

针对音频分类的跨表征可迁移对抗样本生成方法

田梓琳¹, 龙云飞¹, 田野², 张立国¹

(1. 哈尔滨工程大学计算机科学与技术学院, 黑龙江 哈尔滨 150000; 2. 西安电子科技大学杭州研究院, 浙江 杭州 310000)

摘要: 对抗样本是检测深度神经网络漏洞的重要手段。可迁移性是对抗样本的一种特有属性, 赋予其欺骗未知深度神经网络模型的能力。针对音频分类任务的迁移攻击方法引起了广泛讨论。然而, 现有的针对音频分类的攻击方法大多数只考虑单一的音频表征, 忽略了音频对抗样本从一维时域波形到二维梅尔谱图的可迁移性。以波形分类模型作为替代模型生成的对抗样本无法成功攻击以梅尔谱图作为输入的模式。针对这一问题, 提出针对音频分类的跨表征可迁移对抗样本生成方法, 通过线性组合两个以不同表征为输入的替代模型的损失训练生成网络, 使其能够生成具有跨表征和跨模型可迁移性的音频对抗样本。同时, 通过减少两个替代模型之间的预测差异, 进一步增强对抗样本跨表征可迁移性。此外, 利用生成网络输出对抗性的训练替代模型, 同步优化替代模型鲁棒性来提高对抗样本的跨模型可迁移性。在 UrbanSound8k 和 ShipsEar 数据集上的大量实验结果表明, 所提方法可以提高对抗样本的跨表征可迁移性, 并显著优于现有的迁移攻击。

关键词: 对抗样本; 迁移攻击; 音频分类攻击; 跨表征

中图分类号: TP393

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2025022

Cross-representation transferable adversarial examples generation for audio classification

TIAN Zilin¹, LONG Yunfei¹, TIAN Ye², ZHANG Ligu¹

1. College of Computer Science and Technology, Harbin Engineering University, Harbin 150000, China

2. Hangzhou Research Institute of Xidian University, Hangzhou 310000, China

Abstract: Adversarial examples have been used as an important tool for detecting vulnerabilities in deep neural networks. The unique property of transferability enables them to deceive black-box deep neural network models. Transfer-based attack methods for audio classification have recently attracted extensive discussion. However, in the

收稿日期: 2024-08-27; 修回日期: 2025-01-16

通信作者: 张立国, zhangliguo@hrbeu.edu.cn

基金项目: 国家自然科学基金 (52071111), 海南省自然科学基金 (623CXTD394), 国家资助博士后研究人员计划 (GZC20232046)

Foundation Items: The National Natural Science Foundation of China (52071111), Hainan Provincial Natural Science Foundation of China (623CXTD394), National Sponsored Postdoctoral Researcher Program (GZC20232046)

引用格式: 田梓琳, 龙云飞, 田野, 等. 针对音频分类的跨表征可迁移对抗样本生成方法[J]. 网络与信息安全学报, 2025, 11(2): 125-135.

Citation Format: TIAN Z L, LONG Y F, TIAN Y, et al. Cross-representation transferable adversarial examples generation for audio classification[J]. Chinese Journal of Network and Information Security, 2025, 11(2): 125-135.

audio domain, existing transfer-based methods overlooked the transferability of adversarial audio examples from 1D waveform to 2D Mel-spectrogram representation. It was observed that adversarial examples generated against a raw-waveform surrogate model could not successfully attack Mel-spectrogram target models. A method for generating cross-representation transferable adversarial examples for audio classification was proposed. Specifically, losses from two pre-trained surrogate models accepting different representations were linearly combined to train generators for crafting adversarial audio examples with high cross-representation and cross-model transferability. Furthermore, the predictive discrepancies between the two surrogate models were minimized to enhance cross-representation transferability. Additionally, the generator's output was used to train these surrogate models adversarially, which significantly enhanced the robustness of the surrogate models and consequently boosted the cross-model transferability of adversarial examples. Extensive experiments were carried out on the Urban Sound8k and ShipsEar datasets. The results demonstrate that this approach could improve the cross-representation transferability of adversarial examples and significantly outperform existing transfer-based attacks.

Keywords: adversarial example, transfer-based attack, audio classification attack, cross-representation

0 引言

深度神经网络 (deep neural network, DNN) 在各种计算听觉任务上表现出了卓越的性能, 尤其是在音频分类任务上, 如音乐流派分类^[1]、环境声音分类^[2]等。然而, DNN 很容易受到对抗样本的误导^[3-6]。对抗样本是指通过在原始样本上添加人眼或人耳难以察觉的特定扰动所形成的样本, 能够导致 DNN 做出错误的预测^[7]。优化扰动生成对抗样本的过程被称为对抗攻击。对抗攻击起源于图像领域, 根据模型结构的掌握程度可大致分为两类: 白盒攻击和黑盒攻击。在白盒攻击中, 目标模型的结构和参数是完全已知的。当目标模型的信息有限时, 对抗攻击通常被称为黑盒攻击。在黑盒攻击中, 迁移攻击受到广泛关注, 它不需要目标模型任何信息, 在现实中更加实用^[8-11]。迁移攻击基于替代模型生成对抗样本, 利用对抗样本的迁移性攻击其他黑盒模型。从攻击目标上可分为有目标攻击和无目标攻击, 有目标攻击是诱导目标模型做出特定的错误分类, 无目标攻击则对分类种类不做要求。尽管现有工作已经证明音频对抗攻击的可行性^[12-14], 但这些工作通常依赖为图像领域设计的攻击方法。关于高可迁移性音频对抗样本的研究尚不多见。

音频数据在不同结构的分类模型中往往被预处理为不同表征^[15-17]。例如, 一些学者设计一维卷积神经网络以一维时域波形作为输入, 进行端到端的目标识别^[18-19], 也有学者将音频的二维谱图视作常规图像, 构建二维卷积神经网络进行视觉分

析^[20-21]。不同输入模态给音频分类的对抗攻击带来了挑战。音频对抗样本的理想情况是音频不仅在被处理为一维波形时被错误分类, 而且在被转换成二维梅尔谱图后仍然被预测错误。这意味着提高音频对抗样本的可迁移性的重点在于确保其不同表征都具有欺骗分类模型能力, 本文将这一能力定义为跨表征可迁移性。而目前的实际情况是, 针对以一维波形为输入的分类模型设计的对抗样本, 在攻击以梅尔谱图为输入的分类模型时, 效果并不理想。尽管此前有方法证明了音频对抗样本跨表征迁移性的存在^[22], 但并未提出有效的增强方法, 尚未完全解决音频对抗攻击的跨表征迁移能力低问题。

