分类号 TP311 密级 公开
UDC 编号

雲南大學

硕士研究生学位论文

题 目 基于伪造缺陷与语义对比的 deepfake 检测

Title Deepfake Detection Based on Forgery Defects and Semantic Comparison

学院（所、中心） 软件学院

专 业 名 称 软 件 工 程

研 究 方 向 深 度 伪 造 检 测 与 图 像 取 证

研究生姓名 汪高健 学号 12019202394

导 师 姓 名 崔晓晖 职称 教授

2022 年 6 月

摘要

伪造图像或视频的检测一直是计算机视觉与信息安全领域的研究重点。随着诸如自动编码器和生成对抗网络等深度生成模型的显著发展与进步，人们现在可以轻松地使用开源工具甚至移动应用操纵图像和视频。最为知名的 deepfake 是指基于深度学习制作的虚假多媒体内容，即深度伪造。鉴于 deepfake 较低的制作门槛且高质量的伪造效果，这些技术很容易被恶意者滥用，包括传播虚假新闻，色情报复，金融欺诈甚至影响政治舆论。因此，行之有效的 deepfake 检测技术是迫切而重要的需求。

虽然 deepfake 检测算法得到了许多关注和研究，但当前的方法仍存在如下局限性：第一，一些局部操纵方法伪造的和高质量的 deepfake 仅存在细微的伪造伪影，导致大多数检测方法的有效性不足。第二，检测算法在现实应用中的鲁棒性不强，如上传至社交媒体平台的伪造图像或视频通常经过不同程度压缩的场景。第三，现有的检测模型通常直接使用数据训练深度神经网络进行分类学习，虽然它们在数据集内的测试结果可观，但在跨不同数据集或伪造方法进行检测时的泛化性严重不足。为了应对这些挑战，旨在提升 deepfake 检测的有效性，泛化性和鲁棒性，本文基于伪造缺陷与语义对比对 deepfake 检测进行了研究，并提出了两个创新性的方法。

针对检测不同 deepfake 操纵方法的有效性以及在压缩环境下的鲁棒性，本文提出了一个基于空域与频域中局部伪造缺陷的 deepfake 检测算法，命名为局部伪影感知的 deepfake 检测网络（LA-Net）。该算法的研究旨在放大真实图像与伪造图像在空域和频域中隐含的局部差异，而不是特定的外观特征。该工作创新性地设计了两个模块：局部风格提取模块（LSEB）在空域中对局部风格进行编码，从而提取更具判别性的特征；Patch-wise 频谱交互注意模块（PFSCA）在频域中交互地挖掘幅度谱和相位谱上的局部伪造伪影。通过双分支的深度学习网络，该算法同时在空域和频域中捕获细微的局部伪造缺陷。在不同操纵方法和压缩场景下的大量实验证明了提出的 deepfake 检测方法的有效性和鲁棒性。

为了提升 deepfake 检测的泛化性，本文工作关注到了各种 deepfake 伪造算法主要是对人脸所蕴含的内容和信息进行操纵。以人脸语义内容存在的伪造缺陷为突破口，本文创新性地提出了一种基于语义分割与对比分类的检测算法。该方法将人脸语义内

容和非语义内容进行分割，并针对语义特征进行额外地对比学习。语义分割使能在语义内容中学习不同 deepfake 之间共同的伪造缺陷。对比学习强迫编码空间中同类特征更紧凑且异类特征的间隔更大，这进一步提升了模型泛化性。通过大量的跨数据集测试实验和对比评估，证明了提出的方法能显著改善当前基于深度卷积神经网络的 deepfake 检测模型的泛化能力。

关键词：Deepfake 检测；伪造视频检测；图像取证；特征提取；深度学习

ABSTRACT

Forged images or videos detection has always been a research focus in the field of computer vision and information security. With the remarkable development of deep generative models such as, autoencoders and generative adversarial networks, people can easily manipulate images and videos with open-source tools and even mobile apps. The most well-known deepfake refers to manipulating multimedia content via deep learning. Given low barriers to production and high-quality forgery effects of deepfakes, these techniques are easily abused by malicious users, including spreading fake news, pornographic revenge, financial fraud and even influencing political opinion. Therefore, it is an urgent and important demand to develop effective deepfake detection methods.

Although deepfake detection algorithms have received a lot of attention and research, current methods still have the following limitations: First, some local manipulations and high-quality deepfakes only have subtle forgery artifacts, which limits the effectiveness of most detection methods. Second, detection algorithms are not robust in real-world applications, such as scenarios where fake images or videos uploaded to social media platforms are compressed to varying degrees. Third, existing detection models usually directly use data to train deep neural networks for classification learning. Although they have achieved impressive test results within datasets, their generalization is seriously weak across different datasets or forgery methods. To tackle these challenges and improve the effectiveness, generalization, and robustness of deepfake detection, this thesis studies deepfake detection based on forgery defects and semantic comparison and then proposes two novel methods.

Aiming at improve the effectiveness of detecting different deepfake manipulation methods and the robustness in compressed scenarios, this thesis proposes a deepfake detection algorithm based on local forgery defects in the spatial and frequency domains, named Local Artifact-aware Deepfake Detection Network (LA-Net). The research of this algorithm aims to amplify the local differences implied in the spatial and frequency domains between real images and fake images, rather than specific appearance features. In this work, two modules are innovatively designed: Local Style Extraction Block (LSEB) encodes local styles in the spatial domain to extract more discriminative features; Patch-wise Frequency Spectrum Cross Attention (PFSCA) module interactively mine local forgery artifacts on amplitude and phase

spectrum from the frequency domain. Through a two-branch deep learning network, the model captures subtle local forgery defects in both the spatial and frequency domains. Extensive experiments under different manipulation methods and compression scenarios demonstrate the effectiveness and robustness of the proposed deepfake detection method.

To improve the generalization of deepfake detection, this thesis focuses on the manipulation of human face content and information by various deepfake forgery algorithms. Taking the forgery defect of facial semantic content as a breakthrough, this thesis proposes a novel detection algorithm based on semantic segmentation and contrastive classification. The proposed method segments facial semantic content and non-semantic content, and then performs additional contrastive learning for semantic features. Semantic segmentation enables the learning of common forgery defects among different deepfakes in semantic content. Contrastive learning further improves the generalizability of the model by forcing features of the same class to be put together while increasing the distance between different classes. Through extensive cross-dataset test experiments and comparative evaluations, we demonstrated the proposed method can significantly improve the generalization ability of current deepfake detection models based on deep convolutional neural networks.

Key words: Deepfake detection; Forged video detection; Image forensics; Feature extraction; Deep learning

目 录

第一章 绪论

1.1 研究背景及意义

自从数字视觉媒体出现以来，人们就一直以各种目的来操纵它们。通常，操纵多媒体内容需要领域专家的专业知识，而且相当耗时和费力，例如使用专业软件 Adobe Photoshop 编辑照片或使用 Adobe Lightroom 修饰照片。在电影领域，视频的操纵通常需要非常复杂的视觉特效(VFX)，当涉及到重建动画面部与现实的面部肌肉运动和表情时，动作捕捉技术通常借助于高速跟踪标记，如知名电影《阿凡达》。随着诸如变分自动编码器(VAE)[1]和生成对抗网络(GAN)[2]等深度生成模型的显著发展与进步，现在普通用户也可以轻易生成身份不存在于真实世界的逼真面孔；或执行面部操纵如将视频中的人脸替换成另一个人的，这常被研究界称为 deepfake(深度伪造)。图 1 展示了生成的虚假人脸和人脸替换的效果，这些逼真的伪造图像让人难以区分。

图 1：(a)合成的虚假人脸，(b)人脸替换

Deepfake 技术本身是中立性质的，取决于使用者的意图。例如，将其用于视频会议场景[3]提升与会者的体验感。在最近几年，基于深度伪造技术的手机应用程序被广泛用于制作娱乐视频，并在社交媒体平台上病毒式传播，如 Zao 和 Reface。2021 年 3 月份，中国著名的抖音短视频平台充满了基于面部重现技术的“蚂蚁呀嘿”特效的娱乐视频。与此同时，一款名为 Avatarify 的手机应用程序在 Apple Store 中国区登顶，该程序提供了人脸操纵等深度伪造功能，然而很快被下架。这是因为深度伪造技术可能会被恶意者

利用，给个人、社会甚至国家带来消极影响[4]。例如，制造报复他人性质的色情图像或视频；将其用于政客发表不当言论或出席不适场所，在人们不知情的情况下操纵并干预选举过程；利用深度伪造技术破解人脸识别系统并进行金融欺诈等。这些都是恶意使用 deepfake 技术的例子。

许多国家的立法和管理机构都在用新的政策、法规来应对恶意深度伪造的扩散。2019 年，美国参议员提出了法案“S.2065 - Deepfake Report Act of 2019”，要求国土安全部科技局定时报告数字内容伪造技术的状况。其中数字内容伪造是利用新兴技术，包括人工智能和机器学习算法，以误导为目的制造或操纵音频、视觉或文本内容。中国国家互联网信息办公室等三部门颁发了《网络音视频信息服务管理规定》，针对网络视频的“深度伪造”进行管控，并已于 2020 年 1 月 1 日起施行。同时中国政府将恶意的深度伪造或虚假新闻定为刑事犯罪，许多国家也纷纷效仿。

深度生成模型仍处于蓬勃发展阶段，合成的虚假图像和视频的逼真程度不断提高，所以在互联网上有效地检测 deepfake 是一个极其困难的任务。这些深度伪造很难仅凭人眼去辨别，而传统的图像取证方法在对抗新兴的深度伪造时同样面临挑战。对此，近年来许多科技公司和学术团队投入研究，旨在从真实的图像或视频中鉴别出虚假的内容，即 deepfake 检测。虽然各种检测方法被提出，但仍有许多关键问题需要解决，如检测质量更好的 deepfake 的算法有效性，在不同数据质量场景下检测模型的鲁棒性，检测未知伪造方法的泛化性等。因此，开发出有行之效的检测模型或系统既是信息安全的研究热点，更是迫切且至关重要的需求。

1.2 国内外研究现状

近年来，研究人员不断开发各种技术与模型，以鉴别图像或视频的真实性。由深度学习生成的虚假图像或视频，直接的方法是也通过深度学习进行检测。使用或设计各种深度神经网络(deep neural network)一直是主流的 deepfake 检测方法。例如，Two-stream 方法[5]使用双流卷积神经网络(CNN)检测伪造图像，其中 CNN 是基于 GoogLeNet 的 InceptionV3 模型。MesoNet 模型[6]是一种基于 CNN 的针对图像介观属性的 deepfake 检测方法，并提出了使用传统卷积层的 meso4 和基于 Inception 模块的 mesoInception4。Multi-task 方法[7]使用 CNN 模型同时进行伪造图像检测与操纵区域分割的任务。Capsule

工作[8]研究了基于胶囊结构的 VGG19 网络作为 deepfake 分类模型。但这些方法都存在不同的局限性。比如缺乏泛化能力，即模型只为特定的一种图像操纵方法或单一数据集而设计的，不能推广到其它操纵方法或数据集。此外，它们缺乏可解释性，也就是说，我们很难理解什么信息被用于区分真假。所以一些工作研究了视频或图像中明显的视觉缺陷，如 deepfake 视频中的人脸缺乏闭眼的帧[9]；异常的头部姿势[10]；换脸的过程中留下的伪影[11]。这类方法同样只能针对特定场景进行检测，而且这些明显的缺陷已被改进的伪造方法[12]修复。

另一方面，泛化性受到了广泛的关注。Cozzolino 等人[13]设计了一个基于自动编码器结构的分类器，它可以跨不同的数据集检测虚假图像，但这些数据集的伪造方法较为相似。北京大学研究人员[14]提出的 Face X-ray 通过检测换脸中普遍存在的边界伪影以提升 deepfake 检测的泛化性，但是这种边界伪影很容易受到数据压缩的影响而弱化，这限制了其模型的鲁棒性。Wang 等人[15]通过实验表明，使用一个包含大量类别图像的 ProGAN 数据集，并通过适当的预处理和后处理训练一个卷积神经网络模型 ResNet 可以显著提升泛化能力，但是在检测换脸和面部重现等操纵方法伪造的 deepfake 上的有效性并未被表明。从这些研究中可以看出，提高泛化能力可以被认为是找到不同伪造方法生成的虚假图像和真实图像之间的普遍性差异，而目前的研究仍有局限性。

真实图像和虚假图像在频域上的差异是 deepfake 检测的另一突破口，许多研究[16][17][18]发现深度生成模型在合成模拟数据时普遍存在上采样步骤，并识别上采样结构导致生成图像在频率成分上的异常和频谱存在的棋盘状伪影。这些基于频域的方法通常利用傅里叶变换或离散余弦变换将图像从时域转换到频域进行频谱分析，并以此进行检测。Liu 等人[19]进一步证明了相位谱比振幅谱对上采样更敏感，并结合相位信息进行 deepfake 检测。然而对于一些没有上采样步骤的伪造方法，如 Face2Face[20]和 NeuralTextures[21]等面部重现方法主要在局部区域篡改人脸，频域异常不会在频谱中得到明显地暴露。对此，国内研究者提出的 F3- Net[22]面部伪造检测模型充分考虑了局部频域信息的重要性，它由两个频率感知分支组成，并在高压缩视频中实现了最先进的性能，但是仅挖掘过于细微的频域模式会受到过拟合的影响而不能保证泛化能力。

最新的一些检测方法受启发于 deepfake 难以模拟视频中潜在的生物信号。FakeCatcher 方法[23]提取了 6 种不同的生物信号，利用空间和时间相关性来验证摄像机

拍摄的真实视频。研究表明，心率可以用来检测假视频，但是从视频中获取心率是一项耗时的任务。Fernandes 等人[24]利用神经常微分方程对原始视频进行训练，对测试视频的心率进行检测。DeepFakesON-Phys[25]同样利用心率进行 deepfake 检测，使用远程光学体积描记术(rPPG)观察人类皮肤细微的颜色变化来说明血流的存在，并提出了一种卷积注意网络从视频中提取时空信息来检测 deepfake 视频。虽然合成微妙的生物学特征对于当前的深度生成模型是一个挑战，但基于生物信息的 deepfake 检测模型在准确性和泛化性上仍有待发展。

调研[26][27]与实验研究结果表明，当前的 deepfake 检测方法或模型仍面临如下挑战，同时这也是本文研究工作的目标：deepfake 检测器的有效性，即能否在最先进和反复优化过的 deepfake 上保持可观的检测效果；deepfake 检测器的泛化性，即检测模型能否推广到训练时未使用的数据集或伪造方法。近年来，一系列的研究正朝着这一目标努力，以开发出更通用的检测方法。然而过去的很多工作仅仅是在伪造效果相对较差的 deepfake 视频数据集上进行评估；deepfake 检测器的鲁棒性。在现实世界中，deepfake 很容易受到各种恶意攻击或无意扰动的影响，如图像或视频压缩、模糊处理等。

1.3 本文的研究内容与创新

本文的主要研究内容针对了现有的 deepfake 检测算法或模型的局限性，旨在提升检测的有效性、鲁棒性和泛化性，主要创新如下：

（1）现有的 deepfake 检测方法大多是基于图像的全局特征进行二分类，然而一些 deepfake 操纵方法只执行小规模的人脸篡改，这对全面地捕获微妙和局部的伪造伪影提出了挑战，需要提升模型检测不同操纵方法的有效性和在数据高压缩环境下的鲁棒性。对此，本文提出了一个基于空域与频域中局部伪造缺陷的 deepfake 检测算法，命名为局部伪影感知的 deepfake 检测网络，LA-Net(Local Artifact-aware Deepfake Detection Network)。该算法的研究旨在放大真实面孔与伪造面孔在空域和频域之间隐含的局部差异，而不是特定的外观特征。LA-Net 创新性地设计了局部风格提取模块(LSEB)，通过度量浅层特征映射之间的两两相关性对局部风格信息进行编码，从而在空间域中提取更具判别性的特征。此外，本文的研究发现细微的伪造缺陷可以进一步暴露在切片为 patch 的局部频谱中，并在幅度谱和相位谱中表现出不同的线索。根据幅度和相位信息的互补

