中图分类号：TP391　　单位代号：10280
密　　级：公开　　学　　号：

上海大学

硕士学位论文

SHANGHAI UNIVERSITY
MASTER'S DISSERTATION

题目	基于扰动嵌入的对抗样本和不可学样本生成方法研究

作　　者：任行东
学科专业：通信与信息系统
导　　师：孙广玲
完成日期：2025 年 5 月

姓　　名：任行东　　学号：22721277

论文题目：基于扰动嵌入的对抗样本和不可学样本生成方法研究

上海大学

本论文经答辩委员会全体委员审查，确认符合上海大学硕士学位论文质量要求。

答辩委员会签名：

主　　席：

委　　员：

导　　师：

答辩日期：　　年　　月　　日

姓　　名： 任行东　　学号：22721277

论文题目：基于扰动嵌入的对抗样本和不可学样本生成方法研究

上海大学学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师指导下，独立进行研究工作所取得的成果。除了文中特别加以标注和致谢的内容外，论文中不包含其他人已发表或撰写过的研究成果。参与同一工作的其他研究者对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名：

日 期：　　年　　月　　日

上海大学学位论文使用授权说明

本人完全了解上海大学有关保留、使用学位论文的规定，即：学校有权保留论文及送交论文复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容。

（**保密论文在解密后应遵守此规定**）

学位论文作者签名：　　导师签名：

日 期：　　年　　月　　日　　日 期：　　年　　月　　日

上海大学工学硕士学位论文

基于扰动嵌入的对抗样本和不可学样本生成方法研究

作　　者：任行东

学科专业：通信与信息系统

导　　师：22721277

上海大学通信与信息工程学院

2025 年 5 月

A Dissertation Submitted to Shanghai University for the Degree of Master in Engineering

Research on Adversarial and Unlearnable Examples Generation Methods Based on Perturbation Embedding

Candidate： **Ren Xingdong**

Major： **Communication and Information System**

Supervisor： **Sun Guangling**

School of Communication and Information Engineering

Shanghai University

May, 2025

摘 要

深度学习技术的快速普及在加速智能化进程的同时，亦使数据隐私保护面临严峻挑战。深度神经网络的全生命周期中存在两类核心隐私泄露风险：在模型应用阶段，恶意攻击者可借助已部署的深度模型对用户数据进行高精度分析，进而导致敏感信息泄露；在模型训练阶段，第三方可能非法收集并利用用户数据训练模型，导致数据隐私泄露和所有权失控。针对上述风险，本文系统研究对抗样本（Adversarial Examples）与不可学样本（Unlearnable Examples）两类防护技术，分别构建模型推理与训练环节的隐私保护屏障。

尽管对抗样本与不可学样本技术为解决上述隐私保护问题提供了重要思路，但现有方法仍存在显著局限性：对抗样本技术通过注入细微扰动干扰模型推理以抵御恶意数据分析，但其基于像素空间直接叠加扰动的生成范式易导致视觉伪影，且迁移性不足的缺陷也限制了它的应用；不可学样本技术通过嵌入训练阶段扰动破坏模型学习过程以阻止非法数据利用，但现有方法难以平衡不可学性与隐蔽性，制约其实际应用价值。为此，本文提出两项创新工作：

1）提出了一种基于扰动嵌入的对抗样本生成方法（Perturbation Embedding based Adversarial Example, PtEm-AE），通过深度网络隐写技术将通用扰动嵌入图像中，在保持视觉隐蔽性的同时增强攻击性，以提升隐私保护能力。与传统叠加扰动的方法不同，PtEm-AE 利用编码网络完成扰动的嵌入。其生成流程分为三阶段：首先通过编码网络学习扰动嵌入模式，其次迭代更新通用扰动增强攻击效力，最后将优化后的扰动嵌入新样本中生成对抗样本。此外，本文分别从样本层面与数据分布层面提出两项互补的对抗攻击损失。样本层面借鉴对比学习思想，通过计算原始图像与对抗样本的相似度矩阵以增大其在特征空间的距离；数据分布层面则基于 KL 散度最大化两者的特征分布差异以提升迁移性；为兼顾攻击性与隐蔽性，引入视觉隐蔽性损失以保障对抗样本的视觉质量。与传统隐写技术不同，PtEm-AE 以攻击成功率作为隐写的有效性验证指标，当对抗样本成功误导模型时即证明扰动已被有效嵌入。

2）提出了一种基于扰动嵌入的不可学样本生成方法（Perturbation Embedding based Unlearnable Example, PtEm-UE），旨在有效防止隐私数据被深度模型学习与利用。该方法采用动态优化代理模型策略：第一阶段交替优化扰动嵌入网络与代理模型，并引入对比学习机制以引导模型聚焦于扰动特征；第二阶段冻结嵌入网络，交替优化样本扰动与代理模型以增强不可学性。PtEm-UE 的损失分为三项：隐蔽性损失确保视觉一致性；对比学习损失通过误导特征对齐过程，削弱模型对语义特征的建模能力以干扰自监督学习；聚类紧疏损失通过约束类内紧密度与类间分离度干扰监督学习。这样的损失设计确保生成的不可学样本能够同时有效干扰监督学习与自监督学习过程，从而增强对隐私数据的保护能力。此外，本文将传统隐写的解码验证机制转换为“训练失效即隐写成功”的判据范式，摆脱了显式解码约束，使优化目标聚焦于隐蔽性与不可学性的协同增强。

综上，本文针对深度神经网络在训练与应用阶段面临的数据隐私泄露风险，构建了覆盖模型全生命周期的防护体系。提出基于扰动嵌入的对抗样本和不可学样本生成方法，提升了对抗样本和不可学样本的隐蔽性和有效性。实验结果表明，这些方法在多种不同的场景下均展现出优异性能，为数据隐私保护提供了新的技术路径。

关键词：对抗样本；不可学样本；扰动嵌入；对比学习；数据隐私保护

ABSTRACT

With the rapid proliferation of deep learning technologies accelerating the development of intelligent systems, data privacy protection is facing increasingly severe challenges. Throughout the lifecycle of deep neural networks, there exist two critical types of privacy leakage risks. In the inference stage, malicious attackers can leverage deployed models to conduct high-precision analysis on user data, resulting in the disclosure of sensitive information. In the training stage, third parties may illegally collect and exploit user data to train models, leading to privacy breaches and data ownership loss. To address these risks, this dissertation systematically investigates two types of defense techniques—adversarial examples and unlearnable examples—to construct privacy protection mechanisms in both inference and training phases.

Although adversarial and unlearnable examples offer promising directions for privacy protection, existing approaches still face significant limitations. Adversarial examples inject subtle perturbations to disrupt model inference and resist malicious data analysis; however, conventional pixel-space additive perturbations often introduce perceptible artifacts and suffer from poor transferability, limiting their practical deployment. Unlearnable examples embed perturbations during training to prevent unauthorized data utilization, but current methods struggle to balance unlearnability and imperceptibility, hindering their applicability. To overcome these limitations, two novel approaches are proposed:

1. A perturbation embedding based adversarial example generation method (PtEm-AE) is introduced, which leverages steganographic techniques to embed universal perturbations into images, enhancing attack effectiveness while maintaining visual imperceptibility for improved privacy protection. Unlike traditional additive perturbation methods, PtEm-AE employs an encoder network to embed perturbations. The generation process involves three stages: learning perturbation embedding patterns via the encoder, iteratively optimizing universal perturbations to improve attack strength, and embedding the optimized perturbations into new samples to generate adversarial examples. Furthermore, two complementary loss functions are proposed from both the instance and distribution levels: at the instance level, a contrastive loss is used to enlarge the feature distance between original and adversarial samples; at the distribution level,

a KL divergence loss maximizes the discrepancy between feature distributions of the two sets to enhance transferability. A visual imperceptibility loss is also introduced to balance attack success with visual quality. Unlike traditional steganography, PtEm-AE validates embedding effectiveness through model attack success rates, where successful model deception confirms effective perturbation embedding.

2. A perturbation embedding based unlearnable example generation method (PtEm-UE) is proposed to prevent private data from being effectively learned by deep models. This method adopts a dynamically optimized surrogate feature extractor. In the first stage, the perturbation embedding network and the surrogate model are jointly optimized with a contrastive loss to force attention on perturbation features. In the second stage, the embedding network is frozen, and the sample-specific perturbations and surrogate model are iteratively optimized to further enhance unlearnability. The loss function consists of three components: a visual imperceptibility loss to maintain appearance consistency, a contrastive loss to mislead feature alignment and interfere with semantic modeling in self-supervised learning, and a clustering separation loss that constrains intra-class compactness and inter-class separability to disrupt supervised learning. Such a loss design ensures that the generated unlearnable examples can disrupt both supervised and self-supervised learning, thereby strengthening privacy protection. In addition, the traditional decoding-based verification mechanism in steganography is reformulated as a new paradigm: "training failure implies steganographic success," removing the reliance on explicit decoding and focusing optimization on both imperceptibility and unlearnability.

In summary, this dissertation addresses the data privacy risks faced by deep neural networks during both training and inference, and constructs a comprehensive protection framework covering the model's full lifecycle. The proposed PtEm-AE and PtEm-UE methods significantly improve the imperceptibility and effectiveness of adversarial and unlearnable examples. Experimental results demonstrate strong performance across various scenarios, offering a novel technical pathway for data privacy protection.

Keywords: Adversarial Examples; Unlearnable Examples; Perturbation Embedding; Contrastive Learning; Data Privacy Protection

目　录

第一章　绪论

1.1 课题研究目的和意义

深度学习作为机器学习领域的重要分支，通过构建包含多层非线性变换的深度神经网络（Deep Neural Networks, DNNs）来实现对数据的高层次抽象与表征学习。自 2006 年 Hinton 等人提出深度置信网络并成功突破深层网络梯度消失瓶颈以来，深度神经网络经历了三个关键发展阶段[1]：2006 至 2012 年以全连接网络为主的初步探索期；2012 年至 2017 年以 AlexNet[2]在 ImageNet[3]大规模视觉识别竞赛中超越人类视觉水平为标志的快速发展期；以及 2017 年至今以 Transformer[4]架构为核心的多模态融合期。在图像处理领域，深度神经网络已从基础感知向高级认知迈进：例如基于深度网络的 DeepLesion[5]医疗影像系统灵敏度达到 97.9%；Waymo[6]自动驾驶检测 mAP 值突破 85%。自然语言处理被 Transformer 架构革新，推动了如 GPT[7, 8]、DeepSeek[9, 10]等模型的快速崛起和广泛应用。产业应用中金融客服意图识别准确率达 98.6%，法律文书 BLEU 分数超 82[11]。此外，深度强化学习在复杂决策中表现卓越，例如 AlphaGo Zero[12]仅通过 80 小时的自弈便超越了人类，而宇树科技的 G1[13]机器人在复杂地形环境中的稳定行走率达到 95%，这些数据充分表明了深度神经网络的理论成熟性、技术通用性和实际应用可行性。

在深度学习的技术发展进程中，监督学习与无监督学习代表了两种不同的数据驱动范式[14]。监督学习通过人工标注监督信号，推动 ImageNet 分类 Top-5 准确率达 90.5%，但其依赖大规模标注的特性引发显著瓶颈：ImageNet 数据集 1400 万标注消耗 22000 人一年的工作量，自动驾驶多模态标注成本甚至达 30 美元/帧，是常规任务 15 倍，对高质量标注的需求导致 89%医学算法因标注不足无法临床落地，工业检测模型因标注损失 35%产能。与之相对的无监督学习的技术突破呈现清晰的演进路径：初期基于聚类分析与主成分分解的传统无监督学习方法受限于线性可分假设，在 CIFAR-10 数据集上仅取得 59.7%的分类准确率；随后发展的自监督学习方法通过代理任务（如旋转预测、拼图复原）构造伪监督信号，

将 ResNet-50 表征能力提升至监督学习基准的 78.4%，但其受限于预设任务与目标域的语义鸿沟；最终，对比学习通过样本关系建模实现范式跃迁——通过优化特征空间中的对比损失函数，在 ImageNet 线性评估协议下达到 70.1%的 Top-1 准确率（较无监督基线提升 38.4 个百分点），同时将标注需求压缩至监督学习的 1%以下。对比学习在数据效率与泛化能力上实现双重突破，基于对比学习的预训练表征迁移至医疗影像、自动驾驶等领域时错误率降低 41%，标注成本压缩至 1/100[15]。这种类人的高效表征学习机制，正推动 AI 突破标注依赖的认知瓶颈。

然而，深度学习技术的迅猛发展也带来了一系列不可忽视的数据隐私和伦理挑战：**在模型训练阶段**，海量数据的需求催生了诸如隐蔽爬取、授权滥用等非法数据收集行为，导致诸如用户的生物特征、医疗记录等敏感信息在未经过用户知情同意的情况下被纳入训练数据集，使知情同意原则遭到系统性的破坏。一旦这些使用用户隐私数据训练的模型应用于现实场景，便可能带来严重的风险。根据深度学习的基本原理，模型训练过程中学习到的知识被编码在网络权重参数中。已有研究表明，在特定条件下可通过逆向工程方法从这些权重参数中部分重构出原始训练数据。例如，早在 2015 年，Fredrikson 等[16]研究者便通过模型的逆向工程，仅利用人脸识别 API 的置信度输出，成功重构了输入人脸的原始图像（PSNR>28dB）。类似的，谷歌和 DeepMind 等[17]团队的安全研究人员发现，现行的图像生成模型，如 Stable Diffusion[18]、Imagen[19]等，也存在类似的隐私泄露问题。他们通过巧妙设计的两阶段数据提取攻击，成功提取了超过 100 张与训练数据几乎完全一致的图像，其中包括个人可识别照片、商标、Logo 等。**在模型应用阶段**，随着高性能模型的逐步普及，其双刃剑效应愈加明显。例如，人脸识别、语音合成等技术可能被滥用于身份追踪、声纹伪造等恶意应用场景。医疗诊断、邮件分析等模型可能会反推患者的病史、心理状态等敏感信息。例如，人脸识别系统（如 DeepFace[20]）通过提取 128 维深度特征，在 LFW 数据集[21]上可实现高达 99.63%的验证准确率。恶意攻击者可以利用这一技术建立跨平台的身份关联，斯坦福大学的研究[22]表明，结合社交媒体的照片与公共监控数据，个人身份的重识别成功率可高达 91%。2025 年曝光的“精准通客户管理平台”通过分析用户邮

件和社交动态中的图文特征，标注出如“抑郁倾向”、“消费冲动”等标签，用于高利贷、赌博平台的定向推销。图 1.1 展示了上文所讨论的两类数据隐私风险。

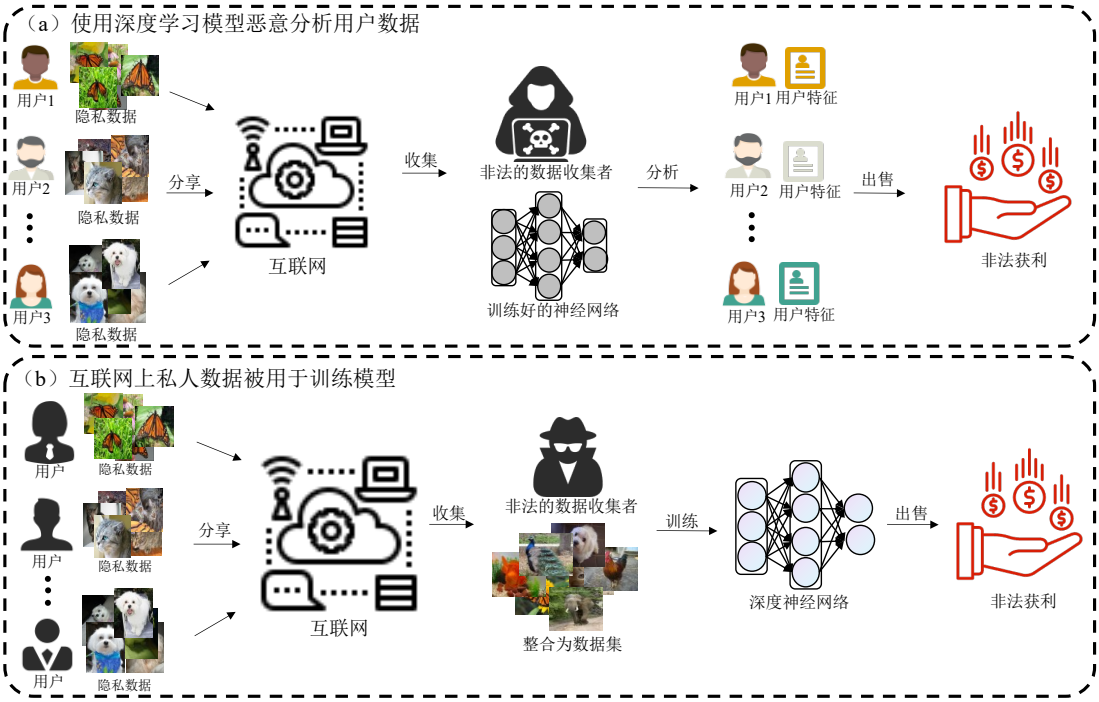


图 1.1 用户数据可能遭受的两类风险，（a）模型应用阶段的数据安全风险，（b）模型训练阶段的数据安全风险

Figure 1.1 Two Types of Risks to User Data, (a) data security risks in model deployment, (b) data security risks in model training

针对**非法利用深度学习模型进行隐私数据分析**的严峻挑战（图 1.1（a）），研究者提出了以对抗样本[23]技术为核心的防御手段。该技术通过在原始数据中注入人眼不可察觉的细微扰动，破坏模型对隐私特征的有效提取：Sharif 等人[24]通过在眼镜框区域嵌入特定噪声，使人脸识别模型的准确率从 99.6%骤降至 3.2%；类似地，Jin 等人[25]开发的 TextFooler 算法通过同义词替换生成对抗文本，使 BERT[26]模型的情感分析准确率下降 47%，同时保持 92%的人类语义连贯性评分。该技术的核心机制在于高维流形扰动优化：通过求解约束空间中的梯度优化问题，使扰动后的数据在模型的特征空间内产生定向偏移。例如，Szegedy 等人[27]通过实验证明，向 ImageNet 图像添加噪声后，ResNet[28]的错误率从 7.8%急剧上升至 89.3%。根据对目标模型的认知程度，对抗攻击可分为白盒攻击和黑盒攻击两大范式。白盒攻击要求掌握模型的内部参数和架构，例如，基于梯度符号法（FGSM[29]）快速生成扰动；而黑盒攻击仅通过模型的输入输出交互进行逆向破解，例

如，研究者采用自然进化策略（NES[30]）在 2000 次查询内达成 78%的攻击成功率。两者的核心区别在于攻击成本和隐蔽性——白盒攻击的精度较高，但依赖于内部信息的泄露；黑盒攻击的计算代价较高，但更贴近实际应用。值得注意的是，传统对抗样本需要对每个输入单独进行优化，而通用对抗扰动（Universal Adversarial Perturbations, UAP[31]）则通过生成一个单一扰动模板，实现跨样本的攻击。其原理类似于“万能钥匙”——通过分析模型决策边界的共性（如 Dezfooli 使用奇异值分解提取 ImageNet 数据集中跨越性扰动模式），生成可批量应用的噪声模板，并且在 VGG[32]，ResNet[28]等不同模型上达到了平均 82%攻击成功率。最新的研究[33]进一步突破了模态限制，基于元学习框架生成的扰动可以同时攻击视觉和文本模型，将跨任务的攻击成功率提高至 65%，显著降低了对抗攻击形式的隐私保护方法的实施复杂性。

针对**非法采集隐私数据并用于训练模型**的潜在风险（图 1.1（b）），数据可用性攻击作为一种主动防御机制，受到了学术界的广泛关注。其核心目标是通过数据层面的干预，使受保护数据无法有效地参与到模型的训练过程中，或显著降低训练出的模型的泛化性能，从而实现对隐私数据的主动隔离和权益保全。目前，主要的技术路径包括数据投毒[34]（Data Poisoning）和不可学样本[35]（Unlearnable Examples）两种机制。数据投毒攻击的目标是通过向训练数据中注入恶意样本（如标签翻转、特征污染等），系统性地误导模型的学习，进而导致模型出现偏差、准确性下降或漏洞等问题——例如，在医疗影像中植入 5%的中毒样本，可导致诊断模型的 AUC 值下降 0.35[36]；而在文本分类任务中，特定词频的扰动可使 BERT 模型的 F1 值降低 47%[37]。不可学样本防御则采用被动阻断策略，其核心机制是通过构造误差最小化噪声，约束训练过程中的损失最小化，从而迫使注入噪声的梯度方向与原始数据的梯度正交，进而产生梯度抵消效应，使数据在特征空间内失去可学习性。Huang 等人[35]的实验证明，在 CIFAR-10 数据集[38]上添加此类噪声后，ResNet-18 模型的测试准确率从 94.3%骤降至 14.7%，且通过对抗训练仅能恢复 65%的性能。这两类技术的主要区别在于——数据中毒攻击依赖攻击者持续参与模型的训练过程，而不可学样本通过人眼不可察觉的方式，保

障了正常的数据分享需求，并有效防止非法数据收集者使用这些数据，从而达到数据保护的目的。

随着技术的发展，数据安全的攻防体系呈现出动态协同演化的特征。在模型应用阶段，现有的对抗样本防御方法多基于线性叠加的扰动生成范式（如 FGSM [29]、PGD[39]等梯度优化方法），尽管能够生成对抗样本，但在视觉隐蔽性和迁移性方面依然面临显著瓶颈。对于训练阶段的数据保护，虽然差分隐私、数据投毒和不可学样本等技术已在一定程度上实现了隐私保护，但这些方法的迁移性和隐蔽性仍然存在较大的提升空间。因此，如何更有效地在数据使用的两个阶段进行保护，依然是值得深入探讨的问题。