基于上述分析, 本文提出了针对音频分类的跨表征可迁移对抗样本生成 (cross representation transferable adversarial examples generation, CRT-GAN) 方法。首先, 将原始音频通过生成网络输出对抗扰动, 并将其重新添加至原始音频得到对抗样本。利用两个接收不同表征的预训练替代模型作为判别网络, 同时减少对抗损失以及替代模型的预测差异来优化生成网络, 提高对抗样本的跨表征可迁移能力。其次, 与现有的迁移攻击使用固定替代模型不同, CRT-GAN 利用生成网络的输出对抗性地训练这些替代模型, 显著增强了替代模型的防御能力, 从而提高了对抗样本的跨模型可迁移性。图 1 给出 CRT-GAN 方法生成的音频对抗样本。在 UrbanSound^[23]和 ShipsEar^[24]数据集上进行的实验表明, 与现有攻击方法相比, 本文方法通过提高音频对抗样本跨表征攻击能力, 展现更高的可迁移性, 实现更高的攻击成功率。

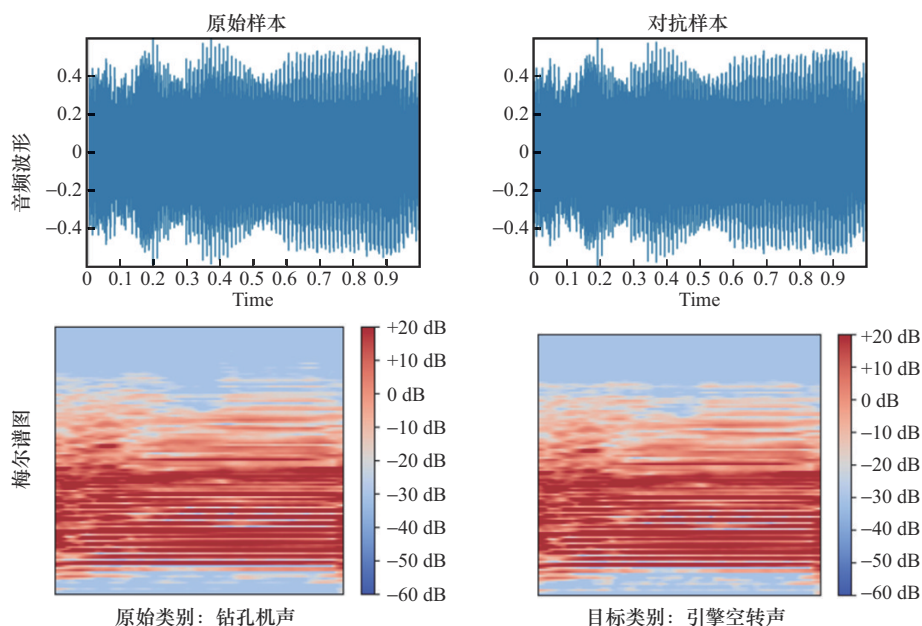


图1 CRT-GAN方法生成的音频对抗样本

Figure 1 Original example and adversarial example generated by CRT-GAN

1 相关工作

1.1 迁移攻击

自Goodfellow等^[25]提出基于梯度的攻击方法(fast gradient sign method, FGSM), Kurakin等^[26]进一步将FGSM的单步扰动生成方法细化为迭代方法,并提出迭代的FGSM攻击(iterative fast gradient sign method, I-FGSM)。Dong等^[8]利用替代模型的梯度方向迭代优化标准目标函数,提出动量迭代快速梯度符号法(momentum iterative fast gradient sign method, MI-FGSM)。这些工作在图像对抗攻击中是有效的,但在应用于音频领域时表现不佳。除了利用梯度信息之外,还可以通过输入变换来增强对抗样本的可迁移性。多样化输入法^[10]、平移不变法^[9]和尺度不变法^[27]都可以与MI-FGSM结合,进一步提高对抗样本的可迁移性。但是图像的旋转、平移等输入变换方法无法在音频上实现,因此很难提升音频对抗样本的可迁移性。

1.2 生成式攻击

生成式攻击可以看作迁移攻击的一个分支,它使用生成网络来生成对抗扰动或对抗样本。与上述迭代方法相比,生成式攻击在攻击大规模数据集时表现出更高的效率,并且可以在学习数据分布时找到对抗攻击方式。Poursaeed等^[28]首先提出了一种生成模型,用于产生扰动以使图像被

错误分类。Naseer等^[29]通过最大化源分布和目标分布之间的一致性,在有目标攻击中实现了最先进的性能。Zhang等^[30]从数据和模型的角度进一步增强了对抗样本的可迁移性。Zhao等^[11]最小化替代模型的最大模型差异(minimize the maximum model discrepancy, M3D)来训练生成网络,生成具有跨模型可迁移性的对抗样本。这些工作通过生成网络,将原始样本映射到能攻击不同的黑盒模型的对抗扰动或对抗样本。然而上述方法应用于音频分类模型时效果不佳,且所生成的对抗样本的迁移性较低。

2 背景介绍

2.1 符号定义

音频分类模型 $D: X \rightarrow Y$ 将输入音频变量 X 映射到分类变量 Y 。任意一个原始样本 $x \in X$,其真实标签为 $y \in Y$, D 可以以高置信度对 x 进行分类,即期望预测标签 $\arg \max D(x) = y$ 。对抗攻击的目标是为任意音频 x 生成对抗样本 x^{adv} ,并满足 $D(x^{\text{adv}}) \neq D(x) = y$ 以及 $\|x^{\text{adv}} - x\|_v < \epsilon$,其中, v 表示测量 x 和 x^{adv} 之间距离的范数, ϵ 表示允许的最大距离。

2.2 有目标的生成式攻击

有目标的生成式攻击可以生成指定预测类别

的对抗样本。任意原始音频 x ，生成对抗扰动的过程可以表示为 $\mu = G(x)$ ，其中， G 表示经过训练的生成网络。对抗样本表示为 $x^{\text{adv}} = x + \mu$ 。有目标的生成式攻击目的是生成 x^{adv} 来误导黑盒模型 D 将其识别成目标类别 y' ，即 $\arg\max D(x^{\text{adv}}) = y'$ 。由于黑盒模型 D 是不可访问的，现有的工作通常基于替代模型 $\hat{D}$ 生成对抗样本。由于对抗样本具有可迁移性，这些对抗样本对黑盒模型 D 有较强的攻击性。使用替代模型生成指定类别的对抗样本的目标可以描述为：

$$\begin{aligned} x^{\text{adv}} &= \arg\min \mathcal{L}(\hat{D}(x^{\text{adv}}), y') \\ \text{s.t. } \|x^{\text{adv}} - x\|_{\infty} &\leq \epsilon \end{aligned} \quad (1)$$