性，为 LA-Net 设计了 Patch-wise 频谱交互注意模块(PFSCA)，用于捕获频域中局部相关的不一致性。通过在不同操纵方法和压缩场景下进行 deepfake 检测的大量实验证明了提出的方法的有效性和鲁棒性。

（2）为了提升 deepfake 检测模型的泛化性，本文提出了基于面部语义分割与对比分类的 deepfake 检测模型 FSCM。该研究基于伪造人脸虽然在不同的数据集和生成算法中迥然不同，但被操纵的主体内容是人脸语义或者特定的五官区域的现象，创新性地提出将人脸语义内容和非语义内容进行分割，并针对语义特征进行额外地对比学习。语义分割使能在语义内容中学习 deepfake 中更共同的伪造缺陷，对比学习通过强迫编码空间中同类特征更紧凑且异类特征的间隔更大进一步提升模型泛化性。数据集内和跨 6 个数据集（DeepfakeTIMIT、UADFV、FF++、DFD、DFDC、Celeb-DF）的实验评估结果证明了该模型对 deepfake 检测的有效性和以及显著改进的泛化能力。

1.4 论文的结构与组织

本文一共由 6 个章节组成，结构与内容安排如下：

第一章的绪论首先介绍了 deepfake 检测的研究背景，从技术与社会影响的角度阐明了课题的必要性和研究意义；然后描述了国内外 deepfake 检测研究的现状，通过对当前 deepfake 检测算法的归纳分析表明了现存的问题与面临的挑战；接着概述了本文提出的结合伪造缺陷与与语义对比的 deepfake 检测方法；最后说明了本文的结构与章节内容。

第二章主要介绍了相关技术的理论与研究。第一节阐述了两个知名的深度生成模型，自动编码器和生成对抗网络及其用于 deepfake 伪造的原理；为了明确 deepfake 内容带来的挑战以及当前检测算法的局限性，第二节对 deepfake 的伪造算法进行了介绍，第三节总结了 deepfake 检测相关的研究与方法。

第三章介绍了研究工作涉及的代表性 deepfake 数据集与后继实验前的数据预处理方法，以及本文提出的算法所采用的骨干网络 Xception 模型。

第四章提出了基于空域与频域中局部伪造缺陷的 deepfake 检测算法。首先对 deepfake 在空域的全局和局部伪造缺陷进行了分析，并设计了局部风格提取模块和空域分支网络；然后对频域的全局和局部伪造缺陷进行了分析，并设计了 patch-wise 频谱交

互注意模块和频域分支网络；接着介绍了提出的算法模型 LA-Net 的总体结构与细节；最后通过实验表明了本章方法检测不同伪造方法的有效性和在不同压缩场景中的鲁棒性。

第五章提出了基于语义分割与对比分类的 deepfake 检测算法。首先对人脸语义分割的必要性和具体方法进行了介绍；然后说明了监督对比学习的原理并结合面部语义特征，同时进行对比学习和分类学习；接着介绍了提出的算法模型 FSCM 的总体结构与细节；最后通过对比分析与跨数据集检测的实验评估证明了本章方法的有效性和显著提升的泛化能力。

第六章对全文进行了总结并展望了未来的研究工作与计划。

第二章 相关技术研究与概述

本章将主要介绍本文涉及的相关理论与技术。第一部分首先阐述了用于 deepfake 伪造的深度生成模型的原理。第二部分对 deepfake 伪造的内容和代表性技术进行了分析，以了解对检测任务带来的挑战，第三部分归纳并总结了 deepfake 检测相关的研究与方法。

2.1 深度生成模型

在过去的几年中，深度学习在生成建模任务中取得了显著的进展与应用，如 NVIDIA 公司提出的 StyleGAN[28]能创建超逼真的人脸图像，以及 OpenAI 的 GPT-2[29]语言模型能在给出简短的介绍后完成一段文本内容。但深度生成模型伪造的 deepfake 内容也引发了信息的信任和安全问题。本节将主要介绍目前最为基本和知名的两个深度生成模型，自动编码器(autoencoder)和生成对抗网络(generative adversarial network)，以了解 deepfake 伪造背后的基础与原理。

2.1.1 自动编码器及其换脸原理

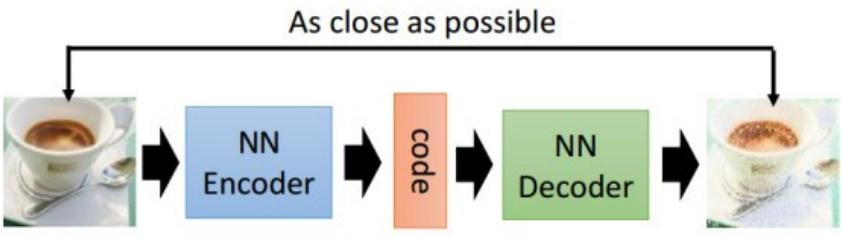

图 2：自动编码器的结构

自动编码器是一个由两部分组成的神经网络，如图 2 所示：第一部分是将高维输入数据压缩成低维向量编码表示的编码器(encoder)网络；第二部分是将给定的向量编码表

示还原到原始域的解码器(decoder)网络。自动编码器的训练目标是为网络寻找权重，使得原始输入和经过编码器-解码器后的重构输出之间的损失最小化。编码向量是将原始图像压缩到一个低维的潜在空间，其思想是通过选择潜在空间中的任意点，然后通过解码器生成新的图像。所以解码器的目标是学会如何将潜在空间中的点转换为可接受的图像。

由于在自动编码器中，编码向量与输入数据的映射是一个确定性的过程，即每个图像被直接映射到潜在空间中的一个点，所以自动编码器的生成能力有限。为了解决这个问题，变分自动编码器(variational autoencoder)[1]将每个图像映射到潜在空间的一个点周围的多元正态分布$N(\mu, \varepsilon^2)$中，如图 3 所示。变分自动编码器通过正态分布同时引入数据信息和噪声。通过模型中引入的随机性和约束潜在空间中点的分布，变分自动编码器可以用不同的编码向量生成更多的新数据。

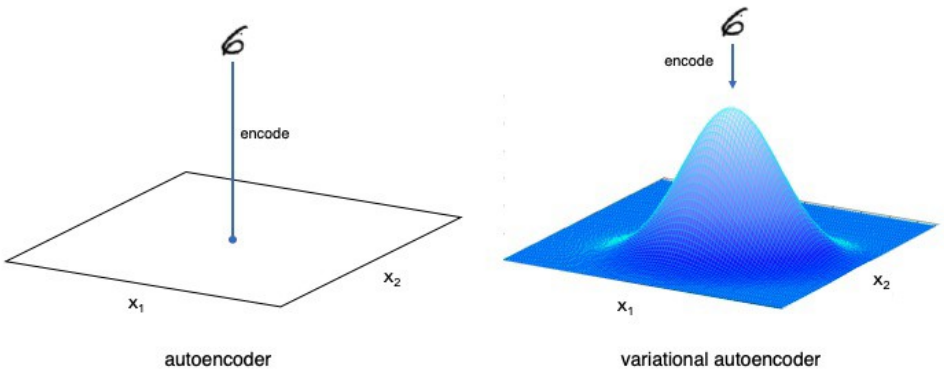

图 3：自动编码器和变分自动编码器在潜在空间中的数据映射

一些创建 deepfake 的应用如 fakeapp，它们的人脸交换模型便是基于自动编码器或变分自动编码器，其基本过程和原理如图 4 所示：共享的编码器对人脸 A 和 B 进行训练并提取潜在特征向量，然后由各自的解码器分别完成对应的人脸重构。压缩后的潜在向量可以理解为人脸的宏观信息，如表情和背景。解码器用于从潜在向量恢复特定的面部特征，如鼻子细节。经过训练后，将共享编码器从 A 中提取的潜在向量输入 B 的解码器，最后使能 A 人脸被 B 人脸替换。

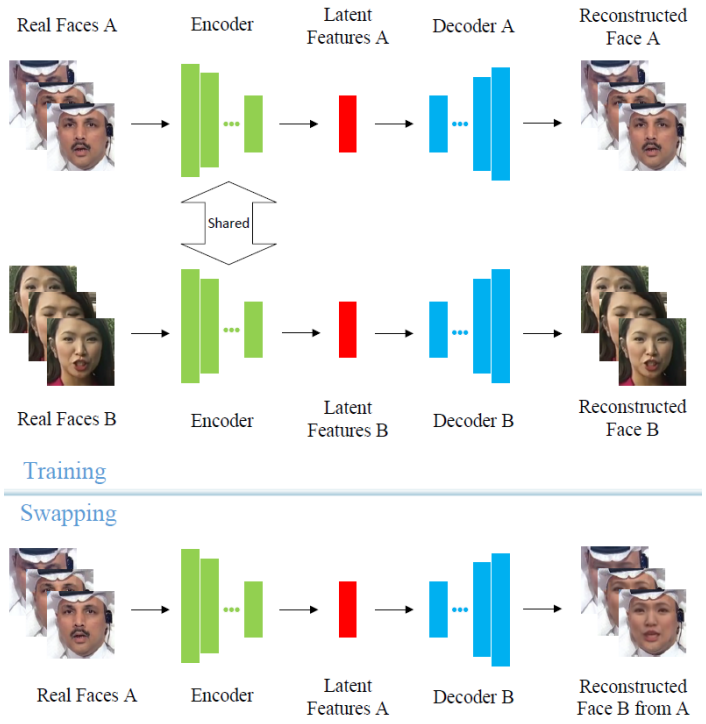

图 4：基于自动编码器的换脸过程

2.1.2 生成对抗网络与人脸生成原理

2014 年，Ian Goodfellow 等人[2]首次提出了生成对抗网络(GAN)，将深度生成模型的领域研究推向了更大的高度。GAN 可以做两个组件之间的零和博弈，即生成器和判别器。生成器试图将随机噪声转换为看起来像是从训练数据集采样的观察结果，判别器则试图预测某个观察结果是来自训练数据还是生成器的伪造结果。生成对抗网络的基本结构如图 5 所示。

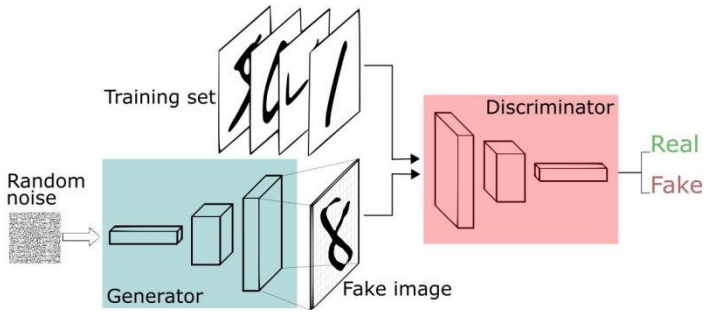

图 5：生成对抗网络的结构

假设$E(*)$代表分布函数的期望，$P_{data}(x)$表示训练数据即真实样本的分布，$P_{noise}(x)$是噪声分布，G和D分别代表生成器和判别器，那么生成对抗网络的目标函数可以表示为：

$$min_G max_D V(D,G) = E_{x\sim P_{data}(x)}[logD(x)] + E_{x\sim P_{noise}(z)}[log\,(1-D(G(z)))]. \quad (2\text{-}1)$$

在训练过程开始时，生成器输出有噪声的图像，而判别器接近于随机预测。生成对抗网络学习过程的关键在于如何交替训练两个组件，当生成器变得更擅长欺骗判别器时，判别器需要得到更新以保持其正确分类数据的能力，这反过来促使生成器更新网络参数以欺骗鉴别器。经过充分迭代训练后，判别器无法区分生成的数据和真实数据，这意味着生成器能很好的拟合训练数据的分布甚至生成逼真的伪造内容。

如今，基于生成对抗网络的人脸生成方法日益成熟。图 6 展示了合成人脸的进步，可以看出如今的 PGGAN[30]和 StyleGNA[28]足以生成人眼难以辨别的图像。PGGAN 通过逐步深化网络层的数量来生成分辨率更高的图像，最终生成分辨率为 1024×1024 的高分辨率图像。StyleGAN 基于 ProGAN 修改网络中每个层级的输入，在控制该级别表示的视觉特征的同时不影响其他级，从而实现由粗粒度到细节的优化。StyleGAN2 进一步修复了 StyleGAN 生成图像的伪影并提高视觉质量。

图 6：生成对抗网络合成的人脸

2.2 Deepfake 伪造技术简介

Deepfake 一词可以追溯到 2017 年底，当时一位名为 deepfakes 的 Reddit 用户利用深度神经网络改变了色情视频的人脸。现在，deepfake 一般指由各种深度学习技术生成的虚假面部图像或视频，主流的技术方法包括 FakeApp，Faceswap[31]，Face2Face[20]和

NeuralTextures[21]等。了解伪造原理有助于设计更有效的检测模型，其中 FakeApp 基于自动编码器进行换脸，它的原理已在 2.1.1 小节进行了介绍，所以本节介绍了其它 deepfake 伪造技术。

2.2.1 Faceswap 算法

Faceswap[31]是一种基于计算机图形学的方法，它使用人脸对齐、高斯牛顿优化和图像融合等进行人脸交换。该算法首先获取输入图像的人脸区域及其坐标。然后利用 3D 模型最小化投影形状和局部坐标之间的差值来将输入图像反投影到目标图像，其中高斯牛顿法被用于实现混合形状权值和仿射变换的最小化。最后渲染模型通过使用 alpha 混合和颜色校正技术与输入图像融合。这个算法相比基于深度生成模型的方法更为轻量，但存在分辨率低、伪造伪影和篡改痕迹较为明显等缺点。

2.2.2 Faceswap-GAN 算法

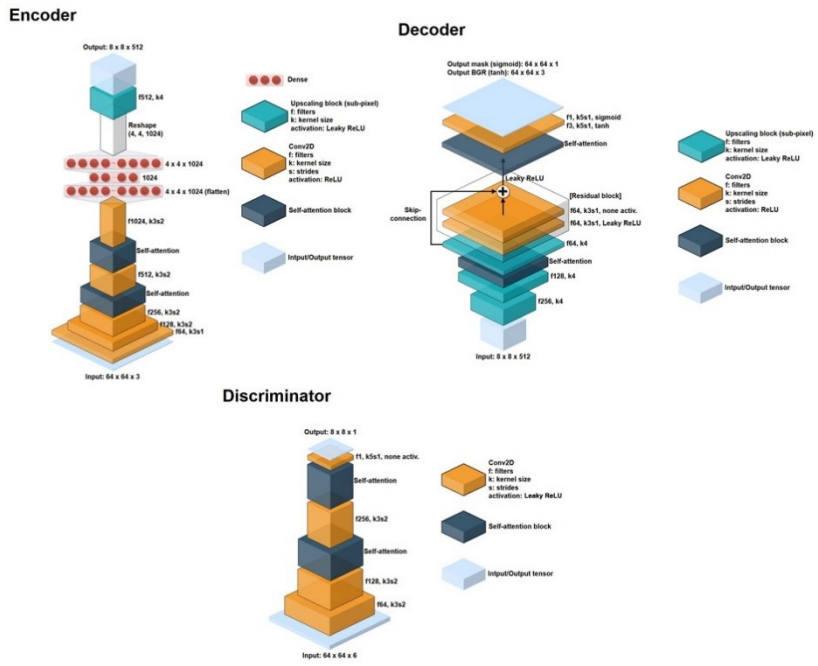

图 7：Faceswap-GAN 网络结构

基于图 4 所示的自动编码器架构，Faceswap-GAN 进一步引入了对抗损失和感知损

失。该模型的编码器，解码器和判别器的结构细节如图 7 所示，其生成网络采用了 CycleGAN。感知损失通过改进眼球方向确保替换人脸与输入人脸之间的一致性，并平滑掩码(mask)中的伪影提高真实度，其中注意力掩码(mask)用于协助模型对遮挡，伪影和肤色的处理与优化。Faceswap-GAN 在转换视频时使用 MTCNN 和卡尔曼滤波器进行人脸跟踪与对齐，MTCNN 能实现更稳定的人脸检测和对齐，卡尔曼滤波器则平滑帧上的边界框位置，并消除换脸产生的抖动。总而言之，该算法进一步提升了 deepfake 人脸的伪造质量。