1.2 国内外研究现状

针对深度神经网络应用全生命周期中的数据隐私防护需求，对抗样本（Adversarial Examples）与不可学样本（Unlearnable Examples）分别从不同维度构建了防御机制，二者的技术路径具有显著差异与内在关联。从风险场景来看，不可学样本聚焦于模型训练阶段的数据滥用风险，其核心在于通过扰动破坏训练数据的可学习性，使得非法收集的隐私数据无法被有效用于模型训练，从而在数据源头上实现隐私防护；而对抗样本则面向模型部署阶段的推理滥用风险，通过构造对模型决策边界具有强干扰性的扰动，使得已训练完成的模型无法从隐私数据中提取有效信息，从而在数据应用层面形成保护屏障。在技术机理层面，对抗样本通常采用单循环误差最大化优化策略，通过最大化模型预测误差来求解对抗扰动；不可学样本则采用双循环交替优化框架，通过同步优化扰动参数与模型参数，使扰动既能保持视觉隐蔽性，又能最小化数据的可学习价值。二者的核心联系在于均通过向原始数据注入扰动实现防御目标，但在扰动生成范式和优化目标上存在本质差异：对抗样本的扰动旨在欺骗已收敛的模型决策，而不可学样本的扰动则致力于破坏模型的训练动态。本文系统性研究这两类技术，旨在构建覆盖模型训练与应用双阶段的全链条隐私防护体系，为解决深度神经网络生命周期中的数据安全问题提供理论支撑与方法论指导。

（1）对抗样本相关工作：

近年来，样本针对性对抗扰动方法层出不穷。早期的工作中，Goodfellow 等人提出了快速梯度符号法，通过单次梯度计算便可生成扰动；随后，Carlini 与 Wagner[40]设计了更为复杂的优化方法，在保证攻击成功率的同时尽量减少扰动幅度；而 DeepFool[41]则采用了计算决策边界距离的方法来求解最小扰动。此外，还有迭代型 I-FGSM[42]、Momentum Iterative Method[43]以及 One-Pixel Attack[44]等方法，在不同场景下均展示了较高的攻击效果。尽管这些样本针对性对抗扰动方法能针对每个图像生成专属扰动，但由于需要为每个输入样本单独求解，计算效率较低，难以达成大规模应用。为了解决上述问题，研究者们开始探索通用对抗扰动，即寻找一种单一扰动，能够对某一数据分布中的大多数样本产生攻击效果。Dezfooli 等人[31]首次证明了通用对抗扰动的存在，并提出了一种基于多样本迭代更新扰动向量的方法，使得生成的扰动在多个样本上均能有效地干扰分类器。此后，Mopuri 等人[45]等进一步提出了数据无关的通用对抗扰动生成策略，试图在不依赖特定数据分布的情况下提高扰动的普适性；同时，还有研究采用多任务学习的方法来增强通用扰动在不同任务间的适应性。虽然通用扰动大大降低了每个样本单独求解的计算成本，并在一定程度上提升了攻击的泛化性，但它们通常直接将优化后的扰动像素叠加到原始图像上，导致图像出现明显的伪影，严重影响了视觉效果。在改善视觉效果方面，Xiao 等人[46]提出利用局部几何变换生成对抗样本的方法，通过对图像局部区域进行连续变形，在保持整体感知质量的同时实现了有效攻击。此外，也有部分工作[47, 48]利用生成对抗网络的优势，通过设计特定结构的生成器来学习生成具有较高自然度的对抗样本，既能迷惑识别模型，又能在视觉上保持较高的真实性。然而，无论是针对单个样本的方法还是通用扰动方法，都存在各自的局限性：单个样本对抗攻击方法能够针对特定图像生成高效的扰动，但因计算量大、泛化性不足而限制了大规模应用；而通用对抗扰动虽然在计算效率和适用性上有明显优势，但其直接像素叠加的策略常常导致视觉效果不佳，从而在实际使用中面临隐蔽性和真实感的挑战。

（2）不可学样本相关工作：

深度学习模型的训练往往依赖大量的数据，因此，如何防护非法的数据获取行为已成为研究者们关注的重点之一。2019 年，Shen 等人[49]提出了 TensorClog

技术，是一种专门用于保护数据隐私的中毒攻击方法，通过生成有毒数据来训练模型，导致了模型的推理精度下降，甚至可能引发梯度消失，影响模型训练进程。随后，Fowl 等人[50]在 2021 年提出了一种基于对抗样本的对抗中毒攻击技术，利用在图像中加入与目标信息相冲突的元素来降低模型的性能。研究表明，这种对抗中毒方法在预训练模型中效果尤为显著，且表现优于传统的数据中毒方法。此外，为了加强对抗攻击的隐私保护功能，Fowl 等人发布了中毒版本的 ImageNet-P 数据集。类似的，许多研究者尝试将对抗攻击用于隐私数据保护。2017 年，Koh 等人[51]提出的 Error-Maximizing 方法通过计算影响函数，揭示了不同数据样本对模型预测的影响，并据此生成不可区分的视觉攻击训练集。2019 年，Feng 等人[52]使用类似自编码器的网络结构，在训练集上生成对抗性扰动，使得模型在测试时表现出最差的分类性能。2021 年，Huang 等人[35]提出了不可学样本概念，利用模型参数与扰动的交替优化来生成损失最小化扰动，从而使得生成样本既视觉上无差异，又能使模型错误判断该样本没有学习价值，进而达到保护隐私数据的目的。然而，不可学样本仍然面临一些挑战。2021 年，Liu 等人[53]发现，不可学样本的扰动主要集中在图像的三原色通道上，易受到灰度滤波的破解，因此提出 ULEO-GrayAugs 方法，提前使用灰度滤波器处理样本以避免扰动集中在某一通道。随后，Fu 等人[54]发现这些扰动在鲁棒性上表现较差，并通过加入对抗损失来提升扰动的稳定性，从而提高不可学样本的效能。2022 年，Yu 等人[55]深入探讨了不可学样本的扰动机制，提出扰动呈现线性可分性，进而可以通过直接合成这种线性可分扰动来简化并增强其效果。2022 年，Ren 等人[56]指出，不可学样本在不同数据集和训练设置下的迁移性较差，因此提出基于类可分离性判别器的不可学样本生成策略，旨在通过增强线性可分性来改善不可学样本的迁移效果。2023 年，He 等人[57]进一步发现，监督学习中的不可学样本并不适用于自监督算法，于是提出了专门针对自监督算法的对比中毒方法，生成对自监督学习有效的不可学样本。与此同时，Zhang 等人[58]提出了基于聚类的不可学集群技术，这种技术不依赖标签信息，能够生成适用于不同训练场景的不可学样本。此外，越来越多的研究者将不可学样本应用于各类不同领域，如人脸图像[59]、时间序列[60]、医疗数据[61]、音频数据[62]以及三维点云数据[63]等领域，进一步拓展了该技术的应用范围。

1.3 本文研究内容

本文依托作者硕士期间参与的科研项目，重点研究深度神经网络应用各阶段的数据保护方法。为了保护深度神经网络应用阶段的数据安全，本文提出了基于扰动嵌入的对抗样本生成方法，以保护图像数据不被非法的深度模型使用者轻易的分析。此方法相较于常见的对抗样本生成方法来说具有更好的视觉隐蔽性。其次，为了保护人们分享在互联网的隐私数据不被非法用于训练神经网络，本文提出了基于扰动嵌入的不可学样本生成方法。该方法生成的不可学样本不仅能有效阻碍了模型从隐私数据中获取信息，而且保持了较高视觉保真度，不影响日常生活的使用。上述两种算法具体如下：

基于扰动嵌入的对抗样本生成方法：本文提出了基于扰动嵌入的对抗样本生成框架 PtEm-AE(Perturbation Embedding based Adversarial Example)，不同于现有的直接在像素上以加和的方式添加扰动的对抗样本生成方法，PtEm-AE 借助深度网络隐写技术，利用编码网络完成通用扰动的嵌入。此外，本文分别从样本层面与数据分布层面提出了两种互补的对抗攻击损失函数，以系统性地增强所生成对抗样本的攻击性与迁移性。样本层面本文借鉴了对比学习的思想，计算一批原始图像及其对应对抗样本之间的相似度矩阵，来增强原始图像与对抗样本在特征空间中的距离。数据分布层面，计算原始图像与对抗样本各自特征的相似度矩阵，并通过归一化操作将其转化为概率分布。最后，使用 KL 散度增大这两个分布之间的差异，从而促进模型学习到能够有效改变样本分布的扰动。为兼顾攻击性与隐蔽性，引入视觉隐蔽性损失以保障对抗样本的视觉质量。

基于扰动嵌入的不可学样本生成方法：本文介绍了一种保护隐私数据不能被用于训练模型的方法 PtEm-UE(Perturbation Embedding based Unlearnable Example)，提出了基于扰动嵌入的不可学样本生成框架。该框架的训练过程分为两步，首先训练一个扰动嵌入网络。该网络利用隐写技术将扰动嵌入到原始图像中生成不可学样本。同时引入与之交替训练的代理模型模拟真实的不可学样本使用场景，引导扰动嵌入网络生成具备初步不可学性的不可学样本。第二步，寻找不可学特性最强的扰动。在此阶段，扰动嵌入网络的参数将被冻结，而将每个样本对应的

扰动作为优化目标，交替优化扰动和代理模型寻找数据保护效果最强的扰动。此外，在损失设计方面，本文综合考虑对比学习和监督学习场景设计了针对性的损失函数保障生成的不可学样本同时对对比学习和监督学习有效。

1.4 论文的组织结构

全文共有五个章节，其中各个章节的内容安排如下图 1.2 所示：

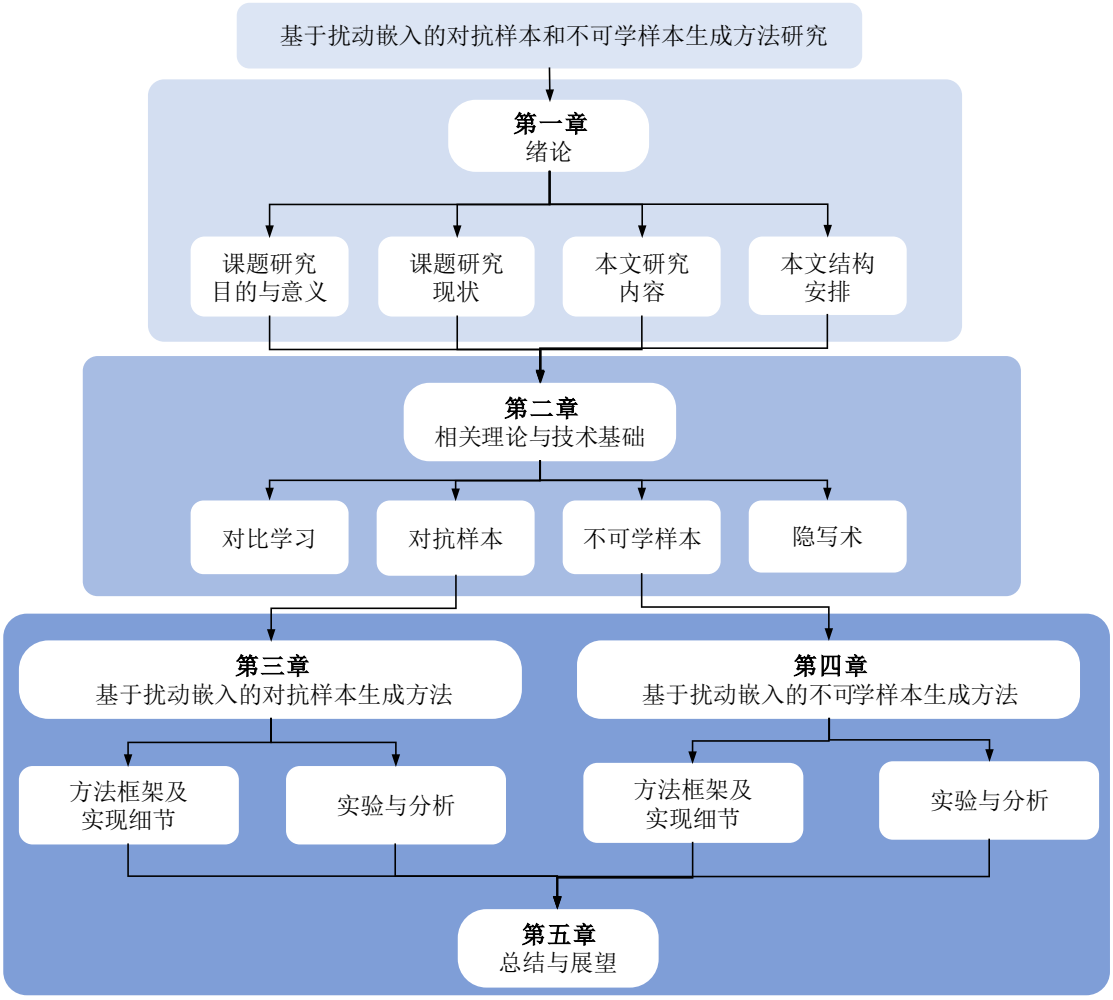


图 1.2 论文的结构

Figure 1.2 Structure of Dissertation

第一章为绪论，介绍了深度神经网络的研究背景及存在的数据安全隐患，阐明了数据隐私保护具有重要意义；总结和分析现有方法的优点和不足；最后对本文的主要研究内容和结构安排进行介绍。

第二章为相关理论与技术基础。首先介绍了当前主流的深度学习网络及对应的训练算法。然后介绍了本文对抗样本和不可学样本都应用到的相关技术——隐写术。接着介绍了现有数据保护的场景以及关键技术。最后对文中使用的数据集进行了简要介绍。

第三章为基于扰动嵌入的对抗样本生成方法。阐述和分析了所提出的对抗样本生成框架中的扰动嵌入网络的实现细节和损失设计。然后在多种数据集和模型结构上对其有效性、迁移性和鲁棒性进行了全面的评估和分析。

第四章为基于扰动嵌入的不可学样本生成方法。首先介绍了所设计的不可学样本生成框架的实现细节，并在多个数据集、模型和算法上进行全面的分析和评估，并与现有方法进行对比，阐明该方法的优势。

第五章为总结与展望，系统回顾全文核心贡献与方法论框架，批判性分析研究局限性，并基于当前不足前瞻性探讨未来发展趋势。

第二章　相关理论与技术基础

2.1　相关模型与算法

在深度学习技术演进中，监督学习始终是驱动突破的核心引擎。如下图 2.1 所示，通过建立输入数据与人工标注标签的精确映射，该范式在 ImageNet 图像分类任务中创造了从 AlexNet[64]（2012 年 Top-5 准确率 84.7%）到 Swin Transformer[65]（2022 年 Top-5 准确率 90.5%）的技术奇迹，并在目标检测、语义分割等领域催生了 Faster R-CNN[66]、Mask R-CNN[67]等里程碑模型。然而，该范式面临三重核心挑战：其一，标注成本随数据维度指数增长，其二，模型易受标注噪声干扰，其三，过拟合风险在数据有限场景下显著加剧，这种强标注依赖性使模型陷入“数据饥渴”困境：当训练数据规模突破千万级后，准确率提升呈现显著边际效应，且专业领域模型泛化能力急剧衰减。

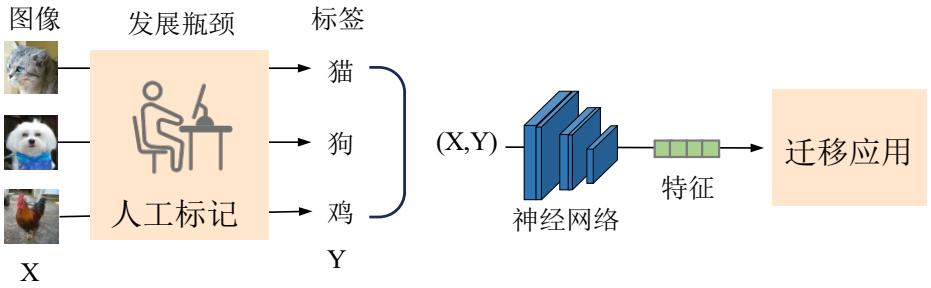


图 2.1 监督学习总体流程

Figure 2.1 The General Process of Supervised Learning

无监督学习通过重构数据价值挖掘范式打破僵局，其进化轨迹呈现三级跨越：早期无监督方法（聚类/PCA）受限于线性假设，在 CIFAR-10 数据集上分类准确率不足 60%；代理任务阶段（旋转预测/拼图复原）通过构造伪标签将 ResNet-50 表征能力提升至监督学习的 78.4%，但仍存在语义偏差；直到对比学习实现范式重构，通过样本关系建模开启新纪元。工业界验证了该技术路线的普适性：蚂蚁金服[68]通过对比学习构建 10 亿用户跨模态画像，使金融推荐点击率提升 14 个百分点；特斯拉[69]利用对比增强生成 200 万帧极端场景，将自动驾驶 Corner Case

检测率从 73%提升至 89%。这些实践昭示着，当监督学习陷入标注成本与模型泛化的二律背反时，对比学习正通过数据关系的量子化挖掘，重塑人工智能的燃料供给体系。本文的第一个工作借鉴了对比学习的思想，设计了针对对抗样本的损失函数；第二部分则对现有常见的对比学习算法进行了分析，并设计了相应的不可学样本。

2.1.1 卷积神经网络

卷积神经网络（Convolutional Neural Networks, CNN）通过仿生学启发的层次化特征抽象机制，彻底变革了计算机视觉任务的范式。数学上，给定输入特征图$X \in R^{H\times W\times C}$与卷积核$K \in R^{k\times k\times C\times D}$，卷积层的输出可表述为：

$$Y_{i,j,d} = \sum_{c=1}^{C}\sum_{u=-s}^{s}\sum_{v=-s}^{s} X_{i+u,j+v,c}\cdot K_{u+s,v+s,c,d} + b_d \tag{2.1}$$

其中$s = \lfloor k/2 \rfloor$，b_d为偏置项。该操作通过限制卷积核的局部连接性与跨位置参数复用，将全连接网络的参数量从$O((HWC)^2)$降至$O(k^2CD)$，解决了高维数据处理的维数灾难问题。LeCun 等人提出的 LeNet-5[70]首次验证了 CNN 在手写数字识别中的有效性。2012 年，Krizhevsky 等人提出的 AlexNet[64]通过多项革新推动CNN进入深度化时代：采用ReLU激活函数（$f(x) = max(0, x)$）替代Sigmoid，其非饱和特性使梯度衰减率降低至传统函数的 1/4，训练速度提升 6 倍；引入 Dropout（随机失活率$p = 0.5$）与数据增强策略，通过特征空间扰动将 ImageNet Top-5 错误率从 26.2%压缩至 16.4%；同时，双 GPU 流水线设计突破单卡显存限制，首次实现大规模 CNN 的并行化训练，为后续分布式训练框架奠定硬件基础。此后，CNN 架构设计沿着深度扩展、效率提升与多尺度融合三条轴线持续演进。经典的卷积网络结构如下图 2.2 所示，本文实验所使用的主干网络包括经典的 VGGNet[32]、ResNet[28]和 DenseNet[71]。

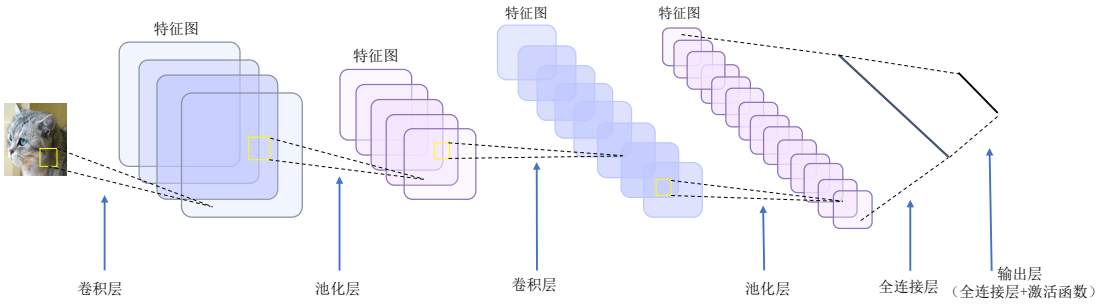


图 2.2 卷积神经网络经典结构

Figure 2.2 Classic Architecture of Convolutional Neural Networks

VGGNet(Visual Geometry Group Network)通过系统性实验验证了小尺度卷积核堆叠的深度网络设计范式，这一成果对现代卷积神经网络架构设计产生了深远影响。其核心贡献不仅在于提出使用连续 3×3 卷积核替代大尺度卷积核的可行性，更从数学上证明了这种设计的参数效率优势：通过级联两个 3×3 卷积层可获得与单个 5×5 卷积层等效的感受野，同时参数量减少至后者的 9/25，显著缓解了深层网络的过拟合风险。VGG-16 作为其典型实现，采用 13 个卷积层与 3 个全连接层的级联架构，配合 ReLU 激活函数与 L2 正则化策略，在 ImageNet 数据集上实现 Top-5 分类误差 7.3%的突破性性能。值得注意的是，其分层特征提取机制为后续目标检测与语义分割任务中的特征金字塔网络提供了理论基础。

ResNet(Residual Neural Network)如下图 2.3 所示，通过残差学习框架重构了深度神经网络的优化空间，为解决网络深度增加导致的梯度衰减与表征瓶颈问题提供了理论突破。其创新性残差模块可形式化定义为：$H(x) = F(x, W_i) + W(x)$ 其中，$F(x)$为待学习的残差函数，$W(x)$为维度匹配函数。该设计通过引入跨层恒等连接，使反向传播过程中的梯度可直接绕过非线性变换层传递，从而理论上保证深层网络的训练稳定性。数学分析表明，残差结构将传统网络的优化目标从学习复杂函数$H(x)$转换为学习残差$F(x) = H(x) - x$，显著降低了参数空间的搜索复杂度。特别地，ResNet-50 采用的“Bottleneck”模块（1×1→3×3→1×1 卷积序列）通过降维与升维操作减少计算量，其参数效率较标准残差块提升约 35%，为平衡特征提取效能与计算开销提供了实验基础。其与传统 CNN 中的普通连接对比如图 2.4 所示。ResNet 系列包含多个不同深度的版本，如 ResNet-18、ResNet-34、ResNet-50、ResNet-101、ResNet-152 等，适用于不同任务复杂性和计算资源场景。本文实验主要使用了相对轻量的 ResNet-18，中等深度的 ResNet-34、ResNet-