其中，无穷范数 ℓ_{∞} 用于约束扰动的大小， $\mathcal{L}$ 是分类损失，即交叉熵损失。

3 本文方法

3.1 跨表征可迁移存在性分析

梅尔谱图是音频数据的一种视觉表示，整合了频率和时间维度信息。它的基础是短时傅里叶变换（short-time Fourier transform, STFT），即把一段音频信号分帧、加窗，再对每一帧做快速傅里叶变换（fast Fourier transform, FFT），最后把每一帧的结果沿频域堆叠起来，得到二维信号形式。梅尔谱图为了更贴近人类感知，基于梅尔刻度对频率信号做非线性映射：

$$m(\omega) = 2595 \lg\left(1 + \frac{\omega}{700}\right) \quad (2)$$

其中， $m(\omega)$ 是给定音频频率 ω 的梅尔频率，执行对数压缩以减小动态范围。所以，将音频转换为梅尔谱图的过程可以表示为：

$$M(x) = \log m(\omega) |\text{STFT}(\omega, t, x)|^2 + \epsilon \quad (3)$$

其中， t 表示段的长度， ϵ 是为了避免取零的对数而添加的一个小常数。显然， $|\text{STFT}(\omega, t, x)|^2$ 是非线性的，所以无法进行精确反演，即从梅尔谱图无法准确获得原始的音频信号。然而，这种映射是可微的，所以可以计算 $M(x)$ 相对于 x 的梯度。

假设 $S = m(\omega) |\text{STFT}(\omega, t, x)|^2 + \epsilon$ ， x 的梯度可以用链式法，则计算如下：

$$\nabla x = \frac{\partial M(x)}{\partial S} \frac{\partial S}{\partial |\text{STFT}(\omega, t, x)|^2} \frac{\partial |\text{STFT}(\omega, t, x)|^2}{\partial x} \quad (4)$$

其中， $\frac{\partial M(x)}{\partial S}$ 和 $\frac{\partial S}{\partial |\text{STFT}(\omega, t, x)|^2}$ 均可导，所以

本文重点讨论 $\frac{\partial |\text{STFT}(\omega, t, x)|^2}{\partial x}$ 的可导性。STFT 中的映射为：

$$\text{STFT}(\omega, t, x) = \sum_n x(n) \omega(n-t) e^{-j2\pi\omega n} \quad (5)$$

其中， $\omega(n-t)$ 是平移的窗口函数。对于每一个固定的 t 和 ω ， $\text{STFT}(\omega, t, x)$ 关于 $x(n)$ 是线性变换的，且是可微的。进一步考虑：

$$|\text{STFT}(\omega, t, x)|^2 = (\Re)^2 + (\Im)^2 \quad (6)$$

其中， $\Re$ 和 $\Im$ 分别代表 $\text{STFT}(\omega, t, x)$ 中的实部和虚部。对于复数 $\text{STFT}(\omega, t, x)$ 来说，取实部、虚部、平方及相加都是标准的可微运算。因此，从梅尔谱图到时域波形的整个过程，每部分都是可微的，允许梯度向后传播，通过梯度迭代优化可以获得期望的音频样本。

3.2 跨表征的可迁移对抗攻击

本文考虑两种情况的音频分类模型： D_w 直接采用一维音频信号的原始波形 x 作为输入，而 D_m 接收经过短时傅里叶变换和梅尔滤波获得的二维梅尔谱图 $M(x)$ ， D_w 和 D_m 的输入均可表示为 x 。跨表征的可迁移对抗攻击目的是生成满足以下条件的对抗音频 x^{adv} 。

$$\begin{aligned} \arg\max_{D_w} (x^{\text{adv}}) &= D_m(x^{\text{adv}}) \neq y \\ \text{s.t. } \|x^{\text{adv}} - x\|_{\infty} &\leq \epsilon \end{aligned} \quad (7)$$

本文期望生成的对抗样本因具有表征之间的高可迁移性，从而表现出模型之间的更高可迁移性，即对抗样本可以同时攻击 D_w 和 D_m ，而之前的方法仅可以攻击 D_w 或 D_m 。为了实现上述目标，引入了两个替代模型 $\hat{D}_w$ 和 $\hat{D}_m$ 作为判别网络，对它们进行预训练，以最小化其对原始音频的分类损失。然后，可以通过以下目标获得跨表征可迁移性。

$$\min_G \mathcal{L}_a = \mathbb{E} \left[\ell_{ce}(\hat{D}_w(x^{\text{adv}}), y') + \ell_{ce}(\hat{D}_m(x^{\text{adv}}), y') \right] \quad (8)$$

$$\min_G \mathcal{L}_c = \mathbb{E} \left[d(\hat{D}_w(x^{\text{adv}}), \hat{D}_m(x^{\text{adv}})) \right] \quad (9)$$

其中, $\mathbb{E}$ 表示期望, ℓ_{ce} 表示交叉熵损失, $d(\cdot, \cdot)$ 是 $\hat{D}_w$ 和 $\hat{D}_m$ 之间的差异度量, 使用 ℓ_1 距离作为度量标准, $\mathcal{L}_a$ 优化生成网络以误导 $\hat{D}_w$ 和 $\hat{D}_m$ 将对抗样本分类到指定类别, 而 $\mathcal{L}_c$ 最小化 $\hat{D}_w$ 和 $\hat{D}_m$ 输出之间的差异。生成网络利用这些损失来训练对抗扰动, 产生具有高跨表征可迁移性的对抗样本。

为增强对抗样本的跨模型可迁移性, 使用对抗样本训练判别网络 $\hat{D}_w$ 和 $\hat{D}_m$ 。

$$\min_{\hat{D}_w, \hat{D}_m} \mathcal{L}_{at} = \mathbb{E} \left[\ell_{ce}(\hat{D}_w(x^{\text{adv}}), y) + \ell_{ce}(\hat{D}_m(x^{\text{adv}}), y) \right] \quad (10)$$

