# 基于频域方法的伪造检测算法文献阅读

小组成员：马子睿，王韫，王家安

我们针对伪造检测这个子领域进行了文献阅读。随着深度学习的迅猛发展，利用深度学习伪造或生成图像的技术愈发成熟，有些生成图像甚至达到了以假乱真的程度。这也引起了人们对安全和伦理的关注，带来了人们对图像的信任危机。由于近年来的论文绝大多数都是以深度学习作为主要的解决办法，而传统的图像处理方法一般作为一种辅助的处理和研究手段。我们找到了三篇基于频域的伪造检测方法进行了阅读和总结。三篇文献分别为:

1. Detecting and Simulating Artifacts in GAN Fake Images. In WIFS 2019,被引数：577
2. Spatial-Phase Shallow Learning: Rethinking Face Forgery Detection

in Frequency Domain. In CVPR 2021,被引数：440

1. Generalizing Face Forgery Detection with High-frequency Features. In CVPR 2021，被引数：401

尽管三篇文章都是从频域出发对伪造检测任务进行分析，但三篇文章并没有互相引用的关系。因此，我们将分别对三篇文章进行介绍。

# GAN伪图像中伪影的检测与模拟

## 研究背景和意义

基于机器学习的方法，比如那些基于生成对抗网络（GAN）的方法，已经使得创建接近真实的假图像比以前更加可行，并使娱乐和教育中许多有趣的应用成为可能。一些由最新的GAN模型生成的高分辨率图像与人类观众的真实图像很难区分，这也引起了人们对安全和伦理的关注，因为将视觉媒体视为可信内容的传统观点已不再有效。作为一种部分的补救措施，开发一种自动工具来区分真实和GAN生成的图像将提供巨大的价值。

一个典型的设计真实的方法。GAN假图像分类器是指从一个或多个预先训练过的GAN模型中收集大量的GAN生成的图像，并训练一个二值分类器。不幸的是，在现实世界的应用程序中，本文通常无法访问攻击者所使用的特定模型。这对检测器的泛化性提出了挑战。因此，开发一种可以在所有GAN的变体模型生成的图像上都有效的检测器是解决这一问题的关键。

## **问题描述**

在以往的研究中，部分方法尝试在某个特定的GAN模型上提取特征用来训练分类器，这些基于机器学习的方法都需要由一个或多个预先训练过的GAN模型生成足够的训练图像，以确保分类器的泛化能力。

在现实世界的应用程序中，通常不可能访问用于生成假图像的预先训练好的GAN模型。因此，本文研究了如何通过理解、检测和模拟GAN生成管道中产生的伪影，来消除在训练GAN伪图像分类器时访问预训练模型的要求。

## **解决方案**

### **3.1 在GAN管道中的上采样伪影的研究**

GAN模型包含两个主要组件：鉴别器和生成器。

生成器包含两个组件，编码器和解码器。该编码器包含几个下采样层试图从输入的图像中提取的高级别的信息，并生成一个低分辨率的特征张量。解码器则包含几个上采样层，以低分辨率特征张量为输入，输出高分辨率图像。

不同的GAN模型中使用的上采样模块是一致的。文献中最常用的两个上采样模块分别是转置卷积和最近邻插值。这两个上采样器都可以被表述为一个简单的流程，如图2所示。给定一个低分辨率的特征张量作为输入，上采样器将水平和垂直的分辨率都增加了m倍。上采样器在低分辨率特征张量中的每行/列之后插入一个零行/列，并应用卷积操作，以便为“零插入”的位置分配适当的值。转置卷积与最近邻插值的差异是，如图2所示，转置卷积中的卷积核是可学习的，而在最近邻插值中的卷积核是固定的。

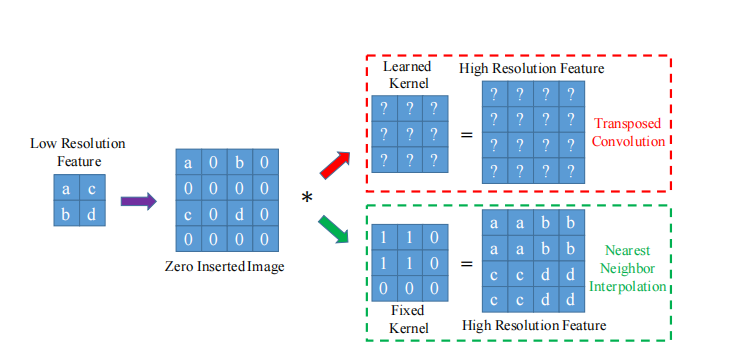


图2

转置卷积的上采样伪影称为“棋盘伪影”，本文在频域的进一步分析。根据离散傅里叶变换（DFT）的性质，如图3所示，在低分辨率图像中插入零，相当于在最终高分辨率图像的频谱的高频部分上复制原始低分辨率图像的频谱的多个副本。

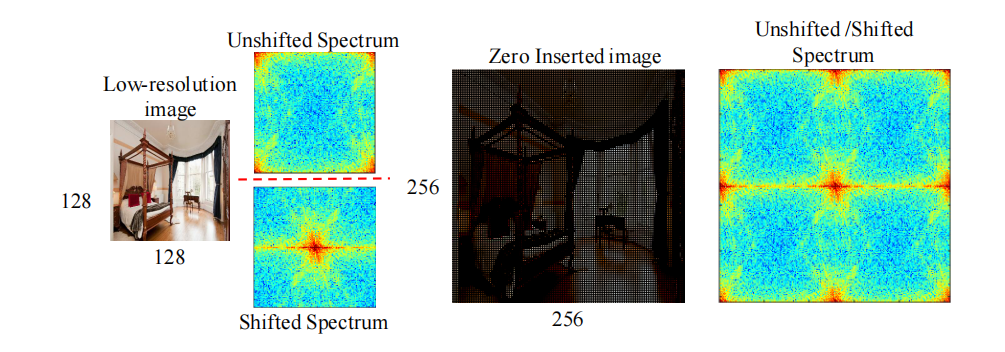


图3

如果上采样器是转置卷积，则不能保证学习到的卷积核是低通的。因此，在许多图像中仍然可以观察到棋盘格伪影。本文在图4中展示了一个例子，其中最左边的两幅图像是一个真实的人脸图像及其频谱图。右边的图像是由真实图像生成的假图像及其频谱图。棋盘格伪影在红色框中突出显示。在频谱中，在宽度和高度的1/4和3/4处有明亮的斑点，这与生成器中两个连续的上采样模块产生的伪影相对应。它是频域中的棋盘式伪影。

