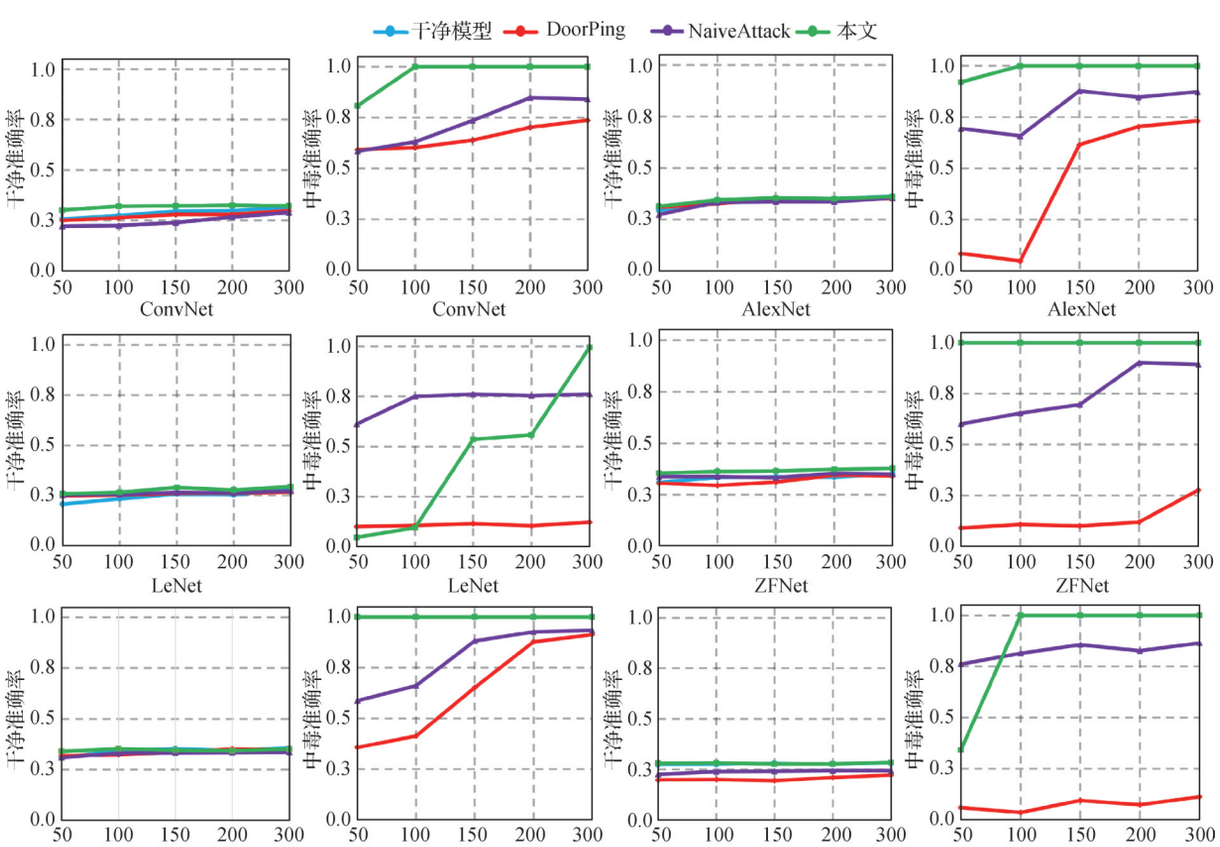

图 10　在不同浓缩轮数下的干净准确率和中毒准确率

Fig. 10　The poisoned accuracy and clean accuracy under different epochs of condensation

表6　在不同超参数α下的干净准确率和中毒准确率

Table 6　The poisoned accuracy and clean accuracy under different hyperparameter α

