文章编号:1007－6735(2025)05－0590－13　　DOI: 10.13255/j.cnki.jusst.20240923001

基于区域选择的医学图像可逆对抗样本

张永东[1]，　黄　霖[2]，　周　亮[3]，　陈立范[4]，　王宏杰[4]，　孔　平[4]
（1. 上海理工大学 健康科学与工程学院，上海 200093；2. 上海理工大学 光电信息与计算机工程学院，上海 200093；
3. 上海健康医学院 协同科研中心，上海 201318；4. 上海健康医学院 文理教学部，上海 201318）

摘要：针对现有可逆对抗样本生成方法在医学图像中适用性较差的问题，提出一种基于区域选择的医学图像可逆对抗样本生成方法。具体地，通过利用生成对抗网络生成精确而微小的对抗扰动，并结合区域分割模块，选择性地将扰动限制到对诊断结果没有影响的区域，确保医学图像能拥有高攻击性能的同时仍能保持其临床可用性。此外，为了保证原始图像能够无损地得到恢复，利用可逆信息隐藏算法将添加的对抗扰动可逆地嵌入到图像之中，以保证授权机构能够合法和安全地使用。为了防止可逆嵌入的溢出问题，在进行嵌入前对这些扰动采用无损压缩技术进行压缩，以降低对抗扰动的大小，减轻因为嵌入带来的图像攻击能力和视觉质量的下降。实验结果表明，本文方法生成的可逆对抗样本平均攻击成功率、峰值信噪比和结构相似性分别为 89.27%、29.54 dB 和 0.7922，比几种经典的可逆对抗样本生成方法的性能更加优秀。

关键词：可逆对抗样本；医学图像；生成对抗网络；可逆信息隐藏；对抗扰动
中图分类号：TP 391　　**文献标志码：**A

Reversible adversarial examples for medical images based on region selection

ZHANG Yongdong[1], HUANG Lin[2], ZHOU Liang[3], CHEN Lifan[4], WANG Hongjie[4], KONG Ping[4]
(1. *School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China*; 2. *School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China*; 3. *Collaborative Innovation Center, Shanghai University of Medicine and Health Sciences, Shanghai 201318, China*; 4. *College of Arts and Science, Shanghai University of Medicine and Health Sciences, Shanghai 201318, China*)

Abstract: To address the poor applicability of existing reversible adversarial example generation methods in medical imaging, a region selection-based reversible adversarial example generation method

收稿日期: 2024-09-23
基金项目: 国家自然科学基金资助项目(61902239)；上海市浦江人才计划(22PJD031)
第一作者: 张永东(1998—)，男，硕士研究生．研究方向：医学图像处理．E-mail：223332624@st.usst.edu.cn
通信作者: 孔　平(1979—)，女，教授．研究方向：智能信息处理．E-mail：kongp@sumhs.edu.cn

引文格式: 张永东，黄霖，周亮，等．基于区域选择的医学图像可逆对抗样本[J]．上海理工大学学报，2025，47(5)：590-602.
Citation: ZHANG Yongdong, HUANG Lin, ZHOU Liang, et al. Reversible adversarial examples for medical images based on region selection[J]. Journal of University of Shanghai for Science and Technology, 2025, 47(5): 590-602.

for medical images was proposed. Specifically, a generative adversarial network was used to generate precise and subtle adversarial perturbations. Combined with a region segmentation module, these perturbations were selectively constrained to regions that have no impact on diagnostic results. This ensured that the medical images retain high attack performance while maintaining clinical usability. Furthermore, to enable lossless recovery of the original images, a reversible data hiding algorithm was employed to embed the adversarial perturbations in a reversible manner, allowing authorized institutions to legally and securely recover the original data. To prevent overflow issues during reversible embedding, lossless compression was applied to the perturbations before embedding, reducing their size and minimizing the degradation of both attack effectiveness and visual quality caused by the embedding process. Experimental results demonstrate that the proposed method achieves an average attack success rate of 89.27%, a peak signal-to-noise ratio of 29.54 dB, and a structural similarity index of 0.792 2, outperforming several classical reversible adversarial example generation methods in overall performance.

Keywords: *reversible adversarial example*; *medical image*; *generative adversarial network*; *reversible data hiding*; *adversarial perturbation*

医疗影像设备是现代医学领域中的重要医疗设备之一，通过借助于 X 射线、电磁场、超声波等与人体相互作用，把人体内部组织及器官的相关信息以影像方式予以可视化呈现，供临床医生进行诊断。医疗互联网作为互联网在医疗领域的一个重要分支，其通过将医疗影像设备直接或间接连接到医疗信息网络来提高就医的便利性和效率，极大地缓解医疗压力。但是，直接将这些带有隐私信息的医疗图像暴露在公共网络之中可能会被未经授权的机构或者个人所捕获，导致患者隐私数据的泄露和造成潜在的危险。另外，这些数据由于获取难度大、获取成本高，也被认为是一种重要的资产，未经授权的使用将会侵犯数据拥有者的合法权益。因此，当前阶段迫切需要一种医学图像隐私保护方法，以适应医疗互联网的高速发展和变革所带来的威胁与挑战。

目前常用的医学图像隐私保护方法有图像加密[1-2]、图像隐写[3-4]和可逆信息隐藏[5-6]（reversible data hiding，RDH）。一般来说，图像加密使用复杂的加密算法对图像进行加密，该算法会生成类似噪声且难以理解的图像，拥有密钥的授权接收者可以利用密钥恢复原始图像。然而，这些类似噪声的图像虽然有很高的安全性，但同时也丧失了视觉可用性。一些加密方法[7-8]企图在视觉可用性和安全性上取得平衡，但是其视觉质量仍然较差，可用性仍然较低，并且也很容易引起怀疑和攻击。与加密不同，图像隐写将病历或图像等隐私数据隐藏在载体图像中，使其不可见。由于嵌入隐私数据的含密图像与载体图像在视觉上保持高度一致，因而不会被非通信方注意到。但这种方式不仅隐藏能力有限，还会对载体图像造成不可逆的损伤，这对于敏感且珍贵的医学图像来说是不可接受的。RDH 也称为可逆水印，它不仅能够在载体图像中嵌入隐私数据，进行隐私信息传递，还能进行图像完整性认证，以检测第三方的攻击和篡改。但是，RDH 无法抵抗深度神经网络（deep neural network, DNN）的捕获和分析医学图像的隐私信息。

最近的研究表明[9]，在原始图像中添加精心设计的扰动所生成的对抗样本（adversarial example，AE）能够欺骗 DNN，使之识别出错，尤其针对医学图像效果明显[10]。由于 AE 与原始图像之间非常相似，一些研究将隐私图像转换成 AE 以规避 DNN 的恶意捕获和分析[11-12]。这些方法在需要多方共享图像的场景下具有很好的应用价值，但生成的 AE 不能无损恢复，限制了它的应用。Liu 等[13]提出了可逆对抗样本（reversible adversarial example，RAE）的概念，并结合 RDH 技术和经典的对抗攻击算法首次生成 RAE。Yin 等[14]通过可逆图像转换（reversible image transformation，RIT）技术有效地减少了需要嵌入的信息量，提升了攻击成功率（attack success rate，ASR）和视觉质量。Yin 等[15]

又提出了一种利用类映射技术和通道转换的可逆对抗攻击算法，仅针对某个通道生成扰动，并将其限制在对模型影响较大的区域，以此来进一步提升图像的视觉质量。但是以上几种RAE生成方法均是基于自然图像所提出的，没有考虑医学图像的特点。医学图像可以分为感兴趣区域(region of interest，ROI)和不感兴趣区域(region of non-interest，NROI)，大多数临床有用信息均分布在ROI。并且，相比于自然图像，医学图像更加敏感，轻微的改变可能会造成误诊等严重后果。虽然文献[16]提出一种医学图像RAE生成方法，但是通过将可视水印嵌入图像之中来加强图像数据集版权保护，尽管能拥有较高的攻击能力，却大大降低了图像的临床可用性。因此，对医学图像直接使用现有的RAE方法虽然可以防止未经授权的使用与分析，但是临床可用性仍然较低。

