计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
62(10): 2455−2467, 2025

针对 SAM 下游模型脆弱模块的对抗迁移攻击

丁 熠 林能健 蒋昀陶 钟宇浩 曹明生
(网络与数据安全四川省重点实验室(电子科技大学) 成都 610054)
(yi.ding@uestc.edu.cn)

Adversarial Transfer Attacks on Weak Block of SAM Downstream Models

Ding Yi, Lin Nengjian, Jiang Yuntao, Zhong Yuhao, and Cao Mingsheng
(*Network and Data Security Key Laboratory of Sichuan Province* (*University of Electronic Science and Technology of China*), *Chengdu* 610054)

Abstract As a common visual foundation model, the segment anything model (SAM) has been widely applied in various image segmentation tasks, but it exhibits vulnerability in the face of adversarial attacks. A adversarial transfer attack method FSGR (fragile section gradient robustness) has been proposed for vulnerable modules of SAM downstream models, which can effectively generate adversarial samples to attack SAM downstream models without understanding the details of fine-tuning in downstream. The core of FSGR lies in its ability to identify the most vulnerable parts and apply targeted attacks to these weak points. This method designs a strategy based on feature similarity to identify the most vulnerable module in SAM encoder and enhance the attack effectiveness in a targeted manner, and introduces gradient robust loss to eliminate the gradient difference between the target model and downstream task models. FSGR combines vulnerability layer attacks and gradient robust loss to achieve relative performance improvement on multiple datasets. The experimental results show that FSGR significantly reduces model performance in transfer attacks on various fine-tuning models, such as medical segmentation, shadow segmentation, and camouflage segmentation, demonstrating its correctness and practicality. Compared with baseline methods, FSGR not only performs well in attack success rate, but also achieves relative performance improvement by combining vulnerability layer attacks and gradient robust loss.

Key words image segmentation; adversarial attack; transfer-based attack; feature similarity; model robustness

摘 要 SAM(segment anything model)作为一种通用的视觉基础模型,已被广泛应用于多种图像分割任务,但其在对抗性攻击面前表现出脆弱性.提出一种针对 SAM 下游模型脆弱模块的对抗迁移攻击方法 FSGR(fragile section gradient robustness).该方法在无需知晓下游微调细节的前提下,可有效生成对抗样本,实现对 SAM 下游模型的攻击.该方法运用“脆弱层精准定位+局部强化迁移”策略,通过特征相似度筛

收稿日期: 2025-06-04; **修回日期**: 2025-07-18
基金项目: 国家自然科学基金项目(62027827, 62372083); 四川省科技计划项目(2023YFS0020, 2023YFS0197, 2023YFG0148); 电子科技大学医工交叉联合基金(ZYGX2021YGLH212, ZYGX2022YGRH012); 成都医学院 2024 年第一批临床科学研究基金重点项目(24LHXX1-01, 24LHXX1-02); 四川省卫生健康委员会 2023 年科学技术项目——临床研究专项基金(23LCYJ011); 教育部残障人士智能康复与无障碍重点实验室(长春大学)项目(2024KFJJ004)
This work was supported by the National Natural Science Foundation of China (62027827, 62372083), the Sichuan Science and Technology Support Plan (2023YFS0020, 2023YFS0197, 2023YFG0148), the Medico-Engineering Cooperation Funds from University of Electronic Science and Technology of China (ZYGX2021YGLH212, ZYGX2022YGRH012), the Chengdu Medical College 2024 First Batch Clinical Scientific Research Fund Key Research Project (24LHXX1-01, 24LHXX1-02), the Sichuan Provincial Health Commission 2023 Science and Technology Project—Clinical Research Special Fund (23LCYJ011), and the Project of Key Laboratory of Intelligent Rehabilitation and Barrier-free for the Disabled (Changchun University) Ministry of Education (2024KFJJ004).
通信作者: 曹明生(Cms@uestc.edu.cn)

选出跨任务共享且最易被激活的模块,针对性地强化攻击效果;同时,引入梯度稳健损失以消除目标模型与下游任务模型间的梯度差异. FSGR 方法融合了脆弱层攻击与梯度稳健损失机制,在多个数据集上均实现了相对性能的提升. 实验结果表明,FSGR 在多种微调模型(如医学分割、阴影分割和伪装分割)的迁移攻击中显著降低了模型性能,证明了其正确性和实用性. 与基线方法相比,FSGR 不仅在攻击成功率上表现出色,还通过结合脆弱层攻击和梯度稳健损失,实现了相对性能的提升.

关键词 图像分割;对抗攻击;迁移攻击;特征相似度;模型鲁棒性

中图法分类号 TP311.13; TP309

DOI: 10.7544/issn1000-1239.202550388 **CSTR:** 32373.14.issn1000-1239.202550388

深度学习在图像分割领域取得了显著的进展,通过利用卷积神经网络(CNN)和 Transformer 等架构,能够提取图像中的特征并进行像素级分类. 其中 U-Net[1]、DeepLab[2] 系列和 Mask R-CNN[3] 等模型已经得到了广泛应用. 在医学影像分析中,U-Net 常用于辅助诊断以分割病变区域;在自动驾驶领域,Mask R-CNN 用于检测道路上的车辆和行人. 随着深度学习技术的发展和下游任务的多样化应用,图像分割在多个领域展现出了巨大的潜力和价值.

SAM(segment anything model)是由 Meta AI 开发的一种开创性的视觉基础模型,专门用于图像分割[4]. 它使用来自海量视觉语料库的超过 10 亿个掩码进行训练,能够在各种场景中分割各种物体和结构. SAM 展示了出色的零样本能力和跨各种任务的广泛泛化能力,彻底改变了图像分割领域.

SAM 的开放可访问性使其成为一个有前景的基础模型,作为微调某些领域和下游应用的分析模型的起点,例如医学分割[5]、3D 物体分割[6]、伪装物体分割[7]、透明物体分割[8]、航拍图像分割[9] 和高质量分割[10]. 随着研究的深入,SAM 在面对对抗性攻击时暴露出严重的安全问题. 通过在干净输入中加入精心制作的、几乎不可察觉的对抗性扰动,攻击者可以以高成功率误导训练有素的模型,即使攻击者只有有限的信息可用.

对抗攻击的研究最早可以追溯到 2015 年,Goodfellow 等人[11] 发现深度神经网络对对抗样本的脆弱性. 随后,2016 年,Moosavi-Dezfooli 等人[12] 提出了 DeepFool 方法,通过计算出最小的必要扰动来构建对抗样本. 此后,对抗攻击相关研究逐渐兴起并持续发展,目前已形成黑盒攻击与白盒攻击两大成熟分支. 二者的核心区别在于攻击者是否能够完全访问目标模型[13],这类研究的进展促使研究者开始思考如何提升模型的对抗鲁棒性[14]. 在实际生产应用场景中,由于攻击者难以获取目标模型的完整信息,白盒攻击的威胁程度显著降低;而基于迁移性的黑盒攻击仅需较少信息即可对现实应用构成实质性威胁,这一特性引发了学界与业界对开源模型微调安全性的高度关注.

大语言模型的成功使得对抗攻击领域被推广到了大模型上[15]. 随着 SAM 等视觉基础模型的出现,针对 SAM 的对抗样本研究也逐渐展开. 例如,2023 年提出的 Attack-SAM[16] 是一种基于 mask removal 的对抗攻击方法,并在白盒和黑盒 2 种情况下对 SAM 进行评估. 2024 年提出的 ASAM[17] 方法通过对抗性微调来增强 SAM 的性能,该方法通过使用稳定的扩散模型生成更能代表自然变化的对抗实例,从而在不损害其固有泛化能力的前提下提升了 SAM 的性能. 2024 年提出的 UMI-GRAT[18] 方法创新性地揭示了模型中固有的脆弱性,并探索了攻击者在不访问下游任务特定数据集来训练一个紧密对齐的替代模型的情况下,对任何受害模型发动对抗性攻击的能力.

尽管 GRAT 通过其独特的设计,成功地从模型中提取出内在的脆弱性,在实际攻击过程中并未直接针对这些脆弱模块进行针对性的攻击. 这种设计使得 GRAT 在攻击策略上具有一定的间接性.

