利用自混合图像检测深度伪造

枫东京大学盐原俊彦山崎

*{盐原，山崎}@cvm.t.u-tokyo.ac.jp*

# 摘要

*在本文中，我们提出了新的合成训练数据称为自混合图像(SBIs)来检测深度假货。通过混合来自单个原始图像的伪源图像和目标图像来生成SBI，再现常见的伪造伪像(例如，混合源图像和目标图像之间的边界和统计不一致性)。SBIs背后的关键思想是，更一般和几乎不可识别的假样本鼓励分类器学习通用和健壮的表示，而不会过度适应特定于操作的伪像。通过遵循标准的交叉数据集和交叉操作协议，我们在FF++、CDF、DFD、DFDC、DFDCP和FFIW数据集上将我们的方法与最先进的方法进行了比较。大量实验表明，该方法提高了模型对未知操作和场景的泛化能力。特别地，在DFDC和DFDCP上，现有方法存在训练集和测试集之间的领域差距，我们的方法在跨数据集评估中分别优于基线4.90%和11.78%。代码可从以下网址获得*[*https://github.com/mapooon/SelfBlendedImages*](https://github.com/mapooon/SelfBlendedImages)*。*

# 介绍

生成性对抗网络最近的快速发展[[11](#_bookmark37),[26](#_bookmark52),[32](#_bookmark58),[33](#_bookmark59),[46](#_bookmark72),[52](#_bookmark78),[64](#_bookmark90)计算机视觉已经使生成逼真的面部图像成为可能。特别是，被称为deepfake的操纵主体的身份、表情或属性的技术被用于娱乐目的，例如智能手机应用程序或电影；然而，它们也可能被用于恶意目的，例如制造假新闻或伪造证据。因此，视觉社区正在积极研究deepfake检测技术。

大多数以前的检测方法[[9](#_bookmark35),[17](#_bookmark43),[27](#_bookmark53),[31](#_bookmark57),[37](#_bookmark63),[49](#_bookmark75),[54](#_bookmark80),[65](#_bookmark91)]在数据集内的情况下表现良好，在这种情况下，他们可以检测他们在训练中学习到的伪造；然而，一些研究表明[16](#_bookmark42),[22](#_bookmark48),[34](#_bookmark60),[62](#_bookmark88)]已经发现在跨数据集的场景中检测性能显著下降



募捐

混合图像

基础图像

基础图像

自混合图像

转换

转换

源图像

目标图像

我们的方法融合了同一张脸

地标匹配

源图像

目标图像

**以前的方法混合不同的面**

图一。假样品合成概述。先前的方法混合两个不同的面部，并基于所选源图像和目标图像之间的间隙产生伪像。相比之下，我们的方法从一个单一的图像混合轻微变化的脸，并通过变换积极地产生伪像。在这个例子中，我们对源图像应用了颜色抖动、锐化、调整大小和平移，而没有对目标图像应用任何变换。

样本是通过未知的操作伪造的。

这个问题最有效的解决方案之一是用合成数据训练模型，这鼓励模型学习用于深度假冒检测的通用表示。例如，面部区域被模糊以再现GAN合成的源图像的质量下降[[42](#_bookmark68)]，从两个原始图像对中生成混合图像，以再现混合伪像[[40](#_bookmark66),[66](#_bookmark92)].然而，deepfakes的质量多年来一直在提高，这导致前一种方法在最近的基准测试中失败[[43](#_bookmark69),[53](#_bookmark79)].尽管后一种方法在一些数据集上表现良好[[2](#_bookmark27),[43](#_bookmark69)]，更具挑战性的数据集中的低质量视频[[20](#_bookmark45),[21](#_bookmark47)]藏物很难被识别-

DF的AUC分别为99.99%、99.88%、99.91%和98.79%，



分别为F2F、FS和NT。虽然FF++上的性能已经饱和，但我们的方法仍然优于整个FF++上的最新水平(99.64%对99.11%)。

* 1. 地标(b)边界(c)颜色(d)频率

**图二。锻造面上的典型人工制品。我们将伪影分为四种类型，(a)界标不匹配，(b)混合边界，(c)颜色不匹配，以及(d)频率不一致。**

由于高压缩或极端暴露导致它们具有不可接受的检测性能。

在本文中，我们提出了新的合成训练数据称为自混合图像(SBIs)来检测深度假货。我们的方法和以前的方法的概述[[40](#_bookmark66),[66](#_bookmark92)]如图所示。[1](#_bookmark1)。关键思想是，包含普通人脸伪造痕迹的更难识别的伪造样本鼓励模型学习更通用和更鲁棒的人脸伪造检测表示。我们分析了伪造的人脸，并定义了四个典型的源于先前作品的人工制品(例如，混合边界[[40](#_bookmark66)]，源特征不一致[[66](#_bookmark92)]，以及频域中的统计异常[[14](#_bookmark40)])如图所示。[2](#_bookmark2)。为了基于我们的关键思想合成这些关联，我们开发了源-目标生成器(STG)和掩码生成器(MG)。STG使用简单的图像处理从单个棱镜图像生成伪源和目标图像对，MG从输入图像的面部标志生成各种混合遮罩。通过将源图像和目标图像与掩模混合，我们获得SBI。用SBIs训练鼓励模型学习一般表示，因为模型学习我们在STG中主动生成的伪造痕迹。此外，我们的方法在计算成本方面提高了训练效率。鉴于先前成功的作品[[40](#_bookmark66),[66](#_bookmark92)]使用最近界标搜索来选择源-目标对，这在计算上是昂贵的，在没有这个过程的情况下生成SBI。因此，我们的方法不会遭受大数据集大小的问题。

我们按照两个评估协议评估我们的方法，跨数据集评估和交叉操作评估。在跨数据集评估中，我们在FF++上训练我们的模型[53](#_bookmark79)]并在CDF上对其进行评估[[43](#_bookmark69)]，DFD [[2](#_bookmark27)]，DFDC [[20](#_bookmark45)]，DFDCP [[21](#_bookmark47)]，以及FFIW [[68](#_bookmark94)].这种实验设置类似于真实的检测场景，其中防御者暴露于看不见的领域。尽管简单，但我们的方法在所有测试集上都超过或至少相当于最先进的方法。特别是，在DFDC和DFDCP上，以前的方法在训练集和测试集之间存在域间隙，我们的方法优于最先进的无监督基线[[66](#_bookmark92)]分别下降4.90%和11.78个百分点。在交叉操作评估中，我们评估了我们的模型对FF++的不可见操作方法的通用性；DF [[4](#_bookmark30)]，F2F [[57](#_bookmark83)]，FS [[6](#_bookmark32)]，以及NT [[56](#_bookmark82)].我们的方法实现了

# 相关著作

Deepfake检测。虽然已经引入了许多检测方法，但是研究的主要课题是最佳神经网络结构的开发(例如，有效的浅层网络[[9](#_bookmark35)]，多任务自动编码器[[22](#_bookmark48),[49](#_bookmark75)]，胶囊网[[50](#_bookmark76)]，递归卷积网-

作品[[27](#_bookmark53),[54](#_bookmark80)]，以及注意力网络[[17](#_bookmark43),[65](#_bookmark91)]).一些

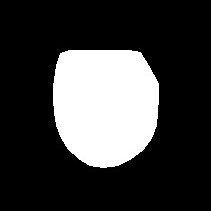
研究[[24](#_bookmark50),[38](#_bookmark64),[44](#_bookmark70),[45](#_bookmark71),[51](#_bookmark77)]专注于频域，以更有效地捕捉伪造痕迹。这些方法在高度压缩的视频上取得了令人印象深刻的性能。另一个值得注意的方向是专注于特定的表示(例如，头部姿势[[63](#_bookmark89)]，眨眼[[31](#_bookmark57),[41](#_bookmark67)]，嘴部动作[[28](#_bookmark54)]，神经元行为[[61](#_bookmark87)]，光流[[10](#_bookmark36)]，以及隐写分析功能[[25](#_bookmark51)]).面部x光[[40](#_bookmark66)]引入了一种基于改变的面部和背景图像之间的边界的面部表示。PCL [[66](#_bookmark92)]测量输入图像的逐块相似性，以检测源图像和目标图像之间的不一致。

训练数据合成。尽管大多数现有方法在检测已知操纵方面表现良好，但一些研究[[16](#_bookmark42),[22](#_bookmark48),[34](#_bookmark60),[62](#_bookmark88)]已经发现，这些方法不能推广到由未知操作伪造的假脸，因为它们倾向于过度适合在训练中看到的方法特定的假像。解决这个问题最有效的方法之一是用合成数据训练模型；这鼓励模型学习用于人脸伪造检测的通用特征。FWA [[42](#_bookmark68)]专注于GAN合成的人脸和自然人脸之间的质量差距，并通过模糊面部区域在真实图像上重新生成。然而，deepfake技术多年来一直在改进，这种方法在最近的基准上检测伪造失败[2](#_bookmark27),[53](#_bookmark79)].BI [[40](#_bookmark66)]和I2G [[66](#_bookmark92)]被引入以生成混合人脸，该混合人脸从具有相似面部标志的两个原始图像对中再现混合伪像。

这些混合图像作为假样本可以很好地训练更通用的检测模型；然而，一些问题仍然存在。首先，因为这些混合伪影依赖于由界标匹配选择的源和目标图像对，所以不规则交换[[58](#_bookmark84)]有时会出现在生成的图像中。这些简单的样本可能会阻止模型学习健壮的表示。第二，因为这些方法被引入来学习定向表示，即BI中的混合边界和I2G中的源特征一致性，所以为鲁棒的深度伪造检测而学习的特征可能仅仅对于混合图像中的伪影是不够的。