50 模型和更复杂的 ResNet-101，以便在不同资源环境下进行对比分析。

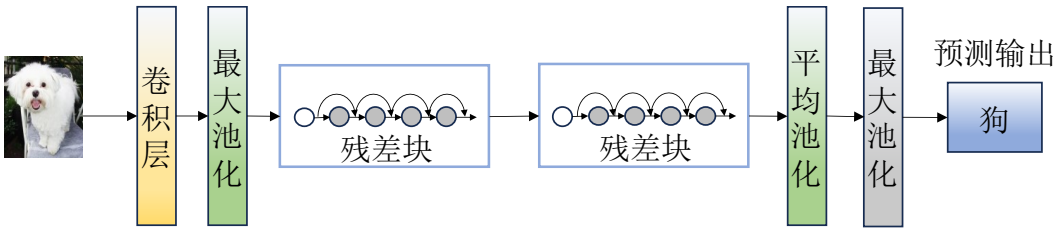


图 2.3 ResNet 模型结构

Figure 2.3 ResNet Model Architecture

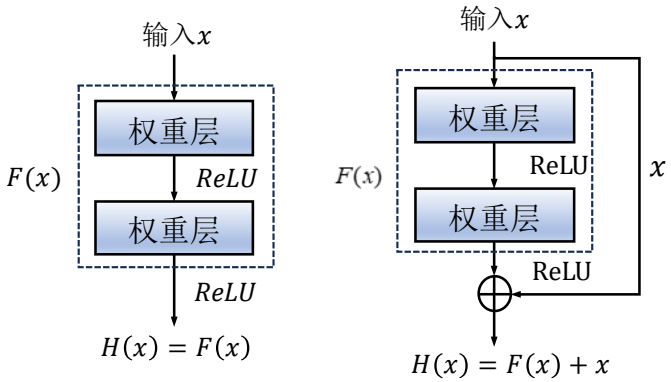


图 2.4 传统卷积的普通连接与 ResNet 模型残差连接的对比

Figure 2.4 Comparison of Traditional Convolutional Connections and Residual Connections

DenseNet(Densely Connected Convolutional Networks) 提出了如下图 2.5 所示的跨层密集连接机制，重新定义了特征复用的拓扑范式。其第l层的输出可表述为：$x_l = H_l([x_0, x_1, \cdots, x_{l-1}])$其中，[·]表示沿通道维度的拼接操作，$H_l(\cdot)$为包含批量归一化、ReLU 激活与 3×3 卷积的复合函数。这种密集连接模式通过最大化特征图复用率，使网络在参数量减少 50%的情况下达到与 ResNet 相当的分类性能。数学分析表明，DenseNet 的特征梯度可沿所有前置层路径反向传播，其梯度幅值满足：

$$\frac{\partial L}{\partial x_0} = \sum_{l=1}^{L} \frac{\partial L}{\partial x_l} \prod_{i=1}^{l} \frac{\partial x_i}{\partial x_{i-1}} \tag{2.2}$$

这种梯度分散效应显著缓解了梯度消失问题，但同时也导致计算图复杂度随深度呈$O(L^2)$增长。为此，DenseNet 引入过渡层，通过 1×1 卷积压缩特征通道数并结合 2×2 平均池化实施空间下采样，最终将计算量控制在 ResNet 的 60%~70%范围内。

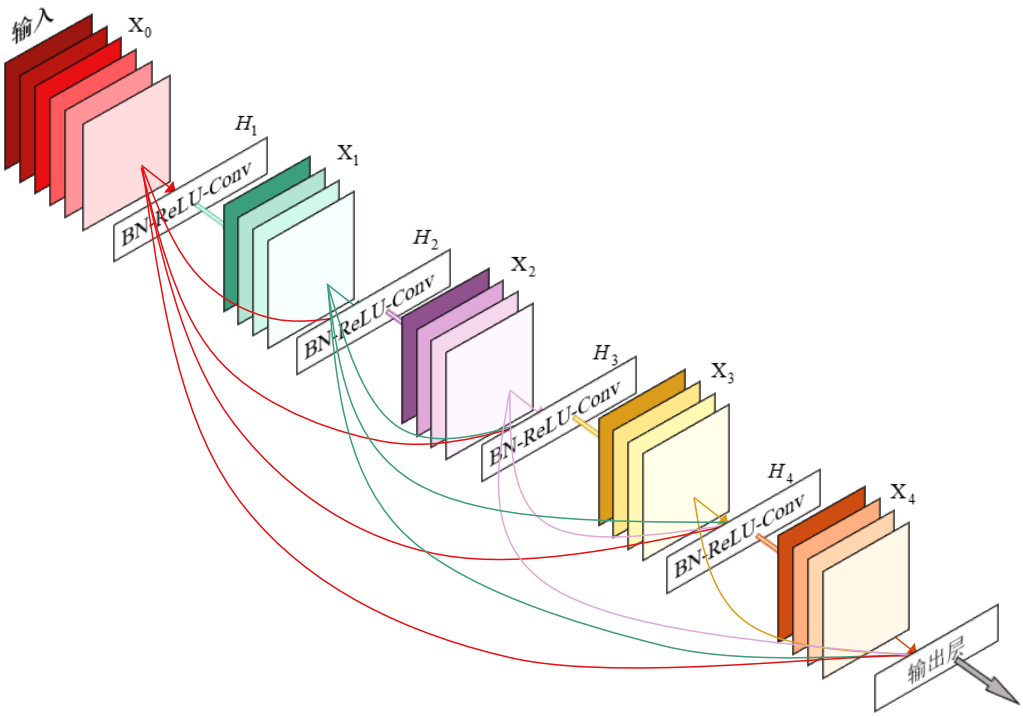


图 2.5 DenseNet 模型首次提出的密集连接机制

Figure 2.5 DenseNet Model's Novel Dense Connection Mechanism

2.1.2 对比学习

在深度学习的技术发展进程中，监督学习与无监督学习代表了两种不同的数据驱动范式。监督学习作为机器学习领域的核心范式，其本质是通过带标签数据集训练模型以建立输入特征与目标输出之间的映射函数。当前，监督学习正与自监督预训练、对比学习等技术深度融合，逐步演进为“预训练-微调”的新型范式，持续推动人工智能系统的实用化进程。对比学习是自监督学习中的一种具有代表性的方法，其核心思想是通过最大化正样本对之间的相似性，并最小化负样本对之间的相似性，从而有效地提取数据的特征表示。在训练过程中，卷积网络的分类头通常被去除，网络仅保留主干部分用于特征提取，并根据具体算法添加映射头，将特征映射到低维空间以提高特征的判别性。以下是本文第二个工作所探讨的几种常见的对比学习算法及其原理介绍：

SimCLR（Simple Contrastive Learning of Representations）

SimCLR 是由 Google Brain 于 2019 年提出的对比学习方法，它的整体框架

如图 2.6 所示，通过简单而高效的数据增强技术生成正负样本对，并通过编码器网络和投影头将样本映射到低维空间。其损失函数为对比损失，定义为：

$$L(i,j) = -log\frac{exp\left(\frac{sim(z_i,z_j)}{\tau}\right)}{\sum_k^{2N} exp\left(\frac{sim(z_i,z_k)}{\tau}\right)} \tag{2.3}$$

其中$sim(z_i,z_j)$是样本i和样本j的相似度，τ是温度参数，z_i和z_j是输入样本的特征表示。SimCLR 通过批量处理生成$2N$个增强样本，使同一原始样本的增强版本靠得更近，同时远离其他样本，从而有效提升特征表示的质量和模型的泛化能力。

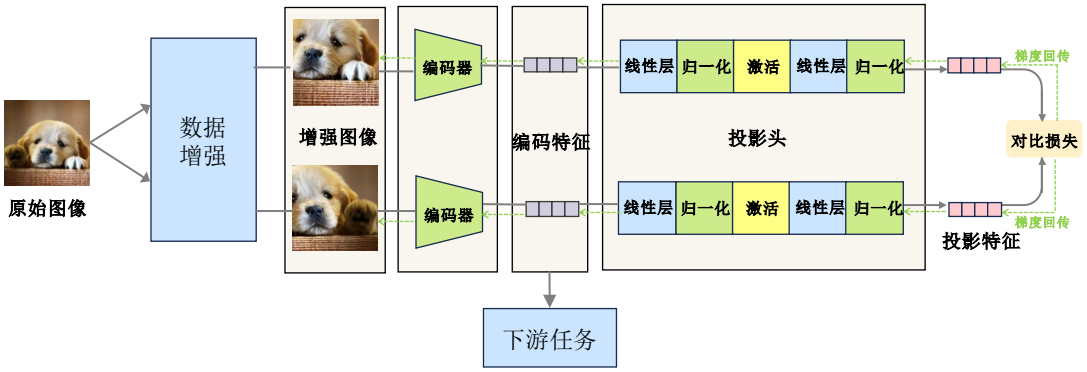


图 2.6 SimCLR 算法流程图

Figure 2.6 SimCLR Algorithm Flowchart

MoCo v2（Momentum Contrast v2）

SimCLR 框架的核心缺陷源于其对比学习机制对大容量负样本的依赖性，这迫使模型训练时必须采用超常规批量尺寸（典型配置为 4096）以维持表征判别力。为实现该批量规模的技术可行性，研究者不得不引入 LARS 优化器并实施跨 GPU 分布式批归一化处理，由此形成的多重技术准入门槛严重制约了该方法的工程适用性。如下图 2.7 所示，MoCo 框架通过构建静态特征库（memory bank）实现历史样本特征的持续性存储，但该机制存在特征时效性衰减缺陷——存储特征因参数迭代滞后产生表征漂移。针对此局限，MoCo v2 创新性地融合 SimCLR 的数据增强策略与双编码器架构：在线编码器通过梯度反传更新，动量编码器采用指数移动平均渐进同步，二者构成动态特征协同系统。改进后的队列式存储机制结合动量编码器的延迟更新特性，使负样本特征兼具时间一致性和对比差异性，最终将 ImageNet 线性评估精度提升 6.2 个百分点。

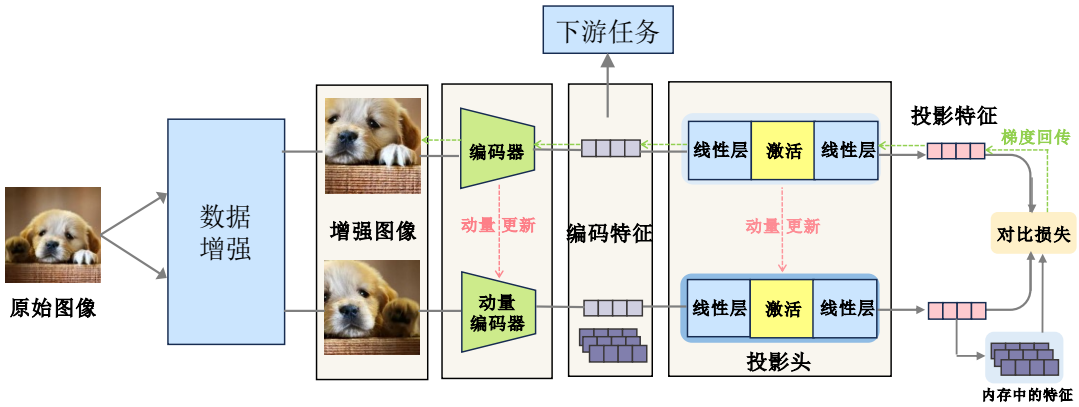


图 2.7 MoCo v2 算法流程图

Figure 2.7 MoCo v2 Algorithm Flowchart

BYOL（Bootstrap Your Own Latent）

尽管 SimCLR 依赖超大批量维持负样本规模，而 MoCo 系列通过动量编码器与队列存储实现历史特征复用，但二者本质上仍受限于对比学习对负样本的强依赖性。如下图 2.8 所示，而 DeepMind 提出的 BYOL 则突破性构建了非对称双流架构：在线网络（online model）通过梯度更新，驱动预测头学习目标网络（target model）的动量特征，而目标网络仅通过$\theta_{target} \leftarrow \eta\theta_{target} + (1-\eta)\theta_{online}$实现参数滞后更新。该方法摒弃了显式负样本构造，仅通过正样本对的特征预测一致性（MSE 损失：$\| q(z_{online}) - z_{target} \|^2$）实现自监督学习，在 ImageNet-1K 上达到 74.3%线性评估精度的同时，将训练内存消耗降低 63%。这一创新不仅规避了 SimCLR 的批量敏感性与 MoCo 的特征时效性衰减问题，更通过动态教学机制揭示了自监督学习中负样本非必要性的理论可能性，为后续研究开辟了新范式。

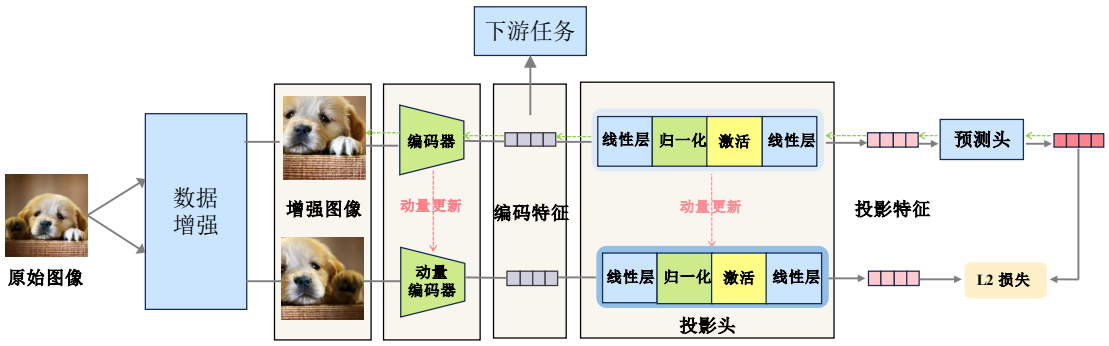


图 2.8 BYOL 算法流程图

Figure 2.8 BYOL Algorithm Flowchart

SimSiam (Simple Siamese Representation Learning)

SimSiam[72]可视为 BYOL 架构的极简主义演化——在保留双分支编码器与预测头（projection head）的基础上，通过参数共享的对称编码器架构替代动量更新机制。如下图 2.9 所示，SimSiam 的核心创新在于“梯度截断”（stop-gradient）操作：在线网络通过反向传播更新参数，而目标网络直接镜像在线网络的瞬时参数快照（$\theta_{target} \leftarrow \theta_{online}$，$\theta_{online}$和$\theta_{target}$其中分别为在线网络和目标网络的参数），同时阻断梯度流以预防模型坍缩。这种设计在 ImageNet-1K 上实现 71.3% 线性评估精度（较 BYOL 仅下降 3%），却无需动量编码器或历史特征存储队列，将 GPU 内存占用再压缩 41%。该框架揭示了自监督学习中动量更新的非必要性，通过参数同步截断与对称约束构建隐式对比空间，既继承了 SimCLR 的架构简洁性，又克服了 MoCo 系列对动态存储的工程依赖，为边缘设备部署提供了新的可行性路径。

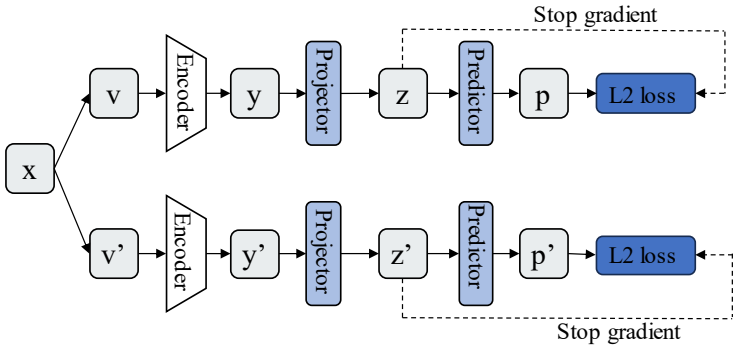


图 2.9 SimSiam 算法流程图

Figure 2.9 SimSiam Algorithm Flowchart

2.2 基于深度网络的隐写方法

隐写术通过将信息隐蔽嵌入载体数据实现秘密通信，其技术演进始终围绕隐蔽性、容量与鲁棒性的平衡展开。传统方法从空域 LSB 替换逐步发展为频域统计融合（如 F5 算法[73]），但手工设计特征难以应对现代检测算法。近年来，基于深度学习的隐写方法通过端到端特征学习重构嵌入范式：利用编码器-解码器架构实现自适应信息隐藏，噪声层增强对抗信道干扰的鲁棒性，扩散模型突破载体依赖实现语义级嵌入。受深度网络隐写思想启发，本文提出新型扰动嵌入机制：

不同于传统隐写需独立编解码过程，本文将扰动视为“隐写信息”，利用扰动嵌入网络将其嵌入原始图像中（而非直接像素叠加），在保障视觉质量的同时实现攻击效力。关键差异在于：传统隐写需确保信息可提取性，而本文以攻击效果作为扰动嵌入的隐式验证——当对抗样本成功误导模型或不可学样本阻断模型训练时，即证明扰动被有效隐写。这一设计摆脱了显式解码约束，使优化目标聚焦于隐蔽性与攻击效能的协同提升，实验表明其 SSIM 指标较像素叠加方法最高提升 0.1，进一步验证了该技术路线的优越性。

2017 年，Hayes 等人[74]提出了以图像藏文本的隐写框架——HayesGAN，这是首个利用编码网络实现图像隐藏信息的深度学习方法。该框架将秘密信息与载体图像一起输入到编码器中，编码器直接输出含有隐藏信息的图像；接着，解码器接收含密图像并解码出其中的秘密信息。为了确保隐藏图像在视觉上的真实性和安全性，Hayes 等人还引入了判别器进行分析和检测，确保含密图像能够成功通过检测。为提高隐写的鲁棒性，Zhu 等人提出了一种名为 HiDDeN[75]深度图像隐写方式的改进方法。他们在隐写网络训练过程中，加入了噪声层，模拟了在实际传输过程中图像可能遇到的噪声攻击、压缩等退化情况。该方法通过将噪声影响下的含密图像输入解码网络中提取秘密信息，有效提升了隐写的鲁棒性，并为后续隐写方法的提升提供了思路。为了进一步增加隐写容量，SteganoGAN[76]隐写模型被提出，采用密集连接来缓解梯度消失问题，在训练过程中使用多个损失函数同时优化解码器和评估网络。SteganoGAN 成功地将任意数据嵌入到各种自然场景中提取的图像中，同时能够有效绕过传统的检测工具。北京大学团队进一步将扩散模型引入图像隐写领域，提出了一种无载体隐写框架 CRoSS[77]。该方法利用 Stable Diffusion[18]的生成能力和 DDIM[78]确定性采样特性，通过双文本描述将隐藏信息嵌入到生成的容器图像中，而无需依赖传统的载体图像。解码时，通过逆向扩散过程恢复隐藏信息。这一方法显著提升了对抗检测算法的安全性，并在图像压缩、噪声干扰等退化条件下表现出更强的鲁棒性。该方法不仅提高了隐写的安全性，而且解决了传统方法在面对噪声或压缩时的性能下降问题。

2.3 对抗样本

近年来，随着深度学习模型在多个领域的广泛应用，针对非法利用这些模型进行隐私数据分析的挑战变得愈发严峻。例如，人脸识别系统的生物特征滥采、自然语言处理模型对敏感文本的解析等，这些问题已引起学术界和工业界的广泛关注。为此，研究者提出了对抗样本技术作为核心防御手段。该技术通过向原始数据中注入人眼不可察觉的细微扰动，显著降低模型对隐私特征的敏感性，从而有效阻止攻击者使用深度神经网络进行隐私数据分析。其理论依据来自于深度神经网络在高维空间中的线性脆弱性，通过精心设计的扰动，能够在保持人类感知不可区分性的同时，显著改变模型的输出。目前，主流的对抗样本生成方法可分为两大类：一类是样本针对性对抗扰动（Sample-specific Adversarial Perturbation），通过逐样本优化的方式构造针对性扰动；另一类是样本无关对抗扰动（Sample-agnostic Adversarial Perturbation）也即通用对抗扰动（Universal Adversarial Perturbation, UAP），UAP 通过训练一个适用于所有样本的全局或局部扰动，使其能够跨样本、跨数据分布地破坏模型的特征提取能力。与样本针对性对抗扰动相比，通用对抗扰动具有显著的计算效率优势，只需单次生成即可泛化至未知样本。这两类方法在防御机理上互为补充：前者通过精确的局部优化实现强对抗性，后者则依赖数据分布的全局特性实现高效的普适性，二者共同构成了对抗样本技术在隐私保护领域的核心方法论基础。本文的第一个研究工作便围绕通用对抗扰动展开。

2.3.1 样本针对性对抗扰动

近年来，样本针对性对抗扰动生成方法已成为保护隐私数据的关键技术，其研究脉络已从经典的理论基础逐步扩展至多维度优化领域。早期的奠基性工作由 Goodfellow 等人[79]提出，他们首次通过快速梯度符号法展示了深度神经网络在面对梯度扰动时的脆弱性，为对抗样本生成奠定了理论基础。在此基础上，Madry 等人[39]提出的投影梯度下降法（PGD）通过多步迭代优化和随机初始化策略，显著提高了对抗攻击的鲁棒性，成为对抗训练的黄金基准。Kurakin 等人[43]提出的迭代快速梯度符号法以及 Moosavi 等人[41]提出的 DeepFool 算法，分别从物理攻

击的可行性和最小扰动优化的角度拓展了对抗样本的应用边界。随着研究的深入，对抗攻击逐渐迁移到更为复杂的场景，Dong 等人[43]引入动量机制来增强扰动的迁移性，从而提高黑盒攻击的效率；而 Suya 等人[80]提出了一种混合批量攻击方法，将传输攻击和优化攻击相结合，利用局部模型生成候选对抗样本，并通过优化在有限查询的情况下进一步提升了黑盒攻击的效率。同年，Xiao 等人[47]提出了基于局部几何变换（如平滑的图像形变）的对抗样本生成方法，该方法在保持图像感知质量的同时，显著提高了对抗样本的真实性和鲁棒性。