网络	α	FashionMNIST		CIFAR10		STL10		SVHN	
		干净准确率	中毒准确率	干净准确率	中毒准确率	干净准确率	中毒准确率	干净准确率	中毒准确率
ConvNet	0.005	0.721±0.007	1.000±0.000	0.317±0.013	1.000±0.000	0.323±0.003	1.000±0.000	0.161±0.020	1.000±0.000
	0.006	0.711±0.004	1.000±0.000	0.317±0.008	1.000±0.000	0.322±0.006	1.000±0.000	0.173±0.013	1.000±0.000
	0.007	0.712±0.005	1.000±0.000	0.322±0.011	1.000±0.000	0.321±0.007	1.000±0.000	0.161±0.009	1.000±0.000
AlexNet	0.005	0.701±0.006	1.000±0.000	0.374±0.018	1.000±0.000	0.373±0.012	1.000±0.000	0.511±0.027	1.000±0.000
	0.006	0.723±0.016	1.000±0.000	0.375±0.016	1.000±0.000	0.369±0.011	1.000±0.000	0.527±0.035	0.956±0.142
	0.007	0.729±0.012	1.000±0.000	0.368±0.019	1.000±0.000	0.365±0.009	1.000±0.000	0.515±0.036	0.999±0.005
LeNet	0.005	0.690±0.017	1.000±0.000	0.300±0.017	1.000±0.000	0.302±0.016	1.000±0.000	0.259±0.099	1.000±0.000
	0.006	0.694±0.015	1.000±0.000	0.306±0.023	1.000±0.000	0.295±0.025	1.000±0.000	0.256±0.100	1.000±0.000
	0.007	0.691±0.01	1.000±0.000	0.308±0.016	1.000±0.000	0.306±0.025	1.000±0.000	0.251±0.098	1.000±0.000
ZFNet	0.005	0.740±0.012	1.000±0.000	0.364±0.016	1.000±0.000	0.368±0.018	0.999±0.001	0.533±0.028	0.962±0.169
	0.006	0.741±0.011	1.000±0.000	0.369±0.015	1.000±0.000	0.366±0.015	1.000±0.000	0.528±0.039	1.000±0.000
	0.007	0.735±0.013	1.000±0.000	0.378±0.019	1.000±0.000	0.366±0.013	1.000±0.000	0.512±0.021	1.000±0.000
CaffeNet	0.005	0.728±0.008	1.000±0.000	0.352±0.014	1.000±0.000	0.350±0.012	1.000±0.000	0.481±0.038	1.000±0.000
	0.006	0.728±0.008	1.000±0.000	0.362±0.013	1.000±0.000	0.355±0.011	1.000±0.000	0.460±0.033	1.000±0.000
	0.007	0.723±0.010	1.000±0.000	0.345±0.014	1.000±0.000	0.354±0.10	1.000±0.000	0.445±0.034	1.000±0.000
VGG11	0.5	0.692±0.012	1.000±0.000	0.278±0.006	1.000±0.000	0.265±0.008	1.000±0.000	0.315±0.010	1.000±0.000
	0.8	0.678±0.015	1.000±0.000	0.257±0.007	0.993±0.008	0.268±0.007	1.000±0.000	0.305±0.009	1.000±0.000
	1	0.689±0.015	1.000±0.000	0.258±0.006	1.000±0.000	0.278±0.007	1.000±0.000	0.284±0.012	1.000±0.000

实数据的数量对攻击成功率的影响。本文从真实数据集提取不同比例(0.2～1)的数据子集用于数据浓缩,并在图11中展示了干净准确率和中毒准确率。该方法不受原始真实数据数量的影响,保持100%的中毒准确率。而现有数据浓缩后门攻击方法则随着真实数据数量的增加出现2%左右的中毒准确率提升。这表明本文方法在保证中毒准确率前提下,可以通过减少真实数据的数量以降低浓缩过程中的计算成本。

3.8　防御

现有的防御方法主要分为3个级别(Xu等,2021),即模型级别(Li等,2021b;Wang等,2019;Liu等,2018),用于检测模型是否遭受后门攻击;数据级别(Tang等,2021;Huang等,2023;Chen等,2019),用于检测训练集中是否含有中毒数据;输入级别(Xie等,2024;Cho等,2020;Kwon,2025;Zeng等,2021),在模型部署后,检测输入数据是为中毒数据。本文在CIFAR10数据集上评估了该方法面对这3种级别的后门防御方法时的逃逸效果。其中,针对每种级别的防御,本文选择了3种具有代表性的防御方法。

3.8.1　模型级别

1)NAD(neural attention distillation)(Li等,2021b)。NAD利用知识蒸馏的方法,引导后门学生模型与干净教师模型的中间层表征进行对齐,从而去除学生模型中的隐藏后门。在实验中,本文从干净训练数据集中选择了一个占比为0.05的数据子集用于知识蒸馏,并在表7中列出了学生模型的干净准确率和中毒准确率。NAD处理后的后门模型依旧保持100%的中毒准确率。NAD防御方法无法抵抗提出的后门攻击方法。

2)Neural Cleanse(Wang等,2019)。Neural Cleanse通过反向工程为模型的每个输出类别生成一个触发器,并利用中位数绝对值偏差计算每个触发器的异常值。若存在异常值大于2的触发器说明模型被植入后门。图12展示了Neural Cleanse生成的异常值