2.2.3 Face2Face 面部重现

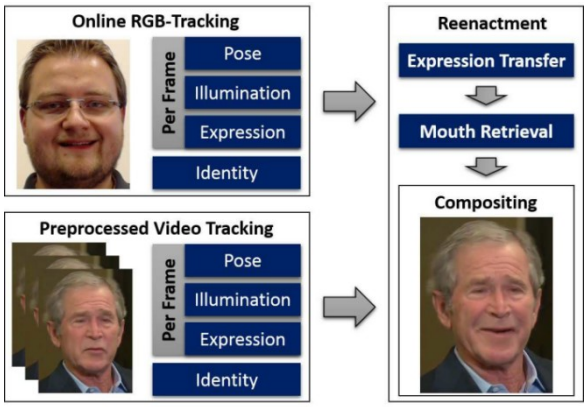

图 8：Face2Face 操纵流程与效果

Face2Face 是 Thies 等人[20]提出的一种面部重现技术，常用于 deepfake 内容中的表情操纵。该方法可以实时地将源视频中的人脸表情迁移到目标视频，同时保持目标人脸的身份不变。Face2Face 基于 3D 人脸模型进行拟合，操纵流程和效果如图 8 所示。它首先利用光流一致性特征来跟踪源视频和目标视频的面部表情。然后通过变形传递函数将源脸的特征参数包括姿势、光照和表情快速地转移到目标脸。为了确保精确的匹配，该算法不但从目标序列中检索并扭曲获得最合适的口腔嘴型，同时与目标视频重新进行了光照渲染。Face2Face 算法在操纵人脸上表现出了高效性、真实性以及良好的嘴部形状和光照效果。

2.2.4 NeuralTextures 渲染方法

NeuralTextures[21]是一种可被用于表情重现的渲染方法，它为目标人脸从原始视频数据中学习神经纹理和渲染网络。神经纹理和渲染网络都是端到端的训练方式，其中神经纹理是学习到的特征映射并结合 3D 内容表示，使能在 3D 空间中显式地控制输出。该方法可用于重新渲染或操纵视频内容。相比 Face2Face，NuralTextures 的操纵人脸在嘴部区域表现的更多样化，有着更丰富的细节和更少的篡改伪影。在知名的 deepfake 数据集 FaceForensics++中，该方法操纵的人脸仅篡改了面部表情对应的嘴部区域，保持其它区域如眼部不变。这种小规模操纵的伪造人脸是 deepfake 检测中更难的任务，尤其在视频高压缩场景下。

2.3 Deepfake 检测方法研究

随着 deepfake 内容越来越受到人们的关注，研究机构和各大公司提出了各种方法来应对日益严峻的威胁。然而，改进的合成算法、小规模篡改和高压缩场景仍然对现有检测方法的有效性、泛化性和鲁棒性提出挑战。这一小节将归纳并分析 deepfake 检测相关的研究与方法。

2.3.1 基于空域的检测方法

早期的面部操纵技术容易产生明显的篡改伪影，这启发了许多探索空间域异常的检测方法，如 deepfake 中缺乏眨眼的帧[9]、异常的头部姿势[10]、面部交换过程留下的扭曲伪影[11]和视觉伪影[32]。然而，改进的生成算法显著减少了这些可见的伪影，导致这些方法的检测性能大幅下降[12]。Face x-ray[14]进一步观察了换脸的混合步骤并定位边界，通过对人脸融合拼接步骤的建模提升了检测能力和泛化性能。然而，在高度压缩的视频场景中，它的性能受到弱化的换脸边界的影响而下降。

得益于深度学习的特征自动提取能力，许多基于卷积神经网络(convolutional neural networks)的方法都以 RGB 像素为输入，将人脸伪造检测视为普通二值分类问题。MesoNet[6]引入了两种基于图像介观特性的 CNN 体系结构来检测伪造人脸。Nguyen[8]等人将胶囊网络与预先训练的 VGG19 模型相结合，用于人脸取证。Rossler 等人[33]发布

了 deepfake 操纵检测基准 FaceForensics++(FF++)，并展示了 Xception[34]作为骨干网络对 deepfake 的有效性。考虑到视频流中的时间信息，一些研究工作采用了循环神经网络模型，如 LSTM[35][36]和 GRU[37]。

不同于盲目地直接使用深度学习的分类模型作为检测器，最近的一系列方法侧重于结合伪造缺陷设计模型，这同样也是本文的研究思路。阿里巴巴公司的研究者们[38]提出利用使用双流分支网络同时提取 RGB 特征和噪声信息用于伪造检测。Dang 等人[39]提出通过可学习的注意图来突出伪造区域，同时进行篡改定位检测。Liu 等人[40]利用全局纹理信息来检测完全由 GANs 生成的图像，但在对局部篡改的 deepfake 的检测性能可能较差。然而，大多数基于空间的检测方法都容易受到压缩设置的影响，因为微妙的伪造痕迹对质量敏感。此外，伪造伪影在色彩空间中呈现出迥然不同的形式，这也对这些模型的泛化能力提出了挑战。因此一些研究工作转而挖掘 deepfake 的频域线索。

2.3.2 基于频域的检测方法

在数字信号处理和计算机视觉领域，频域分析一直是一种经典而强大的方法，它也被用于图像伪造的检测。实际上，深度生成模型的上采样结构会引入频域异常，如高频分量的分布差异[42]和棋盘伪影[17]。Durall 等人[18]提出通过平均离散傅里叶变换的振幅谱再利用分布差异来检测虚假图像。AutoGAN 模型[16]模拟了 GAN 在频域产生的棋盘伪影，并将傅里叶变换获得的频谱输入检测器。Frank 等人[17]通过离散余弦变换将图像从空间域转换到频域，分析各种 GAN 生成图像的频率伪影。这些方法对于完全由 GAN 合成的图像检测确实是有效的。然而，Face2Face[20]和 NeuralTextures[21]等 deepfake 操纵方法并不会在全局频域引起明显的异常，所以这些方法检测小规模篡改的 deepfakes 性能大大下降。对此，F3-Net[22]设计了频率感知分解模块并利用局部频率统计在频域中挖掘细微的伪造模式，并在高压缩视频场景报告了领先的 deepfake 检测结果，但它在未见数据集评估中的泛化性能表现不佳。

为了综合且互补地提取特征，同时考虑空域和频域信息逐渐成为deepfake检测模型研究的主流。SSTNet[43]利用改进的Xception和LSTM提取空间特征、隐写分析和时间特征，检测经过处理的人脸图像。Two-branch RN[44]结合了空间域和频域的信息，并利用

拉普拉斯高斯算子(LOG)增强了多频段的频率。然而，使用固定的过滤器或手工提取的固定特征限制了可判别性。最近，2021年CVPR会议论文SPSL[19]观察到上采样结构会导致deepfake与真实图像在相位谱上的差异比在振幅谱上更大，并提出将相位信息与浅层特征图相结合，提高了人脸伪造检测的泛化能力。然而，SPSL也难以在全局相位谱中捕获局部操作方法引起的小尺度频域伪影，限制了其检测性能和鲁棒性。

2.4 本章小结

本章首先对自动编码器和生成对抗网络的机理进行了详细的介绍，以及这些深度生成模型如何实现 deepfake 内容操纵。然后进一步介绍了 deepfake 相关的伪造技术。通过对 deepfake 生成与篡改原理的理解，为后文的检测算法设计提供了突破口与思路。最后对现有的 deepfake 检测方法进行了分析研究，阐明了当前研究的局限性和面临的挑战，同时为本文在 deepfake 检测上的研究明确了目标。

第三章 实验数据与检测模型骨干网络

为了促进 deepfake 检测的研究，许多深度伪造视频基准数据集已经发布。另一方面，Xcpetion 作为经典而强力的深度卷积神经网络，因其有效性而被许多 deepfake 检测算法的模型用作骨干网络。所以本章介绍了研究工作涉及的代表性 deepfake 数据集与后继实验前的数据预处理方法，以及本文提出的算法所采用的骨干网络模型 Xception。

3.1 实验数据集

3.1.1 数据集简介

UADFV 数据集[9]包括 49 个原始视频和 49 个对应的 deepfake 伪造视频，其中视频是使用 FakeAPP 的自动编码器模型生成的。

DeepfakeTIMIT 数据集[45]基于 faceswap-GAN 对 Vid-TIMIT 数据集生成了 320 个 deepfake 视频，并根据 64×64 像素和 128×128 像素分别提供了 LQ 版本 HQ 版本。本文的实验对该数据集选用了伪造质量更好的 HQ 版本。

FaceForensics++[33]由 1000 个来自 YouTube 的原始视频和 4000 个分别由 DeepFakes(DF)、Face2Face(F2F)、FaceSwap(FS)、NeuralTextures(NT)四种常用的人脸操纵方法伪造的视频组成。FF++中的每个视频根据压缩级别分为三个版本：c0 未经压缩(RAW)、c23 轻度压缩(high-quality、HQ)和 c40 重度压缩(low-quality、LQ)。

DFD[46]是 Google 公司的 Jigsaw 等人发布的 deepfake 检测数据集。该数据集基于 28 个不同性别、年龄和种族的参与者，由 363 个原始视频生成了 3068 个 deepfake 视频。合成算法的细节尚未披露。在论文的实验中选用了未经压的 RAW 版本。

DFDC[47]是 Facebook 主半的 Deepfake Detection Challenge 大规模 deepfake 检测竞赛所提供的数据集。该数据集基于 66 个不同性别、年龄和种族的参与者，由 1131 个原始视频生成了 4113 个 deepfake。通过在任意的背景录制视频，其场景和种类相比 DFD 更为丰富。

Celeb-DF[12]是纽约州立大学和中国科学院大学的 Li 等人发布的高质量 deepfake 数据集，包含了 5639 个的 deepfake 名人视频。Celeb-DF 视频的平均长度为 13 秒，帧率为 30 FPS (frame per second)。

3.1.2 数据集对比与分析

表 1 总结了 deepfake 数据集的数量、操纵方法等信息，其中 UADFV，DeepfakeTIMIT 和 FaceForenscics++被分为第一代 deepfake 视频数据集，而第二代则包括 DFD、DFDC 和 Celeb-DF。

表 1：Deepfake 数据集信息

数据集	真实视频数/帧数	伪造视频数/帧数(操纵方法)	发布时期
UADFV[9]	49/17.3k	49/17.3k(FakeAPP)	2018
DeepfakeTIMIT[45]	320/34.0k	320/34.0k(Faceswap-GAN)	2018
FaceForensics++[33]	1000/509.9k	1000/509.9k(DeepFakes) 1000/509.9k(Face2Face) 1000/509.9k(FaceSwap) 1000/509.9k(NeuralTextures)	2019
DFD[46]	363/315.4k	3068/2242.7k(Unknown deepfake)	2019
DFDC[47]	1131/488.4k	4113/1783.3k(Unknown)	2019
Celeb-DF[12]	590/225.4k	5639/2116.8k(Improved Deepfake)	2020

第二代 deepfake 数据集的改进主要针对了 deepfake 视频中合成人脸分辨率低、合成人脸与原始视频颜色不匹配、人脸掩码不精确以及每帧之间的时间抖动等问题。升级的生成算法和处理技术显著改进了视觉质量，在大幅提高视频数量的同时扩展了多样性，对 deepfake 的有效检测带来了更严峻的挑战。

3.2 数据预处理

遵循现存的 deepfake 检测研究工作，本文将伪造检测作为帧级别的二值分类问题。由于 deepfake 视频数据集的规模庞大，提取全部图像进行训练会带来巨大的时间和计算

开销。而且临近的帧之间存在着大量的冗余信息，所以有必要合理地提取视频中的帧。本文利用了计算机视觉中知名的 opencv-python 库进行视频取帧，首先通过调用 cv2.VideoCapture 函数读取数据集中的视频，然后通过 for 循环控制帧的提取数量与间隔。

考虑到 deepfake 伪造算法主要篡改的是人脸信息，而人物背景中的无用信息会干扰检测模型对判别性特征的学习。为了提高检测的准确性和效率，本文同样与大多数检测方法进一致，仅保留检测到的人脸图像作为实验数据。人脸检测使用了 Dlib 工具包，它提供了机器学习、图像处理等诸多算法和工具，被广泛用于工业和学术界。具体地，首先运行了一个人脸检测器 face_detector，然后识别图像中的人脸特征点并返回坐标值。注意在本文实验中，为了提升效率，人脸检测直接对提取的帧进行，并仅保存最后检测并对齐的人脸图像，其分辨率为 256×256。视觉流程如图 9 所示。

图 9：数据处理流程

3.3 Xcepton 网络

Xception[34]是由 Google 研究人员开发的一种基于深度可分离卷积(depthwise separable convolution)的深度卷积神经网络，是图像识别的强力模型。鉴于 Xception 在 deepfake 检测任务上的出色性能，本文第四章和第五章提出的方法将其作为了骨干网络。

3.3.1 深度可分离卷积

如图 10 所示，原始的深度可分离卷积是 depthwise 卷积后跟一个 pointwise 卷积。其中 depthwise 卷积是在通道方向上进行 n×n 空间卷积，在图中有 5 个通道，所以进行

了 5 个 n×n 空间卷积。而 pointwise 卷积实际上是利用 1×1 卷积来改变通道（channel）维度。与传统卷积相比，由于不需要在所有通道上执行卷积，这意味着更少的连接数和计算开销。

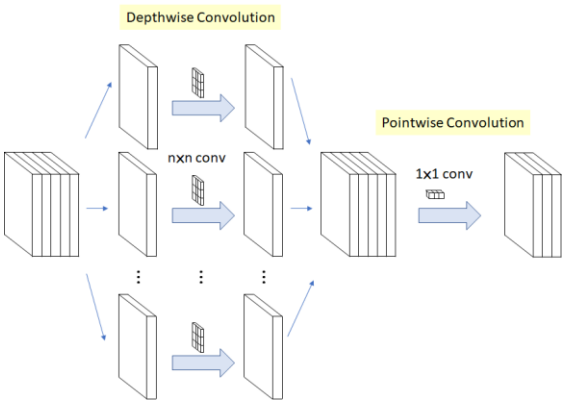

图 10：深度可分离卷积

Xception 假设跨通道相关性和空间相关性可以完全分开映射，采用了改进的深度可分离卷积用作 Inception 模块的“极端”版本，如图 11 所示。改进的深度可分离卷积是 pointwise 卷积后跟一个 depthwise 卷积。这种修改方式受到了 Inception-v3 中的 inception 模块启发，该模块在 n×n 空间卷积之前先进行 1×1 卷积。主要区别在于：原始的深度可分离卷积的实现是先在通道方向上进行空间卷积，然后执行 1×1 卷积，而改进的深度可分离卷积首先执行 1×1 卷积，然后在通道上进行空间卷积；在最初的 Inception 模块中，第一次操作后使用 Relu 函数进行非线性激活，而 Xception 中改进的深度可分离卷积没有中间的非线性激活。作者对改进的深度可分离卷积测试了不同的激活单元，并表明不采用任何中间激活的 Xception 具有更高的准确度和更快的收敛速度。

