*** 在 CVPR 2020 上接受 ***

观看你的上卷积:基于 CNN 的生成深度神经网络是无法重现光谱分布

理查德·杜拉尔1,3 , 玛格丽特·库珀2 , 贾尼斯·库珀1,4

1Competence Center High Performance Computing, Fraunhofer ITWM, Kaiserslautern, 德国
2 德国曼海姆大学数据和网络科学组
3 IWR,海德堡大学,德国
4 德国奥芬堡大学机器学习与分析研究所

抽象的

生成式卷积深度神经网络,例如流行的 GAN 架构,依赖于基于卷积的上采样方法来生成非标量输出,如图像或视频序列。在本文中,我们展示了常见的上采样方法,即上卷积或转置卷积,导致此类模型无法正确再现自然训练数据的光谱分布。这种效果独立于底层架构,我们证明它可用于轻松检测生成的数据,如深度造假,在公共基准测试中准确率高达 100%。为了克服当前生成模型的这一缺点,我们建议在训练优化目标中添加一个新的频谱正则化项。我们表明,这种方法不仅允许训练避免高频错误的光谱一致的 GAN。此外,我们还表明,频谱的正确近似对生成网络的训练稳定性和输出质量有积极影响。

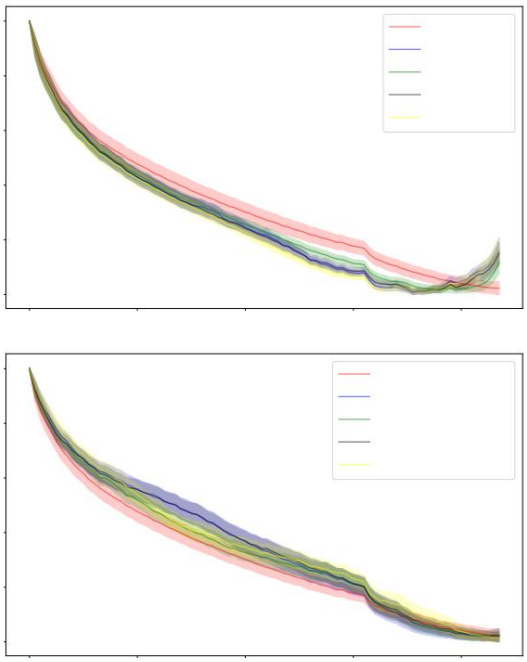


图 1:常见的上卷积方法正在将严重的光谱失真引入生成的图像中。上图显示了真实图像和 GAN 生成图像的功率谱（参见第 2.1 节)方位角积分后的统计数据（均值和方差）。对 CelebA [34] 数据集的评估,这里所有的 GAN（DCGAN [47]、DRA GAN [32]、LSGAN [37]、WGAN-GP [20]）都使用“转置卷积”（见第 2.2 节)上采样。

底部:与上述相同实验的结果,在 GAN 训练期间添加了我们提出的光谱损失。

一、简介

生成卷积深度神经网络最近被广泛用于计算机视觉任务:生成逼真的图像 [29、6]、图像到图像 [45、26、61、9、42、30] 和文本图像转换 [48、11、58、59]、风格迁移 [27、60、61、25]、图像修复 [45、54、33、26、56]、迁移学习 [5、10、15] 甚至用于训练语义分割任务 [35、53],仅举几例。

最突出的生成神经网络架构是生成对抗网络 (GAN) [18] 和变分自动编码器 (VAE) [46]。两种基本方法都试图近似联合国的潜在空间模型

来自训练数据样本的 derlying（图像)分布。
给定这样一个潜在空间模型,我们可以画出新的（arti

官方)在各个维度上采样和操作它们的语义属性。虽然 GAN 和 VAE 方法都以许多不同的变体形式发表,例如具有不同的损失函数 [18、4、20]、不同的潜在空间约束 [41、13、13、21、30] 或各种深度神经网络工作（DNN) 生成器网络的拓扑 [47, 43],所有这些方法都必须遵循基本的数据生成原则:它们必须将样本从低维（通常是 1D)和低分辨率潜在空间转换为高分辨率（2D图像)输出空间。因此,这些生成神经网络必须提供某种（可学习的)放大属性。

虽然所有这些生成方法都通过优化某些损失函数来引导模型参数的学习,但最常用的损失只关注输出图像空间的属性,例如使用卷积神经网络 (CNN) 作为判别器网络适用于图像生成 GAN 中的隐式损失。

这种方法已被证明足以生成视觉声音输出,并且能够在某种程度上捕获图像空间中的数据（图像)分布。

然而,众所周知,放大操作会改变信号的频谱特性 [28],导致输出中出现高频失真。

在本文中,我们研究了生成器网络中常用的上采样技术的影响。图 1 的上图说明了我们最初实验的结果,支持我们的工作假设,即当前的生成网络无法再现光谱分布。图 1 还表明,这种影响与实际发电机网络无关。

1.1.相关工作

1.1.1 Deepfake检测

我们展示了我们的发现对 Deepfake 检测任务的实际影响。术语 deepfake [22, 8] 描述了最近的现象,即人们通过深度生成神经网络 [7] 滥用人工面部生成的进步来制作名人和政客的虚假图像内容。

由于此类造假的潜在社会影响,深度造假检测已成为其自身的重要研究课题。文献中报道的大多数方法,如 [38、3、57],本身都依赖于 CNN,因此需要大量带注释的训练数据。同样,[24] 引入了具有对比损失函数的深度伪造鉴别器,并且 [19] 通过在 CNN 之上采用递归神经网络 (RNN) 来合并时域信息。

1.1.2 GAN 稳定化

正则化 GAN 以促进更稳定的训练并避免模式崩溃最近引起了一些关注。而 [40] 通过展开来稳定 GAN 训练为了鉴别器的优化,[50] 提出了通过噪声进行正则化以及一种有效的基于梯度的方法。最近在 [16] 中提出了一种基于八度卷积的稳定 GAN 训练。这些方法都没有考虑正则化的频谱。然而,最近,[17] 中提出了带宽受限的 CNN,用于使用压缩模型进行图像分类。在 [55] 中,首次观察暗示了功率谱对模型鲁棒性的重要性,再次用于图像分类。相比之下,我们建议利用对 GAN 生成的频谱的观察来训练稳定。

1.2.投稿

我们工作的贡献可以总结如下:

- 我们通过实验证明了当前生成神经网络架构无法正确近似训练数据的光谱分布。
- 我们利用这些光谱失真为生成的图像和视频提出了一个非常简单但高度准确的检测器,即在公共基准上达到 100% 准确度的 DeepFake 检测器。
- 我们的理论分析和进一步的实验表明,常用的上采样单元,即上卷积,正在引起观察到的效果。
- 我们提出了一种新颖的光谱正则化项,它能够补偿光谱失真。
- 我们还通过实验表明,在 GAN 训练中使用光谱正则化会导致更稳定的模型并提高视觉输出质量。

本文的其余部分组织如下:第 2 节介绍了常见的放大方法并分析了它们对图像光谱特性的负面影响。

在第 3 节中,我们介绍了一种新的光谱损失,它允许训练能够补偿放大误差并生成正确光谱分布的生成网络。

我们在第 4 节中使用公共基准上的当前架构评估我们的方法。

2. 上卷积的光谱效应

2.1.在 DFT 功率谱上使用方位角积分分析图像的光谱分布

为了分析对光谱分布的影响,我们依赖于简单但具有特征的一维表示

傅立叶功率谱。我们从大小为 M×N 的 2D（图像)数据 I 的离散傅里叶变换 F 计算该光谱表示，

$$F(I)(k,) = \sum_{米=0}^{M-1} \sum_{n=0}^{N-1} 和^{-2\pi i \cdot jk} 我^{-2\pi i \cdot j} \overline{N} \cdot I(m, n), \quad (1)$$

对于 k = 0, . . . , M − 1, = 0, . . . , N - 1,

通过径向频率 ϕ 上的方位角积分

$$AI(\omega k) = \int_0^{2页} F(I)(\omega k \cdot \cos(\phi), \omega k \cdot \sin(\phi))\, 2d\phi$$

