成绩	

西北大学

本科毕业论文（设计）

题目：深伪图片识别模型研究

学生姓名 叶湛博

学　　号 2021112120

指导教师 汤战勇

院　　系 软件学院

专　　业 软件工程

年　　级 2021

教务处制

二〇二五年六月

诚信声明

本人郑重声明：本人所呈交的毕业论文（设计），是在导师的指导下独立进行研究所取得的成果。毕业论文（设计）中凡引用他人已经发表或未发表的成果、数据、观点等，均已明确注明出处。除文中已经注明引用的内容外，不包含任何其他个人或集体已经发表或在网上发表的论文。

特此声明。

论文作者签名：_____

日　期：　2025 年 6 月 7 日

摘　　要

深度伪造（Deepfake）是一种结合了 Deep Learning（深度学习）与 Fake（伪造）的技术，它利用先进的人工智能算法和深度学习模型，对图像中的人脸、物体和背景等进行篡改或生成，以达到以假乱真的效果。这些技术最初被应用于娱乐行业，如电影特效制作、游戏开发等，为观众带来了前所未有的视觉体验。然而，随着技术的普及和应用范围的扩大，深度伪造技术逐渐被不法分子利用，产生了许多负面影响。一个典型的 Deepfake 示例是 2022 年俄乌冲突中，被发布的泽连斯基（Volodymyr Zelensky）的 Deepake 投降演讲，这是 Deefake 第一次出现在大规模的信息战中。这个演讲由俄方使用开源模型结合在社交媒体上收集到的数据制作，它引起了人们对身份盗用、假冒以及错误信息在社交媒体上传播的担忧。

现有的基于频率的检测方法都依赖于 GAN 管道上采样过程中引入的特殊频率来检测伪造。但随着技术的快速发展，每一代模型都有自己独特的超参数和特定模式。因此，这些检测器在学习频域方面表现出缺乏熟练程度，并且过度拟合训练数据中存在的频域和范式，这导致了其在未知模型上的表现平平。

为了解决这个问题，开发可以在有限的训练数据下有效识别未见的深度伪造图像的模型，本文设计了一种新的频率感知方法，称为 Frequency Perceptron，通过学习不同频域的范式以提高深度伪造分类器的通用性。该方法使得检测器忽略不同生成器的特殊信噪，持续关注高频信息。该方法的核心是一个频域处理模块，它通过在快速傅里叶变换（FFT）和反快速傅里叶变换（iFFT）之间插入卷积层，来处理振幅和相位频谱，从而避免了分类器对于训练数据中图像的特定频域的过拟合现象。涉及数十个不同生成器的实验证明该方法的有效性，但是在部分模式如低频信噪处理、通道注意力机制等方面还有亟待优化。

关键词：深度伪造　伪造识别　频域

Abstract

Deepfake is a technology that combines Deep Learning and Fake. It utilizes advanced artificial intelligence algorithms and deep learning models to tamper with or generate faces, objects, and backgrounds in images to achieve an effect that is indistinguishable from the real thing. These technologies were initially applied in the entertainment industry, such as film special effects production and game development, bringing unprecedented visual experiences to audiences. However, with the popularization of technology and the expansion of its application scope, deepfake technology has gradually been exploited by lawbreakers, causing many negative impacts. A typical example of Deepfake was Volodymyr Zelensky's Deepake surrender speech that was released during the Russia-Ukraine conflict in 2022. This was the first time that Deefake appeared in a large-scale information war. This speech was produced by the Russian side using open-source models combined with data collected on social media. It has raised concerns about identity theft, counterfeiting and the spread of misinformation on social media.

The existing frequency-based detection methods all rely on the special frequencies introduced during the sampling process on the GAN pipeline to detect forgery. However, with the rapid development of technology, each generation of models has its own unique hyperparameters and specific patterns. Therefore, these detectors show a lack of proficiency in learning the frequency domain and overfit the frequency domains and paradigms existing in the training data, which leads to their mediocre performance on unknown models.

To solve this problem and develop a model that can effectively identify unseen deepfake images under limited training data, this paper designs a new frequency-aware method, called Frequency Perceptron, which improves the universality of the deepfake classifier by learning the paradigms of different frequency domains. This method enables the detector to ignore the special signal-to-noise of different generators and continuously focus on high-frequency

information. The core of this method is a frequency-domain processing module. It processes the amplitude and phase spectra by inserting convolutional layers between the Fast Fourier Transform (FFT) and the Inverse Fast Fourier Transform (iFFT), thereby avoiding the overfitting phenomenon of the classifier to the specific frequency-domain of the images in the training data. Experiments involving dozens of different generators have proved the effectiveness of this method, but there is still an urgent need for optimization in some modes such as low-frequency signal-to-noise processing and channel attention mechanisms.

