Deepfake检测研究报告

张瀚文，2201212865，信息工程学院

2022年11月29日

## 1 研究背景

随着深度学习技术的快速发展，特别是生成式对抗网络(GAN)，导致了Deepfake的产生。Deepfake算法可以创建虚假的图像和视频，人类难以将它们与真实的图像和视频区分开来，对个人隐私和国家安全都构成了巨大的威胁。伴随着Deepfake技术不断迭代，检测合成人脸的挑战也越来越多。Deepfake检测通常被认为是一个二元分类问题，包括Deepfake图像检测和Deepfake视频检测，其检测方法并不完全相同，但大体上可以分为人为视觉特征提取和深度学习两类。

在人为视觉特征方面，S. McCloskey等人提出使用简单的颜色特征和SVM分类器来检测Deepfake图像，而Alexandros Haliassos等人利用读唇语学习到的丰富表征，实现了对未见伪造类型的最先进的泛化，同时对各种常见损坏的鲁棒性明显高于其他方法。

近期，使用深度学习进行Deepfake检测取得了很大的进展。为了避免专注于特定的篡改取证，实现鲁棒的篡改检测，Zhou.P等人提出了一种双流网络结构来捕获篡改伪迹证据和局部噪声残差证据。H. Dang, 等人在检测中引入注意力机制，并将其与流行的CNN模型相结合，如exception-Net和VGG16，最终取得了令人印象深刻的结果，表明了注意力机制的重要性。Guoet等人提出一个新的CNN模型——SCnet，采用了分层特征提取块的结构，能够自动学习图像数据的高级取证特征。Li等人的工作Face X-ray尝试定位目标和原始人脸之间的混合边界，而不是捕捉特定操作的合成伪影，且可以在没有假图像的情况下进行训练。N. Yu等人则利用GAN生成Deepfake图像时留下特殊的“指纹”来进行检测。此外，Joel.Frank等人的工作首次在频域考虑假脸检测问题，揭示了频谱不同的核心原因是上采样操作，也暗示了现在GAN从低维隐空间映射到高维空间的结构性问题。

长期以来，解决Deepfake检测最常见的方式是学习一个二分类模型来进行真/假判别。但如今由于人脸编辑App的流行，我们可以非常方便地对人脸图片进行多步序列DeepFake篡改。为解决此类问题，来自南洋理工大学的研究人员在ECCV2022的工作中提出了检测并还原Deepfake篡改序列（Seq-DeepFake）任务，贡献了首个大规模的Seq-DeepFake数据集，并提出了一个简单且有效的Seq-DeepFake Transformer作为此新课题很好的起始方案。