在对抗性攻击领域,特别是在针对基于 Transformer 架构的模型时,识别和利用模型中的脆弱模块是提高攻击成功率(attack success rate, ASR)的关键. 脆弱模块在处理输入数据时表现出较高的敏感性,使得微小变化可能导致模型输出的显著变化,因此被视为模型的脆弱点,攻击者可以利用这些脆弱点来生成更有效的对抗性样本. 本文提出了一种针对 SAM 下游模型脆弱模块的对抗迁移攻击方法 FSGR(fragile section gradient robustness),寻找 Transformer 模块中最易受攻击的模块;并引入了梯度稳健损失,通过拟合原模型与下游模型的梯度差异,在扰动迭代过程中更新微调引起的梯度偏差. 该攻击有效地误导了 SAM 及其下游任务微调模型. 综上所述,本研究的主要贡

献有如下 4 点：

1）提出一种针对 SAM 下游模型脆弱模块的对抗迁移攻击方法，该方法在无需知晓下游微调细节的前提下，可有效生成对抗样本以实现对 SAM 下游模型的攻击.

2）设计基于特征相似度的识别策略，定位 SAM 编码器中抗攻击能力最弱的 Transformer 模块，从而针对性地强化攻击效果.

3）将本文提出的脆弱层攻击与现有 GRAT 方法相结合，并通过实验验证了该融合方法的有效性.

4）通过大量实验验证，在医学分割、阴影分割、伪装分割等多种下游任务中均证实了所提方法的有效性，且能显著降低模型性能.

1 相关工作

1.1 基础模型 SAM

SAM 通过一个通用的分割框架，适应不同的分割任务，而无需针对每个任务重新设计模型. SAM 模型的图像编码器将输入图像映射为紧凑的特征嵌入，为后续分割提供语义基础；提示编码器把类别标签、边界框等提示信息转换为与图像特征对齐的嵌入向量；掩码解码器融合 2 种嵌入后输出最终的分割掩码. 通过这一流程，SAM 在大量自然图像上展示了“零样本”泛化能力.

1.2 模型微调

对 SAM 模型的微调使其在医学图像分割、自然图像分割或伪装物体分割等场景中取得更好的表现. SAMed[5] 采用 LoRA[19] 技术仅对掩码解码器进行轻量级微调，在 3D 与 2D 医学影像上均取得了优于原始 SAM 的效果. Shadow-SAM 与 Camouflaged-SAM 同属 SAM-Adapter[7] 族，在 ViT 各层插入小型 Adapter，将领域先验或视觉提示作为附加信息，与大规模通用知识结合，显著提升挑战性任务上的精度. 然而，这些方法需要为每个下游任务保存一组增量权重，当任务数量增多时存储开销可观.

1.3 对抗性攻击

对抗性攻击是一种针对深度学习模型的攻击方法，其核心思想是通过在输入数据中添加精心设计的微小扰动，使得模型的输出发生错误.

作为一个简单却十分经典的对抗性攻击方法，FGSM[11] 通过一步梯度更新生成对抗性样本. 然而，扰动方向的单步估计往往过于粗糙，生成的对抗样本容易落入局部极值. I-FGSM[20] 将单步更新拆分为多次小步迭代，显著提升了攻击成功率，但计算量随步数线性增长. MI-FGSM[21] 进一步在 I-FGSM 的迭代框架中引入动量项以稳定更新方向和抑制振荡. 然而，动量系数与迭代步长的耦合增加了超参数调优复杂度. DMI-FGSM[22] 提出“输入多样性”策略，每一次迭代都先对输入进行随机缩放、随机填充或随机旋转，再计算梯度并应用动量更新，显著提升了跨模型、跨分辨率甚至跨任务的迁移能力. PGN[23] 通过最大化损失函数在输入邻域内的平坦性，使扰动方向在多个局部极大值间保持鲁棒，显著增强了黑盒攻击的可迁移性. BSR[24] 在迭代攻击过程中对特征图按块随机打乱并旋转，打破网络对局部纹理的依赖，从而提升对抗样本在异构模型间的迁移能力. ILPD[25] 在生成对抗扰动时逐步降低中间层特征的扰动强度，既保留高层语义失真又抑制低层纹理异常，兼顾了视觉隐蔽性与跨模型攻击成功率.

现有方法虽在提升对抗迁移性方面取得一定进展，但仍存在两大局限：1）这些方法均将整个网络视为均匀受扰的“黑箱”，在像素级或特征级施加全局无差别扰动，却忽视了 SAM 下游任务中冻结层梯度较小而微调头梯度较大的特性，导致多数噪声被浪费在难以扰动的区域，而对下游特征起关键作用的少数模块反而未得到充分扰动；2）这些方法对跨任务迁移的假设过于理想化，默认下游模型与上游 SAM 保持一致的通道、分辨率及权重分布，然而当下游采用 Adapter、LoRA 或分辨率缩放等调整措施时，攻击效果会急剧衰减. 为此，本文提出“脆弱层精准定位+局部强化迁移”策略：首先通过特征相似度筛选出跨任务共享且最易被激活的模块，再集中对该狭窄子空间施加扰动，从而在未知下游参数的条件下实现高效迁移攻击.

2 问题定义

以通用对抗攻击框架为例，设 f 为深度学习模型，x 为输入数据，y 为模型的预测输出. 对抗性攻击的目标是生成一个对抗性样本 x_{adv}，使得 $f(x_{\text{adv}}) \neq f(x)$. 对抗性扰动 δ 是在输入数据 x 上添加的微小变化，使得模型的输出发生错误，其生成通常通过优化以下目标函数来实现：

$$\delta = \arg\max_{\delta \in B_\varepsilon(x)} \mathrm{L}(x+\delta, \theta), \tag{1}$$

其中，$B_\varepsilon(x)$ 是以 x 为中心、ε 为半径的 ℓ_p-范数球，限制扰动的大小；θ 是模型参数.

通过迭代更新扰动δ, 逐步优化对抗性样本x_{adv}:

$$\delta_{t+1}=clip_{B_\varepsilon(x)}\left(\delta_t+\alpha\cdot\nabla_{\delta_t}L(x+\delta_t,\theta)\right), \tag{2}$$

其中, α是步长, $clip_{B_\varepsilon(x)}$用于将扰动投影到ℓ_p-范数球内, 确保扰动的大小不超过ϵ.

最终得到的对抗性样本x_{adv}可以表示为:

$$x_{\text{adv}}=x+\delta. \tag{3}$$

作为针对 SAM 模型的对抗迁移攻击, 本文的目标是基于 SAM 来产生扰动δ, 以最大化扰动图像在微调模型上的损失.

给定θ为模型参数, f_θ为基础模型 SAM 在一般数据集D上预训练的编码器, f_θ^k表示对于输入图像x在编码器第k层的 Transformer 模块, y^k为该模块输出的特征图. $f_{\theta'}$为经过下游数据集D'微调的 SAM 下游模型的编码器, $\mathcal{L}$是模型损失. AT是对抗攻击策略, δ为攻击过程中施加的扰动, ε为扰动预算, 即$\|\delta\|\leqslant\varepsilon$.

模型训练目标是迭代更新参数θ来最小化损失:

$$\theta=\arg\min\mathbb{E}\left[\mathcal{L}(f_\theta(x))\right],\ \text{s.t.}\{x\in D\}. \tag{4}$$

$$\theta'=\arg\min\mathrm{E}\left[\mathcal{L}(f_{\theta'}(x))\right],\ \theta\xrightarrow{\text{initial}}\theta',\ \text{s.t.}\{x\in D'\}. \tag{5}$$

针对模型的对抗攻击, 攻击者的目标是在扰动预算内找到最大化损失的扰动:

$$\delta=\arg\max\mathbb{E}\left[\mathcal{L}(f_\theta(x+\delta))\right],\ \text{s.t.}\{x\in D,\|\delta\|\leqslant\varepsilon\}. \tag{6}$$

此时的攻击策略和原模型紧密一致, 但在针对 SAM 的下游模型对抗迁移攻击中, 攻击者无法得知下游数据集和微调模型参数, 使得攻击策略的设计具有挑战性. 进一步分析对 SAM 下游模型的攻击, 微调过程中模型集成了一些原有的知识和微调后的知识之间存在一些相互联系. 考虑到这种联系, 一种直观的想法是设计一种攻击策略, 基于原模型来产生扰动δ, 去利用f_θ和$f_{\theta'}$的共同信息, 最大化扰动后图像在微调模型上的损失, 下游模型攻击如图 1 所示.