Keywords： Deepfake; Forgery Identification; Frequency

目录

1 绪论

1.1 研究背景和意义

1.1.1 选题背景

随着数字技术的飞速发展，尤其是深度学习算法的不断进步，深度伪造技术（Deepfake）应运而生，并迅速成为数字媒体领域的一个重要研究热点。深度伪造技术主要基于生成对抗网络（GANs）和其他深度学习模型，能够生成高度逼真的虚假图片、视频和音频内容。这些技术最初被应用于娱乐行业，如电影特效制作、游戏开发等，为观众带来了前所未有的视觉体验。然而，随着技术的普及和应用范围的扩大，深度伪造技术逐渐被不法分子利用，产生了许多负面影响。

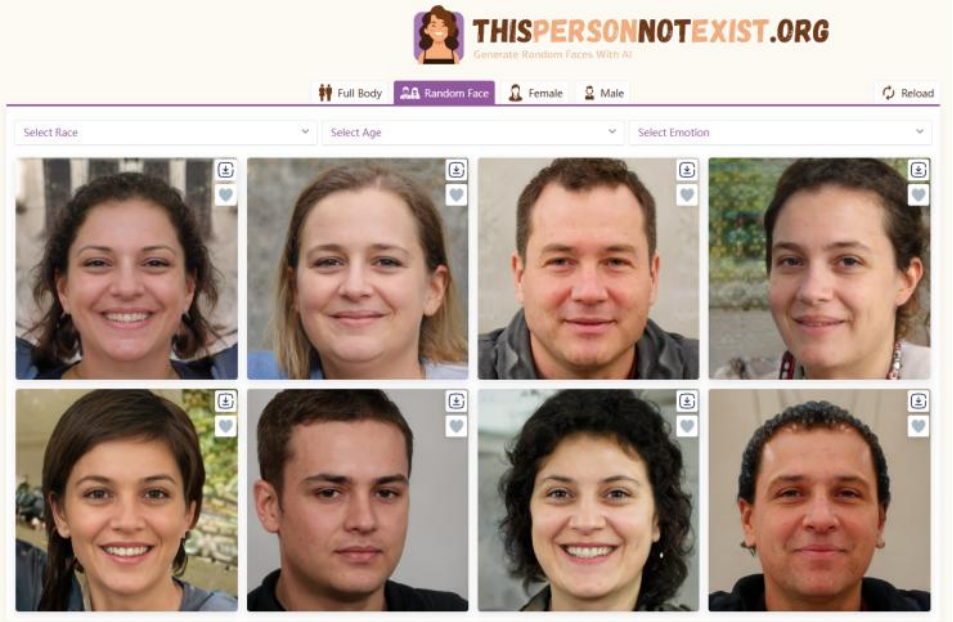

图 1-1 thispersonnotexist 网站，其中的人像均为深度伪造图片

在社交媒体和网络平台上，深度伪造技术被用于生成虚假的新闻图片、名人丑闻图片以及政治人物的虚假言论图片等，这些内容往往难以与真实图片区分开来，容易误导公众，引发社会恐慌和信任危机；在司法领域，深度伪造图片可能被用作伪证，干扰司法公正；在信息安全领域，深度伪造技术也可能被用于网络攻击，如通过生成虚假的用户身份图片或系统界面图片，欺骗用户输入敏感信息，从而窃取个人隐私和企业机密。

在数字化时代，信息安全是国家安全的重要组成部分。深度伪造图片的出现，给信息安全带来了新的挑战。通过研究深度伪造图片识别模型，可以有效识别和防范这些虚假图片，保护个人、企业和国家的信息安全。

1.1.2 选题意义

深度伪造图片识别模型的研究，是人工智能领域的一个重要课题。随着生成对抗网络（GANs）技术的快速发展，合成图像的生成变得越来越容易，其质量和逼真度也不断提高。这种现象对社会带来了潜在的负面影响，例如虚假信息传播、隐私侵犯和信任危机等。因此，如何有效检测这些伪造图像成为一个亟待解决的问题。

现有的伪造图像检测方法大多依赖于图像的空间信息或频率信息。其中，基于频率信息的方法通过分析 GAN 生成过程中引入的频率级伪影来检测伪造图像。然而，随着合成技术的不断进步，不同生成模型产生的伪影特征越来越多样化，导致现有检测器在面对未见过的生成模型时泛化能力不足，容易过拟合到训练数据中的特定伪影，从而在未知数据上表现不佳。因此，研究在有限的训练数据下有效识别未知的深度伪造图像的模型具有重要的理论和实践意义。

1.2 国内外研究现状和现状

本节首先对深度伪造图像检测的相关研究现状进行介绍，总结当前领域的发展前景和仍需解决的问题。

1.2.1 国外研究现状

近年来，深度伪造技术在国际上引起了广泛关注。在生成方面，国外研究学者 Goodfellow 等人提出的生成对抗网络（GANs）不仅能够生成高度逼真的虚假图片，其通过生成器和判别器进行对抗训练的思想，更是为后续许多模型和算法打下了基础。

在检测方面，国外研究者主要基于深度学习模型，如卷积神经网络（CNNs）、Transformer 等。例如，Afchar 等人提出的 MesoNet 模型，通过学习图片的局部特征，能够有效检测深度伪造图片。此外，Jia 等人提出的 PM-DETR 模型，利用 Transformer 的注意力机制，能够检测出深度伪造图片中的细微差异。

1.2.2 国内研究现状

在国内，深度伪造技术的研究也得到了广泛关注。在生成方面，国内研究者提出了扩散模型深度学习技术，能够生成高质量的虚假图片，甚至可以模拟特定人物的表情和动作。在检测方面，国内研究者开创性的提出的 Xception 模型，通过学习图片的复杂空间特征，能够有效检测深度伪造图片。此外，国内研究者还提出了基于多模态数据的检测方法，利用多模态信息提高检测的准确性。