## **2 论文阅读**

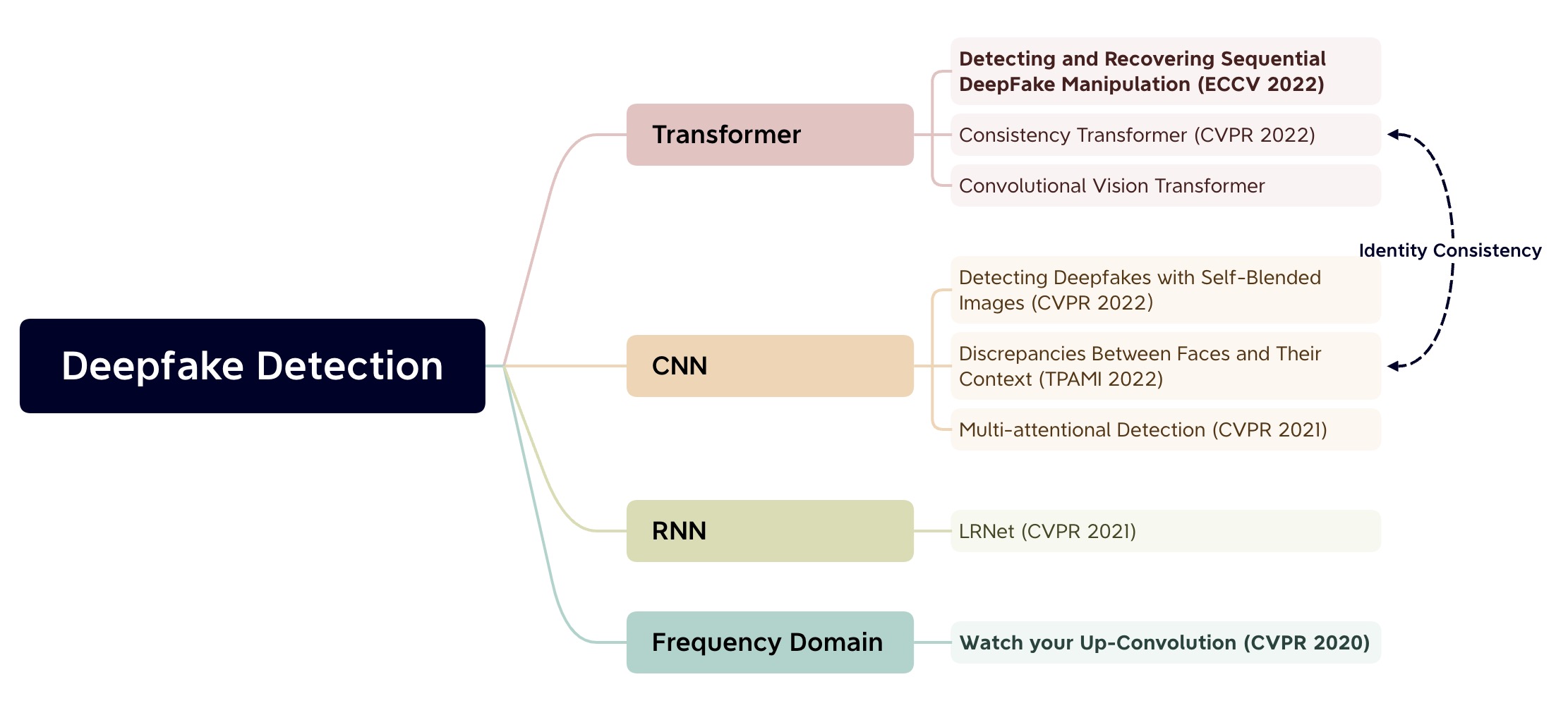


图1 Deepfake检测论文方向归纳

### 2.1 基于Transformer的方法

#### 2.1.1 Detecting and Recovering Sequential DeepFake Manipulation（精读）

这篇论文首次提出检测并还原DeepFake篡改序列（Seq-DeepFake）的任务。相比于现有基于二分类 (真/假) 的DeepFake检测，作者提出DeepFake篡改序列检测这个研究课题，建立并开源了全球首个Seq-DeepFake数据集。为了更全面的研究此问题，本文考虑了两种不同的篡改方法，即人脸部件序列篡改和人脸属性序列篡改。

论文指出，现有的人脸编辑算法基本基于GAN，这会导致在编辑一种人脸部件/属性后，会间接影响其他人脸部件/属性的间关系。也就是说，我们可以在空间篡改痕迹的基础上捕捉到序列篡改痕迹, 并最终据此检测出DeepFake篡改序列。根据上述观察，作者认为可以把Seq-DeepFake看成是一种特殊的image-to-sequence problem，并提出Seq-DeepFake Transformer来检测篡改序列。根据检测出的篡改序列，我们可以逆序还原出原始人脸。