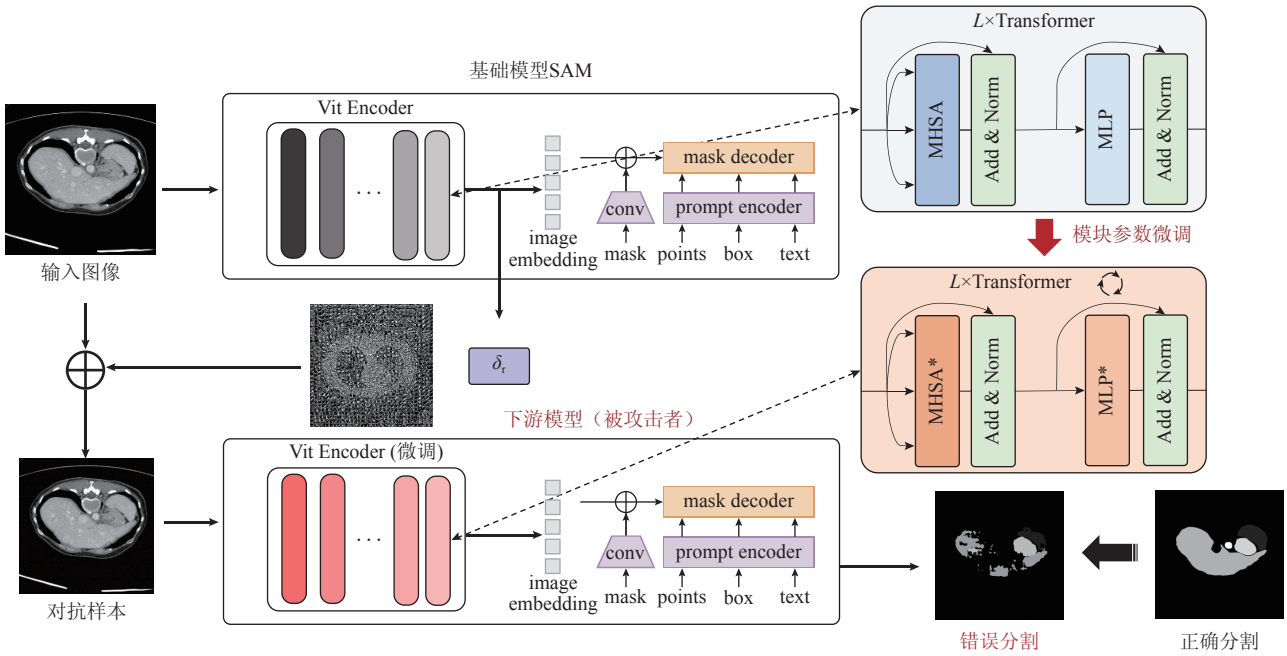


Fig. 1 Example diagram of downstream migration attack

图 1 下游迁移攻击示例图

$$\delta=AT(f_\theta,x),\ \text{s.t.}\{\|\delta\|\leqslant\varepsilon\}, \tag{7}$$

$$\max\mathbb{E}\left[\mathcal{L}(f_{\theta'}(x+\delta))\right],\ \text{s.t.}\{\delta=AT(f_\theta,x)\}. \tag{8}$$

损失$\mathcal{L}$通常为交叉熵损失$\mathcal{L}_{\text{OR}}$, 即对每个像素做K分类, 然后对所有像素求平均.

$$\mathcal{L}_{\text{OR}}=-\frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{k=1}^{K}y_{h,w,j}\log_{p_{h,w,k}}, \tag{9}$$

其中, H和W分别为输出特征图的高和宽, K为类别数(含背景), $y_{h,w,j}\in\{0,1\}$为像素(h,w)在类别k上的 ground-truth 标签, $p_{h,w,k}\in[0,1]$为模型对像素(h,w)在类别k的预测概率(Softmax 输出).

3 方法概述

3.1 对抗迁移攻击框架

本框架基于脆弱模块和梯度稳健损失迭代计算

梯度，拟合微调模型带来的偏差，具体来说，就是提高生成扰动的泛化性. 为了实现这个目标，框架引入了 2 个损失模块：一是脆弱损失攻击模块，通过计算编码器每层的特征图相似度，找出前 N 层最脆弱的 Transformer 模块并针对其发起攻击；二是梯度稳健损失模块，通过加入扰动迭代过程中更新微调引起的梯度偏差，使攻击结果更具稳健性.

具体过程如算法 1 和图 2 所示，框架以 MI-FGSM 为基线，引入 2 个新的损失模块迭代优化扰动，提升跨模型迁移能力. 1）脆弱损失 $\mathcal{L}_{\mathrm{FS}}$. 度量各层特征相似度，锁定最脆弱的 N 个 Transformer 模块精准攻击；2）梯度稳健损失 $\mathcal{L}_{\mathrm{GR}}$. 迭代修正微调导致的梯度偏差，增强扰动稳健性.

算法 1. 通过 $\mathcal{L}_{\mathrm{FS}}$ 和 $\mathcal{L}_{\mathrm{GR}}$ 生成对抗样本.

输入：SAM 开源基础模型 f_θ，自然图像数据集 D，任务特定图像 x，通用步长 η，攻击迭代 T_{a}，攻击步长 α，扰动预算 ε；

输出：扰动图像 x^{r}.

① 初始化 $\delta=\varepsilon^0$；

② for $t\leftarrow 1$ to T_{a} do

③ $x^{\mathrm{r}}=x+\delta$；

④ $g_{\mathrm{s}}\leftarrow\nabla(\mathcal{L}_{\mathrm{OR}}(f_\theta(x),f_\theta(x^{\mathrm{r}}))+\mathcal{L}_{\mathrm{FS}}(f_\theta(x),f_\theta(x^{\mathrm{r}}))+\mathcal{L}_{\mathrm{GR}}(f_\theta(x),f_\theta(x^{\mathrm{r}})))$；

⑤ $\delta=clip_{B_\epsilon}[\delta+\alpha\cdot sign(g_{\mathrm{s}})]$；

⑥ end for

⑦ $x^{\mathrm{r}}=x+\delta$.

算法 1 中，ε^0 代表 $N(\varepsilon;\mu,\sigma^2 I)$，即各向同性高斯噪声；均值 μ 等于扰动预算 ε. $\mathcal{L}_{\mathrm{OR}}$ 表示下游模型的原始损失函数，本文中采用的是交叉熵损失函数. *clip* 操作会将扰动 δ 的范围限制在 $(-\varepsilon,+\varepsilon)$. 我们将 T_{a} 设置为 10，通过最大化损失来更新对抗性扰动 δ.

3.2 脆弱损失攻击模块（FS）

传统 FGSM、MI-FGSM 沿梯度反方向扰动，在微调模型中会因方向失配而失效.

借鉴 Zheng 等人[26]DTA 的思想，我们直接攻击编码器中最脆弱的 Transformer 模块，以提升对抗样本的跨模型泛化能力.

这种策略首先定义了特征相似度 F，计算扰动图像 x_{r} 与输入图像 x 在编码器第 k 层的 Transformer 模块输出的特征图 y_{r}^k 和 y^k 的余弦相似度.

$$y^k=f_\theta^k(y^{k-1}),\quad y^0=x, \tag{10}$$

$$F^k=\cos(y^k,y_{\mathrm{r}}^k). \tag{11}$$

余弦相似度最小的模块最脆弱，原因有 2 个：1）差异大表明对抗样本在该层的特征空间中已经与干净样本“分道扬镳”，从而更容易误导下游模型；2）这种差异可在微调后保留，提升了迁移成功率.

本文找出前 N 个特征相似度最小的模块 $m^k\in M$，提出损失 $\mathcal{L}_{\mathrm{FS}}$ 为这些模块的特征相似度负数之和，扰动更新的目标是尽可能增大这些模块输出后的 $f_\theta^k(x)$ 与 $f_\theta^k(x^{\mathrm{r}})$ 的差异.

$$\mathcal{L}_{\mathrm{FS}}=\sum_{m^k\in M}-\cos\left(FS^k\right). \tag{12}$$

3.3 梯度稳健损失模块（GR）

除了寻找脆弱的 Transformer 模块来增强 AT 外，本文参考了另一种扰动生成方法，该思想由 Xia 等人[18]提出，通过拟合原模型与下游模型的梯度差异，在扰