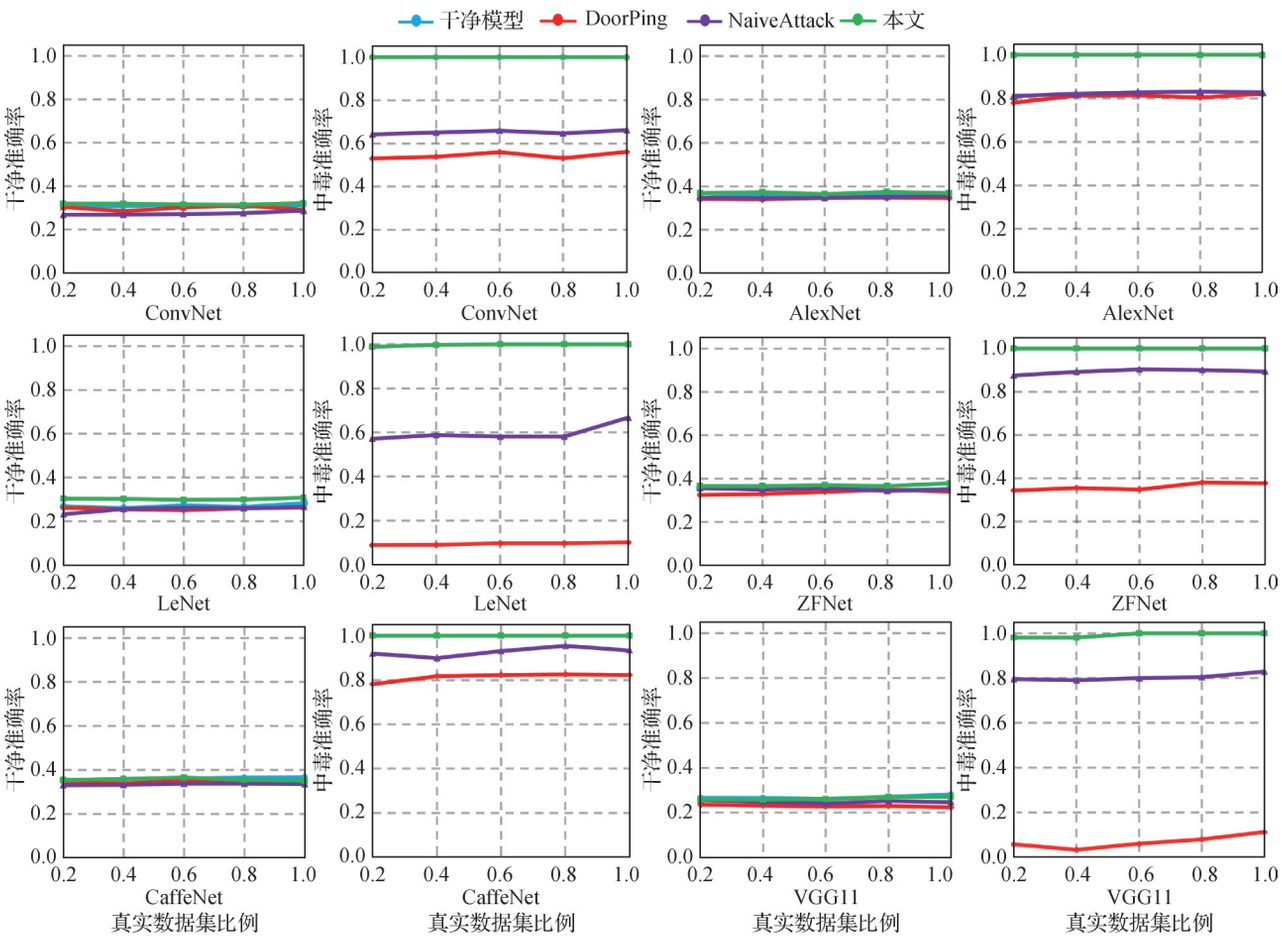

图 11　在不同真实数据集比例下的干净准确率和中毒准确率

Fig. 11　The clean accuracy and poisoned accuracy under different proportions of real datasets

表 7　NAD 处理后的中毒准确率和干净准确率

Table 7　The poisoned accuracy and clean accuracy of the proposed method after NAD processing

网络	FashionMNIST		CIFAR10		STL10		SVHN	
	干净准确率	中毒准确率	干净准确率	中毒准确率	干净准确率	中毒准确率	干净准确率	中毒准确率
ConvNet	0.1	1.0	0.1	1.0	0.1	1.0	0.1	1.0
AlexNet	0.1	1.0	0.1	1.0	0.1	1.0	0.07	1.0

大小。Neural Cleanse 生成的异常值均小于 2，这表明 Neural Cleanse 防御方法无法检测到所提出的后门攻击。

3）FP（fine-pruning）（Liu 等，2018）。FP 结合模型剪枝和模型微调移除模型中的后门。表 8 列出了模型经过精细微调后的干净准确率和中毒准确率。实验结果表明，FP 防御方法无法抵抗提出的后门攻击方法。

3.8.2　数据级别

1）SCAn（statistical analysis of DNNs）（Tang 等，2021）。SCAn 先利用 EM（expectation-maximization）算法构建训练数据的高斯混合模型，然后利用似然比检验计算训练数据和干净数据之间的分布差异。当检验统计量大于 7.389，说明训练数据中混入中毒数据。如图 13 所示，SCAn 生成的检验统计量均小于 7.389。所提出的后门攻击方法逃逸 SCAn 防御方法。

2）CD（cognitive distillation）（Huang 等，2023）。CD 在模型分类过程中发现中毒数据的判别特征往往较小。基于这一发现，CD 对每个训练数据学习一个掩码来掩盖数据中的非关键特征，然后根据掩码大小去除那些掩码较大的训练数据，从而去除训练数据中的中毒数据。图 14 展示了中毒浓缩数据和