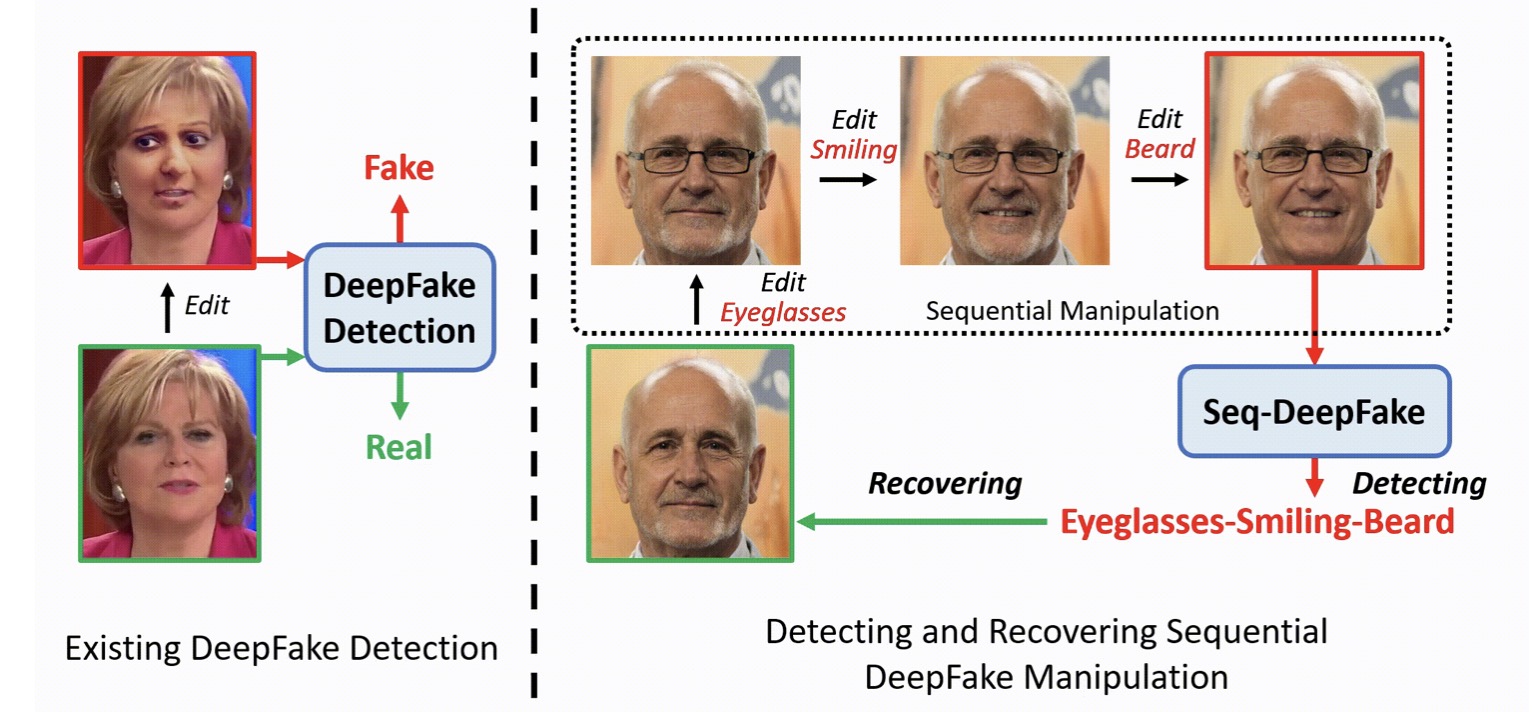


图2 DeepFake篡改序列检测任务

#### 2.1.2 Protecting Celebrities from DeepFake with Identity Consistency Transformer

这篇论文中，作者提出了身份一致性Transformer，这是一种新的人脸伪造检测方法，专注于高层语义，特别是身份信息，并通过发现内外人脸区域的身份不一致来检测可疑人脸。身份一致性变换器结合了一致性损失，用于身份一致性确定。身份一致性Transformer在不同的数据集上以及现实世界中各种类型的图像退化形式上都表现出优异的泛化能力。

#### 2.1.3 Deepfake Video Detection Using Convolutional Vision Transformer

这篇论文中，作者提出用于Deepfakes 检测的Convolutional Vision Transformer，其包含两个组件：卷积神经网络和Vision Transformer (ViT)。其中，CNN提取可学习的特征，而 ViT将学习到的特征作为输入并使用注意力机制对其进行分类。其主要贡献是在ViT架构中添加了一个CNN模块，并在DFDC数据集上取得了有竞争力的结果。

### 2.2 基于CNN的方法

#### 2.2.1 Detecting Deepfakes with Self-Blended Images

这篇论文的目标是构造一类Deepfake detection训练集，作者提出了一种称为自混合图像（SBI）的新型合成训练数据来检测深度伪造。SBI是通过从单个原始图像中混合伪造源图像和目标图像，再现常见的伪造伪迹（例如，混合边缘、源图像与目标图像之间的统计不一致性）。SBI背后的关键思想是，更通用且难以识别的假样本鼓励分类器学习泛化且健壮的表征，而不会过拟合于特定操作的伪迹。检测模型整体框架基于EfficientNet。

#### 2.2.2 Discrepancies Between Faces and Their Context

这篇论文提出了一种识别假图像的新信号，该信号是基于比较内部面部内区域（直接操作的区域）和外轮廓区域（我们所知的所有面部操作方法都不会改变外部轮廓），与2.1.2中身份一致性Transformer模型在思路上有相似之处。其中，表示面部和外轮廓的两个向量是通过训练两个独立的人脸识别网络（基于XceptionNet）获得的：一个网络基于面部内区域识别身份，另一个网络基于外轮廓区域识别身份。

#### 2.2.3 Multi-attentional Deepfake Detection

这篇论文将Deepfake检测表述为一个细粒度的分类问题，并提出了一种新的多注意力的Deepfake检测网络，使用多注意力机制捕获人脸的多个区域的特征。模型整体框架基于XceptionNet，在主干上新增了纹理增强模块、注意力模块、注意力pooling层。

### 2.3 基于RNN的方法

#### 2.3.1 Improving Deepfakes Detection through Precise Geometric Features

论文提出了LRNet框架，通过对精确的几何特征进行时间建模来检测Deepfakes视频。LRNet由四个部分组成：人脸预处理模块、校准模块、特征嵌入过程和RNN分类过程。它通过检测异常的面部运动模式和时间不连续性来暴露被操纵的面部，并且整个框架只需要训练RNNs部分。与以前的方法相比，LRNet更轻量，更容易训练，并且在检测高度压缩或噪声污染的视频中表现出鲁棒性。

### 2.4 基于频域 (Frequency Domain) 的方法

#### 2.4.1 Watch your Up-Convolution（精读）

深度生成模型从中间隐特征生成图像有两种方式：up-convolution 和 transposed convolution。论文在自编码器上使用不同的上采样方式重构真实图像，发现up-convolution方法缺少高频信息，而transposed convolution方法引入了太多的高频信息。基于这样的观察，作者设计了利用频域信息的Deepfake检测方法，如下图所示，左图代表真实图像功率谱和生成图像功率谱，而右图是考虑spectral loss后的功率谱。经过DFT后，将2D Power Spectrum根据不同半径池化成1D Power Spectrum，最后使用SVM进行分类。