对于 k = 0, . . . , 米/2 - 1 , (2)

假设正方形 images1 。图 2 给出了该处理步骤的示意图。

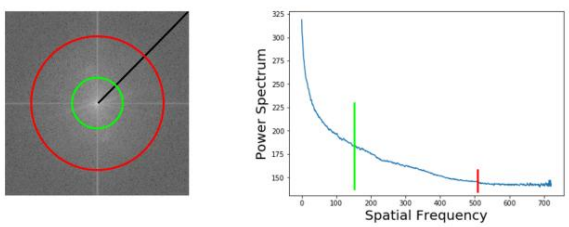


图 2:方位角积分 (AI) 示例。 （左)图像的二维功率谱。 （右)一维功率谱:每个频率分量都是二维频谱的径向积分（红色和绿色示例）。

2.2.生成 DNN 中的向上卷积

像 GAN 这样的生成神经架构从非常低维的潜在空间产生高维输出，例如图像。因此,所有这些方法都需要在通过网络传播数据时使用某种放大机制。文献和流行的实现框架（如 TensorFlow [2] 和 PyTorch [44])中最常用的两种放大技术如图 3 所示:通过插值进行的向上卷积 (up+conv) 和转置卷积(transconv)) 。

我们使用一个非常简单的自动编码器 (AE) 设置（参见图 4)来初步研究上卷积单元对上采样后二维图像的光谱特性的影响。图 5 显示了两种方法对频谱的不同但巨大的影响。图 6 给出了重建图像的定性结果,并表明频谱中的错误与视觉外观有关。

1→ M = N。我们知道这种表示法是滥用的,因为 F(I) 是离散的。然而，完全正确的离散符号只会使我们工作的一个方面过于复杂。 https://github.com/cc-hpc-itwm/UpConv 上提供了 AI 的离散实现。

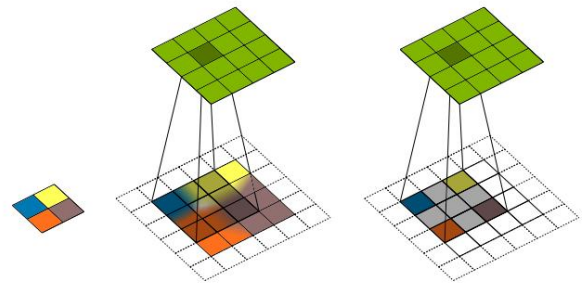


图 3:两个最常见的上卷积单元的示意图概述。左图:低分辨率输入图像（此处为 2×2） ;中心:通过插值进行上卷积(up+conv) - 输入通过插值（双线性或最近邻)进行缩放,然后与标准可学习滤波器内核（大小为 3 × 3)进行卷积以形成 5x5 输出（绿色） ,右:转置卷积（transconv） 输入用“钉床”方案填充（灰色网格点为零） ,然后与标准滤波器核卷积形成 5 × 5 输出（绿色）。

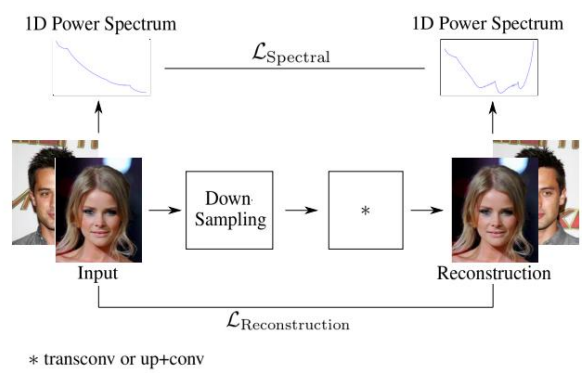


图 4:用于演示图 5 中向上卷积效果的简单自动编码器 (AE) 设置的示意图,仅使用标准 MSE 重建损失（底部)在真实图像上训练 AE。

我们将输入缩小两倍,然后使用不同的上卷积方法重建原始图像大小。在第 3 节中,我们使用额外的光谱损失（顶部)来补偿光谱失真（见图 7）

.

2.3.理论分析

对于理论分析,我们在不失一般性的情况下考虑一维信号 a 及其离散傅立叶变换 a^ 的情况

$$一个 k = \sum_{j=0}^{N-1} 和^{-2\pi i \cdot jk} \overline{N} \cdot 还有, \quad 对于 k = 0, . . . , N − 1. \quad (3)$$

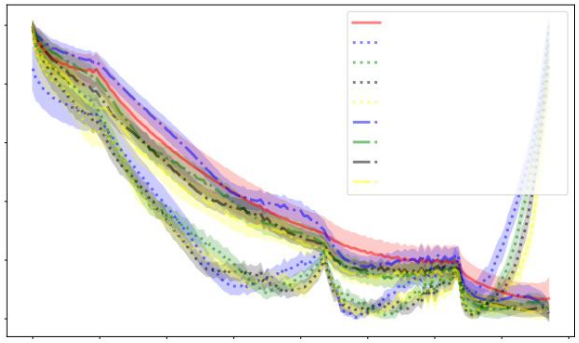


图 5:单个向上卷积单元（设置参见图 4)对输出图像的频谱（方位角积分)的影响。两种上卷积方法对输出的光谱分布都有巨大影响。转置卷积增加了大量的高频噪声,而基于插值的方法（up+conv)则缺乏高频。

图 6:在我们简单的 AE 设置中,光谱失真对图像输出的影响。左:原始图像;中心： AE 输出图像;右:过滤后的差异图像。

第一行显示了在（up+conv)情况下丢失高频的模糊效果;底行显示由（transconv)引起的高频伪影。

如果我们想将 a 的空间分辨率增加 2 倍,我们得到

$$\hat{a}^{\uparrow}_{k} = \sum_{j=0}^{2\cdot N-1} b_j \, e^{-2\pi i \frac{jk}{2\cdot N}} \cdot a_{\lfloor j/2 \rfloor} \tag{4}$$

$$= \sum_{j=0}^{N-1} e^{-2\pi i \cdot \frac{2\cdot jk}{2\cdot N}} \cdot a_j + \sum_{j=0}^{N-1} e^{-2\pi i \cdot \frac{2\cdot (j+1)k}{2\cdot N}} \cdot b_j , \tag{5}$$

对于 $k = 0, \ldots, 2N-1$。

其中$b_j = 0$ 用于“钉床”插值（由$a_{j-1}+a_j$ transconv 使用)和b_j = 用于双线性插值（由2用于 up+conv）。

让我们首先考虑$b_j = 0$ 的情况,即“钉床”插值。在那里,方程式中的第二项。 (6) 为零。第一项类似于原始的傅里叶变换,但参数 k 被替换为 k。因此，将空间分辨率增加 2 倍会导致频率轴缩放 2 倍。现在让我们从基于采样理论的角度考虑效果。这是

$$\hat{a}^{\uparrow}_{k} = \sum_{j=0}^{2\cdot N-1} j^{-2\pi i \frac{jk}{2\cdot N}} \cdot a_{\lfloor j/2 \rfloor} \tag{6}$$

$$= \sum_{j=0}^{2\cdot N-1} e^{-2\pi i \frac{jk}{2\cdot N}} \cdot \sum_{t=-\infty}^{\infty} \delta(j - 2t) \tag{7}$$

因为与 Dirac 脉冲梳的逐点乘法仅删除 a = 0 的值。假设一个周期信号并应用卷积定理 [31],我们得到

$$(7) = 2 \cdot \sum_{t=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} e^{-2\pi i \frac{jk}{2\cdot N}} \cdot a_j \, \delta(k-2t) , \tag{8}$$

等于

$$\sum_{t=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} e^{-2\pi i \cdot \frac{jk}{N}} \cdot a_j \, \delta(k-2t) \tag{9}$$

通过方程式。 (6).因此，“钉床上采样”将创建信号的高频副本。要去除这些频率副本,需要对上采样信号进行适当的平滑处理。超出的所有观察到的空间频率都是潜在的上采样伪影。虽然从理论的角度来看这是显而易见的,但我们也在图 8 中实际证明了使用常用的 3 × 3 卷积滤波器不可能校正如此大的频带（假设中高分辨率图像）.