**源-目标生成器**

𝐼𝑡



源-目标增强

应用于源或目标

𝐼𝑆𝐵 = 𝐼𝑠 ⊙ 𝑀 + 𝐼𝑡 ⊙ (1 − 𝑀)

*I:基础图像*

目标图像

源图像

调整大小和平移

𝐼𝑠

*ISB : SBI*



混合

𝑀

**掩模发生器**

地标检测

凸包

面具增强

图3。生成SBI概述。基本图像被输入到源-目标生成器(STG)和掩模生成器(MG)中。STG使用一些图像变换从基础图像生成伪源图像和目标图像，而MG生成混合图像

面具从面部标志和变形，以增加面具的多样性。最后，源图像和目标图像与蒙版混合。

# 自混合图像(SBI)

我们的目标是检测deepfakes上被改变的人脸和背景图像之间的统计不一致性。为了训练更通用和鲁棒的检测器，我们生成由常见伪造痕迹组成的合成伪造样本，这些样本很难识别。我们的主要观察结果是，如果deepfake生成技术继续改进，GAN合成的源图像将在它们的属性(例如，面部标志和像素统计)方面更接近原始目标图像。因此，我们开发了一种合成数据生成管道，其中通过从单个图像中混合伪源图像和目标图像来生成伪图像，从而为模型提供更一般和更困难的人脸伪造检测任务。

为了实现这一点，我们引入了自混合图像(SBI)。如图所示。[3](#_bookmark4)SBI通过三个步骤产生；(1)源-目标生成器生成用于混合的伪源和目标图像。源图像和目标图像被增强以产生它们之间的统计不一致性(例如，颜色和频率)。源图像也被调整大小和平移以再现混合边界和界标不匹配。(2)掩模生成器生成具有一些变形的灰度掩模图像。(3)我们将源图像和目标图像与掩模混合以获得SBI。虽然SBI代的一般流程如图2所示。[3](#_bookmark4)，我们展示了Alg中的伪代码。[1](#_bookmark6) 其中该过程与图1中的过程略有不同。[3](#_bookmark4) 为了训练效率(例如，在预处理中提取面部标志，但不在训练中提取)。我们生成假样本的流水线具有恒定的运行时间，而不管数据集的大小

ods [[40](#_bookmark66),[66](#_bookmark92)]在预处理中的运行时间为O(NK ),

用于生成SBIs的算法1伪代码

输入:大小为(H，W，3)的基本图像I，面部标志L

大小为(81，2)

输出:自混合图像的大小为(高，宽，3)

1: def (I) : ▷源-目标增强

T

2: I颜色变换(I)

←

3: I频率变换(I)

←

**4:返回I**

5:如果均匀(最小值= 0，最大值= 1) < 0.5:

6:是，它(一)，我

← T

**7:否则:**

8: Is，It ← I，T (I)

9: Is，p ← RandomResizeTranslate(Is) ▷ p:参数

10: L ← LandmarkTransform(L)

11: M ←凸壳(L)

12:M←parameterizedresize translate(M，p)

13: M ← MaskDeform(M)

14: r ←均匀({0.25，0.5，0.75，1，1，1})

15: M ← rM

16: ISB ←是⊙M+It⊙(1m)

其中N和K分别是视频的数量和每个视频的帧数[1](#_bookmark7)。

## 3.1.源-目标生成器

给定输入图像I，STG通过复制I初始化伪源图像和目标图像。为了在源图像和目标图像之间产生统计不一致，STG随机地对它们中的任何一个应用一些图像变换。这里，我们随机移动输入的RGB通道、色调、饱和度、亮度和对比度的值

1因为[Because[40](#_bookmark66),[66](#_bookmark92)]不公开，我们只定性讨论。



**图4。原始图像的样本(顶行)及其SBI(底行)。**

图像作为颜色转换。然后我们向下采样或者

将输入图像锐化为频率变换。

我们在图2中展示了一些有代表性的SBI例子。[4](#_bookmark8)。虽然SBI的目的不是为了伪造，但是它们包含了伪造人脸中看到的伪影。

## 3.4.SBI培训

一旦生成了SBI，我们就可以训练任何二进制分类器，不管它是否是为深度伪造检测而设计的。给定输入图像X = [x0，x1，，xN 1]

的大小(N，H，W，3)和相应的二进制标签

大小为N的T = [t0，t1，，tN-1]，分类器根据二进制交叉熵损失L优化如下:

l = 1σ{ t log F(x)+(1t)log(1 F(x))}，(2)

*n1*

普通

*我*

*我*

*我*

*我*

*i=0*

为了再现混合边界和地标不匹配，STG调整源图像的大小。设I的高度和宽度分别为H和W。我们定义了

尺寸调整后的图像的高度Hr和宽度Wr为Hr = uhH和Wr = uwW，其中uh和uw从范围[umin，umax]中的连续均匀分布U[umin，umax]中独立采样。调整后的图像被零填充或居中裁剪，以具有与原始图像相同的大小。

然后，STG转换调整大小的源图像。我们将平移向量t = [th，tw]定义为th = vhH和tw = vwW，其中vh和vw独立于U[vmin，vmax]进行采样。

## 屏蔽发生器(毫克)

MG提供灰度掩模图像来混合源图像和目标图像。为此，MG将标志检测器应用于输入图像以预测面部区域，并通过从预测的面部标志计算凸包来初始化掩模。然后用BI [[40](#_bookmark66)].为了增加混合遮罩的多样性，遮罩的形状和混合比率被随机改变。首先，通过弹性变形使掩模变形，如在[[66](#_bookmark92)].第二，用两个不同参数的高斯滤波器平滑掩模。在第一次平滑之后，小于1的像素值被改变为0。这意味着，如果第一高斯滤波器的核大小大于第二高斯滤波器的核大小，则掩模被侵蚀，反之则被扩大。最后，MG改变源图像的混合比例。这可以通过将掩模图像乘以

常数r ∈ (0，1]。这里，我们均匀地从

{0.25, 0.5, 0.75, 1, 1, 1}.

## 混合

通过用混合掩模M混合源图像Is和目标图像It，我们获得自混合图像ISB为

ISB = Is⊙M+It⊙(1m)。(1)

其中F(x)是x为“假”的概率。我们输入“真实”的目标图像，而不是使用基础图像

鼓励模型只关注SBI上的工件。因为MG提供了混合遮罩，所以我们也可以采用基于遮罩的多任务学习[[40](#_bookmark66),[49](#_bookmark75),[66](#_bookmark92)].

# 实验

## 实施细节

预处理。我们采用Dlib [[35](#_bookmark61)]和RetinaFace [[19](#_bookmark46)]分别从每个视频帧中提取面部标志和边界框。我们使用81个面部标志形状预测器[1](#_bookmark28)]在Dlib。对于从边界框计算的脸部的宽度和高度，脸部区域被裁剪为4–20%的随机边距用于训练，12.5%的固定值用于推断。请注意，推断过程中不需要地标；因此，我们只在推理时使用RetinaFace。

源-目标增强。对于颜色和频率变换，我们采用广泛使用的图像处理工具箱中的RGBShift、HueSaturationValue、RandomBrightnessContrast、Downscale和Sharpen[12](#_bookmark38)].

培训。我们采用最先进的卷积网络架构EfficientNet-b4 [[55](#_bookmark81)] (EFNB4)在ImageNet上进行了预先培训[[18](#_bookmark44)]作为分类器，并用SAM [[23](#_bookmark49)]优化器。批量大小和学习率分别设置为32和0.001。为了训练，我们对每个视频只采样八帧。如果在一帧中检测到两个或更多的面部，则提取具有最大边界框的面部。每一批由真实图像和它们的SBI组成，并且相同的增强被应用于每个真实图像和它的SBI。我们还使用了一些数据增强，即图像压缩、RGBShift、HueSaturationValue和随机亮度对比度。

模型验证。考虑到实际情况，在没有额外评估数据集的情况下验证模型是很重要的。我们使用一个由真实的

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 方法输入类型训练集测试集AUC (%) | | | | | | | | | | |
|  |  | 真实的 | 骗子 |  | 累积分布函数（Cumulative Distribution Function的缩写） | 数据流程图 | 有机产品认证中心 | DFDCP | FFIW |  |
| DSP-FWA [[42](#_bookmark68)  面部x光+ BI [[40](#_bookmark66) | 框架框架 | ✓  ✓ | ✓ |  | 69.30  - | - 93.47 | -  - | - 71.15 | -  - |  |
| 面部x光+ BI [[40](#_bookmark66) | 基本框架 | ✓ ✓ - | | | | 95.40 | - | 80.92 | - | |
| LRL [[14](#_bookmark40) | 基本框架 | ✓ ✓ 78.26 | | | | 89.24 | - | 76.53 | - | |
| FRDM [[45](#_bookmark71) | 基本框架 | ✓ ✓ 79.4 | | | | 91.9 | - | 79.7 | - | |
| PCL + I2G [[66](#_bookmark92) | 基本框架 | * 90.03 | | | | **99.07** | 67.52 | 74.37 | - | |
| 双枝[[48](#_bookmark74) | 录像 | ✓ ✓ 76.65 | | | | - | - | - | - | |
| 大坝[[68](#_bookmark94) | 录像 | ✓ ✓ 75.3 | | | | - | - | 72.8 | - | |
| LipForensics [[28](#_bookmark54) | 录像 | ✓ ✓ 82.4 | | | | - | - | - | - | |
| FTCN [[67](#_bookmark93) | 录像 | ✓ ✓ 86.9 | | | | 94.40∗ | 71.00∗ | 74.0 | 74.47∗ | |
| efn B4+SBI(我们的) | 基本框架 | * **93.18** | | | | 97.56 | **72.42** | **86.15** | **84.83** | |