与此同时，Shamsabadi 等人[81]则提出了 ColorFool 方法，通过语义引导的局部颜色篡改，在保持攻击效能的同时提升了视觉自然性。值得关注的是，Duan 等人[82]提出了 AdvDrop 方法，通过选择性丢弃输入中的高频信息，揭示了模型对特定特征的依赖性，为对抗攻击提供了逆向解释的新视角。当前的研究趋势表明，对抗样本生成方法已经从单纯的误差最大化，向隐蔽性、物理可实现性及跨模型泛化能力等多维度目标演进，并与防御技术的动态博弈共同推动着隐私保护与模型鲁棒性理论的协同发展。

2.3.2 通用对抗扰动

通用对抗扰动作为对抗样本技术的重要分支，近年来因其输入无关的普适攻击特性而备受关注。早期的经典研究由 Moosavi 等人[31]提出，首次通过迭代优化生成单一扰动向量，使其能够泛化至不同输入样本，成功揭示了深度模型在高维特征空间中的共享脆弱性。该方法基于数据分布的统计特性，通过最大化扰动对分类边界的累积偏移量，实现了对 ImageNet 数据集的平均攻击成功率超过 80%，为后续的研究奠定了理论基础。随后，Dong 等人[43]通过引入动量机制，优化了梯度方向的稳定性，显著提高了扰动的迁移性，为黑盒攻击效率的提升提供了新思路。2019 年，Brown 等人[83]提出的 Adversarial Patch 则突破了数字攻击的局限，验证了物理可实现的对抗补丁的可行性，通过打印扰动图像并通过摄像头拍摄，仍能成功欺骗模型。进入 2020 年后，研究逐渐集中于生成效率与隐蔽性的优化。Xu 等人[84]利用残差网络构建生成器，将随机噪声映射为通用扰动，在 CIFAR-10 数据集上实现了 89%的攻击成功率，标志着生成式方法在 UAP 领域

的崛起。

在 2022 年，Kang 等人[85]提出了通道激活抑制方法，通过动态调整模型内部特征响应来提升鲁棒性，从而间接推动了 UAP 生成需更精细的模型结构适配。近年来，UAP 的研究在隐蔽性、生成效率及泛化方面取得了显著突破。例如，Ye 等人[86]利用了神经崩塌现象创造具有较强迁移性的对抗扰动。通过攻击发生 NC 的层，FG-UAP 能够生成更强且更有效的对抗扰动，将自然图像的特征聚集到一个新的方向从而提升生成对抗扰动的泛化性。这些进展不仅延续了经典 UAP 的全局扰动特性，也在生成效率、隐蔽性及泛化性上实现了方法论的创新，为隐私保护和模型安全测试提供了新的技术路径。

2.4 不可学样本

在防止数据被恶意用于模型训练方面，主流的方法一直依赖于数据投毒技术来干扰模型的学习过程。这些方法通常通过在数据集中添加噪声或注入恶意样本（例如标签翻转、特征污染），迫使模型在训练过程中产生错误或不确定的结果，从而降低数据的有效利用率。尽管这种方式在一定程度上有效保护了数据隐私和安全，但其缺点在于对数据本身的干扰较为显著。近年来，研究者提出了一种全新的数据隐私保护思路——不可学样本。这种方法与传统的数据投毒不同，不仅能够通过交替优化模型参数和扰动，破坏模型在监督学习中的有效学习，而且已被扩展到自监督（如对比学习）领域。不可学样本在增强数据保护效果的同时，具有更好的鲁棒性和视觉隐蔽性，为防止未经授权的模型训练提供了全新的解决思路。

2.4.1 面向监督学习的不可学样本

2021 年，Huang 等人[35]提出不可学样本（Unlearnable Examples）的概念，通过向训练数据注入特定噪声，使模型无法有效学习其内在特征。其创新性体现在将传统对抗攻击的单一优化目标扩展为双级优化框架：在扰动生成过程中，同时优化模型参数和扰动，使得添加扰动后的样本既能最小化模型损失，又能在视觉上保持隐蔽性，为数据隐私保护提供了新的解决方案。具体而言，给定原始样

本x及其标签y，扰动生成需满足约束条件$\|\delta\|_p \leq \varepsilon$（即扰动在$p$范数下的幅度不超过阈值$\varepsilon$，以保证生成的不可样本的视觉隐蔽性）。构造的优化目标函数可表述为：$min_\theta min_\delta\, L(f'(x+\delta;\,\theta), y)$其中$f'(\cdot)$为用于生成噪声的基准模型，$L$为交叉熵损失函数。为实现这一目标，Huang 等人设计了交替优化策略：首先固定模型参数θ，通过投影梯度下降更新扰动δ；随后固定扰动δ，通过常规训练更新模型参数θ。此后，Liu 等人[53]发现上述方法生成的不可学样本的扰动主要集中在三个颜色通道，仅需要使用简单的灰度滤波技术便可以轻松的消除这一扰动使不可学样本变得可学，于是他并进一步提出了 ULEO-GrayAugs 算法以解决这一问题。此外，Fang 等人[87]提出了一种从数据分布角度审视数据隐私的创新方法，通过生成可迁移的不可学样本，增强了数据保护能力。Liu 等人[88]则进一步提出了稳定误差最小化噪声技术，这一方法通过训练防御性噪声以对抗随机扰动，增强了不可学样本的鲁棒性，从而在防止未经授权模型训练方面发挥了更大的作用。

2.4.2 面向对比学习的不可学样本

尽管不可学样本在监督学习中已取得显著进展，但其在自监督学习中的应用仍面临一定挑战。2023 年，He 等人[57]提出对比中毒方法，将不可学样本的概念扩展至自监督学习领域。具体而言，首先选择一种自监督算法（如 SimCLR、MoCo 或 BYOL），然后通过双重最小优化对编码器和扰动进行交替优化。研究表明，对比中毒方法生成的不可学样本具有更强的鲁棒性，并且采用动量编码器（如 MoCo 和 BYOL）的模型，相较于不包含动量编码器的模型（如 SimCLR），表现出更强的防御能力。通过动量编码器生成的不可学样本也能更好地迁移到其他自监督算法中。后续的 TUE[56]框架则揭示了现有不可学样本在训练方式和数据集迁移性上的局限性，并提出基于类间可分性判别的不可学样本生成框架，显著提高了不可学样本的迁移性。与传统方法不同，TUE 不仅能保持在监督学习和自监督学习中的不可学效果，还能够跨数据集迁移，显著提高了不可学样本的实用性和效率。

2.5 本文相关数据集

本文实验采用了三种常见的图像分类数据集，包括 ImageNet[3]、CIFAR-10[38]以及 MS COCO[89]。

ImageNet：作为计算机视觉领域最具影响力和重要性的图像数据集之一，ImageNet 由斯坦福大学的研究人员创建，旨在用于视觉对象识别和目标定位挑战。该数据集包含超过 1400 万张经过人工精细标注的图像，涵盖了来自 1000 个不同类别的各种真实世界场景和物体。每个类别中的图像都经过详细的标注，确保数据的高质量与准确性。ImageNet 不仅为深度学习模型的训练提供了丰富的资源，而且成为了图像分类和目标检测等多项任务的标准评估基准，其影响深远，推动了计算机视觉技术的迅猛发展。

CIFAR-10：CIFAR-10 是由加拿大高级研究所 University of Toronto 创建的图像分类数据集。该数据集包含 60,000 张彩色图像，涵盖 10 个不同的类别，每个类别有 6,000 张图像。CIFAR-10 数据集的训练集包含 50,000 张图像，测试集包含 10,000 张图像。数据集中的图像在不同的角度、背景和光照条件下展现了图像样本的多样性，这使得其在评估图像分类算法的性能时，具有一定的挑战性。CIFAR-10 广泛用于学术研究，是图像分类领域中常见的测试数据集。

MS COCO：MS COCO（Microsoft Common Objects in Context[89]）是一个广泛用于目标检测、图像分割和图文问答的图像数据集。该数据集包含了超过 33 万张图像，其中 80 万个物体实例被精确地标注。MS COCO 的数据覆盖了 80 个不同类别的物体，涵盖日常生活中的常见物体和场景。不同于传统的数据集，MS COCO 提供了更为复杂的场景背景和物体之间的上下文关系，适合于多种计算机视觉任务的研究，尤其是在目标检测、语义分割和图像描述生成等任务中。其多样性和复杂性使得它成为了图像理解任务中的一个重要标准数据集，为相关研究提供了宝贵的资源。

2.6 本章小结

本章详细介绍了论文实验部分涉及的模型及训练算法和所使用的经典卷积

神经网络。在自监督学习中，通过对比学习，模型能够通过增强样本的方式提取有效的特征表示，从而提升泛化能力和任务性能。我们分析了几种主流的对比学习算法，如 SimCLR、MoCo v2、BYOL 和 SimSiam，并对比了它们在特征学习中的优势和局限性。除此之外，本章还涵盖了基于深度网络的隐写技术，强调了在隐写术中的图像与信息隐藏方法的演进，尤其是通过深度神经网络提高隐写技术的鲁棒性与隐蔽性。本章还讨论了对抗样本的生成机制，重点介绍了通用对抗扰动及其在隐私保护中的应用。最后，介绍了不可学样本的概念及其在监督学习与自监督学习中的不同实现，阐述了如何通过扰动与优化算法阻止模型学习敏感数据。本章的研究为后续章节的实验设计与方法分析奠定了坚实的理论基础。

第三章　基于扰动嵌入的对抗样本生成方法

3.1　引言

随着深度神经网络在医疗诊断、社交推荐、自然语言处理等领域的广泛应用，其潜在的隐私泄露风险引发了学术界的高度关注。研究证实，攻击者可能滥用深度神经网络的分析能力，从用户数据中挖掘出敏感隐私信息：人脸识别系统可通过高精度特征提取实现跨平台身份关联；社交平台推荐算法可被恶意操纵来预测用户收入水平等敏感属性；自然语言处理模型甚至能通过语义解析暴露文本中的机密内容。此类技术滥用使得原本旨在提升社会效率的人工智能系统，意外演变为隐私泄露的隐蔽通道。在此背景下，对抗样本技术逐渐发展为隐私保护领域的一种重要防护机制。其核心机理是通过在原始数据中嵌入不可感知的细微扰动，系统性地降低深度网络对隐私特征的提取能力。例如，Huang 等人[90]首次提出通过引入人耳难以察觉的噪声来防止他人滥用语音转换技术。而 Salman 等人[91]的研究则提出了通过在图像中嵌入微小、不可察觉的对抗扰动，来保护用户图像免受恶意扩散模型编辑的方法，从而有效防止未经授权的图像编辑。

尽管对抗样本生成技术已取得显著进展，但传统方法普遍受限于需为每个输入独立生成特定扰动的单样本攻击范式，导致计算效率低下且跨样本泛化能力薄弱。为此，Moosavi 等人[31]开创性地提出通用对抗扰动（Universal Adversarial Perturbations, UAP），通过迭代优化生成与输入无关的全局扰动向量，在 ImageNet 数据集上实现超过 80%的平均攻击成功率，揭示了深度模型在高维特征空间的共享脆弱性，该优化框架启发了后续大量衍生研究。然而，现有方法多采用像素空间线性叠加方式，导致生成样本存在显著视觉伪影，严重破坏图像可用性，同时面临跨模型迁移效率受限的瓶颈。近年来研究重点转向隐蔽性与泛化性的协同优化，其中 Ye 团队[86]提出的 FG-UAP 框架通过利用神经崩塌现象，在模型特征坍缩层构建对抗子空间，使生成的扰动在保持高攻击迁移性，为突破传统 UAP 的优化约束提供了新思路。这些进展不仅深化了对深度模型脆弱性的理论认知，也

为隐私保护与模型安全测试开辟了创新技术路径。

与此同时，深度网络隐写技术近年来也取得了显著进展。Hayes 等人[74]提出利用生成对抗网络生成高质量载体图像，并结合传统隐写算法实现信息隐藏。随后，Zhu 等人[75]提出了基于卷积神经网络的端到端信息隐藏和提取方法。受到这些研究的启发，本文提出了一种基于扰动嵌入的新型对抗样本生成方法(Perturbation Embedding based Adversarial Example, PtEm-AE)。该方法通过引入深度网络隐写技术，在不同图像上高效嵌入扰动，从而有效保障生成对抗样本的视觉质量。同时，本文还设计了两种互补的对抗损失函数，分别作用于样本层面和数据分布层面，进一步提升了攻击效果和泛化能力。

3.2 方法框架和实现细节

3.2.1 框架结构

如图 3.1 所示，本方案的训练过程分为三个阶段：

第一阶段：训练扰动嵌入网络

本阶段采用扰动嵌入网络生成对抗样本，以预训练代理模型模拟目标网络的特征提取过程。具体而言，将原始图像与高斯采样的随机噪声共同输入扰动嵌入网络，通过特征编码实现扰动信息的自适应嵌入。为协调对抗攻击效能与视觉隐蔽性，设计多维约束优化框架：基于代理模型特征空间构建对抗损失函数驱动扰动生成，引入样本对比损失扩大特征差异以增强分类决策干扰，同时结合分布差异损失提升扰动迁移性，并通过视觉隐蔽性损失严格控制对抗样本与原始图像的视觉一致性以维持视觉保真度。通过上述多目标协同优化，在确保攻击有效性的同时实现对抗样本的高隐蔽性与强迁移性。

第二阶段：通用扰动优化

在固定扰动嵌入网络参数的基础上，本阶段专注于对扰动参数进行定向优化。以随机初始化的扰动向量为起点，采用迭代优化算法搜索具有最大攻击效能的扰动参数。此过程使用投影梯度下降（Projected Gradient Descent, PGD）方法的数学框架，通过梯度上升策略逐步调整扰动参数，使其最大化代理模型的预测误差。

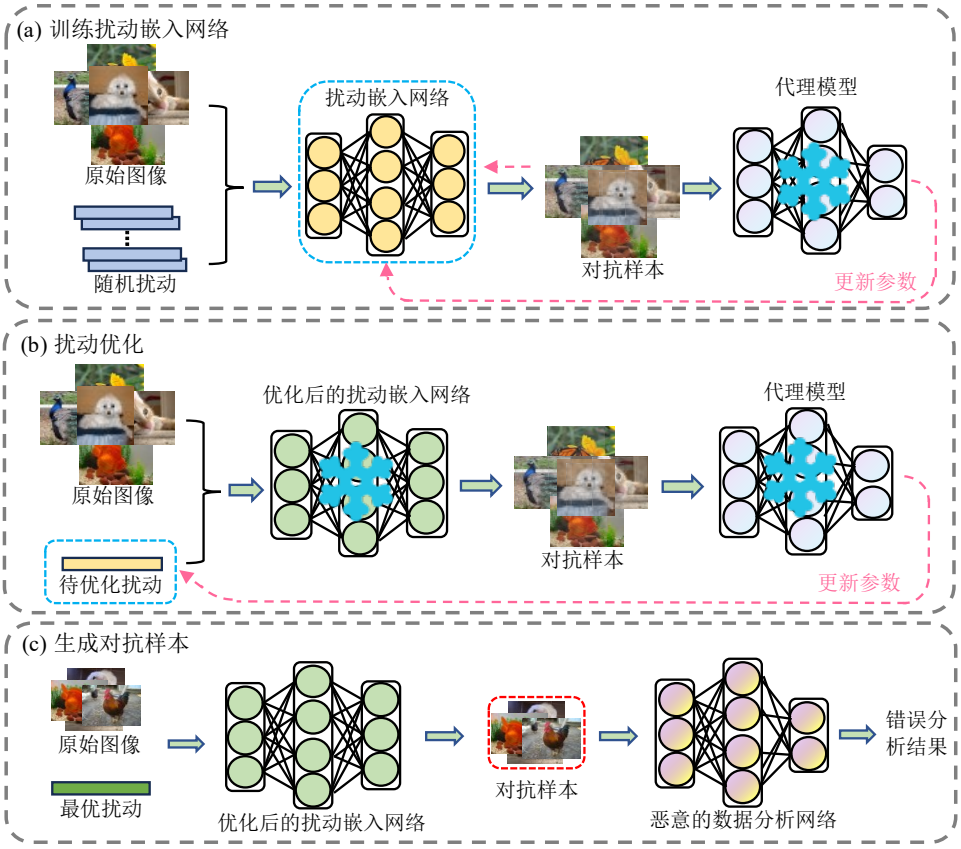


图 3.1 基于扰动嵌入的对抗样本生成方法的总体框架

Figure 3.1 Overview of Adversarial Example Generation Based on Perturbation Embedding

第三阶段：对抗样本生成

基于前两阶段训练完成的扰动嵌入网络与优化后的扰动参数，本阶段构建端到端的对抗样本生成系统。将原始数据集输入固定参数的扰动嵌入网络，结合优化所得的扰动向量，即可高效生成与原始图像一一对应的高攻击性对抗样本集合。

3.2.2 网络细节

扰动嵌入网络：

扰动嵌入网络利用隐写技术将扰动信息以不可见的方式嵌入图像中，生成对抗样本。该模块的结构主要由两种卷积块组成：普通卷积（Conv）和特殊卷积（ConvBNSiLU）。ConvBNSiLU 由卷积层（Conv）、批标准化层（Batch Normalization，BN）和 SiLU 激活函数组成，多个 ConvBNSiLU 堆叠形成 ConvBlock，用于提取图像特征。

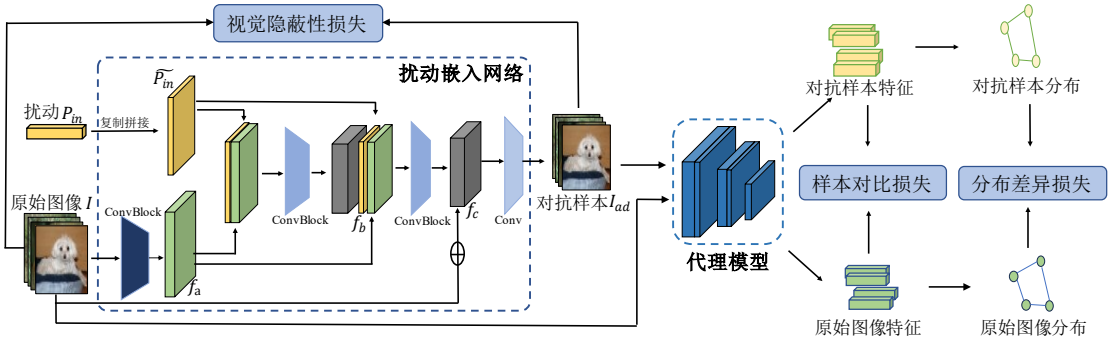


图 3.2 PtEm-AE 模型细节与损失设计

Figure 3.2 Details of the PtEm-AE Model and Loss Design

编码器的输入由原始图像I和扰动张量P_{in}组成。训练数据集表示为$D_s = \{I^i\}_{i=1}^N$，其中$I \in R^{C\times H\times W}$为原始图像，$P_{in}^L \in (-1,1)$是长度为$L$的扰动张量，元素为在区间(−1,1)内服从正态分布的随机浮点数。编码器接收原始图像I和扰动张量P_{in}，生成视觉上与原始图像不可区分的对抗样本I_{ad}。

首先，扰动张量P_{in}在空间上进行复制和拼接，使其形状与图像特征对齐，得到$\widetilde{P_{in}}$。接下来，编码器通过普通卷积从原始图像I中提取初步特征f_a，并将其与扰动张量$\widetilde{P_{in}}$在通道维度上拼接，得到复合特征并通过 ConvBlock 进一步处理，得到特征f_b。然后，特征f_a、f_b和$\widetilde{P_{in}}$再次进行融合，并通过 ConvBlock 处理得到最终特征f_c。最后，特征f_c与原始图像I进行融合，经过普通卷积生成与原始图像尺寸相同的对抗样本I_{ad}。这一过程确保了生成的对抗样本在视觉上与原始图像几乎没有差异，同时嵌入了扰动信息，从而可以有效干扰后续模型的预测。

代理模型：

为了在黑盒场景下有效生成对抗样本，PtEm-AE 假设存在一个目标网络作为代理模型，用于提取图像特征。考虑到代理模型与实际攻击目标可能存在差异，我们设计了相应的损失函数以弥补这种不确定性。具体而言，通过最大化原始图像和对抗样本特征之间的距离，并增加原始图像与对抗样本各自分布之间的差异，PtEm-AE 能够扰乱图像特征，从而赋予对抗样本攻击特性。这两种互补的损失设计确保即使在黑盒环境下，生成的对抗样本仍能有效干扰目标网络。本章实验中使用预训练的 ResNet-50 作为默认的代理模型结构。

3.2.3 损失设计

PtEm-AE 框架通过集成视觉隐蔽损失、样本对比损失与分布差异损失三项核心组件构建联合损失函数。这些损失项协同工作，确保生成的对抗样本既保持图像质量又具备强攻击性。在第一阶段的扰动嵌入过程中，三项损失共同指导网络将扰动以隐蔽方式融入原始图像。第二阶段的优化则专注攻击效果：仅使用样本对比和分布差异两项损失进行对抗性优化，同时通过数值截断技术控制扰动幅度。本框架的创新点在于重新定义隐写的验证标准——传统方法需要保证隐藏信息可被准确提取，而我们以实际攻击效果作为隐写成功的证据：当对抗样本成功欺骗模型时，即证明扰动已有效嵌入。这种设计思路摆脱了传统隐写方法的束缚，使模型能够更灵活地平衡隐蔽性和攻击力的双重需求。

视觉隐蔽性损失

为确保对抗样本I_{ad}与原始图像I的视觉一致性，避免因图像质量劣化影响实际应用，本章设计了视觉隐蔽性损失函数L_I。该函数通过约束原始图像与对抗样本的像素空间距离，构建基于均方误差的相似性度量准则，其数学表达为：

$$L_I = \frac{1}{C \times H \times W} \| I - I_{ad} \|_2^2 \tag{3.1}$$

该损失函数通过反向传播优化对抗扰动嵌入过程，最小化两样本在视觉层面的表征差异，有效保持对抗样本的视觉隐蔽性，确保其在保留对抗攻击效力的同时维持正常视觉认知功能以避免影响图像的正常使用。