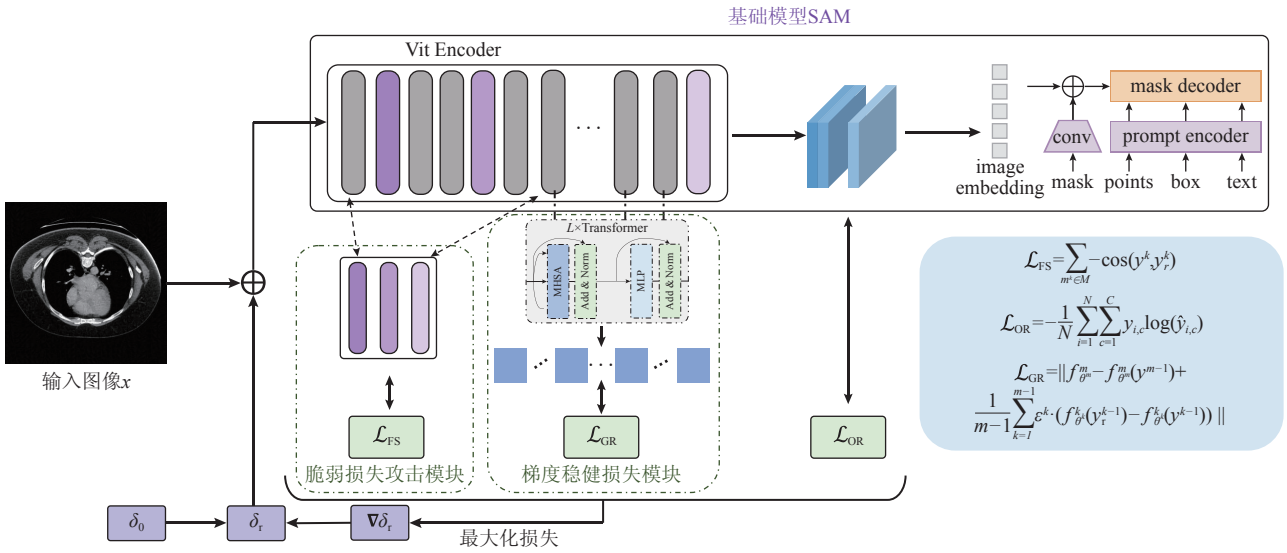


Fig. 2 Attack process

图 2 攻击过程

动迭代过程中更新微调引起的梯度偏差来增强对抗样本的稳健性.

SAM 的图像编码器由m个顺序连接的模块f_θ^k组成, 第k个模块的参数为θ^k, 经微调后的参数为θ_{r}^k, 差值为$\Delta\theta_{\mathrm{r}}^k$, 其中$h_{\Delta\theta_{\mathrm{r}}^k}^k$是由差值$\Delta\theta_{\mathrm{r}}^k$带来的对特征的更新.

$$y_{\mathrm{r}}^k = f_{\theta^k+\Delta\theta_{\mathrm{r}}^k}^k\left(y_{\mathrm{r}}^{k-1}\right) = f_{\theta^k}^k\left(y_{\mathrm{r}}^{k-1}\right) + h_{\Delta\theta_{\mathrm{r}}^k}^k\left(y_{\mathrm{r}}^{k-1}\right). \tag{13}$$

设δ_{s}为通过在基础模型上最大化损失$\mathcal{L}$生成得到的扰动, $\Delta\delta_{\mathrm{s}}$为每步扰动迭代的差值, δ_{r}为在下游模型上最大化损失$\mathcal{L}$生成得到的扰动, $\Delta\delta_{\mathrm{r}}$为该扰动的每步扰动迭代的差值, 在扰动迭代过程中更新微调引起的梯度偏差可以表示为:

$$\Delta\delta_{\mathrm{r}} - \Delta\delta_{\mathrm{s}} = \nabla L\left(y_{\mathrm{r}}^m\right)\cdot\left(\prod_{k=1}^{m}\left(\nabla f_{\theta^k}^k\left(y_{\mathrm{r}}^{k-1}\right) + \nabla h_{\Delta\theta_{\mathrm{r}}^k}^k\left(y_{\mathrm{r}}^{k-1}\right)\right) - \prod_{k=1}^{m}\nabla f_{\theta^k}^k\left(y_{\mathrm{r}}^{k-1}\right)\right). \tag{14}$$

为了减轻微调引起的偏差, 提高产生对抗性扰动δ_{s}的有效性, 引入了一种梯度稳健损失$\mathcal{L}_{\mathrm{GR}}$, 旨在通过基于梯度的噪声增强来缓解式(14)的偏差. 将$N(\varepsilon;\mu,\sigma^2 I)$表示为均值$\mu$和方差$\sigma^2$的各向同性高斯噪声, 其维数与$\nabla h_{\Delta\theta_{\mathrm{r}}}$相同.

$$\prod_{k=1}^{m}\left(\nabla f_{\theta^k}^k\left(y_{\mathrm{r}}^{k-1}\right) + \nabla h_{\Delta\theta_{\mathrm{r}}^k}^k\left(y_{\mathrm{r}}^{k-1}\right)\right) \Rightarrow \prod_{k=1}^{m}\left(\nabla f_{\theta^k}^k\left(y_{\mathrm{r}}^{k-1}\right) + \varepsilon^k\cdot\nabla f_{\theta^k}^k\left(y_{\mathrm{r}}^{k-1}\right)\right). \tag{15}$$

忽略式(15)中高阶不确定项, 基于噪声增强梯度的每步扰动迭代的差值$\Delta\delta_{\mathrm{s}}^*$为:

$$\Delta\delta_{\mathrm{s}}^* = \prod_{k=1}^{m}\left(\nabla f_{\theta^k}^k\left(y_{\mathrm{r}}^{k-1}\right)\right) + \sum_{i=1}^{m}\varepsilon^i\cdot\prod_{k=1}^{i-1}\left(\nabla f_{\theta^k}^k\left(y_{\mathrm{r}}^{k-1}\right)\right). \tag{16}$$

依据式(16)中的基于噪声增强梯度的每步扰动迭代的差值$\Delta\delta_{\mathrm{s}}^*$, 相应的梯度稳健损失$\mathcal{L}_{\mathrm{GR}}$定义为:

$$\mathcal{L}_{\mathrm{GR}} = \left\| f_{\theta^m}^m\left(y_{\mathrm{r}}^{m-1}\right) - f_{\theta^m}^m\left(y^{m-1}\right) + \frac{1}{m-1}\sum_{k=1}^{m-1}\varepsilon^k\cdot\left(f_{\theta^k}^k\left(y_{\mathrm{r}}^{k-1}\right) - f_{\theta^k}^k\left(y^{k-1}\right)\right)\right\|. \tag{17}$$

3.4 复杂度分析

下面给出算法 1(见 3.1 节)在单张图像一次完整攻击迭代中的时间复杂度推导, 详细符号描述参见表 1.

算法 1 中的每一步计算复杂度见表 2.

因此, 单张图像一次迭代的总时间复杂度为$O(L\cdot P\cdot C^2)$.

MI-FGSM 单次迭代仅 1 次前向加 1 次反向, 复杂度同样为$O(L\cdot P\cdot C^2)$. 本算法额外多了n个模块的特征相似度计算$O(n\cdot P\cdot C)$(低阶项)和梯度稳健损失的一次额外前向$O(L\cdot P\cdot C^2)$(与主前向同量级), 整体计算开销约为 MI-FGSM 的 2~3 倍.

Table 1 Annotations of Symbols

表 1 符号说明

符号	符号说明
N	batch size(算法 1 对单张图像循环, 故 N=1)
L	ViT 编码器 Transformer 模块总数
$P=H\times W$	单张特征图像素数
C	通道数(hidden dim)
K	分割类别数
T	攻击迭代次数
n	选出的最脆弱模块个数(算法 1 取前 n 个)
ε	扰动预算
α	步长

Table 2 Complexity of Each Step of the Algorithm

表 2 算法每步复杂度

步骤	计算内容	复杂度
前向传播	把图像送入 ViT, 得到所有 L 个模块特征	$O(L\cdot P\cdot C^2)$
计算脆弱模块相似度	对 n 个候选模块计算余弦相似度和排序	$O(n\cdot P\cdot C)$
交叉熵损失	对整个图逐像素计算概率平均	$O(P\cdot K)$
脆弱模块损失	对最脆弱的前 n 个模块求余弦相似度之和	$O(n\cdot P\cdot C)$
梯度稳健损失	额外一次前向+反向求梯度差, 加噪声	$O(L\cdot P\cdot C^2)$
反向传播	计算总损失对输入图像的梯度	$O(L\cdot P\cdot C^2)$

4 实验结果

4.1 实验设置

本节介绍实验使用的模型、数据集、对比的基线方法 baseline、对比使用的评估指标、攻击参数设置与具体实现细节.

1) 模型和数据集

对 SAMs 的下游模型进行了实验, 包括医学分割模型 Medical-SAM[5]、阴影分割模型 Shadow-SAM[7] 和伪装分割模型 Camouflaged-SAM[7]. 数据集包括突触多器官分割数据集[27], 其中包含 3 779 张腹部 CT 扫描 CT-Scans, 标注了 13 种类型的器官, ISTD 数据集[28] 包含 1 870 张阴影图像三组, COD10K 数据集[29] 包含 5 066 张伪装对象图像, CAMO 数据集[30] 包含 1 500 张伪装对象图像. 在 Medical-SAM 中, 图像编码器基于 SAM-Vit-B, 并使用 LoRA[19] 进行微调. 在 Shadow-SAM 和 Camouflaged-SAM 中, 图像编码器基于 SAM-Vit-H, 并使用适配器[31] 进行微调. 这些模型中的解码器都

是完全重新训练的.