国内的研究主要包括：李佳乐（2023）[1] 提出基于解耦动态卷积的检测方法。通过数据增强（随机亮度对比度）预处理图像，引入解耦动态卷积捕获异常特征及其位置，结合通道和空间信息，优化检测效果。彭舒凡（2023）[2] 将人脸伪造检测任务定义为“定位高区分度区域和提取细粒度特征”，通过两个特征提取模块获取区域级和细粒度特征，并利用多分支网络进行学习，提高了检测精度。朱新同等（2021）[3] 使用 Scharr 算子提取 YCbCr 色彩空间中 Cb 和 Cr 分量的图像边缘信息，以及使用 Laplacian 算子提取 RGB 色彩空间中 G 分量的图像边缘二阶梯度信息。彭舒凡等（2022）[4] 在 XceptionNet 的基础上引入通道注意力机制 ECA-Net，形成 XceptionECA 网络。谢菲（2023）[5] 基于 ForgeryNet 数据集，将图像分为四组进行研究，提出了两种特征选择方法：基于数理统计法和基于深伪技术成像规律法。通过这两种方法筛选出与深伪图像高相关性和高稳定性的特征。马喆等（2022）[6] 提出一种基于低层特征的检测方法，通过预处理提取低层特征并使用简单的全连接网络进行分类。黄怿豪（2022）[7] 通过在伪造图像中加入噪声破坏伪影，再利用深度预测滤波技术恢复图像。该方法不仅减少了伪影，还具有处理速度快和域自适应性强的优点。张亚等（2024）[8] 提出了一种基于自动编码器的伪造图像检测方法，结合高斯滤波预处理和注意力机制，有效提升了检测性能。在小样本、多场景数据集上验证了方法的有效性，平均准确率显著优于现有方法。郭歌阳（2022）[9] 该算法针对完全伪造图像的检测，利用 GAN 生成图像在颜色和噪声方面的差异，设计了双分支网络结构。颜色分支提取颜色域特征，噪声分支提取噪声域特征，最后将两分支特征融合进行分类。吴远泸（2022）[10] 提出了一种域不变表征学习（DIRL）方法。该方法通过模拟的方式将多个源域划分为虚拟源域和虚拟目标域，利用特征分布差异最小化网络（FDDM）和最优分类器距离最小化网络（OCDM）学习域不变表征空间。张今吟（2025）[11] 引入 ManTra-Net 架构，构建痕迹特征提取器的三个网络模块：“First Block”、“MiddleBlock”和“LastBlock”，其分别负责初步特征提取、高级语义信息提取和特征精炼。唐玉敏等（2022）[12] 总结了 Jeong 等人提出双边高通滤波器（BiHPF）；Liu 等人提出块混洗学习结合对抗损失算法；Guarnera 等人提出基于 EM 算法的检测方案。杨雨鑫等（2021）[13] 提出的融合传统特征与神经网络的深度伪造检测算法，它结合了神经网络的高检测率和传统特征的可解释性，能够有效检测多种深度伪造方法生成的图像。 张凯等（2025）[14] 提出 ATNet，通过采用对抗训练策略，动态合成对抗伪造样本来扩大样本空间，增强模型对不同伪造算法生成样本的敏感度，提高检测精度和泛化性。李沛（2022）[15] 提出并研究了两种改进的深度伪造检测模型：分级特征全局融合 HVGG19 模型和全局自注意力融合 AVGG19-LSTM 模型。

1.3 拟解决的问题

（1）许多现有的先进检测方法依赖于复杂的网络结构和大量的参数，这导致计算成本高、训练时间长，并且在实际应用中效率较低。

（2）现有的深度伪造检测方法大多在训练和测试时依赖于同一数据源，导致它们在面对未见过的生成模型（GAN）或新类别时泛化表现不佳。

（3）现有的学习策略只是简单的对频域空间特征进行收集处理，对于不同的数据集的特殊频率伪影会产生失真。

1.4 论文主要内容和结构

本文主要分为五个部分，具体如下：

第一章 绪论。 本章主要提出本文的选题背景、选题意义，并对目前深度伪造图片检测领域的研究做简要介绍。在介绍本文基本内容后，再提出本文拟解决的问题。

第二章 相关工作支撑。 本章主要介绍深度伪造图片检测需要的基本理论支撑。对国内外学者的研究、Deepfake 原理、伪造图片频域处理等概念做详细介绍，方便后续检测模型的设计。

第三章 Frequency Perceptron 模型设计。本章主要介绍本文设计的 Frequency Perceptron 模型如何通过频率空间学习，增强伪造检测能力。其中包括高频图像表示（High-Frequency Image）、高频特征表示（High-Frequency Feature）和频率卷积层（Frequency Convolute Layer）三个核心模块。

第四章 实验和结果。 通过设置不同的对照实验方案，来比较国内外现有开源模型和本文设计模型的在泛化场景下的性能，得出结论。

第五章 结论与展望。 通过已经完成的实验结果进行分析，得出模型在检测性能、参数量、泛化性等结果上的提升。同时发现该模型的一些不足之处，对未来本方向的工作研究进行展望。

2 相关工作支撑

2.1 生成对抗网络 GAN

生成对抗网络（Generative Adversarial Networks，GAN）是由 Ian Goodfellow 等人在 2014 年提出的一种深度学习模型，用于生成与真实数据难以区分的合成数据。GAN 的核心思想是通过两个神经网络——生成器（Generator）和判别器（Discriminator）——的对抗训练来生成逼真的数据。