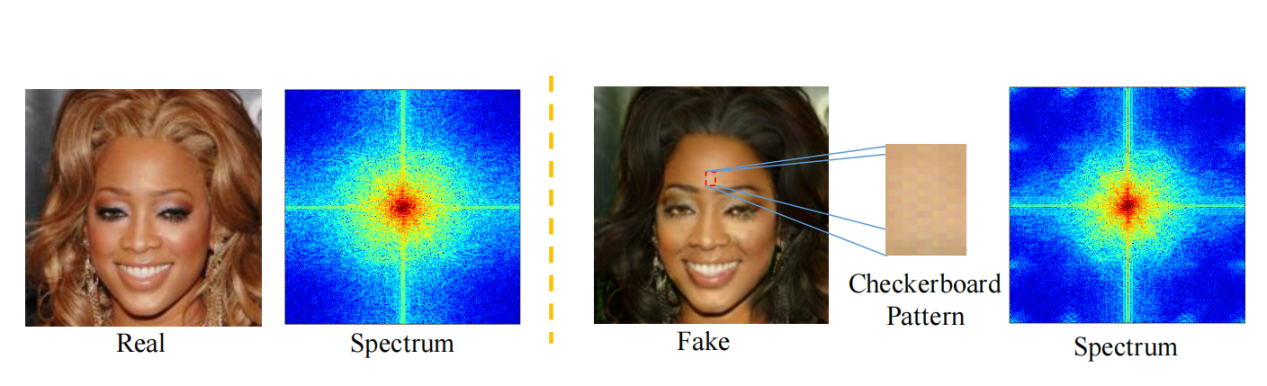


图4

基于上述发现的理论特性，本文使用图像的频谱图作为输入，而不是原始像素来训练GAN伪造图像分类器。

### **3.2 GAN伪影的检测**

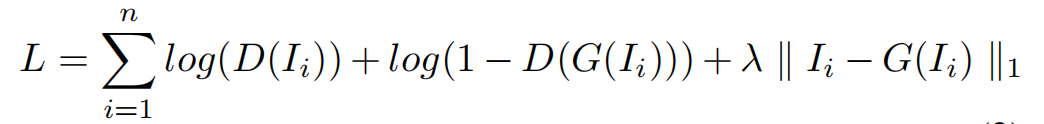
给定一个图像I作为输入，本文将二维DFT应用于每个RGB通道，得到3个通道的频谱图F（丢弃相位信息）。本文计算该频谱的对数log(F)，并将对数后的频谱归一化为[−1,1]。归一化频谱是假图像分类器的输入。因此，分类器的主要目标是揭示在前一节中识别出的伪影，以便将图像分类为是否由GAN模型生成的。基于频谱的分类器比基于像素的分类器具有更好的性能，特别是当训练数据只包含来自一个语义类别的图像时。

### **3.3 GAN伪影的模拟**

为了进一步允许本文在没有假图像的情况下训练分类器，本文提出了AutoGAN，这是一个GAN模拟器，可以在任何图像中合成GAN伪影，而不需要访问任何预先训练过的GAN模型。

针对GAN的结构，设计了一个用于生成GAN伪影的生成器AutoGAN。它将一个真实的图像(I)作为输入，并将其通过一个生成器(G)，该生成器的结构类似于图像生成GAN中使用的生成器。解码器包含上采样模块，如转置卷积或最近邻插值。

AutoGAN的模型结构如图5所示，与其将生成器的输出分布与另一个语义类别的图像相似，生成器的输出与原始图像本身相匹配。其损失函数如下：



D（·）为鉴别器，G（·）为生成器，G（Ii）为Ii作为G（·）输入时的输出，λ为在两种不同的损失之间权衡参数。

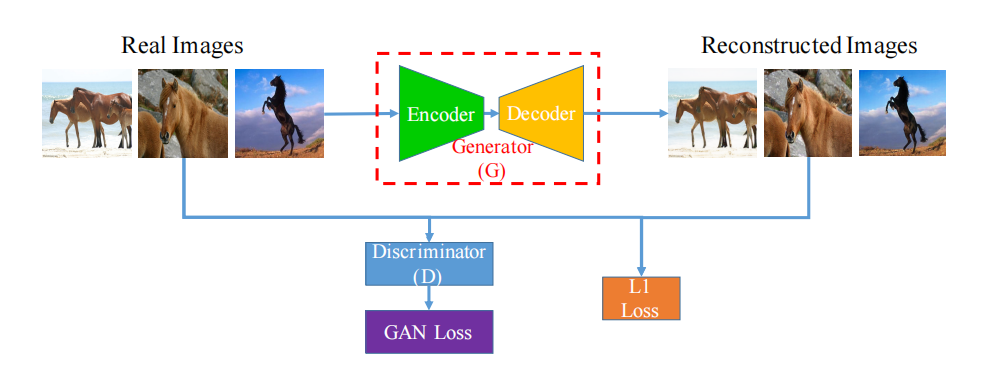


图5

## 实验分析

### **4.1 实验设置**

本文的实验在CycleGAN图像上进行一系列实验，选取了CycleGAN的训练集的14个类别进行训练和测试。本文使用预训练的resnet34作为判别真假的二分类器。以下共有四种设定:

Img：用真实图像和cycleGAN生成的假图像进行训练；

Spec:训练数据与Img相同，但是用频谱作为输入对分类器进行训练；

A-Img：用真实图像和由AutoGAN生成的假图像进行训练；

A-Spec：训练数据与A-Img相同，但是用频谱输入对分类器进行训练。

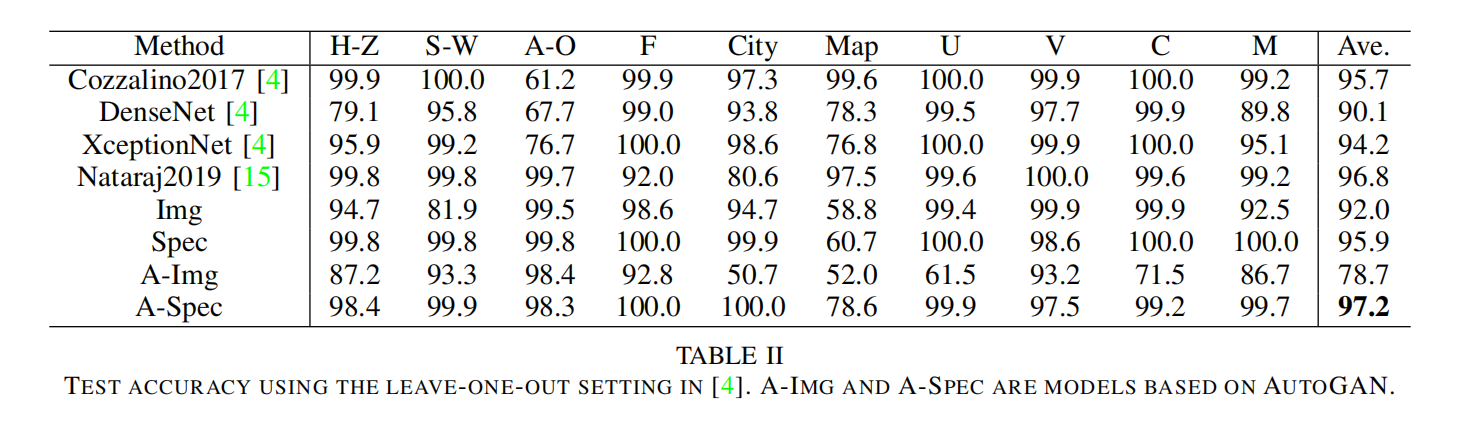
### **4.2 模块有效性**

### **0a86bf3d8eddc9226548f63b3d46e60**

通过在一种类别上进行训练（马或者斑马），基于频谱的分类器（Spec）大大提高了泛化能力（如用horse类别训练的A-Spec相比A-Img提升了28.7）。

AutoGAN的方法在某些类别上效果较好（如马这个类别作为训练），但在其他一些类别上表现较差（如斑马）。这意味着AutoGan要求训练图像的频谱具有足够的多样化以进行泛化。

### **4.3 对比实验**



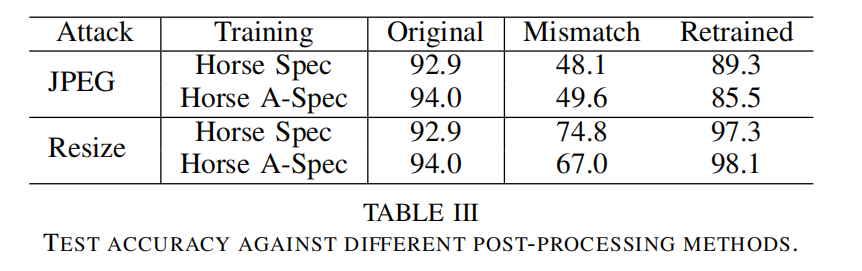
对比其他方法，AutoGAN+频谱图的方法可以在泛化性性能方面达到SOTA。

### **4.4 频带对比**

### 7b255682cd7f298fbaa17a79bc77879

对比使用不同频段的信息的频谱分类器的性能，发现中高频的效果相对更好。这是因为伪影大多出现在中高频段。

### **4.5 后处理影响对比**



对图像进行不同程度的resize处理和jpeg压缩处理，其中mismatch指训练时不做后处理，但是用后处理过的图片进行测试，retrained则是指训练时就对原图进行后处理。结果发现该方法并不能直接泛化后处理的图像，因为resize或jpeg压缩会破坏原先的高频伪影，导致模型失效。

## **结论**

本文研究了GAN管道的上采样器在频域内引起的伪影，以开发鲁棒的GAN伪图像分类器。为了检测这种伪影，本文建议使用频谱代替图像像素作为输入。它极大地提高了分类器的泛化能力。本文进一步提出了AutoGAN，它模拟了公共的GAN管道，并在真实图像中合成了GAN伪影。所提出的AutoGAN允许本文训练一个GAN假图像分类器，而不需要假图像作为训练数据或特定的GAN模型来生成这样的图像。基于AutoGAN频谱的分类器可以很好地推广到具有相似结构的GANs生成的假图像。

## **图像处理方法在本文中的作用**

在研究上采样模块所造成的伪影的机制时，本文发现这是一种low level的特征，很适合用图像处理中，将图像从空域使用傅里叶变换转换到频域再去处理。通过对频域的可视化，我们发现了GAN生成图像在高频区域的伪影，而让模型直接对带有明显伪影的频谱图进行学习也有效的提高了模型的性能。

# 基于空间相位的浅层学习：频域中的人脸伪造检测

## **研究背景和意义**

近年来，生成对抗网络（GAN）和变分自编码器（VAE）等生成技术的快速发展，使得生成高质量伪造人脸成为可能。这些伪造图像不仅可以以假乱真，还被用于非法活动，如制造虚假新闻或政治攻击。这些伪造技术不仅对社会安全构成威胁，还对个人隐私和伦理道德提出了挑战。因此，人脸伪造检测技术成为计算机视觉领域的重要研究方向。

传统的人脸伪造检测方法通常基于空间域特征，利用RGB像素信息来捕捉伪造痕迹。然而，这些方法存在两个主要问题：一是在跨数据集场景中，检测效果不佳，泛化性较弱；二是伪造技术的多样性导致单一空间域信息难以全面覆盖伪造特征。本文提出了一种新的方法——基于空间相位的浅层学习（SPSL），通过结合空间域和频域信息，提升了人脸伪造检测的准确性和泛化性。