2) 评估指标

评估医学分割采用平均 Dice 相似度得分(mDSC)和平均 Hausdorff 距离 (mHD), 评估伪装对象分割采用平均绝对误差(MAE) 和结构相似性(S_α), 以及评估阴影分割的误码率(BER).

3) 比较方法

本文提出的方法与当前的对抗性攻击方法进行比较和评估, 包括基于梯度的攻击 MI-FGSM 和 PGN[23]、基于输入增强的攻击 DMI-FGSM 和 BSR[24], 以及称为 ILPD[25] 的基于中间特征的攻击.

4) 实验细节

使用 MI-FGSM 作为我们的基本攻击方法. 攻击更新迭代T_a设置为 10 和步长$\alpha=2$, 扰动限制在$(\max-\min)\times 10$的范围内, 对于损失$\mathcal{L}_{OR}$, $\mathcal{L}_{FS}$, $\mathcal{L}_{GR}$的系数分别为1,6,1, 计算$\mathcal{L}_{FS}$时取前 4 个最小相似度的模块.

4.2 实验情况

4.2.1 在医学数据集下微调的 SAM 模型 Medical-SAM 上进行迁移攻击

表 3 展示了在医学数据集 CT-Scans 下微调的模型 Medical-SAM 上进行基于转移的对抗攻击的比较结果. 表中“无攻击”行数据反映了未受攻击时模型的原始性能. 接下来的部分呈现了在不同对抗攻击方法下模型的性能表现, 其中本文模型攻击效果最强.

Table 3 Comparative Results of SAM Fine-Tuned in Medical Dataset

表 3 在医学数据集微调的 SAM 的比较结果

攻击方法	mDSC/%	mHD
无攻击	81.88	20.64
MI-FGSM	41.61	74.3
DMI-FGSM	37.99	81.08
PGN	46.03	62.48
BSR	40.91	94.35
ILPD	33.65	65.98
本文模型	4.62	101.14

在未受攻击时, Medical-SAM 模型在 CT-Scans 数据集上达到 81.88% 的 mDSC 和 20.64 的 mHD, 这表明模型在无干扰场景下可以做到高分割精度与拓扑一致性. MI-FGSM、DMI-FGSM、PGN、BSR 和 ILPD 等方法虽能显著降低 mDSC(最低降至 33.65/%), 但 mHD 指标未同步恶化(如 ILPD 仅 65.98), 表明这些方法主要影响区域分割精度, 对拓扑结构的破坏有限. 在本文模型的攻击下, 4.62% 的 mDSC 意味着模型几乎完全丧失分割能力, 远超 ILPD 的 58.9% 降幅; 101.14 的 mHD 表明对抗样本导致分割结果与真实标注的拓扑差异极大(如器官边界断裂、空洞生成), 进一步证实了攻击在结构完整性上的破坏力, 这是现有方法未能触及的深层脆弱性. 本研究提出的脆弱模块攻击方法在攻击强度(mDSC 降幅 94.4%)和结构破坏(mHD 提升 389.6%)上均显著优于现有技术, 揭示了医疗分割模型在对抗环境下的深层脆弱性.

4.2.2 在阴影数据集下微调的 SAM 模型 Shadow-SAM 上进行迁移攻击

表 4 展示了在阴影数据集下微调的 SAM 模型 Shadow-SAM 上进行基于转移的对抗攻击的比较结果.

Table 4 Comparative Results of SAM Fine-Tuned in Shadow Dataset

表 4 在阴影数据集微调的 SAM 的比较结果

攻击方法	BER
无攻击	1.43
MI-FGSM	4.31
DMI-FGSM	4.39
PGN	5.16
BSR	5.2
ILPD	4.4
本文模型	15.28

在 ISTD 数据集上, 本文模型的攻击方法将阴影分割的 BER 从 1.43(无攻击)显著提升至 15.28, 实现了 968% 的相对增幅, 远超现有方法的最佳表现(PGN 的 5.16, 增幅 261%). 这一结果体现了攻击对阴影区域检测精度的毁灭性破坏. 传统攻击(如 MI-FGSM、DMI-FGSM)的 BER 增幅仅 200%~207%, 而本文模型通过梯度优化与特征空间扰动协同策略, 将 BER 推向 15.28 的极端值. 这种差异源于攻击利用梯度稳健损失计算梯度方向, 放大受害者模型对光照变化的敏感性, 以及对抗样本在替代模型中生成的阴影边界扰动, 成功迁移至受害者模型的特征空间, 导致其决策边界偏移.

4.2.3 在伪装数据集下微调的 SAM 模型 Camouflaged-SAM 上进行迁移攻击

表 5 展示了伪装数据集微调的 SAM 模型 Camouflaged-SAM 在基于转移对抗攻击的比较结果.

在伪装物体分割任务中, 数据与自然图像非常相似. 本文模型在 COD10K 数据集上取得了最佳性能, 并且在 CAMO 数据集上与现有的最先进方法(SOTA)相比, 性能也十分接近. 本文模型攻击在 COD10K 和

Table 5 Comparative Results of SAM Fine-Tuned in Fake Datasets

表 5 在伪装数据集微调的 SAM 的比较结果

攻击方法	COD10K 数据集		CAMO 数据集	
	S_α	MAE	S_α	MAE
无攻击	0.883	0.025	0.847	0.070
MI-FGSM	0.372	0.214	0.451	0.211
DMI-FGSM	0.435	0.134	0.442	**0.212**
PGN	0.368	0.23	0.451	0.201
BSR	0.414	0.146	0.443	0.208
ILPD	0.366	0.245	**0.417**	0.200
本文模型	**0.354**	**0.273**	0.423	0.210

注：黑体数值表示攻击效果最强的指标结果.

CAMO 两个数据集上均实现了对模型性能的显著破坏, 但表现出明显的数据集特异性. 在 COD10K 数据集上, 攻击后模型S_α从 0.883 降至 0.354(降幅 59.9%), MAE 从 0.025 升至 0.273(增幅 992%). 这表明对抗扰动通过放大颜色/纹理相似性, 使模型完全混淆目标与背景, MAE 的千倍级增长揭示模型对伪装物体边缘的感知能力崩溃, 导致分割结果出现大面积空洞或错误连接. 而在 CAMO 数据集进行攻击后, S_α从 0.847 降至 0.423(降幅 50.1%), MAE 从 0.070 升至 0.210(增幅 200%). 尽管性能下降仍显著, 但下降幅度小于 COD10K, 原因包括: CAMO 包含更多人工伪装场景(如迷彩服), 其纹理规律性较强, 对抗扰动难以完全掩盖目标特征; Camouflaged-SAM 在 CAMO 微调时可能已学习到部分对抗鲁棒性特征, 削弱了攻击效果.

4.2.4 实验结果定性分析

图 3 以数据集 CT-Scans 为例, 以原图与原标签为基准对比各攻击对医学微调 SAM 分割结果的影响, BSR 边缘误差明显; PGN 整体结构尚存; ILPD 中心偏移; DMI-FGSM 边缘与中心均错; MI-FGSM 错误分散但结构大体完整. 本文模型在所有图像上造成最严重破坏, 模型几乎无法定位目标区域, 验证了其在黑盒场景下的高效迁移与泛化能力.

4.3 消融实验

4.3.1 FS 与 GR 的消融实验

在消融实验中, 我们通过将 FS 和 GR 分别与 2 种基础攻击方法 MI-FGSM 和 PGN 结合, 验证了它们的有效性.

在医学微调 SAM 模型上的消融结果如图 4 所示, 在 MI-FGSM 基础上单独引入 GR 后, Medical-SAM 的 mDSC 从 41.61% 降至 5.75%(降幅 86.2%), mHD 从 74.28 升至 106.98(增幅 44.0%). 这一显著提升表明 GR 通过引入梯度稳健损失有效抑制了替代模型与受害者模型间的梯度域差异, 使对抗扰动在跨任务(医学→自然图像)迁移时仍能保持高攻击性. 例如, GR 模块在 PGN 实验中同样将 mDSC 从 46.03% 降至 14.75%(降幅 67.9%), 验证了其在不同梯度优化策略下的通用性. 单独使用 FS 时, MI-FGSM 的 mDSC 降至 16.84%(降幅 59.5%), mHD 升至 96.58(增幅 30.0%). 这一结果印证了 FS 通过 Transformer 脆弱模块定位实现了攻击的精准聚焦. 值得注意的是, FS 在 PGN 实验中效果稍弱(mDSC 为 17.75%), 可能因 PGN 本身依赖像素级梯度, 与模块级脆弱性分析的协同性较低. 当 FS 与 GR 联合使用时, 攻击效果呈现非线性叠加. 在 MI-FGSM+FS+GR 组合下, mDSC 降至 4.62%(降幅 88.9%), mHD 达 101.14(增幅 36.2%), 超越所有单一模块组合. 这表明 GR 的梯度稳健性与 FS 的模块脆弱性分析形成互补: GR 确保扰动在跨模型迁移时的有效性, 而 FS 则通过定位脆弱模块放大攻击效果. 这一观察结果与本文第 3 节中的分析一致. 在 PGN+FS+GR 组合下, mDSC 17.65% 虽略逊于 MI-FGSM 组合, 但 mHD 降至 70.33(优于 PGN+GR 的 89.47), 说明该组合在拓扑破坏上更具优势. 这归因于 PGN 的像素级扰动与 FS 的模块级定位结合后, 能更均匀地覆盖目标区域, 减少分割结果的碎片化.