表1。对CDF、DFD、DFDC、DFDCP和FFIW进行跨数据集评估。现有方法的结果直接引用自原始论文及其子序列，以便公平比较。粗体和下划线值分别对应最佳值和次佳值。\*表示我们对官方代码的实验。我们的方法在CDF、DFDC、DFDCP和FFIW上优于最先进的方法，并且在没有任何用于deepfake检测的特殊网络架构的情况下在DFD上达到第二好。

视频及其每个时期后的SBI，并在具有最高AUC的五个权重中选择具有最高时期数的权重。因此，在我们的方法中，即使对于模型验证，也不使用经过处理的图像。

推理策略。为了推断，我们对每个视频采样32帧。如果在一帧中检测到两个或两个以上的人脸，则将分类器应用于所有人脸，并将最高的伪造置信度用作该帧的预测置信度。一旦获得了所有帧的预测，我们对它们进行平均以获得视频的预测。为了公平比较，我们使用所有测试集的所有视频进行评估，将所有帧中没有检测到人脸的视频的置信度设置为0.5。

## 实验环境

数据集。我们采用广泛使用的基准人脸取证++ [[53](#_bookmark79)] (FF++)进行训练，遵循约定。它包含1，000个原始视频和4，000个通过四种操纵方法伪造的伪造视频，即，深度伪造[[4](#_bookmark30)] (DF)，Face2Face [[57](#_bookmark83)] (F2F)，FaceSwap [[6](#_bookmark32)] (FS)和神经纹理[[56](#_bookmark82)] (NT)。对于我们的跨数据集评估，我们使用了五个最近的deepfake数据集。名人-DF- v2 [[43](#_bookmark69)] (CDF)对从YouTube下载的名人视频应用更高级的deepfake技术。深度伪造检测[[2](#_bookmark27)] (DFD)提供了数千个由同意的演员生成的deepfake视频。DeepFake检测挑战预览[[21](#_bookmark47)] (DFDCP)和DeepFake检测挑战公开测试集[[20](#_bookmark45)] (DFDC)，随竞赛一同发布[[3](#_bookmark29)]，包含大量受干扰的视频，例如压缩、下采样和噪声。我们进一步在最近的大规模基准FFIW-10K上提供了一个新的跨数据集基线[68](#_bookmark94)]

(FFIW)主要关注多人场景。我们遵循所有数据集的官方训练/测试划分，除了FFIW，在FFIW中我们使用原始验证集作为我们的测试集，因为官方测试集尚未发布。变形人[[39](#_bookmark65)]和DeeperForensics-1.0 [[30](#_bookmark56)]提供复杂的deepfake视频，我们不在我们的跨数据集评估中采用它们，因为它们从FF++的真实视频中生成deepfake，FF ++与训练中使用的域相同。更多统计细节见补充材料。

框架级基线。我们参考五种最先进的帧级检测方法，包括:(1) DSP- FWA [[42](#_bookmark68)提出了一种基于GAN合成源图像质量退化的训练数据生成方法。(2)面部x光片[[40](#_bookmark66)]通过分割源图像和目标图像之间的混合边界来检测深度假货。该模型是用名为BI的合成假样本训练的，这些样本是通过混合来自不同视频的两幅图像生成的。

**(3)局部关系学习[**[14](#_bookmark40)**] (LRL)和(4)融合+ RSA + DCMA +多尺度[**[45](#_bookmark71)**] (FRDM)融合来自RGB和频域的两种不同表示。**

(5)成对自我一致性学习[[66](#_bookmark92)] (PCL)通过测量输入图像补片之间的一致性来检测深度赝品。该模型用类似于BI [[40](#_bookmark66)].

视频级基线。我们进一步将我们的方法与视频级方法进行比较，视频级方法为一些视频帧输出单个标量伪分数。与帧级别的方法不同，视频级别的方法可以检测跨帧的不一致性，尽管它们需要主题以规则的间隔出现多个帧。我们参考了四种最先进的方法，包括:(1)双分支[[48](#_bookmark74)]提出了拉普拉斯算子

方法测试集AUC (%)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | DF | 面对面 | 空军上士 | 西北地区(Northwest Territories) | FF++ |  |
| 面部x光+ BI [[40](#_bookmark66)] | 99.17 | 98.57 | 98.21 | 98.13 | 98.52 |  |
| PCL + I2G [[66](#_bookmark92)] | **100** | 98.97 | 99.86 | 97.63 | 99.11 |  |
| efn B4+SBI(我们的) | 99.99 | **99.88** | **99.91** | **98.79** | **99.64** |  |

**表二。FF++上的交叉操作评估。我们的方法在F2F、FS、NT和整个FF++上取得了最先进的结果。**

高斯核来增强输入图像的频率分量。(2)辨别注意模型[[68](#_bookmark94)] (DAM)建议关注[[60](#_bookmark86)]，用于多人场景。(3)唇取证[[28](#_bookmark54)]使用预先训练的唇读模型检测嘴部运动的时间不一致性[[47](#_bookmark73)].(4)完全时间卷积网络[[67](#_bookmark93)] (FTCN)通过将卷积的空间核大小减少到1来增强时间表示。

评估指标。我们报告了接收机工作特性曲线(AUC)下的视频电平区域，以便与先前的工作进行比较。通常，帧级预测是对视频帧进行平均。我们还在补充材料中提供了平均精度(AP)。

## 跨数据集评估

为了展示我们的方法的通用性，我们进行了跨数据集评估，其中模型在FF++上训练，在其他数据集上评估。尽管许多研究人员已经考虑了这项任务，但是他们每个人在实验中使用的测试集因工作而异，使得全面的比较变得困难。因此，我们仔细检查了以前著作中的实验设置，并把它们汇编成表[1](#_bookmark10)。

与帧级方法的比较。这里，我们将我们的方法与其他帧级方法进行比较[[14](#_bookmark40),[40](#_bookmark66),[42](#_bookmark68),[45](#_bookmark71),[66](#_bookmark92)].我们的方法在CDF、DFDC和DFDCP上分别优于最先进的方法6.08%、5.17%和5.23%，平均提高基线4.58%(87.33%对82.75%)。我们的结果与PCL + I2G相当[66](#_bookmark92)]在DFD (97.56%对99.07%)，伪造的人脸有时会与其他一些原始人脸一起放在一个经过处理的帧中，受试者在整个经过处理的视频中所占的比例比其他测试集要小。因此，我们的方法可以通过将任何对象跟踪过程合并到我们的推理策略中来改进，如在PCL + I2G [[66](#_bookmark92)]，而不是像我们的简单策略那样以相等的间隔从视频中提取帧。

与视频级方法的比较。然后，我们将我们方法与视频级方法进行比较[28](#_bookmark54),[48](#_bookmark74),[67](#_bookmark93),[68](#_bookmark94)].为了更全面的比较，我们对FTCN进行了额外的实验[67](#_bookmark93)]在未经考虑的测试集上，即DFD、DFDC和正式发布的FFIW

方法测试集AUC (%)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | DF | 面对面 | 空军上士 | 西北地区(Northwest Territories) | FF++ |
| 异常+ BI [[40](#_bookmark66)] | 98.95 | 97.86 | 89.29 | 97.29 | 95.85 |
| 例外+ SBIs(我们的) | **99.99** | **99.90** | **98.79** | **98.20** | **99.22** |

**表3。与BI的AUC比较[**[**40**](#_bookmark66)**].**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 累积分布函数（Cumulative Distribution Function的缩写） | 有机产品认证中心 | DFDCP | 平均 |  |
| ResNet-34 + I2G [[66](#_bookmark92)] | 78.18 | 51.72 | 69.93 | 66.61 |  |
| ResNet-34+SBI(我们的) | **87.04** | **66.41** | **82.16** | **78.54** |  |

方法测试集AUC (%)

**表4。与I2G的AUC比较[**[**66**](#_bookmark92)**].**

代码[[7](#_bookmark33)].结果在表中用\*表示[1](#_bookmark10)。我们的方法在CDF、DFD、

DFDC、DFDCP和FFIW，平均提高基线6.68个百分点(86.83%对80.15%)。我们还在用于LipForensics实验的DFDC子集上评估了我们的方法[28](#_bookmark54)]，表现优于竞争对手(76.78%对73.5%)。视频列表可在作者的知识库中获得[[8](#_bookmark34)].