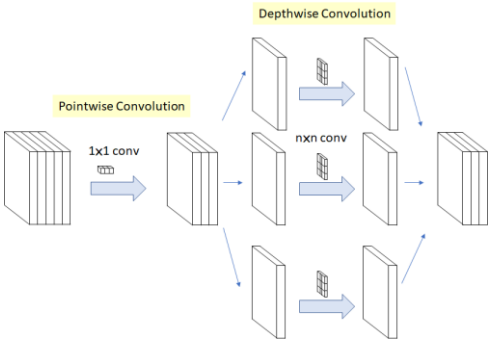

图 11：改进的深度可分离卷积

3.3.2 网络架构

Xception 的网络结构如图 12 所示，它被分为 3 个 flow，包括 Entry flow, 重复 8 次的 Middle flow 和 Exit flow。其中 SeparableConv 就是改进的深度可分离卷积，注意所有卷积层和深度可分离卷积之后都进行批归一化(batch normalization)。此外可以看出，Xception 网络使用了 ResNet[48]提出的残差连接，即图中的“跳层”连接方式。残差学习能在加深网络深度的同时改善梯度消失问题，相比简单叠加网络层更容易优化。

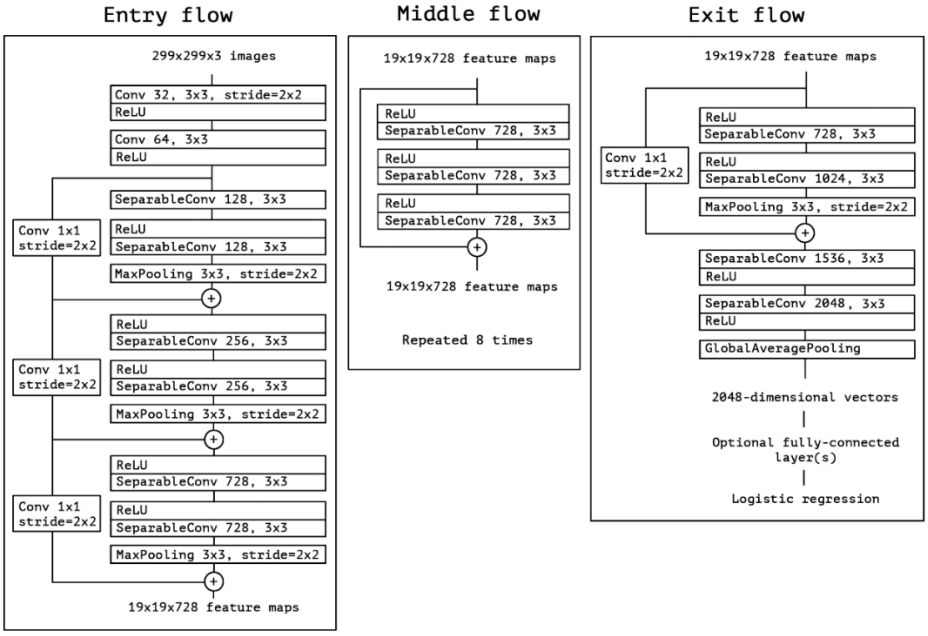

图 12：Xception 的网络结构

3.4 本章小结

本章主要对 deepfake 代表性的数据集以及数据预处理操作进行了介绍，为后继实验与研究评估的理解提供基础。此外，还对本文后继提出的方法中所使用的骨干网络模型 Xception 的原理与结构进行了阐述。

第四章 基于空域与频域中局部伪造缺陷的 deepfake 检测算法

得益于规模越来越大的神经网络在图像识别领域取得的显著成功，各种基于深度神经网络的 deepfake 检测方法被提出[5][6][8][33][35][37]。这些算法模型大多遵循如图 13 所示的流程设计，即直接将待检测的图像输入骨干网络以提取整张图像的全局特征，然后通过二值分类器进行决策。然而，一些先进的伪造方法只在局部进行了小规模篡改，如基于 NeuralTextures[21]技术伪造的人脸仅在嘴部区域被篡改。所以直接利用深度神经网络最后一层输出的全局特征可能难以有效检测一些局部伪造的 deepfake。

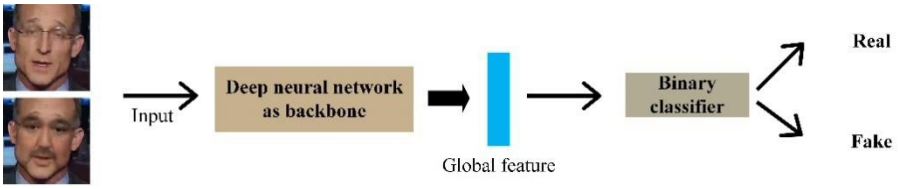

图 13：Deepfake 检测模型常见的流程设计

在另一方面，色彩空间中的伪造伪影很容易受到数据压缩的影响，在高压缩场景下仅利用空域特征的检测模型缺乏鲁棒性。频域信息为挖掘伪造缺陷提供了另一个视角，之前的研究[17][42]表明生成器的上采样结构会导致频域的棋盘状伪影。这些检测方法通常使用傅里叶变换或离散余弦变换来获取频谱，然后将其作为检测器的特征输入。然而，当遇到没有上采样结构的操纵方法时，它们的性能会大幅下降。此外，现有的基于频域的方法大多直接将整个图像转换为频谱进行分析，一些局部被篡改的人脸在全局频域中同样难以显示出可识别的伪影。

综上所述，本章的研究动机主要包括如下三个方面:

（1）除了深度神经网络提取的高层语义特征，还需要在空域中捕获细微的篡改痕迹。并且提取的局部模式要趋于常见的伪造缺陷，而不是特定的操纵方法。

（2）局部伪造方法篡改的有限像素也不会引起明显的频率成分异常，因此在频域

中挖掘的伪造线索也应趋于局部化。

（3）在频域中，仅利用幅度或相位信息是不全面的，鉴于两者的互补性，幅度和相位信息可以被协同利用。

因此，本章提出了一种结合空域与频域中局部伪造缺陷的 deepfake 检测算法，命名为 LA-Net，旨在同时从空域和频域放大真实和伪造人脸之间隐含的局部差异，而不是特定的外观特征。具体而言，本章设计的局部风格提取模块编码浅层特征映射之间的相关性，从而在空间域提取更显著的局部伪造痕迹。此外，本章还观察到，微妙的伪造伪影可以进一步暴露在切分为 patch 后的相位谱和振幅谱中，并显示出不同的线索。根据幅度和相位信息的互补性，本章开发了 Patch-wise 频谱交互注意模块，用于捕捉频域内局部相关的不一致性。4.1 节和 4.2 节分别分析了空域与频域中的全局和局部伪造缺陷，并设计了对应分支的检测算法与模块。4.3 节则介绍了提出的 LA-Net 及其模型架构与细节。4.4 节提供了模型的实验结果与分析。4.5 对本章进行了小结。

4.1 空域伪造缺陷与局部风格提取模块

4.1.1 空域的全局和局部伪造缺陷分析

用于身份互换的 deepfake 伪造方法通常将视频或图像中的人脸替换成另一个人脸，同时保持背景内容不变。如图 14(a)中的两对换脸图像所示，换脸算法的篡改操纵会在整个人脸上产生明显的伪造缺陷，例如眼睛、额头等区域的伪影。因此基于深度神经网络提取的图像全局特征检测换脸 deepfake 是行之有效的。然而另一些伪造算法，如表情互换和属性操纵方法仅篡改了人脸中的一些局部区域，如图 14(b)所示，我们难以从整张图像中观察到明显的伪造伪影。通过进一步观察图像中的局部细节，图 14(b)左列人脸边界中还是存在细微的锯齿状伪影，而右列人脸中仅嘴部区域出现牙齿轮廓丢失的缺陷。所以针对小尺度篡改的 deepfake 进行检测不能仅利用图像在空域中的全局特征，还应进一步结合利用局部伪造缺陷。

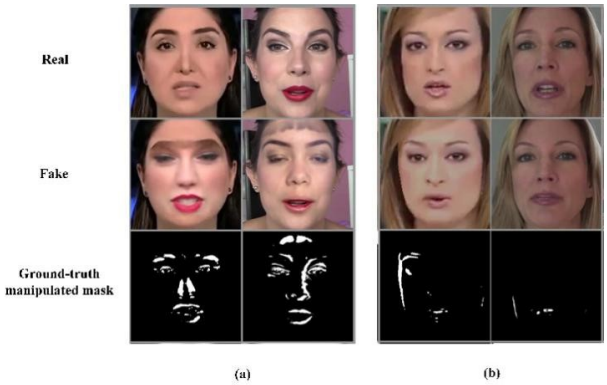

图 14：真实图像，对应的 deepfake 图像与操纵掩模

在 deepfake 检测任务中，许多方法直接使用神经网络最后一层的输出连接分类器进行检测，即利用全局特征进行决策。神经网络的浅层具有较小的感受野，并倾向于在局部区分低级特征如颜色和纹理，而深层则更倾向于高层次的抽象特征。一个具体的例子如图 15 所示，在该实验中使用了一个简单的卷积神经网络并以全连接层作为 deepfake 检测的二分类器，人脸图像经过训练好的神经网络并被输出不同层的特征图（features maps）。可以观察到，网络第一层输出的特征图通常被小规模激活并鉴别眼睛、鼻子和嘴巴等局部区域。而随着网络层次不断加深，特征图的激活区域变得越来越大，这意味着更高层次的全局特征被编码学习。所以对于只修改人脸指定区域的 deepfake 操纵方法而言，在浅层特征图中的伪造缺陷通常具有更强的判别性，这为本节提出的局部风格提取模块带来了启发。

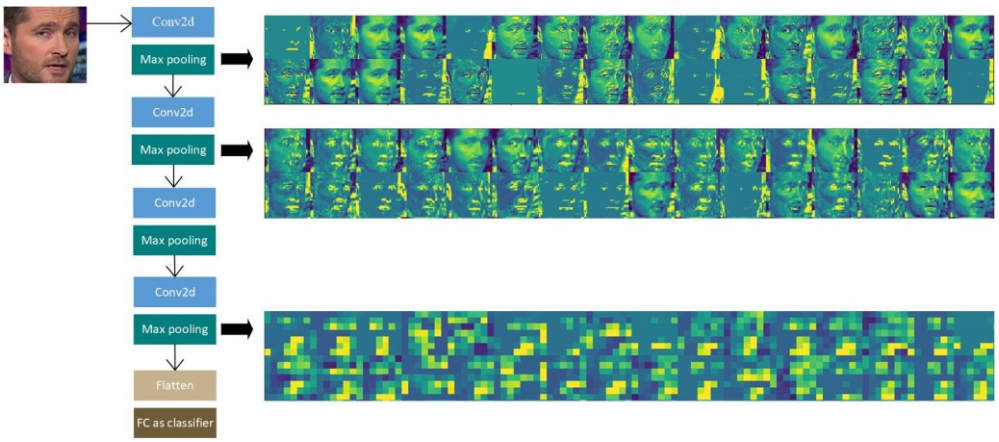

图 15：卷积神经网络中特征图的激活可视化

4.1.2 局部风格提取模块设计

基于章节 4.1.1 的现象观察与实验结果，局部风格提取模块(LSEB)用于提取多尺度的空域局部特征。LSEB 的原理是在骨干网络的浅层计算特征图各自的格拉姆(Gram)矩阵。Gram 矩阵在深度神经网络中通常被用于编码风格属性[49][50]，例如纹理模式。具体地，它对第l层向量化后的特征图F_i^l和F_j^l计算点积：

$$G_{ij}^l = \sum_k F_{ik}^l F_{jk}^l = \begin{bmatrix} {F_{i1}^l}^T F_{j1}^l & \cdots & {F_{i1}^l}^T F_{jk}^l \\ \vdots & \ddots & \vdots \\ {F_{ik}^l}^T F_{j1}^l & \cdots & {F_{ik}^l}^T F_{jk}^l \end{bmatrix}, \tag{4-1}$$

Gram 矩阵G_{ij}^l实质上是一个偏心协方差矩阵，即神经网络第 l 层中第 i 个和第 j 个通道上的特征图的协方差矩阵，但是没有减去均值。Gram 矩阵实际上衡量了特征图两两之间的互相关性：对角线元素提供了不同卷积核各自的响应，其余元素表明了不同卷积核响应之间的相关程度。这种强调相关性的响应模式有助于捕捉不同特征映射之间的伪造纹理，因为 deepfake 中的视觉伪影[32]或人脸扭曲伪影[11]可能在多个局部区域出现相似的模式。

LSEB 的具体结构设计见图 16，它在骨干网络的第 l 层截断特征图$F_l \in R^{H\times W\times C}$作为输入，并将$F_l$优化为特征的向量表示。由于计算出的 Gram 矩阵G_{ij}^l存在冗余关联性且很难在多个语义层被对齐（网络不同层所计算的 Gram 矩阵大小不一致），LSEB 通过全局平均池化层(GAP)聚集每个特征图即G_{ij}^l每行的累积相关性。最后，LSEB 输出局部风格特征$LTF \in R^{c\times 1}$。

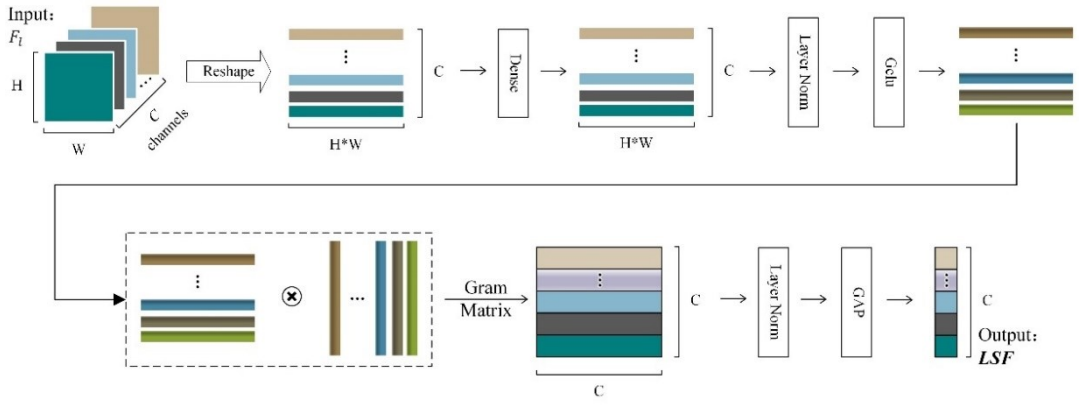

图 16：局部风格提取模块(LSEB)的结构设计

4.1.3 空域分支网络

经典的深度神经网络 Xception 在 deepfake 检测任务中取得了显著的性能[33]，所以其被用作骨干网络。Xception 网络结构由 14 个 block 组成，LSEB 分别在 Xception 的浅层 block1，block2 和 block3 截断感受野较小的特征图，并提取具有判别性的局部纹理特征LSF_{b1}，LSF_{b2}和LSF_{b3}。这些浅层风格特征作为 block14 最后输出的全局语义特征GF的补充，鼓励网络在空域分支网络捕获更微妙的局部纹理特征。

4.2 频域伪造缺陷与 patch-wise 频谱交互注意模块

4.2.1 频域的全局和局部伪造缺陷分析

在数字信号处理和计算机视觉领域，频域分析是一种经典而强大的方法，它同样被用于 deepfake 的检测。Deepfake 通常采用了各种深度生成模型，如生成对抗网络(GAN)和自动编码器。深度生成模型的上采样步骤会导致频域异常，包括如图 17 所示的(a)高频分量差异和(b)棋盘状伪影。有些方法[16][17][18]直接从整个图像中通过离散傅里叶变换(DFT)或离散余弦变换(DCT)提取频谱，然后训练分类器来检测各种 GAN 生成的图像。值得一提的是，大部分方法仅从图像的幅度谱提取频域信息，Liu 等人[19]进一步证明了相位谱比振幅谱更易受上采样步骤的影响并将相位信息用于 deepfake 检测。