在双线性插值的情况下,我们在等式中有b_j =。 (6),这对应于与b_j相邻的 $a_{\lfloor j/2 \rfloor}$ 和$a_{\lfloor j/2 \rfloor+1}$ 的值的平均过滤。

这是等效频谱 a ^到定理的逐点乘法,它抑制了人为的高频。然而,预钟由此产生的频谱在低和卷积会过低。

3. 学习生成正确的光谱分布

我们在上一节中的发现的实验评估及其在检测生成的con-

帐篷（见第 4.1 节），提出一个问题，是否有可能纠正由生成网络中使用的上卷积单元引起的光谱失真。毕竟，通常的网络拓扑结构包含可学习的卷积过滤器，这些过滤器跟在上卷积之后，并有可能纠正此类错误。

3.1.光谱正则化

由于常见的生成网络架构大多专门使用基于图像空间的损失函数，因此无法直接捕获和校正光谱失真。因此，我们建议在生成器损失中添加一个额外的频谱项：

$$L_{final} = L_{Generator} + \lambda \cdot L_{Spectral}, \tag{10}$$

其中 λ 是加权光谱损失影响的超参数。由于我们已经使用方位角积分 AI 测量光谱失真（参见方程式 (2)），并且 AI 是可微分的，因此 LSpectral 的一个简单选择是生成的输出AIout与从真实样本获得的平均AIreal之间的二元交叉熵：

$$L_{Spectral} := -\frac{1个}{(米/2 - 1)} \sum_{我=0}^{米/2-1} AI_{real} \cdot \log(AI_{out}) + (1 - AI_{real}) \cdot \log(1 - AI_{out}) \tag{11}$$

请注意，M 是图像大小，我们使用归一化系数 (AI0)将 0 的值缩放到 [0, 日 1] 的方位角积分。

图 7 显示了将我们的光谱损失添加到第 2.2 节中不同 λ 值的 AE 设置的效果。 2.3，观察到的效果不能通过单个学习的 3 × 3 滤波器来校正，即使对于大值 λ 也是如此。因此，我们需要重新考虑架构参数。

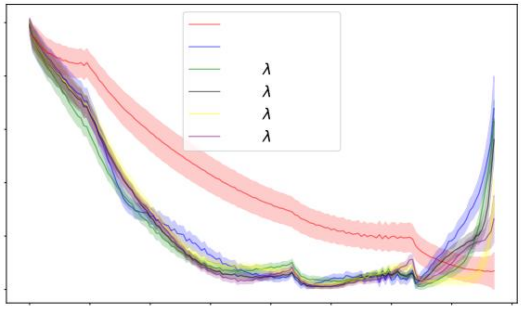


图 7：光谱损失为 λ 的自动编码器 (AE) 结果。
即使光谱损失具有很高的权重，也无法使用单个 3×3 卷积层来校正光谱失真。这一结果与第 2.3 节的发现一致。

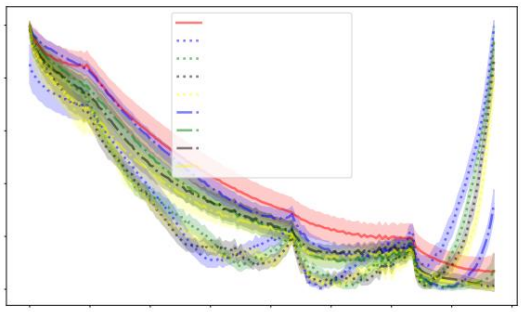


图 8：上采样步骤后卷积滤波器大小导致光谱损失的 AE 结果。结果在很大程度上取决于所选的过滤器大小和卷积层的数量。借助三个可用的 5 × 5 卷积滤波器，AE 可以使用建议的光谱损失大大减少光谱失真。

3.2.向上卷积的过滤器尺寸

在图 8 中，我们评估了第 2.2 节中关于滤波器大小和卷积层数量的 AE 的光谱损失。我们考虑从 3 × 3 到 11 × 11 和 1 或 3 个卷积层的不同解码器滤波器大小。虽然不能通过单个甚至三个 3×3 卷积消除上采样产生的频谱失真，但可以通过学习更多、更大的滤波器时提出的损失来纠正它。

4. 实验评价

我们在三个不同的实验中评估了前面部分的结果，在公共人脸生成数据集上使用了著名的 GAN 架构。第 4.1 节显示，普通人脸生成网络产生的输出为可用于检测人造或“假”图像的强烈光谱失真。在第 4.2 节中，我们表明我们的频谱损失足以补偿相同数据频域中的伪影。最后，我们在第 4.3 节中凭经验表明，谱正则化对 GAN 的训练稳定性也有积极影响。

4.1. Deepfake检测

在本节中，我们展示了由最先进的 GAN 中的上卷积引起的光谱失真可用于轻松识别“假”图像数据。仅使用少量带注释的训练数据，甚至是未经监督的设置，我们就能够以近乎完美的准确度从公共基准中检测生成的人脸。

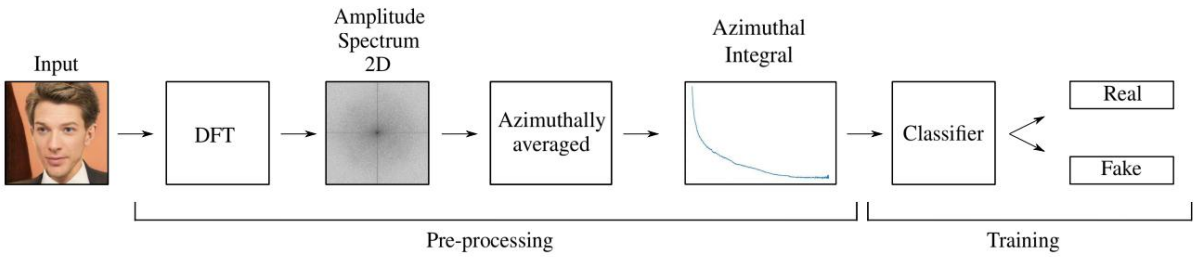


图 9:我们方法的处理管道概述。它包含两个主要块,一个使用 DFT 的特征提取块和一个训练块,其中分类器使用新的转换后的特征来确定人脸是否真实。

请注意,输入图像在 DFT 之前被转换为灰度。

4.1.1 基准

我们在三个不同的面部图像数据集上评估我们的方法,提供具有不同空间分辨率的注释数据:

- FaceForensics++ [49] 包含一个 DeepFake 检测数据集,其中包含 16 个不同场景中 28 个付费演员的 363 个原始视频序列,以及超过 3000 个面部操作视频及其相应的二进制掩码。所有视频都包含一个可跟踪的,主要是没有遮挡的正面,这使得自动篡改方法能够生成逼真的伪造。提取的人脸图像的分辨率各不相同,但通常约为 80 × 80 × 3 像素。

- CelebFaces Attributes (CelebA) 数据集 [34] 包含 202,599 张名人面部图像,面部属性有 40 种变化。人脸图像的尺寸为 178 × 218 × 3,在我们的上下文中可以认为是中等分辨率。

- 为了评估高分辨率 1024 × 1024 × 3 图像,我们提供了新的 Faces-HQ 数据集,它是来自 CelebA-HQ [29]、Flickr Faces-HQ 数据集[2][30] 40k 公开可用图像的注释集合]、100K Faces 项目 [1] 和 www.thispersondoesnotexist.com。

4.1.2 方法

图 9 说明了我们的简单处理管道,通过方位角积分从样本中提取光谱特征（见图 2）,然后使用基本 SVM [51] 分类器3 进行监督和 K-Means [36] 进行无监督假检测。对于每个实验,我们随机选择不同大小的训练集,并使用剩余的数据进行测试。为了处理不同大小的输入图像,我们通过 0 将一维功率谱归一化,将得到的一维特征向量缩系数和放到固定大小。

4.1.3 结果

图 15 显示真实的和“假的”面孔在我们的光谱特征空间的高频范围内形成轮廓分明的簇。表 3 中的实验结果证实,由上采样单元引起的功率谱失真是一个常见问题,可以轻松检测生成的内容。这个简单的指标甚至优于使用大型注释训练集的基于 DNN 的复杂检测方法4