2.1.1 基本架构

（1）生成器（Generator，G）：

输入：随机噪声向量 z（通常从高斯分布中采样）。

输出：合成数据 G(z)，其目标是尽可能接近真实数据的分布。

作用：生成器的目标是生成能够“欺骗”判别器的合成数据，使其被判别为真实数据。

（2）判别器（Discriminator，D）：

输入：真实数据 x 或生成器生成的合成数据 G(z)。

输出：概率值 D(x) 或 D(G(z))，表示输入数据为真实数据的概率。

作用：判别器的目标是区分真实数据和合成数据，输出接近 1 的概率表示输入是真实数据，接近 0 的概率表示输入是合成数据。

2.1.2 训练过程

GAN 的训练过程是一个对抗过程，生成器和判别器交替更新，具体步骤如下：

（1）判别器训练

输入真实数据 x，判别器输出 D(x)，目标是使 D(x) 接近 1。

输入生成器生成的合成数据 G(z)，判别器输出 D(G(z))，目标是使 D(G(z)) 接近 0。

通过最小化损失函数（2-1）更新判别器的参数：

$$\mathcal{L}_D = -\mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] - \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))] \tag{2-1}$$

（2）生成器训练：

生成器的目标是生成能够“欺骗”判别器的合成数据，即使 D(G(z))接近 1。

通过最小化损失函数（2-2）更新生成器的参数：

$$\mathcal{L}_G = -\mathbb{E}_{z\sim p_z(z)}[\log D(G(z))] \tag{2-2}$$

（3）交替更新：

生成器和判别器交替更新，直到达到纳什均衡点，即生成器生成的数据与真实数据难以区分。

2.2 频率伪影

频率伪影（Frequency Artifact）是指由于生成模型的算法特性或数据处理过程导致的图像在频率域中出现的异常特征或模式。

2.2.1 产生原因和意义

在生成高分辨率的伪造图像中，GAN 或者其他生成模型会通过上采样（如反卷积层）将低分辨率的特征图放大，以增加图像尺寸。在这一过程中，某些 GAN 架构（如 StyleGAN）可能会在频率域中引入周期性的模式；并且有研究表明，伪造图像的频率分布可能与真实图像的自然频率分布不一致，在频域中表现为异常增强或不规则分布，这一现象在高频区域尤其显著。

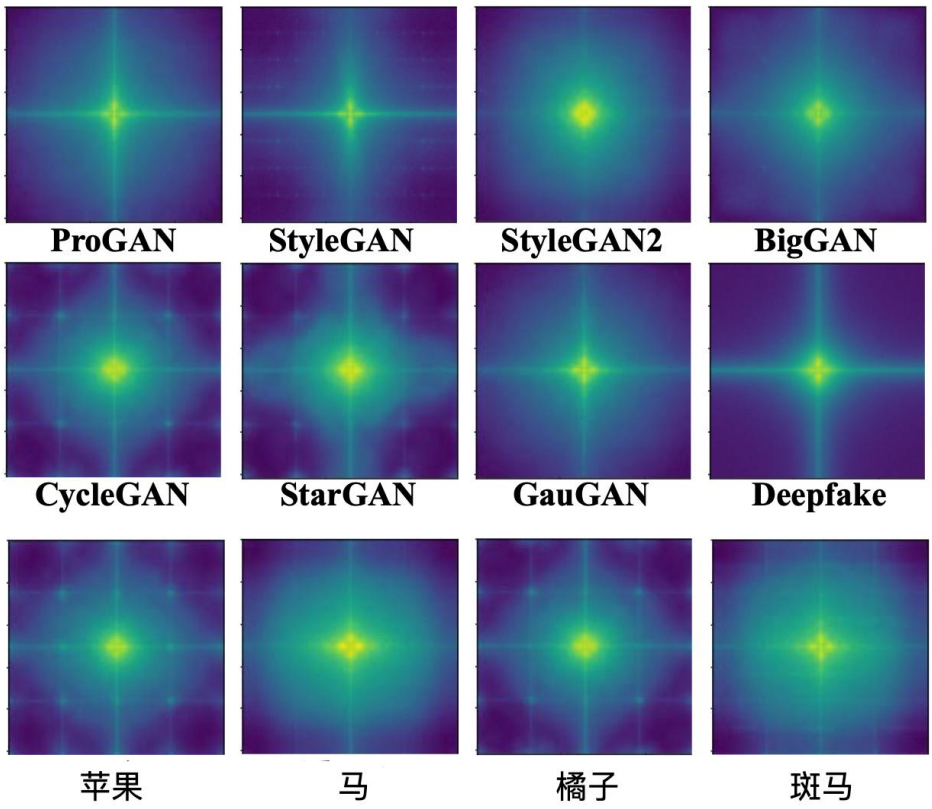

图 2-1 不同模型的伪造图片和真实图片的频率伪影

因此，频率伪影是深度伪造图像中一个关键的检测线索，通过分析图像的该特征，可以有效识别出伪造图像，并提高检测的准确性和鲁棒性。

2.2.2 检测方式

（1）通过分析图像的频率域特征，可以检测出伪造图像中的频率伪影。主流的检测方式如下：
（2）傅里叶变换（FFT）：将图像从空间域转换到频率域，分析其频率谱。
（3）离散余弦变换（DCT）：通过将图像从空间域转换到频率域，可以更清晰地观察到频率伪影。
（4）局部频率统计（LFS）：通过分析图像的局部频率统计特征，可以识别出伪造图像中异常的频率分布。

2.3 FrePGAN

FrePGAN 的应用场景主要集中在 Deepfake 检测领域，其出现为打击伪造图像提供了有效的工具。通过生成频率级别的扰动映射，使生成的图像与真实图像难以区分，从而提升检测器的鲁棒性。FrePGAN 能够帮助识别出那些未在训练集中出现的 GAN 模型生成的图像，从而保护信息的真实性和完整性