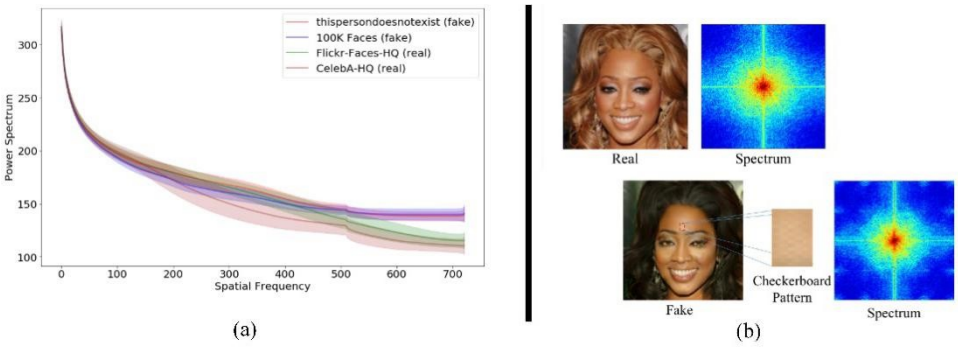

图 17：生成图像的(a)高频分量差异和(b)棋盘状伪影

然而，在一些基于计算机图形学方法处理的人脸[31]和改进的 deepfake 视频中[12][21]，伪造伪影在整个图像的全局频谱上并未被显著暴露。在图 18 中提供了一个视觉例子，

伪造图像仅篡改了原始图像的局部区域，无论是在空域的 RGB 图像上还是在频域的幅度谱和相位谱中，伪造缺陷在全局上表现的极其细微，残差谱则进一步表明了频谱上的差异极其细微。

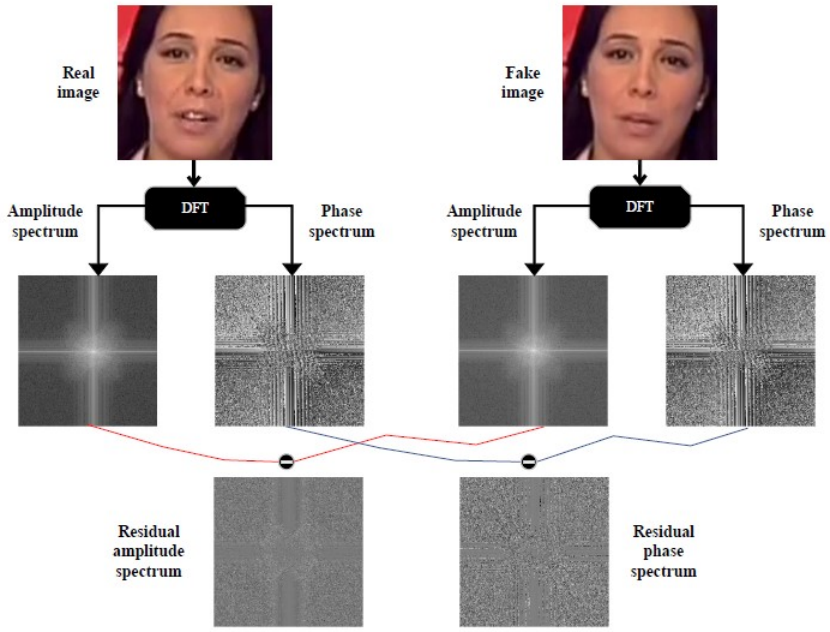

图 18：真实图像与 NeuralTextures 操纵的虚假图像的频谱和残差谱

通过以上观察，进一步探索频域中的局部伪造伪影是必要的。此外，相位谱一般记录图像的结构和位置信息，而振幅谱包含了大部分的纹理和色差信息，振幅信息和相位信息在频域中相互补充，为图像感知起着重要作用[51]。对此，本节提出将图像切片为局部的 patches，然后再进行频域转换获得 patch-wise 相位谱和幅度谱，以在频域中挖掘细微的局部伪造痕迹。具体步骤如图 18 所示，分辨率为 H×W×C 的输入图像 I 首先被转换为灰度图像，即 C=1。因为在后继的实验中使用 RGB 图像显著增加了频域特征的维度，但并没有显著提高性能。然后灰度图像被划分为 $N=HW/P^2$ 个不重叠的 patches，每个 patch 的大小为 P×P×1。最后，每个 patch 即$f_i(x,y)$被独立地进行离散傅里叶变换(DFT)，即

$$F_i(u,v) = \frac{1}{P^2}\sum_{x=0}^{P-1}\sum_{y=0}^{P-1} f_i(x,y)e^{-2\pi j\left(\frac{ux+vy}{P}\right)}, for \quad i = 1,\dots,N. \qquad (4\text{-}2)$$

使用欧拉公式进一步展开可表示为：

$$F_i(u,v) = R_i(u,v) + jI_i(u,v) = |F_i(u,v)|e^{j\varphi_i(u,v)}, \tag{4-3}$$

这里为每个 patch 取得其幅度谱AS_i与相位谱PS_i，其计算如下：

$$AS_i = |F_i(u,v)| = [R_i^2(u,v) + I_i^2(u,v)]^{\frac{1}{2}},$$
$$PS_i = \varphi_i(u,v) = arctan\left[\frac{I_i(u,v)}{R_i(u,v)}\right],$$
$$for \quad i = 1, \dots, N. \tag{4-4}$$

最后获得 patch-wise 幅度谱和相位谱如图 18 所示，可以明显看出，细分为切片的幅度谱和相位谱均比图 19 中全局的幅度谱和相位谱更加多样化。在图 17 中的全局频谱残差只有轻微的像素差异，导致伪造图像难以与真实图像区分开来。而在图 18 中，patch-wise 幅度谱残差图中再次暴露出许多明显的棋盘伪影。进一步比较 patch-wise 幅度谱和相位谱的残差图像，相位谱的像素差异存在于更多的 patch 中，这说明伪造图像和真实图像在相位谱上的不一致性更多。这启发了本节提出的 patch-wise 频谱交互注意模块的设计。

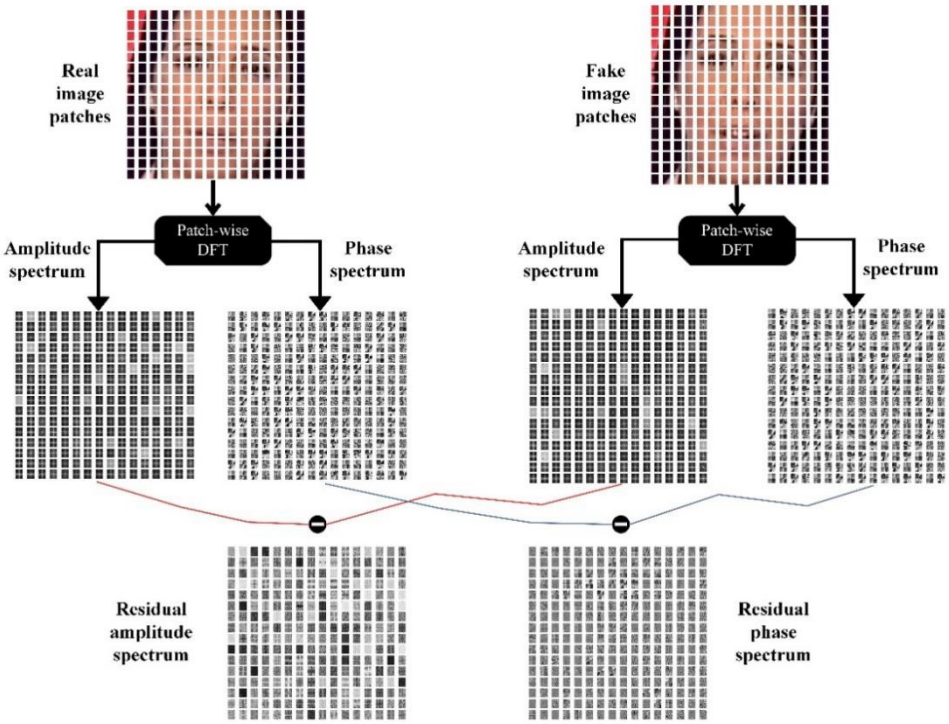

图 19：真实图像与 NeuralTextures 操纵的虚假图像的 patch-wise 频谱和残差谱

4.2.2 频域分支网络

频域分支网络为 deepfake 检测提供局部伪影感知的频域特征。首先，patch-wise 幅度谱$AS_i(i=1,\dots,N)$被平展开，然后通过共享的全连接层被线性投影到 d 维的嵌入$E_{as}\in R^{N\times D}$中，并通过与 position embedding 层的输出相加来保持图像多个 patches 在频域中的位置关系。这里假设 D 与输入维数相同，即 D=P×P。同理，通过另一个线性层将相位谱PS_i投影到相位嵌入$E_{ps}\in R^{N\times D}$中。E_{as}和E_{ps}被输入后继的频域分支，包括 patch-wise 频谱交互注意模块和 MLP-Mixer，最后通过全局平均池化层(GAP)输出幅度特征AF和相位特征 PF。

4.2.3 Patch-wise 频谱交互注意模块设计

Patch-wise 频谱交互注意模块(PFSCA)的设计如图 20 所示，旨在充分利用幅度和相位信息的互补特性。PFSCA 的灵感来自于 transformer[52]的注意力机制，将E_{as}或E_{ps}选择性地作为查询(Q)、键(K)和值(V)。在幅度注意分支中，$K_{as}=V_{as}=LN(E_{as})$，而$Q_{ps}=LN(E_{ps})$，其中LN表示层归一化，并计算矩阵的输出为：

$$A(Q^{ps},K^{as},V^{as})=softmax\left(\frac{Q^{ps}(K^{as})^T}{\sqrt{D_s}}\right)V^{as}, \tag{4-5}$$

其中Q_{ps}和K_{as}用于衡量幅度伪影和相位差异之间的相关性，所以$Q_{ps}(K_{as})^T$作为交互并并通过计算 softmax 作为权重增强幅度嵌入特征E_{as}。注意力的计算使用了多头注意(multi-head attention)[52]：通过 h 个头的线性投影，在 d 维 Q、K、V 上并行计算注意函数即式(4-5)，然后将所有输出连接并再次投影回来。注意，d 设置为 d =h，以保持维数和计算的一致性。该计算过程可以表示为如下：

$$\begin{aligned} MHA(Q^{ps},K^{as},V^{as})&=[A_1,A_2,\dots,A_h]W^O,\\ A_i&=A\left(Q^{ps}W_i^Q,K^{as}W_i^K,V^{as}W_i^V\right). \end{aligned} \tag{4-6}$$

其中$W_i^Q\in R^{D\times d_q},W_i^K\in R^{D\times d_k},W_i^V\in R^{D\times d_v};W^O\in R^{h*d_v\times D}$ 为投影矩阵。该模块使用了 4 个注意力头部，所以$d_q=d_k=d_v=\frac{D_s}{h}$。最后我们通过残差连接获得幅度 token：

$$AT=E_{as}+MHA(Q^{ps},K^{as},V^{as}). \tag{4-7}$$

如图 20 所示，通过相似的计算得到相位 token:

$$PT=E_{ps}+MHA(Q^{as},K^{ps},V^{ps}). \tag{4-8}$$

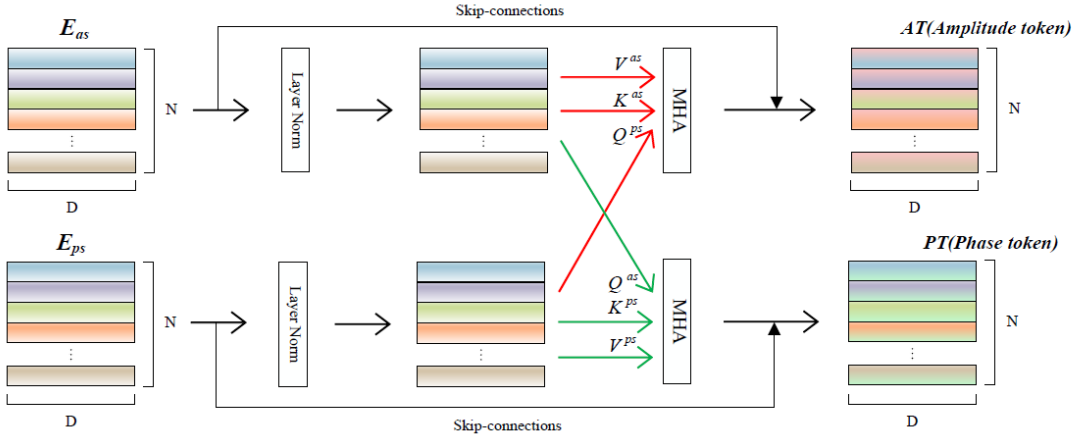

图 20：Patch-wise 频谱交互注意模块(PFSCA)的设计

4.2.4 局部频域伪造特征提取

由于频域不具备自然图像的平移不变性和局部一致性，频谱表示AT和PT与普通的卷积神经网络并不兼容。与其他方法将提取的频域表示再次转换回 RGB 空间输入卷积神经网络不同，本文设计的频域分支的目标是直接利用频率信息，即通过一个双流 MLP-Mixer[53]网络分别以AT和PT作为输入，捕获不同 patch 之间的局部频域伪造特征。

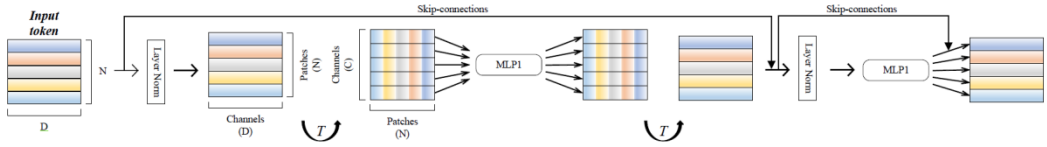

图 21：用于提取频谱伪造特征的 MLP-Mixer[53]网络结构

具体地，幅度 token $AT \in R^{N\times D}$或相位 token $AT \in R^{N\times D}$被输入到一系列的 Mixer 层中。如图 21 所示，MLP-Mixer 的每一层都是相同的，由一个 token-mixing MLP 块和一个 channel-mixing MLP 块组成。Token-mixing MLP 块的计算如下:

$$U = T + W_2[\sigma(W_1(LN(T)^T))]^T, \tag{4-9}$$

其中W_1和W_2为全连接层的线性操作，LN和σ分别表示层归一化[51]和 GELU[54]非线性激活。特征映射 $R^N \rightarrow R^N$ 在所有 patches 之间共享，通过独立地作用于每个通道来混

合不同 patch 的信息。然后，将得到的$U \in R^{N\times D}$输入到 channel-mixing MLP 块：

$$Y = U + W_4\left[\sigma\left(W_3\left(LN(U)\right)\right)\right], \tag{4-10}$$

类似地，$Y \in R^{N\times D}$聚合每个 patch 的不同通道信息，并输入到下一 Mixer 层。

通过 MLP-Mixer 提取局部频域伪造特征的原理如下，channel-mixing 块对每个 patch 中的所有信道即 d 维向量进行相同的线性变换，即参数共享，类似于 1×1 卷积。channel-mixing 块从局部的 patches 中提取有区分性的特征，这种局部表示有利于捕获不同 patch 中伪造的频率模式如棋盘伪影。相反，token-mixing 块在每个通道上对不同的 patch 即 n 维向量进行线性变换，这类似于在通道上使用共享权值进行深度卷积。Mixer 层适合 patch-wise 幅度和相位频谱之间的相关性，为长距离交互提供了一个很好的方式，所以该分支网络补充了卷积神经网路的局部结构感知性。在实验中只使用了两个 Mixer 层，作为检测性能和网络规模之间的折衷。将 AT 和 PT 输入到两个 MLP-Mixer 网络中，最后通过全局平均池化层输出幅度特征 AF 和相位特征 PF，即：

$$\begin{aligned} AF &= GAP\big(MLP - Mixer(AT)\big), \\ PF &= GAP\big(MLP - Mixer(PT)\big). \end{aligned} \tag{4-11}$$