## 交叉操作评估

在真实的检测情况下，防御者通常不知道攻击者的伪造方法。因此，验证模型对各种伪造方法的普适性是非常重要的。遵循[[40](#_bookmark66),[66](#_bookmark92)]，我们在FF++的四种操作方法上评估了我们的模型，即DF、F2F、FS和NT。我们和竞争对手一样使用原始版本进行评估。

桌子[2](#_bookmark13)展示了我们在FF++上的交叉操作评估结果。我们的方法在四个操作上优于或接近于现有的方法(在DF上99.99%，在F2F上99.88%，在FS上99.91%，在NT上98.79%)，并且在整个FF++上实现了最好的性能(99.64%对99.11%)。这一结果表明，我们的方法不仅对deepfakes有效，而且对其他人脸操作也有效。

## 数据质量评估

这里，我们将我们的方法与最先进的合成训练数据进行比较[40](#_bookmark66),[66](#_bookmark92)]，消除分类器差异的影响。为了实现这一点，我们训练了与竞争对手在原始论文中使用的相同的模型和优化器。桌子[3](#_bookmark15) 呈现与BI的比较[[40](#_bookmark66)].我们训练例外[15](#_bookmark41)]与亚当[[36](#_bookmark62)]优化器。我们的方法优于BI [[40](#_bookmark66)]根据AUC对FF++的所有操作方法。特别是，FS的基线从89.29%提高到98.79%。接下来，与I2G的比较结果[[66](#_bookmark92)]在表中给出[4](#_bookmark16)。我们训练ResNet-34 [[29](#_bookmark55)]使用Adam optimizer。我们的方法优于I2G[66](#_bookmark92)]关于CDF，DFDC，

流程测试集AUC (%)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | FF++ | 累积分布函数（Cumulative Distribution Function的缩写） | DFDCP | FFIW | 平均 |  |
| 无来源8月 | 98.58 | **93.59** | 78.06 | 61.11 | 82.84 |  |
| 不含目标8月 | 99.35 | 76.61 | 83.84 | 82.87 | 85.67 |  |
| 不含S-T aug | 89.18 | 70.68 | 85.16 | **88.31** | 83.33 |  |
| 不含资源和运输 | 99.58 | 85.28 | 81.04 | 74.69 | 85.15 |  |
| SBIs(我们的) | **99.64** | 93.18 | **86.15** | 84.83 | **90.95** |  |

**表5。STG各工序的效果。跳过任何进程都会导致致命的性能下降。**

训练集测试集AUC (%)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 数据库ˌ资料库 | #真实 |  | FF++ | 累积分布函数（Cumulative Distribution Function的缩写） | DFDCP | FFIW | 平均 |  |
|  | FF++ | 720 |  | 99.64 | 93.18 | 86.15 | 84.83 | 90.95 |  |
|  | 累积分布函数（Cumulative Distribution Function的缩写） | 622 |  | 98.10 | 93.74 | 81.10 | 77.82 | 87.69 |  |
|  | DFDCP | 737 |  | 98.76 | 90.79 | 88.70 | 81.31 | 89.89 |  |
|  | FFIW | 7090 |  | 99.72 | 95.57 | 78.91 | 88.07 | 90.57 |  |

表6。不同训练数据集的性能。我们的方法在每个训练数据集上都取得了很好的结果。“#Real”表示训练集的真实视频数，不包括验证集的真实视频数。

和DFDCP及其平均值分别下降8.86%、14.69%、12.23%和11.93个百分点。这些结果清楚地表明，无论网络架构如何，我们的方法作为合成训练数据都优于竞争对手。

## 消融术

STG各工序的效果。在STG中，我们使用一些图像处理来生成伪源图像和伪目标图像。相反，因为习得的表征是基于我们在STG中主动提供的工件，所以生成过程的消融实验使得能够在deepfake基准上探索有效的线索。这里，我们在没有一些过程的情况下训练我们的模型，即源增强、目标增强、源-目标增强或调整大小和转换，并在FF++、CDF、DFDCP和FFIW上评估它们。如表中所示[5](#_bookmark17)，source和tar- get增强在检测deepfakes方面确实有效，这两者对于更好的性能都是必要的。我们还观察到，调整大小和平移会再现重要的伪像，因为没有它们性能会很差。通过消融，可以得出结论，不同的线索对于不同数据集上的检测器是有用的，因为它们具有不同的深度假生成过程。

训练数据集的通用性。从实用的角度来看，重要的是表明我们的方法可以在各种真实的人脸数据集上很好地执行。我们在这里用FF++、CDF、DFDCP和FFIW的原始视频中的SBI来训练模型。然后我们在测试集上评估它们。在CDF和FFIW上，我们将原始训练集分成不同的训练/验证集。桌子[6](#_bookmark19) 呈现结果。

架构测试集AUC (%)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | FF++ | 累积分布函数（Cumulative Distribution Function的缩写） | DFDCP | FFIW | 平均 |  |
| ResNet-50 | 97.77 | 90.66 | 82.88 | 79.30 | 87.65 |  |
| ResNet-152 | 98.33 | 90.71 | 85.01 | 76.43 | 87.62 |  |
| 例外 | 99.26 | 90.27 | 78.85 | 76.72 | 86.28 |  |
| 效率网络-b1 | 99.10 | 91.16 | 84.58 | 80.23 | 88.77 |  |
| 高效网络-b4 | **99.64** | **93.18** | **86.15** | **84.83** | **90.95** |  |

**表7。不同网络架构的性能。容量更大的架构往往会带来更好的通用性。**

我们的方法被推广到所有数据集，而没有关键的性能下降。我们观察到FFIW的大数据集大小有助于模型的通用性。但FFIW和DFDCP的视频场景差异导致在DFDCP上性能略有下降；FFIW由从YouTube收集的视频组成，而DFDCP由拍摄招募的受试者制作的视频组成。结果还表明，学习原始视频可以帮助检测与训练相同领域的伪造人脸，即使模型没有学习经过处理的视频，如表中棕色突出显示的分数所示[6](#_bookmark19)，支持我们不采用FaceShifter [[39](#_bookmark65)]和DeeperForensics-1.0 [[30](#_bookmark56)]在跨数据集评估中，如第节所述[4.2](#_bookmark12)。

网络架构的选择。虽然我们采用了EfficientNet-b4 [[55](#_bookmark81)]作为我们的标准分类器，我们的方法可以应用于其他网络体系结构。在这里，我们研究了不同的最新架构的性能，即ResNet-50，-152 [[29](#_bookmark55)]，异常[[15](#_bookmark41)]、EfficientNet-b1和-b4 [[55](#_bookmark81)]用SBIs训练。如表中所示[7](#_bookmark20)，所有架构在FF++、CDF、DFDCP和FFIW上都取得了良好的结果，而没有出现严重的性能下降。值得注意的是，如表中所示，即使我们的方法使用vanilla ResNet-50也优于所有以前在CDF、DFDCP和FFIW上的方法[1](#_bookmark10) 和[7](#_bookmark20)。我们观察到更大的网络倾向于导致更大的通用性，这表明SBIs提供了各种各样的训练样本。

## 定性分析

为了获得定性的见解，我们可视化模型显著图和特征空间。通过分析，我们使用了两个模型；一个在FF++上训练(基线)，另一个在SBIs上训练(我们的模型)。

显著图。为了可视化模型在伪造人脸上的注意力，我们应用Grad- CAM++ [[13](#_bookmark39)]到FF++的操作框架上的模型，即DF、F2F、FS和NT，如图。[5](#_bookmark21)。可以观察到，我们的方法鼓励模型使其注意力比基线更稀疏。这是因为我们的模型检测到了与操作无关的微小伪影，例如混合边界，而基线捕捉到了广泛分布在伪造人脸中的特定于方法的像素分布。