在 Shadow-SAM 的伪装物体分割任务中, GR 与 FS 模块均展现出独立且互补的攻击增强能力, 如图 5 所示. 当单独与 MI-FGSM 结合时, BER 从 4.31 提升至 10.36(增幅 140%); 与 PGN 结合时 BER 从 5.16 提升至 8.07(增幅 56%). 这一结果表明 GR 通过引入梯度稳健损失抑制了基础模型与受害者模型间的域差异, 使对抗扰动在跨任务迁移时仍能保持高攻击性. 通过约束梯度方向一致性, 确保在伪装物体的高频纹理区域(如树叶、岩石)生成的扰动能稳定迁移至目标模型. MI-FGSM+FS 组合使 BER 提升至 7.25(增幅 68%), PGN+FS 组合提升至 9.34(增幅 81%). FS 通过 Transformer 脆弱模块定位, 针对伪装物体与背景高度相似的区域, FS 优先攻击编码器末端的纹理敏感模块, 放大局部特征混淆效果. 当 FS 与 GR 联合使用时, 攻击效果呈现显著协同效应, 在 MI-FGSM+FS+GR 组合下 BER 飙升至 15.28(较基础方法提升 255%), 较 GR 单独使用(10.36)再提升 48%, 较 FS 单独使用(7.25)提升 111%. 这表明 GR 的梯度稳健性与 FS 的模块脆弱性分析形成互补. 在 PGN+FS+GR 组合下, BER 达 17.24(较基础方法提升 234%), 较 GR 单独使

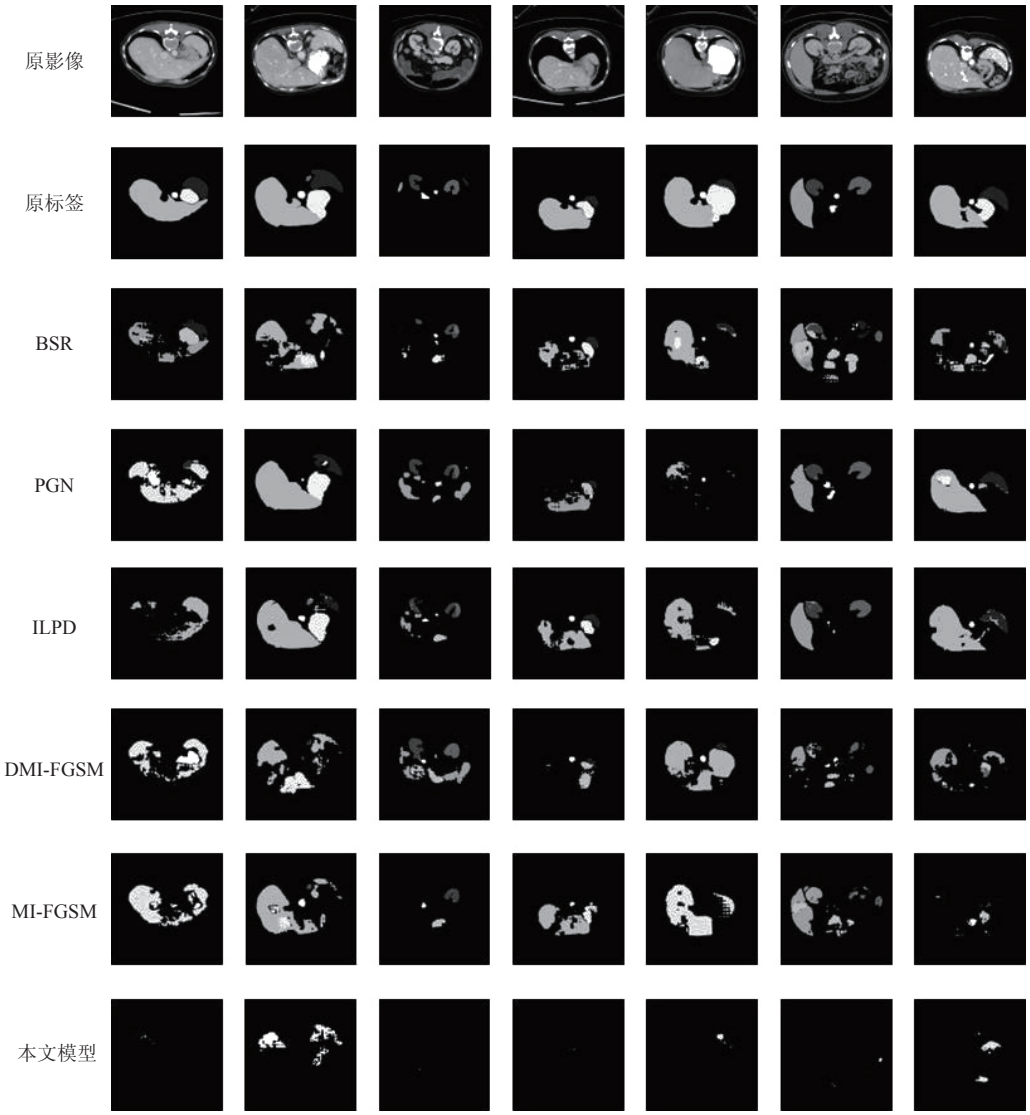
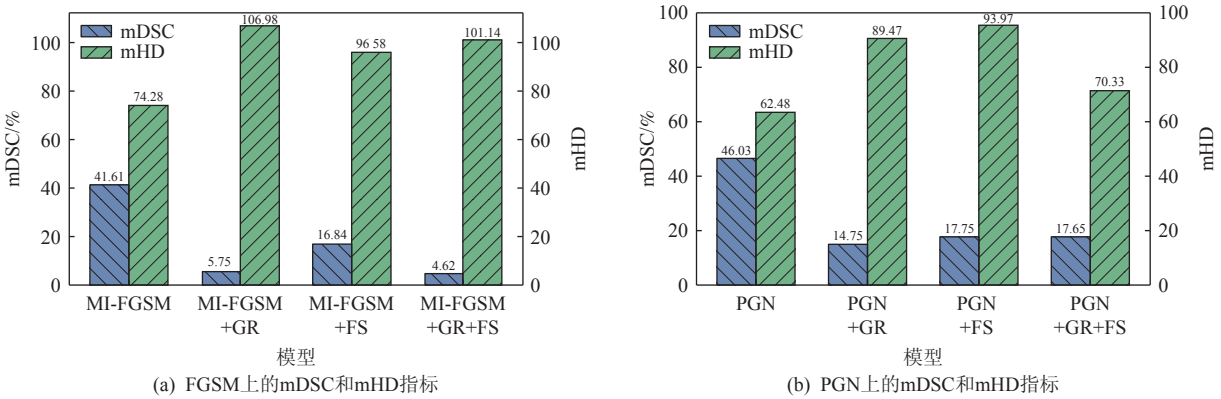


Fig. 3 Visualized attack results

图 3 可视化攻击结果

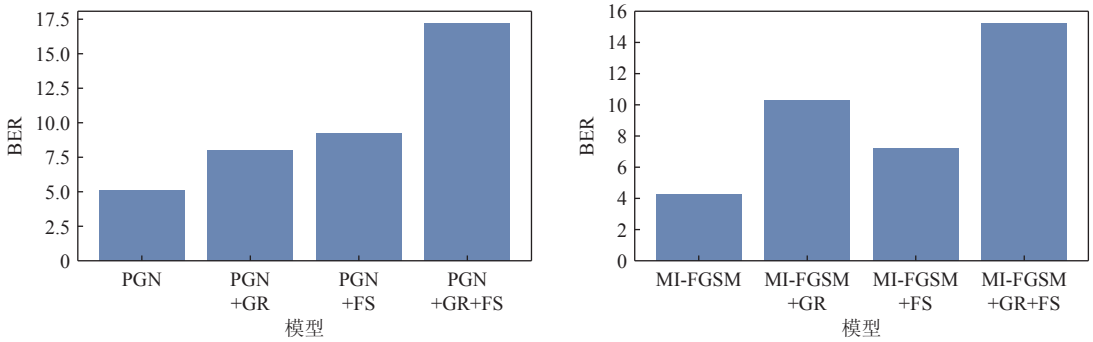


(a) FGSM上的mDSC和mHD指标 (b) PGN上的mDSC和mHD指标

Fig. 4 Experimental results of ablation of SAM model fine-tuned in medical dataset

图 4 在医学数据集微调 SAM 模型的消融实验结果

用(8.07)提升 114%, 较 FS 单独使用(9.34)提升 84%. PGN 的像素级扰动与 FS 的模块级定位结合后, 能更均匀地覆盖目标区域, 减少分割结果的碎片化, 尤其适用于伪装物体边缘模糊的场景.