本文提出一种新的适用于医学图像的可逆对抗攻击算法，以防止未经授权的机构利用神经网络窃取医学图像的隐私数据，同时又保证医学图像临床可用性。为了确保不对ROI添加扰动，一个分割网络首先被用来将原始图像准确地分割为ROI和NROI。然后，一个对抗扰动生成网络被设计来为原始图像的NROI生成对抗扰动。最后，为了确保原始图像的无损恢复，RDH技术被用来将这些对抗扰动可逆地嵌入对抗样本中得到可逆对抗样本。这些可逆对抗样本可以正常地被临床医生使用，但无法被未经授权的DNN训练或分析，从而达到隐私数据保护的目的。对于授权用户，可以通过RDH技术提取嵌入的扰动以实现图像的无损恢复。

1 相关工作

1.1 对抗样本

对抗样本是一类经过精心设计生成的特殊样本，其目的在于使目标机器学习模型产生错误判断。Szegedy等[9]首次揭示了神经网络易受攻击的脆弱性，并提出了对抗样本的概念。设有一个参数为θ的分类模型为

$$\mathcal{F}_\theta : \mathbb{R}^{H\times W\times C} \to C \tag{1}$$

式中：H、W、C分别表示图像的高度、宽度以及通道数；C表示类别标签。模型$\mathcal{F}_\theta$将输入样本$\boldsymbol{X}\in\mathbb{R}^{H\times W\times C}$映射为对应的类别标签。对抗攻击的目标是在相同的模型$\mathcal{F}_\theta$下，寻找最小的扰动$\boldsymbol{\eta}$，使得干净样本$\boldsymbol{X}$与对抗样本$\boldsymbol{X}_{\mathrm{AE}}=\boldsymbol{X}+\boldsymbol{\eta}$产生不一样的输出，即

$$\begin{cases} \min \|\boldsymbol{\eta}\|_2^2 \\ \text{s.t. } \mathcal{F}_\theta(\boldsymbol{X}) \neq \mathcal{F}_\theta(\boldsymbol{X}_{\mathrm{AE}}),\ \|\boldsymbol{\eta}\|_2^2 \leqslant \varepsilon \end{cases} \tag{2}$$

式中：$\|\cdot\|_2$表示L2范数；ε表示抗动强度。

不同的分类标准可以将对抗攻击方法分成不同的类别，从模型的参数是否可以被访问的角度来看，对抗攻击可以分为白盒攻击和黑盒攻击。白盒攻击意味着攻击者可以获得全部想要的信息(包括模型结构和参数)来生成对抗样本，而对于黑盒攻击来说，只能得到预测分数和硬标签。

1.1.1 白盒攻击

a. 基于梯度的攻击。Goodfellow等[17]提出了快速梯度符号法(fast gradient sign method, FGSM)，该方法利用损失函数关于输入的梯度信息，在梯度方向上对输入施加微小扰动，从而最大化模型的损失并生成对抗样本。Kurakin等[18]提出了基本迭代法(basic iterative method, BIM)，这是FGSM的改进版本，通过减小修改幅度和增加修改次数来提升攻击成功率。Madry等[19]对输入样本添加了随机初始化噪声，并使用投影函数而不是裁剪函数来限制像素，提出了投影梯度攻击法(project gradient descent, PGD)。以PGD为例子，它通过多次迭代来调整搜寻路径，以减小损失函数$\mathcal{L}$，可以表述为

$$\boldsymbol{X}_{\mathrm{AE}}^{t+1} = \Pi_{\boldsymbol{X}+\boldsymbol{S}}\left(\boldsymbol{X}_{\mathrm{AE}}^{t} + \alpha \operatorname{sign}\left(\nabla_{\boldsymbol{X}_{\mathrm{AE}}^{t}} \mathcal{L}\left(\theta, \boldsymbol{X}_{\mathrm{AE}}^{t}, y\right)\right)\right) \tag{3}$$

式中：$\boldsymbol{X}_{\mathrm{AE}}^{t}$表示迭代到第$t$次的对抗样本；$\Pi_{\boldsymbol{X}+\boldsymbol{S}}(\cdot)$表示将对抗扰动限制在球面范围内的裁剪函数；$\boldsymbol{S}$表示随机噪声；$\alpha$代表步长；$\theta$和$y$分别表示模型参数和真实分类标签。

b. 基于优化的攻击。Szegedy等[9]首先将对抗攻击转换成优化问题并提出了L-BFGS攻击法，Carlini等[20]对其进行改进，引入了新的变量$\boldsymbol{w}$和使用tanh(·)函数来消除盒约束，可以表述为

$$\min_{\boldsymbol{w}} \left\| \frac{1}{2}(\tanh(\boldsymbol{w})+1) - \boldsymbol{X} \right\|_2^2 + h\mathcal{F}_\theta(\boldsymbol{X}_{\mathrm{AE}}) \tag{4}$$

其中

$$\mathcal{F}_\theta(\boldsymbol{X}_{\mathrm{AE}}) = \max\left(\max_{i\neq y}\{\boldsymbol{Z}(\boldsymbol{X}_{\mathrm{AE}})_i\} - \boldsymbol{Z}(\boldsymbol{X}_{\mathrm{AE}})_y - k\right) \tag{5}$$

式中：h表示平衡损失的超参数；k表示控制置信度的参数；$\boldsymbol{Z}(\cdot)_i$和$\boldsymbol{Z}(\cdot)_y$分别表示第i类别和真实类

别的 logits。

c. 基于生成式的攻击。Baluja 等[21] 首次提出了基于生成式的 ATN 攻击方法。基于此，Xiao 等[22] 提出基于生成对抗网络（generative adversarial network, GAN）的 AdvGAN 对抗攻击，通过使用鉴别器来进一步增强攻击性能。Liu 等[23] 为了提高生成器的鲁棒性，提出了通过两个阶段训练判别器，这也进一步提高了生成样本的质量。Zhang 等[24] 提出一种可恢复的对抗样本生成网络，该方法通过在传统的 GAN 网络上增加原始样本恢复模块来实现，能通过对抗样本的生成与恢复来保护社交网络用户的隐私安全。

1.1.2　黑盒攻击

a. 基于迁移的攻击。Papernot 等[25] 证明相同任务之间在不同模型上有很强的迁移性，因此，他们使用相似样本训练一个替代模型，再对替代模型使用已有的白盒攻击以生成有效的对抗样本，首次实现了目标模型黑盒场景下的攻击。Liu 等[26] 首次对大数据集和大模型的迁移性进行研究，并通过集成模型来生成对抗样本以增强其迁移性。Wang 等[27] 对重要特征进行识别，提出特征重要性感知攻击，该方法通过破坏主导模型决策的关键特征来生成对抗样本。由于这些特征在不同模型中具有较强的不变性，攻击具有较高的迁移性。

b. 基于查询的攻击。Chen 等[28] 首先提出基于查询的黑盒攻击方法，根据零阶优化和有限微分估计目标模型的梯度，利用这个梯度去生成对抗样本以攻击目标模型。然而，反复估计梯度会消耗大量的时间与计算成本。Ilyas 等[29] 提出了一种基于自然进化策略的梯度估计方法，在保持较高攻击率的同时，可以有效降低查询次数。Shi 等[30] 将基于迁移的攻击和基于决策的攻击相结合提出了一种定制化迭代和采样的黑盒对抗攻击，该方法根据每个像素的噪声敏感性进行定制化采样，进一步压缩噪声，同时对状态转移函数进行松弛以获得更高的查询效率。