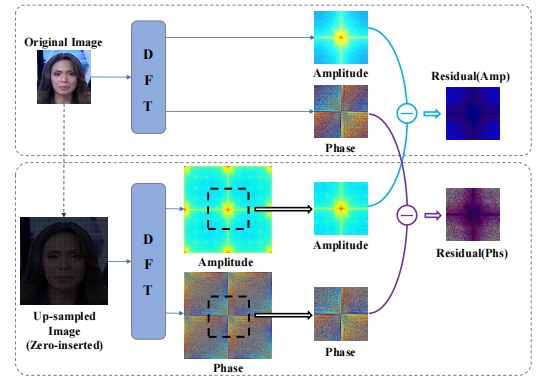
## **问题描述**

人脸伪造通常涉及三个阶段：1）编码源人脸；2）在潜在空间中进行特征交换；3）解码生成目标人脸。在解码过程中，上采样是一项必要步骤，该步骤会在图像的频域中留下独特的痕迹。尤其是伪造图像的相位频谱，其频率分量比幅度频谱更加丰富，且对上采样伪影更为敏感。

在以往的研究中，部分方法尝试利用频域信息检测伪造，但大多集中在幅度频谱上，而忽视了相位频谱蕴含的关键特征。此外，许多基于卷积神经网络（CNN）的检测方法依赖深层网络以提取高层语义信息，但对于人脸伪造检测而言，伪造的关键痕迹往往是局部的低级特征（如纹理和颜色），而非全局的高级语义信息。因此，设计一种结合相位频谱和浅层网络的检测方法成为解决上述问题的核心思路。

## **解决方案**

针对以上问题，本文提出了一种结合空间域和频域特征的检测框架，即基于空间相位的浅层学习（SPSL）。SPSL通过分析频域中的相位信息，捕捉上采样过程中的伪影，同时利用浅层网络专注于局部特征，从而提升检测精度和跨数据集的泛化能力。



3.1 相位频谱特征提取

为了更好地捕捉伪造图像中上采样带来的伪影，本文提出了一种基于相位频谱的特征提取方法。利用离散傅里叶变换（DFT），将图像从空间域转换到频域，并通过逆傅里叶变换（IDFT）重建相位频谱的空间表示。将这一空间表示与原始RGB图像结合，形成四通道输入（RGBP），以提供更多伪造特征。

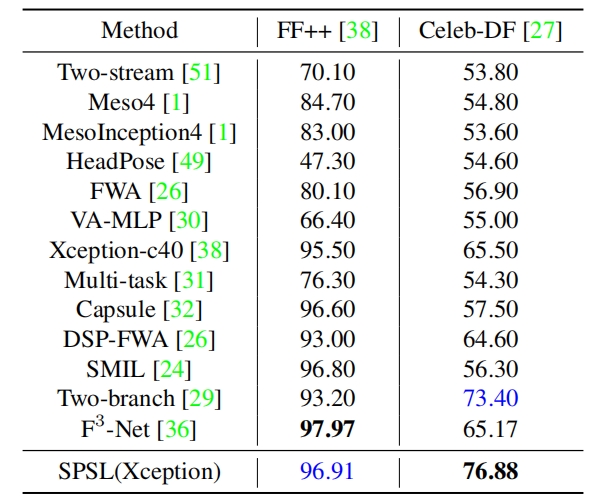
3.2 浅层网络结构  
 考虑到伪造图像的核心特征主要存在于局部区域，本文设计了一种浅层卷积神经网络，通过减少网络深度以减小感受野，专注于局部低级特征的提取。例如，本文在Xception网络的基础上，仅保留了部分浅层卷积模块，以抑制全局语义信息的干扰。

3.3 联合训练策略  
 在网络训练过程中，将空间域和频域信息作为输入，通过联合优化策略提升检测模型的鲁棒性和泛化能力。尤其是在跨数据集测试中，频域信息的引入显著增强了模型捕捉伪造特征的能力。

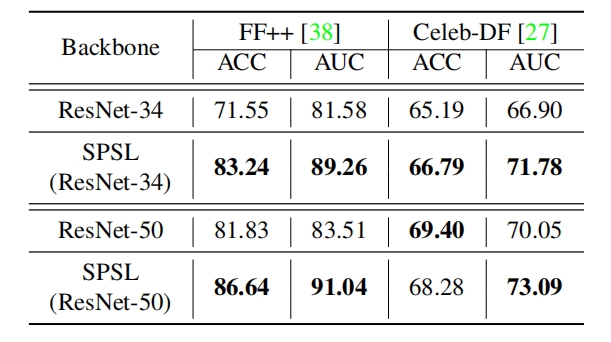
## **实验分析**

**数集与评价指标**  
 本文在三个公开数据集上进行实验：FaceForensics++（FF++）、Celeb-DF和DFDC。这些数据集包含多种伪造方法和多样化的图像质量（高压缩、低压缩、无压缩）。实验采用准确率（ACC）和AUC（受试者工作特征曲线下面积）作为主要评价指标。

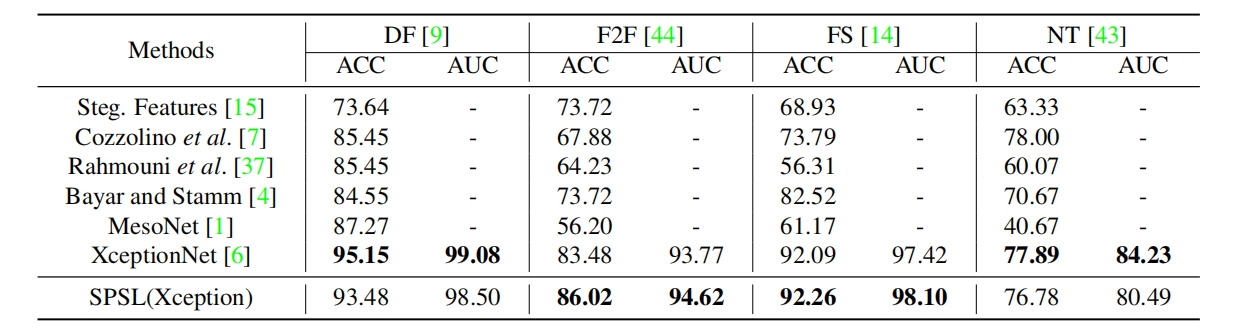
**跨据集测试**  
 为了验证模型的泛化能力，本文在FF++数据集上进行训练，并在Celeb-DF和DFDC数据集上进行测试。结果表明，SPSL在Celeb-DF上取得了76.88%的AUC，显著优于现有方法。此外，在DFDC上的表现也保持在较高水平。