4.3 模型架构与实验设置

基于 4.1 和 4.2 节的理论分析与算法设计，本章提出了一种结合空域与频域的局部伪造缺陷的 deepfake 检测模型，命名为 LA-Net(Local artifact-aware deepfake detection network)。

4.3.1 模型总体结构与细节

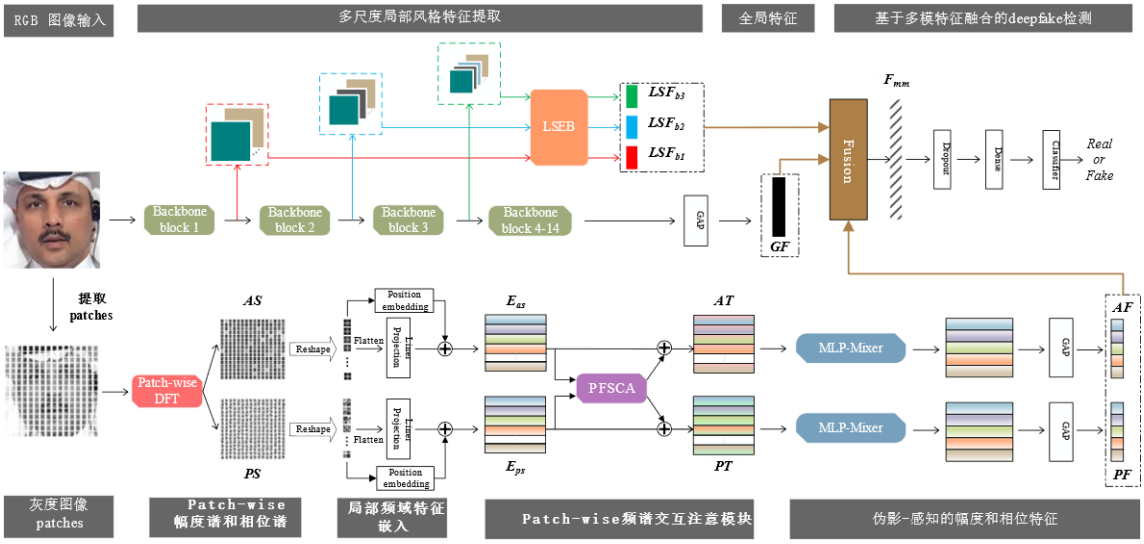

图 22：LA-Net 模型总体架构

LA-Net 是一个端到端的网络，由空域和频域两个分支组成，其总体架构如图 22 所示。模型的输入与输出分别是 RGB 图像和对该图像的检测结果。空域分支从 LSEB 和骨干网络提取浅层特征LSF_{b1}、LSF_{b2}、LSF_{b3}和全局特征GF,。频域分支则提取局部的幅度特征AF和相位特征AF。这些多模态且多尺度的特征被融合为F_{mm}，用于更全面有效的 deepfake 检测：

$$F_{mm} = concatenate(LSF_{b1}, LSF_{b2}, LSF_{b3}, GF, AF, PF), \quad (4\text{-}12)$$

最后F_{mm}通过全连接层输出预测$\hat{y}_i$，并通过交叉熵损失(cross entropy loss)监督二分类网络的训练：

$$L_{ce} = -\frac{1}{N}\sum_{i=1}^{N} y_i. \quad (4\text{-}13)$$

空域分支的骨干网络使用了在 imagenet 上预训练的 Xecption，并且在训练过程中冻结任何网络层以充分学习适用于 deepfake 检测的表征。网络中所有 dropout 层的 drop rate 都设置为 0.5，更多的模型细节和特征规格等如表 2 所示。

4.3.2 实验数据与设置

为了综合评估提出的 LA-Net 在 deepfake 检测上的有效性以及在不同压缩场景下的

鲁棒性，实验使用了大规模的人脸伪造基准数据集 FF++[33]。由于 FF++数据集的基准模型 Xception 已经在 RAW 版本中实现了近乎完美的检测性能，所以本章实验采用了 HQ 和更具挑战性的 LQ 版本。按照基准的预处理方式，我们以 720 个视频进行训练，140 个视频进行验证，剩下的 140 个视频进行测试。每个视频采样 30 帧后，我们使用 Dlib 从帧中提取并对齐人脸，最终获得大小调整为 256×256 的 RGB 图像。遵从于 deepfake 检测任务中常用的评价指标[33][12]，实验报告了帧级别的 ACC (accuracy)和 AUC (area under curve of ROC)评分以合理评估模型性能，以便于与其它先进的检测方法进行对比分析。在训练过程中，batch size 被设置为 16，初始学习率为 0.001 且伴随 0.0001 权值衰减的 Adam 算法[54]被用于网络优化。在实验中，LA-Net 默认被训练 50 个 epochs。如果验证损失在 5 个 epochs 后仍未减少，则学习率将降低到原来的一半。在验证集上获得最高的 ACC 分数的权重被保存为最终模型。

表 2：本章提出的 LA-Net 的模型细节和特征规格

缩写表示	说明	尺寸或大小
I	输入 LA-Net 的 RGB 待检测图像	256×256×3
LSF_{b1}	从 Xception block1 提取的局部风格特征	64
LSF_{b2}	从 Xception block2 提取的局部风格特征	128
LSF_{b3}	从 Xception block3 提取的局部风格特征	256
GF	从 Xception 最后一层（block14）输出的全局特征	1024
P; AS_i; PS_i	切分的 patch 大小	16×16×1
N	切分的 patch 数量	256(HW/P^2)
D	频域嵌入特征	256
E_{as}; E_{ps}; AT; PT	幅度嵌入; 相位嵌入; 幅度 token; 相位 token	256×256 (N×D)
W_1	MLP-Mixer 中 token-mixing block 的第一个全连接层	256(N)
W_2	MLP-Mixer 中 token-mixing block 的第二个全连接层	256(N)
W_3	MLP-Mixer 中 channel-mixing block 的第一个全连接层	256(D)
W_4	MLP-Mixer 中 channel-mixing block 的第二个全连接层	256(D)
AF; PF	幅度特征; 相位特征	256
F_{mm}	融合后的特征	1984

本文的实验设备包括：一张 NVIDIA GeForce RTX 2080Ti 显卡，处理器为 Intel Core i7-9700K，内存为 128GB。软件环境包括 3.7 版本的 Python，深度学习框架采用了 2.2.0 版本的 tensorflow。

4.4 实验结果和对比分析

4.4.1 检测不同伪造方法的有效性

本小节评估了提出的 LA-Net 针对 FF++四种操纵方法的检测性能，在该场景下，模型都在高度压缩的 LQ 版本上进行训练和测试。测试结果如表 3 所示，相比其它方法和 baseline 模型 Xception，LA-Net 在 ACC 和 AUC 指标上均取得了最优，这证明了提出的方法检测不同伪造人脸类型的有效性。先进的检测方法 SPSL 利用上采样步骤产生的全局相位差异，但是一些局部的操纵如 F2F 和 NT，在全局相位谱上不会有明显的差异。为了解决这个问题，LA-Net 中提出的 PFSCA 在局部 patches 中捕获了更显著的振幅伪影和相位差异，这解释了提出的 LA-Net 优于 SPSL 的指标分数。

表 3：FF++数据集检测不同操纵方法的测试 ACC(%)和 AUC(%)

检测模型↓	操纵方法(LQ)→	DF		F2F		FS		NT	
	测试指标→	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
Steg. Features + SVM[55]		73.64	-	73.72	-	68.93	-	63.33	-
LD-CNN[56]		85.45	-	67.88	-	73.79	-	78.00	-
C-Conv[57]		85.45	-	64.23	-	56.31	-	60.07	-
CP-CNN[58]		84.55	-	73.72	-	82.52	-	70.67	-
MesoNet[6]		87.27	-	56.20	-	61.17	-	40.67	-
Xception[34]		95.13	99.24	87.34	93.62	92.42	97.08	77.54	84.51
SPSL[19]		93.48	98.50	86.02	94.62	92.26	98.10	76.78	80.49
LA-Net（ours）		**97.56**	**99.44**	**89.98**	**95.27**	**94.44**	**98.50**	**81.45**	**86.43**

值得说明是，NeuralTextures(NT)操纵方法是最难被检测的，它只修改了面部表情对应的嘴部区域的像素，导致在 RGB 空间和频谱中产生的伪造伪影极其细微。与 Xcpetion 相比，LA-Net 使 ACC 显著提高约 4%，性能改进主要归功于设计的 LSEB 和 PFSCA 在

空间和频域捕获伪造缺陷，并集中于 deepfake 局部暴露的伪影。

4.4.2 检测不同压缩场景的鲁棒性

本小节在 FF++的不同压缩场景评估本章提出的方法，包括 c23 轻度视频压缩的高质量版本和 c40 重度视频压缩的低质量版本，并在表 4 中报告了与之前的检测方法的对比。测试结果表明: (1)在准确率(ACC)上明显优于 Steg、LD-CNN、C-Conv、CPCNN、MesoNet 等早期方法。(2)在不同的质量环境下，LA-Net 的结果优于强力的检测方法 Xception 和 Face X-ray。在 HQ 和 LQ 版本中，LA-Net 的 AUC 显著超过 Face X-ray 算法 12.25%和 28.68%。Face X-ray 依赖于混合边界周围的差异，而严重的压缩设置会导致混合边界的伪影现象减弱，这限制了它的检测性能。而 LA-Net 受益于设计的 PFSCA，如 4.2 节和图 18 所示：尽管压缩视频丢失了许多底纹和频率信息，但在 patch 级别的幅度和相位谱中仍然暴露出明显的伪影。

表 4：FF++数据集检测不同压缩设置的测试 ACC(%)和 AUC(%)

方法↓	数据质量→	HQ(c23)		LQ(c40)	
	测试指标→	ACC	AUC	ACC	AUC
Steg. Features + SVM[55]		70.97	-	55.98	-
LD-CNN[56]		78.45	-	58.69	-
C-Conv[57]		82.97	-	66.84	-
CP-CNN[58]		79.08	-	61.18	-
MesoNet[6]		83.10	-	70.47	-
Xception[34]		95.73	96.30	86.86	89.30
Xception-ELA[59]		93.86	94.80	79.63	82.90
DSP-FWA[11]		-	57.50	-	62.30
Face X-ray[14]		-	87.40	-	61.60
Two-branch[44]		96.43	98.70	86.34	86.59
SPSL[19]		91.5	95.32	81.57	82.82
F3-Net[22]		97.52	98.10	**90.43**	**93.30**
LA-Net(ours)		**97.89**	**99.65**	87 .19	89.44

LA-Net 在 FF++的 HQ 压缩设置中取得了领先地位，并在 LQ 版本上取得了具有竞争力的结果，表明了提出的方法在不同压缩场景进行检测的鲁棒性。Two-branch[44]和

SPSL[19]也融合了空域和频域的特征以检测伪造人脸图像。提出的方法在几乎所有指标上都优于它们，这表明所提出的 LSEB 提取浅层风格表示的有效性。LA-Net 在 LQ 环境下得到的结果要差于 F3-Net[22]，因为 F3-Net 是专门针对高压缩 deepfake 视频检测设计的，而 LA-Net 中设计的 LSEB 提取风格表征，如浅层纹理信息，对高压缩率很敏感，但它可以提高局部伪造如 F2F 和 NT 的检测性能。

4.4.3 检测模型的复杂度对比

考虑到 LA-Net 同时在空域和频域引入了多个模块，本节比较了其与现有模型的复杂度，以衡量资源消耗和计算量。表 5 报告了 LA-Net 和现有 deepfake 检测方法在参数(parameters，Param)，每秒浮点操作(floating-point operations per second，Flops)和总读写内存(total MemR+W)上的对比。比较模型包括基线网络 VGG19，EfficientNet-B4，ResNet50 和 Xception，这些模型被广泛作为各种 deepfake 检测方法的骨干网络。本节还用官方发布的模型重现了最先进的 DSP-FWA[11]方法。所有模型在相同的实验环境下进行评估，图像的输入尺寸为 256×256×3。

表 5：LA-Net 与其它方法的效率比较

模型	Param	Flops	MemR+W
VGG19	20.6M	25.6 G	860MB
EfficientNet-B4	19.5M	4.0G	222MB
ResNet50	25.7M	10.9G	376MB
Xception	22.9M	11.9G	405MB
DSP-FWA (ResNet101-based)	42.6M	10.8G	487MB
LA-Net (Xception-based)	30.5M	13.4G	446MB

如表 5 所示，由于没有针对 deepfake 检测任务引入模块或重新设计框架，基线网络的参数，Flops 和内存成本都更少，但这另一方面也限制了这些朴素 CNN 的检测性能。与基线网络 Xception 相比，LA-Net 增加了约 7.6M 的网络参数，但仍远低于 DSP-FWA[11]。额外的参数主要来自本章提出的 LSEB，PFSCA 和 MLP-Mixer 模块。但相比 Xception 骨干网络，LA-Net 并没有显著增加 Flops 和 Total MemR+W 的大小，并且在时间和空间复杂度上远小于 VGG19，这表明本文方法在计算和内存开销方面是可以接受的。

4.5 本章小结

本章提出了基于空域与频域中局部伪造缺陷的 deepfake 检测算法。在 deepfake 图像空域中的观察与分析结果表明，局部的伪造特征相比于全局特征更具判别性。对此，本章展示了设计的局部风格提取模块和空域分支网络，提升对小规模篡改 deepfake 的有效性。然后，本章在 deepfake 图像的频域中发现了一个关键的现象，即微妙的伪造伪影可以进一步暴露在切分为 patch 图像的相位谱和振幅谱中，并表现出不同的线索。对此，本章进一步设计了 patch-wise 频谱交互注意模块和频域分支网络，用于捕获频域内的局部相关不一致性。最终提出的算法模型以双分支结构的网络学习 deepfake 判别性的伪造特征。实验评估与对比分析表明了该工作对不同操纵方法进行检测的有效性提升，以及在不同压缩场景下的鲁棒性提升。

第五章 基于语义分割与对比分类的 deepfake 检测算法

第四章的研究挖掘了 deepfake 相对真实图像存在的伪造痕迹，并结合深度学习模型提取局部伪影特征，实现了有效鲁棒的检测算法。而本章的观察与研究动机来自于伪造算法通常对人脸语义进行操纵，关注人脸语义内容存在的伪造缺陷以设计更泛化的 deepfake 检测算法。本章研究了人脸语义与背景的分割用于分类学习，同时对语义内容特征设计了额外的对比学习。在大量数据集上的实验证明了本章方法对 deepfake 检测的有效性和改进的泛化能力。