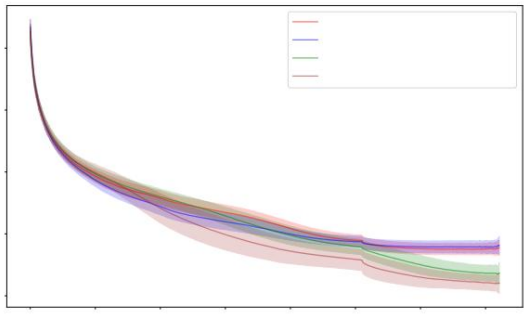


图 10:来自每个 Faces-HQ 子数据集的 1000 个样本的 AI（一维功率谱)统计数据（均值和方差）。
显然,真实和“假”图像可以通过它们的 AI 表示来区分。

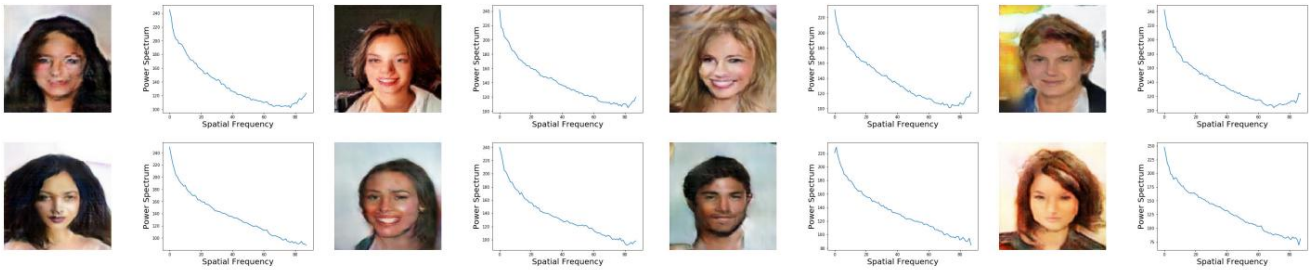
4.2.应用光谱正则化

在本节中,我们评估正则化方法在 CelebA 基准上的有效性,就像之前的实验一样。基于我们的理论分析（见第 2.3 节)和第 3 节中的第一个 AE 实验,我们扩展

2Faces-HQ 数据有一个 https://cutt.ly/6enDLYG。另请参阅 [14]。大小为 19GB。下载:

3SVM的超参数可以在源码中找到

4 注:[57] 报告的所有其他方法的结果。方法的直接比较可能有偏差,因为 [57] 使用相同的真实数据,但使用不同的 GAN 独立生成假数据。



(一)DCGAN。 (b) 德拉甘。 (c) LSGAN。 (d) WGAN。

图 11:来自不同类型 GAN 的样本及其一维功率谱。顶行:由标准拓扑生成的样本。底行:由标准拓扑和我们的光谱正则化技术产生的样本。

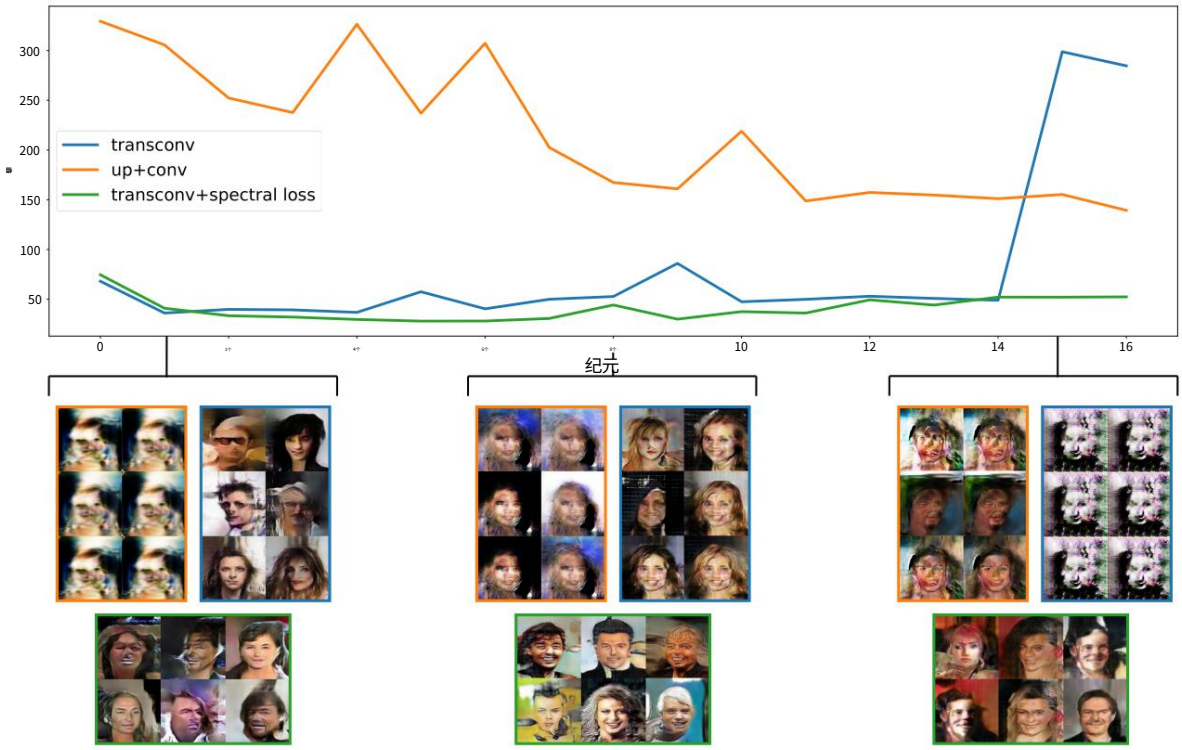


图 12:在整个训练过程中,CelebA 上 DCGAN 基线的 FID 值与 GAN 输出之间的相关性。
低 FID 分数对应于多样化但视觉上健全的面部图像输出。高 FID 分数表示输出质量差和“模式崩溃”场景,其中所有生成的图像都绑定到原始分布的非常狭窄的子空间。

现有的 GAN 架构有两种方式:首先,我们在生成器损失中添加一个频谱损失项（见等式（11））。我们使用来自数据集的 1000 个未注释的真实样本来估计AIreal,这是计算光谱损失所需的（见等式（11））。其次,我们将最后一个上卷积单元之后的卷积层更改为三个内核大小为 5 × 5 的滤波器层。图 1 的底部图表显示了该实验与原始 GAN 架构的直接比较结果。图 11 给出了在没有和使用我们建议的正则化的情况下产生的几个定性结果。

4.3.频谱正则化的积极影响

通过对频谱进行正则化,我们获得了生成合成图像的直接好处,这些合成图像不仅看起来逼真,而且还模仿了频域中的行为。

这样,我们离真实分布的样本图像又近了一步。此外,这种正则化还有一个有趣的副作用。在我们的实验中,我们注意到具有光谱损失项的 GAN 在避免“模式崩溃”[18] 和更好的收敛方面似乎更加稳定。众所周知,GAN

数据集	80%（训练）- 20%（测试)方			
	法 # 样本受监督 不受监督 1000 100 20 2000			
Faces-总部	我们的		100%	82%
Faces-总部	我们的		100%	81%
Faces-总部	我们的		100%	75%
名人A	我们的		100%	96%
名人 [57]		100000	99.43%	-
名人 [39]		100000	86.61%	-
FaceForensics++ oursA		2000	85%	-
FaceForensics++ oursB		2000	90%	-

表 1:测试精度。我们的方法在不同的数据设置下使用 SVM（监督)和 k-means（无监督）。
A) 在单帧上进行评估。 B) 通过单帧检测的多数表决对完整视频序列的准确性。

可能会受到具有挑战性和不稳定的训练过程的影响,并且几乎没有理论可以解释这种现象。这使得尝试新的生成器变体或将它们用于新领域变得极其困难,这极大地限制了它们的适用性。