2.3.1 工作原理

FrePGAN 的核心思想是通过在频率级别上引入扰动，使得生成的图像在频率域上具有与真实图像相似的特性。这种方法可以有效减少过拟合问题，使得检测器能够更好地泛化到未知的 GAN 模型上。

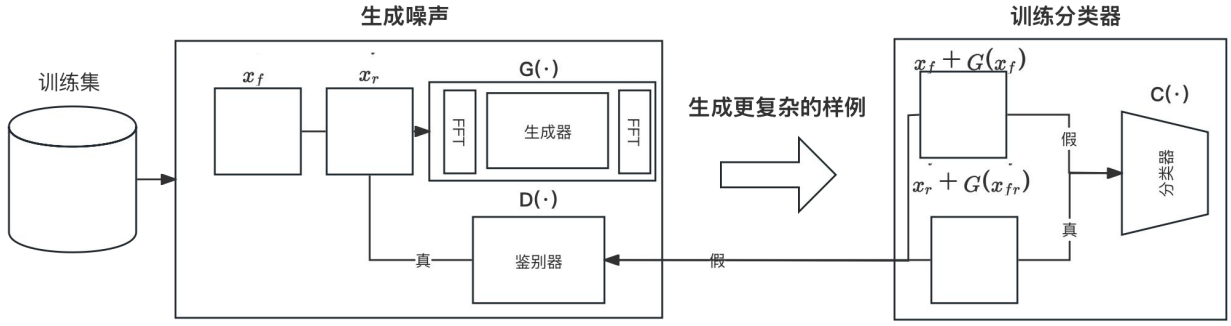

图 2-2 FrePGAN 流程图

FrePGAN 的实现主要依赖于三个关键的模块：
（1）频率扰动映射生成器：FrePGAN 生成频率级别的扰动映射，这些映射使得生成的图像在频率域上与真实图像更加相似。
（2）对抗训练模块：在训练过程中，生成器和扰动映射相互对抗，生成器努力生成更真实的图像，而扰动映射则试图揭示这些图像的伪造痕迹。
（3）优化检测器：通过不断的训练和优化，FrePGAN 能够提高检测器对各种 GAN 模型的泛化能力，使其能够识别出不同 GAN 模型生成的图像。

2.3.2 缺点

FrePGAN 通过在训练过程中引入频率级扰动图，试图在未知的生成对抗网络（GAN）模型上实现更好的泛化能力。然而，这种设计在频率域中生成和处理扰动图，无疑增加了计算复杂度。在实时应用中，这种额外的计算负担可能会限制其实际部署。同时，这种设计在一定程度上牺牲了对已训练 GAN 模型的检测性能。换言之，虽然 FrePGAN 能够更好地检测未见过的伪造图像，但在面对训练集中已知模型生成的图像时，其检测精度反而会有所下降。

3 Frequency Perceptron 模型设计

本节介绍了 Frequency Perceptron 的频域学习网络，一个通用的深度伪造检测器。该检测器能在训练数据有限的情况下，准确识别伪造图像。Frequency Perceptron 通过在卷积中融入频域学习模块，提高对未知数据源的泛化能力。主要模块包括：从图像中提取高频成分的高频图像模块、强调特征空间高频信息的高频特征模块、以及在频域中学习特征的频域卷积层。这些模块协同工作，使检测器更关注高频细节，减少对于已知数据集的过拟合。

3.1 基本流程

Frequency Perceptron 频域学习网络，通过在卷积神经网络中融入频域学习插件模块，促使分类器在频域中进行操作，从而提高检测器对未见数据源的泛化能力。如图 3-1 所示，展示了 Frequency Perceptron 模型的整体流程图，该方法通过频域学习来减轻源特定依赖性。

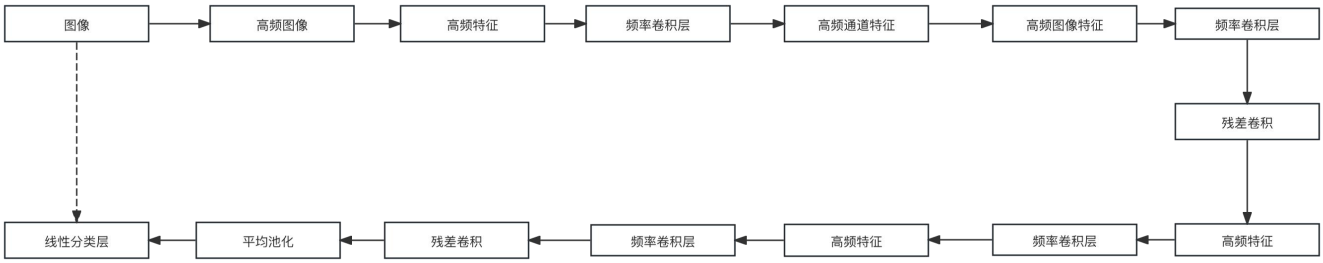

图 3-1 Frequency Perceptron 流程图

3.2 关键模块

Frequency Perceptron 的核心目标是增强检测器对未见数据源的泛化能力。为此，本文设计了一个频域学习网络，通过在 CNN 分类器中融入频域学习插件模块，迫使分类器在频域中进行操作，其中传统频域网络和 Frequency Perceptron 区别如图 3-2 所示。