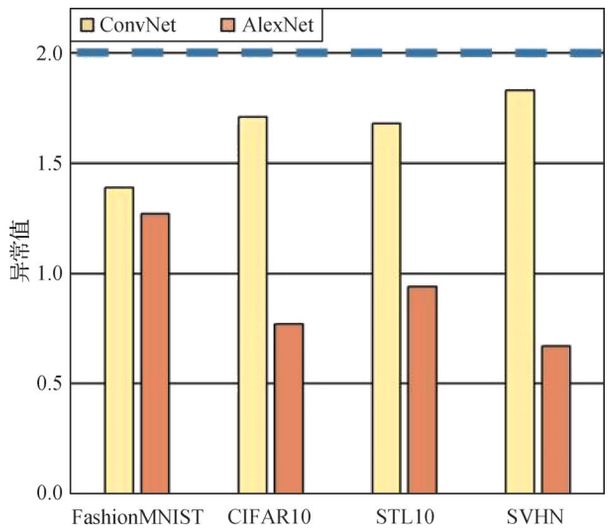

图12　Neural Cleanse生成的异常值

Fig. 12　Anomaly indices produced by Neural Cleanse

干净浓缩数据的掩码L2范数。干净浓缩数据的掩码通常大于中毒浓缩数据的掩码，这表明CD防御方法无法区分中毒浓缩数据与干净浓缩数据。

3）AC（activation clustering）（Liu等，2018）。AC先利用聚类算法将训练数据分为两个簇，然后利用剪影分数计算两个簇的距离。若两个簇距离较远说明训练数据中存在中毒数据。为了检验所提出的方法是否能够逃逸AC防御方法，本文为中毒浓缩数据的每个类计算一个剪影分数，然后取其中的最大值并与干净浓缩数据的最大值和平均值进行比较。如图15所示，中毒浓缩数据的最大剪影分数均小于干净浓缩数据的最大剪影分数，并且接近干净浓缩数据的平均剪影分数。AC防御方法无法检测到所提出的后门攻击。

表8　FP处理后的中毒准确率和干净准确率

Table 8　The poisoned accuracy and clean accuracy of the proposed method after FP processing

网络	FashionMNIST		CIFAR10		STL10		SVHN	
	干净准确率	中毒准确率	干净准确率	中毒准确率	干净准确率	中毒准确率	干净准确率	中毒准确率
ConvNet	0.716	1.000	0.313	1.000	0.331	1.000	0.233	1.000
AlexNet	0.755	1.000	0.364	1.000	0.370	1.000	0.495	1.000

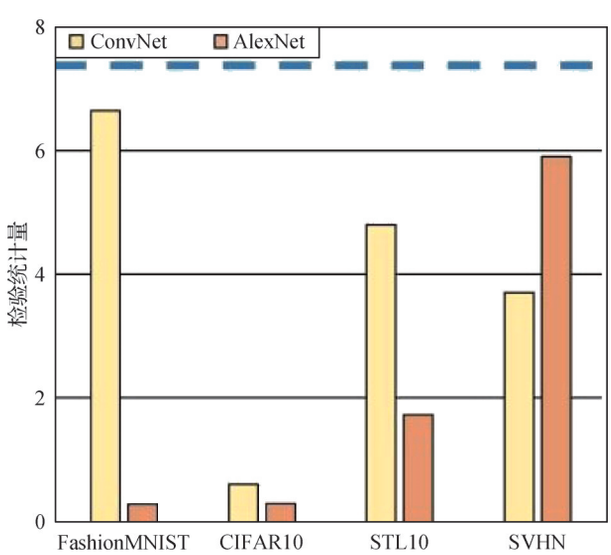

图13　SCAn生成的检验统计量

Fig. 13　The test statistic generated by SCAn

3.8.3　输入级别

1）Badexpert（Xie等，2024）。Badexpert先使用错误标记的干净数据对后门模型做微调，使其在保留后门的同时遗忘目标任务，随后根据输入数据在该后门专家模型上的置信度检测其是否为中毒数据。如表9所示，Badexpert将99%的中毒测试数据识别为干净数据。Badexpert无法有效移除输入数据中存在的中毒数据。

2）去噪自编码器（Cho等，2020）和去触发器自编码器（Kwon，2025）。表10列出了该方法应用去噪自编码器和去触发器自编码器后的干净准确率和中毒准确率。经过这些自编码器处理后的中毒数据依旧激活模型中的隐藏后门。这些防御方法未能去除中毒数据中的触发器。

3）Frequency detection（Zeng等，2021）。Frequency detection利用离散余弦变化（discrete cosine transform，DCT）频域转化方法检测输入数据中是否存在高频伪影。若输入数据存在高频伪影说明该数据为中毒数据。如表11所示，在CIFAR10、STL10和SVHN数据集上，Frequency detection未能检测出测试集中存在的中毒数据。在FashionMNIST数据集中，由于该数据集是一种单通道的黑白图像，当触发器叠加到干净数据时，不可避免地使其产生高频伪影，导致所提出的后门攻击方法无法有效逃逸该防御方法。