为了研究光谱正则化对 GAN 训练的影响,我们进行了一系列实验。通过使用一组不同的基线架构,我们评估了光谱正则化的稳定性,提供了 CelebA 数据集的定量结果。我们的评估指标是 Frechet Inception Distance (FID) [23],它使用在 ImageNet [12] 上预训练的 Inception-v3 [52] 网络从中间层提取特征。

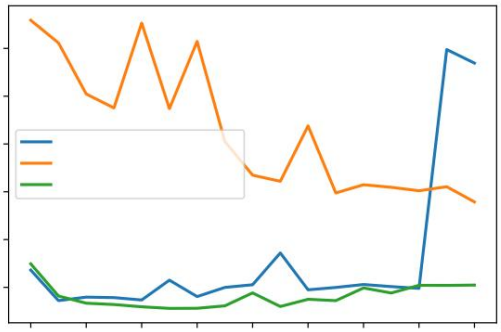


图 13:在有和没有光谱损失（此处 λ = 2)的情况下,DC GAN 基线在训练时间内的 FID（越低越好）。
虽然 DCGAN 的 up+conv 变体未能改善,但 transconv 版本中训练时间的 FID 分数正在收敛但不稳定。只有我们的光谱损失变体能够获得低而稳定的 FID 分数。

图 13 和 14 显示了 FID 在训练时期的演变,使用具有不同上卷积单元的基线 GAN 实现和具有光谱损失的相应版本。这些结果显示出明显的积极

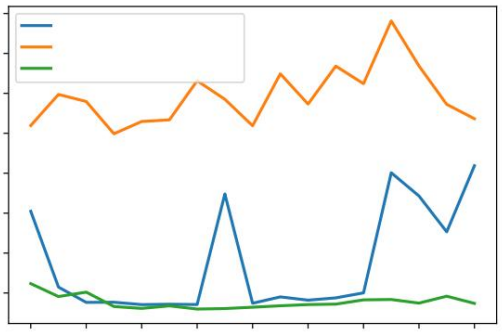


图 14:在有和没有光谱损失的情况下,LS GAN 基线在训练时间内的 FID（越低越好）(此处 λ = 0.5）。至于 DCGANS,是 LSGAN 的 up+conv 变体未能在训练时间内提高 FID 分数。 transconv 版本正在收敛但不稳定。同样,只有我们的光谱损失变体能够获得低而稳定的 FID 分数。

在 FID 度量方面的效果,其中频谱正则化 k.pdf 在整个训练过程中具有稳定且低的 FID,而未正则化的 GAN 往往会“崩溃”。图 12 可视化了高 FID 值与失败的 GAN 图像生成之间的相关性。

五、讨论与结论

我们展示了常见的“最先进”卷积生成网络,如流行的 GAN 图像生成器,无法近似真实数据的光谱分布。

这一发现具有很强的实际意义:这不仅可以用来轻松识别生成的样本,还意味着所有训练数据生成或迁移学习的方法都存在根本性缺陷,不能指望当前的方法能够近似真实数据分布正确。然而,我们表明有一些简单的方法可以解决这个问题:通过将我们提出的频谱正则化添加到生成器损失函数并将最终生成器卷积的滤波器大小增加到至少 5 × 5,我们能够补偿频谱误差.在实验上,我们发现了强烈的迹象表明谱正则化对 GAN 的训练稳定性有非常积极的影响。虽然这种现象需要进一步的理论研究,但从直觉上讲这是有道理的,因为众所周知,高频噪声会对基于 CNN 的鉴别器网络产生强烈影响,这可能会导致生成器过度拟合。

可用的源代码： https ://
github.com/cc-hpc-itwm/UpConv

参考

[1] 生成了 100,000 张面孔。 https://生成。相片/。

[2] M. Abadi、A. Agarwal、P. Barham、E. Brevdo、Z. Chen、C. Citro、GS Corrado、A. Davis、J. Dean、M. Devin、S. Ghe mawat、I. Goodfellow , A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mane, ´ R. Monga, S. Moore, D. Murray, C. Olah、M. Schuster、J. Shlens、B. Steiner、I. Sutskever、K. Talwar、P. Tucker、V. Vanhoucke、V. Vasudevan、F. Viegas、O. Vinyals、P. Warden、 M. Wattenberg、M. Wicke、Y. Yu 和 X. Zheng。 Tensor Flow:异构系统上的大规模机器学习,2015 年。软件可从 tensorflow.org 获得。

[3] D. Afchar、V. Nozick、J. Yamagishi 和 I. Echizen。 Mesonet:一个紧凑的面部视频伪造检测网络。 2018 年 IEEE 国际信息取证与安全研讨会 (WIFS),第 1-7 页。 IEEE,2018 年。

[4] M. Arjovsky、S. Chintala 和 L. Bottou。 Wasserstein 甘。 arXiv 预印本 arXiv:1701.07875, 2017。

[5] S. Bartunov 和 D. Vetrov。带有生成匹配网络的小样本生成建模。在人工智能和统计国际会议上,第 670-678 页,2018 年。

[6] A. Brock、J. Donahue 和 K. Simonyan。用于高保真自然图像合成的大规模 gan 训练。 arXiv 预印本 arXiv:1809.11096, 2018。

[7] M. Brundage、S. Avin、J. Clark、H. Toner、P. Eckersley、B. Garfinkel、A. Dafoe、P. Scharre、T. Zeitzoff、B. Filar 等。 人工智能的恶意使用:预测、预防和缓解。 arXiv 预印本 arXiv:1802.07228, 2018。

[8] R. Chesney 和 D. Citron。 Deepfakes 和新的虚假信息战争:即将到来的后真相地缘政治时代。 For eign Aff., 98:147, 2019.

[9] Y. Choi, M. Choi, M. Kim, J.-W. Ha、S. Kim 和 J. Choo。 Stargan:用于多域图像到图像翻译的统一生成对抗网络。在 IEEE 计算机视觉和模式识别会议记录中,第 8789-8797 页,2018 年。

[10] L. Clouatre 和 M. Demers。 Figr:使用爬行动物生成少量图像。 arXiv 预印本 arXiv:1901.02199, 2019。

[11] B. Dai、S. Fidler、R. Urtasun 和 D. Lin。通过条件 gan 实现多样化和自然的图像描述。在 IEEE 计算机视觉国际会议论文集中,第 2970-2979 页,2017 年。

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei Fei. Imagenet: A large-scale hierarchical image database. 2009 年 IEEE 计算机视觉和模式识别会议,第 248-255 页。 IEEE,2009 年。

[13] J. Donahue、P. Krahenb ¨ uhl 和 T. Darrell。对抗特征 真正的学习。 arXiv 预印本 arXiv:1605.09782, 2016。

[14] R. Durall、M. Keuper、F.-J。 Pfreundt 和 J. Keuper。使用简单的功能取消掩盖 deepfakes。 arXiv 预印本 arXiv:1911.00686, 2019。

[15] R. Durall, F.-J. Pfreundt 和 J. Keuper。致敬翻译半少镜头。 arXiv 预印本 arXiv:1910.03240, 2019。

[16] R. Durall, F.-J. Pfreundt 和 J. Keuper。用八度卷积稳定甘斯。 arXiv 预印本 arXiv:1905.12534, 2019。

[17] A. Dziedzic、J. Paparrizos、S. Krishnan、A. Elmore 和 M. Franklin。卷积神经网络的带限训练和推理。在 K. Chaudhuri 和 R. Salakhut dinov,编辑,第 36 届机器学习国际会议论文集,机器学习研究论文集第 97 卷,第 1745-1754 页,美国加利福尼亚州长滩,2019 年 6 月 9-15 日。PMLR。

[18] I. Goodfellow、J. Pouget-Abadie、M. Mirza、B. Xu、D. Warde-Farley、S. Ozair、A. Courville 和 Y. Bengio。生成对抗网络。在神经信息处理系统的进展中,第 2672–2680 页,2014 年。

[19] D. Guera 和 EJ Delp。使用递归神经网络的 Deepfake 视频检测。 2018年第15届IEEE国际 高级视频和基于信号的监视 (AVSS) 会议,第 1-6 页。 IEEE,2018 年。