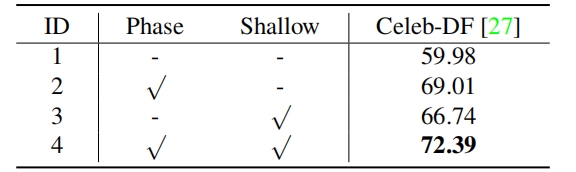
**单数据集性能**  
 在FF++数据集上，SPSL在低质量（LQ）设置下的AUC达到82.82%，与当前最优方法持平甚至更优。实验进一步表明，频域信息的引入显著提升了低质量图像上的检测性能。



**多分类任务**  
 为了验证SPSL在不同伪造方法上的区分能力，本文进行了多分类实验。结果显示，SPSL能够有效区分不同类型的伪造方法，例如DeepFake、Face2Face和NeuralTextures。这表明，相位频谱的独特性对分类任务具有重要贡献。



**对比实验与消融分析**  
 本文通过消融实验验证了相位频谱和浅层网络的独立贡献。相比仅使用RGB图像的基线模型，引入相位频谱后，AUC提高了约9个百分点；结合浅层网络后，进一步提升了3个百分点，最终在Celeb-DF上达到72.39%的AUC。



## **图像处理算法的特色**

**相位频谱的应用**  
 通过分析频域中相位频谱的变化，捕捉上采样伪影。相位频谱对频率成分的敏感性使其成为检测伪造的重要工具。

**逆傅里叶变换的重建技术**

通过IDFT将相位频谱重构为空间域表示，结合RGB图像，实现空间域和频域信息的联合建模。

**浅层积神经网络**

浅层网络的设计减少了感受野，使模型更加关注局部纹理特征，而非全局语义信息，从而提高了伪造检测的精确性。

## **总结与展望**

本文提出了一种结合空间信息和频域信息的创新性方法，通过利用相位频谱和浅层网络结构，在人脸伪造检测任务中取得了显著的性能提升。实验表明，该方法在跨数据集测试和多分类任务中具有出色表现。未来，随着人脸伪造技术的不断进步，如何进一步增强检测方法的鲁棒性与适应性，将是一个重要的研究方向。

# 基于高频特征的面部伪造检测的泛化研究

## 研究背景和意义

随着深度生成技术的突破，面部伪造技术（如Deepfakes、FaceSwap、NeuralTextures等）正变得日益流行。这些技术的快速发展为娱乐行业、教育领域以及虚拟现实技术带来了巨大的机遇。例如，在电影制作中，这些技术可以帮助演员完成无法亲自表演的场景，在教育中，这些技术能够生成生动形象的虚拟教学助手。然而，这些技术也被滥用，用于制造高度逼真的伪造视频或图像，从而引发了一系列社会问题。

虚假内容的传播对公众信任和社会稳定造成了严重威胁。例如，伪造的政治人物演讲可能导致公共恐慌或政策误导，而虚假的名人视频则可能毁坏个人声誉甚至引发法律纠纷。此外，伪造内容的使用还可能对选举公平性、金融安全性和司法系统的公正性构成威胁。因此，研究有效且鲁棒的面部伪造检测技术具有重要的社会意义。

目前，大多数伪造检测方法主要针对“数据库内检测”，即训练和测试伪造图像由相同算法生成。然而，这些方法在跨数据库场景下（训练和测试伪造图像由不同算法生成）的性能显著下降。这是因为不同的伪造技术生成的图像具有显著的分布差异，现有方法无法充分泛化到这些不同分布的数据上。因此，开发一种能够应对多样化伪造生成技术的检测方法是解决这一问题的关键。

## **问题描述**

尽管现有面部伪造检测方法在特定数据库上表现出色，但在实际应用中，这些方法往往会因为泛化能力不足而失效。具体而言，这些方法通常依赖于CNN模型对伪造图像中特定纹理特征的捕捉。然而，这种“纹理偏差”使得模型容易过拟合到某种特定伪造技术生成的纹理，而难以适应其他技术生成的伪造图像。

例如，当训练数据由Face2Face算法生成时，模型可能专注于Face2Face伪造图像中特定的面部区域（如嘴部）的异常纹理（图1所示）。然而，当面对未见过的伪造技术（如Deepfakes或NeuralTextures）时，这些纹理特征可能完全不同，从而导致模型难以有效检测。实验表明，模型在不同伪造技术上的性能差异显著，表明其泛化性能存在明显问题。