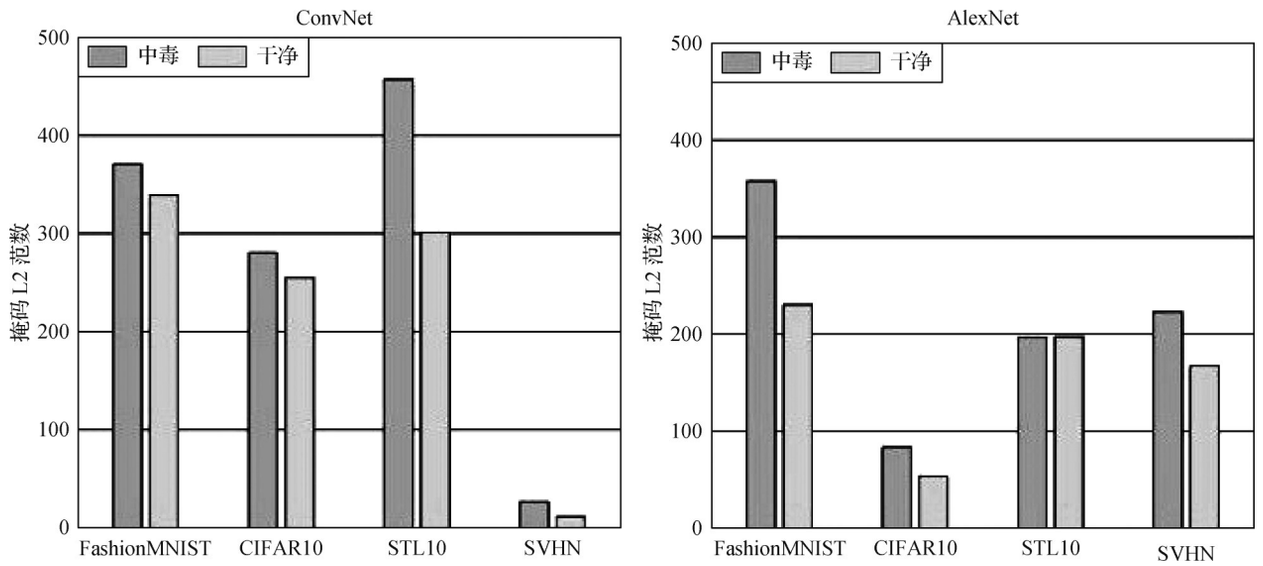

图 14　CD 生成的掩码 L2 范数

Fig. 14　The L2 norm of the mask generated by CD

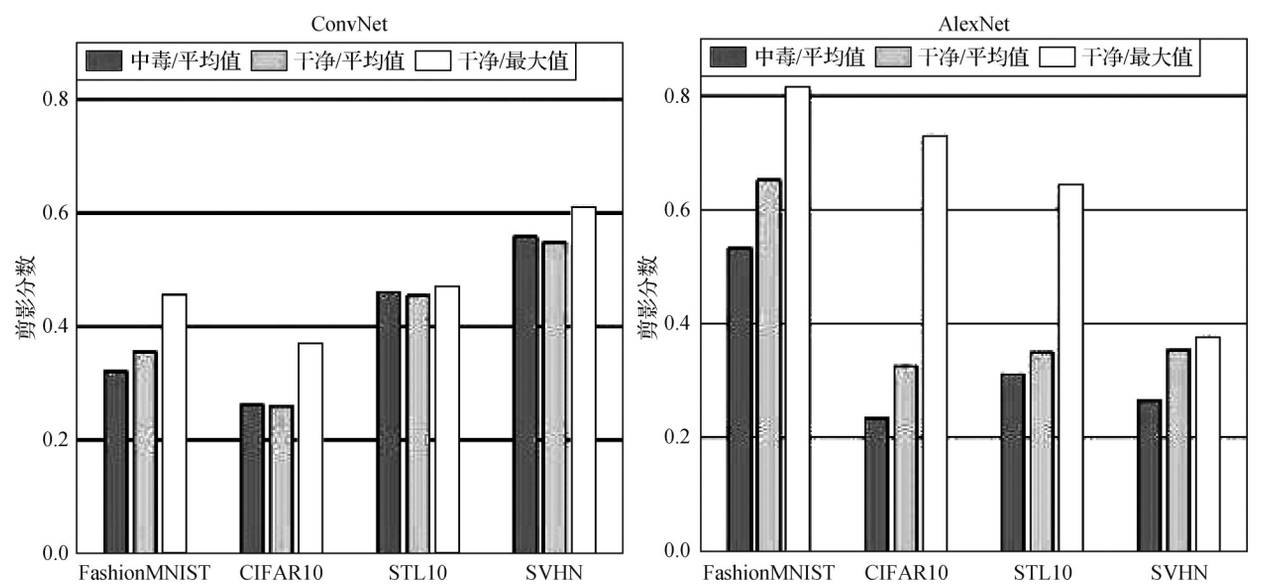

图 15　AC 生成的剪影分数

Fig15　Silhouette scores generated by AC

表 9　Badexpert 在测试数据上的真阳率和假阳率

Table 9　The true positive rate and false positive rate of Badexpert detecting test data

网络	FashionMNIST		CIFAR10		STL10		SVHN	
	真阳率	假阳率	真阳率	假阳率	真阳率	假阳率	真阳率	假阳率
ConvNet	0.99	1.00	0.98	0.99	0.99	1.00	0.99	1.00
AlexNet	0.99	1.00	0.99	1.00	0.99	0.99	0.99	1.00

4　结　论

本文提出一种分离触发器和多重对比的数据浓缩后门攻击方法。触发器与真实数据进行分离。在数据浓缩阶段，触发器单独作为样本与真实数据嵌入浓缩数据，减少真实数据对触发器的干扰。本文对分离的触发器进行了优化，将触发器接近目标类