在FF++上训练的DF

接受过SBI培训

在FF++上训练的F2F

接受过SBI培训

接受FF++培训的FS

接受过SBI培训

在FF++上训练的NT

接受过SBI培训

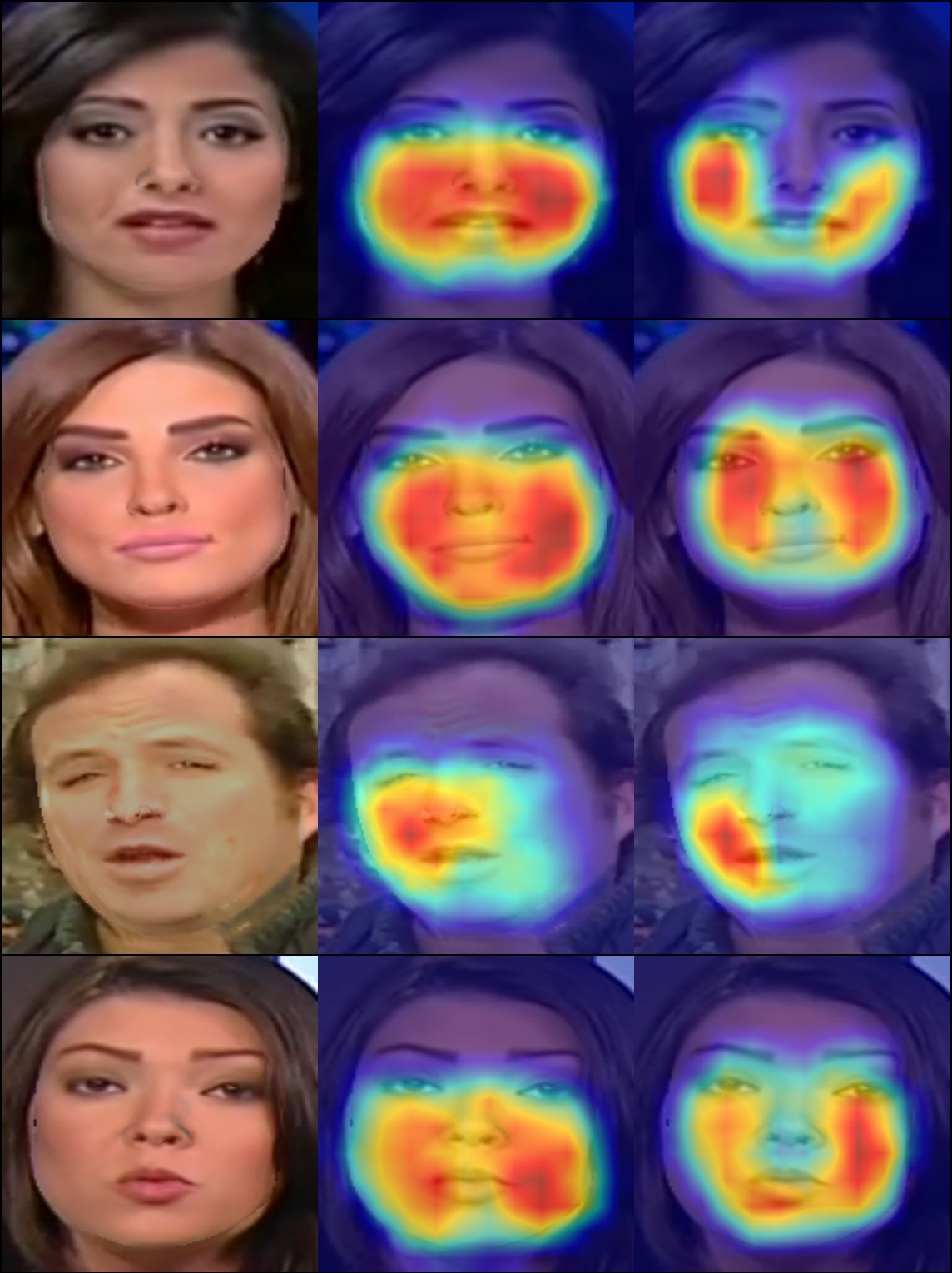
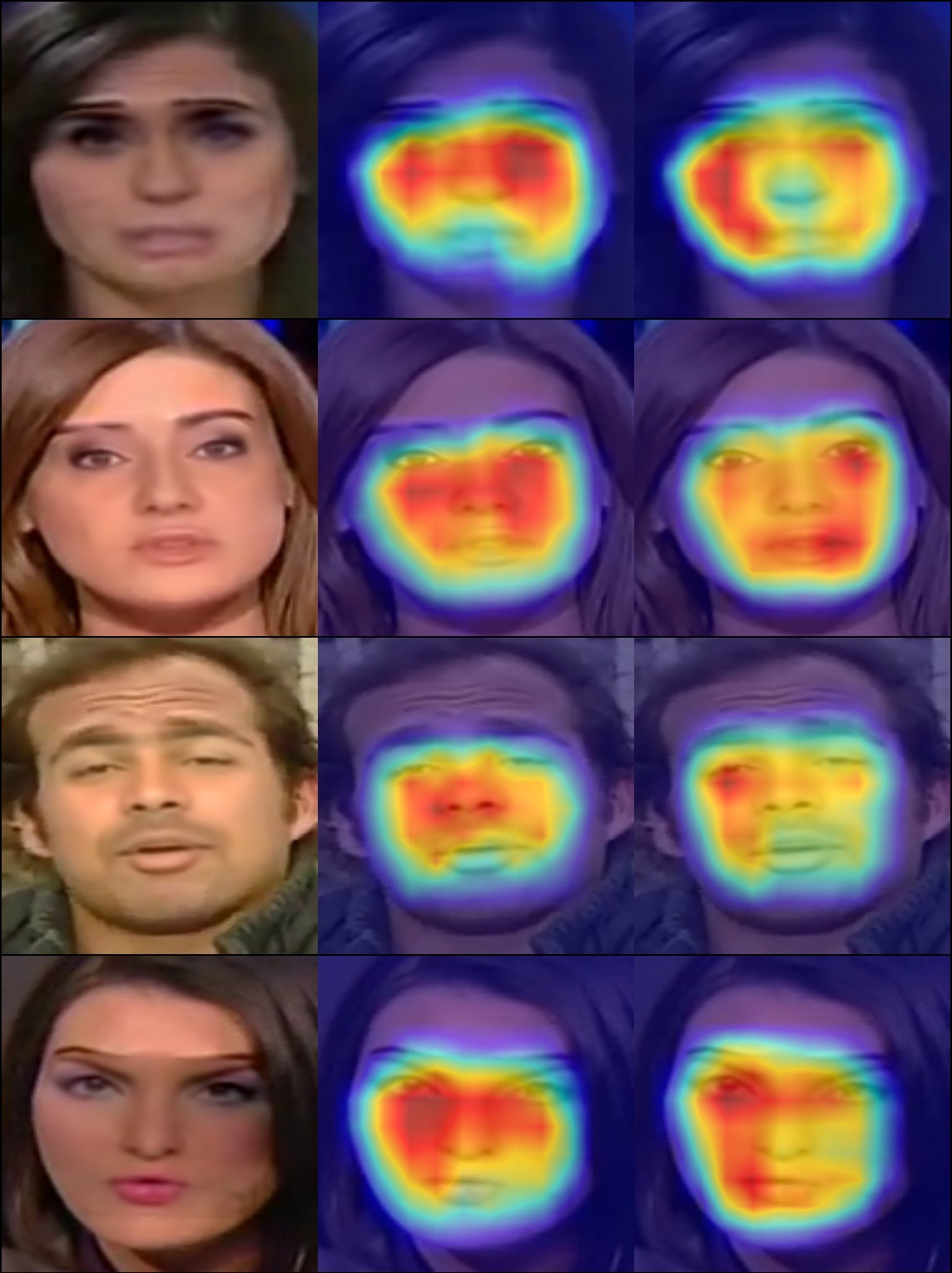
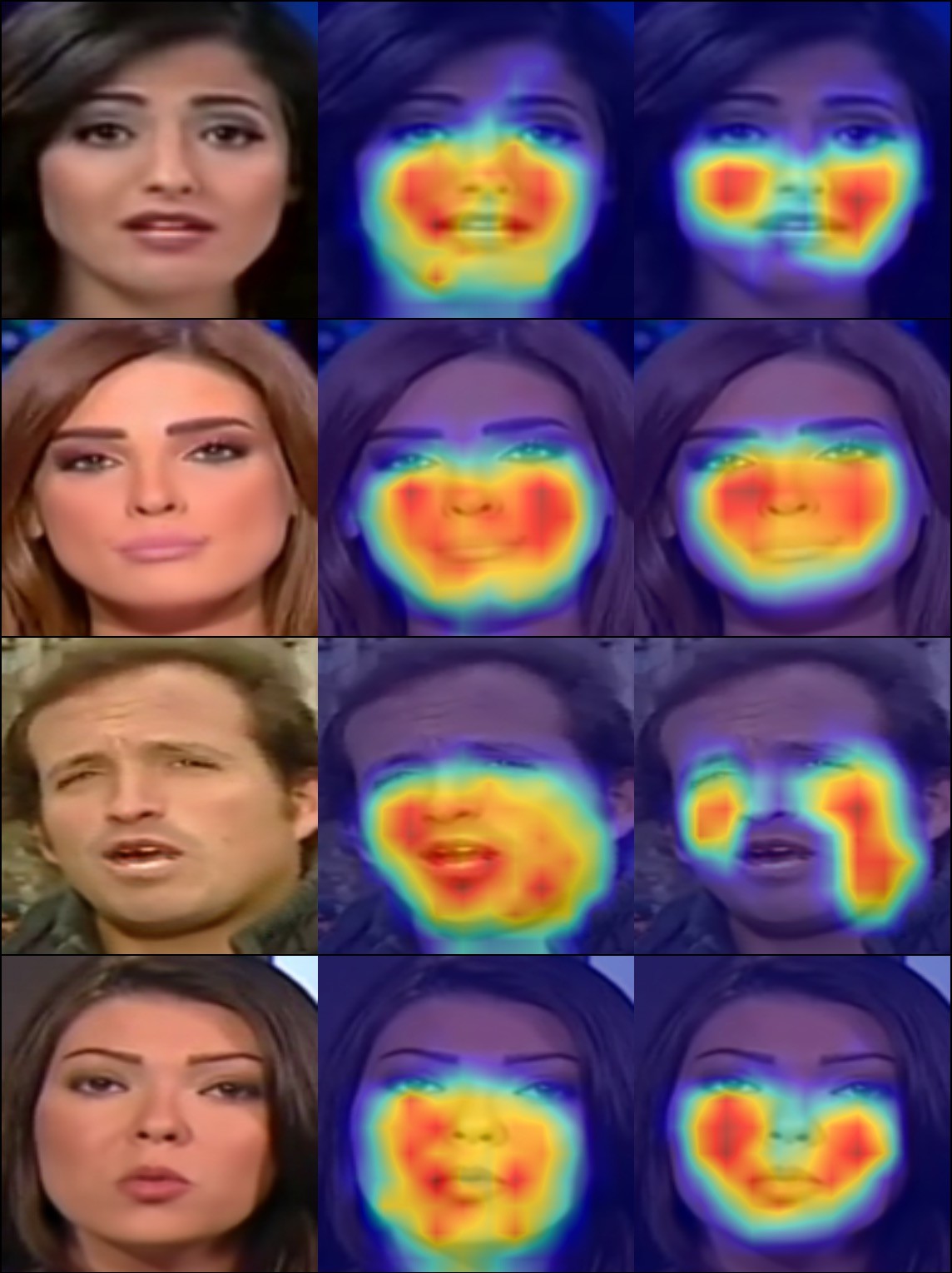
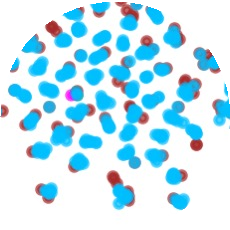
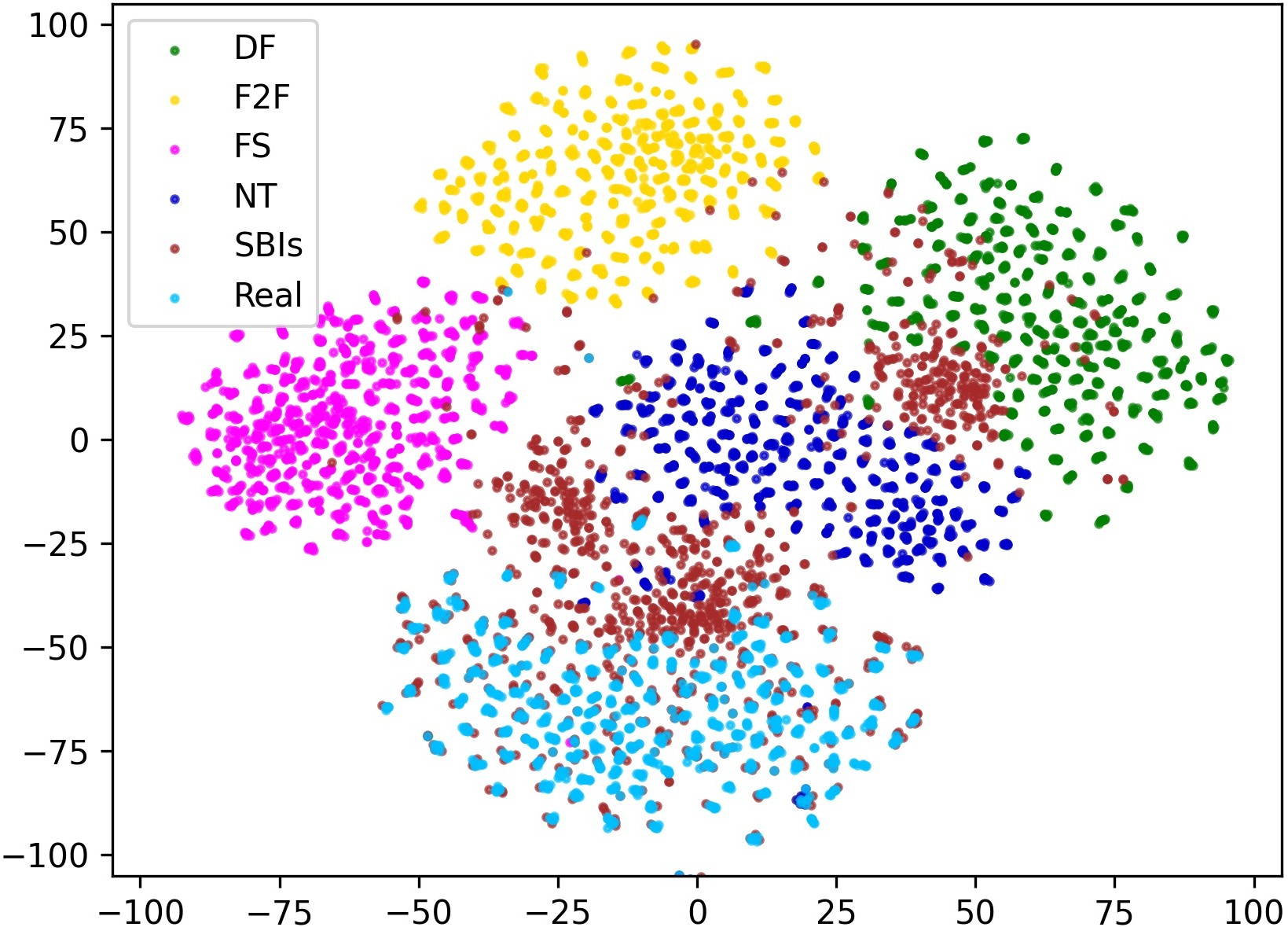
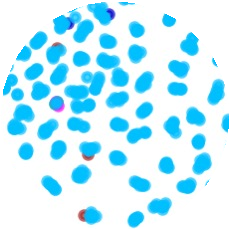
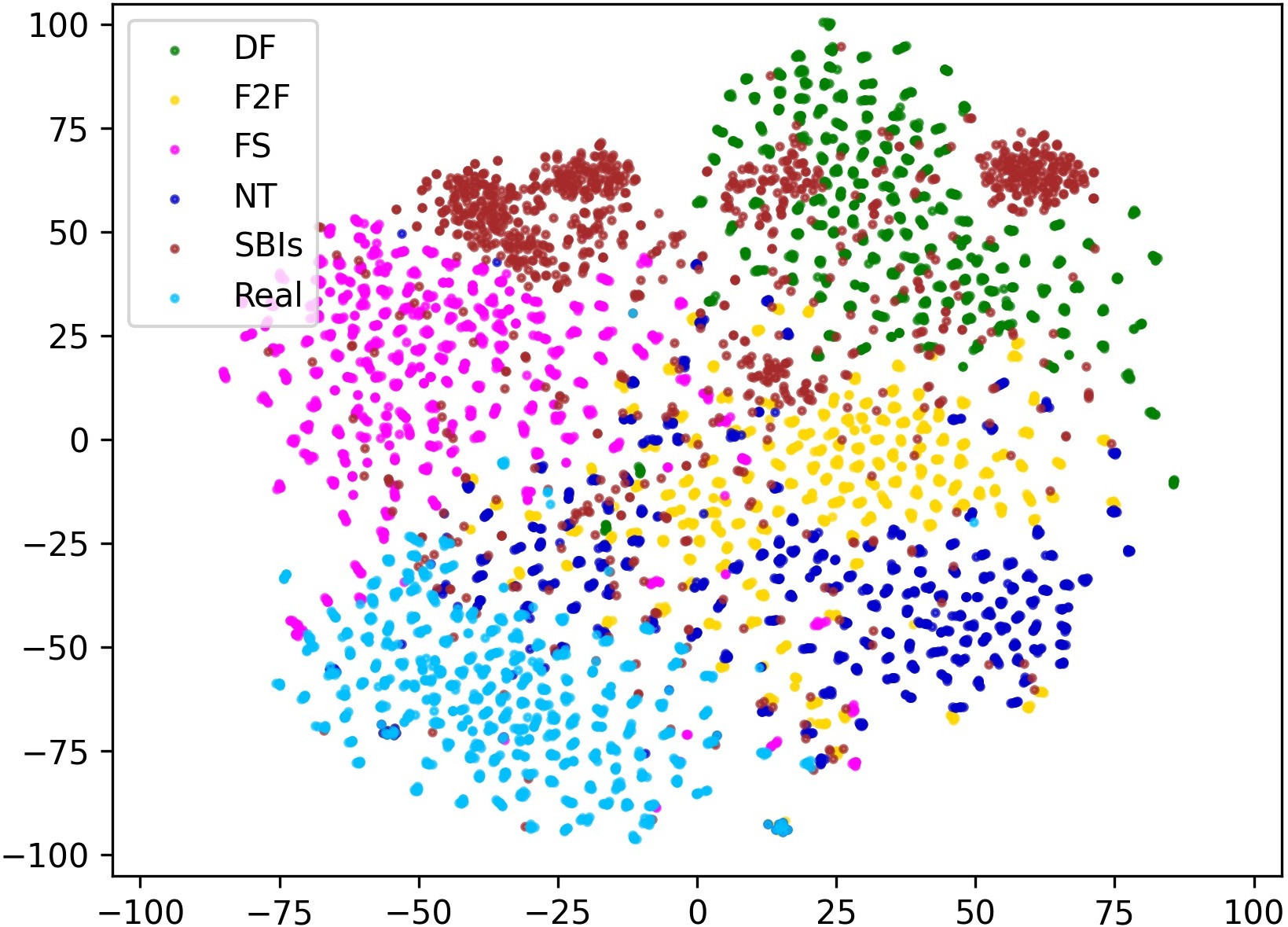
 