(a) PGN上的BER指标　(b) FGSM上的BER指标

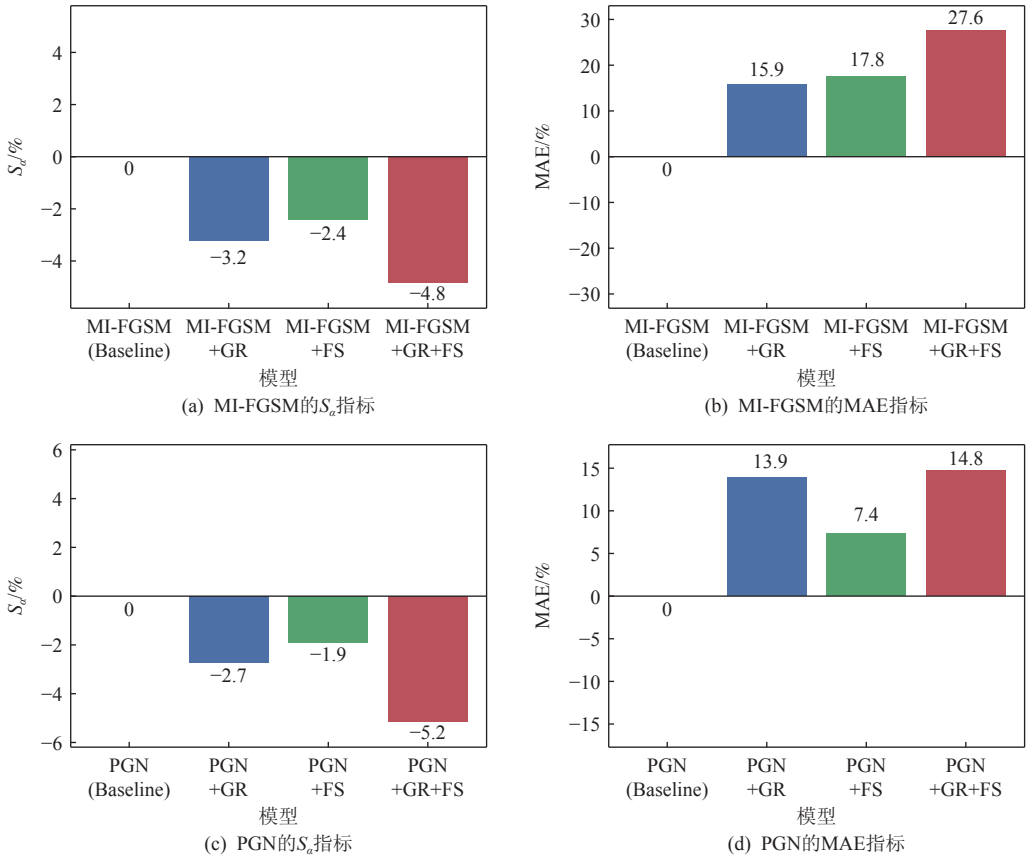
Fig. 5　Experimental results of ablation of SAM model fine-tuned in shadow dataset

图 5　在阴影数据集微调 SAM 模型的消融实验结果

在伪装微调 SAM 模型上的消融结果如图 6 所示. 在 MI-FGSM+GR 和 PGN+GR 组合下，从柱状图中可以看出，引入 GR 模块后，S_α指标明显下降，而 MAE 指标显著上升. 这表明 GR 模块通过优化梯度方向，增强了攻击在不同任务和数据集间的迁移能力，使得攻击更加有效. FS 模块的引入使得S_α指标在 MI-FGSM+FS 时下降了 2.4%，在 PGN+FS 时下降了 1.9%. 而 MAE 指标分别上升了 17.8% 和 7.4%. 实验结果表面 FS 模块通过识别并攻击模型中的脆弱 Transformer 模块，提高了攻击的针对性，从而更有效地干扰下游模型的决策. 当 FS 和 GR 模块同时结合时，S_α指标进一步下降了 4.8%，MAE 指标进一步上升. 这一结果表明，FS 和 GR 模块的协同作用能够更有效地生成对抗样本，对下游模型构成更强的威胁.

4.3.2　针对前 N 个特征相似度最小的模块进行攻击的参数选择

在本节实验中，我们对数据集 CT-Scans 中的部



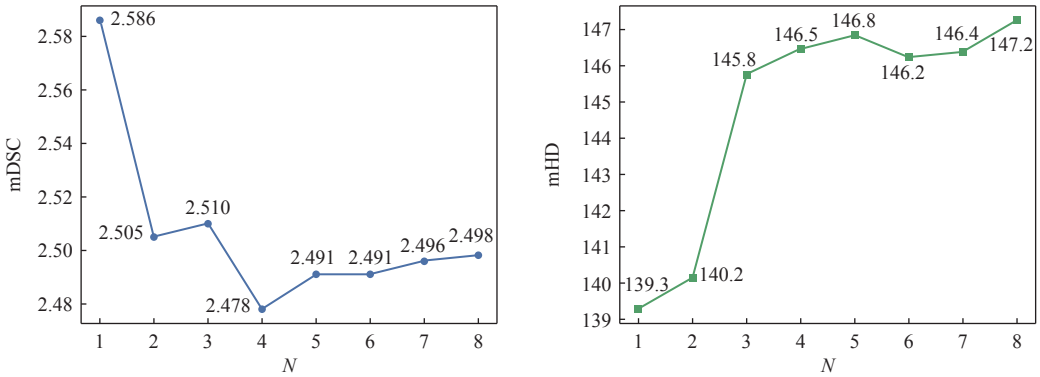
(a) MI-FGSM的S_α指标　(b) MI-FGSM的MAE指标

(c) PGN的S_α指标　(d) PGN的MAE指标

Fig. 6　Experimental results of ablation of SAM model fine-tuned in fake dataset

图 6　在伪装数据集微调 SAM 模型的消融实验结果

分样本进行了攻击，并在医学微调 SAM 模型上评估了不同攻击策略的效果. 实验的核心是针对模型中特征相似度最小的前 N 个模块（$m^k \in M$）进行攻击，通过对 CT-Scans 数据集中样本的攻击实验，我们分析了不同 N 值对模型性能的影响，并据此确定了最优的 N 值，实验结果如图 7 所示. 当 N=1 时，mDSC 从 2.586 显著下降至 2.505，随着 N 的增加，mDSC 在 N=4 时达到最低值 2.478，表明此时攻击对模型性能的破坏达到最大. 在 N=4 之后，mDSC 指标开始缓慢回升，至 N=8 时为 2.498. 这可能意味着当攻击范围扩大至更多模块时，模型的某些部分可能展现出对攻击的一定鲁棒性. mHD 从 N=1 时的 139.3 开始上升，至 N=2 时显著增加至 140.2，随着 N 的增加，mHD 持续上升，至 N=8 时达到 147.2，表明模型精度随着攻击模块数量的增加而持续下降. 在 N=4 时，mDSC 达到最低点，显示出攻击对模型性能的最大破坏. 同时，mHD 在 N=4 后增速放缓，表明增加 N 对模型精度的影响开始减弱. 选择较小的 N 值可以在保持攻击效果的同时，减少计算资源的消耗. 基于 mDSC 和 mHD 的变化趋势，我们选择 N=4 作为最终参数值.