其中, $\mathcal{L}_{at}$ 优化 $\hat{D}_w$ 和 $\hat{D}_m$ 将对抗样本分类为真实类别, 从而提高判别网络鲁棒性并稳定整个训练过程。根据式(8)~式(10)的损失, 本文的跨表征生成攻击框架的优化目标如下。

$$\min_G [\mathcal{L}_a + \mathcal{L}_c] + \min_{\hat{D}_w, \hat{D}_m} \mathcal{L}_{at} \quad (11)$$

优化上述目标函数, 可以学习鲁棒的生成网络 G , 从而生成能在接收不同表征的模型中仍然有效的对抗扰动。

3.3 音频对抗样本生成框架

用于生成音频对抗扰动的生成网络 G 将原始音频 x 和目标标签 y' ($y' \neq y$) 作为输入, 然后将目标标签 y' 编码为与原始音频长度相同的特征 C , 与原始音频同元素相加。通过生成网络以创建与输入音频 x 长度相同的扰动 $\mu = G(x, y')$, 为了保证对抗样本与原始样本之间的距离保持在约束范围内, 对扰动进行约束:

$$\mu = \max(\min(\mu, -\epsilon), \epsilon) \quad (12)$$

最终, 向音频添加对抗扰动以生成对抗样本:

$$x^{\text{adv}} = \mu \oplus x \quad (13)$$

其中, $\oplus$ 表示逐元素加法。

图2给出跨表征音频对抗样本的生成框架。将原始标签进行编码与原始音频相加, 一同输入生成器 G 中, 生成器输出与原始音频相同的扰动, 扰动与原始音频相加即对抗样本, 该对抗样本具有跨表征的攻击能力。整个训练过程如算法1所示。

算法1 有目标攻击的跨表征音频对抗样本生成方法

输入 训练数据 $\{X, Y\}$, 迭代轮次 N , 预训练替代模型 $\hat{D}_w, \hat{D}_m$

输出 生成网络模型 G

1) 初始化 G

2) **for** $n = 1, \dots, N$ **do**

3) 随机采样一小批训练数据 $(x_n, y_n) \subset \{X, Y\}$ 。

4) 将所有音频 x 输入生成网络, 生成对抗扰动 $\mu = G(x_n)$, 然后构建对抗样本 $x_n^{\text{adv}} = \mu \oplus x_n$ 。

5) 将所有对抗样本 x_n^{adv} 正向输入 $\hat{D}_w, \hat{D}_m$ 。

6) 更新 G : 计算损失 $\mathcal{L}_a + \mathcal{L}_c$ 并执行向后传递以更新 G , 此过程保持 $\hat{D}_w$ 和 $\hat{D}_m$ 不更新。

7) 更新 $\hat{D}_w$ 和 $\hat{D}_m$: 计算损失 $\mathcal{L}_{at}$ 并执行向后传递以更新 $\hat{D}_w, \hat{D}_m$, 同时保持 G 不变。

8) **end for**

9) **return** G

3.4 生成网络和判别网络

本文的生成网络和判别网络的设计如下。

1) 生成网络: 采用之前生成攻击^[11,29]中引入的网络架构, 对其进行调整以适应音频输入。生成网

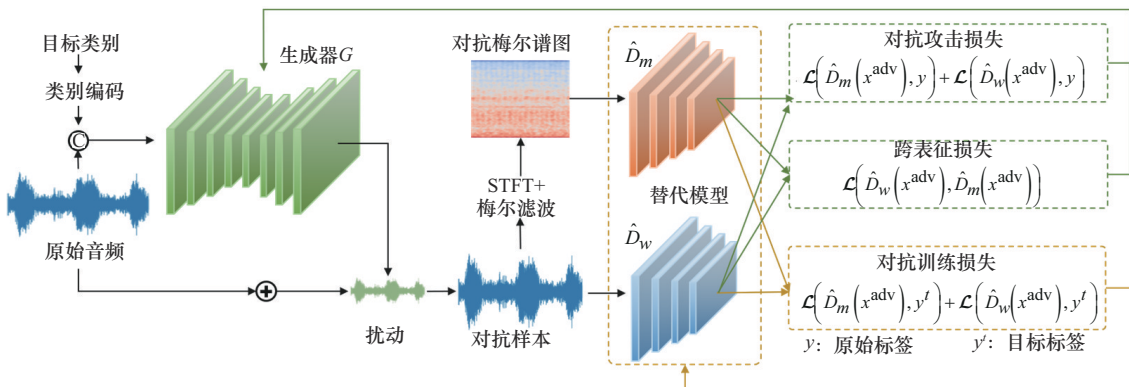


图2 跨表征音频对抗样本的生成框架

Figure 2 Architecture for generating cross-representation audio adversarial examples

络的输出是长度与输入音频大小相同的对抗扰动。

2) 判别网络: 包括一维原始波形分类模型和二维梅尔谱图分类模型。本文使用原始音频预训练波形分类模型 Res-18-1D^[31]和梅尔谱图分类模型 Res-18-2D^[31], 两者都具有 ResNet-18^[31]的模型结构。

4 实验

4.1 实验设置

1) 数据集: 本文使用两个广泛应用的音频分类数据集, 分别是用于环境声音分类的 UrbanSound8k^[23]数据集和用于水声目标识别的 ShipsEar^[24]数据集。UrbanSound8K 数据集包含 8 372 个 4 s 左右的音频片段, 分为 10 个类别 (如空调、汽车鸣笛声等)。水声信号是在海水中传播的声学信号, 一般包括海洋环境噪声、舰船辐射噪声以及海洋生物噪声。ShipsEar 数据集是经典的水下声学数据集。它包含 90 条音频, 分为 11 种船舶噪声 (如拖船和远洋客轮) 类别和 1 种海洋噪声类别, 时长从 4 s 到 10 min。相比空气中传播的声音, 水中声波波长更长, 传播损耗更小。为了便于训练, 本文将所有的音频裁剪为 1 s。从这两种数据集中分别随机选择 1 000 个可以被全部测试模型正确分类的音频样本作为原始音频。

2) 目标模型: 对于一维原始波形分类模型, 选择 ResNet-18 (Res-18-1D)^[31]、ResNet-50 (Res-50-1D)^[31]、DenseNet-121 (Dense-1D)^[32]和 EfficientNet (Efficient-1D)^[33]作为评估攻击性能的目标模型, 以上模型使用一维卷积来适应音频波形输入。对于二维梅尔谱图分类模型, 使用二维卷积结构 ResNet-18 (Res-18-2D)^[31]、ResNet-50 (Res-50-2D)^[31]、DenseNet-121 (Dense-2D)^[32]和 ShuffleNetV2 (Shuffle-2D)^[34]作为目标模型。