对抗攻击损失

为了提升生成对抗样本的攻击效果，我们设计了对抗攻击损失L_{AD}，从样本层面和数据分布层面分别进行优化。该损失由两个子损失组成：样本对比损失和分布差异损失。

- **样本对比损失：**该损失的设计借鉴了对比学习的经典训练思路[92]，通过计算原始图像及其对应对抗样本之间的相似度矩阵，增强原始图像与对抗样本在特征空间中的距离。具体来说，对于一批原始图像的特征H及其对应的对抗样本的特征H_{ad}，我们计算其相似度矩阵$S = H \cdot {H_{ad}}^T$，其中$S_{i,j}$表示第i个原始图像与第j个对抗样本的相似度。样本对比损失通过交叉熵损

失拉大相似度矩阵对角线之外的值，从而增大原始图像与对抗样本之间的特征距离。其损失定义为：

$$L_{AD1} = -\frac{1}{N}\sum_{i=0}^{N-1} log\left(\frac{exp(S_{i,j})}{\sum_{j=0}^{N-1} exp(S_{i,j})}\right) \tag{3.2}$$

- **分布差异损失：**从数据分布的角度出发，我们设计了分布差异损失。该损失计算原始图像与对抗样本各自特征的相似度矩阵，并通过归一化将其转化为概率分布。最后，使用 KL 散度衡量这两个分布之间的差异，促进模型学习能够有效改变样本分布的扰动。具体来说，我们计算原始图像特征的相似度矩阵$S = H \cdot H^T$和对抗样本特征的相似度矩阵$S_{ad} = H_{ad} \cdot {H_{ad}}^T$，归一化后得到概率分布$P$和$P_{ad}$。分布差异损失则通过 KL 散度度量这两个分布之间的距离：

$$L_{AD2} = KL(P, P_{ad}) \tag{3.3}$$

通过在单个样本特征和整体数据分布两个层面同时施加限制，对抗攻击损失函数保证生成的对抗样本既具备更强的攻击效果，又能适应不同模型，从而全面提升对抗能力。

3.2.4 算法伪代码

PtEm-AE 的训练过程分为两步，分别是训练扰动嵌入网络和优化扰动，具体的训练细节如下表 3.1、3.2 所示：

表 3.1 扰动嵌入网络训练过程

Table 3.1 Training Process of the Perturbation Embedding Network

输入：扰动嵌入网络训练数据集 D，代理模型 F，扰动长度 L，模型学习轮次 T，学习率η_γ

输出：扰动嵌入网络(encoder)

for n=1 **in** T **do:**

I∈**D**，从正态分布中随机采样长度为 L 的扰动张量P_{in}

$I_{ad} = encoder(I, P_{in})$ #获得对抗样本

$H, H_{ad} = F(I), F(I_{ad})$ #获得各自的特征

$S = H \cdot {H_{ad}}^T$ #计算原始图像和对抗样本间的相似度矩阵

$S_{cl} = H \cdot H^T$ ， $S_{ad} = H_{ad} \cdot {H_{ad}}^T$ #计算其各自相似度矩阵

$P, P_{ad} = Normalized(S_{cl}, S_{ad})$ #由相似度计算对应的分布

$L_I = \frac{1}{C \times H \times W} \| I - I_{ad} \|_2^2$ #计算视觉隐蔽性损失

$L_{AD1} = -\frac{1}{N} \sum_{i=0}^{N-1} log\left(\frac{exp(S_{i,j})}{\sum_{j=0}^{N-1} exp(S_{i,j})}\right)$ #计算样本对比损失

$L_{AD2} = KL(P, P_{ad})$ #计算分布差异损失

$L_{total} = \alpha L_{IR} + \beta L_{AD1} + \varphi L_{AD2}$ #计算总损失

$\gamma = \gamma - \eta_\gamma \nabla_\gamma L_{total}$ #模型参数更新

end

表 3.2 扰动优化过程

Table 3.2 Perturbation Optimization Process

输入：通用扰动优化数据集 D，代理模型 F，扰动嵌入模型 encoder，扰动长度 L，扰动优化轮次 T，学习率η_P，随机初始化的扰动P

输出：通用扰动P'

for n=1 **in** T **do**

I∈D

$I_{ad} = encoder(I, P)$ #获得对抗样本

$L_{AD} = L_{AD1}(F(I), F(I_{ad}))) + L_{AD2}(F(I), F(I_{ad})))$ #计算对抗攻击损失

$P = P - \eta_P \nabla_P L_{AD}$ #扰动参数更新

$P = Clip_{-1,1}(P)$ #限制扰动范围

end

3.3 实验分析

3.3.1 实验设置

数据集：本章在 ImageNet、CIFAR-10 和 MS COCO 数据集上进行了实验验证。

- 对于 ImageNet 数据集，从中随机选择了 12 个类别（ImageNet-12），包含约 38k 张图像，用于训练扰动嵌入网络及通用扰动。
- 对于 CIFAR-10 数据集，利用其全部训练图像进行扰动和扰动嵌入网络的训练。
- 对于 MS COCO 数据集从中随机抽取了和 ImageNet-12 同等数量的图像用于验证本章方法的跨数据集有效性。

模型：为全面评估本章方法的性能，选择了多种主流网络结构作为代理模型进行实验，包括 ResNet-50(RN50)、ResNet-101(RN101)、VGG-16、VGG-19、DenseNet-121(DN121)和 DenseNet-161(DN161)。

超参数：扰动嵌入网络的默认训练轮次为 200 轮，扰动的训练轮次为 100 轮。实验平台基于 PyTorch 深度学习框架，并在 NVIDIA RTX 3090 GPU 计算环境下完成。

3.3.2 有效性验证

表 3.3 PtEm-AE 的有效性
Table 3.3 Effectiveness of PtEm-AE

方法	CIFAR-10			ImageNet-12		
	RN50	VGG16	DN121	RN50	VGG16	DN121
Random	14.03	15.62	13.85	11.53	12.97	10.76
UAP	82.6	82.52	77.97	83.51	82.13	80.55
FG-UAP	86.74	**95.17**	84.81	88.24	93.53	87.83
PtEm-AE	**87.75**	94.89	**86.14**	**88.93**	**95.62**	**89.15**

本章针对白盒攻击场景开展对抗样本生成方法的系统性评估。白盒攻击假设攻击者具备对目标模型架构与参数的完全认知权限，从而生成具有强指向性的对抗样本。为量化模型防御能力，本章采用误导成功率（Fooling Rate, FR）作为核

心指标，其定义为对抗样本成功突破模型防御的比例。实验数据显示（表 3.3），本文在主流深度学习模型中的表现显著优于 UAP、FG-UAP 等基准方法。具体而言，在 CIFAR-10 数据集上，ResNet-50、VGG-16 及 DenseNet-121 模型的 FR 值分别达到 87.75%、94.89%和 86.14%；在 ImageNet-12 数据集上，对应指标提升至 88.93%、95.62%和 89.15%。结果表明，本章方法在多个模型上均表现出显著优势，且相比现有对抗样本生成方法，能够更有效突破模型防御，提升攻击性能。此外，实验还表明，向图像中添加随机噪声并不能有效抵御对抗攻击，而本章方法通过全面的损失设计，显著提高了攻击成功率，进一步验证了其在白盒攻击场景下的优越性。

3.3.3 迁移性实验

（1）跨模型迁移性评估：

黑盒攻击是指攻击者无法访问目标模型的内部结构和参数信息，只能通过输入和输出进行推测与攻击。在这种场景下，攻击者需要通过观察模型的输入输出，推断其行为特征，从而设计有效的攻击策略。为了验证本章方法在黑盒攻击下的有效性，我们在多个模型上进行了实验，评估了不同方法的攻击成功率，并用使用粗体表示迁移性最高的方法。如下表 3.4 所示，本章方法在跨模型迁移性方面优于 UAP 和 FG-UAP。以 ResNet-50 为源模型时，本章方法在 ResNet-101、VGG-16、VGG-19、DenseNet-121 和 DenseNet-161 上的平均攻击成功率为 50.92%，高于 FG-UAP 的 48.75%和 UAP 的 42.65%。这一结果表明，本章方法生成的对抗扰动在黑盒攻击场景下具有更好的迁移性和攻击能力。综合对比不同源模型的实验，本章方法在各个模型上的迁移效果均优于 UAP 和 FG-UAP，进一步验证了其在黑盒攻击下的优越性。本章方法的优越性能得益于我们设计的全面攻击损失函数，该函数既考虑了样本层面的对比损失，又注重了数据分布层面的差异损失。这一优化设计增强了对抗样本在黑盒环境下的攻击能力，从而取得了较好的实验结果。

表 3.4 跨模型迁移性对比

Table 3.4 Comparison of Cross-Model Transferability

模型	方法	RN50	RN101	VGG16	VGG19	DN121	DN161	平均
RN50	UAP	83.51	52.54	42.24	41.22	39.1	38.16	42.65
	FG-UAP	88.24	54.2	**56.11**	53.91	40.67	38.84	48.74
	PtEm-AE	88.93	**58.1**	50.88	**54.3**	**48.38**	**45.95**	**50.92**
RN101	UAP	52.79	78.58	38.17	37.39	35.11	33.76	39.44
	FG-UAP	53.81	86.17	52.17	53.49	44.72	41.35	49.10
	PtEm-AE	**60.19**	87.57	**58.31**	**57.6**	**47.16**	**45.21**	**53.69**
VGG16	UAP	34.23	32.56	82.13	86.9	29.56	26.26	41.90
	FG-UAP	39.16	37.4	96.53	84.73	35.87	32.74	45.98
	PtEm-AE	**46.77**	**39.52**	95.62	**87.15**	**38.97**	**34.76**	**49.43**
VGG19	UAP	40.03	33.38	**87.56**	84.36	35.6	31.77	45.66
	FG-UAP	44.36	31.88	85.81	94.71	35.85	31.9	45.96
	PtEm-AE	**45.82**	**36.39**	85.14	96.07	**37.89**	**33.12**	**47.67**
DN121	UAP	46.39	43.34	56.37	54.23	80.55	49.42	49.95
	FG-UAP	44.13	44.91	57.87	59.82	87.83	59.67	53.28
	PtEm-AE	**45.1**	**47.46**	**60.17**	**61.29**	89.15	**73.96**	**57.59**
DN161	UAP	42.9	39.01	60.07	60.54	74.97	79.46	55.49
	FG-UAP	41.6	37.76	51.97	50.99	72.11	83.95	50.88
	PtEm-AE	**45.76**	**46.77**	**57.87**	**60.81**	**75.87**	82.96	**57.41**

（2）跨数据集迁移性评估：

为验证本文在跨数据集场景下的迁移攻击效能，本章选取 VGG-16、ResNet50 及 DenseNet121 三类异构深度模型构建迁移攻击实验框架。实验采用 MS COCO 作为对抗样本训练集，ImageNet-12 作为目标测试集，系统评估对抗扰动的跨域泛化能力。为确保实验设计的合理性，通过随机采样构建与 ImageNet-12 等量的 MS COCO 子集作为对抗扰动训练集，并利用 ImageNet-12 作为测试集，评估不同数据集上训练出的对抗扰动的迁移效果，以便与表 3.3 同数据分布下的实验结果进行直接对比。如表 3.5 所示，传统通用对抗扰动（UAP）方法呈现出显著的数据域偏移敏感特性，其跨数据集攻击成功率（65.65%）较同源数据场景下降最高达到了 18.7%。相比之下，FG-UAP 通过捕捉模型特征坍缩方向构建对抗扰动，

该特征空间在不同数据域间具有较高稳定性，因而获得更优的迁移性能。而本文通过两种互补的对抗损失设计，显著提高了对抗样本的泛化性。值得关注的是，FG-UAP 与本方法均利用模型中间层的深层特征表征进行扰动优化，这种基于特征空间泛化性的设计理念，相较于 UAP 单纯依赖末端分类决策的优化范式，能够更有效地突破数据分布差异带来的泛化瓶颈，从而在跨域攻击场景中展现出更强的技术鲁棒性。

表 3.5 跨数据集迁移性对比

Table 3.5 Comparison of Cross-Dataset Transferability

方法	RN50	RN101	VGG16	VGG19	DN121	DN161	平均
UAP	69.45	60.42	76.05	77.82	59.51	50.66	65.652
FG-UAP	79.24	75.7	**83.41**	**88.93**	67.38	64.07	76.422
PtEm-AE	**84.13**	**80.85**	81.94	84.73	**72.35**	**71.95**	**79.325**

3.3.4 鲁棒性实验

为了全面评估我们方法的有效性，我们进一步考察了对抗样本在不同对抗防御机制下的表现，重点验证我们的方法是否能够突破现有的经典防御手段。本章主要考虑了两类对抗防御策略：基于预处理的防御方法和对抗训练。基于预处理的防御方法通常在对抗样本输入模型之前进行特殊处理，试图消除扰动的影响。我们选择了几种经典的预处理防御方法进行实验，包括 JPEG 压缩（质量因子设置为 75）、标准像素偏移（Pixel）、核尺寸为 5×5 的高斯模糊以及不做任何处理（None）的原始图像作为对比参考。此外，我们还采用了经典的对抗训练防御方法，训练了 AdvInc-v3 模型[93]进行对比实验。如下表 3.6 所示，我们的方法在面对基于预处理的防御时展现出较强的鲁棒性，尤其在像素偏移防御下，我们生成的对抗样本几乎不受影响。尽管 JPEG 压缩对我们的攻击成功率产生了一定抑制作用，但与其他防御方法相比，我们的方法仍保持一定的优势。在对抗训练防御下，所有方法的攻击成功率均显著下降，降至约 30%。然而，值得注意的是，尽管对抗训练在某些情况下能够有效防范对抗样本，我们的方法在所有模型下仍然表现出较高的攻击成功率。综上所述，尽管面对不同的对抗防御机制，我们的方法展现出较强的适应性，并在不同模型下的表现始终优于其他现有对抗攻击方

法。

表 3.6 对不同预处理及对抗训的练鲁棒性对比

Table 3.6 Comparison of Robustness under Different Preprocessing and Adversarial Training

模型	方法	None	JPEG	Pixel	高斯	AdvInc-v3	平均
RN50	UAP	83.51	37.84	57.36	42.39	27.52	41.278
	FG-UAP	88.24	40.29	**59.57**	44.35	31.7	43.978
	PtEm-AE	**88.93**	**44.53**	59.12	**47.11**	**34.18**	**46.235**
VGG16	UAP	82.13	47.3	80.12	52.41	24.95	51.195
	FG-UAP	**96.53**	66.14	85.91	63.98	30.96	61.748
	PtEm-AE	95.62	**72.82**	**87.13**	**65.73**	**32.4**	**64.52**
DN121	UAP	80.55	50.72	63.71	48.74	28.57	47.935
	FG-UAP	87.83	54.16	61.97	50.06	30.04	49.058
	PtEm-AE	**89.15**	**56.78**	**66.17**	**50.81**	**31.97**	**51.433**

3.3.5 视觉隐蔽性评估

表 3.7 视觉质量评估

Table 3.7 Visual Quality Evaluation

方法	SSIM	PSNR	LPIPS
UAP	0.784	28.31	0.244
FG-UAP	0.851	30.49	0.213
PtEm-AE	**0.882**	**31.75**	**0.197**

本章采用 SSIM、PSNR 和 LPIPS 指标评估对抗样本的视觉质量。本章创新性地采用深度网络隐写技术实现对抗样本的嵌入生成，有别于传统扰动叠加法，通过扰动嵌入网络完成扰动自私应嵌入，确保对抗样本与原始图像在视觉感知层面保持高度一致。如下表 3.7 所示，本章方法在视觉质量上优于 UAP 和 FG-UAP。具体而言，本章方法的 SSIM、PSNR 和 LPIPS 分别为 0.882、31.75 和 0.197，均优于 FG-UAP（0.851、30.49、0.213）和 UAP（0.784、28.31、0.244）。其中，SSIM（结构相似性指标）用于衡量对抗样本与原始图像的结构相似度，PSNR（峰值信噪比）反映图像重构质量，LPIPS（感知图像相似度）则通过深度网络评估图像的感知相似度。实验结果表明，本章方法生成的对抗样本在视觉质量上具有

显著优势。

3.3.6 消融实验

（1）扰动长度的影响

为了分析扰动长度对攻击效果的影响，本章在 ImageNet12 数据集上对 ResNet-50 模型进行了不同扰动长度的攻击实验。如下图 3.3 所示，扰动长度与攻击成功率呈现显著正相关，但与图像质量呈负相关关系。值得注意的是，本章采用深度网络隐写方法将扰动嵌入原始图像中，与传统的直接叠加扰动方法相比，这种隐写技术在视觉质量上具有显著优势。实验表明：较短的扰动虽具有优异的隐蔽性，但其攻击成功率不足，主要受限于主要因为其包含的信息量有限，无法有效扰乱模型的决策边界，难以突破模型防御；而较长的扰动虽可将攻击成功率提升至 92%以上，却导致图像结构相似度骤降至 0.76 以下，显著增加被防御机制检测的风险。因此，本章选择了 64 作为默认的扰动长度，因为这一长度在攻击效果和隐写效果之间取得了良好的平衡。能够有效提高攻击成功率，同时保持较高的视觉质量，避免了短扰动攻击效果不足和长扰动隐写困难的问题。在实际应用中，选择合适的扰动长度是提升对抗样本攻击性能的关键。

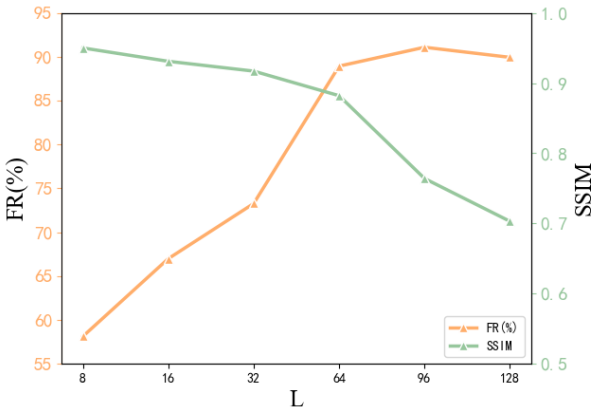


图 3.3 扰动长度对攻击效果与视觉质量的影响

Figure 3.3 Impact of Perturbation Length on Attack Effectiveness and Visual Quality

（2）双阶段优化的必要性

为验证两阶段优化框架的有效性，本研究设计了对比实验分析扰动优化机制的影响：在控制变量条件下，分别测试仅使用扰动嵌入网络嵌入随机扰动生成准对抗样本（Quasi-AE，对应图 3.4 中橙色）与嵌入经过第二阶段优化的扰动生成

的对抗样本（AE，对应图 3.4 中蓝色）的攻击性能。如图 3.4 所示，尽管仅使用第一阶段训练时嵌入的随机噪声，仍能实现一定的攻击效果（平均攻击成功率 83 左右），这是因为第一阶段的训练过程中已经引入了相关的对抗攻击损失，增强了扰动嵌入网络的攻击能力。而使用经过第二阶段梯度优化的扰动所生成样本将攻击成功率显著提升至 87 左右，表明优化后的扰动能更有效地突破模型防御，进一步提高了攻击性能。

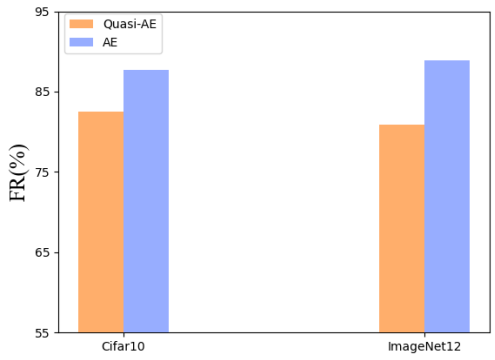


图 3.4 扰动优化的必要性
Figure 3.4 Necessity of Perturbation Optimization

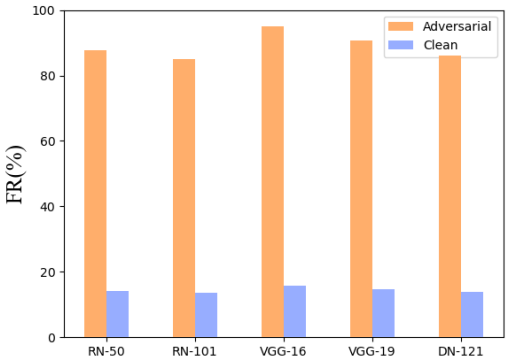


图 3.5 代理模型选择的影响
Figure 3.5 Impact of Different Surrogate Model

（3）代理模型选择的影响

为评估代理模型结构对 PtEm-AE 的影响，我们在 CIFAR-10 数据集的白盒攻击场景下开展多种模型结构的对比实验。选取 ResNet-50(RN-50)、ResNet-101(RN-101)、VGG-16、VGG-19 及 DenseNet121(DN-121)三大主流模型作为测试对象，图 3.5 显示 PtEm-AE 在不同模型上均保持稳定攻击效果。实验结果验证了该方法的模型无关性，其中 ResNet-50 凭借训练效率与内存占用量的综合优势，被确立为默认代理模型。基准实验表明，不同模型在原始图像上的原始性能差异未对攻击效果评估产生显著干扰。

（4）各项对抗损失的作用

为验证本章所提出对抗损失的有效性，我们进一步进行了不同损失设置下的消融实验。具体而言，在 CIFAR-10 和 ImageNet-12 数据集上，针对 ResNet-50 模型，分别测试了仅使用分布差异损失（AD2）、仅使用样本对比损失（AD1）以及两者结合使用（OUR）的对抗攻击效果。如下图 3.6 所示，仅使用样本对比损失虽然能够实现一定的攻击效果，但表现较为有限；而仅使用分布差异损失的攻击效果则更为局限，无法显著提升攻击成功率。然而，当两种损失结合使用时，