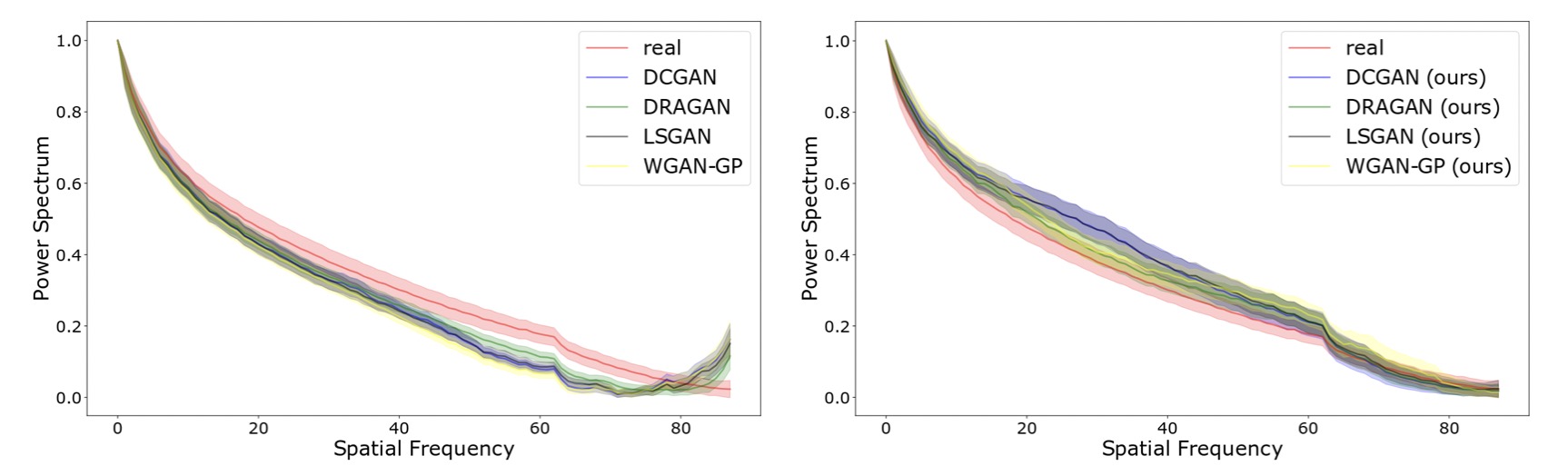


图3 真实图像和生成图像图像功率谱

## 3 实验介绍

对于所选择的两篇精读论文，基于频域方法的Watch your Up-Convolution (CVPR 2020)以及篡改序列检测Detecting and Recovering Sequential DeepFake Manipulation (ECCV 2022)均有开源的官方实现。目前，论文Watch your Up-Convolution所介绍的频域检测方法已基于GitHub仓库<https://github.com/cc-hpc-itwm/DeepFakeDetection.git>成功复现。而用于检测并还原DeepFake篡改序列的Transformer模型Seq-DeepFake Transformer已经根据<https://github.com/rshaojimmy/SeqDeepFake>初步跑通代码，但由于数据集较大，训练需要比较长的时间，暂时仅完成了30个epoch的训练，接下来会继续训练和尝试改进这个模型。

### 3.1 Watch your Up-Convolution的复现

#### 3.1.1 数据集简介

作者在三个数据集上测试了论文所介绍的利用频域信息的Deepfake检测方法，包括：CelebA、FaceForensics++和作者自己建立的Faces-HQ数据集。其中，作者自己建立的Faces-HQ数据集收集了来自CelebA-HQ dataset, Flickr-Faces-HQ dataset, 100K Facesproject和 [www.thispersondoesnotexist.com](http://www.thispersondoesnotexist.com)的高质量图像，共计40K张，而且其中一半是真实人脸，一般是合成假脸，使得该数据集是一个平衡的数据集。

#### 3.1.2 实验过程与结果

根据作者公布的源码，整个Deepfake检测方法的pipline可以总结如下：Create feature -> Loading Features -> Check Spectrum -> Classification. 根据原文，在三个数据集上按照80%-20%的比例划分训练集和测试集，进行实验。使用SVM分类器，其中CelebA上测试准确率可达99%，FaceForensics++上测试准确率可达86%，而在作者自己建立的Faces-HQ上测试准确率达到100%。