1.2　可逆信息隐藏

RDH 算法是一种使用广泛的信息隐藏算法，由于其在接收端能将隐藏的数据和原始图像无损地恢复，而被广泛应用于军事和医疗等数据敏感的领域。一般来说，RDH 方法可分为以下几类：基于无损压缩的方案、基于差值扩展的方案、基于直方图移动的方案。对于基于无损压缩的方案，它通过对像素进行无损压缩为数据嵌入生成冗余空间；对于基于差值扩展的方案，它通过扩展两像素之间的差值来嵌入数据；对于基于直方图移动的方案，数据嵌入可以通过移动图像直方图中的某些分区来实现。由于 RDH 在信息嵌入时只对像素进行很小的改动（通常为 0 或±1），对分类模型的结果影响较小，现有的可逆对抗攻击均使用 RDH 来嵌入对抗扰动或者辅助信息。

1.3　可逆对抗样本

文献 [13] 首次提出可逆对抗样本的概念，其通过现有的对抗攻击算法生成对抗样本，再利用 RDH 算法将对抗扰动可逆地嵌入到对抗样本之中得到可逆对抗样本。由于 RDH 对像素的修改幅度很小，生成的可逆对抗样本在保持攻击性的同时拥有高度不可感知性。并且，如果需要可以无损地将对抗扰动去除而恢复干净的样本，这对于敏感或者珍贵图像是非常有价值的。但是，将所有扰动都嵌入到图像之中可能不大现实，因为这些扰动数据量可能超过 RDH 嵌入容量的限制。文献 [14] 提出了一种基于 RIT 的可逆对抗样本生成算法，该算法利用常见的白盒攻击算法生成中间对抗样本，将这些对抗样本看成目标图像，再利用 RIT 算法将干净样本变成目标对抗样本。虽然该方法避免了对抗扰动的嵌入，但是仍然需要利用 RDH 将大量的辅助信息嵌入到对抗样本之中以保证可逆性，并且增加了时间成本。文献 [15] 提出了一种基于 YUV 颜色空间的新型可逆攻击方法。作者首先将对抗性扰动添加到亮度通道中，然后通过 RDH 将对抗性扰动引起的亮度通道失真嵌入到色度通道中，实现可逆攻击。并且，作者还引入类激活映射技术来缩小扰动区域，从而减少嵌入数据量。

2　本文方法

2.1　应用场景

医学图像由于其获取成本较高且含有大量敏感信息，因此被视为一种重要的数据资产。RAE 通过阻止未经授权的 DNN 捕获与分析，来防止隐私泄露和数据滥用，从而保障数据所有者的权益。为了更好地保护医学图像数据集的同时保证医生临床诊断的可靠性，本文设计了一种新的 RAE 生成方法。图 1 为本文方法的应用场景，原始图

像通过可逆攻击算法生成 RAE 后，可在医疗互联网中安全传输，而不被外部未经授权的机构所捕获与分析。与此同时，对于已授权的机构，通过图像恢复操作将 RAE 恢复成原始图像以供使用。

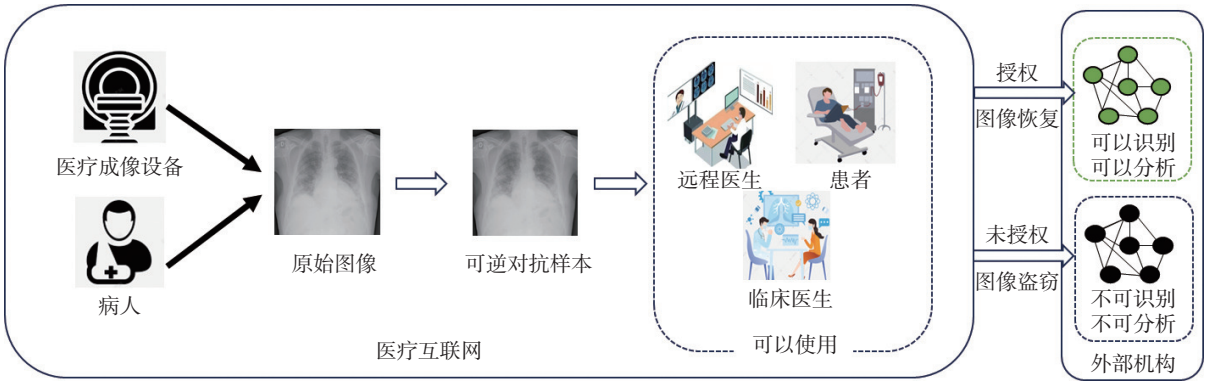

图 1　本文方法的应用场景

Fig.1　Application scenarios of the proposed method

2.2　整体框架

本文方法的框架如图 2 所示，由 3 个部分组成：对抗扰动生成、可逆对抗样本生成和原始图像的恢复。对抗扰动是由可以训练的区域扰动生成器生成的，它能产生小而特定的扰动，这些扰动将与原始图像相加生成对抗样本来误导分类模型。然后，这些扰动将通过 RDH 嵌入算法嵌入到对抗性样本中，以获得可逆对抗性样本。为了减小扰动比特，对这些扰动进行无损压缩。最后，当授权用户想要恢复原始图像时，可以根据 RDH 提取算法提取对抗样本、对抗扰动，从而恢复原始图像。

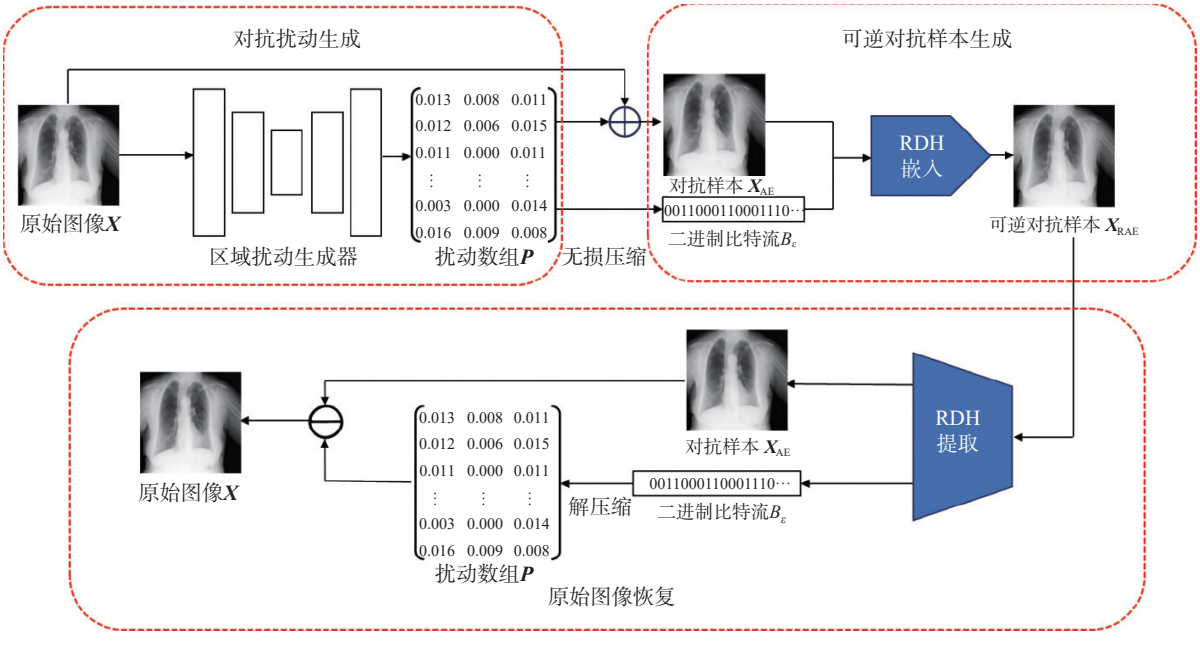

图 2　本文方法的整体框架

Fig.2　Overall framework of the proposed method

2.3　对抗扰动的生成

对抗扰动由基于不感兴趣区域的对抗扰动生成网络 NROI-AdvGAN 生成，如图 3 所示。NROI-AdvGAN 主要包括 4 个模块：区域分割模块U、扰动生成器G、鉴别器D和待攻击目标模型$\mathcal{F}$。