3) 基线: 将 CRT-GAN 与经典的迁移攻击和生成攻击进行比较, 包括 FGSM^[25]、I-FGSM^[26]、MI-FGSM^[8]和 M3D^[11], 并将这些方法根据音频进行了调整。详细细节在超参数中进行了描述, 对比方法均为有目标的攻击方法。

4) 超参数: 将最大扰动设置为 0.01 ($\mathcal{L}_\infty = 0.01$)。对于 MI-FGSM, 迭代次数 $T=10$, 衰减因子为 1.0。对于 M3D, 将其中的二维变换转换为一维变换 (调整大小和一维填充)。FGSM、

IFGSM 和 MI-FGSM 分别使用一维原始波形替代模型 (Res-18-1D) 和二维梅尔谱图替代模型 (Res-18-2D) 进行对抗样本生成。M3D 使用 Res-18-1D 和 Res-50-1D 作为判别网络。CRT-GAN 使用 Res-18-1D 和 Res-18-2D 作为替代模型。实验环境为 GeForce RTX 3090 GPU。

4.2 有目标的攻击效果

本节评估了在不同攻击目标下 CRT-GAN 的攻击成功率。本文在 ShipsEar 数据集上随机选取 1 000 个干净的音频, 分别以挖泥船、渔船、摩托艇、贻贝船、环境噪声、远洋客轮、客船轮渡、领航船、滚装船、帆船、拖网渔船和拖船为目标进行目标攻击, ShipsEar 数据集不同攻击目标的攻击成功率如表 1 所示。

当攻击模型为 Res-18-1D 时, 所有目标的攻击成功率均达到了 94% 以上。与此同时, 当所生成的对抗样本转为梅尔谱图时, 对 Res-18-2D 的攻击成功率也能达到 65% 以上。对于攻击目标摩托艇, 无论是一维原始波形目标模型还是二维梅尔谱图目标模型都有较高的攻击成功率, 平均攻击成功率分别达到了 71.5% 和 60.9%。表 1 中结果说明 CRT-GAN 可以有效地进行有目标攻击, 且所生成的对抗样本拥有跨表征可迁移性, 证明了本文方法的有效性。另外, 观察到某些类别的攻击成功率较低 (如攻击目标为拖船时), 其原因是拖船在目标模型 (如 Dense-1D) 的决策边界所规划的数据空间中处于较边缘的位置, 所以将该类别作为目标时, 成功率较其他目标类别更低。

4.3 与其他方法的对比

本节评估了不同方法攻击不同分类模型的攻击结果。图 3 中展示了不同方法生成的对抗样本的音频波形图 (第 1、3 列) 和梅尔谱图 (第 2、4 列)。原始音频为 UrbanSound8k 数据集中标签为汽车鸣笛声和儿童玩耍声的音频, 攻击的目标均为钻孔机声。

从图 3 中原始标签是汽车鸣笛声的样本可以看出, 尽管对比方法 I-FGSM 和 M3D 可以欺骗目标模型, 但是识别结果不是目标结果。虽然其他方法的梅尔谱图与原始样本差距较小, 但是并不能成功欺骗梅尔谱图为输入的目标模型。而从原始标签是儿童玩耍声的音频样本可以发现, CRT-

表1 ShipsEar数据集不同攻击目标的攻击成功率

Table 1 The attack success rates in different attack targets on ShipsEar dataset

攻击目标	一维原始波形目标模型攻击成功率				平均成功率	二维梅尔谱图目标模型攻击成功率				平均成功率
	Res-18-1D	Res-50-1D	Efficient-1D	Dense-1D		Res-18-2D	Res-50-2D	Shuffle-2D	Dense-2D	
挖泥船	100.0%	91.7%	60.5%	6.6%	64.7%	87.1%	24.5%	16.8%	12.3%	35.2%
渔船	100.0%	33.6%	25.1%	9.3%	42.0%	87.0%	14.5%	17.7%	9.8%	32.3%
摩托艇	100.0%	77.6%	56.0%	52.4%	71.5%	90.8%	60.7%	38.5%	53.5%	60.9%
贻贝船	99.1%	66.0%	56.1%	36.9%	64.5%	82.0%	43.2%	34.8%	32.3%	48.1%
环境噪声	100.0%	56.2%	47.6%	42.6%	61.6%	65.9%	52.1%	16.5%	18.2%	38.2%
远洋客轮	100.0%	41.2%	7.7%	14.6%	40.9%	83.4%	35.9%	5.3%	3.9%	32.1%
客船轮渡	99.9%	34.7%	5.9%	44.0%	46.1%	82.9%	46.9%	38.8%	42.3%	52.7%
领航船	99.5%	38.9%	52.7%	9.9%	50.3%	78.5%	45.5%	13.1%	8.8%	36.5%
滚装船	99.1%	77.4%	61.4%	41.4%	69.8%	86.9%	54.9%	56.6%	63.1%	65.4%
帆船	100.0%	37.8%	45.7%	10.1%	48.4%	73.5%	56.0%	43.8%	38.0%	52.8%
拖网渔船	94.6%	34.4%	11.2%	14.3%	38.6%	81.7%	42.7%	15.6%	39.5%	44.9%
拖船	99.7%	41.8%	23.1%	2.4%	41.8%	80.9%	30.6%	17.4%	36.8%	41.4%

GAN生成的对抗样本隐蔽性更高。无论是一维波形还是二维梅尔谱图，本文方法生成的对抗样本扰动几乎不可察觉，且可以同时欺骗音频和梅尔谱图分类模型。

另外，在表2中展示了各种方法制作的对抗样本在UrbanSound8k和ShipsEar数据集上的攻击成功率。其中，攻击目标为真实标签的下一类别（如真实标签为第一类别，目标类别为第二类别）。可以发现，目前的方法无法同时兼顾两种表征目标模型的攻击成功率。当替代模型为一维音频波形分类模型时，虽然可以在替代模型上达到较高的攻击成功率，但将所生成的对抗样本转化成梅尔谱图时，攻击成功率大幅降低。同理，替代模型为二维梅尔谱图分类模型时，在音频波形分类模型的攻击成功率也很低。