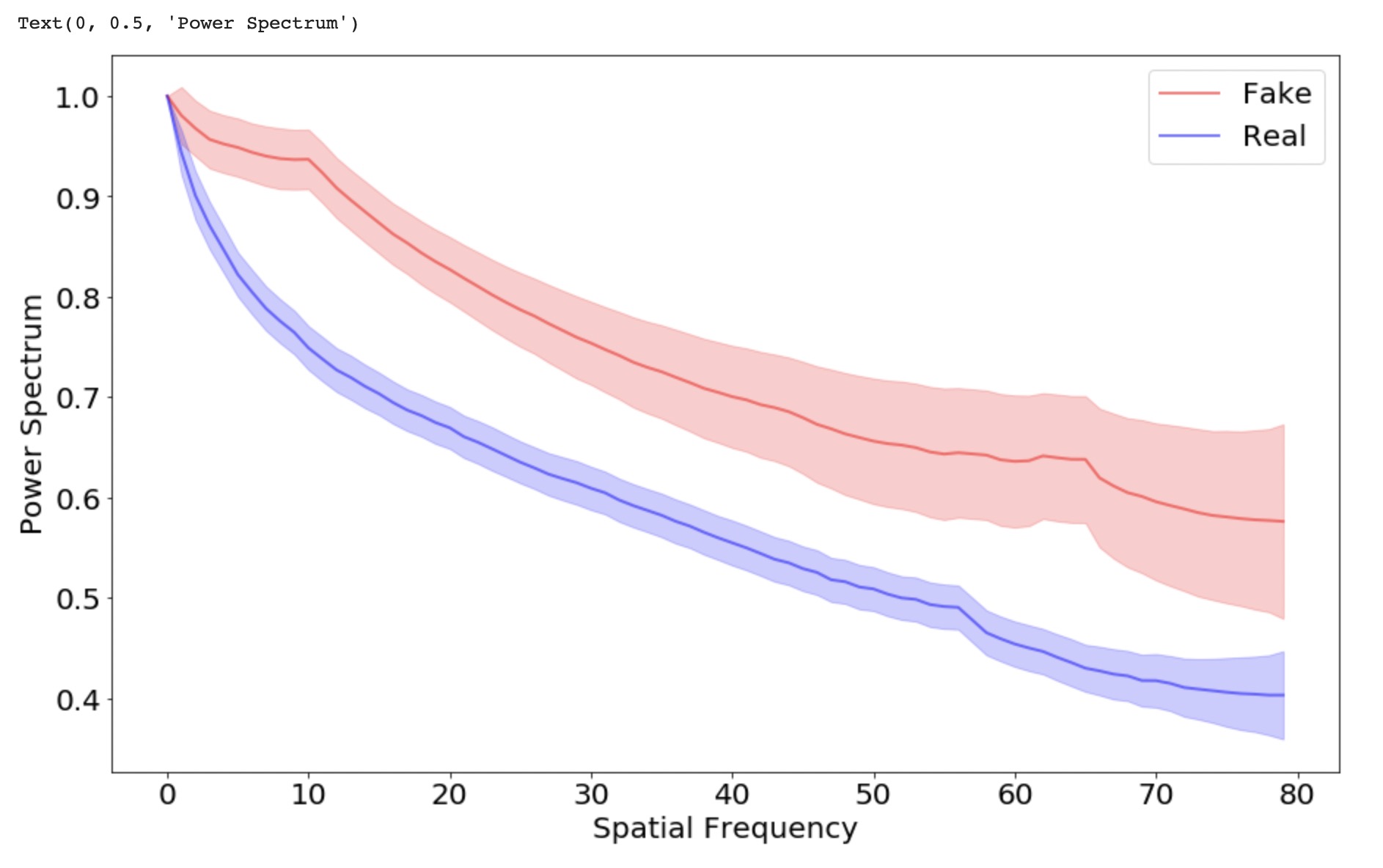


图4 CelebA数据集功率谱

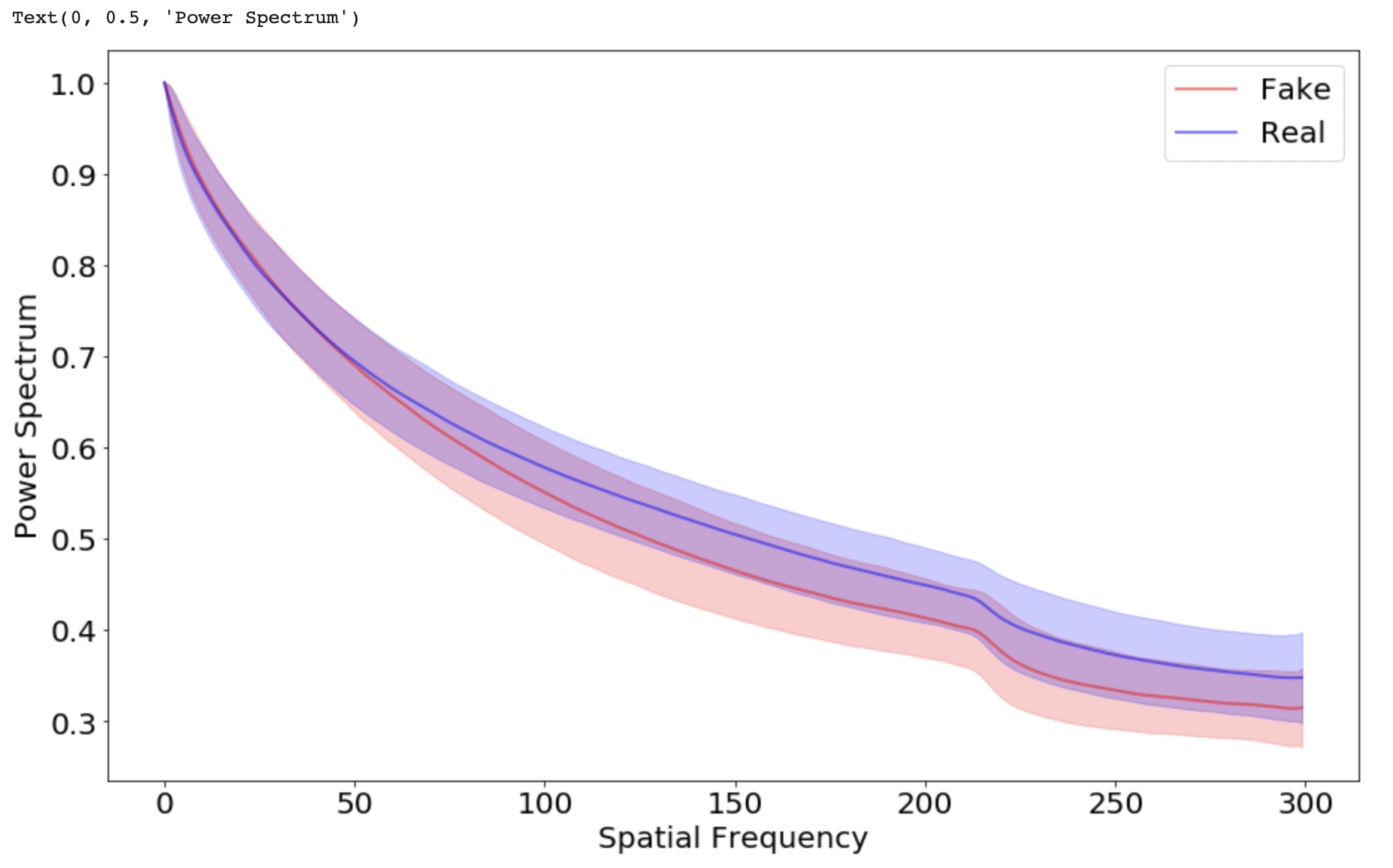


图5 FaceForensics++数据集功率谱

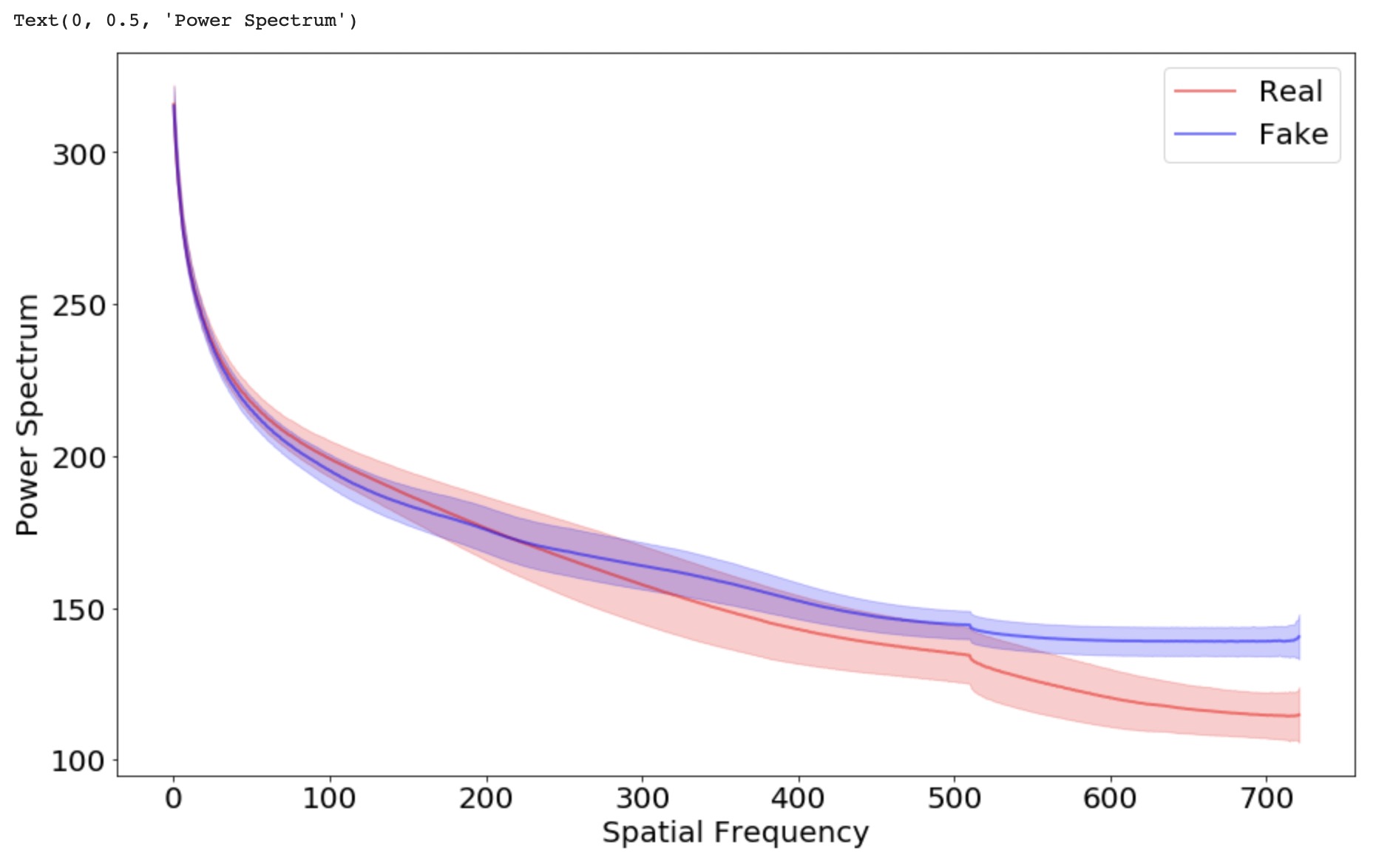


图6 Faces-HQ数据集的功率谱

#### 3.1.3 改进思路

作者所使用的Deepfake检测方法，基于频域所提取的特征二分类，优点是模型轻量，训练、测试的时间都比较短。但在测试中，也反映出了一些局限性。可以看出，对于质量较高的CelebA数据集数据集，检测的效果是比较理想的，但在FaceForensics++上，则出现了明显的准确率下降。出现这种情况，一个比较重要的原因是，视频压缩、以及任何采样、缩放、裁剪都会对频谱信息造成破坏，使得检测方法的性能大大下降。此外，作者的源码中没有使用人脸检测模块，因此，对于一些复杂的场景，如果人脸不是主要的输入部分，方法同样可能失效。