图5。基线和我们的模型的显著图可视化。基线捕捉广泛存在于伪造人脸中的方法特定的伪像，而我们的模型检测独立于操作的次要伪像。最好是彩色的。



* + 1. 接受FF++培训



* + 1. 接受过SBI培训

图6。基线(a)和我们的模型(b)的特征空间可视化。基线不能将真实图像与SBI区分开(因为特征向量落入相同的特征空间),而我们的模型不仅成功地将真实图像与SBI区分开，而且成功地将伪造图像区分开。最好是彩色的。

特征空间。然后我们应用t-SNE [[59](#_bookmark85)]可视化到模型最后一层的特征向量。我们再次强调，基线很容易识别伪造的人脸，因为它们是在训练中看到的，我们的目标是将真实的人脸与其他人分开，而不是对操纵的类型进行分类。如图所示。[6](#_bookmark23)基线不能从真实图像中区分SBI，尽管它包含-

训练中常见的四种手法。另一方面，我们的模型不仅区分了SBI，还区分了伪造的人脸和真实的人脸。我们还观察到SBI分布在特征空间的四个操作中。这些结果表明SBI是训练人脸伪造检测器的通用合成数据。

# 限制

尽管我们在交叉数据集和交叉操作评估中的结果预期是有益的，但我们观察到我们方法的一些局限性。首先，与其他帧级方法类似，我们的模型无法捕捉视频帧之间的时间不一致性。因此，具有较少空间伪影的复杂的深度伪造生成技术可以通过我们的检测器。此外，我们的方法在整幅图像合成上表现不佳，因为我们将“伪图像”定义为人脸区域或背景被操纵的图像。我们在从FFHQ数据集和Style- GAN [[33](#_bookmark59)]合成，其AUC仅为69.11%。

# 结论

本文提出了一种新的合成训练数据——自混合图像(SBIs ),它基于这样一种思想，即更一般的和难以识别的伪样本鼓励分类器学习更一般的和鲁棒的表示。通过混合伪源图像和目标图像来生成SBI，伪源图像和目标图像是从单个真实图像稍微变换的，以再现伪造伪像。使用SBIs，我们可以在没有伪造人脸图像的情况下训练探测器。大量的实验表明，我们的方法优于当前最先进的方法，并且适用于不同的网络结构和训练数据集。

# 附加实验

为了显示我们提出的方法的有效性和通用性，我们使用BI [[40](#_bookmark66)].我们按照作者的代码实现它[[5](#_bookmark31)].