表10　去噪自编码器和去触发器自编码器处理后的干净准确率和中毒准确率

Table 10　The clean accuracy and poisoned accuracy after processing by de-noising autoencoder and de-triggering autoencoder

类别	网络	FashionMNIST		CIFAR10		STL10		SVHN	
		干净准确率	中毒准确率	干净准确率	中毒准确率	干净准确率	中毒准确率	干净准确率	中毒准确率
去噪自编码器	ConvNet	0.690	1.000	0.317	1.000	0.300	1.000	0.154	1.000
	AlexNet	0.714	1.000	0.348	1.000	0.342	1.000	0.351	0.999
去触发器自编码器	ConvNet	0.690	1.000	0.28	1.000	0.220	1.000	0.162	1.000
	AlexNet	0.723	1.000	0.302	1.000	0.261	1.000	0.404	1.000

表11　Frequency detection测试数据上的真阳率和假阳率

Table 11　The true positive rate and false positive rate of Frequency detecting test data

网络	FashionMNIST		CIFAR10		STL10		SVHN	
	真阳率	假阳率	真阳率	假阳率	真阳率	假阳率	真阳率	假阳率
ConvNet	0.998	0.000	1.000	1.000	0.830	1.000	0.956	0.996
AlexNet	0.998	0.000	1.000	1.000	0.830	1.000	0.956	1.000

真实数据的特征，从而提高触发器的嵌入效果。在触发器的学习过程中，本文对触发器进行了分区放大预处理来增加触发器像素的数量，使其获取大量的梯度用于指导学习。在数据浓缩阶段，通过多重对比将目标类浓缩与触发器特征投影在相同空间，将非目标类浓缩数据与触发器特征进行分离，以去除非目标类浓缩数据中混杂的触发器特征。所提出的方法在5种数据集和6种网络框架下均达到100%攻击成功率。同时，人眼难以区分中毒浓缩数据和干净浓缩数据。研究人员针对基于数据浓缩的联邦学习方法的研究已取得一些研究成果（Hu等，2022；Song等，2023；Xiong等，2023）。由于中毒浓缩数据无法直接应用于联邦学习，基于数据浓缩的联邦学习后门攻击方法成为未来的研究工作。

参考文献（References）

Agarwal P K, Har-Peled S and Varadarajan K R. 2004. Approximating extent measures of points. Journal of the ACM (JACM), 51(4): 606-635 [DOI: 10.1145/1008731.1008736]

Aljundi R, Lin M, Goujaud B and Bengio Y. 2019. Gradient based sample selection for online continual learning//Proceedings of the 33rd Conference on Neural Information Processing Systems. Vancouver, Canada: NeurIPS: 11816-11825

Cazenavette G, Wang T Z, Torralba A, Efros A A and Zhu J Y. 2022. Dataset distillation by matching training trajectories//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 4749-4758 [DOI: 10.1109/CVPRW56347.2022.00521]

Chen B, Carvalho W, Baracaldo N, Ludwig H, Edwards B, Lee T, et al. 2019. Detecting backdoor attacks on deep neural networks by activation clustering//Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu, USA: CEUR-WS.org

Chen X Y, Liu C, Li B, Lu K and Song D. 2017. Targeted Backdoor Attacks on deep learning systems using data poisoning [EB/OL]. [2025-03-10]. https://arxiv.org/pdf/1712.05526.pdf

Chen Y T, Welling M and Smola A J. 2010. Super-samples from kernel herding//Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence. Catalina Island, USA: AUAI Press

Cho S, Jun T J, Oh B and Kim D. 2020. DAPAS: denoising autoencoder to prevent adversarial attack in semantic segmentation//Proceedings of 2020 International Joint Conference on Neural Networks. Glasgow, UK: IEEE: 1-8 [DOI: 10.1109/IJCNN48605.2020.9207291]

Coates A, Ng A and Lee H. 2011. An analysis of single-layer networks in unsupervised feature learning//Proceedings of the 14th International Conference on Artificial Intelligence and Statistics. Fort Lauderdale, USA: JMLR.org: 215-223

Gidaris S and Komodakis N. 2018. Dynamic few-shot visual learning without forgetting//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 4367-4375 [DOI: 10.1109/CVPR.2018.00459]

Girshick R, Donahue J, Darrell T and Malik J. 2014. Rich feature hier-

archies for accurate object detection and semantic segmentation// Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE: 580-587 [DOI: 10.1109/CVPR.2014.81]

Gong X L, Wang Z Y, Chen Y J, Xue M, Wang Q and Shen C. 2023. Kaleidoscope: physical backdoor attacks against deep neural networks with RGB filters. IEEE Transactions on Dependable and Secure Computing, 20 (6): 4993-5004 [DOI: 10.1109/TDSC.2023.3239225]

Gu T Y, Dolan-Gavitt B and Garg S. 2017. BadNets: identifying vulnerabilities in the machine learning model supply chain [EB/OL]. [2025-03-10]. https://arxiv.org/pdf/1708.06733.pdf