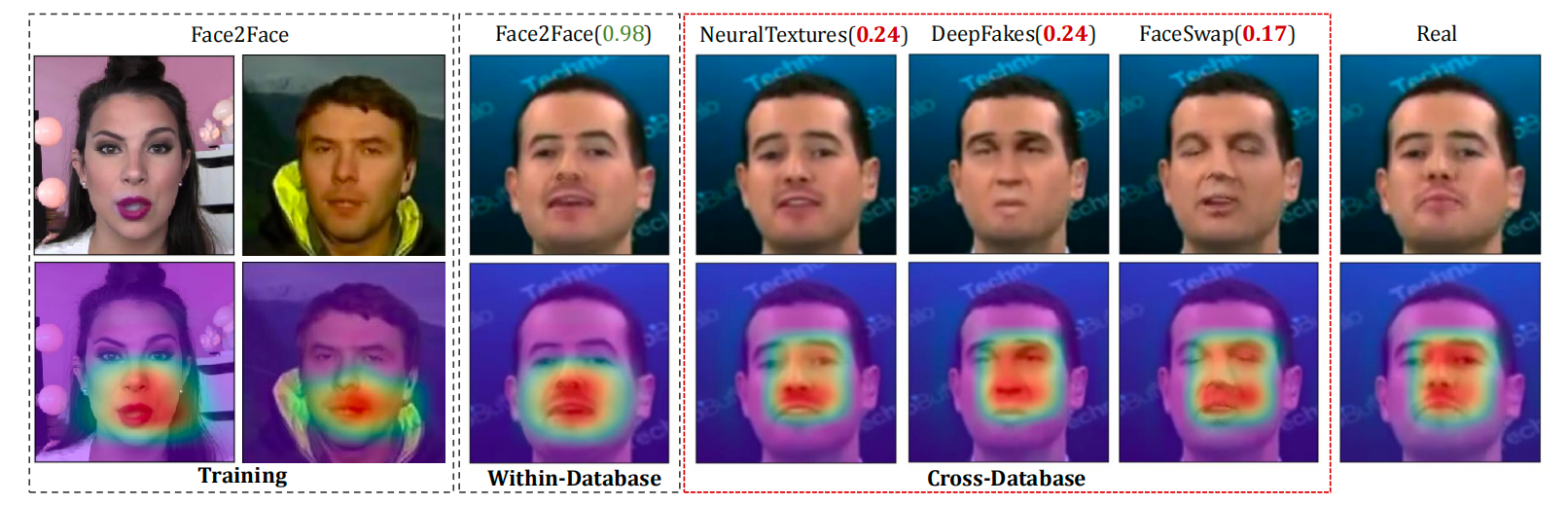


图1

此外，伪造图像的生成通常涉及复杂的处理过程，包括生成虚假面部图像以及将其与真实图像进行融合。这种处理会在伪造区域与背景之间引入不一致性，例如边界模糊或色彩失真。然而，现有方法并未充分利用这些跨区域的不一致性，导致检测能力不足。因此，需要设计一种既能够捕捉伪造纹理特征，又能够利用伪造区域与真实背景之间统计差异的检测方法，以提高模型的泛化能力。

## **解决方案**

### **3.1 多尺度高频特征提取模块**

高频噪声能够突出图像中的局部异常特征，例如伪造区域的边界或融合痕迹。为了充分利用这些特征，本文通过SRM高通滤波器从图像及其多层低级特征中提取高频信号。与传统方法只从输入图像中提取噪声不同，本文在多尺度特征图上应用高通滤波器，以捕捉更加丰富的高频信息。这些多尺度特征能够有效地突出伪造区域的边界。