5.1 人脸语义分割

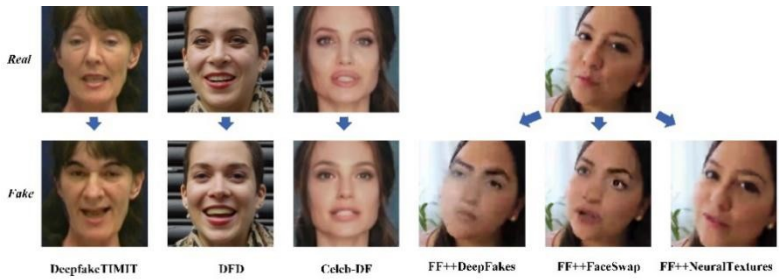

图 23：不同 deepfake 伪造方法对人脸图像的操纵效果

从 deepfake 生成者的角度来看，伪造算法的效果通常是实现对人脸所蕴含的内容和信息进行操纵以达到伪造目的。换脸算法如 DeepFakes 和 FaceSwap 将原始人脸替换成另一个人脸的同时保持背景不变，表情互换如 Face2Face 在不改变人脸身份的情况下控制人脸五官以实现表情效果，人脸属性操纵如 NeuralTextures 篡改嘴部区域。如图 22 所示，伪造的人脸虽然在不同的数据集和生成算法中迥然不同，但被操纵的主体内容是人脸语义或者特定的五官区域。另一方面，虽然伪造人脸图像中的背景和大部分面部皮肤区域差别不大，但一些不完美的伪造图像在这些区域仍可能存在缺陷，如图 23 中的

DeepFakes 也存在强烈的换脸边界伪影。基于这个观察结果，本章提出的 deepfake 检测模型旨在对人脸语义内容进行对比学习，以提取更本质的伪造特征，同时仍然保留非面部语义特征作为另一个分支补充。

人脸语义分割的实现同样基于 Dlib 开源库，如图 24 所示，其具体实现步骤为：

（1）对于输入的一张人脸图像，首先运行人脸检测器为人脸提取 68 个关键点；

（2）根据关键点坐标将人脸划分出 6 个 bounding box 作为语义区域：嘴部（49-68），右眉（18-22），左眉（23-27），右眼（37-42），左眼（43-48）和鼻子（28-36）；

（3）保留人脸图像在语义区域的像素值，而在非语义区域则用黑色像素即 0 值进行填充，获得人脸语义图像；

（4）保留人脸图像在非语义区域的像素值，而在语义区域则用黑色像素即 0 值进行填充，获得人脸非语义图像。

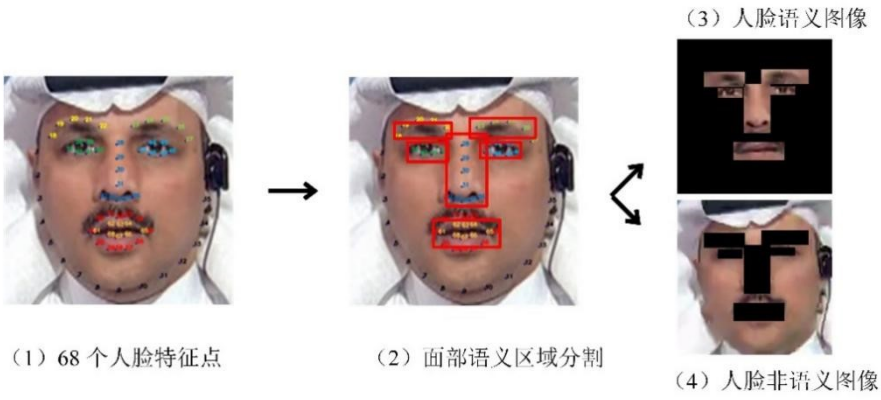

图 24：人脸语义分割的实现

5.2 面部语义监督对比学习

在深度学习应用于包括 deepfake 检测任务等分类场景时，交叉熵（cross entropy）是最广泛使用的损失函数(loss function)。由于该损失函数只关注样本是否正确分类，但没有显式地考虑类内紧凑度和类间分散度，所以其监督的分类模型可能存在较差的决策边界，这导致有限的鲁棒性和泛化性。最近，对比学习的研究工作取得了重大进展，它

们的共同思想是选定一个锚(anchor)样本，在嵌入空间中推近锚和正样本的距离，同时推远锚和负样本的距离。

在如图 25 左侧所示的自监督对比学习场景中，由于避免使用标签信息，正样本通常来自于锚样本的数据增强形式，负样本从 mini-batch 中随机选取。而图 24 右侧所示的监督对比学习则直接使用类别标签来衡量相似度，来自同一个类的样本为正样本，来自不同类别的样本为负样本。考虑到伪造人脸与真实人脸的高度相似性以及 deepfake 基准数据集中的标签信息可以被有效地利用于控制距离，所以本章介绍了监督对比损失[60]用于学习面部语义特征。

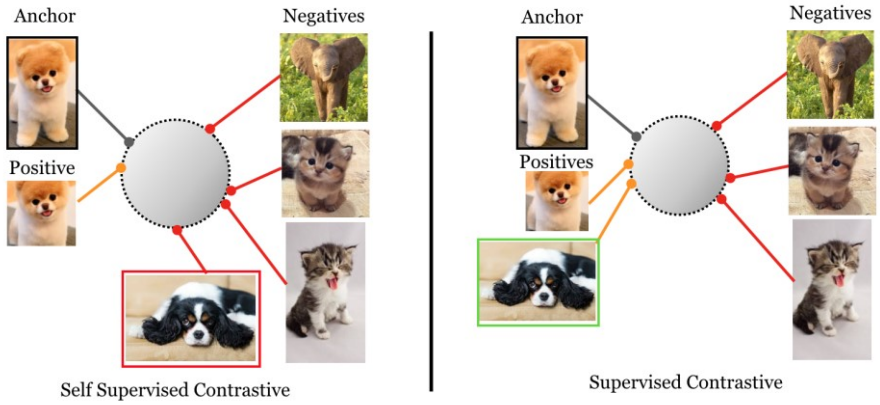

图 25：自监督对比学习与监督对比学习

基于 4.1 节中 deepfake 伪造人脸的被操纵区域主要是语义内容的现象，本章提出方法的目标是在特征空间中将同一类别(真实或伪造)的人脸语义图像拉进距离，并推远真实人脸语义与伪造人脸语义的距离。针对骨干网络从人脸语义图像提取的特征，通过全连接层(dense layer)投影到嵌入空间$z \in R^p$。投影后的输出被归一化，并使用内积来衡量嵌入空间中的距离。如果 batch size 为 N，用于训练模型的监督对比损失为：

$$L_{SC} = \sum_{i=1}^{N} \frac{-1}{|\{z_i^0\}|} \sum_{z_p \in \{z_i^0\}} log \frac{\exp(z_i \cdot z_p / \tau)}{\sum_{z_n \in \{z_i^1\}} \exp(z_i \cdot z_n / \tau)}, \quad (5\text{-}1)$$

其中 0 和 1 用于表示图像的类别即真实或伪造，· 表示点乘（内积），温度标量$\tau \in R^+$是一个超参数。$|\{z_i^0\}|$表示正例的数量。i 指示用于对比的锚，p 是与锚类别相同的正样本，n 表示不同类别的反样本。在这样一个有限的嵌入空间中，监督对比缺失明确地约

束类内紧凑度并提高类间分散度，鼓励网络从人脸语义中学习具备判别性且更泛化的特征表示。

5.3 模型架构与实现

本章提出的基于面部语义分割与对比分类的 deepfake 检测模型命名为 FSCM（Facial Semantic Contrastive-Classification Model），其网络架构如图 26 所示。

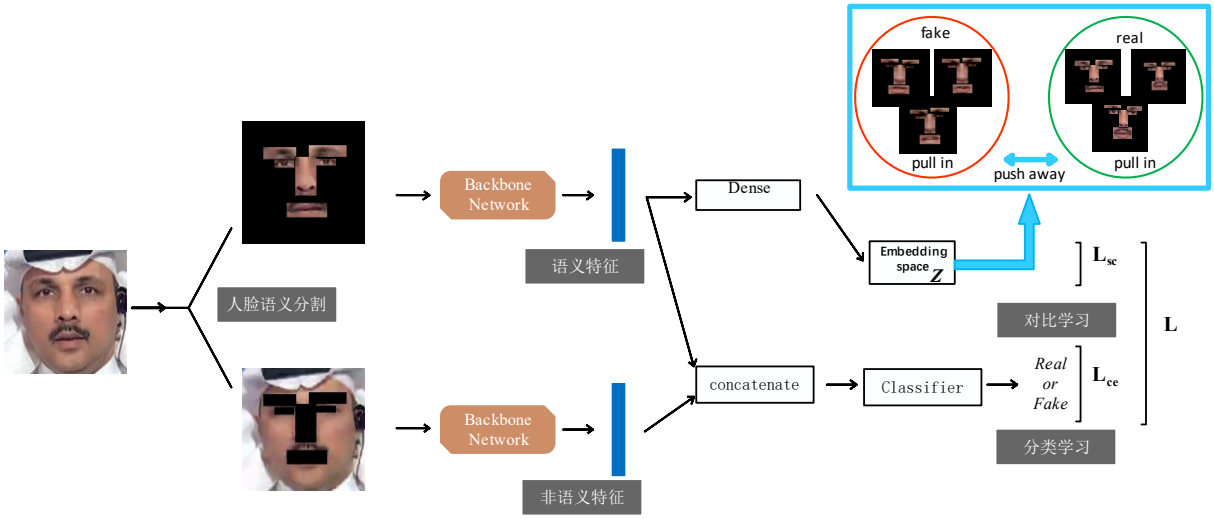

图 26：FSCM(Facial Semantic Contrastive-Classification Model)网络架构

此端到端的模型接受 RGB 人脸图像作为输入，首先根据小节 5.1 提出的面部语义分割方法将人脸分割为语义图像和非语义图像。然后语义图像和非语义图像分别被输入各自的骨干网络(Backbone Network)，并提取出语义特征和非语义特征。FSCM 的骨干网络采用 Xception[34]进行特征提取，其输出的语义特征和非语义特征均是 block 14 输出的 2048 维向量。语义特征被一个额外的全连接层(dense layer)投影到一个嵌入空间 z 中，通过计算监督对比损失即式(5-1)鼓励面部语义分支网络学习到更具备鲁棒性和泛化性的表征。嵌入空间 z 的大小为 128，即通过 128 个节点的全连接层进行投影。同时，语义特征与非语义特征被串联融合，并通过全连接层以交叉熵损失学习 deepfake 分类器：

$$L_{ce} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i). \quad (5\text{-}2)$$

为了兼顾检测能力与泛化能力，模型通过结合对比损失L_{sc}和分类损失L_{ce}同时学习嵌入特征表示和分类器，所以最终的损失函数为：

$$L = \alpha L_{SC} + (1 - \alpha)L_{ce}, \tag{5-3}$$

其中 α 是权重超参数。

5.4 实验设置与结果分析

5.4.1 数据集与实验设置

为了评估所提出的方法对 deepfake 检测的有效性和鲁棒性，6 个被广泛使用的第一代和第二代 deepfake 基准数据集被用于实验。第一代包括 DeepfakeTIMIT、UADFV 和 FF++，第二代数据集为 DFD、DFDC 和 Celeb-DF，细节信息已在章节 3.1 介绍。每个数据集分成 80% 用于训练，其余用于测试，且训练集 20%的数据用作验证集。对于 Celeb-DF 的测试集划分则参照了官方指定的标签[12]。数据预处理方式如章节 3.2 所述。

实验报告了帧级别的 AUC（Area Under Curve of ROC）评分以合理评估模型性能，同时便于与其它先进的检测方法进行对比分析。在训练过程中，式（5-1）中的温度超参数τ设置为 0.1，损失函数（5-3）中的权重超参数 α 被设置为 0.5。batch size 被设置为 12，包含 6 张真实和 6 张虚假图像，以指导对比学习。初始学习率为 0.001 且伴随 0.0001 权值衰减的 Adam 算法被用于网络优化。在实验中，FSCM 默认被训练 50 个 epochs。如果验证损失在 5 个 epochs 后仍未减少，则学习率将降低到原来的一半。在验证集上获得最高的 ACC 分数的权重被保存为最终模型。

5.4.2 对比分析与有效性评估

本小节对所提出的基于面部语义分割与对比分类的 deepfake 检测模型 FSCM 进行了有效性评估并与其它方法进行了对比，数据集内和跨数据集测试的结果如表 6 所示。实验中的对比方法包括：Capsule 模型[8]基于胶囊结构结合 VGG19 网络作为骨干架构进行 deepfake 分类；DeepfakeUCL[61]是基于无监督对比学习的深度伪造检测方法；Xception[34]作为经典而强力的图像分类深度网络，是 deepfake 检测任务的主流骨干模型；Xception+Tri[62]进一步在伪造特征提取阶段介绍了三重损失函数(triplet loss function)的

使用。

观察在 FaceForensics++数据集上训练的模型：当在 FaceForensics++上测试时，提出的 FSCM 的 AUC 分数高于 Capsule 和 DeepfakeUCL，仅比 Xception 和 Xception+Tri 分别差 1.2%和 1.4%；然而在跨数据集测试 UADFV 时，FSCM 相比 Xception 和 Xception+Tri 分别大幅提升 13.8%和 19.8%的 AUC；在跨数据集测试 Celeb-DF 时提出的 FSCM 同样获得了最优的检测指标。当采用 Celeb-DF 作为训练集时，出现了类似的实验结果，即提出的 FSCM 与现有方法在数据集内的测试场景性能相差不大，但在跨数据集测试场景的性能获得显著提升。

本章实验揭示了跨数据集测试场景比数据集内测试场景更具挑战性。从现实世界场景和应用部署的角度出发，待检测的内容往往不知道其具体的 deepfake 操纵方法，所以训练后的模型在跨数据集上的泛化性能更值得被重视。

表 6：FF++数据集与 Celeb-DF 数据集的测试结果对比，指标为 AUC(%)

方法	训练数据集	测试数据集		
		FF++	UADFV	Celeb-DF
Capsule	FF++	96.6	61.3	57.5
DeepfakeUCL	FF++	93.0	67.5	56.8
Xception	FF++	99.7	80.4	48.2
Xception+Tri.	FF++	**99.9**	74.3	61.7
FSCM(ours)	FF++	98.5	**94.2**	**68.3**
DeepfakeUCL	Celeb-DF	58.9	85.6	90.5
Xception	Celeb-DF	64.1	93.3	99.5
Xception+Tri.	Celeb-DF	60.2	88.9	**99.9**
FSCM(ours)	Celeb-DF	**67.8**	**96.2**	99.8

5.4.3 跨数据集检测的泛化性评估

上一小节的实验结果表明尽管直接使用深度卷积神经网络如 Xception 作为 deepfake 检测模型也能取得出色的域内性能，但是如果外推到训练未见的数据集，即跨数据集测试模型的场景，朴素的深度卷积神经网络的检测性能大幅下降。为了进一步验证提出的方法对泛化能力带来的改进，本小节进行了跨数据集评估实验。在此设置中，baseline 模

型和提出的 FSCM 均在 DeepfakTIMIT(HQ)数据集上进行训练，同时在 UADFV、FF++_DF(LQ)、DFD、DFDC 和 Celeb-DF 数据集进行测试来评估泛化性。这是一个具有挑战性的任务，因为 DeepfakeTIMIT 是第一代 deepfake 数据集，伪造的质量较差导致易于检测且包含的数据（帧数）较少，而其他数据集的数量更大且质量更好。VGG19[63]、ResNet50[48]、Xception[34]和 EfficientNet-B0[64]等深度神经网络经常用作 deepfake 检测器，因此它们被选作 baseline。Baseline 模型遵循常规的部署方式，即接受整张原始人脸图像作为输入并以交叉熵损失训练二分类模型。

对比的 baseline 和 FSCM 的泛化结果如表 6 所示。可以看出，几乎所有的模型在域内测试均表现出接近完美的检测性能。事实上，这些深度卷积神经网络确实有效地提取了数据集内的判别特征，但是也仅限于域内数据。在跨数据集的域外测试场景下，baseline 的 AUC 分数均在 50%波动即盲目猜测的水平，表明它们仅学习到了特定于 DeepfakeTIMIT 数据集的判别模式，严重缺乏泛化能力。另一方面，所提出的 FSCM 在跨不同数据集测试的 AUC 分数均为最高，表明提出的方法有效地挖掘了 deepfake 共同的伪造缺陷和强大的泛化性能。