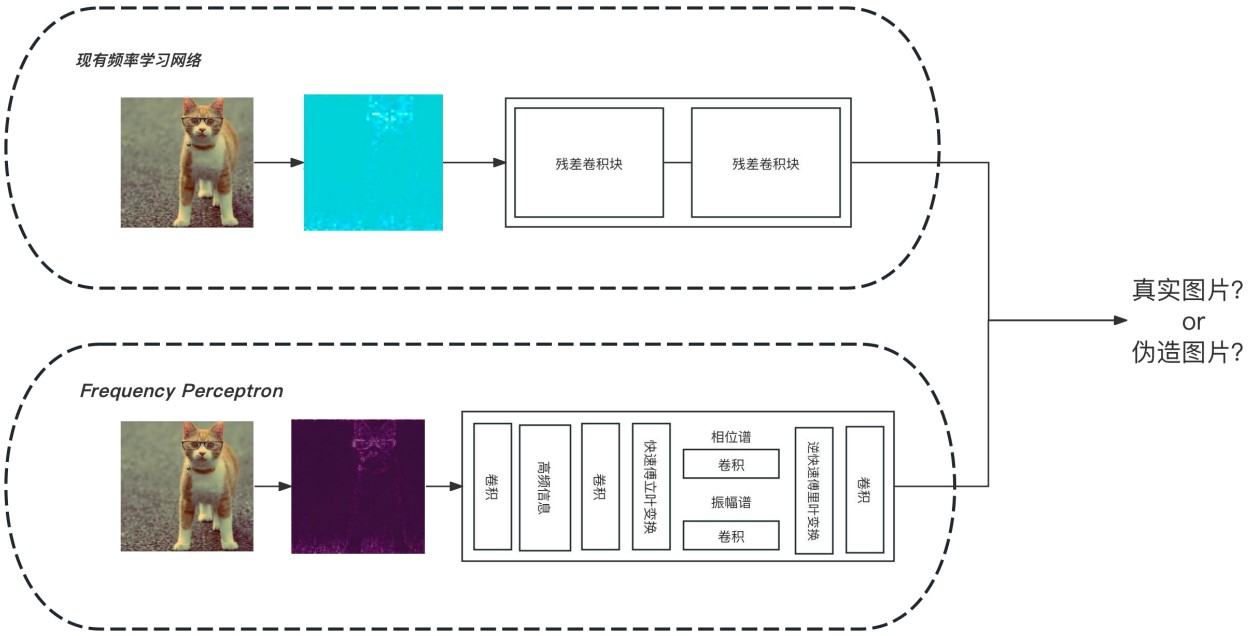

图 3-2 传统频域网络和 Frequency Perceptron 区别

3.2.1 高频图像（High-Frequency Image）

从图像中提取高频成分作为检测器的输入。具体来说，将训练图像转换到频域，使用高通滤波器提取高频成分，然后再将这些高频成分转换回图像空间。这样可以确保检测器重点关注图像中包含的相关高频信息。

- 作用：提取图像的高频成分，作为检测器的输入。
- 实现步骤：
  - 将训练图像 $x \in \mathrm{R}^{W \times H \times 3}$ 转换到频域，使用快速傅里叶变换（FFT）。
  - 通过高通滤波器 B_h 提取高频成分 f_h 。
  - 将高频成分转换回图像空间，得到 x_h ，即图像空间中的高频成分。

代码 3-1 High-Frequency Image 模块

```
def high_frequency_image(self, x, scale=5):
    x = torch.fft.fft2(x, norm="ortho")
    x = torch.fft.fftshift(x, dim=[-2, -1])
    _ ,_ , h, w = x.shape
    x[:, :, h//2-h//scale:h//2+h//scale, w//2-w//scale:w//2+w//scale] = 0.
    x = torch.fft.ifftshift(x, dim=[-2, -1])
    x = torch.fft.ifft2(x, norm="ortho")
    x = torch.real(x)
    x = F.relu(x)
    return x
```

3.2.2 高频特征（High-Frequency Feature）

在特征提取过程中，模型始终优先关注特征空间中的高频信息，以减少对训练源特定模式的过拟合。模型对卷积层的输出分别在空间维度和通道维度上进行频域转换，应用高通滤波器提取高频信息，再将提取的高频信息转换回特征空间。这种做法可以使检测器更关注于特征中的高频细节，从而提高其泛化能力。

- 作用：在特征空间中强调高频信息，减少对训练源特定模式的过拟合。
- 实现步骤：
  - 对第 k 个卷积层的输出 $M_k \in \mathrm{R}^{H\times W\times C}$，分别在空间维度（W, H）和通道维度 C 上应用 FFT。
  - 移动零频分量到中心，应用高通滤波器 B_h 提取高频信息。
  - 将提取的高频信息通过逆 FFT（iFFT）转换回特征空间，得到 $\mathrm{M}_{\mathrm{k}}^{\mathrm{h}}$。

代码 3-2 High-Frequency Feature 模块

```
def high_frequency_feature(self, x, scale=5):
    x = torch.fft.fft(x, dim=1, norm="ortho")
    x = torch.fft.fftshift(x, dim=1)
    _, c, _, _ = x.shape
    x[:, c//2-c//scale:c//2+c//scale, :, :] = 0.
    x = torch.fft.ifftshift(x, dim=1)
    x = torch.fft.ifft(x, dim=1, norm="ortho")
    x = torch.real(x)
    x = F.relu(x)
    return x
```