Hu S Y, Goetz J, Malik K, Zhan H Y, Liu Z and Liu Y. 2022. FedSynth: gradient compression via synthetic data in federated learning [EB/OL]. [2025-03-10]. https://arxiv.org/pdf/2204.01273.pdf

Huang H X, Ma X J, Erfani S M and Bailey J. 2023. Distilling cognitive backdoor patterns within an image//Proceedings of the 11th International Conference on Learning Representations. Kigali, Rwanda: OpenReview.net

Jiang W B, Li H W, Xu G W and Zhang T W. 2023. Color backdoor: a robust poisoning attack in color space//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 8133-8142 [DOI: 10.1109/CVPR52729.2023.00786]

Krizhevsky A. 2019. Learning multiple layers of features from tiny images [EB/OL]. [2025-03-10]. https://www.cs.utoronto.ca/kriz/learning-features-2009-TR.pdf

Krizhevsky A, Sutskever I and Hinton G E. 2011. ImageNet classification with deep convolutional neural networks//Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe, USA: Curran Associates Inc.: 1097-1105

Kwon H. 2025. Defending deep neural networks against backdoor attack by using de-trigger autoencoder. Journal of IEEE Access, 13: 11159-11169 [DOI: 10.1109/ACCESS.2021.3086529]

LeCun Y, Bottou L, Bengio Y and Haffner P. 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11): 2278-2324 [DOI: 10.1109/5.726791]

Li S F, Xue M H, Zhao B Z H, Zhu H J and Zhang X P. 2021a. Invisible backdoor attacks on deep neural networks via steganography and regularization. IEEE Transactions on Dependable and Secure Computing, 18 (5): 2088-2105 [DOI: 10.1109/TDSC.2020.3021407]

Li Y G, Lyu X X, Koren N, Lyu L J, Li B and Ma X J. 2021b. Neural Attention Distillation: erasing backdoor triggers from deep neural networks//Proceedings of the 9th International Conference on Learning Representations. [s.l.]: OpenReview.net

Liu K, Dolan-Gavitt B and Garg S. 2018. Fine-pruning: defending against backdooring attacks on deep neural networks//Proceedings of the 21st International Symposium on Research in Attacks, Intrusions, and Defenses. Heraklion, Greece: Springer: 273-294 [DOI: 10.1007/978-3-030-00470-5_13]

Liu Q, Qiu Y L, Zhou T Q, Xu M, Qin J H, Ma W T, et al. 2025. Mitigating cross-modal retrieval violations with privacy-preserving backdoor learning. IEEE Transactions on Circuits and Systems for Video Technology, 35 (3): 2526-2540 [DOI: 10.1109/TCSVT.2024.3489886]

Liu Q, Zhou T Q, Cai Z P, Yuan Y, Xu M, Qin J H, et al. 2023a. Turning backdoors for efficient privacy protection against image retrieval violations. Information Processing and Management, 60(5): #103471 [DOI: 10.1016/j.ipm.2023.103471]

Liu Y G, Li Z, Backes M, Shen Y and Zhang Y. 2023b. Backdoor attacks against dataset distillation//Proceedings of the 30th Annual Network and Distributed System Security Symposium. San Diego, USA: The Internet Society, #24287 [DOI: 10.14722/ndss.2023.24287]

Mielikäinen J. 2006. LSB matching revisited. IEEE Signal Processing Letters, 13(5): 285-287 [DOI: 10.1109/LSP.2006.870357]

Netzer Y, Wang T, Coates A, Bissacco A, Wu B and Ng A Y. 2011. Reading digits in natural images with unsupervised feature learning//Proceedings of NIPS Workshop on Deep Learning and Unsupervised Feature Learning. Granada, Canada: [s.n.]

Nguyen T, Chen Z R and Lee J. 2021a. Dataset meta-learning from kernel ridge-regression//Proceedings of the 9th International Conference on Learning Representations. [s.l.]: OpenReview.net

Nguyen T, Novak R, Xiao L C and Lee J. 2021b. Dataset distillation with infinitely wide convolutional networks//Proceedings of the 35th International Conference on Neural Information Processing Systems. [s.l.]: Curran Associates Inc.: #397

Nguyen T A and Tran A T. 2021. WaNet-imperceptible warping-based backdoor attack//Proceedings of the 9th International Conference on Learning Representations. [s.l.]: OpenReview.net

Rebuffi S A, Kolesnikov A, Sperl G and Lampert C H. 2017. iCaRL: incremental classifier and representation learning//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 5533-5542 [DOI: 10.1109/CVPR.2017.587]

Salem A, Wen R, Backes M, Ma S Q and Zhang Y. 2022. Dynamic backdoor attacks against machine learning models//Proceedings of the 7th IEEE European Symposium on Security and Privacy. Genoa, Italy: IEEE: 703-718 [DOI: 10.1109/EuroSP53844.2022.00049]

Sener O and Savarese S. 2018. Active learning for convolutional neural networks: a core-set approach//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada: OpenReview.net