区域分割模块U用来分割图像以得到需要的掩码，如图 3 所示。其中，掩码的黑色区域元素被设置为 0，白色区域为 1。文献 [15] 通过类映射技术来确定分类网络关注的图像区域，以此生成掩码来生成局部区域扰动，有效地减少对抗扰动的数量。但是，在医学图像上这些区域可能是 ROI，也可能是 NROI。如图 4 所示，图中颜色较

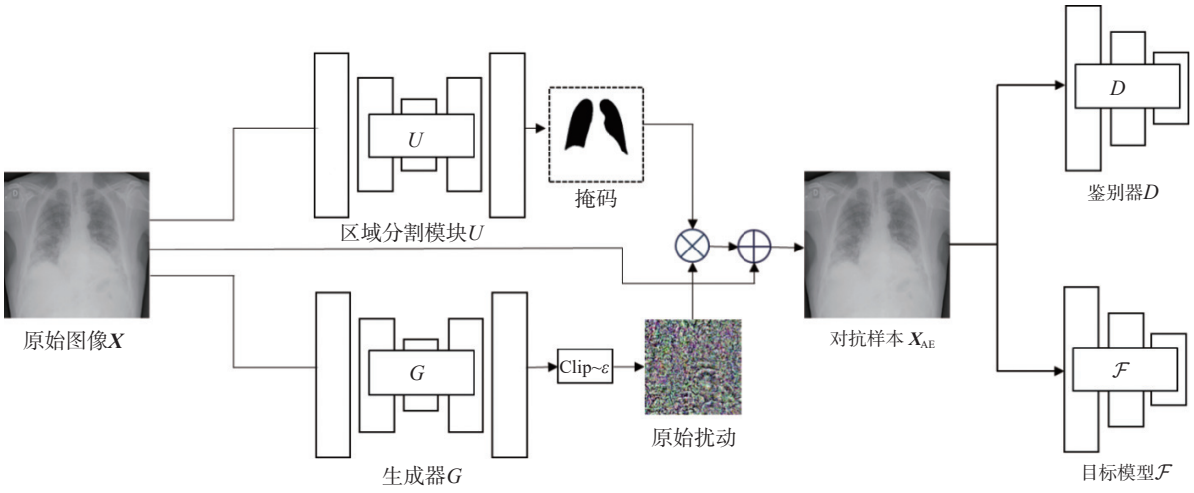

图 3　基于不感兴趣区域的对抗扰动生成网络（NROI-AdvGAN）

Fig.3　NROI-based adversarial perturbation generation network（NROI-AdvGAN）

深的区域表示模型更关注的区域，这些区域并不全是 ROI 或 NROI。为了确保 ROI 不被修改，本文方法采用 UNet 网络[31] 对原始图像进行分割，以生成掩码图像，其结构如图 5 所示。

扰动生成器G能根据输入的原始图像$\boldsymbol{X}$的特征，生成误导目标模型$\mathcal{F}$预测结果出错的对抗扰动，参考文献 [24] 的编码器−瓶颈模块−解码器结构、网络结构、卷积核大小和输出通道如图 6 所示。编码器由 4 个卷积层、归一化层和 ReLU 激活函数组成，旨在从原始图像中提取特征。相应地，解码器利用转置卷积层、归一化和激活函数将特征映射到与图像大小相同的扰动。为了增强编码器与解码器的特征共享能力，在编码器和解码器之间添加了跳跃连接和卷积注意力模块[32]，它能够通过结合通道注意力和空间注意力，为模型提供更加全面的特征提取能力。瓶颈模块被添加在编码器与解码器之间来进一步增加生成器的特征学习能力，瓶颈模块由多个残差块组成。最后，使用Clip ~ ε操作将生成的扰动限制在 $[-\varepsilon,\varepsilon]$ 之间。

鉴别器D首先通过 5 层卷积处理提取输入图像的多层次特征。每层卷积处理通常包括：卷积层、ReLU 激活函数、最大池化层。之后，将输出特征图展平成一维向量，通过一个全连接层将其转换为 512 维的特征向量。最后，再通过一个全连接层和 Sigmoid 激活函数将 512 维特征向量映射到 1 维，得到最终结果。5 个卷积核的大小分别为 7、5、3、3、3，输出通道数量为 8、16、32、64、128。鉴别器的主要作用是区分对抗性样本和干净样本。随着鉴别器鉴别能力的增加，它可以利用干净图像和对抗样本之间的区别来指导生成器G生成扰动，这样可以确保生成的对抗性样本在高维特征上尽可能地与干净样本相似。