基于以上几点观察，我认为可能的改进实验方案有：（1）在所有操作开始前，添加人脸检测模块，使得人脸为主要输入部分，并观察在FaceForensics++上是否有精度提升；（2）对数据集中的图像做裁剪、缩放等破坏性操作，观察功率谱的变化，尝试仔频域上寻找到更加鲁棒的检测特征。

### 3.2 Seq-DeepFake Transformer的复现

#### 3.2.1 数据集简介

本论文所使用的数据集是作者自己建立的Sequential DeepFake数据集。作者贡献了第一个大规模的Sequential DeepFake数据集Seq-Deepfake，其中包括85k个顺序操作的人脸图像，每个图像都标注其的真实操作序列。数据集中图像是根据以下两种不同的面部操作方法生成的：（1）顺序面部组件操作（基于CelebAMask-HQ和StyleMapGAN），（2）顺序面部属性操作（基于FFHQ和Talk-To-Edit），共具有28/26种类型的操作序列。所有操作序列的长度从1~5不等。数据集中的每张图像都标注了长度为5的列表，指示了真实操作序列。

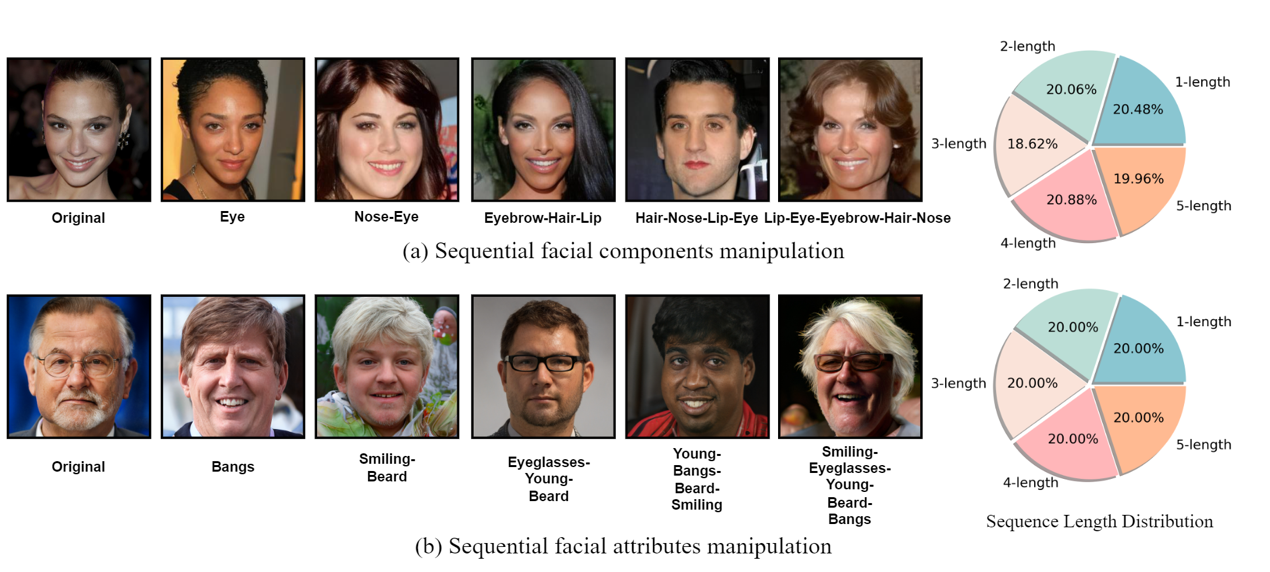


图7 Seq-Deepfake数据集介绍

#### 3.2.2 前期准备与训练

**下载数据集**：在<https://github.com/rshaojimmy/SeqDeepFake>仓库找到数据集的下载入口，按照指示下载即可。下载后，先不要在本地解压。

**运行环境**：在colab在线平台训练。首先，需要在colab上准备好运行的环境，按照作者要求，所需的package如下，在colab平台上使用pip命令安装。除此之外，还需要把数据集（压缩）、整个源码仓库先上传到谷歌云盘，再在colab挂载云盘。