CRT-GAN所生成的对抗样本可以同时在这两种表征目标模型中保持较高的攻击成功率。特别是，本文方法生成的对抗样本，在梅尔谱图分类网络上的攻击成功率高于所有对比方法。在UrbanSound8k数据集上，平均攻击成功率与FGSM、I-FGSM、MI-FGSM（替代模型为Res-18-1D）和M3D相比，分别提高了31.5%、31.3%、31.8%和18.1%；在ShipsEar数据集上，比FGSM、I-FGSM、MI-FGSM（替代模型为Res-18-2D）分别提高了29.4%、11.0%、6.6%和

21.9%。这些结果表明了本文方法能有效提高对抗样本的跨模型和跨表征可迁移性。

4.4 消融实验

1) 特征C：为了更好地理解特征C对跨表征迁移性的影响，本文在ShipsEar数据集上进行了消融实验。特征C由目标标签 y' 编码形成，它与原始音频长度相同。表3给出不同训练情况的消融实验结果。实验中攻击目标为真实标签的下一个类别。实验结果表明，无论是在一维还是二维分类模型上，特征C的使用明显提升了攻击成功率，增强了对抗样本的跨模型和跨表征可迁移性。

2) 替代模型：为了进一步验证本文方法的有效性，使用不同的替代模型作为生成网络G的判别网络进行训练。表4给出不同替代模型的消融实验结果。使用UrbanSound数据集，且攻击目标同样为真实标签的下一个类别。由表4可知，当目标模型与判别网络结构相同时，攻击成功率更高。同时，生成网络 G_1 和 G_2 生成的对抗样本都拥有跨模型和跨表征的可迁移能力。

3) 扰动 ϵ ：图4展示了不同扰动大小对攻击成功率的影响。由图4中可知，扰动越大，攻击成功率越高，为了兼顾隐蔽性，本文将最大扰动设为0.01。

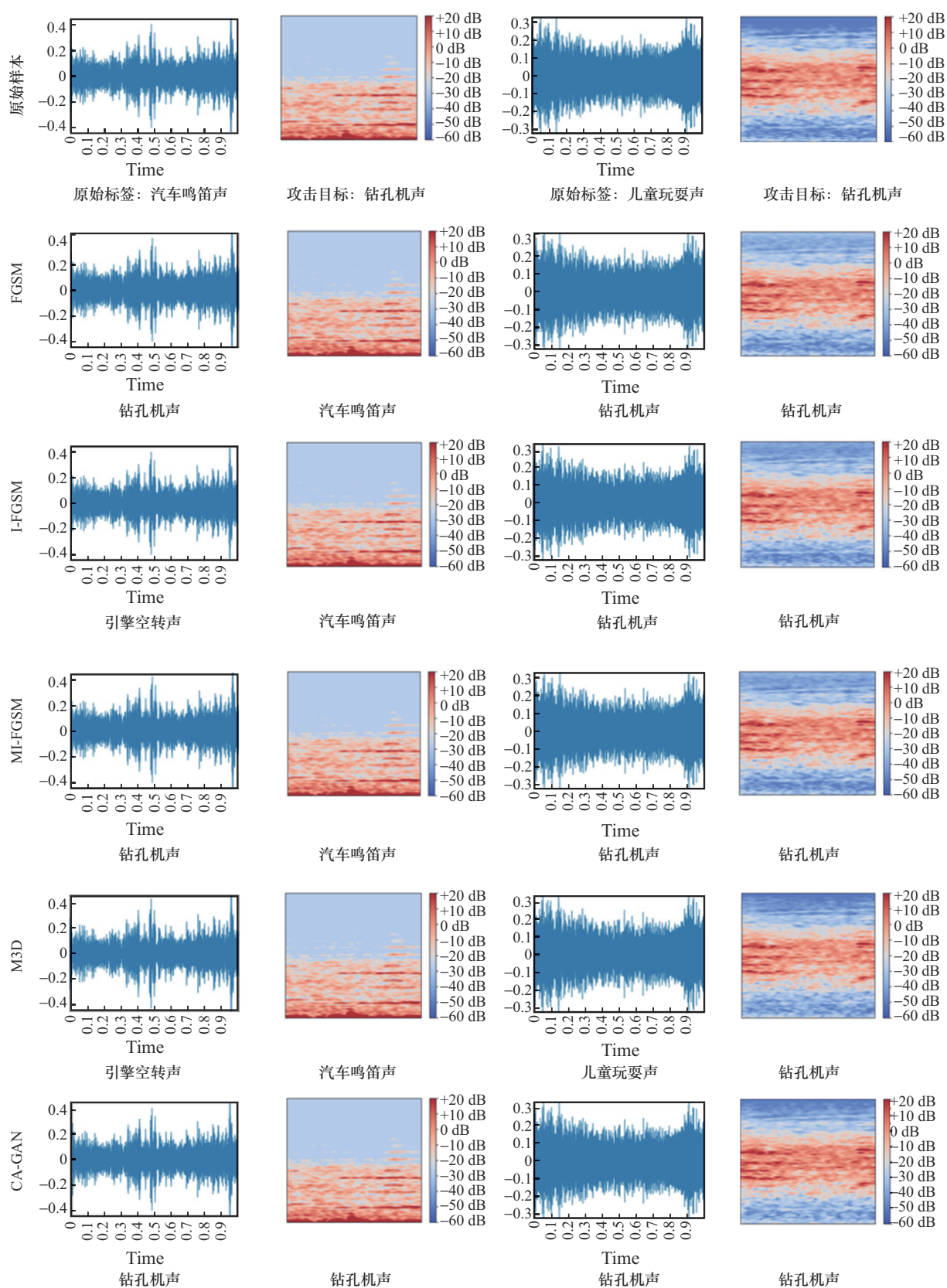


图3 不同方法生成的音频对抗样本的波形图和梅尔谱图

Figure 3 Waveform and mel-spectrogram of audio adversarial examples generated by different methods.