2.4　损失函数

区域分割模块U用来得到所需要的掩码，掩码

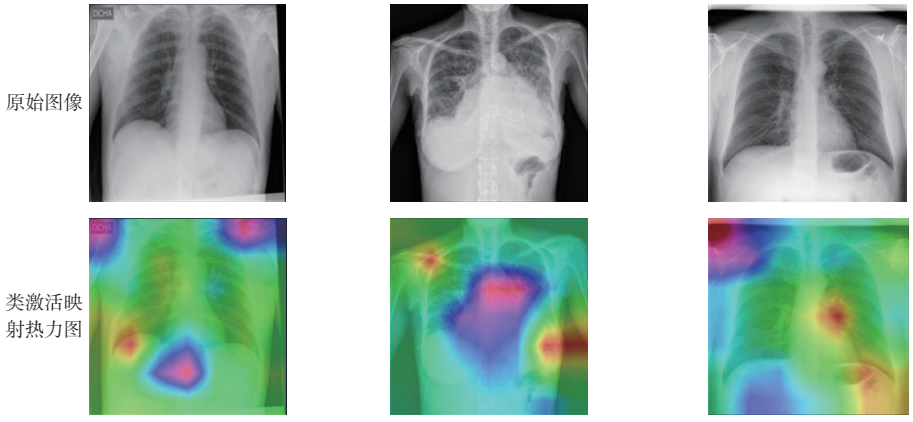

图 4　通过类激活映射得到的热力图

Fig.4　Heat maps obtained through class activation mapping

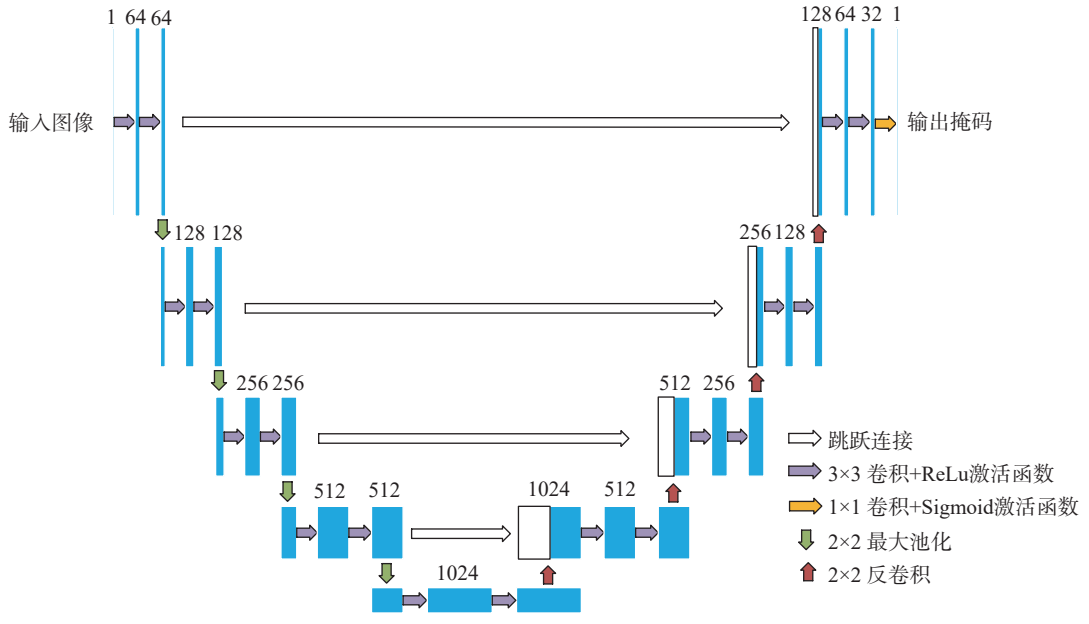

图 5 区域分割模块的网络结构

Fig.5 Network structure of regional segmentation module

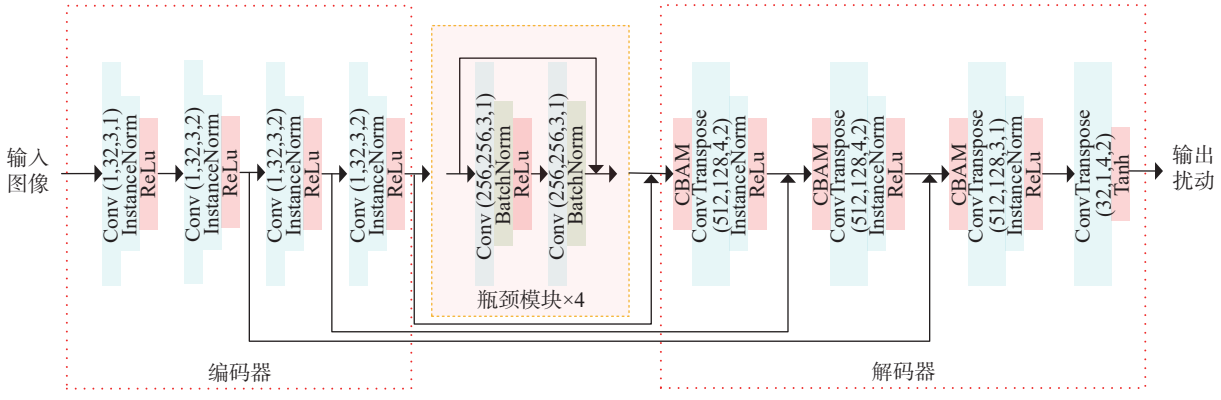

图 6 生成器的网络结构

Fig.6 Network structure of generator

元素只有 0 和 1 两个值，可以看作一种特殊的二分类任务，其损失函数为

$$\mathcal{L}_U = -\frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}[y_{h,w}\ln\widehat{y}_{h,w} + (1-y_{h,w})\ln(1-\widehat{y}_{h,w})] \tag{6}$$

式中：$y_{h,w}$表示位置(h,w)像素的标签值，$y_{h,w}\in\{0,1\}$；$\widehat{y}_{h,w}$表示通过分割网络得到的对应像素的预测概率。

生成器G的损失函数包括以下部分：为使生成器产生的对抗扰动能误导目标模型，增加攻击成功率的分类损失$\mathcal{L}_{\text{cls}}$；用于指导生成器产生以假乱真的对抗样本，以骗过鉴别器的对抗损失$\mathcal{L}_{\text{adv}}$；使对抗样本与干净样本之间视觉质量差异更小的视觉损失$\mathcal{L}_{\text{mse}}$和$\mathcal{L}_{\text{ssim}}$。生成器$G$的损失函数为

$$\mathcal{L}_G = \lambda_1\mathcal{L}_{\text{cls}} + \lambda_2\mathcal{L}_{\text{adv}} + \lambda_3\mathcal{L}_{\text{mse}} + \lambda_4(1-\mathcal{L}_{\text{ssim}}) \tag{7}$$

其中，

$$\mathcal{L}_{\text{cls}} = -\ln\left(\text{softmax}\left(\boldsymbol{Z}\left(\widehat{\boldsymbol{X}}\right)\right)_{i^*}\right) \tag{8}$$

$$\mathcal{L}_{\text{adv}} = -\ln\left(1-D\left(\widehat{\boldsymbol{X}}\right)\right) \tag{9}$$

$$\mathcal{L}_{\text{mse}} = \frac{1}{HW}\|\widehat{\boldsymbol{X}}-\boldsymbol{X}\|_2^2 \tag{10}$$

$$\mathcal{L}_{\text{ssim}} = \frac{\left(2\mu_{\boldsymbol{X}}\mu_{\widehat{\boldsymbol{X}}}+C_1\right)\left(2\sigma_{\boldsymbol{X}\widehat{\boldsymbol{X}}}+C_2\right)}{(\mu_{\boldsymbol{X}}^2+\mu_{\widehat{\boldsymbol{X}}}^2+C_1)(\sigma_{\boldsymbol{X}}^2+\sigma_{\widehat{\boldsymbol{X}}}^2+C_2)} \tag{11}$$

式中：λ_1、λ_2、λ_3、λ_4是超参数；$\widehat{\boldsymbol{X}}$表示添加对抗扰动的样本；$i^* = \underset{i\neq y}{\text{argmax}}\,\mathcal{F}(\widehat{\boldsymbol{X}})_i$表示对于样本$\widehat{\boldsymbol{X}}$模型$\mathcal{F}$输出与真实标签不相同的最大概率的类别标签，即第二大概率标签；$\text{softmax}\left(\boldsymbol{Z}\left(\widehat{\boldsymbol{X}}\right)\right)_{i^*}$表示对输出向量$\boldsymbol{Z}(\widehat{\boldsymbol{X}})$归一化后，类别$i^*$对应的概率；$\mu_{\boldsymbol{X}}$和

$\sigma_{\boldsymbol{X}}$分别表示$\boldsymbol{X}$的平均值和标准差，$\sigma_{\boldsymbol{X}\widehat{\boldsymbol{X}}}$表示$X$和$\widehat{\boldsymbol{X}}$的协方差；$C_1$和$C_2$为防止分母为 0 的常数，本实验中$C_1 = 6.502\,5$，$C_2 = 58.525\,5$。

鉴别器D的损失函数可以描述为

$$\mathcal{L}_D = \mathcal{L}_{\text{real}} + \mathcal{L}_{\text{fake}} \tag{12}$$

其中，

$$\mathcal{L}_{\text{real}} = -\ln \widehat{y} \tag{13}$$

$$\mathcal{L}_{\text{fake}} = -\ln(1-\widehat{y}) \tag{14}$$

式中：$\mathcal{L}_{\text{real}}$和$\mathcal{L}_{\text{fake}}$分别用来学习对抗样本和干净样本的特征。通过损失函数$\mathcal{L}_D$，鉴别器能够很好地区分干净样本和对抗样本。

2.5　可逆对抗样本的生成

原始图像$\boldsymbol{X}$经过区域扰动生成器生成区域扰动数组$\boldsymbol{P}$，并添加到原始图像上生成对抗样本$\boldsymbol{X}_{\text{AE}}$。然后，根据文献 [33] 的基于直方图平移的 RDH 算法，将这些扰动可逆地嵌入到对抗样本中，生成可逆对抗样本$\boldsymbol{X}_{\text{RAE}}$。一般来说，直接将这些扰动转换成二进制比特流的数量是非常庞大的，直接嵌入不仅降低图像质量还会影响攻击成功率。因此，选择基于游程编码的无损压缩算法对其进行压缩，游程编码适用于较少元素重复出现的情况，用元素出现的次数和元素本身组成的数组来代替原始数组，如“4444401112222”通过游程编码后可表示为：“5(4)1(0)3(1)4(2)”。为了降低元素种类从而增加压缩效率，对扰动数组$\boldsymbol{P}$的元素求绝对值，得到数组$\boldsymbol{P}'$。通过符号函数sign(·)来获得每个元素的符号，并得到符号数组$\boldsymbol{Q}$：

$$Q_{i,j} = \begin{cases} 1, & \text{sign}\left(P_{i,j}\right) < 0 \\ 0, & \text{sign}\left(P_{i,j}\right) \geqslant 0 \end{cases} \tag{15}$$