3.2.3 频域卷积层（Frequency Conv Layer）

模型不仅将频率信息用作伪影表示，还引入频域学习作为一种增强检测器泛化能力的策略。模型将卷积层的特征图从特征空间转换到频域，在幅值谱和相位谱上应用卷积层进行学习，然后将学习到的频谱信息转换回特征空间。这个过程有助于提高检测器对频域特征的敏感性。

- 作用：通过在频域中学习，提高检测器对频域特征的敏感性。
- 实现步骤：
  - 将卷积层的特征图 $M_k \in \mathrm{R}^{W\times H\times C}$ 转换到频域。
  - 在幅值谱和相位谱上分别应用卷积层 $L\mathrm{conv}$，学习频域中的特征。
  - 将学习到的频谱信息通过 iFFT 转换回特征空间，得到新的特征图 $\widetilde{\mathrm{M}}_{\mathrm{k}}$。

代码 3-3 Frequency Convolute 模块

```
def frequency_convolute_layer(self, x):
    x = torch.fft.fft2(x, norm="ortho")
    x = torch.fft.fftshift(x, dim=[-2, -1])
    x = torch.complex(self.realconv1(x.real), self.imagconv1(x.imag))
    x = torch.fft.ifftshift(x, dim=[-2, -1])
    x = torch.fft.ifft2(x, norm="ortho")
    x = torch.real(x)
    x = F.relu(x)
    return x
```

3.2.4 协同工作

总的来说，High-Frequency Image 负责从图像中提取高频信息，作为检测器的输入，确保模型关注图像中的高频细节；High-Frequency Feature 和 Frequency Conv Layer 则负责在特征提取过程中，持续强调高频信息，迫使模型学习更具泛化能力的特征表示。模块之间的连接方式如代码（3-4）所示。

代码 3-4 forward 方法

```
def forward(self, x):
    x = self.high_frequency_image(x)
    x = F.relu(F.conv2d(x, self.weight1, self.bias1, 1, 0))
    x = self.high_frequency_feature(x)
    x = self.frequency_convolute_layer(x)
    x = self.high_frequency_image(x)
    x = F.relu(F.conv2d(x, self.weight2, self.bias2, 2, 0))
    x = self.high_frequency_feature(x)
    x = self.frequency_convolute_layer(x)
    x = self.maxpool(x)
    x = self.layer1(x)
    x = self.high_frequency_image(x)
    x = F.relu(F.conv2d(x, self.weight3, self.bias3, 1, 0))
    x = self.frequency_convolute_layer(x)
    x = self.high_frequency_image(x)
    x = F.relu(F.conv2d(x, self.weight4, self.bias4, 2, 0))
    x = self.frequency_convolute_layer(x)
    x = self.layer2(x)
    x = self.avgpool(x)
    x = self.fc1(x.view(-1))
    return x
```

4 实验和结果

在本节中，我会对 Frequency Perceptron 的效果进行全面评估，内容涵盖了训练集、实现细节、测试集等。接下来将进一步阐述这些方面，以便全面了解 Frequency Perceptron 的能力和有效性。

4.1 实验设计

实验基于 PyTorch 框架来实现不同的模型方法和 Frequency Perceptron，利用Nvidia GeForce RTX 3080 来增强计算能力。对于快速傅里叶变换，我选择了 PyTorch 库中的 torch.fft.fftn 函数。

4.1.1 训练和评估

在训练过程中，我们使用初始学习率为 1e-3 的 AdamW 优化器。BatchSize大小设置为 32，训练模型 100 个 epoch，通过 AdamW 的 betas 来实现学习率衰减策略。我选择使用平均精度分数和准确度作为衡量不同模型方法有效性能的主要指标。这些指标对不同辨别方法的性能提供了一个全面且有效的标准参考。

4.1.2 训练集

为了确保比较的一致基础，我们使用ForenSynths 的训练集来训练检测器，与基线对齐。训练集由 20 个不同的类别组成，每个类别包括使用StarGAN生成的 18,000 个合成图像，以及来自LSUN 数据集的相同数量的真实图像。为了排除偶然因素和模型训练导致的误差，我们使用相同的超参数重复三组实验，将平均值作为最终的评估结果。

4.1.3 测试数据集

为了评估该方法在真实场景中的泛化能力，我采用了不同的 GAN 模型生成的不同图像。首先， 我从 ForenSynths 的测试集中挑选了 5 个 GANs 模型的数据作为基础。为了增加野生场景的不可预测性，我又额外收集了这些 GANs 生成的图像来扩展数据集。真实图像则从 6

个不同数据集中采样得来。最终，保证了一共有 30K 张测试图像，每个数据集中真假图像的数量相等。

4.2 结果

我们比较了之前的四种模型方法：FrePGAN、F3Net、LGrad、Ojha。结果如表 1 所示。提出的 Frequency Perceptron 方法在平均准确率方面超过了测试的其他方法。与常见的 LGrad 和 Ojha方法相比，Frequency Perceptron 有了很大的改进，平均 Acc比这两种方法分别高出 1.9% 和 4.5%。此外，Frequency Perceptron 的参数量约为 190 万，明显少于 Ojha 拥有的 3.04 亿个参数和 LGrad 拥有的 4600 万个参数，参数量的巨大差异可以被转化为显著的性能增益。这一结果强调了 Frequency Perceptron方法的效率和有效性，表明可以用更少的参数实现更优的性能，这是现实应用中的一个关键优势。 与更早的方法 FrePGAN 和 F3Net 相比，Frequency Perceptron 方法在现实场景中实现了更好的性能。结果证实了所提出方法的泛化能力和可行性。