表 6：DeepFakeTIMIT(DT)训练的模型跨数据集测试的 AUC(%)分数

模型	训练数据	测试数据					
		DT	UADFV	FF++_DF(LQ)	DFD	DFDC	Celeb-DF
VGG19	DT	98.9	49.8	50.2	49.9	50.8	51.6
ResNet50		**99.9**	48.5	53.2	53.2	53.4	51.3
Xception		**99.9**	50.3	50.0	56.4	51.6	50.2
EfficientNet-B0		**99.9**	50.0	50.0	54.1	54.1	53.2
FSCM(ours)		**99.9**	**64.3**	**78.5**	**81.9**	**65.1**	**54.3**

值得说明的是，FSCM 的骨干网络采用了 Xception，在 UADFV、FF++_DF(LQ)、DFD、DFDC 和 Celeb-DF 数据集分别提升了 14.3%，28.5%，25.5%，13.5%和 4.1%的 AUC 分数。如此大幅度的泛化性能提升主要受益于提出的方法所做的改进，同时也是与 baseline 模型 Xception 的不同之处：(1)将人脸语义内容和非语义内容进行分割，在语义

内容中学习 deepfake 更共同的伪造缺陷；(2)针对语义特征进行额外地对比学习，强迫编码空间中同类特征更紧凑，异类特征间隔更大。本节跨数据集评估实验证明了本章方法对 deepfake 检测的有效性和改进的泛化能力。

5.5 本章小结

本章工作关注到了各种 deepfake 伪造算法主要是对人脸所蕴含的内容和信息进行操纵。旨在提升模型检测不同 deepfake 时的可迁移性，以人脸语义内容存在的伪造缺陷为突破口，本章提出了一种基于语义分割与对比分类的检测算法。该工作的核心思想是解耦合对人脸图像中语义内容特征和非语义内容特征的分类学习，并介绍了监督对比学习的结合方法，鼓励检测模型从人脸语义中提取更具备判别性和泛化性的特征表示。本章进行了大量的跨数据集测试实验，通过对比评估证明了提出的方法能显著改善当前基于深度卷积神经网络的 deepfake 检测模型的泛化能力。

第六章 总结与展望

6.1 工作总结

深度学习方法驱动的图像与视频伪造对信息安全带来了巨大威胁，deepfake 检测技术逐渐成为研究热点与应用需求。通过对当前主流方法的分析与研究，针对现有研究工作的局限性，本文创新性地提出了两个结合伪造缺陷与语义对比的 deepfake 检测算法。

针对目前检测算法提取的伪造缺陷特征不足以有效地鉴别伪影较少的局部操纵 deepfake，以及在视频高压缩场景下模型的鲁棒性不强等问题，本文提出了一种基于空域与频域中局部伪造痕迹的 deepfake 检测算法。基于实验观察与分析，该模型精心研究设计了两个算法模块：局部风格提取模块识别更具鉴别性的局部纹理特征，频谱交互注意模块用于捕获幅度和相位信息局部不一致性。通过对 deepfake 在局部上暴露的伪影交互地进行特征编码，提出的双分支网络能同时在空域和频域上挖掘到细微的伪造缺陷。大量的评估实验和对比结果表明了该算法对伪影较少且检测难度较大的操纵方法的有效性，以及相较常规的深度分类模型在不同压缩环境下的鲁棒性。

针对基于深度神经网络的 deepfake 检测模型受到过拟合影响且检测未见伪造数据的泛化能力极差的问题，本文介绍了一种基于语义分割与对比分类的 deepfake 检测算法，以监督深度学习分类模型识别到泛化性更强的伪造模式。该方法受启发于各种 deepfake 伪造算法操纵的主体内容是人脸语义或者特定的五官区域，并提出将人脸图像进行语义分割。通过解耦合语义内容特征和非语义内容特征，鼓励模型学习到 deepfake 的本质缺陷，此外，通过结合监督对比学习和分类学习，在特征空间中将同一类别（真实或伪造）的人脸语义图像拉进距离，并推远真实人脸语义与伪造人脸语义的距离，进一步提升了检测网络的泛化能力。数据集内和跨 6 个数据集的实验评估结果证明了该模型对 deepfake 检测的有效性和以及大幅改进的泛化性。提出的方法为应用场景下的通用 deepfake 检测提供了新的思路。

6.2 未来展望

本文从伪造缺陷和语义对比出发，对 deepfake 检测方法在有效性、鲁棒性和泛化性上的局限性进行了研究与改进。然而 deepfake 检测方法或模型仍面临如下挑战，同时这也是未来研究工作待解决的问题。

（1）伪造技术与检测技术仍在博弈的过程中不断发展，需要在最先进和反复优化的 deepfake 伪造内容上保持有效的检测能力。

（2）在现实场景中，deepfake 很容易受到各种恶意攻击或无意扰动的影响，如图像或视频压缩、添加噪声、模糊处理等。现有的研究提出了各种方法应对压缩和模糊等常见的后处理操作，且大多数方法使用深度神经网络作为检测模型。深度神经网络很容易被难以察觉的对抗噪声攻击，然而当前的 deepfake 检测研究极少对此评估。

（3）在法医和取证场景下，需要提供具备可解释性的检测结果。然而现有的基于深度学习的检测模型难以给出具备证据能力的检测解释。

（4）现有的方法在检测常见的 deepfake 数据集时普遍报告了可观的检测指标。然而研究论文或报告的结果不能完全代表在实际场景中的 deepfake 检测性能。因此，deepfake 对社区和学术界来说仍然是一个强大的威胁，需要开发更具备实用性的检测方法。

参考文献

[1] Kingma D P, Welling M. Auto-encoding variational bayes[J]. arXiv preprint arXiv:1312.6114, 2013.

[2] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.

[3] Wang T C, Mallya A, Liu M Y. One-Shot Free-View Neural Talking-Head Synthesis for Video Conferencing[J]. arXiv preprint arXiv:2011.15126, 2020.

[4] Kwok A O J, Koh S G M. Deepfake: a social construction of technology perspective[J]. Current Issues in Tourism, 2021, 24(13): 1798-1802.

[5] Zhou P, Han X, Morariu V I, et al. Two-stream neural networks for tampered face detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, 2017: 1831-1839.

[6] Afchar D, Nozick V, Yamagishi J, et al. Mesonet: a compact facial video forgery detection network[C]//2018 IEEE International Workshop on Information Forensics and Security (WIFS). IEEE, 2018: 1-7.

[7] Nguyen H H, Fang F, Yamagishi J, et al. Multi-task learning for detecting and segmenting manipulated facial images and videos[J]. arXiv preprint arXiv:1906.06876, 2019.

[8] Nguyen H H, Yamagishi J, Echizen I. Capsule-forensics: Using capsule networks to detect forged images and videos[C]//ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019: 2307-2311.

[9] Li Y, Chang M C, Lyu S. In ictu oculi: Exposing ai created fake videos by detecting eye blinking[C]//2018 IEEE International Workshop on Information Forensics and Security (WIFS). IEEE, 2018: 1-7.

[10] Yang X, Li Y, Lyu S. Exposing deep fakes using inconsistent head poses[C]//ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019: 8261-8265.

[11] Li Y, Lyu S. Exposing deepfake videos by detecting face warping artifacts[J]. arXiv preprint arXiv:1811.00656, 2018.

[12] Li Y, Yang X, Sun P, et al. Celeb-df: A large-scale challenging dataset for deepfake forensics[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 3207-3216.

[13] Cozzolino D, Thies J, Rössler A, et al. Forensictransfer: Weakly-supervised domain adaptation for forgery detection[J]. arXiv preprint arXiv:1812.02510, 2018.

[14] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, FangWen, and Baining Guo. Face x-

ray for more general face forgery detection[C]. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5001–5010, 2020. 1, 3

[15] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now[C]. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, volume 7, 2020. 1, 3

[16] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in gan fake images[C]. In 2019 IEEE International Workshop on Information Forensics and Security (WIFS), pages 1–6. IEEE, 2019. 1, 3

[17] Frank J, Eisenhofer T, Schönherr L, et al. Leveraging frequency analysis for deep fake image recognition[C]//International Conference on Machine Learning. PMLR, 2020: 3247-3258.

[18] Ricard Durall, Margret Keuper, Franz-Josef Pfreundt, and Janis Keuper. Unmasking deepfakes with simple features. arXiv preprint arXiv:1911.00686, 2019. 1, 3

[19] Liu H, Li X, Zhou W, et al. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 772-781.

[20] Thies J, Zollhofer M, Stamminger M, et al. Face2face: Real-time face capture and reenactment of rgb videos[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2387-2395.

[21] Thies J, Zollhöfer M, Nießner M. Deferred neural rendering: Image synthesis using neural textures[J]. ACM Transactions on Graphics (TOG), 2019, 38(4): 1-12.

[22] Qian Y, Yin G, Sheng L, et al. Thinking in frequency: Face forgery detection by mining frequency-aware clues[C]//European Conference on Computer Vision. Springer, Cham, 2020: 86-103.

[23] Aybars Ciftci U, Demir I. FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals[J]. arXiv e-prints, 2019: arXiv: 1901.02212.

[24] Fernandes S, Raj S, Ortiz E, et al. Predicting heart rate variations of deepfake videos using neural ode[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. 2019: 0-0.

[25] Qi H, Guo Q, Juefei-Xu F, et al. Deeprhythm: Exposing deepfakes with attentional visual heartbeat rhythms[C]//Proceedings of the 28th ACM International Conference on Multimedia. 2020: 4318-4327.

[26] Tolosana R, Vera-Rodriguez R, Fierrez J, et al. Deepfakes and beyond: A survey of face manipulation and fake detection[J]. Information Fusion, 2020, 64: 131-148.

[27] Juefei-Xu F, Wang R, Huang Y, et al. Countering Malicious DeepFakes: Survey, Battleground, and Horizon[J]. arXiv preprint arXiv:2103.00218, 2021.

[28] Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks[C]//Proceedings of the IEEE/CVF Conference On Computer Vision and Pattern Recognition. 2019: 4401-4410.

[29] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[J]. Advances in neural information processing systems, 2020, 33: 1877-1901.

[30] Karras T, Aila T, Laine S, et al. Progressive growing of gans for improved quality, stability, and variation[J]. arXiv preprint arXiv:1710.10196, 2017.

[31] FaceSwap, https://github.com/MarekKowalski/FaceSwap. Accessed:2021-08.

[32] Matern F, Riess C, Stamminger M. Exploiting visual artifacts to expose deepfakes and face manipulations[C]//2019 IEEE Winter Applications of Computer Vision Workshops (WACVW). IEEE, 2019: 83-92.

[33] Rossler A, Cozzolino D, Verdoliva L, et al. Faceforensics++: Learning to detect manipulated facial images[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 1-11.

[34] Chollet F. Xception: Deep learning with depthwise separable convolutions[C]//Proceedings of the IEEE Conference on Computer Cision and Pattern Recognition. 2017: 1251-1258.

[35] Güera D, Delp E J. Deepfake video detection using recurrent neural networks[C]//2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS). IEEE, 2018: 1-6.

[36] 陈鹏,梁涛,刘锦,戴娇,韩冀中.融合全局时序和局部空间特征的伪造人脸视频检测方法[J].信息安全学报,2020,5(02):73-83.DOI:10.19363/J.cnki.cn10-1380/tn.2020.02.06.

[37] Sabir E, Cheng J, Jaiswal A, et al. Recurrent convolutional strategies for face manipulation detection in videos[J]. Interfaces (GUI), 2019, 3(1): 80-87.

[38] 李旭嵘,于鲲.一种基于双流网络的 Deepfakes 检测技术[J].信息安全学报,2020,5(02):84-91.DOI:10.19363/J.cnki.cn10-1380/tn.2020.02.07.

[39] Dang H, Liu F, Stehouwer J, et al. On the detection of digital face manipulation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition. 2020: 5781-5790.

[40] Liu Z, Qi X, Torr P H S. Global texture enhancement for fake face detection in the wild[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 8060-8069.

[41] Zhao H, Zhou W, Chen D, et al. Multi-attentional deepfake detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 2185-2194.

[42] Durall R, Keuper M, Keuper J. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 7890-7899.

[43] Wu X, Xie Z, Gao Y T, et al. Sstnet: Detecting manipulated faces through spatial, steganalysis and temporal features[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020: 2952-2956.

[44] Masi I, Killekar A, Mascarenhas R M, et al. Two-branch recurrent network for isolating deepfakes in videos[C]//European Conference on Computer Vision. Springer, Cham, 2020: 667-684.

[45] Korshunov P, Marcel S. Deepfakes: a new threat to face recognition? assessment and detection[J]. arXiv preprint arXiv:1812.08685, 2018.

[46] Dufour N, Gully A. Contributing data to deepfake detection research[J]. Google AI Blog, 2019, 1(2): 3.

[47] Dolhansky B, Howes R, Pflaum B, et al. The deepfake detection challenge (dfdc) preview dataset[J]. arXiv preprint arXiv:1910.08854, 2019.

[48] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference On Computer Vision and Pattern Recognition. 2016: 770-778.

[49] Gatys L A, Ecker A S, Bethge M. Texture synthesis using convolutional neural networks[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1. 2015: 262-270.

[50] Gatys L A, Ecker A S, Bethge M. Image style transfer using convolutional neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2414-2423.

[51] Morgan M J, Ross J, Hayes A. The relative importance of local phase and local amplitude in patchwise image reconstruction[J]. Biological cybernetics, 1991, 65(2): 113-119.

[52] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017: 6000-6010.

[53] Tolstikhin I, Houlsby N, Kolesnikov A, et al. MLP-Mixer: An all-MLP Architecture for Vision[J]. arXiv preprint arXiv:2105.01601, 2021.

[54] Loshchilov I, Hutter F. Decoupled weight decay regularization[J]. arXiv preprint arXiv:1711.05101, 2017.

[55] Fridrich J, Kodovsky J. Rich models for steganalysis of digital images[J]. IEEE Transactions on information Forensics and Security, 2012, 7(3): 868-882.

[56] Cozzolino D, Poggi G, Verdoliva L. Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection[C]//Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security. 2017: 159-164.

[57] Bayar B, Stamm M C. A deep learning approach to universal image manipulation detection using a new convolutional layer[C]//Proceedings of the 4th ACM Workshop On Information Hiding and Multimedia Security. 2016: 5-10.

[58] Rahmouni N, Nozick V, Yamagishi J, et al. Distinguishing computer graphics from natural images using convolution neural networks[C]//2017 IEEE Workshop On Information Forensics and Security (WIFS). IEEE, 2017: 1-6.

[59] Gunawan T S, Hanafiah S A M, Kartiwi M, et al. Development of photo forensics algorithm by detecting photoshop manipulation using error level analysis[J]. Indonesian Journal of Electrical Engineering and Computer Science, 2017, 7(1): 131-137.

[60] Khosla P, Teterwak P, Wang C, et al. Supervised contrastive learning[J]. Advances in Neural Information Processing Systems, 2020, 33: 18661-18673.

[61] Fung S, Lu X, Zhang C, et al. DeepfakeUCL: Deepfake Detection via Unsupervised Contrastive Learning[C]//2021 International Joint Conference On Neural Networks (IJCNN). IEEE, 2021: 1-8.

[62] Feng D, Lu X, Lin X. Deep detection for face manipulation[C]//International Conference On Neural Information Processing. Springer, Cham, 2020: 316-323.

[63] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[64] Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks[C]//International Conference On Machine Learning. PMLR, 2019: 6105-6114.