**地标相似度。在BI中，100幅图像与**

方法测试集AUC (%)

CDF DFD DFDC DFDCP FFIW

(a)地标相似性的影响

BI1-100(原装)69.40 97.50 66.55 68.71 85.69

BI1-20 71.44 97.27 65.27 68.80 85.35

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 基准图像的面部标志被用作候选 | (Joi的影响 | nt-SBI培训 | 还有毕 |  | |
| 源图像。我们训练效率网-B4 [[55](#_bookmark81)  所有100个(BI1-100)和前20个(BI1-20)。如Ta-所示 | (我们的)BI | **93.18 97.56**  69.40 97.50 | **72.42**  66.55 | **86.15**  68.71 | 84.83  **85.69** |
| brotherhood of locomotive engineers 火车司机兄弟会[8](#_bookmark25) | SBIs + BI | 89.36 98.34 | 71.87 | 82.92 | 83.53 |
| 在CDF和DFDCP的原始BI上接受培训，并达到标准  DFD，DFDC和FFIW。这个结果表明， | (c)来源-目标扩大的影响 | | 在…的地位 | 双性恋的 |  |

至少在地标相似度上，相似度低的简易样本对模型通用性没有贡献。

SBIs和BI的联合培训。为了探索更一般的deepfake检测的最佳实践，我们在SBI和BI的联合数据集上训练EFNB4，每个数据集都以0.5的概率进行采样。如表所示，联合训练的结果基本上低于我们提出的SBIs[8](#_bookmark25)(二)。

将源-目标增强应用于商业智能。为了证明我们的两个想法(1)混合身份图像和(2)增强源和目标图像的优越性，我们将源-目标增强合并到BI中。如表中所示[8](#_bookmark25)(c)在五个测试集中的四个上，结果低于SBIs，尽管它们比原始BI的结果好，这表明我们的两个想法对于一般的deepfake检测都是重要的。

# 综合结果

我们提供了我们方法的综合结果，包括接收机工作特性曲线下的视频电平面积(AUC)和平均精度(AP)。我们还描述了真假视频的数量。我们还在FaceShifter上评估了我们的方法[[39](#_bookmark65)] (FSh)和Deep- erForensics1.0 [[30](#_bookmark56)] (DF1.0)数据集。在DF1.0上，我们遵循惯例使用FF++的c23(轻度压缩)真实视频。结果在表中给出[9](#_bookmark26)。

**表8。附加实验。**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SBIs(我们的) | **93.18 97.56** | **72.42** | **86.15** | 84.83 |
| 双性恋的 | 69.40 97.50 | 66.55 | 68.71 | **85.69** |
| 双w/ S-T Aug | 78.99 99.04 | 71.98 | 74.25 | 83.14 |

### 测试集指标

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 数据库ˌ资料库 | #真实 | #假的 |  | AUC(%) | AP(%) |  |
|  | DF [[4](#_bookmark30) | 140 | 140 |  | 99.99 | 99.99 |  |
|  | F2F [[57](#_bookmark83) | 140 | 140 |  | 99.88 | 99.89 |  |
|  | FS [[6](#_bookmark32) | 140 | 140 |  | 99.91 | 99.91 |  |
|  | 新台币[[56](#_bookmark82) | 140 | 140 |  | 98.79 | 99.15 |  |
|  | FF++ [[53](#_bookmark79) | 140 | 560 |  | 99.64 | 99.92 |  |
|  | DFD [[2](#_bookmark27) | 363 | 3068 |  | 97.56 | 99.70 |  |
|  | FSh [[39](#_bookmark65) | 140 | 140 |  | 98.27 | 98.24 |  |
|  | DF1.0 [[30](#_bookmark56) | 140 | 140 |  | 83.14 | 85.06 |  |
|  | CDF [[43](#_bookmark69) | 178 | 340 |  | 93.18 | 96.35 |  |
|  | DFDC [[20](#_bookmark45) | 2500 | 2500 |  | 72.42 | 75.17 |  |
|  | DFDCP [[21](#_bookmark47) | 276 | 501 |  | 86.15 | 91.37 |  |
|  | FFIW [[68](#_bookmark94) | 250 | 250 |  | 84.83 | 84.30 |  |

**表9。综合结果和统计细节。**

# 参考

* 1. 81面部标志形状预测器。[https://github.com/codeniko/shape \_ predictor \_ 81 \_ face \_landmarks](https://github.com/codeniko/shape_predictor_81_face_landmarks)。访问时间:2021-11-13。[4](#_bookmark9)
  2. 为deepfake检测研究贡献数据。[https : / / ai。谷歌博客。com / 2019 /09 /贡献数据到深度假-detection.html](https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html)。访问时间:2021-11-13。[1](#_bookmark0),[2](#_bookmark3),[5](#_bookmark11),[9](#_bookmark24)
  3. Deepfake检测挑战。[https://www.kaggle.伪造-检测-挑战](https://www.kaggle.com/c/deepfake-detection-challenge)。成功日期:2021年11月13日。[5](#_bookmark11)
  4. Deepfakes。<https://github.com/deepfakes/faceswap>。访问时间:2021-11-13。[2](#_bookmark3),[5](#_bookmark11),[9](#_bookmark24)
  5. 面部x光。[https : / / github。com / AlgoHunt /Face-Xray](https://github.com/AlgoHunt/Face-Xray)。访问时间:2021-11-13。[9](#_bookmark24)
  6. 面部交换。[https : / / github。com /MarekKowalski/FaceSwap/](https://github.com/MarekKowalski/FaceSwap/)。访问时间:2021-11-13。[2](#_bookmark3),[5](#_bookmark11),[9](#_bookmark24)
  7. Ftcn。<https://github.com/yinglinzheng/FTCN>。

访问时间:2021-11-13。[6](#_bookmark14)

* 1. 唇语鉴识。<https://github.com/ahaliassos/LipForensics>。访问时间:2021-11-13。[6](#_bookmark14)
  2. D.Afchar、V. Nozick、J. Yamagishi和I .越前。Mesonet:一个紧凑的面部视频伪造检测网络。在WIFS，第1–7页，2018。[1](#_bookmark0),[2](#_bookmark3)
  3. 艾琳·阿梅里尼、莱昂纳多·加尔特里、罗伯托·卡尔德利和阿尔-贝尔托·德尔·宾博。基于cnn的光流视频深度防伪检测。在ICCV，2019年。[2](#_bookmark3)
  4. 马丁·阿约夫斯基，苏密特·钦塔拉和勒昂·博图。沃瑟斯坦生成对抗网络。在ICML，2017年第214-223页。[1](#_bookmark0)
  5. 亚历山大·巴斯拉耶夫、弗拉迪米尔·伊格洛维科夫、尤金·赫维德-陈亚、亚历克斯·帕利诺夫、米哈伊尔·德鲁日宁和亚历山大·巴甫洛夫·加里宁。相册:快速灵活的图像增强。信息，11:125，02 2020。[4](#_bookmark9)
  6. Aditya Chattopadhay，Anirban Sarkar，Prantik Howlader和梵持N . Balasubramanian。Grad-cam++:深度卷积网络的通用化的基于梯度的可视化解释。在WACV中，第839–847页，2018。[7](#_bookmark18)
  7. 、姚太平、、丁守洪、、嵇蓉蓉。用于人脸伪造检测的局部关系学习。在AAAI，第35卷，第1081-1088页，2021年。[2](#_bookmark3),[5](#_bookmark11),[6](#_bookmark14)
  8. 弗朗索瓦.乔莱。例外:具有深度可分卷积的深度学习。在CVPR，第1251-1258页，2017。[6](#_bookmark14),[7](#_bookmark18)
  9. 达维德·科佐利诺、胡斯图斯·提斯、安德烈亚斯·罗斯勒、克里斯蒂安·里斯、马蒂亚斯·涅纳和路易莎·维尔多瓦。forensitrans-fer:用于伪造检测的弱监督域适应。arXiv:1812.02510，2018。[1](#_bookmark0),[2](#_bookmark3)
  10. 郝当、柳峰、乔尔·斯特霍沃、刘晓明和安尼尔·K·贾恩。数字人脸操作的检测。

在2020年的CVPR。[1](#_bookmark0),[2](#_bookmark3)