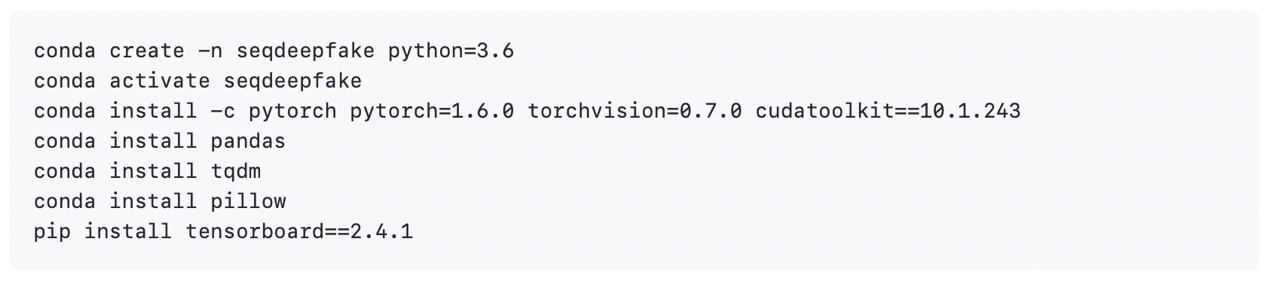


图8 运行环境

**训练**：通过在colab输入sh train.sh来开始训练。每个epoch耗时接近9min，训练30个epoch后，结果如下图所示。

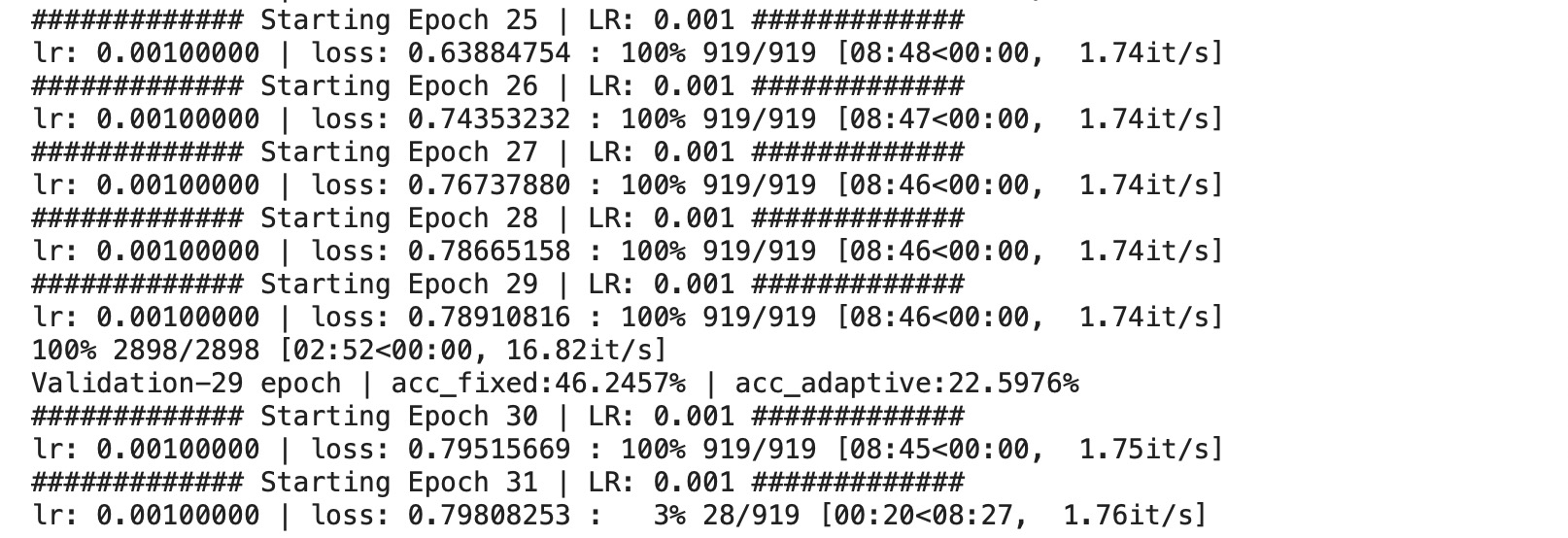


图9 训练Seq-DeepFake Transformer

#### 3.2.3 改进思路

Seq-DeepFake实际上是一种特殊的image-to-sequence problem，所以其中一个可以尝试的方向是，用其他（Transformer等）结构，替换目前使用的Seq-DeepFake Transformer，观察精度是否改进。