攻击效果显著提升，表明样本层面和数据分布层面的损失具有互补性，能够更有效地增强对抗攻击性能。这一结果充分验证了本章提出的对抗损失设计的合理性和必要性。

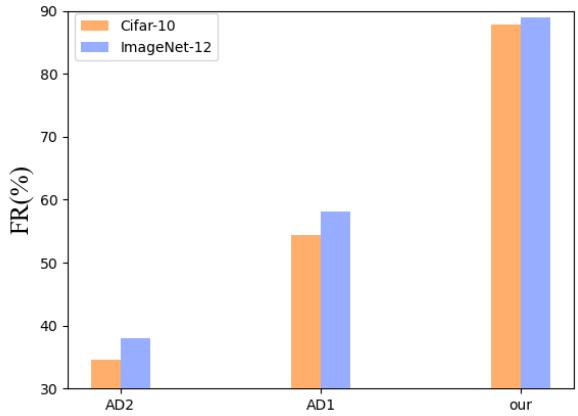


图 3.6 两种对抗攻击损失的对攻击效果的影响
Figure 3.6 Impact of Two Adversarial Attack Losses on Attack Effectiveness

3.3.7 对抗样本的扰动

图 3.7 对抗样本及其扰动
Figure 3.7 Adversarial examples and their noise

如图 3.7 所示，本章通过可视化分析揭示了 PtEm-AE 生成的对抗扰动特性。基于训练完成的扰动嵌入网络，我们选取 5 类不同的原始图像，生成对抗样本后通过图像相减提取嵌入噪声。实验表明，虽然采用通用扰动策略，但扰动嵌入网络会依据输入图像的纹理特征，自适应性选择最优区域嵌入扰动。这种动态嵌入机制使得噪声分布与图像局部结构呈现高度相关性。通过深度网络隐写技术实现的像素级自适应调节，使得对抗扰动与原始图像形成视觉一致性，相较于传统固定模式的对抗噪声，PtEm-AE 的扰动既保持了跨样本的通用攻击能力，又提高了生成的对抗样本的视觉质量。这种双重特性确保对抗样本既能有效破坏模型推理，又能规避数据筛选机制的检测，为实际应用提供了关键保障。

3.4 本章小结

本章提出基于扰动嵌入的对抗样本生成框架 PtEm-AE，通过双阶段优化机制和全面的损失设计保证扰动的高效嵌入。第一阶段采用扰动嵌入网络与预训练代理模型的协同优化，结合视觉隐蔽性损失与对抗攻击损失生成初始对抗样本；第二阶段通过投影梯度下降算法优化扰动参数，最大化模型预测误差；最后构建端到端生成流程，实现对抗样本的批量生产。此外，本文创新性地以模型攻击效果替代传统隐写解码验证标准，将优化目标聚焦于隐蔽性与攻击性的协同提升。实验结果表明，本章方法在多个数据集和模型上均获得优良表现，且在视觉质量和攻击成功率上取得了显著提升。该方法为对抗样本生成提供了一种新的思路，具有广泛的应用前景。

第四章　基于扰动嵌入的不可学样本生成方法

4.1 引言

近年来，自监督学习凭借强大的特征表征能力，在人脸识别、目标检测与姿态估计等诸多领域展现出突破性进展。然而，这种技术突破高度仰赖海量训练数据，导致互联网数据采集行为日趋普遍，由此引发的未授权数据使用问题已对隐私保护与版权安全构成严峻挑战。为应对这一困境，学术界提出了可用性攻击的防御思路，通过干扰未授权模型的训练过程实现数据保护。Fowl 团队[50]开创的无差别中毒攻击率先采用对抗扰动注入策略来破坏分类模型的性能表现。随后，Huang 等人[35]提出了“不可学样本”（Unlearnable Examples）技术通过设计人眼不可察觉的精细扰动，有效抑制深度学习模型的特征提取能力，使得在干净测试集上的性能显著下降。值得关注的是，这些针对监督学习设计的防护手段仍存在漏洞——非法数据采集者可通过对比学习算法对受污染数据进行预训练并获取有效表征。为此，He 等人[57]将不可学习样本扩展至对比学习领域，提出对比数据投毒（CP）方法以增强自监督学习场景的防护能力。与此同时，Ren 等人[56]提出的可迁移不可学习样本（TUE）通过深度挖掘数据分布特性，构造出跨模型、跨数据集通用的防御样本。尽管现有方法展现出一定潜力，但仍面临双重挑战：其一，防护效能的局限性导致生成样本难以提供强有力的数据保护；其二，样本视觉质量的显著下降严重制约其实际应用价值。为突破这些瓶颈，本文创新性地提出（Perturbation Embedding based Unlearnable Example, PtEm-UE），区别于传统直接叠加扰动的范式，我们构建经过专门训练的扰动嵌入网络，将优化后的扰动隐式嵌入原始图像。该方法采用两阶段优化框架——第一阶段专注于学习扰动嵌入机制，确保扰动可靠嵌入的同时误导代理模型关注扰动特征；第二阶段着重进行扰动优化，持续强化其不可学习特性。经过完整训练流程处理的不可学样本既能有效抵御非法模型的训练使用，又能在人眼视觉感知层面保持优良的视觉质量。

4.2 方法框架与实现细节

4.2.1 框架结构

在本章中，我们将概述 PtEm-UE 框架的设计与工作流程。图 4.1 中展示了这一框架的各个训练阶段，首先是第一阶段，涉及扰动嵌入网络和代理模型之间的交替优化。此阶段的核心任务是训练扰动嵌入网络，使其能够将扰动有效嵌入到原始图像中。具体而言，在这一阶段，扰动嵌入网络的目标是学习如何将随机扰动嵌入到数据中，并与代理模型交替优化，从而确保生成的对抗样本具有基本的不可学性。需要特别指出的是，在此阶段，由于扰动本身尚未经过优化，随机扰动被用作其替代品。此外，通过将对比学习损失整合到训练目标中，扰动嵌入模型能够有效误导代理模型，使其将注意力集中在这些随机扰动上，以提高不可学样本的不可学性。

接下来的第二阶段，如图 4.1(b)所示，涉及扰动和代理模型之间的交替优化。在这一阶段，第一阶段学习到的扰动嵌入网络被固定，而扰动本身则进入优化过程，目的是最大限度地提升其不可学习性。具体来说，本阶段的目标是找到一个能够使对比学习损失和聚类紧疏损失最小化的扰动，从而增强生成的不可学样本的不可学性。第二阶段的优化过程借鉴了 PGD（Projected Gradient Descent）思想，通过迭代优化扰动，最大化其不可学性。这一优化过程确保扰动能有效地干扰模型的训练，使得生成的不可学样本中嵌入的扰动在特征空间中逐渐主导正常样本的特征，最终确保其具备足够的不可学性，从而有效地防止模型的学习和训练。

在这两个阶段的优化完成后，嵌入了优化扰动的原始图像最终转变为不可学样本。这些不可学样本随后可以被非法数据收集者访问。数据收集者可以将这些不可学样本用作整个训练集，或者将其与其他收集到的正常样本混合，并利用对比学习或传统的监督学习方法对模型进行训练。

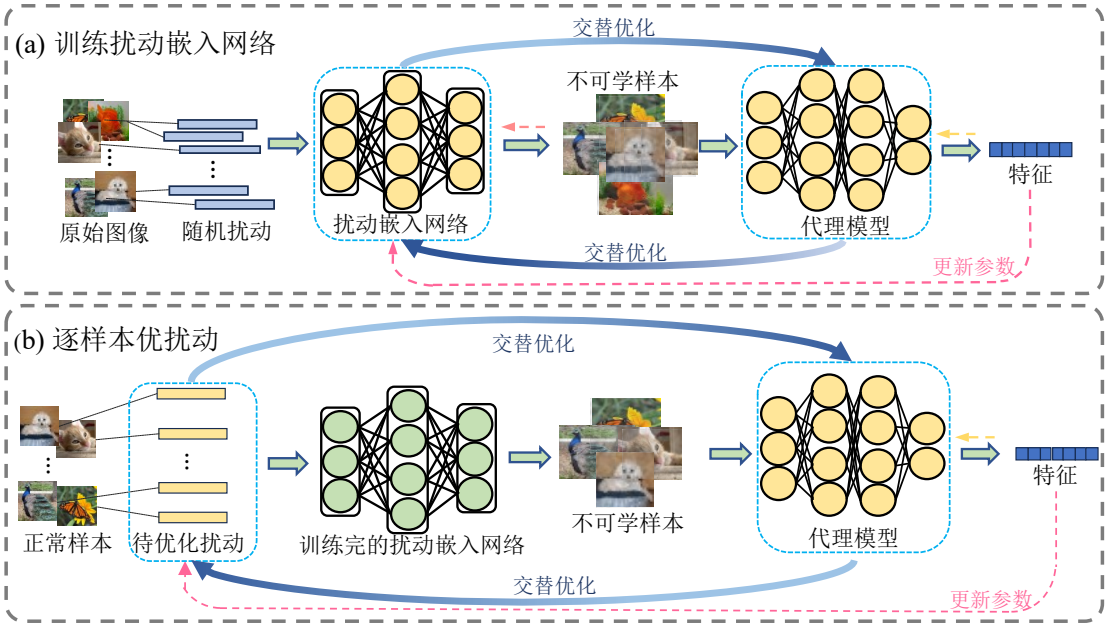


图 4.1 基于扰动嵌入的不可学样本生成方法的总体框架

Figure 4.1 Overview of Unlearnable Example Generation Based on Perturbation Embedding

4.2.2 网络细节

如图 4.2 所示，PtEm-UE 框架包含扰动嵌入网络与代理模型两大核心模块，其与第三章所述对抗样本生成框架 PtEm-AE 的差异源于**防御场景与优化范式**差异：首先，PtEm-AE 采用预训练的代理模型模拟模型部署阶段的恶意数据分析场景，通过干扰模型推理过程阻断敏感信息提取；而 PtEm-UE 则构建**未初始化的代理模型与扰动嵌入网络的交替优化机制**，通过动态更新代理模型参数模拟模型训练初期的特征学习过程，保证生成的不可学样本无法用于模型训练。其次，二者在**损失函数设计上存在本质分异**：PtEm-AE 通过数据分布差异损失增大对抗样本与原始图像的全局分布一致性以提升其泛化性；PtEm-UE 则采用样本级对比损失最大化扰动对局部特征空间的污染强度，通过对比学习强制模型关注扰动而非真实语义。再者，**扰动生成机制差异**：PtEm-AE 针对固定模型参数的静态攻击场景，采用全局扰动实现批量样本攻击；而 PtEm-UE 需阻断复杂训练动态下的特征学习过程，故为**每个样本生成针对性扰动**，通过扰动与代理模型参数的迭代优化提升其不可学性。二者差异本质源于任务复杂度：前者为单目标优化，后者为动态优化。最后，**泛化性需求差异**：PtEm-AE 需确保生成的对抗扰动对跨数

据分布具备攻击效力；而 PtEm-UE 的扰动优化完全面向训练集不可学性，其有效性通过模型在扰动数据上的训练失败直接验证，无需考虑测试集泛化，这也直接导致了二者的评估指标有本质区别。上述差异从防御阶段、代理模型状态与扰动作用粒度三个维度界定了两类技术的边界。具体网络细节如下：

扰动嵌入网络：扰动嵌入网络负责将扰动以隐写的方式嵌入到原始图像中。它主要由三种模块组成：特征提取模块 ConvBlock、空间注意力融合模块 Fuse、普通卷积 Conv。其中特征提取模块 ConvBlock 模块由多个特征提取层（ConvBNSiLU）堆叠而成，每个特征提取层集成了卷积层（Conv）、批归一化层（Batch Normalization，BN）、SiLU 激活函数层，是基本的特征提取单元。Fuse 模块通过通道注意力机制自适应地调整输入特征图中不同通道的重要性。该模块首先采用全局平均池化和最大池化操作提取输入特征图的全局信息，分别捕捉通道的平均值和最大值。然后，通过两个 1x1 卷积层对池化结果进行通道数压缩与恢复，并使用 ReLU 激活函数引入非线性。接着，将两个池化操作的输出相加，生成通道注意力特征图，并通过 Sigmoid 激活函数将其归一化为 0 到 1 之间的权重。最后，利用这些权重对原始输入特征图进行加权，从而强化重要通道的特征。

形式上，编码器将原始图像$I \in R^{C\times H\times W}$和随机扰动$P_{rnd}^{L}$（具有随机值的长度为$L$的张量）作为输入。原始图像$I$首先由 ConvBlock 处理，生成一个 64 维的中间特征表示。随机扰动P_{rnd}^{L}被复制和扩展，以匹配中间特征的维数，然后与之连接。随后融合特征经过特征提取看 ConvBlock 的压缩、与 Fuse 模块的注意力增强后与原来的图像特征和扰动特征再次融合并经过相同的处理得到最终的融合特征表示，并与原始图像融合并经过卷积层处理得到最终的嵌入了扰动的不可学样本。具体流程见下图 4.2。

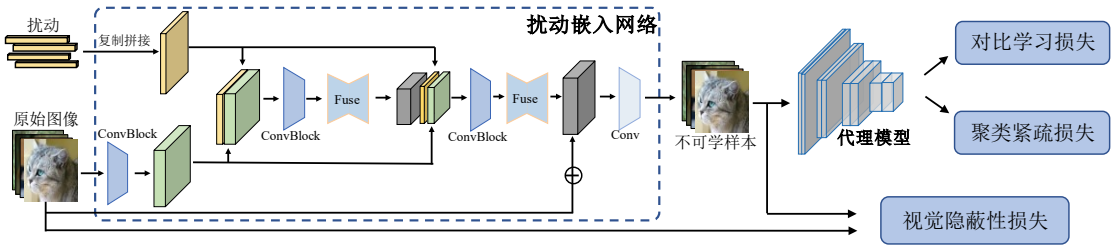


图 4.2 PtEm-UE 模型细节与损失设计

Figure 4.2 Details of the PtEm-UE Model and Loss Design

代理模型：PtEm-UE 利用深度网络隐写技术将扰动嵌入原始图像生成不可学样本，其核心在于迫使模型训练时持续关注扰动而忽略真实特征：由于原始图像不含扰动且使用不可学样本训练的模型无法提取有效特征，导致下游任务性能显著下降，从而实现数据保护。为实现扰动与模型注意力的动态博弈，我们采用未经训练的 ResNet-18 作为默认的代理模型，通过交替优化扰动嵌入网络与代理模型模拟真实训练场景。我们交替优化代理模型和扰动嵌入网络，使得扰动嵌入网络学会如何嵌入扰动并保持扰动对代理模型具有吸引力。同样的在扰动优化的过程中，我们固定扰动嵌入网络而交替优化扰动本身和代理模型以寻找最优的扰动。

4.2.3 损失设置

PtEm-UE 框架通过三重损失机制实现不可学样本的生成优化，包含视觉隐蔽性损失、对比学习损失和聚类紧疏损失，三者协同作用平衡生成的不可学样本的隐蔽性与不可学效果。在扰动嵌入网络的优化过程中，三项损失共同调控网络参数，确保扰动以不可察觉方式融入原始图像。在扰动的优化过程中，仅采用后两项损失联合优化并引入扰动幅度截断控制，在增强不可学性的同时维持视觉合理性。本方案突破性地重构了隐写的验证范式：不同于传统方法依赖信息提取准确率，我们以模型训练失效作为隐写有效性判据——当不可学样本成功破坏模型训练过程，即证明扰动嵌入达到预期效果。这种逆向验证机制摆脱了传统隐写任务的固有约束，实现了隐蔽性与不可学效果的动态平衡。

视觉隐蔽性损失：为保障生成的不可学样本I_{UE}与原始图像I的感知一致性，本章构建了像素级约束函数L_I。通过建立基于均方误差的视觉差异评估，该损失项数学定义为：

$$L_I = \frac{1}{C \times H \times W} \| I - I_{UE} \|_2^2 \tag{4.1}$$

其中C、H、W分别表示图像通道数、高度和宽度维度。该约束通过梯度反传优化扰动嵌入过程，在像素空间最小化原始图像与不可学样本的欧氏距离，有效防止因视觉失真影响实际应用的可靠性，确保嵌入的扰动保持高度隐蔽性。

对比学习损失：我们希望不可学样本中的扰动特征占据主导以盖过原始图像的有意义特征，为了达成这一目标，我们使用生成的不可学样本模拟训练代理模

型，并根据不同的对比学习算法计算对比损失，朝着最小化该损失的方向优化网络和扰动确保不可学样本的有效性。本章采用了四种对比学习算法（SimCLR，MoCo，BYOL, SimSiam）的对比损失。统一地表示为:

$$L_{CL}(I_{UE}) = CL(F(I_{UE})) \tag{4.2}$$

其中CL指代不同的对比学习算法，F表示代理模型，这个损失旨在模拟真实的不可学样本用于对比学习训练的场景。

聚类紧疏损失：本章提出基于类别感知的聚类紧疏损失，通过引入监督信息优化特征空间拓扑结构，其核心由类内离散度D_{intra}与类间排斥度D_{inter}共同构成。对于给定不可学样本I_{UE}及其类别标签Y，定义特征空间中第i类样本的类中心为$c_i = \frac{1}{N_i}\sum_{k=1}^{N_i} f_k^{(i)}$，$f_k^{(i)}$为第$i$类第$k$个样本的特征向量，$N_i$为第$i$类样本总数。类内离散度描述同类样本与类中心的分布紧密度$D_{intra}^{(i)} = \frac{1}{N_i}\sum_{k=1}^{N_i}\left\| f_k^{(i)} - c_i \right\|_2^2$，类间排斥度则通过异类中心距离量化特征可区分性$D_{inter}^{(i,j)} = \left\| c_i - c_j \right\|_2^2$。最终构建的损失函数为:

$$L_{cluster}(I_{UE}, Y) = \sum_{i \neq j} \frac{D_{intra}^{(i)} + D_{intra}^{(j)}}{D_{inter}^{(i,j)}} \tag{4.3}$$

该设计通过最小化同类特征离散度与最大化异类中心间距的联合优化，迫使模型在关注扰动模式时破坏原始特征判别性，从而驱动扰动在混淆模型特征提取能力的同时保持对监督信号的敏感性。

4.2.4 算法伪代码

PtEm-UE 的训练过程分为两个过程，分别是训练扰动嵌入网络和逐样本优化扰动，具体的训练细节如下表所示：

表 4.1 扰动嵌入网络训练过程

Table 4.1 Training Process of the Perturbation Embedding Network

输入：扰动嵌入网络训练数据集 D，扰动嵌入网络 E，代理模型 F，模型学习轮次 T，批次大小T_E, T_F，学习率η_E, η_F

输出：训练完备的扰动嵌入网络

for n=1 **in** T **do:**

    #更新扰动嵌入网络参数，冻结代理模型参数

    for index **in** T_E **do:**

        从训练集中采样一批数据I，从正态分布中采样随机扰动P_{in}

        $I_{UE} = E(I,\ P_{in})$ #将扰动嵌入正常样本

        $L_I = \sum_{I_{NE} \in D}(\left\|\varphi_j(I_{NE}), \varphi_j(I_{UE})\right\|_2)$ #计算隐蔽性损失

        $L_{CL} = CL(F(I_{UE}))$ #计算对比学习损失

        $L_{cluster}(I_{UE},\ Y) = \sum_{i \neq j} \frac{D_{intra}^{(i)} + D_{intra}^{(j)}}{D_{inter}^{(i,j)}}$ #计算聚类紧疏损失

        $L_{total} = \alpha L_I + \beta L_{CL} + \gamma L_{cluster}$

        $\theta_E \leftarrow \theta_E - \eta_E \nabla_{\theta_E} L_{total}$

    end

    #冻结扰动嵌入网络参数，更新代理模型参数

    for index **in** T_E **do:**

        从生成的不可学样本采样一批数据I_{UE}

        $\theta_F \leftarrow \theta_F - \eta_F \nabla_{\theta_F} CL(I_{UE})$

    end

end

表 4.2 扰动优化过程

Table 4.2 Training Process of the Perturbation

输入：扰动优化数据集 D，训练完备的扰动嵌入网络 E，代理模型 F，随机扰动比特长度 L，模型学习轮次 T，批次大小T_P, T_F，学习率η_P, η_F

输出：最优扰动 P

for n=1 **in** T **do:**

    #更新扰动，冻结代理模型参数

    for index **in** T_E **do:**

        从训练集中采样一批数据I

        $I_{UE} = E(I,\ P_{in})$ #将扰动嵌入正常样本

        $L_{CL} = CL(F(I_{UE}))$ #计算对比学习损失

        $L_{cluster}(I_{UE},\ Y) = \sum_{i \neq j} \frac{D_{intra}^{(i)} + D_{intra}^{(j)}}{D_{inter}^{(i,j)}}$ #计算聚类紧疏损失

        $L_{total} = \alpha L_{CL} + \beta L_{cluster}$

        $P \leftarrow P - \eta_P \nabla_{\theta_P} L_{total}$

    end

    #冻结扰动参数，更新代理模型参数

    for index **in** T_E **do:**

        从生成的不可学样本采样一批数据I_{UE}

        $\theta_F \leftarrow \theta_F - \eta_F \nabla_{\theta_F} CL(I_{UE})$

    end

end

4.3 实验分析

4.3.1 实验设置

数据集：在实验中，我们使用了两个广泛使用的数据集：ImageNet 和 CIFAR-10。对于 ImageNet，我们随机选择了 12 个类别，包含约 38,000 张图像，称为 ImageNet-12。对于 CIFAR-10，我们使用整个数据集。

模型与算法：为了全面评估本文方法对不同代理模型和训练算法的有效性，我们使用四种对比学习算法进行了实验：SimCLR、MoCo v2、BYOL 和 SimSiam。

此外，我们还测试了三种不同的模型架构：ResNet-18(RN18)、ResNet-34(RN34)、DenseNet121(DN121)和 VGG-16。实验中选定 ResNet-18 为代理模型和数据收集者使用的模型的默认架构。

超参数：所有实验均使用 PyTorch 实现，并在单个 RTX3090 GPU 上进行实验。扰动优化 100 轮，而代理模型模型训练 200 轮。我们将 CIFAR-10 和 ImageNet-12 的批量大小分别设置为 512 和 32，学习率为 0.001。默认情况下数据收集者使用的数据集中全部都是不可学样本，扰动长度设置为 64。