[20] I. Gulrajani、F. Ahmed、M. Arjovsky、V. Dumoulin 和 AC Courville。改进了 wasserstein gans 的训练。在神经信息处理系统的进展中,第 5767-5777 页,2017 年。

[21] S. Gurumurthy、R. Kiran Sarvadevabhatla 和 R. Venkatesh Babu。 Deligan:用于多样化和有限数据的生成对抗网络。在 IEEE 计算机视觉和模式识别会议记录中,第 166–174 页,2017 年。

[22] D.哈里斯。 Deepfakes:虚假色情就在这里,法律无法保护你。 Duke L. & Tech。牧师,2018 年 17:99。

[23] M. Heusel、H. Ramsauer、T. Unterthiner、B. Nessler 和 S. Hochreiter。通过两个时间尺度更新规则训练的 Gans 收敛于局部纳什均衡。在神经信息处理系统的进展中,第 6626–6637 页,2017 年。

[24] C.-C。许,C.-Y。李和 Y.-X。庄。学习在野外检测假人脸图像。 2018 年计算机、消费者和控制国际研讨会 (IS3C),第 388-391 页。 IEEE,2018 年。

[25] X. Huang, M.-Y.刘、S. Belongie 和 J. Kautz。多模态无监督图像到图像的翻译。在欧洲计算机视觉会议 (ECCV) 会议记录中,第 172–189 页,2018 年。

[26] S. Iizuka、E. Simo-Serra 和 H. Ishikawa。全局和局部一致的图像补全。 ACM 图形交易 (ToG),36(4):107,2017 年。

[27] P.伊索拉,J.-Y。 Zhu、T. Zhou 和 AA Efros。使用条件对抗网络进行图像到图像的翻译。在 IEEE 计算机视觉和模式识别会议记录中,第 1125-1134 页,2017 年。

[28] AK 耆那教。数字图像处理基础。新泽西州恩格尔伍德悬崖:Prentice Hall,1989 年。

[29] T. Karras、T. Aila、S. Laine 和 J. Lehtinen。甘斯的渐进生长以提高质量、稳定性和变异性。 arXiv 预印本 arXiv:1710.10196, 2017。

[30] T. Karras、S. Laine 和 T. Aila。用于生成对抗网络的基于样式的生成器架构。在 IEEE 计算机视觉和模式识别会议论文集中,第 4401-4410 页,2019 年。

[31] Y. Katznelson。调和分析简介。凸轮
桥梁大学出版社,2004。

[32] N. Kodali、J. Abernethy、J. Hays 和 Z. Kira。关于甘斯的收敛性和稳定性。 arXiv 预印本 arXiv:1705.07215, 2017。

[33] Y. Li、S. Liu、J. Yang 和 M.-H。阳。生成人脸补全。在 IEEE 计算机视觉和模式识别会议记录中,第 3911–3919 页,2017 年。

[34] Z. Liu、P. Luo、X. Wang 和 X. Tang。深度学习面对野外的贡献。在 IEEE 计算机视觉国际会议论文集中,第 3730-3738 页,2015 年。

[35] P. Luc、C. Couprie、S. Chintala 和 J. Verbeek。使用对抗网络的语义分割。 arXiv 预印本 arXiv:1611.08408, 2016。

[36] J. MacQueen 等人。多变量观测值分类和分析的一些方法。伯克利第五届数理统计与概率研讨会论文集,第 1 卷,第 281-297 页。美国加利福尼亚州奥克兰,1967 年。

[37] X. Mao、Q. Li、H. Xie、RY Lau、Z. Wang 和 S. Paul Smol ley。最小二乘生成对抗网络。在 IEEE 计算机视觉国际会议论文集中,第 2794-2802 页,2017 年。

[38] F. Marra、D. Gragnaniello、D. Cozzolino 和 L. Verdo liva。通过社交网络检测 gan 生成的假图像。 2018 年 IEEE 多媒体信息处理和检索会议 (MIPR),第 384-389 页。

IEEE,2018 年。

[39] F. Marra、D. Gragnaniello、L. Verdoliva 和 G. Poggi。甘斯会留下人工指纹吗? CoRR,abs/1812.11842,2018 年。

[40] L. Metz、B. Poole、D. Pfau 和 J. Sohl-Dickstein。展开
生成对抗网络,2016 年。

[41] M. Mirza 和 S. Osindero。条件生成对抗网络。 arXiv 预印本 arXiv:1411.1784, 2014。

[42] S. Mo、M. Cho 和 J. Shin。实例感知图像到图像的转换。在国际学习表示会议上,2019 年。

[43] A. Nguyen、A. Dosovitskiy、J. Yosinski、T. Brox 和 J. Clune。通过深度生成器网络为神经网络中的神经元合成首选输入。在神经信息处理系统的进展中,第 3387–3395 页,2016 年。

[44] A. Paszke、S. Gross、S. Chintala、G. Chanan、E. Yang、Z. De Vito、Z. Lin、A. Desmaison、L. Antiga 和 A. Lerer。 Pytorch 中的自动微分。 2017.

[45] D. Pathak、P. Krahenbuhl、J. Donahue、T. Darrell 和 AA
埃夫罗斯。上下文编码器:通过修复进行特征学习。在 IEEE 计算机视觉和模式识别会议记录中,第 2536-2544 页,2016 年。

[46] Y. Pu、Z. Gan、R. Henao、X. Yuan、C. Li、A. Stevens 和 L. Carin。用于深度学习图像、标签和说明的变分自动编码器。在神经信息处理系统的进展中,第 2352–2360 页,2016 年。

[47] A. Radford、L. Metz 和 S. Chintala。使用深度卷积生成对抗网络进行无监督表示学习。 arXiv 预印本 arXiv:1511.06434, 2015。

[48] S. Reed、Z. Akata、X. Yan、L. Logeswaran、B. Schiele 和 H. Lee。生成对抗文本到图像合成。 arXiv 预印本 arXiv:1605.05396, 2016。
..

[49] A. Rossler、D. Cozzolino、L. Verdoliva、C. Riess、J. Thies 和 M. Nießner。 FaceForensics++:学习检测被操纵的面部图像。在计算机视觉国际会议 (ICCV) 中,2019 年。

[50] K. Roth、A. Lucchi、S. Nowozin 和 T. Hofmann。在 NIPS 2017,2017 年 05 日。

[51] B. Scholkopf 和 AJ Smola。 Learning with kernels: support vector machines, regularization, optimization, and be yond。麻省理工学院出版社,2001 年。

[52] C. Szegedy、V. Vanhoucke、S. Ioffe、J. Shlens 和 Z. War。
重新思考计算机视觉的初始架构。在 IEEE 计算机视觉和模式识别会议记录中,第 2818-2826 页,2016 年。

[53] Y. Xue、T. Xu、H. Zhang、LR Long 和 X. Huang。 Segan:用于医疗的多尺度 l 1 损失的对抗网络
图像分割。神经信息学,16(3-4):383–392, 2018。

[54] RA Yeh、C. Chen、T. Yian Lim、AG Schwing、M. Hasegawa-Johnson 和 MN Do。具有深度生成模型的绘画中的语义图像。在 IEEE 计算机视觉和模式识别会议记录中,第 5485-5493 页,2017 年。

[55] D. Yin、RG Lopes、J. Shlens、ED Cubuk 和 J. Gilmer。计算机视觉中模型鲁棒性的傅里叶视角。
CoRR,abs/1906.08988,2019 年。

[56] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and TS Huang.具有上下文注意力的生成图像修复。在 IEEE 计算机视觉和模式识别会议论文集中,第 5505-5514 页,2018 年。

[57] N. Yu、L. Davis 和 M. Fritz。将假图像归因于甘斯:学习和分析甘斯指纹。在国际计算机视觉会议 (ICCV) 上,2019 年 10 月。

[58] H. Zhang、T. Xu、H. Li、S. Zhang、X. Wang、X. Huang 和 DN Metaxas。Stackgan:文本到具有堆叠生成对抗网络的逼真图像合成。在 IEEE 计算机视觉国际会议论文集中,第 5907–5915 页,2017 年。