表2 各种方法制作的对抗样本在UrbanSound8k和ShipsEar数据集上的攻击成功率

Table 2 The attack success rates of adversarial examples crafted by various method on the UrbanSound8k and ShipsEar datasets

数据集	替代模型	攻击方法	音频波形分类模型攻击成功率				平均成功率	梅尔谱图分类模型攻击成功率				平均成功率
			Res-18-1D	Res-50-1D	Efficient-1D	Dense-1D		Res-18-2D	Res-50-2D	Shuffle-2D	Dense-2D	
Urban-Sound8k	Res-18-1D	FGSM	21.6%	11.7%	7.4%	8.1%	12.2%	1.8%	1.8%	2.8%	2.6%	2.3%
		I-FGSM	84.8%	11.7%	6.0%	9.8%	28.1%	2.1%	1.8%	3.3%	2.6%	2.5%
		MI-FGSM	92.5%	9.7%	7.4%	8.6%	29.6%	1.5%	1.9%	2.5%	2.2%	2.0%
	Res-18-2D	FGSM	7.9%	12.6%	5.4%	7.2%	8.3%	13.5%	7.9%	6.6%	3.8%	8.0%
		I-FGSM	4.7%	7.0%	4.4%	5.0%	5.3%	87.1%	31.6%	5.9%	5.2%	32.4%
		MI-FGSM	5.9%	9.6%	4.3%	4.0%	6.0%	72.4%	34.1%	5.9%	5.4%	29.5%
	Res-18-1D Res-50-1D	M3D	73.5%	17.6%	12.5%	11.3%	28.7%	33.3%	12.1%	9.6%	7.8%	15.7%
	Res-18-1D Res-18-2D	CRT-GAN	100.0%	21.1%	14.1%	11.4%	36.7%	81.5%	23.3%	17.7%	12.5%	33.8%
	Res-18-1D	FGSM	11.0%	1.0%	1.4%	3.2%	4.15%	0.3%	0.1%	0.5%	0.3%	0.3%
		I-FGSM	66.5%	7.1%	1.8%	7.8%	20.8%	0.1%	0.0%	0.8%	0.2%	0.3%
		MI-FGSM	95.6%	6.5%	1.4%	7.4%	27.7%	0.0%	0.0%	0.7%	0.5%	0.3%
ShipsEar	Res-18-2D	FGSM	0.5%	0.7%	1.0%	4.8%	1.75%	20.5%	0.8%	1.6%	7.3%	7.6%
		I-FGSM	1.5%	1.9%	1.2%	5.6%	2.55%	77.7%	16.0%	2.9%	7.2%	26.0%
		MI-FGSM	2.6%	3.4%	5.2%	5.5%	4.2%	87.3%	20.9%	3.9%	9.5%	30.4%
	Res-18-1D Res-50-1D	M3D	63.9%	27.8%	21.1%	12.1%	31.2%	28.9%	13.4%	9.9%	8.1%	15.1%
	Res-18-1D Res-18-2D	CRT-GAN	96.9%	45.1%	33.9%	17.8%	48.4%	77.4%	32.2%	19.8%	18.4%	37.0%

表3 不同训练情况的消融实验结果

Table 3 The result of ablation study in different training case

特征 C	一维音频波形目标模型				二维梅尔谱图目标模型			
	Res-18-1D	Res-50-1D	Efficient-1D	Dense-1D	Res-18-2D	Res-50-2D	Shuffle-2D	Dense-2D
添加	96.9	45.1	33.9	17.8	77.4	32.2	19.8	18.4
不添加	87.3	33.2	27.1	12.4	66.1	29.3	14.5	14.7

表4 不同替代模型的消融实验结果

Table 4 The result of ablation study for different surrogate models

替代模型	一维音频波形目标模型				二维梅尔谱图目标模型			
	Res-18-1D	Res-50-1D	Efficient-1D	Dense-1D	Res-18-2D	Res-50-2D	Shuffle-2D	Dense-2D
Res-18-1D&Res-18-2D	96.9	45.1	33.9	17.8	77.4	32.2	19.8	18.4
Dense-1D&Dense-2D	23.2	23.9	17.4	96.5	15.8	14.3	17.5	98.6

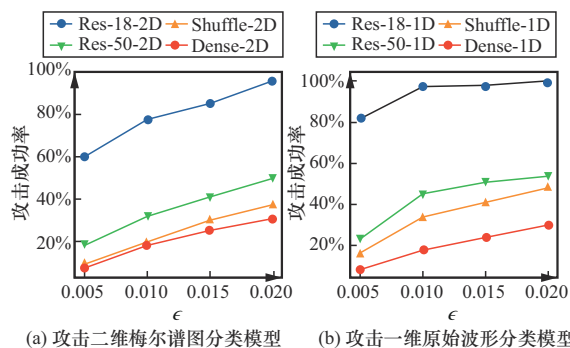


图4 不同扰动大小对攻击成功率的影响

Figure 4 The impact of different perturbation on attack success rate

5 结束语

本文提出了针对音频分类任务的跨表征可迁移对抗样本生成方法，用于生成具有跨表征迁移能力的音频对抗样本。跨表征可迁移性意味着所生成的对抗样本在一维原始波形表征目标模型中被错误分类，且在迁移到二维梅尔谱图表征后仍然被错误分类。因此，使用两个预训练替代模型（分别使用原始波形和梅尔谱图两种表征作为输入）来训练生成网络，同时减少对抗损失以及替

代模型的预测差异来优化生成网络, 提高对抗样本的跨表征可迁移能力。在 UrbanSound8k 数据集和 ShipsEar 数据集上的大量实验结果表明, 与现有的迁移攻击相比, 本文方法在一维音频波形模型和二维梅尔谱图模型上都拥有更高的攻击成功率, 增强了音频对抗样本的可迁移性。下一步研究, 拟从生成器架构优化、频域特征解耦等方面, 降低生成对抗音频扰动所需的计算资源消耗, 使其更适用于实际场景。

参考文献:

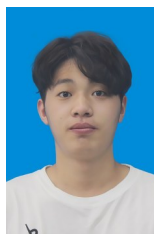
- [1] PRABHAKAR S K, LEE S W. Holistic approaches to music genre classification using efficient transfer and deep learning techniques[J]. Expert Systems with Applications, 2023, 211: 118636.
- [2] DEMIR F, ABDULLAH D, SENGUR A. A new deep CNN model for environmental sound classification[J]. IEEE Access, 2020, 8: 66529-66537.
- [3] 黄诗瑀, 叶锋, 黄添强, 等. 人脸伪造与检测中的对抗攻防综述[J]. 网络与信息安全学报, 2023, 9(04): 1-15.
HUANG S Y, YE F, HUANG T Q, et al. Survey on adversarial attacks and defense of face forgery and detection[J]. Chinese Journal of Network and Information Security, 2023, 9(4): 1-15.
- [4] 林点, 潘理, 易平. 面向图像识别的卷积神经网络鲁棒性研究进展[J]. 网络与信息安全学报, 2022, 8(3): 111-122.
LIN D, PAN L, YI P. Research on the robustness of convolutional neural networks in image recognition[J]. Chinese Journal of Network and Information Security, 2022, 8(3): 111-122.
- [5] 刘西蒙, 谢乐辉, 王耀鹏, 等. 深度学习中的对抗攻击与防御[J]. 网络与信息安全学报, 2020, 6(5): 36-53.
LIU X M, XIE L H, WANG Y P, et al. Adversarial attacks and defenses in deep learning[J]. Chinese Journal of Network and Information Security, 2020, 6(5): 36-53.
- [6] 陈晋音, 吴长安, 郑海斌. 基于 softmax 激活变换的对抗防御方法[J]. 网络与信息安全学报, 2022, 8(2): 48-63.
CHEN J Y, WU C A, ZHENG H B. Novel defense based on softmax activation transformation[J]. Chinese Journal of Network and Information Security, 2022, 8(2): 48-63.
- [7] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[J]. arXiv preprint arXiv:1312.6199, 2013.
- [8] DONG Y P, LIAO F Z, PANG T Y, et al. Boosting adversarial attacks with momentum[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 9185-9193.
- [9] DONG Y P, PANG T Y, SU H, et al. Evading defenses to transferable adversarial examples by translation-invariant attacks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 4312-4321.
- [10] XIE C H, ZHANG Z S, ZHOU Y X, et al. Improving transferability of adversarial examples with input diversity[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 2730-2739.
- [11] ZHAO A Q, CHU T, LIU Y H, et al. Minimizing maximum model discrepancy for transferable black-box targeted attacks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 8153-8162.
- [12] 何媛媛, 胡淼, 彭祖权, 等. 语音识别对抗攻击方法综述[J]. 华中科技大学学报(自然科学版), 2023, 51(2): 10-18.
HE Y Y, HU M, PENG Z Q, et al. Survey of adversarial attacks on speech recognition[J]. Journal of Huazhong University of Science and Technology (Natural Science Edition), 2023, 51(2): 10-18.
- [13] KIM H, PARK J, LEE J. Generating transferable adversarial examples for speech classification[J]. Pattern Recognition, 2023, 137: 109286.
- [14] 李珠海, 郭武. 基于自注意力机制的音频对抗样本生成方法[J]. 数据采集与处理, 2024, 39(2): 416-423.
LI Z H, GUO W. Audio adversarial examples generation method based on self-attention mechanism [J]. Journal of Data Acquisition & Processing, 2024, 39(2): 416-423.
- [15] LI Y, CAO W, XIE W, et al. Few-shot class-incremental audio classification using dynamically expanded classifier with self-attention modified prototypes[J]. IEEE Transactions on Multimedia, 2023, 26: 1346-1360.
- [16] ASHRAF M, ABID F, DIN I U, et al. A hybrid CNN and RNN variant model for music classification[J]. Applied Sciences, 2023, 13(3): 1476.
- [17] RADHA K, BANSAL M, PACHORI R B. Speech and speaker recognition using raw waveform modeling for adult and children's speech: a comprehensive review[J]. Engineering Applications of Artificial Intelligence, 2024, 131: 107661.
- [18] ABDALLAH H B, HENRY C J, RAMANNA S. 1-Dimensional polynomial neural networks for audio signal related problems[J]. Knowledge-Based Systems, 2022, 240: 108174.
- [19] PAN Z S, WANG B, ZHANG R B, et al. MIML-GAN: A GAN-based algorithm for multi-instance multi-label learning on overlapping signal waveform recognition[J]. IEEE Transactions on Signal Processing, 2023, 71: 859-872.
- [20] ZHAO Z P, LI Q F, ZHANG Z X, et al. Combining a parallel 2D CNN with a self-attention dilated residual network for CTC-based discrete speech emotion recognition[J]. Neural Networks, 2021, 141: 52-60.
- [21] CHENG M, WANG W Q, ZHANG Y C, et al. Target-speaker voice activity detection via sequence-to-sequence prediction[C]//ICASSP 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2023: 1-5.
- [22] ZHANG L G, TIAN Z, LONG Y F, et al. Cross-modal and cross-medium adversarial attack for audio[C]//Proceedings of the 31st ACM International Conference on Multimedia. 2023: 444-453.
- [23] SALAMON J, JACOBY C, BELLO J. 2014. A dataset and taxonomy for urban sound research[C]//Proceedings of the 22nd ACM International Conference on Multimedia. 2014: 1041-1044.
- [24] SANTOS-DOMINGUEZ D, TORRES-GUIJARRO S, CARDENAL-LOPZE A, et al. ShipsEar: an underwater vessel noise database[J]. Applied Acoustics, 2016, 113: 64-69.

- [25] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[C]//Proceedings of the IEEE International Conference on Learning Representations, 2015: 2931-2944.
- [26] KURAKIN A, GOODFELLOW I J, BENGIO S. Adversarial examples in the physical world[M]. Artificial Intelligence Safety and Security. Chapman and Hall/CRC, 2018: 99-112.
- [27] LIN J, SONG C, HE K, et al. Nesterov accelerated gradient and scale invariance for adversarial attacks[J]. arXiv preprint arXiv: 1908.06281, 2019.
- [28] POURSAEED O, KATSMAN I, GAO B, et al. Generative adversarial perturbations[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 4422-4431.
- [29] NASEER M, KHAN S, HAYAT M, et al. On generating transferable targeted perturbations[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 7708-7717.
- [30] ZHANG Q L, LI X D, CHEN Y F, et al. Beyond imagenet attack: Towards crafting adversarial examples for black-box domains[J]. arXiv preprint arXiv:2201.11528, 2022.
- [31] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.
- [32] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 4700-4708.
- [33] TAN M X, LE Q V. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks[C]//International Conference on Machine Learning (PMLR). 2019: 6105-6114.
- [34] MA N, ZHANG X, ZHENG H T, et al. 2018. Shufflenet v2: Practical guidelines for efficient CNN architecture design[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 116-131.

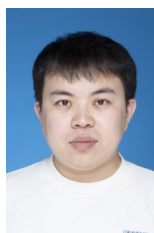
[作者简介]



田梓琳（1997-），女，河北唐山人，哈尔滨工程大学博士生，主要研究方向为对抗攻击、音频处理。



龙云飞（1998-），男，黑龙江哈尔滨人，哈尔滨工程大学博士生，主要研究方向为对抗攻击、鲁棒性分析。



田野（1992-），男，黑龙江哈尔滨人，西安电子科技大学博士后，主要研究方向多模态信号处理、智能安全。



张立国（1981-），男，黑龙江哈尔滨人，哈尔滨工程大学教授、博士生导师，主要研究方向为数字图像处理、多媒体数据处理。