4.3.2 有效性验证

表 4.3 PtEm-UE 的有效性

Table 4.3 Effectiveness of PtEm-UE

数据集	方法	自监督学习				监督学习
		SimCLR	MoCo v2	BYOL	SimSiam	
CIFAR-10	Clean	91.39	91.54	92.67	90.67	94.76
	Gaussian	90.28	90.05	91.21	88.13	91.55
	TUE	48.08	51.63	55.70	70.53	**10.61**
	CP	**38.65**	47.04	51.47	48.62	94.53
	PtEm-UE	40.05	**46.07**	**48.05**	**46.81**	29.07
ImageNet-12	Clean	85.17	86.98	86.38	85.82	89.62
	Gaussian	83.65	84.96	84.98	83.91	87.17
	TUE	47.06	48.21	50.13	55.97	**9.94**
	CP	45.87	46.87	48.67	50.71	74.71
	PtEm-UE	**44.60**	**46.55**	**47.38**	**48.62**	17.65

本实验评估了 PtEm-UE 生成的不可学样本的有效性，结果如表 4.3 所示。本章假设实验中代理模型的结构、训练算法、训练数据集与真实数据收集者保持一致，以确保评估的合理性。使用代理模型在不可学样本上训练后，在测试集上的分类准确率（ACC）作为衡量指标。本章选取 TUE 和 CP 作为主要对比方法，因为它们代表了对比学习领域不可学样本的两种主要技术路线：TUE 通过分类可分性判别增强扰动的可转移性，从而保护不同算法下的数据；而 CP 则实现了针对对比学习框架的无差别中毒攻击。这两种方法均在该领域取得了显著的成功并

得到了广泛关注。本章还展示了 TUE、CP 以及两种基线方法（原始图像和添加高斯噪声的图像）的结果。标为“Clean”和“Gaussian”的行分别表示正常样本和高斯噪声样本的 ACC 值，较低的 ACC 值代表更好的保护性能，表格中每种情况的最低 ACC 值已用粗体标注。实验结果表明，大多数情况下 PtEm-UE 的保护性能优于其他方法。值得注意的是，TUE 在自监督学习下的表现不如 PtEm-UE 和 CP，进一步凸显了我们方法的优势。尽管在监督学习中，TUE 方法的 ACC 值较 PtEm-UE 更低（大约 10%），但相对于 CP 的 80%左右，PtEm-UE 的 20%依然具有显著优势。事实上，20%的 ACC 值在真实场景中足以满足保护需求，并远优于 CP。尤其是 20%和 10%这两个准确率，考虑到实际应用中的保护需求，已经能够提供充分的防护。最后，实验还表明，高斯噪声样本的 ACC 值几乎与正常样本相同，表明高斯噪声并不能有效防止数据被学习，进一步证明了 PtEm-UE 在不可学样本保护中的有效性。

4.3.3 迁移性实验

（1）跨训练算法有效性：

实验验证了 PtEm-UE 生成不可学样本的防护效能（表 4.4）。基于代理模型与真实数据收集者架构、训练流程对齐的假设，我们以测试集分类准确率(ACC)作为评估指标。对比 TUE 和 CP 两种主流方案：TUE 通过类可分性增强扰动可迁移性，CP 则实现对比学习的无差别攻击。实验数据显示，PtEm-UE 在多数场景下取得最优保护效果（粗体标注）。在自监督学习中，TUE 的 ACC 值落后于 PtEm-UE 与 CP；监督学习下 PtEm-UE 的 20%AC 值虽高于 TUE 的 10%，但显著优于 CP 的 80%。实际应用中，20%的准确率已具备充分防护能力，而高斯噪声样本的 ACC 值与正常样本持平，证明其无法有效阻断模型学习。综上所述，PtEm-UE 在迁移性测试中表现出色，无论是在不同算法的对比学习任务下，还是在保护效果方面，相较于现有方法，PtEm-UE 都展现了强大的优势，完全能够满足真实场景中的保护需求。

表 4.4 跨训练算法迁移性对比

Table 4.4 Comparison of Cross-Training Algorithm Transferability

方法	算法	无监督学习				监督学习
		SimCLR	MoCo v2	BYOL	SimSiam	
Clean	-	85.17	86.98	86.38	85.82	89.62
TUE	SimCLR	47.06	71.07	75.39	59.69	**9.94**
	MoCo v2	70.42	48.21	72.18	61.06	**10.55**
	BYOL	68.95	71.82	50.13	59.81	**11.04**
	SimSiam	71.62	69.83	73.17	55.97	**9.83**
	AVG	70.33	70.91	73.58	60.18	**10.34**
CP	SimCLR	45.87	69.92	73.41	**48.22**	74.71
	MoCo v2	69.97	46.87	**71.68**	53.31	74.27
	BYOL	**64.02**	62.97	48.67	54.94	73.4
	SimSiam	66.37	67.09	69.23	50.71	71.34
	AVG	66.79	66.66	71.44	52.16	73.43
PtEm-UE	SimCLR	44.60	**67.82**	**70.04**	50.75	17.65
	MoCo v2	**68.97**	46.55	72.34	**49.78**	17.72
	BYOL	64.14	**60.84**	47.38	**49.94**	16.3
	SimSiam	**63.82**	**65.24**	**67.03**	48.62	14.77
	AVG	**65.64**	**64.63**	**69.8**	**50.16**	16.61

（2）跨模型有效性实验：

本实验进一步评估了 PtEm-UE 生成的不可学样本在不同模型架构间的迁移性，结果参见表 4.5。类似于跨算法的迁移性，我们探讨当数据利用者采用的模型架构与生成不可学样本时的采用的代理模型架构不同时，这些不可学样本是否仍然有效，即跨模型架构迁移性。我们在 CIFAR-10 数据集上使用 SimCLR 算法和 ResNet-18 模型生成不可学样本，随后评估不同模型架构（包括 ResNet-18(RN-18）、ResNet-34(RN-34)、DenseNet-121(DN-121)和 VGG-16）在 CIFAR-10 数据集下游分类任务中的表现，并与正常样本及 CP 和 TUE 方法进行全面对比。最低的 ACC 值以粗体标注。实验结果表明，除 DN-121 外，PtEm-UE 在不同模型架构间展现了良好的迁移性，并优于 CP。

表 4.5 跨模型迁移性评估
Table 4.5 Evaluation of Cross-Model Transferability

方法	RN-18	RN-34	DN-121	VGG-16
Clean	91.39	91.42	92.17	90.53
TUE	48.08	50.31	48.26	39.15
CP	**38.65**	49.49	**44.34**	37.6
PtEm-UE	40.05	**47.75**	46.05	**33.91**

4.3.4 视觉隐蔽性评估

本实验采用了三个常用指标（PSNR、SSIM 和 LPIPS）评估不可学样本的视觉质量。这些指标综合评价了不可学样本的视觉保真度，衡量了原始图像与扰动图像在像素级和感知级的差异。现有方法（如 TUE 和 CP）通过优化随机扰动并直接添加到原图像像素中，而 PtEm-UE 则利用深度网络隐写技术将扰动无缝融入数据中。这种方式确保扰动与图像内容融合，而非简单叠加，从而显著提升视觉质量。如表 4.6 所示，PtEm-UE 的 PSNR（31.87）和 SSIM（0.842）均最高，LPIPS（0.194）最低，明显优于 TUE 和 CP。这些指标均是通过从 ImageNet-12 每个类别随机抽取 10 张图像计算的平均值，保证了评估的客观性，充分证明了 PtEm-UE 在有效保护数据同时可保持较高的视觉质量。

表 4.6 视觉质量评估
Table 4.6 Visual Quality Evaluation

方法	SSIM	PSNR	LPIPS
TUE	0.814	30.46	0.325
CP	0.741	29.02	0.386
PtEm-UE	**0.842**	**31.87**	**0.194**

4.3.5 对预处理的鲁棒性

不可学样本的真实应用过程中，数据收集者可能会对收集到的数据进行一些预处理以增强训练出的模型的鲁棒性。在这个实验中，我们使用了多种预处理方法对 CIFAR-10 数据集上生成的不可学样本进行实验，以评估了它们是否能忽略

这些处理方法保持不可学特性。实验中使用了几种常见的技术：对于随机噪声，控制噪声的标准差为 8/255 或 64/255；对于高斯平滑，调节高斯核的大小，分别使用 3×3 或 15×15 的核；对于 Cutout 方法，采用 16×16 的正方形区域进行切割。针对矩阵补全方法，使用了像素丢弃概率为 0.25，并采用了 USVT 算法[94]重建缺失的像素，奇异值被截断 50%。各种预处理的视觉效果如下图 4.3 所示，实验结果如上表 4.7 所示，多数场景下我们的方法取得了最好的效果，而所有的方法对矩阵补全操作几乎都失效了，这也一定程度上暴露了不可学样本的局限性。

表 4.7 对不同预处理方法的鲁棒性对比

Table 4.7 Comparison of Robustness under Different Preprocessing Methods

预处理	Clean	TUE	CP	PtEm-UE
无处理	91.39	48.08	**38.65**	40.05
随机噪声 8/255	90.49	56.38	55.31	**55.2**
随机噪声 64/255	88.97	77.51	74.49	**72.43**
高斯平滑 k=3	91.21	50.47	49.72	**48.04**
高斯平滑 k=15	89.86	59.29	57.45	**55.39**
随机裁剪	92.17	**50.11**	50.78	51.25
矩阵补全	88.75	87.2	85.92	**84.57**

4.3.6 消融实验

（1）扰动长度的影响

本实验探讨了扰动长度对 PtEm-UE 性能的影响，将扰动长度范围设置为 8 至 128，其余参数保持不变。结果如图 4.4 所示，当扰动长度增加到 64 时，ACC 值明显下降，说明较长扰动提高了样本的不可学习性。然而，进一步增加扰动长度至 128 后，ACC 值反而升高，表明更长的扰动对扰动嵌入网络的嵌入能力提出了更高的挑战。这一趋势在 CIFAR-10 和 ImageNet-12 数据集中均得到验证，提示了扰动长度与嵌入效果之间存在一定的权衡关系。基于以上观察，本章选取了长度为 64 的扰动用于所有实验。

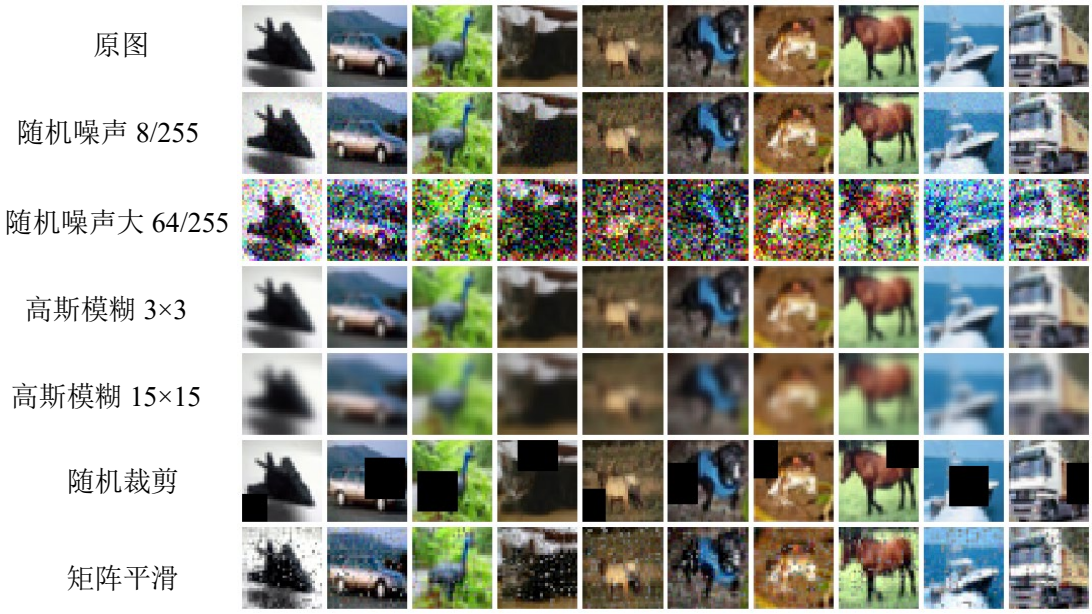


图 4.3 Cifar-10 上不同预处理的效果展示
Figure 4.3 Effect of Different Preprocessing Methods on CIFAR-10

（2）代理模型度选择

为评估模型结构对 PtEm-UE 防护效能的影响，本章在 SimCLR 自监督框架下构建跨模型结构对比实验（CIFAR-10 数据集）。选取 ResNet-18(RN-18)、ResNet-34(RN-34)、DenseNet121(DN-121)及 VGG-16 四种典型模型进行测试，图 4.5 显示 PtEm-UE 的防护效能在不同模型中保持稳定波动范围。通过权衡模型复杂度与防护性能，最终确定 ResNet-18 为默认代理模型。基准实验表明，不同架构在干净数据集上的原生性能差异未对防护效能评估产生显著偏差，印证了方法的架构鲁棒性。这种设计选择既保证了对抗扰动生成效率，又通过适度的模型容量避免了过拟合风险。

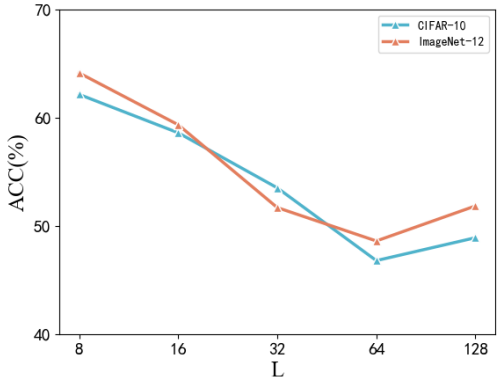


图 4.4 不同扰动长度影响
Figure 4.4 Impact of Different Perturbation Lengths

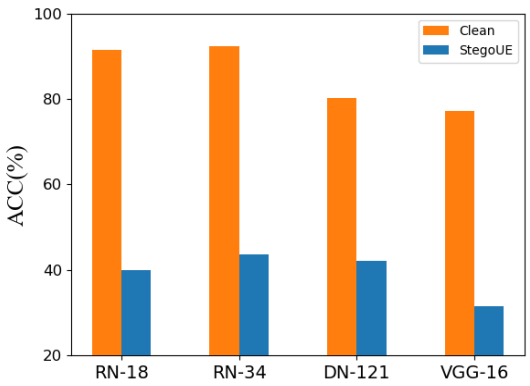


图 4.5 不同代理模型的影响
Figure 4.5 Impact of Different Surrogate Model

（3）代理模型参与训练的作用

对抗样本的生成通常只需进行单层优化，即仅通过优化输入扰动以导致模型产生误分类。然而，不可学样本的生成则涉及双层优化结构：内层优化负责生成输入扰动及相应的扰动嵌入网络，外层优化则专注于调整代理模型本身参数。外层优化的作用尤其关键，它确保生成的扰动不仅仅具有干扰性，而且能够有效破坏模型的训练过程，使模型优先学习到扰动特征而非数据固有的真实特征，从而导致训练数据的不可学性。

如下图 4.6 所示，我们在训练的不同阶段固定代理模型，实验结果进一步验证了外层优化对于不可学样本生成的必要性。外层优化通过明确的目标导向，确保生成的扰动并非单纯的随机噪声，而是一种针对模型学习机制设计的系统性干扰。如果省略外层优化步骤，模型仍可能从扰动数据中提取有用的信息，尤其是在扰动与数据标签存在一定关联的情况下，扰动的有效性会显著降低。因此，缺乏外层优化的扰动很可能难以达到真正意义上的不可学状态，仅表现为较弱的干扰效应。综上，外层优化是确保不可学样本真正具备阻止模型有效学习能力的决定性因素。省略该步骤可能导致生成的扰动过弱，无法实现数据的不可学性，进而削弱不可学样本在实际应用中的有效性。

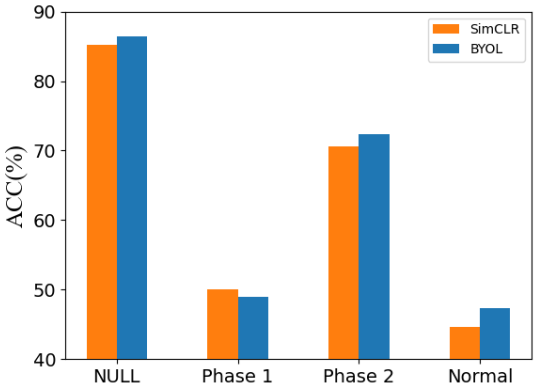


图 4.6 代理模型是否参与训练的影响
Figure 4.6 Impact of Whether the Surrogate Model Participates in Training

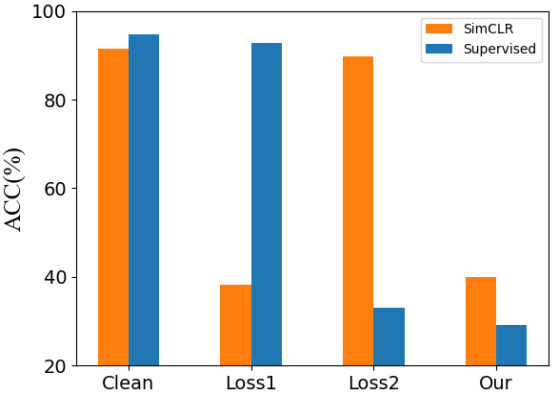


图 4.7 不同损失设置的影响
Figure 4.7 Impact of Different Loss

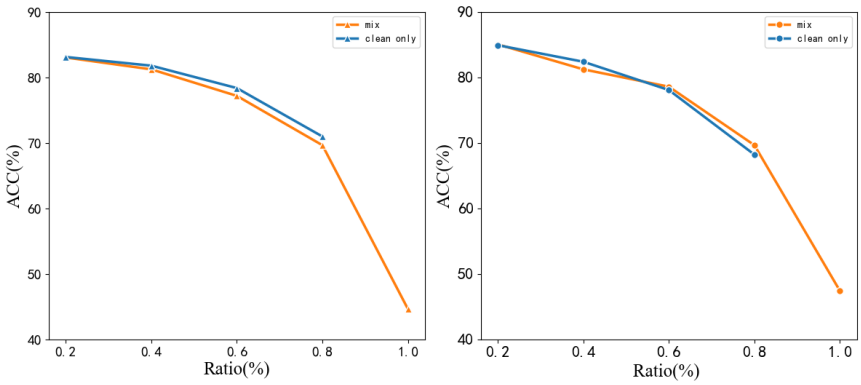
（4）不同损失设置的影响

为验证本章所提出两种损失的有效性，我们进一步设计了在不同损失设置下的消融实验。具体而言，在 CIFAR-10 数据集上，针对 ResNet-18 模型，分别在 SimCLR 对比学习算法和传统监督学习算法中，测试了仅使用对比学习损失

（Loss1）、仅使用聚类紧疏损失（Loss2）以及两者结合（Our）三种设置的效果。如图 4.7 所示，单独使用对比学习损失或聚类紧疏损失均能在各自的训练范式下发挥一定作用。然而当生成的不可学样本迁移至不同的范式后完全丧失了保护能力，当两种损失联合使用时，不可学效果显著增强，生成的不可学样本在监督与无监督场景下均表现出良好的不可学效果。该结果充分验证了本章提出损失设计方案的合理性与必要性。

（5）不可学样本占比的影响

在真实的不可学样本应用场景中，数据收集者极有可能收集到的数据并不来自单一信源，所以不可学样本的占比很难达到理想的 100%。因此，针对不可学样本占比的实验显得尤为必要。本实验通过对 ImageNet-12 数据集上生成的不可学样本占比进行测试，实验范围从最低的 20%到最高的 100%，并分别在两类不同的对比学习算法下进行实验。实验结果如图 4.8 所示，棕色折线表示仅使用原始图像训练的结果，而蓝色折线则表示同时使用原始图像和不可学样本进行训练。实验结果表明，不可学样本并没有对模型的学习产生显著影响。即使数据集中包含大量不可学样本（例如 80%），只要剩余的原始图像占据足够的比例，模型的表现仍然与仅使用这些原始图像训练的模型相似。这表明，虽然不可学样本本身不能为模型提供有效的训练信息，它们对模型训练的影响较小。事实上，模型的最终表现主要依赖于原始图像的数量，而不可学样本的存在仅起到保护作用，确保对比学习框架的完整性，并有效地防止模型对这些不可学样本的学习。因此，增加不可学样本的比例，前提是原始图像数量足够时，模型性能仍能得到保持。



（a）SimCLR　　（b）BYOL

图 4.8 不同对比学习算法下训练集中不同不可学样本占比的影响, (a)SimCLR (b)BYOL

Figure 4.8 Impact of Different Proportions of Unlearnable Examples in the Training Set under Various Contrastive Learning Algorithms, (a)SimCLR, (b) BYOL

4.3.7 不可学样本的扰动

如图 4.9 所示，本文通过可视化分析揭示了 PtEm-UE 生成的对抗扰动特性。基于训练完成的扰动嵌入网络，我们选取 6 张不同类别的原始图像，生成不可学样本后计算差值得到嵌入噪声。实验表明，PtEm-UE 利用深网络度隐写技术将扰动嵌入原始图像中，扰动嵌入网络会依据输入图像的纹理特征，自适应性选择最优区域嵌入扰动。这种动态嵌入机制使得噪声分布与图像局部结构呈现高度相关性，例如，在蝴蝶翅膀鳞片纹理处呈现精细条纹模式，在背景虚化区域则转为弥散形态。这种隐写的方式确保生成的不可学样本既能有效破坏模型训练，又能保持较高的视觉质量。

图 4.9 不可学样本及其扰动
Figure 4.9 Unlearnable examples and their noise

4.4 本章小结

本章提出基于扰动嵌入的不可学样本生成框架 PtEm-UE，通过两阶段交替优化机制实现数据训练阶段的主动防护。第一阶段采用扰动嵌入网络与代理模型的交替优化，迫使代理模型关注随机扰动；第二阶段迭代更新扰动增强其不可学性。设计了两种互补的损失函数，对比学习损失通过误导特征对齐过程，削弱模型对语义特征的建模能力以干扰自监督学习，而聚类损失通过约束类内紧密度与类间分离度干扰监督学习保证生成的不可学样本同时适用于监督学习和自监督