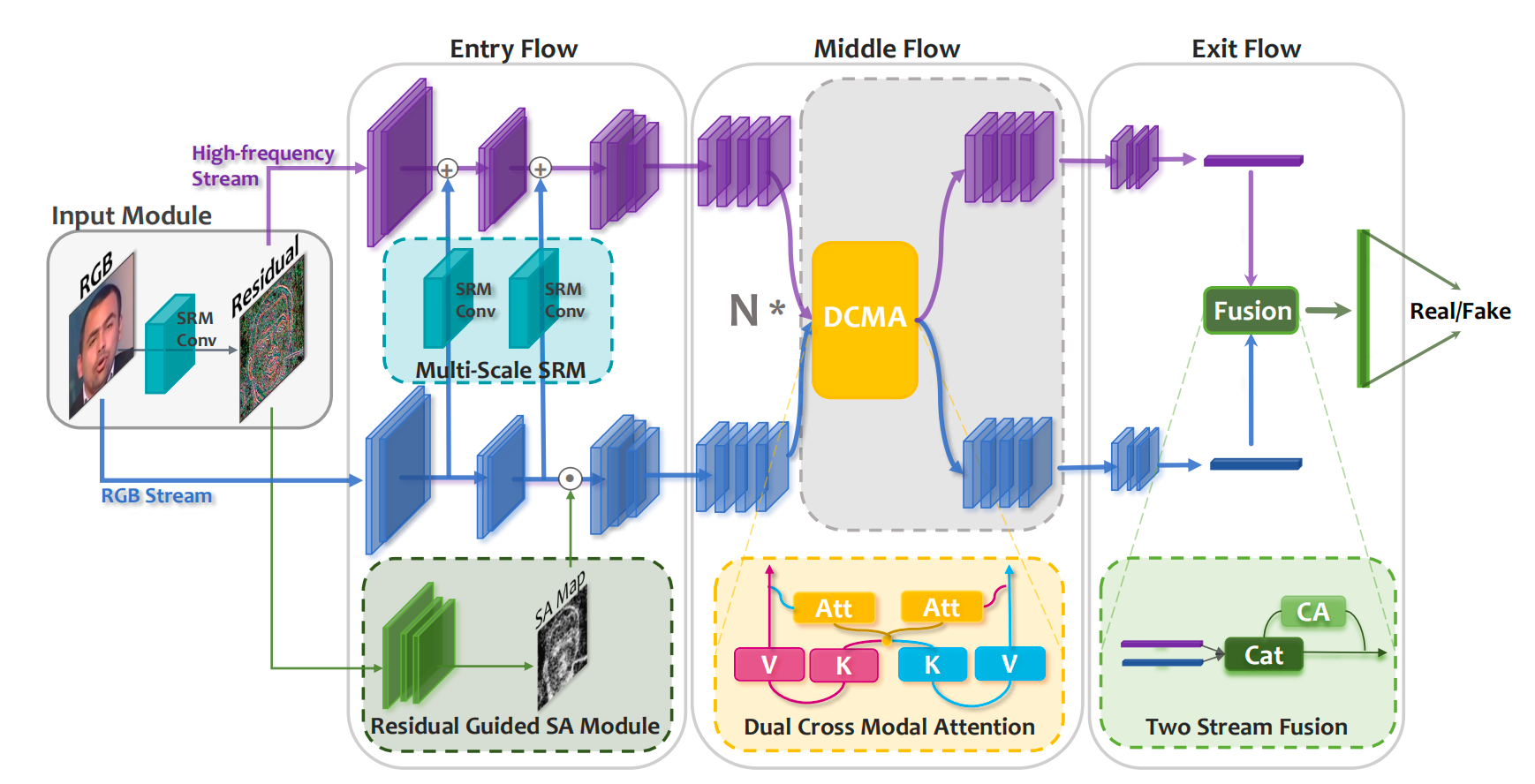


图2

### **3.2 残差引导空间注意模块**

为了更有效地关注伪造区域，本文设计了一种利用高频噪声生成空间注意力图的方法。这种注意力图能够从噪声视角引导模型，使其专注于伪造痕迹，同时抑制真实背景的干扰。如图3所示，注意力模块通过结合高频噪声生成注意力权重，在伪造边界区域显示出更强的响应。实验发现，这种基于高频特征的注意力机制能够显著提升伪造检测性能，尤其是在伪造区域与背景对比不明显的情况下。

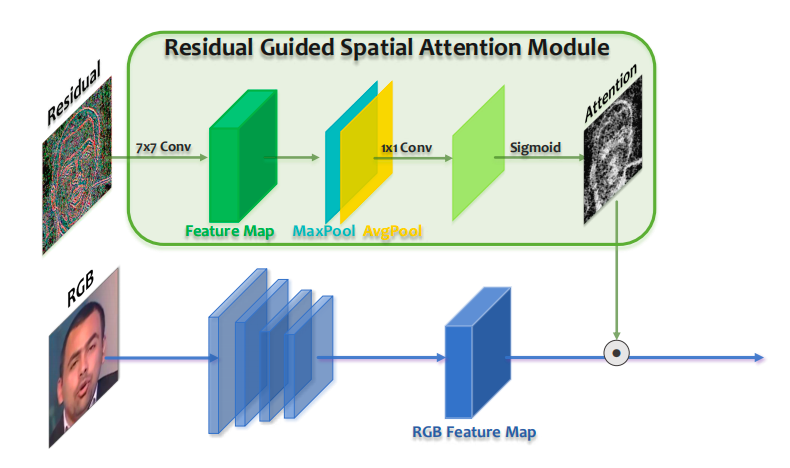


图3

### **3.3 跨模态双注意力模块**

伪造检测涉及低频纹理信息（如颜色分布）和高频噪声信息（如边界异常）的结合。本文提出的跨模态双注意力模块通过建立低频纹理和高频噪声之间的交互关系，实现信息的互补增强。如图4所示，该模块通过自注意力机制捕捉跨模态特征的相关性，同时利用交叉注意力机制将高频信息反向融合到低频特征中。这种设计不仅提升了每种特征的表示能力，还在两个模态之间建立了紧密的联系，使得模型能够在跨数据库场景中表现更加鲁棒。

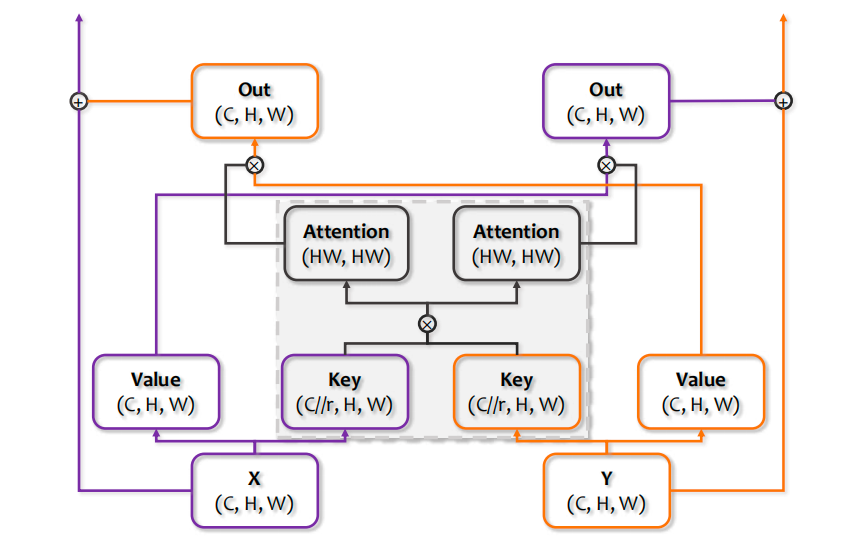


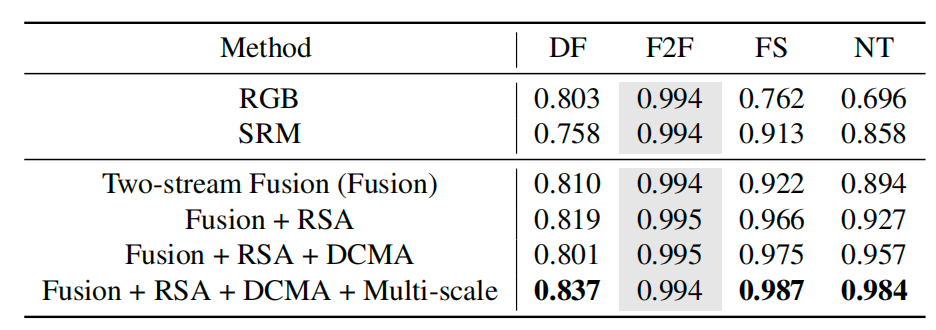
图4

## 实验分析

### **4.1 实验设置**

本文的方法在多个数据库上进行了广泛的实验，包括FaceForensics++（FF++）、DeepfakeDetection（DFD）、CelebDF、DFDC和DeeperForensics-1.0（DF1.0）。这些数据库涵盖了不同的伪造生成技术和多样化的场景，具有很高的评估价值。实验分为“数据库内评估”和“跨数据库评估”两个部分，全面考察了方法在不同场景下的性能。