式中，$P_{i,j}$和$Q_{i,j}$分别代表对应数组第i行第j列的元素。

为了保证可逆性，需要将数组$\boldsymbol{P}'$与$\boldsymbol{Q}$均通过游程编码压缩并转换成二进制比特流B_ε，利用 RDH 嵌入算法分别嵌入到$\boldsymbol{X}_{\text{AE}}$中，得到$\boldsymbol{X}_{\text{RAE}}$。基于直方图平移的 RDH 嵌入算法如下所示：

$$X_{\text{RAE}}^{i,j} = \begin{cases} X_{\text{AE}}^{i,j}, & X_{\text{AE}}^{i,j} < a \\ X_{\text{AE}}^{i,j} + m, & X_{\text{AE}}^{i,j} = a \\ X_{\text{AE}}^{i,j} + 1, & X_{\text{AE}}^{i,j} > a \end{cases} \tag{16}$$

式中：$X_{\text{AE}}^{i,j}$表示图像$\boldsymbol{X}_{\text{AE}}$第$i$行第$j$列像素值；$m \in \{0,1\}$；$a$为图像$\boldsymbol{X}_{\text{AE}}$像素频次最高的像素值。

2.6　原始图像的恢复

授权用户通过相应的 RDH 提取算法，按照式(17)将嵌入的二进制比特流B_ε完整地提取出来，按照式(18)将对抗样本$\boldsymbol{X}_{\text{AE}}$进行恢复。

$$m = \begin{cases} 0, & X_{\text{RAE}}^{i,j} = a \\ 1, & X_{\text{RAE}}^{i,j} = a+1 \end{cases} \tag{17}$$

$$X_{\text{AE}}^{i,j} = \begin{cases} X_{\text{RAE}}^{i,j}, & X_{\text{RAE}}^{i,j} \leqslant a \\ X_{\text{RAE}}^{i,j} - 1, & X_{\text{RAE}}^{i,j} = a+1 \end{cases} \tag{18}$$

将提取的二进制比特流B_ε按照逆编码操作进行恢复，得到扰动数组$\boldsymbol{P}$。最后，将对抗样本$\boldsymbol{X}_{\text{AE}}$与扰动数组$\boldsymbol{P}$相减即可得到原始图像。

3　实验与分析

3.1　实验设置

数据集与目标模型：本实验选择肺部三分类数据集[34-35]作为实验数据集，选择 10 000 张来训练分类模型，并从正确分类的图像中选择 8 000 张构建训练集、700 张构建验证集和 700 张构建测试集，这些灰度图像大小被调整为 224×224。将 ResNet50 和 VGG16 作为目标攻击模型，目标模型的分类性能分别为 96.69% 和 94.46%。本文没有特别说明的情况下，分类网络均为 ResNet50。

实验细节：所有模型均在 PyTorch 框架上进行训练，CPU 为 3.60 GHz Intel(R) Core(TM) i9 10900x，GPU 为英伟达 Titan V。训练中批量大小设置为 1，初始学习率设置为 0.001，在训练过程中，学习率随着验证集的损失而调整，最小学习率设置为 0.000 1。超参数λ_1=10，λ_3=10，λ_4=1，为了加速收敛，刚开始设置λ_2=0，当验证集攻击成功率达到 80% 后，将λ_2设置为 1。

3.2　区域分割模块的性能评估

区域分割模块用来分割临床感兴趣区域，以阻止对感兴趣区域的修改。为了直观地显示分割效果，随机挑选 4 张医学图像，对其原始图像、真实标签和分割结果进行可视化展示，如图 7 所示，分割结果与真实标签基本吻合。为了客观地评估分割模型的分割效果，选择常见的 Dice 相似性系数(dice similarity coefficient，DSC)和像素准确率(pixel accuracy，PA)进行测试。DSC 计算分割图像与标签图像之间重叠区域面积之比，PA 是指被网络预测正确的像素点占图像所有像素点的

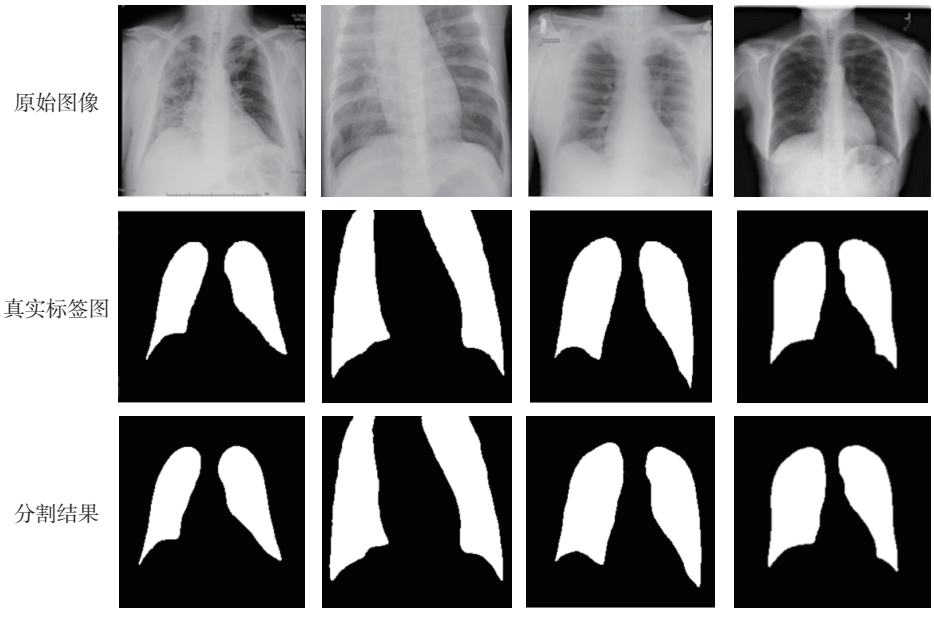

图 7 分割结果

Fig.7 Results of segmentation

比值，DSC 与 PA 越接近于 1，表明模型分割效果越好。经计算，本实验中 DSC 为 0.9568，PA 为 0.9634。

3.3 攻击性能和视觉质量评估

本文 RAE 表示使用本文方法得到的可逆对抗样本，本文 AE 表示本文方法得到的对抗样本。首先，与 3 种传统的 AE 生成方法进行了比较，包括 FGSM[17]、BIM[18]、PGD[19]，BIM 和 PGD 的迭代次数设置为 10，步长设置为 1，实验结果如表 1 所示。增加扰动幅度可以为大多数对比方法带来攻击能力的提升，但是会导致视觉质量降低。当扰动ε=8/255 时，本文方法 ASR、峰值信噪比（peak signal-to-noise ratio，PSNR）和结构相似性（structural similarity index，SSIM）分别为 96.17%、31.83 dB 和 0.8311。虽然略低于传统的 BIM 和 PGD 白盒攻击，但是传统方法是对图像中所有区域添加扰动，扰动过大会导致感兴趣区域的失真较为明显，而对于本文方法，扰动只添加在不感兴趣区域，对于实际临床的可用性几乎没有影响。

表 2 是与现有基于 RDH 和基于 RIT 的 RAE 生成方法进行对比（ε=8/255），基于 RIT 的方法中目标图像是由 FGSM 所生成。实验结果显示，本文方法在攻击成功率和视觉质量方面均表现出显著优势，ASR、PSNR、SSIM 分别达到 89.27%、29.54 dB 和 0.7922。相比之下，基于 RDH 和 RIT 的可逆对抗样本生成方法，ASR 分别提升了 0.75% 和 4.42%；PSNR 分别提升了 1.26 dB和 5.44 dB；SSIM 分别提升了 0.1821 和 0.1416。这是因为本文方法只对部分区域生成扰动，扰动数量较少，并且经过无损压缩进一步减少需要嵌入的比特数量，从而保证了较高的攻击能力和图像质量，而对比方法则是对整体图像添加扰动或者重新排序，从而导致攻击能力和视觉质量的下降。