[59] H. Zhang、T. Xu、H. Li、S. Zhang、X. Wang、X. Huang 和 DN Metaxas。Stackgan++:具有堆叠生成对抗网络的真实图像合成。 IEEE 交易模式分析和机器智能,41(8):1947–1962, 2018。

[60] J.-Y。 Zhu、T. Park、P. Isola 和 AA Efros。使用循环一致的对抗性网络进行不成对的图像到图像的转换。在 IEEE 计算机视觉国际会议论文集中,第 2223–2232 页,2017 年。

[61] J.-Y。 Zhu、R. Zhang、D. Pathak、T. Darrell、AA Efros、O. Wang 和 E. Shechtman。迈向多模态图像到图像的翻译。在神经信息处理系统的进展中,第 465-476 页,2017 年。

补充材料

我们论文的补充材料包含有关所呈现实验的更多详细信息,以及一些可能有助于更好地理解上卷积单元的光谱特性的支持实验。

6. 使用光谱失真探测深度假货

在本节中,我们提供了更详细的结果
论文第 4.1 节中介绍的实验。

6.1.有关所用数据集的更多详细信息

6.1.1 人脸总部

据我们所知,目前没有公共数据集提供带有注释的假面和真面的高分辨率图像。因此,我们从已建立的来源创建了自己的数据集,称为 Faces-HQ5 。

为了拥有足够多的面孔,我们选择下载和标记来自 CelebA-HQ 数据集 [29]、Flickr-Faces-HQ 数据集 [30]、100K Faces 项目 [1] 和 www 的可用图像.thispersondoesnotexist.com.我们总共收集了 4 万张高质量图像,其中一半是真人,另一半是假人。表 2 包含摘要。

Training Setting:我们将转换后的数据分为训练集和测试集,20%用于测试阶段,其余80%作为训练集。然后,我们用训练数据训练分类器,最后评估测试集的准确性。

	# of samples	类别标签	
	10000	10000	10000
CelebA-HQ数据集[29]	10000	真实的	0
Flickr-Faces-HQ数据集[30]		真实的	0
100K Faces 项目 [1]		伪造的	1
www.thispersondoesnotexist.com		伪造的	1

表 2:Faces-HQ 数据集结构。

6.1.2 A 类

CelebFaces Attributes (CelebA) 数据集 [34] 由 202,599 张名人面部图像组成,具有 40 种面部属性变化。人脸图像的尺寸为 178x218x3,在我们的模型中可以认为是中等分辨率

语境。

训练设置:虽然我们可以直接使用来自 CelebA 数据集的真实图像,但我们需要自己生成假样本。因此我们使用真实数据集训练一个 DCGAN [47]、一个 DRAGAN [32]、一个 LSGAN [37] 和一个 WGAN-GP [20] 来生成逼真的假图像。我们将数据集拆分为 162,770 张用于训练的图像和 39,829 张用于测试的图像,我们将初始 178x218x3 大小的图像裁剪并调整为 128x128x3。一旦模型被训练好,我们就可以在中分辨率尺度上进行分类实验。

6.1.3 人脸取证++

FaceForensics++ [49] 是图像取证数据集的集合,包含已使用不同的自动面部处理方法修改的视频序列。

一个子集是 DeepFakeDetection 数据集,它包含来自 16 个不同场景的 28 个付费演员的 363 个原始序列,以及超过 3000 个使用 DeepFakes 及其相应的二进制掩码的操纵视频。所有视频都包含一个可追踪的、主要是没有遮挡的正面脸,这使得自动篡改方法能够生成逼真的伪造品。

训练设置:该数据集使用的管道与 Faces-HQ 数据集和 CelebA 相同,但有一个额外的块。由于 DeepFakeDetection 数据集包含视频,我们首先需要提取帧,然后从中裁剪内面。由于视频场景内容不同,这些裁剪后的人脸大小不一。因此,我们对 1D 功率谱进行插值

到固定大小 (300) 并将其归一化,将其除以 0 频率分量。 日

6.2.实验结果

6.2.1 光谱分布

下面的图 15、16 和 17 显示了所有数据集的光谱 (AI) 分布。在所有这三种情况下,很明显分类器应该能够区分真假样本。此外,根据我们的理论分析（参见本文第 2.3 节）,可以假设使用的 Face-HQ 和 FaceForensics++ 数据集中的生成器使用了基于 up+conv 的向上卷积或连续模糊了生成的图像（由于下降在高频）。基于 CelebA 的假货使用 transconv。

图 18 给出了一些额外的数据示例及其相应的 FaceForensics++ 数据的光谱属性。

6.2.2 T-SNE 评估

图 19 显示了我们的 AI 特征的聚类属性。
很明显,分类器在区分两个类（真实的和假的)时应该没有问题。

5Faces-HQ 数据有一个 https://cutt.ly/6enDLYG 大小为 19GB。下载:

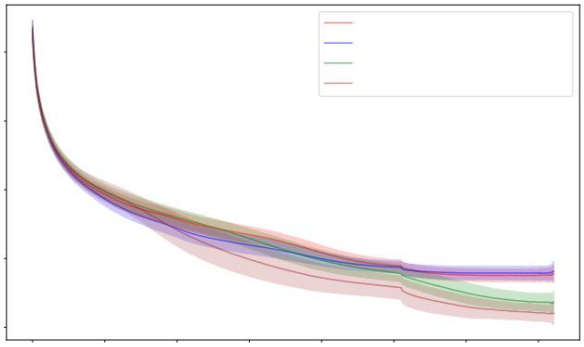


图 15:Faces-HQ 数据集的统计数据（均值和方差）。

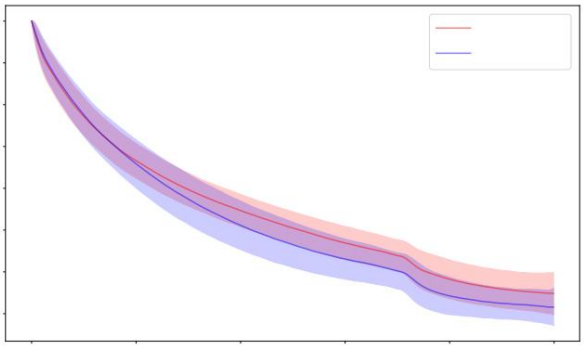


图 16:FaceForen sics++、DeepFakeDetection 数据集的统计数据（均值和方差）。

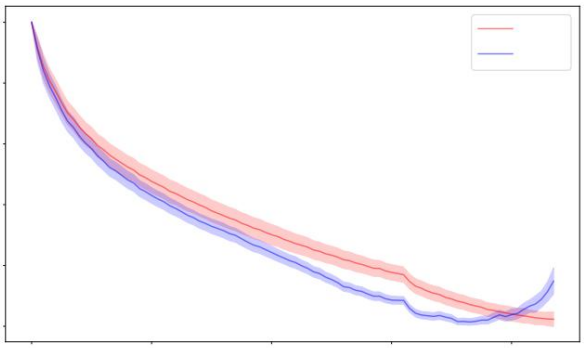


图 17:CelebA 数据集的统计数据（均值和方差）:不同 GAN 方案（DCGAN、DRAGAN、LSGAN 和 WGAN-GP)生成的图像的平均值。

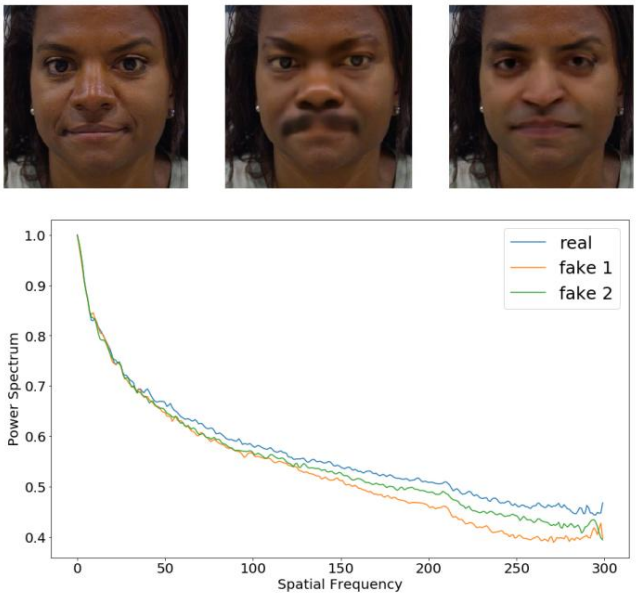


图 18:FaceForensics++ 数据。上图:一张真人脸（左)和两张 Deepfake 人脸的示例,假脸 1（中)和假脸 2（右）。请注意,修改仅影响内面。底部:来自先前图像的归一化和插值一维功率谱。