(a) N对mDSC的影响　　(b) N对mHD的影响

Fig. 7　Experimental results under different N conditions

图 7　使用不同的 N 的情况下的实验结果

5　结　　论

深度学习模型在生产应用场景的安全性一直是个值得关注的问题，在面对对抗性攻击时，模型的脆弱性可能被恶意利用. 本文针对 SAM 下游模型的脆弱模块提出了一种对抗迁移攻击方法 FSGR，通过识别和攻击 Transformer 模块中的脆弱点，并结合梯度稳健损失增强攻击的稳健性，有效地误导了 SAM 及其下游任务微调模型. 实验结果表明，该方法在多种下游任务中表现出色，显著降低了模型性能，验证了其有效性. 然而，该方法依赖于对预训练模型的分析，尚未系统刻画下游任务在结构重参数化、数据分布偏移及微调策略差异等维度上引入的额外复杂性，从而可能无法充分泛化至全部下游变体. 此外，虽然该方法在攻击成功率上表现出色，但在实际应用中，如何平衡攻击效果与计算成本仍是一个需要解决的问题. 未来的研究可以进一步探索更多模型架构中的脆弱模块，并开发更强大的防御策略. 同时优化攻击算法以提高其在实际场景中的适用性和效率也是一个能够继续开发的挑战.

作者贡献声明：丁熠提出想法，设计了整体实验框架，并指导了研究方向；林能健负责算法思路和实验方法并撰写部分论文；蒋昀陶负责查阅并收集信息、撰写部分论文；钟宇浩为研究项目的总体规划和协调做出了贡献；曹明生对研究提出意见并指导论文修改.

参　考　文　献

[1] Ronneberger O, Fischer P, Brox T . U-Net: Convolutional networks for biomedical image segmentation[G]//LNCS 9351: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015. Cham: Springer, 2015. DOI:10.1007/978-3-319−24574−4_28

[2] Chen L C, Papandreou G, Kokkinos I, et al. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834−848

[3] He Kaiming, Gkioxari G, Dollár P, et al. Mask R-CNN[C]//Proc of the IEEE Int Conf on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2017: 2980−2988

[4] Kirillov A, Mintun E, Ravi N, et al. Segment anything[C]//Proc of the IEEE/CVF Int Conf on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2023: 3992−4003

[5] Zhang Kaidong, Liu Dong. Customized segment anything model for medical image segmentation[J]. arXiv preprint, arXiv: 2304.13785, 2023

[6] Cen Jiazhong, Zhou Zanwei, Fang Jiemin, et al. Segment anything in 3D with NERFS[C]//Proc of the Advances in Neural Information Processing Systems 36 (NeurIPS'2023). New York: Curran Associates, 2023, 26: 25971−25990

[7] Chen Tianrun, Zhu Lanyun, Ding Chaotao, et al. Sam-Adapter: Adapting segment anything in underperformed scenes[C]//Proc of the IEEE/CVF Int Conf on Computer Vision Workshops (ICCVW). Piscataway, NJ: IEEE, 2023: 3359−3367

[8] Han Dongsheng, Zhang Chaoning, Qiao Yu, et al. Segment anything model (SAM) meets glass: Mirror and transparent objects cannot be easily detected[J]. arXiv preprint, arXiv: 2305.00278, 2023

[9] Ren S, Luzi F, Lahrichi S, et al. Segment anything, from space?[C]//Proc of the IEEE/CVF Winter Conf on Applications of Computer Vision (WACV). Piscataway, NJ: IEEE, 2024: 8340−8350

[10] Ke Lei, Ye Mingqiao, Danelljan M, et al. Segment anything in high quality[C]//Proc of the Advances in Neural Information Processing Systems 37 (NeurIPS'2023). New York: Curran Associates, 2023, 36: 29914−29934

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

[12] Moosavi-Dezfooli S M, Fawzi A, Frossard P. DeepFool: A simple and accurate method to fool deep neural networks[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2016: 2574−2582

[13] Qin Zhen, Zhuang Tianming, Zhu Guosong, et al. Survey of security attack and defense strategies for artificial intelligence model[J]. Journal of Computer Research and Development, 2024, 61(10): 2627−2648 (in Chinese)
(秦臻, 庄添铭, 朱国淞, 等. 面向人工智能模型的安全攻击和防御策略综述[J]. 计算机研究与发展, 2024, 61(10): 2627−2648)

[14] Li Zituo, Sun Jianbin, Yang Kewei, et al. A review of adversarial robustness evaluation for image classification[J]. Journal of Computer Research and Development, 2022, 59(10): 2164−2189 (in Chinese)
(李自拓, 孙建彬, 杨克巍, 等. 面向图像分类的对抗鲁棒性评估综述[J]. 计算机研究与发展, 2022, 59(10): 2164−2189)

[15] Tai Jianwei, Yang Shuangning, Wang Jiajia, et al. Survey of adversarial attacks and defenses for large language models[J]. Journal of Computer Research and Development, 2025, 62(3): 563−588 (in Chinese)
(台建玮, 杨双宁, 王佳佳, 等. 大语言模型对抗性攻击与防御综述[J]. 计算机研究与发展, 2025, 62(3): 563−588)

[16] Zhang Chenshuang, Zhang Chaoning, Kang T, et al. Attack-SAM: Towards evaluating adversarial robustness of segment anything model[J]. arXiv preprint, arXiv: 2305.00866, 2023

[17] Li Bo, Xiao Haoke, Tang Lv. ASAM: Boosting segment anything model with adversarial tuning[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2024: 3699−3710

[18] Xia Song, Yang Wenhan, Yu Yi, et al. Transferable adversarial attacks on SAM and its downstream models[J]. Advances in Neural Information Processing Systems 2024, 37: 87545−87568

[19] Hu E J, Shen Yelong, Wallis P, et al. LoRA: Low-rank adaptation of large language models[C]//Proc of the Int Conf on Learning Representations (ICLR), 2022, 1(2): 3

[20] Kurakin A, Goodfellow I J, Bengio S. Adversarial Examples in the Physical World[M]//Artificial inTelligence Safty and Security. Chapman and Hall/CRC, 2018: 99−112

[21] Dong Yinpeng, Liao Fangzhou, Pang Tianyu, et al. Boosting adversarial attacks with momentum[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2018: 9185−9193

[22] Xie Cihang, Zhang Zhishuai, Zhou Yuyin, et al. Improving transferability of adversarial examples with input diversity[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2019: 2725−2734

[23] Ge Zhijin, Liu Hongying, Wang Xiaosen, et al. Boosting adversarial transferability by achieving flat local maxima[J]. Advances in Neural Information Processing Systems, 2023, 36: 70141−70161

[24] Wang Kunyu, He Xuanran, Wang Wenxuan, et al. Boosting adversarial transferability by block shuffle and rotation[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2024: 24336−24346

[25] Li Qizhang, Guo Yiwen, Zuo Wangmeng, et al. Improving adversarial transferability via intermediate-level perturbation decay[J]. Advances in Neural Information Processing Systems, 2023, 36: 32900−32912

[26] Zheng Weijie, Ma Xingjun, Huang Hanxun, et al. Downstream transfer attack: Adversarial attacks on downstream models with pre-trained vision transformers[J]. arXiv preprint, arXiv: 2408.01705, 2024

[27] Landman B, Xu Zhoubing, Igelsias J, et al. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge[C]//Proc of MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge. Munich: MICCAI, 2015, 5: 12

[28] Wang Jifeng, Li Xiang, Yang Jian. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition(CVPR). Piscataway, NJ: IEEE, 2018: 1788−1797

[29] Fan Dengping, Ji Gepeng, Sun Guolei, et al. Camouflaged object detection[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2020: 2774−2784

[30] Le T N, Nguyen T V, Nie Zhongliang, et al. Anabranch network for camouflaged object segmentation[J]. Computer Vision and Image Understanding, 2019, 184(7): 45−56

[31] Houlsby N, Giurgiu A, Jastrzebski S, et al. Parameter-efficient transfer learning for NLP[C]//Proc of the Int Conf on Machine Learning. Long Beach: PMLR, 2019: 2790−2799



Ding Yi, born in 1985. PhD, professor, PhD supervisor. His main research interests include medical image processing and computer-aided diagnosis.

丁 熠, 1985 年生. 博士, 教授, 博士生导师. 主要研究方向为医学图像处理、计算机辅助诊断.



Lin Nengjian, born in 2001. PhD candidate. His main research interests include artificial intelligence, information security, and image processing.

林能健, 2001 年生. 博士研究生. 主要研究方向为人工智能、信息安全、图像处理.



Jiang Yuntao, born in 2002. Master candidate. His main research interests include adversarial example and neural network security.

蒋昀陶, 2002 年生. 硕士研究生. 主要研究方向为对抗样本、神经网络安全.



Zhong Yuhao, born in 2000. PhD candidate. His main research interests include artificial intelligence and medical image processing.

钟宇浩, 2000 年生. 博士研究生. 主要研究方向为人工智能、医学图像处理.



Cao Mingsheng, born in 1986. PhD, associate professor, master supervisor. His main research interests include artificial intelligence, information security, and privacy protection.

曹明生, 1986 年生. 博士, 副研究员, 硕士生导师. 主要研究方向为人工智能、信息安全、隐私保护.