表 4.1 不同辨别模型在测试集上的性能评估

测试方法（模型参数量）	测试集样例的生成模型										平均	
	CycleGAN		StyleGAN		ProGAN		GauGAN		Deepfake			
	准确率	平均精度分数	准确率	平均精度分数	准确率	平均精度分数	准确率	平均精度分数	准确率	平均精度分数	准确率	平均精度分数
FrePGAN（大于 500 M）	66.53%	68.2%	80.7%	90.73%	97.83%	99.56%	55.46%	57.26%	67.83%	84.66%	73.67%	80.08%
F3Net（48.9 M）	98.07%	99.97%	87.80%	99.67%	78.10%	86.00%	58.07%	58.83%	61.53%	81.37%	79.07%	87.20%
LGrad（46.6 M）	99.70%	99.97%	95.20%	99.73%	85.30%	94.37%	72.33%	81.43%	61.77%	71.20%	86.20%	92.13%
Ojha（304.1M）	99.50%	100.0%	81.67%	97.33%	92.30%	99.83%	92.50%	100.0%	80.87%	90.43%	88.80%	97.97%
Frequency Perceptron（1.9 M）	99.07%	99.97%	90.87%	99.10%	96.20%	99.50%	91.07%	99.00%	79.80%	93.60%	90.70%	93.00%

5 结论与展望

Frequency Perceptron 是一种轻量级模型，可以通用的检测由各种生成模型创建的伪造图像。与其他传统模型相比，该方法将频率分析融入网络结构中，使得检测器更关注高频细节，从而减少对于已知数据的过拟合。结合实验结果，Frequency Perceptron 用更少的参数就能达到更优水平，说明该方法的有效性。

另一方面，该方法主要关注的是由 GAN 生成的假图像，随着科学技术水平的提高，出现了很多优秀的非 GAN 结构生图网络，如 diffusion 模型。Frequency Perceptron 是否也适用于该类模型伪造的图像仍有待研究和挖掘。

参考文献

[1] 李佳乐.图像深度伪造及其检测技术研究[D].华侨大学,2023.DOI:10.27155/d.cnki.ghqiu.2023.001124.

[2] 彭舒凡.基于深度学习的人脸伪造图像检测研究[D].中国人民公安大学,2023.DOI:10.27634/d.cnki.gzrgu.2023.000322.

[3] 朱新同,唐云祁,耿鹏志.基于特征融合的篡改与深度伪造图像检测算法[J].信息网络安全,2021,21(08):70-81.

[4] 彭舒凡,蔡满春,刘晓文,等.基于图像细粒度特征的深度伪造检测算法[J].信息网络安全,2022,22(11):77-84.

[5] 谢菲.基于深度学习融合多维识别特征的深度伪造图像检测研究[D].中国人民公安大学,2023.DOI:10.27634/d.cnki.gzrgu.2023.000290.

[6] 马喆,周华兵.采用低层特征的深度伪造图像检测方法[J].软件导刊,2022,21(01):238-242.

[7] 黄怿豪.深度伪造图像的精细化与检测研究[D].华东师范大学,2022.DOI:10.27149/d.cnki.ghdsu.2022.000833.

[8] 张亚,金鑫,江倩,等.基于自动编码器的深度伪造图像检测方法[J].计算机应用,2021,41(10):2985-2990.

[9] 郭歌阳.面向生成对抗网络的深度伪造图像检测算法研究[D].西安理工大学,2022.DOI:10.27398/d.cnki.gxalu.2022.001126.

[10] 吴远泸.基于深度学习的图像伪造检测方法研究[D].华南理工大学,2022.DOI:10.27151/d.cnki.ghnlu.2022.000475.

[11] 张今吟.基于深度卷积神经网络的多源网络图像伪造检测方法[J].网络安全技术与应用,2025,(02):36-38.

[12] 唐玉敏,范菁,曲金帅.深度伪造生成与检测研究综述[J].计算机工程与应用,2022,58(23):56-66.

[13] 杨雨鑫,周欣,熊淑华,等.融合传统特征与神经网络的深度伪造检测算法[J].信息技术与网络安全,2021,40(02):33-38+44.DOI:10.19358/j.issn.2096-5133.2021.02.006.

[14] 张凯,范智贤.基于对抗训练的改进人脸伪造检测方法[J/OL].重庆工商大学学报(自然科学版),1-8[2025-03-05].http://kns.cnki.net/kcms/detail/50.1155.N.20231211.1539.002.html.

[15] 李沛.基于卷积神经网络的深度伪造检测算法研究[D].西北大学,2022.DOI:10.27405/d.cnki.gxbdu.2022.000689.

[16] Zhao, Hanqing et al. “Multi-attentional Deepfake Detection.” 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021): 2185-2194.

[17] Tsigos, Konstantinos et al. “Improving the Perturbation-Based Explanation of Deepfake Detectors Through the Use of Adversarially-Generated Samples.” (2025).

[18] Stamnas, Sotirios and Victor Sanchez. “DiffFake: Exposing Deepfakes using Differential Anomaly Detection.” (2025).

[19] Yasir, Siddiqui Muhammad and Hyun Kim. “Lightweight Deepfake Detection Based on Multi-Feature Fusion.” Applied Sciences (2025): n. pag.

[20] Chen, Yunzhuo et al. “Deepfake Detection with Spatio-Temporal Consistency and Attention.” 2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA) (2022): 1-8.