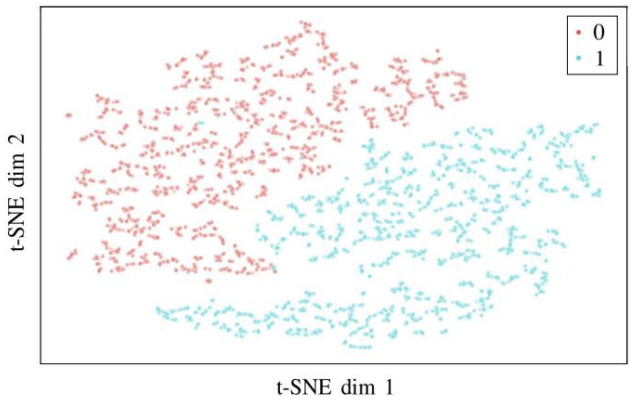


图 19:Faces-HQ 数据集随机子集上一维功率谱的 T-SNE 可视化。我们使用 4 次复杂度和 4000 次迭代来生成绘图。

6.2.3 检测结果取决于数量可用样品

在本节中,我们展示了 DeepFake 检测任务的一些额外结果（论文中的表 1）。在表 3、4 和 5 中,我们关注训练期间可用数据样本数量的影响。如论文所示,我们的方法在无监督环境中运行良好,并且只需 16 个带注释的训练样本即可在有监督环境中实现 100% 的分类准确率。

	80%（训练）- 20%（测试）		
# samples	SVM	Logistic Reg.	K均值
4000	100%	100%	82%
1000	100%	100%	82%
100	100%	100%	81%
20	100%	100%	75%

表 3:Faces-HQ:在不同数据设置下使用 SVM、逻辑回归和 k-means 测试准确性。

	80%（训练）- 20%（测试）		
# samples	SVM	Logistic Reg.	K均值
2000	100%	100%	96%
100	100%	95%	100%
20	100%	85%	100%

表 4:CelebA:使用 SVM、逻辑回归和 k-means 测试准确性。

	80%（训练）- 20%（测试）	
# samples	SVM	Logistic Reg.
2000	85%	78%
1000	82%	76%
200	77%	73%
20	66%	76%

表 5:FaceForensics++:在不同数据设置下使用 SVM 分类器和逻辑回归分类器测试准确性。在单帧上评估。

7. 自动编码器的频谱正则化

在第二部分中,我们展示了一些额外的结果来自我们的 AE 实验（参见论文的图 4）。

7.1.训练期间的损失

图 20 显示了对具有 3 个卷积层和 3 个内核大小为 5 × 5 的滤波器的解码器在使用和不使用频谱正则化的情况下的损失评估（参见论文中的等式 10 和 11）。

这些结果表明,光谱正则化对 AE 的收敛和生成的输出图像的质量（根据 MSE)也有积极影响。

7.2.光谱正则化的影响

图 21 显示了光谱正则化对 AE 问题的影响。我们可以注意到 transconv 和 up+conv 如何在频谱域上受到不同行为的影响,特别是在高频分量中。然而,在应用我们的光谱正则化技术后,结果更接近真实的一维功率谱分布,生成的图像更接近真实分布。

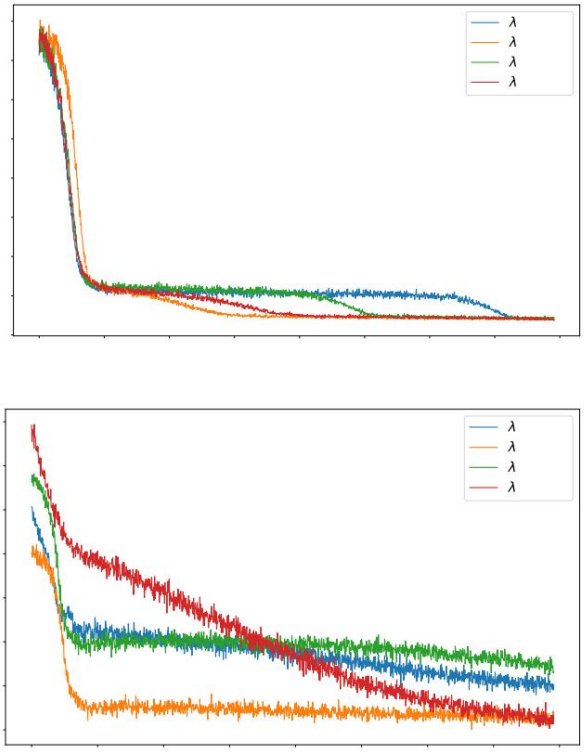


图 20:从我们的 AE定义Lfinal的不同损失的演变。上图:训练 (LReconstruction) 期间的均方误差 (MSE)。底部:训练期间的二元交叉熵损失 (BCE) (LSpectral)。

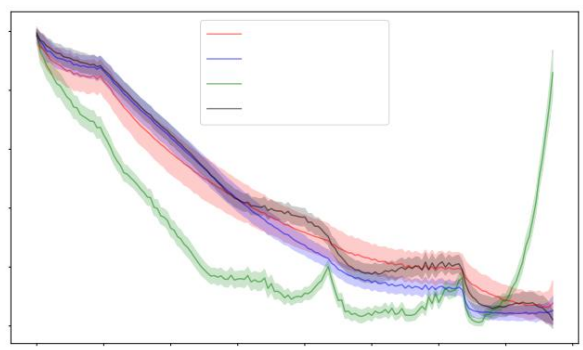


图 21:基线（transconv 和 up+conv)和具有光谱损失的提案（已校正)的 AE 结果。修正后的 AE 在最后一个 transconv 层之后有 3 个额外的卷积层。每层有 32 个大小为 5x5 且 λ = 0.5 的过滤器

7.3.不同拓扑的影响

在本实验中,我们评估了不同拓扑设计选择的影响。图 22 显示了统计数据

某些拓扑的光谱分布：

- 真实：来自 CelebA 的原始人脸图像
- DCGAN v1：一种DCGAN 拓扑结构，在最后两个向上卷积之后具有频谱正则化和一个卷积层（32 个5x5 滤波器）。
- DCGAN v2：具有频谱正则化和两个卷积层（32 个 5x5 滤波器)的 DCGAN 拓扑，在最后一个上卷积之后。
- DCGAN v3：一种DCGAN 拓扑结构，在每个向上卷积之后具有频谱正则化和一个卷积层（32 个5x5 滤波器）。
- DCGAN v4：具有频谱正则化和最后一个上卷积后三个卷积层（32 5x5 滤波器)的 DCGAN 拓扑。

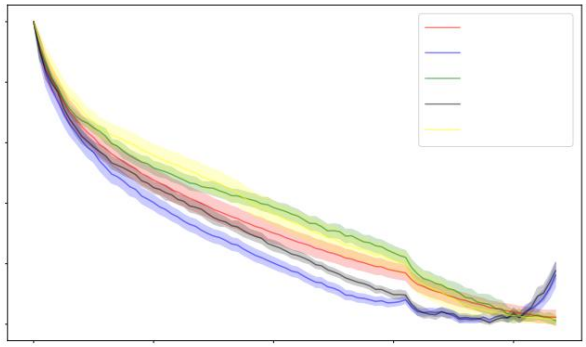


图 22：应用于 DCGAN 的不同拓扑结构的 AE 结果。每个版本都在其 DCGAN 结构中加入了不同数量的卷积层。

经过理论分析和粗略的拓扑搜索验证后，我们得出结论，为了利用谱正则化，在最后一个上卷积之后添加 3 个 5x5 卷积层就足够了。