Simonyan K and Zisserman A. 2015. Very deep convolutional networks for large-scale image recognition//Proceedings of the 3rd Interna-

tional Conference on Learning Representations. San Diego, USA: ICLR

Song R, Liu D, Chen D Z, Festag A, Trinitis C, Schulz M, et al. 2023. Federated learning via decentralized dataset distillation in resource-constrained edge environments//Proceedings of 2023 International Joint Conference on Neural Networks. Gold Coast, Australia: IEEE: 1-10 [DOI: 10.1109/IJCNN54540.2023.10191879]

Tang D, Wang X F, Tang H X and Zhang K H. 2021. Demon in the variant: statistical analysis of DNNs for robust backdoor contamination detection//Proceedings of 30th USENIX Security Symposium. Berkeley, USA: USENIX Association: 1541-1558

Toneva M, Sordoni A, Combes R T D, Trischler A, Bengio Y and Gordon G J. 2019. An empirical study of example forgetting during deep neural network learning//Proceedings of the 7th International Conference on Learning Representations. New Orleans, USA: OpenReview.net

Voulodimos A, Doulamis N, Doulamis A and Protopapadakis E. 2018. Deep learning for computer vision: a brief review. Computational Intelligence and Neuroscience, 2018: #7068349 [DOI: 10.1155/2018/7068349]

Wang B L, Yao Y S, Shan S, Li H Y, Viswanath B, Zheng H T, et al. 2019. Neural Cleanse: identifying and mitigating backdoor attacks in neural networks//Proceedings of 2019 IEEE Symposium on Security and Privacy. San Francisco, USA: IEEE: 707-723 [DOI: 10.1109/SP.2019.00031]

Wang K, Zhao B, Peng X Y, Zhu Z, Yang S, Wang S, et al. 2022. CAFE: learning to condense dataset by aligning features//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 12186-12195 [DOI: 10.1109/CVPR52688.2022.01188]

Wang T Z, Zhu J Y, Torralba A and Efros A A. 2020. Dataset distillation [EB/OL]. [2025-03-10]. https://arxiv.org/pdf/1811.10959.pdf

Xiao H, Rasul K and Vollgraf R. 2017. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms [EB/OL]. [2025-03-10]. https://arxiv.org/pdf/1708.07747.pdf

Xie T H, Qi X Y, He P, Li Y M, Wang J T and Mittal P. 2024. BaDExpert: extracting backdoor functionality for accurate backdoor input detection//Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria: OpenReview.net

Xiong Y H, Wang R C, Cheng M H, Yu F and Hsieh C J. 2023. FedDM: iterative distribution matching for communication-efficient federated learning//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 16323-16332 [DOI: 10.1109/CVPR52729.2023.01566]

Xu X J, Wang Q, Li H C, Borisov N, Gunter C A and Li B. 2021. Detecting AI Trojans using meta neural analysis//Proceedings of 2021 IEEE Symposium on Security and Privacy. San Francisco, USA: IEEE: 103-120 [DOI: 10.1109/SP40001.2021.00034]

Yao Y S, Li H Y, Zheng H T and Zhao B Y. 2019. Latent backdoor attacks on deep neural networks//Proceedings of 2019 ACM SIGSAC Conference on Computer and Communications Security. London, UK: ACM: 2041-2055 [DOI: 10.1145/3319535.3354209]

Young T, Hazarika D, Poria S and Cambria E. 2018. Recent trends in deep learning based natural language processing [Review Article]. IEEE Computational Intelligence Magazine, 13(3): 55-75 [DOI: 10.1109/MCI.2018.2840738]

Zeiler M D and Fergus R. 2014. Visualizing and understanding convolutional networks//Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer: 818-833 [DOI: 10.1007/978-3-319-10590-1_53]

Zeng Y, Park W, Mao Z M and Jia R X. 2021. Rethinking the backdoor attacks' triggers: a frequency perspective//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 16453-16461 [DOI: 10.1109/ICCV48922.2021.01616]

Zhao B and Bilen H. 2023. Dataset condensation with distribution matching//Proceedings of 2023 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 6503-6512 [DOI: 10.1109/WACV56688.2023.00645]

Zhao B, Mopuri K R and Bilen H. 2021. Dataset condensation with gradient matching//Proceedings of the 9th International Conference on Learning Representations. [s.l.]: OpenReview.net

Zhu S W, Luo G, Wei P, Li S, Zhang X P and Qian Z X. 2023. Image-imperceptible backdoor attacks. Journal of Image and Graphics, 28(3): 864-877 (朱淑雯, 罗戈, 韦平, 李晟, 张新鹏, 钱振兴. 2023. 隐蔽图像后门攻击. 中国图象图形学报, 28(3): 864-877) [DOI: 10.11834/jig.220550]

作者简介

蒋桂政,男,硕士研究生,主要研究方向为后门攻击、数据浓缩、联邦学习。E-mail:gzjiang@mail.dhu.edu.cn

黄荣,通信作者,男,副教授,主要研究方向为AI安全、深度学习与机器视觉、多媒体信息安全。

E-mail:rong.huang@dhu.edu.cn

刘浩,男,副教授,主要研究方向为多媒体信号处理与编码、深度学习。E-mail:liuhao@duh.edu.cn

蒋学芹,男,教授,主要研究方向为信息安全与编码、深度学习与图信号处理。E-mail:xqjiang@dhu.edu.cn