### **4.2 模块有效性**

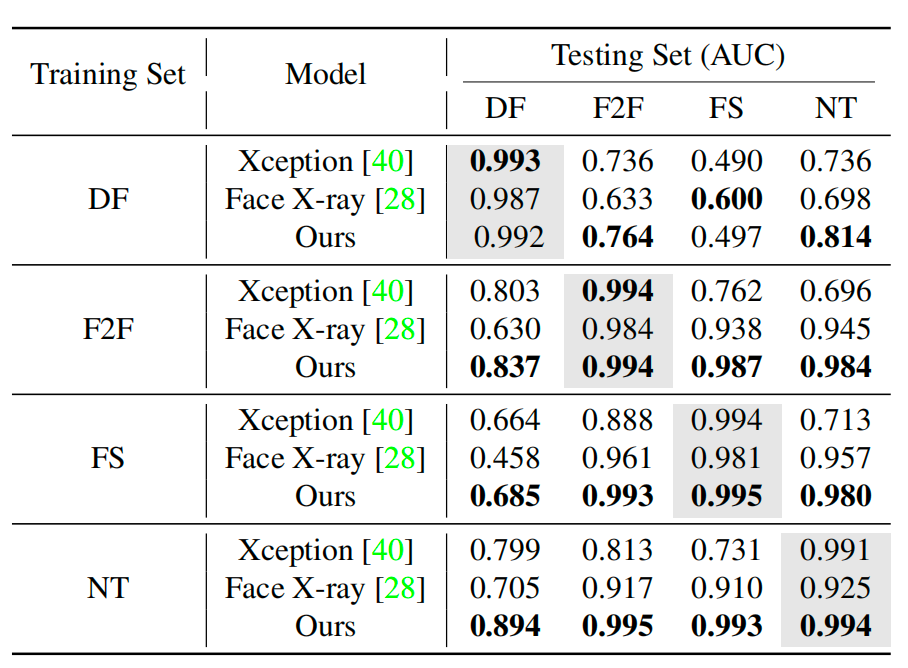


通过在FF++数据库上的消融实验，验证了每个模块对整体性能的贡献。

* 基础模型仅使用RGB特征时，虽然在数据库内检测（如F2F）上表现优异（AUC高达0.994），但在跨伪造技术（如NT和FS）时性能显著下降（AUC分别为0.696和0.762）。
* 使用高频特征（SRM）替换RGB后，模型的跨技术性能有所提升（例如FS的AUC从0.762提高到0.913）。
* 引入残差引导空间注意模块和跨模态双注意力模块后，模型性能进一步提升，跨技术场景中（如FS和NT）的AUC分别达到了0.987和0.984。

这表明每个模块均对提升模型的泛化能力有独立的贡献。

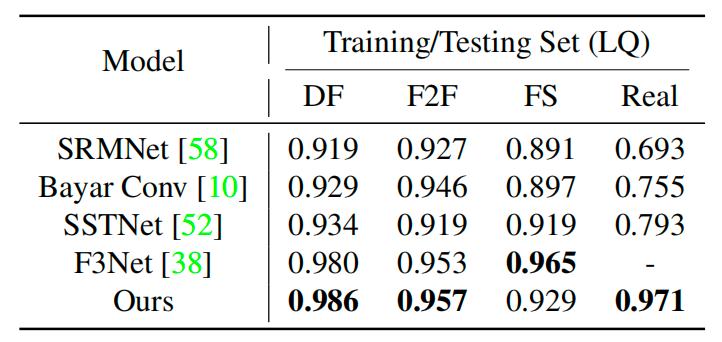
### **4.3 跨数据库性能**



跨数据库评估是衡量模型泛化能力的关键指标。

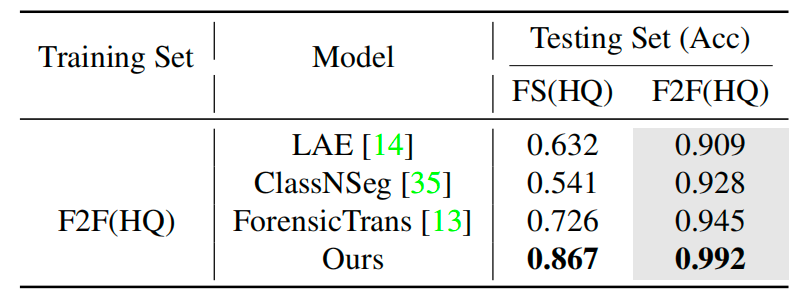
* 当在FF++中的F2F技术上训练并在其他伪造技术（如DF、FS、NT）上测试时，本文方法的性能显著优于对比方法（如Xception和Face X-ray）。例如，在NT上的AUC达到了0.984，而Xception和Face X-ray的AUC分别仅为0.696和0.945。
* 这种性能提升归功于本文方法有效结合了多模态特征，并利用了高频噪声揭示伪造区域的统计差异。

### **4.4 高频特征对比**



与其他高频特征方法（如F3Net、SRMNet等）相比，本文的方法在视频级伪造检测中表现出更高的鲁棒性。在FF++数据库（高压缩版本）上的测试显示，本文方法的检测精度在多个伪造技术上均优于SRMNet和Bayar Conv。例如，在Deepfakes上的精度达到了0.986，显著高于SRMNet的0.919。

### **4.5 多任务学习对比**



与采用多任务学习的伪造检测方法（如ForensicTrans）相比，本文的方法在跨数据库评估中的表现更为突出。例如，在训练于F2F并测试于FS时，本文方法的检测精度达到了0.867，而ForensicTrans仅为0.726。同时，本文的方法不依赖额外标注数据，具有更高的实际应用价值。

## **结论**

本文从现有伪造检测方法的泛化问题出发，提出了一种结合高频特征的检测方法。通过多尺度高频特征提取、残差引导注意力和跨模态交互等机制，本文的方法有效提升了检测模型的鲁棒性和泛化能力。在多个数据库上的实验结果表明，该方法不仅在检测精度上超越了主流方法，同时在跨数据库场景中展现出了显著的优势。这一研究为解决伪造检测中的泛化性难题提供了新思路，同时也为应对未来更加复杂的伪造技术奠定了基础。