* 1. 邓佳，魏东，理查德·索彻，李·，和。Imagenet:一个大规模分层图像数据库。在CVPR，2009年第248-255页。[4](#_bookmark9)
  2. 邓建康，，周宇翔，于，艾琳科特西亚和斯特凡诺斯扎费里欧。视网膜面部:野外单一阶段密集面部定位。arXiv:1905.00641，2019。[4](#_bookmark9)
  3. Brian Dolhansky，Joanna Bitton，Ben Pflaum，Jikuo Lu，Russ，，和Cristian Canton Ferrer。deepfake检测挑战数据集。arXiv:2006.07397，2020。[1](#_bookmark0),[2](#_bookmark3),[5](#_bookmark11),[9](#_bookmark24)
  4. 布莱恩·多尔汉斯基、拉斯·霍维斯、本·普弗洛姆、妮可·巴兰和克里斯蒂安·坎顿·费勒。deepfake检测挑战(dfdc)预览数据集。arXiv:1910.08854，2019。[1](#_bookmark0),[2](#_bookmark3),[5](#_bookmark11),[9](#_bookmark24)
  5. 杜、婆、李和。利用位置感知自动编码器进行可推广的深度伪造检测。在ACM CIKM，第325–334页，2020。[1](#_bookmark0),[2](#_bookmark3)
  6. Pierre Foret、Ariel Kleiner、Hossein Mobahi和Behnam Neyshabur。有效提高泛化能力的清晰度感知最小化。2021年在ICLR。[4](#_bookmark9)
  7. 乔尔·弗兰克、托尔斯滕·艾森霍夫、丽娅·斯科尔、阿斯嘉·菲斯彻、多萝西娅·科洛萨和托尔斯滕·霍尔茨。利用频率分析进行深度伪造图像识别。在ICML，第3247–3258页，2020。[2](#_bookmark3)
  8. 米罗斯拉夫·戈尔扬和杰西卡·弗里德里希。用于彩色图像隐写分析的Cfa感知特征。SPIE，9409，2015。[2](#_bookmark3)
  9. 伊恩·古德菲勒、让·普吉-阿巴迪、迈赫迪·米尔扎、徐炳、戴维·沃德-法利、谢尔吉尔·奥泽尔、亚伦·库维尔和约舒阿·本吉奥。生成对抗网络。在NIPS，2014。[1](#_bookmark0)
  10. D.顾时代和E. J .德尔普。基于递归神经网络的深度伪造视频检测。在AVSS，第1–6页，2018。[1](#_bookmark0),[2](#_bookmark3)
  11. Alexandros Haliassos、Konstantinos Vougioukas、Stavros Petridis和Maja Pantic。嘴唇不会说谎:一种通用和鲁棒的人脸伪造检测方法。在CVPR，第5039–5049页，2021页。[2](#_bookmark3),[5](#_bookmark11),[6](#_bookmark14)
  12. 何、、、任、。用于图像识别的深度残差学习。在CVPR，第770–778页，2016。[6](#_bookmark14),[7](#_bookmark18)
  13. 、任丽、吴、和陈变阿来。DeeperForensics-1.0:用于真实世界人脸伪造检测的大规模数据集。在2020年的CVPR。[5](#_bookmark11),[7](#_bookmark18),[9](#_bookmark24)
  14. T.郑，s .金和k .金。Deepvision:使用人眼眨眼模式的Deepfakes检测。IEEE接入，8:83144–83154，2020。[1](#_bookmark0),[2](#_bookmark3)
  15. Tero Karras、Timo Aila、Samuli Laine和Jaakko Lehtinen。为了提高质量、稳定性和多样性而逐步种植甘蔗。在2018年的ICLR。[1](#_bookmark0)
  16. 泰罗·卡拉斯、萨穆利·莱恩和蒂莫·艾拉。一种基于风格的生成对抗网络生成器体系结构。在CVPR，第4401–4410页，2019。[1](#_bookmark0),[8](#_bookmark22)
  17. Ali Khodabakhsh、Raghavendra Ramachandra、Kiran Raja、Pankaj Wasnik和Christoph Busch。伪脸检测方法:可以推广吗？在BIOSIG，第1–6页，2018。[1](#_bookmark0),[2](#_bookmark3)
  18. 戴维斯·金。一个机器学习工具包。JMLR，10:1755–1758，2009年。[4](#_bookmark9)
  19. 迪德里克·P·金马和吉米·巴。亚当:一种随机优化方法。2015年在ICLR。[6](#_bookmark14)
  20. Prabhat Kumar、Mayank Vatsa和Richa Singh。检测视频中的人脸再现。2020年在WACV。[1](#_bookmark0)
  21. 、谢、、王中原和。单中心损失监督的频率感知鉴别特征学习用于人脸伪造检测。在CVPR，第6458–6467页，2021页。[2](#_bookmark3)
  22. 李凌志、鲍健民、杨昊、和。推进用于伪造检测的高保真身份交换。在2020年的CVPR。[5](#_bookmark11),[7](#_bookmark18),[9](#_bookmark24)
  23. 李凌志，鲍健民，，郝阳，，郭柏宁。面部x光，用于更一般的面部伪造检测。在2020年的CVPR。[1](#_bookmark0),[2](#_bookmark3),[3](#_bookmark5),[4](#_bookmark9),[5](#_bookmark11),[6](#_bookmark14),[9](#_bookmark24)
  24. Y.李，张，吕。在ictu oculi:通过检测眨眼来曝光人工制作的虚假视频。在2018年的WIFS。[2](#_bookmark3)
  25. 李月尊和吕四维。通过检测面部扭曲伪影来曝光深度伪造视频。2019年在CVPR举行的研讨会。[1](#_bookmark0),[2](#_bookmark3),[5](#_bookmark11),[6](#_bookmark14)
  26. 李月尊，，，祁红刚，吕四维。Celeb-df:用于深度伪造取证的大规模挑战性数据集。在2020年的CVPR。[1](#_bookmark0),[2](#_bookmark3),[5](#_bookmark11),[9](#_bookmark24)
  27. 、李晓丹、、、元和、、张伟明、俞能海。空间相位浅层学习:频域人脸伪造检测的再思考。在CVPR，第772-781页，2021。[2](#_bookmark3)
  28. 、张永、严俊池和刘威。基于高频特征的通用人脸伪造检测。

在CVPR，第16317-16326页，2021页。[2](#_bookmark3),[5](#_bookmark11),[6](#_bookmark14)

* 1. 、李晴、谢浩然、刘延刚、王振和斯蒂芬·保罗·斯莫利。最小平方生成对抗网络。在ICCV，第2813–2821页，2017。[1](#_bookmark0)
  2. 布拉斯·马丁内斯、马平川、斯塔夫罗斯·佩特里迪斯和马娅·潘蒂奇。使用时间卷积网络的唇读。

在ICASSP中，第6319–6323页，2020页。[6](#_bookmark14)

* 1. Iacopo Masi、Aditya Killekar、Royston Marian Mascaren- has、Shenoy Pratik Gurudatt和Wael AbdAlmageed。用于分离视频中deepfakes的二分支递归网络。

在ECCV，第667-684页，2020年。[5](#_bookmark11),[6](#_bookmark14)

* 1. Huy H. Nguyen，Fuming Fang，Junichi Yamagishi和Isao Echizen。用于检测和分割被操纵的面部图像和视频的多任务学习。在BTAS，2019年。[1](#_bookmark0),[2](#_bookmark3),[4](#_bookmark9)
  2. 阮惠赫，山本俊一和伊佐越前。使用胶囊网络检测虚假图像和视频。arXiv:1910.12467，2019。[2](#_bookmark3)
  3. 钱宇阳，，尹，卢生，，邵静。在频率中思考:通过挖掘频率感知线索的人脸伪造检测。在ECCV，第86-103页，2020年。[2](#_bookmark3)
  4. 亚历克·拉德福德，卢克·梅斯，和索史密斯·钦塔拉。深度卷积广义对抗网络的无监督表示学习。arXiv:1511.06434，2015。[1](#_bookmark0)
  5. 安德烈斯·罗斯勒、达维德·科佐利诺、路易莎·弗多里瓦、克里斯-田·里斯、贾斯特斯·提斯和马蒂亚斯·尼斯纳。学习检测被操纵的面部图像。在ICCV，2019年。[1](#_bookmark0),[2](#_bookmark3),[5](#_bookmark11),[9](#_bookmark24)
  6. Ekraam Sabir、Jiaxin Cheng、Ayush Jaiswal、Wael AbdAl- mageed、Iacopo Masi和P. Natarajan。视频中人脸操作检测的递归卷积策略。2019年在CVPR举行的研讨会。[1](#_bookmark0),[2](#_bookmark3)
  7. 谭明星和郭乐。反思卷积神经网络的模型缩放。在ICML，第6105–6114页，2019。[4](#_bookmark9),[7](#_bookmark18),[9](#_bookmark24)
  8. 胡斯图斯·提斯、迈克尔·佐勒霍·费尔和马蒂亚斯·尼纳尔。首选神经渲染:使用神经特征的图像合成。ACM TOG，2019。[2](#_bookmark3),[5](#_bookmark11),[9](#_bookmark24)
  9. J.提斯先生、佐勒霍·费尔先生、斯塔明格尔先生、西奥博尔特先生和内尔先生。Face2face:实时人脸捕捉和重现rgb视频。在CVPR，第2387-2395页，2016。[2](#_bookmark3),[5](#_bookmark11),[9](#_bookmark24)
  10. 罗青和刘燕。深度假冒检测人工智能模型的公平性检验。在IJCAI，第567-574页，2021。[2](#_bookmark3)
  11. 劳伦斯·范德·马腾和杰弗里·辛顿。使用t-sne可视化数据。JMLR，9(86):2579–2605，2008。[8](#_bookmark22)
  12. Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszko- reit、Llion Jones、Aidan Gomez、ukasz Kaiser和Illia Polosukhin。你需要的只是关注。在NIPS中，第5998–6008页，2017年。[6](#_bookmark14)
  13. 、费利克斯·许觉飞、、、谢、黄一浩、和。Fakespotter:一个简单而强大的识别人工智能合成假脸的基线。在IJCAI，第3444-3451页，2020。[2](#_bookmark3)
  14. 宣新生、博鹏、王巍和董晶。论gan图像取证的泛化。在CCBR，第134–141页，2019。[1](#_bookmark0),[2](#_bookmark3)
  15. X.杨，李，吕。使用不一致的头部姿势来揭露深层伪装。在ICASSP中，第8261–8265页，2019年。[2](#_bookmark3)
  16. 张寒、伊恩·古德菲勒、迪米特里斯·梅塔克萨斯和奥格斯-图斯·奥登纳。自我关注生成对抗网络。在ICML，第7354–7363页，2019。[1](#_bookmark0)
  17. 、周文波、、、张伟明、俞能海。多注意力深度防伪检测。2021年在CVPR。[1](#_bookmark0),[2](#_bookmark3)
  18. 赵天辰，，，徐，，，熊，。deepfake检测的学习自洽性。在ICCV，第15023–15033页，2021页。[1](#_bookmark0),[2](#_bookmark3),[3](#_bookmark5),[4](#_bookmark9),[5](#_bookmark11),[6](#_bookmark14)
  19. 郑英林、鲍健民、、曾鸣和。探索更一般的视频人脸伪造检测的时间一致性。在ICCV，第15044–15054页，2021页。[5](#_bookmark11),[6](#_bookmark14)
  20. 周，王，梁志远，沈建兵。在野外面对法医。2021年在CVPR。[2](#_bookmark3),[5](#_bookmark11),[6](#_bookmark14),[9](#_bookmark24)