表 1 几种 AE 生成方法在不同扰动强度下的 ASR、PSNR 以及 SSIM

Tab.1 ASR, PSNR, and SSIM of several AE generation methods under different perturbation intensities

方法		ε=1/255	ε=2/255	ε=4/255	ε=8/255
文献[17]	ASR/%	77.58	94.39	99.11	95.72
	PSNR/dB	48.21	42.20	36.20	30.20
	SSIM	0.9918	0.9685	0.8899	0.6948
文献[18]	ASR/%	95.57	100	100	100
	PSNR/dB	49.07	43.36	38.26	34.73
	SSIM	0.9934	0.9765	0.9318	0.8720
文献[19]	ASR/%	95.42	100	100	100
	PSNR/dB	48.99	43.42	39.24	33.59
	SSIM	0.9933	0.9768	0.9298	0.8277
本文（AE）	ASR/%	38.92	54.15	90.14	96.17
	PSNR/dB	48.81	43.69	37.45	31.83
	SSIM	0.9942	0.9457	0.9322	0.8311

为了更直观地进行比较，随机挑选 6 张图像进行可视化展示。如图 8 所示，尽管扰动幅度较小（ε=4/255），其他几种方法在感兴趣区域都产生

表 2　几种 RAE 生成方法的 ASR、PSNR 以及 SSIM

Tab.2　ASR, PSNR, and SSIM of several RAE generation methods

方法	ASR/%	PSNR/dB	SSIM
文献[13]	88.52	28.28	0.6101
文献[14]	84.85	24.10	0.6506
本文(RAE)	89.27	29.54	0.7922

了较为明显的失真，尤其是那些亮度较低的平滑区域，这些细微的改变可能会导致临床医生产生错误的判断。本文方法生成的 AE 未对感兴趣区域添加任何扰动，完整地保留了感兴趣区域的特征。尽管在生成 RAE 时可能会嵌入一些压缩后的扰动信息，但由于可逆数据隐藏的高度不可感知性，AE 与 RAE 在视觉上几乎不可区分，其临床可用性几乎不受影响。

图 8　几种方法生成的 AE 和 RAE 可视化图像

Fig.8　Visualization images of AE and RAE generated by several methods

3.4　计算复杂度比较

计算效率对于对抗攻击算法来说是一个很重要的指标，计算复杂度被用来评估算法的效率。首先，计算了在不同扰动下训练 NROI-AdvGAN 模型所需要的时间：当扰动ε=8/255 时，模型训练收敛时间约为 12 h；当ε=4/255 时，模型收敛时间约为 14 h。其次，还评估了本文方法和现有的 AE 及 RAE 生成方法在测试集上的单张图像平均执行时间，如表 3 所示。结果表明，无论是生成 AE 还是 RAE，本文方法的执行时间均显著优于其他对比方法。具体而言，本文方法生成 AE 的平均执行时间仅为 0.005 s，生成 RAE 的平均执行时间为 1.891 s，与基于 RDH 和基于 RIT 的方法相比，分别缩短了 39.53% 和 54.82%。

3.5　迁移攻击性

对抗样本迁移攻击性能够帮助评估对抗攻击在实际应用中的威胁程度。通过迁移攻击性实验，可以了解在不同模型之间，对抗样本的攻击

表 3 比较几种方法生成 AE 和 RAE 所消耗的平均时间

Tab.3 Comparison of the average time consumed by several methods to generate AE and RAE

s

AE					RAE		
文献[17]	文献[18]	文献[19]	文献[20]	本文（AE）	文献[13]	文献[14]	本文（RAE）
0.012	0.114	0.110	0.758	0.005	3.127	4.186	1.891

效果。实际应用中，攻击者通常无法获取目标模型的详细信息，因此，迁移性尤为关键。高迁移攻击成功率表明生成的对抗样本对未知模型也能进行有效攻击，使得其发生错误判断进而达到数据保护的目的。ResNet50 目标网络下生成的对抗样本迁移攻击成功率如表 4 所示。各种方法在相同类型的 ResNet50 上均保持了较高的迁移攻击性能，而在不同类型的 VGG16 模型上效果则相对较差。此外，尽管本文方法仅对局部区域进行攻击，其迁移性能与现有的几种先进方法基本相当。

表 4 在 ResNet50 目标网络下各方法的迁移攻击成功率

Tab.4 Transfer attack success rate of different methods on the ResNet50 target network

%

迁移模型	文献[17]	文献[18]	文献[13]	文献[14]	本文（AE）	本文（RAE）
ResNet50	95.23	96.85	94.85	90.83	93.79	90.49
VGG16	14.16	9.73	14.89	15.42	14.14	13.88

3.6 主观评价

在医学影像处理领域，主观评价是由医生通过查看影像片子得出的专业判断。这种评价方式在临床诊断中至关重要，因为医生依靠经验和专业知识能够识别出更加重要的细微差别。通过主观评价的方式，能够更好地反映实际临床环境中的诊断情况。本研究邀请 3 位专业的放射科医生对几种方法生成的 RAE 进行判断，对所有医生能有把握作出判断并正确识别的图像进行统计来计算主观评价正确率（subjective evaluation accuracy rate, SEAR）。由于数据集标签是人为标注的，可能有误，因此，事前将原始图像判断标签与实际标签不符的图像剔除，以免影响实验结果的准确性。实验结果表明，本文方法所生成的可逆对抗样本 SEAR 为 100%，显著优于基于 RDH（96.18%）和 RIT（94.23%）的方法，分别高 3.82% 和 5.77%。这一结果证明，本文方法生成的可逆对抗样本能完全保持医学图像的临床可用性，不会对实际诊断造成任何影响。

3.7 消融实验

区域分割模块用于生成掩码以将扰动添加到不感兴趣区域，从而防止感兴趣区域非期望失真。表 5 针对有无区域分割模块的情况分别进行了指标评价，实验结果显示，在没有区域分割模块的情况下，生成的 RAE 具有最高的攻击成功率，达到了 97.86%，但这同时导致了图像质量的显著下降。这是因为在缺乏模块的情况下，对抗扰动被施加到整幅图像上，降低了整体图像质量。此外，在没有区域分割模块的情况下，大量对抗扰动被添加到 ROI，严重影响了医生临床诊断结果的可靠性。相较于使用区域分割模块的情况，SEAR 下降了 5.75%，这对于需要严谨的临床诊断而言是不可接受的。

表 5 区域分割模块对 AE 的影响

Tab.5 Impact of the regional segmentation module on AE

区域分割模块	ASR/%	PSNR/dB	SSIM	SEAR/%
×	97.86	30.26	0.760 5	94.25
√	96.28	31.80	0.830 8	100.00

在 NROI-AdvGAN 框架中，鉴别器、损失函数 $\mathcal{L}_{\text{mse}}$ 和 $\mathcal{L}_{\text{ssim}}$ 旨在保证生成的对抗样本保持高攻击力的同时具备更优越的图像质量。其中，$\mathcal{L}_{\text{mse}}$ 损失函数从整体视觉内容中约束对抗性样本与原始图像之间的相似性，但其在保证图像局部细节方面可能存在不足。因此，添加了 $\mathcal{L}_{\text{ssim}}$ 损失函数来补充图像细节，从而提升图像的结构相似性。实验结果如表 6 所示，显示出鉴别器使攻击成功率从 94.25% 提升到 96.17%，同时也增强了图像的视觉质量。此外，使用 $\mathcal{L}_{\text{mse}}$ 和 $\mathcal{L}_{\text{ssim}}$ 损失函数也能进一步提高攻击成功率和图像质量。总之，通过联合使用鉴别器、损失函数 $\mathcal{L}_{\text{mse}}$ 和 $\mathcal{L}_{\text{ssim}}$ 能很好地提升生成的对抗样本攻击性能和图像视觉质量。