学习。此外，本章创新地将传统隐写的解码验证机制转换为“训练失效即隐写成功”的判据范式，摆脱了显式解码约束，使优化目标聚焦于隐蔽性与不可学效果的协同提升，有效平衡了隐蔽性与不可学性的矛盾需求，为数据滥用防护提供了新的技术路径。

第五章　总结与展望

5.1　总结

本文针对深度神经网络在数据全生命周期中的安全威胁，提出了两个创新性防护框架：基于通用扰动的对抗样本生成方法 PtEm-AE 与样本针对性的不可学样本生成方法 PtEm-UE，分别应对模型推理阶段与训练阶段的数据滥用问题，形成了全面的数据安全防御机制。

对抗样本生成框架 PtEm-AE 采用三阶段生成流程：第一阶段通过扰动嵌入网络学习扰动嵌入模式，第二阶段迭代更新扰动增强其攻击性，第三阶段使用优化后的扰动和嵌入网络生成每个原始图像对应的对抗样本。与传统隐写技术不同，本文以模型攻击成功率作为隐写有效性的核心验证指标——当对抗样本成功误导模型时，即证明扰动已被有效嵌入。在损失函数设计上，本文分别从样本层面与数据分布层面提出了两种互补的损失函数，以系统性地增强所生成对抗样本的攻击性与迁移性。实验结果验证了该方法在多个数据集与模型上的优异攻击效果与鲁棒性。

不可学样本生成框架 PtEm-UE 则采用两阶段交替优化策略：第一阶段交替优化扰动嵌入网络与代理模型，通过对比学习损失迫使代理模型关注扰动特征；第二阶段冻结扰动嵌入网络，交替优化扰动和代理模型，针对每个样本优化特异性扰动，并利用聚类紧疏损失引入监督信号。该框架创新性地将传统隐写的“编码-解码”验证机制转换为“扰动嵌入-训练阻断”范式：当不可学样本阻断模型训练时，即表明扰动嵌入达到预期效果。其中，对比学习损失通过误导特征对齐过程，削弱模型对语义特征的建模能力以干扰自监督学习，而聚类损失通过约束类内紧密度与类间分离度干扰监督学习。实验表明，PtEm-UE 在多个数据集上表现优越，能够有效保护数据隐私并保持较高的视觉质量。

整体而言，本文针对深度神经网络的数据隐私泄露风险，构建覆盖模型训练与应用全生命周期的防护体系，为数据隐私保护提供了新的技术方案。

5.2 展望

尽管本文提出的技术方案在数据安全保护方面取得了显著进展，仍然存在多个值得进一步深入研究的课题。首先，就对抗样本的泛化性而言，目前的研究主要聚焦于特定数据集与特定模型结构，缺乏对跨差异较大的模型架构的泛化能力的全面探讨。未来的研究可致力于增强对抗样本在不同任务和多种网络结构间的泛化性能，以进一步扩展其在实际场景中的应用潜力，或引入元学习框架增强对抗扰动的迁移性。

其次，在不可学样本的生成方面，目前的实验结果表明，当不可学样本占比降低时，其对模型学习的干扰效果相应减弱。因此，未来研究可进一步优化不可学样本的生成策略，以在低比例条件下依然保持较强的不可学习特性。此外，可以考虑构建数据价值评估模型，例如使用 Shapley 值量化样本对模型训练的贡献度，进而实现精准定向防护。

综上所述，未来的研究方向可进一步关注对抗样本的迁移性以及不可学样本在低占比条件下的鲁棒性优化，以持续提高深度学习数据安全防护技术的完备性与实用性，为深度学习环境下的数据隐私保护提供更加坚实的理论基础与实用方法。

参考文献

［1］ SHAH M, SUREJA N. A comprehensive review of bias in deep learning models: Methods, impacts, and future directions [J]. Archives of Computational Methods in Engineering, 2025, 32(1): 255-67.

[2] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks [J]. Communications of the ACM, 2017, 60(6): 84-90.

[3] DENG J, DONG W, SOCHER R, et al. Imagenet: A large-scale hierarchical image database[C]// Proceedings of the 2009 IEEE conference on computer vision and pattern recognition. 2009: 248-55.

[4] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [J]. Advances in neural information processing systems, 2017, 30.

[5] YAN K, WANG X, LU L, SUMMERS R M. DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning [J]. Journal of Medical Imaging, 2018, 36501: 1.

[6] HWANG J-J, XU R, LIN H, et al. Emma: End-to-end multimodal model for autonomous driving [J]. arXiv preprint arXiv:241023262, 2024.

[7] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners [J]. OpenAI blog, 2019, 1(8): 9.

[8] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners [J]. Advances in neural information processing systems, 2020, 33: 1877-901.

[9] LIU A, FENG B, WANG B, et al. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model [J]. arXiv preprint arXiv:240504434, 2024.

[10] GUO D, YANG D, ZHANG H, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning [J]. arXiv preprint arXiv:250112948, 2025.

[11] 张雪娟. 人工智能在金融风险管理中的应用与挑战 [J]. 全国流通经济, 2025, (08): 177-80.

[12] SILVER D, SCHRITTWIESER J, SIMONYAN K, et al. Mastering the game of go without human knowledge [J]. nature, 2017, 550(7676): 354-9.

[13] HE T, GAO J, XIAO W, et al. ASAP: Aligning simulation and real-world physics for learning agile humanoid whole-body skills [J]. arXiv preprint arXiv:250201143, 2025.

[14] MOUJAHID A, DORNAIKA F. Advanced unsupervised learning: a comprehensive overview of multi-view clustering techniques [J]. Artificial Intelligence Review, 2025, 58(8): 1-52.

[15] SESTINO A, KAHLAWI A, DE MAURO A. Decoding the data economy: a literature review of its impact on business, society and digital transformation [J]. European Journal of Innovation Management, 2025, 28(2): 298-323.

[16] FREDRIKSON M, JHA S, RISTENPART T. Model inversion attacks that exploit confidence information and basic countermeasures[C]// Proceedings of the Proceedings of the 22nd ACM SIGSAC conference on computer and communications security. 2015: 1322-33.

[17] CARLINI N, HAYES J, NASR M, et al. Extracting training data from diffusion models[C]// Proceedings of the 32nd USENIX Security Symposium (USENIX Security 23). 2023: 5253-70.

[18] ROMBACH R, BLATTMANN A, LORENZ D, et al. High-resolution image synthesis with latent diffusion models[C]// Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 10684-95.

[19] SAHARIA C, CHAN W, SAXENA S, et al. Photorealistic text-to-image diffusion models with deep language understanding [J]. Advances in neural information processing systems, 2022, 35: 36479-94.

[20] TAIGMAN Y, YANG M, RANZATO M A, WOLF L. Deepface: Closing the gap to human-level performance in face verification[C]// Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 1701-8.

[21] HUANG G B, MATTAR M, BERG T, LEARNED-MILLER E. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments[C]// Proceedings of the Workshop on faces in'Real-Life'Images: detection, alignment, and recognition. 2008.

[22] ZHANG S, FENG Y, BAUER L, et al. “Did you know this camera tracks your mood?”: Understanding Privacy Expectations and Preferences in the Age of Video Analytics [J]. Proceedings on Privacy Enhancing Technologies, 2021, 2: 282-304.

[23] BUCKNER C. Understanding adversarial examples requires a theory of artefacts for deep learning [J]. Nature Machine Intelligence, 2020, 2(12): 731-6.

[24] SHARIF M, BHAGAVATULA S, BAUER L, REITER M K. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition[C]// Proceedings of the Proceedings of the 2016 acm sigsac conference on computer and communications security. 2016: 1528-40.

[25] JIN D, JIN Z, ZHOU J T, SZOLOVITS P. Is bert really robust? a strong baseline for natural language attack on text classification and entailment[C]// Proceedings of the Proceedings of the AAAI conference on artificial intelligence. 2020: 8018-25.

[26] DEVLIN J, CHANG M-W, LEE K, TOUTANOVA K. Bert: Pre-training of deep bidirectional transformers for language understanding[C]// Proceedings of

the Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). 2019: 4171-86.

[27] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks [J]. arXiv preprint arXiv:13126199, 2013.

[28] HE K, ZHANG X, REN S, SUN J. Deep residual learning for image recognition[C]// Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-8.

[29] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples [J]. stat, 2015, 1050: 20.

[30] ILYAS A, ENGSTROM L, ATHALYE A, LIN J. Black-box adversarial attacks with limited queries and information[C]// Proceedings of the International conference on machine learning. 2018: 2137-46.

[31] MOOSAVI-DEZFOOLI S-M, FAWZI A, FAWZI O, FROSSARD P. Universal adversarial perturbations[C]// Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1765-73.

[32] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]// Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015). 2015.

[33] YIN Z, YE M, ZHANG T, et al. Vlattack: Multimodal adversarial attacks on vision-language tasks via pre-trained models [J]. Advances in Neural Information Processing Systems, 2023, 36: 52936-56.

[34] FAN J, YAN Q, LI M, et al. A survey on data poisoning attacks and defenses[C]// Proceedings of the 2022 7th IEEE International Conference on Data Science in Cyberspace (DSC). 2022: 48-55.

[35] HUANG H, MA X, ERFANI S M, et al. Unlearnable Examples: Making Personal Data Unexploitable[C]// Proceedings of the International Conference on Learning Representations.

[36] CHEN Y, SHEN C, SHEN Y, et al. Amplifying membership exposure via data poisoning [J]. Advances in Neural Information Processing Systems, 2022, 35: 29830-44.

[37] BROSCHEIT S, DO Q, GASPERS J. Distributionally robust finetuning BERT for covariate drift in spoken language understanding[C]// Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022: 1970-85.

[38] KRIZHEVSKY A. Learning Multiple Layers of Features from Tiny Images [J]. Master's thesis, University of Tront, 2009.

[39] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks [J]. arXiv preprint arXiv:170606083, 2017.

[40] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]// Proceedings of the 2017 ieee symposium on security and privacy (sp). 2017: 39-57.

[41] MOOSAVI-DEZFOOLI S-M, FAWZI A, FROSSARD P. Deepfool: a simple and accurate method to fool deep neural networks[C]// Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2574-82.

[42] KURAKIN A, GOODFELLOW I J, BENGIO S. Adversarial Machine Learning at Scale[C]// Proceedings of the International Conference on Learning Representations. 2017.

[43] DONG Y, LIAO F, PANG T, et al. Boosting adversarial attacks with momentum[C]// Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 9185-93.

[44] SU J, VARGAS D V, SAKURAI K. One pixel attack for fooling deep neural networks [J]. IEEE Transactions on Evolutionary Computation, 2019, 23(5): 828-41.

[45] MOPURI K R, GARG U, BABU R V. Fast feature fool: A data independent approach to universal adversarial perturbations [J]. arXiv preprint arXiv:170705572, 2017.

[46] XIAO C, ZHU J-Y, LI B, et al. Spatially Transformed Adversarial Examples[C]// Proceedings of the International Conference on Learning Representations. 2018.

[47] XIAO C, LI B, ZHU J Y, et al. Generating adversarial examples with adversarial networks[C]// Proceedings of the 27th International Joint Conference on Artificial Intelligence, IJCAI 2018. 2018: 3905-11.

[48] BAI T, ZHAO J, ZHU J, et al. Ai-gan: Attack-inspired generation of adversarial examples[C]// Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP). 2021: 2543-7.

[49] SHEN J, ZHU X, MA D. TensorClog: An imperceptible poisoning attack on deep neural network applications [J]. IEEE Access, 2019, 7: 41498-506.

[50] FOWL L, GOLDBLUM M, CHIANG P-Y, et al. Adversarial examples make strong poisons [J]. Advances in Neural Information Processing Systems, 2021, 34: 30339-51.

[51] KOH P W, LIANG P. Understanding black-box predictions via influence functions[C]// Proceedings of the International conference on machine learning. 2017: 1885-94.

[52] FENG J, CAI Q-Z, ZHOU Z-H. Learning to confuse: Generating training time adversarial data with auto-encoder [J]. Advances in Neural Information Processing Systems, 2019, 32.

[53] LIU Z, ZHAO Z, KOLMUS A, et al. Going grayscale: The road to understanding and improving unlearnable examples [J]. arXiv preprint arXiv:211113244, 2021.

[54] FU S, HE F, LIU Y, et al. Robust Unlearnable Examples: Protecting Data Privacy Against Adversarial Learning[C]// Proceedings of the International Conference on Learning Representations.

[55] YU D, ZHANG H, CHEN W, et al. Availability attacks create shortcuts[C]// Proceedings of the Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2022: 2367-76.

[56] REN J, XU H, WAN Y, et al. Transferable Unlearnable Examples[C]// Proceedings of the The Eleventh International Conference on Learning Representations.

[57] HE H, ZHA K, KATABI D. Indiscriminate Poisoning Attacks on Unsupervised Contrastive Learning[C]// Proceedings of the The Eleventh International Conference on Learning Representations.

[58] ZHANG J, MA X, YI Q, et al. Unlearnable clusters: Towards label-agnostic unlearnable examples[C]// Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 3984-93.

[59] ZHANG Z, ZHANG J, ZHANG K, et al. Segue: Side-information Guided Generative Unlearnable Examples for Facial Privacy Protection in Real World [J]. CoRR, 2023.

[60] JIANG Y, MA X, ERFANI S M, BAILEY J. Unlearnable examples for time series[C]// Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining. 2024: 213-25.

[61] SUN W, LIU Y, YAN Z, et al. Medical Unlearnable Examples: Securing Medical Data from Unauthorized Training via Sparsity-Aware Local Masking[C]// Proceedings of the ICML 2024 Next Generation of AI Safety Workshop.

[62] GOKUL V, DUBNOV S. Poscuda: Position based convolution for unlearnable audio datasets [J]. arXiv preprint arXiv:240102135, 2024.

[63] WANG X, LI M, LIU W, et al. Unlearnable 3D point clouds: Class-wise transformation is all you need [J]. Advances in Neural Information Processing Systems, 2024, 37: 99404-32.

[64] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks [J]. Advances in neural information processing systems, 2012, 25.

[65] LIU Z, LIN Y, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]// Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision. 2021: 10012-22.

[66] REN S, HE K, GIRSHICK R, SUN J. Faster R-CNN: Towards real-time object detection with region proposal networks [J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 39(6): 1137-49.

[67] HE K, GKIOXARI G, DOLLáR P, GIRSHICK R. Mask r-cnn[C]// Proceedings of the Proceedings of the IEEE international conference on computer vision. 2017: 2961-9.

[68] LIU Y, ZHU S, XIA J, et al. End-to-end learnable clustering for intent learning in recommendation [J]. Advances in Neural Information Processing Systems, 2024, 37: 5913-49.

[69] BREITENSTEIN J, TERMöHLEN J-A, LIPINSKI D, FINGSCHEIDT T. Corner cases for visual perception in automated driving: some guidance on detection approaches [J]. arXiv preprint arXiv:210205897, 2021.

[70] LECUN Y, BOTTOU L, BENGIO Y, HAFFNER P. Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86(11): 2278-324.

[71] HUANG G, LIU Z, VAN DER MAATEN L, WEINBERGER K Q. Densely connected convolutional networks[C]// Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4700-8.

[72] CHEN X, HE K. Exploring simple siamese representation learning[C]// Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 15750-8.

[73] WESTFELD A. F5—A Steganographic Algorithm [J]. Springer Berlin Heidelberg, 2001, 2137: 289.

[74] HAYES J, DANEZIS G. Generating steganographic images via adversarial training[C]// Proceedings of the Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017: 1951-60.

[75] ZHU J, KAPLAN R, JOHNSON J, FEI-FEI L. Hidden: Hiding data with deep networks[C]// Proceedings of the Proceedings of the European conference on computer vision (ECCV). 2018: 657-72.

[76] ZHANG K A, CUESTA-INFANTE A, XU L, VEERAMACHANENI K. SteganoGAN: High capacity image steganography with GANs [J]. arXiv preprint arXiv:190103892, 2019.

[77] YU J, ZHANG X, XU Y, ZHANG J. Cross: Diffusion model makes controllable, robust and secure image steganography [J]. Advances in Neural Information Processing Systems, 2023, 36: 80730-43.

[78] SONG J, MENG C, ERMON S. Denoising Diffusion Implicit Models[C]// Proceedings of the International Conference on Learning Representations. 2021.

[79] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples [J]. arXiv preprint arXiv:14126572, 2014.

[80] SUYA F, CHI J, EVANS D, TIAN Y. Hybrid Batch Attacks: Finding Black-box Adversarial Examples with Limited Queries[C]// Proceedings of the USENIX Security Symposium. 2020.

[81] SHAMSABADI A S, SANCHEZ-MATILLA R, CAVALLARO A. Colorfool: Semantic adversarial colorization[C]// Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 1151-60.

[82] DUAN R, CHEN Y, NIU D, et al. Advdrop: Adversarial attack to dnns by dropping information[C]// Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision. 2021: 7506-15.

[83] BROWN T B, MANé D, ROY A, et al. Adversarial patch [J]. arXiv preprint arXiv:171209665, 2017.

[84] XU J, LIU H, WU D, et al. Generating universal adversarial perturbation with ResNet [J]. Information Sciences, 2020, 537: 302-12.

[85] KANG X, SONG B, WANG D, CAI X. Crafting universal adversarial perturbations with output vectors [J]. Neurocomputing, 2022, 501: 294-305.

[86] YE Z, CHENG X, HUANG X. FG-UAP: Feature-gathering universal adversarial perturbation[C]// Proceedings of the 2023 International Joint Conference on Neural Networks (IJCNN). 2023: 1-8.

[87] FANG B, LI B, WU S, et al. Towards generalizable data protection with transferable unlearnable examples [J]. arXiv preprint arXiv:230511191, 2023.

[88] LIU Y, XU K, CHEN X, SUN L. Stable unlearnable example: Enhancing the robustness of unlearnable examples via stable error-minimizing noise[C]// Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence. 2024: 3783-91.

[89] LIN T-Y, MAIRE M, BELONGIE S, et al. Microsoft coco: Common objects in context[C]// Proceedings of the Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13. 2014: 740-55.

[90] HUANG C-Y, LIN Y Y, LEE H-Y, LEE L-S. Defending your voice: Adversarial attack on voice conversion[C]// Proceedings of the 2021 IEEE Spoken Language Technology Workshop (SLT). 2021: 552-9.

[91] SALMAN H, KHADDAJ A, LECLERC G, et al. Raising the Cost of Malicious AI-Powered Image Editing[C]// Proceedings of the International Conference on Machine Learning. 2023: 29894-918.

[92] CHEN T, KORNBLITH S, NOROUZI M, HINTON G. A simple framework for contrastive learning of visual representations[C]// Proceedings of the International conference on machine learning. 2020: 1597-607.

[93] SZEGEDY C, IOFFE S, VANHOUCKE V, ALEMI A. Inception-v4, inception-resnet and the impact of residual connections on learning[C]// Proceedings of the Proceedings of the AAAI conference on artificial intelligence. 2017.

[94] CHATTERJEE S. Matrix estimation by universal singular value thresholding [J]. THE ANNALS, 2015, 43(1): 177-214.

攻读硕士学位期间取得的研究成果

一、论文

[1] **任行东**，孙广玲. StegoAE：一种基于扰动嵌入的对抗样本生成方法[J]. 工业控制计算机,2025. (本人为第一作者，已录用)

[2] Zhou, L., **Ren, X.**, Qian, C., & Sun, G. (2024). TraceGuard: Fine-Tuning Pre-Trained Model by Using Stego Images to Trace Its User. Mathematics, 12(21), 3333. (本人为共同一作，已发表)

[3] **Xingdong Ren**, Hanyang Qian, Yinggui Wang, Guangling Sun. StegoUE: Generation of Unlearnable Examples Using Steganography for Contrastive Learning[J]. Journal of Electronic Imaging (本人为第一作者，All Reviewers Assigned).

[4] **Xingdong Ren**, Haojie Liu, Hanzhou Wu, Yinggui Wang, Xiaofeng Lu, Guangling Sun. StegoGuard: Secrets Encoder and Decoder Act as Fingerprint of Self-supervised Pre-trained Model[J]. IEEE Internet of Things Journal (本人为第一作者，Minor Revision).

二、专利

[1] **任行东**, 孙广玲, 陆小锋. 一种基于隐空间隐写技术的通用对抗扰动生成方法[P].申请受理号: 2024117465283.(本人为第一作者，已受理)

致　谢

行文至此，篇章将尽。常言道时光荏苒，不觉间，三年研究生生涯如白驹过隙，转瞬即逝。在这漫长而又短暂的研究生旅途中，我得以窥见知识的海洋，结识了众多杰出的同窗与师长。借此机会，我要向所有在这段旅程中给予我陪伴与支持的人表达我最深切的谢意。

恩师如山，教诲如泉。首先，我要向我的导师孙广玲老师致以崇高的敬意。从论文选题到成文定稿，孙老师始终给予我细致入微的指导与无私的帮助，提出的每一条修改建议都如金玉良言，使我受益匪浅。此外，感谢吴汉舟、王子驰老师，在科研与工程项目上对我的耐心引导和不厌其烦的解答疑惑，他们细心严谨、实事求是和精益求精的工作精神是我今后前行的榜样。

父母如天，恩重如山。感谢我亲爱的父母，他们对我无私的奉献与支持，是我一路走来最坚实的依靠。二十多年来，他们始终在背后默默支持我，尊重并鼓励我的每一个选择。正是站在他们的肩膀上，我得以放眼更宽广的世界。

同窗如友，相助如亲。感谢实验室的同学们，你们在项目研究和论文撰写中的无私协助；感谢那些在我学术道路上给予我支持与鼓励的每一位同学，你们的陪伴与关怀，是我前行路上的温暖阳光；也感谢我的室友们，你们的包容、关怀与支持，让我在异乡也能感受到家的温馨。

最后，衷心感谢所有评阅本论文以及参加答辩的各位专家与教授，感谢您们在百忙之中抽出宝贵时间审阅我的论文，并提出宝贵的意见和建议。