表 6 NROI-AdvGAN 中鉴别器、损失函数 $\mathcal{L}_{\text{mse}}$ 和 $\mathcal{L}_{\text{ssim}}$ 的消融实验

Tab.6 Ablation experiment for discriminator, loss function $\mathcal{L}_{\text{mse}}$ and $\mathcal{L}_{\text{ssim}}$ in NROI-AdvGAN

鉴别器	$\mathcal{L}_{\text{mse}}$	$\mathcal{L}_{\text{ssim}}$	ASR/%	PSNR/dB	SSIM
×	×	×	92.15	30.10	0.812 8
×	√	×	94.25	31.53	0.822 1
√	√	×	96.11	31.71	0.825 6
√	√	√	96.17	31.83	0.831 1

4　结　论

提出了一种新的适用于医学图像的可逆对抗样本生成方法，利用 ROI 和 NROI 信息分布差异性，通过所提出的 NROI-AdvGAN 扰动生成网络生成针对 NROI 的对抗扰动，并利用可逆信息隐藏技术将这些扰动可逆地嵌入到图像之中以保证原始图像的无损恢复。同时，为了缓解可逆嵌入带来的图像攻击能力下降与失真，采用了游程编码压缩，进一步缩小扰动产生的二进制比特流。值得注意的是，由于本文方法仅对 NROI 添加扰动，生成的对抗样本对于临床使用几乎没有任何影响。相比于现有的经典可逆对抗攻击算法，本文方法不仅在各种客观评价上取得了很好的结果，更重要的是在主观评价上取得了显著的优势。未来可以尝试设计新的扰动生成网络与损失函数，产生多分类模型通用对抗扰动以增强其迁移性。

参考文献：

[1] DING Y, TAN F Y, QIN Z, et al. DeepKeyGen: a deep learning-based stream cipher generator for medical image encryption and decryption[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(9): 4915−4929.

[2] KONG P, LI A, GUO D D, et al. Joint lossless compression and encryption for medical images[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(6): 4786−4799.

[3] PING P, WEI P, FU D Y, et al. IMIH: imperceptible medical image hiding for secure healthcare[J]. IEEE Transactions on Dependable and Secure Computing, 2024, 21(5): 4652−4667.

[4] PENG Y X, FU C, ZHENG Y, et al. Medical steganography: enhanced security and image quality, and new S-Q assessment[J]. Signal Processing, 2024, 223: 109546.

[5] WANG Y M, XIONG G Q, HE W G. High-capacity reversible data hiding in encrypted images based on pixel-value-ordering and histogram shifting[J]. Expert Systems with Applications, 2023, 211: 118600.

[6] GAO G Y, TONG S K, XIA Z H, et al. Reversible data hiding with automatic contrast enhancement for medical images[J]. Signal Processing, 2021, 178: 107817.

[7] ZHANG X, WANG Y F, MA J H, et al. QAPP: a quality-aware and privacy-preserving medical image release scheme[J]. Information Fusion, 2022, 88: 281−295.

[8] 周娜, 成茗, 贾孟霖, 等. 基于缩略图加密和分布式存储的医学图像隐私保护 [J]. 计算机应用, 2023, 43(10): 3149−3155.

[9] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[C]//Proceedings of the 2nd International Conference on Learning Representations. Banff: ICLR, 2014.

[10] FINLAYSON S G, BOWERS J D, ITO J, et al. Adversarial attacks on medical machine learning[J]. Science, 2019, 363(6433): 1287−1289.

[11] HU S S, LIU X G, ZHANG Y C, et al. Protecting facial privacy: generating adversarial identity masks via style-robust makeup transfer[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 14994–15003.

[12] GHAFOURIAN M, FIERREZ J, GOMEZ L F, et al. Toward face biometric de-identification using adversarial examples[C]//Proceedings of 2023 IEEE 47th Annual Computers, Software, and Applications Conference. Torino: IEEE, 2023: 723–728.

[13] LIU J Y, ZHANG W M, FUKUCHI K, et al. Unauthorized AI cannot recognize me: reversible adversarial example[J]. Pattern Recognition, 2023, 134: 109048.

[14] YIN Z X, WANG H, CHEN L, et al. Reversible adversarial attack based on reversible image transformation[EB/OL]. (2019-11-06).https://doi.org/10.48550/arXiv.1911.02360

[15] YIN Z X, CHEN L, LYU W L, et al. Reversible attack based on adversarial perturbation and reversible data hiding in YUV colorspace[J]. Pattern Recognition Letters, 2023, 166: 1−7.

[16] CHEN Z, CHAI X L, GAN Z H, et al. RAE-VWP: a reversible adversarial example-based privacy and copyright protection method of medical images for internet of medical things[J]. IEEE Internet of Things Journal, 2024, 11(11): 20757−20768.

[17] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[C]//Proceedings of the 3rd International Conference on Learning Representations. San Diego: ICLR, 2015.

[18] KURAKIN A, GOODFELLOW I J, BENGIO S. Adversarial examples in the physical world[C]// Proceedings of the 5th International Conference on Learning Representations. Toulon: ICLR, 2017.

[19] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[C]//Proceedings of the 6th International Conference on Learning Representations. Vancouver: ICLR, 2018.

[20] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]//Proceedings of 2017 IEEE Symposium on Security and Privacy. San Jose: IEEE,

2017: 39–57.

[21] BALUJA S, FISCHER I. Learning to attack: adversarial transformation networks[C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans: AAAI, 2018: 328.

[22] XIAO C W, LI B, ZHU J Y, et al. Generating adversarial examples with adversarial networks[C]//Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm: AAAI, 2018: 3905–3911.

[23] LIU X Q, HSIEH C J. Rob-GAN: generator, discriminator, and adversarial attacker[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 11226–11235.

[24] ZHANG J W, WANG J W, WANG H, et al. Self-recoverable adversarial examples: a new effective protection mechanism in social networks[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(2): 562−574.

[25] PAPERNOT N, MCDANIEL P, GOODFELLOW I. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples[EB/OL]. (2016-05-24).https://doi.org/10.48550/arXiv.1605.07277

[26] LIU Y P, CHEN X Y, LIU C, et al. Delving into transferable adversarial examples and black-box attacks[C]//Proceedings of the 5th International Conference on Learning Representations. Toulon: ICLR, 2017.

[27] WANG Z B, GUO H C, ZHANG Z F, et al. Feature importance-aware transferable adversarial attacks[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 7619–7628.

[28] CHEN P Y, ZHANG H, SHARMA Y, et al. ZOO: zeroth order optimization based black-box attacks to deep neural networks without training substitute models[C]//Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. Dallas: ACM, 2017: 15–26.

[29] ILYAS A, ENGSTROM L, ATHALYE A, et al. Black-box adversarial attacks with limited queries and information[C]//Proceedings of the 35th International Conference on Machine Learning. Stockholm: PMLR, 2018: 2137–2146.

[30] SHI Y C, HAN Y H, HU Q H, et al. Query-efficient black-box adversarial attack with customized iteration and sampling[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(2): 2226−2245.

[31] RONNEBERGER O, FISCHER P, BROX T. U-Net: convolutional networks for biomedical image segmentation[C]//Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention. Munich: Springer, 2015: 234–241.

[32] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module[C]//Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018: 3–19.

[33] NI Z C, SHI Y Q, ANSARI N, et al. Reversible data hiding[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2006, 16(3): 354−362.

[34] CHOWDHURY M E H, RAHMAN T, KHANDAKAR A, et al. Can AI help in screening viral and COVID-19 pneumonia?[J]. IEEE Access, 2020, 8: 132665−132676.

[35] RAHMAN T, KHANDAKAR A, QIBLAWEY Y, et al. Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images[J]. Computers in Biology and Medicine, 2021, 132: 104319.

（编辑：丁红艺）