通过提高 Deepfakes 检测的效率和鲁棒性精确的几何特征

Zekun Sun1 Yujie Han1 Zeyu Hua1 Na Ruan1 Weijia Jia1,2,3
上海交通大学计算机科学与工程系,上海,中国
北京师范大学人工智能与未来网络研究所 (BNU Zhuhai), 广东, 中国
北京师范大学-香港浸会大学联合国际学院人工智能与多模态数据处理重点实验室

抽象的

Deepfakes是恶意技术的一个分支,将目标人脸移植到视频中的原始人脸,造成侵犯版权、混淆信息甚至引起公众恐慌等严重问题。之前对 Deepfakes 视频检测的努力主要集中在外观特征上,这些特征有被如此复杂的操作绕过的风险,也导致高模型复杂性和对噪声的敏感性。此外,如何挖掘被操纵视频的时间特征并加以利用仍然是一个悬而未决的问题。我们提出了一个名为 LRNet 的高效且稳健的框架,用于通过对精确几何特征的时间建模来检测 Deepfakes 视频。

设计了一种新颖的校准模块来增强几何特征的精度,使其更具辨别力,并构建了双流递归神经网络（RNN)以充分利用时间特征。

与以前的方法相比,我们提出的方法重量更轻,更容易训练。此外,我们的方法在检测高度压缩或噪声损坏的视频方面表现出稳健性。我们的模型在 FaceForensics++ 数据集上达到了 0.999 AUC。同时,当面对高度压缩的视频时,它的性能会略有下降 (-0.042 AUC)。 1

一、简介

由于最近自动编码器和生成对抗网络 (GAN) [9] 的改进,合成视频变得前所未有的生动,人类或机器都难以区分。 Deepfakes 是其中最明目张胆的模型,它可以在视频中改变一个人的身份。由于面部视频包含敏感的个人信息,滥用这些技术将成为一种威胁。伪造演讲的来临巴拉克奥巴马[24]和篡改著名女演员的色情视频[27]在互联网上引起了极大关注。除了名人,普通人也可能成为 Deepfakes 的受害者,因为社交平台上有大量的视频剪辑和可免费获取的 Deepfakes 实现。因此,如何检测Deepfakes视频成为当务之急。

到目前为止,Deepfakes 检测方法大致可分为两类。第一种类型主要关注单个帧中的缺陷[22、15、21、18、25、23、16] 。

第二种类型考虑了时间特征[17, 2, 26]。上面提到的一些方法主要针对 Deepfakes 技术的非本质缺陷（例如异常眨眼或不同颜色的虹膜），这反过来又刺激了 Deepfakes 视频合成的改进。

在 Deepfakes 生成和检测技术之间的军备竞赛的背景下,需要遇到几个挑战。首先,先进的操纵方法促使检测器揭示 Deepfakes 视频的内在特征,这些特征不容易被伪装。其次,探测器应该更强大,使它们能够在现实世界的探测任务中表现出色。例如,许多模型[5, 1, 16]在压缩视频上出现了严重的性能下降,这降低了它们在应用中的有效性。第三,应考虑模型的简单性。当前的检测方法严重依赖强大的深度卷积神经网络 (DCNN) 或数据增强技能,这需要难以承受的培训成本。它们也不利于繁殖。

我们有一个关键的观察,虽然被操纵的面部视频在单帧中显示出高保真度,但它们仍然揭示了一些微妙但不自然的表情或面部器官的运动。这是Deepfakes技术的先天缺陷,因为伪造视频是逐帧生成的,对个人和个人都没有强限制。

1Github： https ://github.com/fredericksz k/LRNet

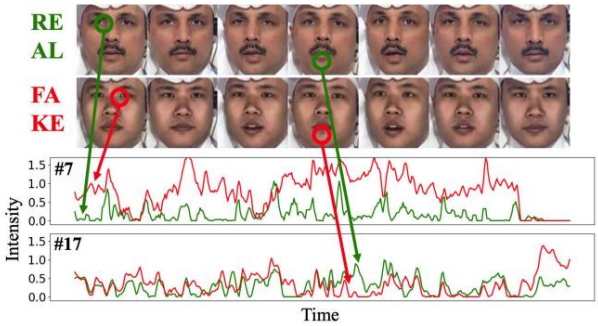

图 1.原始视频序列和 Deepfakes 视频序列的动作单元 (AU) 强度分析。所有表示构成面部表情的各个面部肌肉的运动。我们选择了两个最强烈的动作单元:#7（收紧盖子)和#17（下巴提升器）。正如我们所看到的,虽然假序列过于逼真以至于无法从外观上区分出来,但我们仍然可以在一些细微的表情上分辨出它们的差异,即使这两个视频中的面孔正在执行完全相同的动作。

行为模式和时间连续性（如图 1 所示）
1).为了更好地捕捉这些“时间伪影”,同时考虑到模型的稳健性和简单性,我们选择几何特征,例如面部器官的形状和位置。它们可以更明确地建模面部动态行为。面部地标是一组勾勒出标志性面部部位轮廓的点,足以描述几何信息并适用于我们的框架。

以前的工作已经展示了几何特征（尤其是面部标志)在暴露合成面部图像或视频中的潜力[33、34、2]。然而,他们利用手工制作的或复杂的基于相关性的特征进行分类,这对于捕获所有可区分的动态属性来说并不是最佳的。相比之下,我们设计了一个双流递归神经网络来提取地标序列上的深度时间特征。此外,他们都没有考虑不精确的面部特征点带来的影响,这可能不利于获得更有意义的特征。我们设计了一个新颖的地标校准模块,通过减少抖动来增强几何特征的判别能力,这使我们能够可靠地结合几何和深度时间特征,并构建我们的检测框架,称为地标递归网络（LRNet）。

我们的框架 LRNet 实现了优势互补。一方面,用地标替换人脸图像可以看作是一种有效的数据降维。与其他基于深度学习的模型相比,它不仅减少了模型冗余,而且对视频中的损坏更具不变性。另一方面,深度 RNN 有助于扩大特征空间并提升面部特征点的表达能力。它在成本和性能之间取得了更好的平衡。

我们的贡献可以总结为三个方面：

- 我们提出了一个有效且稳健的框架来对 Deepfakes 视频进行分类,我们在精确的几何特征上对时间特征进行建模。
- 我们引入了一种新颖的即插即用地标校准模块,以提高几何特征的精度和时间建模的有效性,同时使我们的框架更加灵活和可重现。
- 我们进行了大量实验来验证我们方法的效率和稳健性,并探索影响因素。

二、相关工作

2.1. Deepfakes检测

在这一部分中,我们介绍了目前在 Deepfakes 检测领域的进展。

帧级检测。到目前为止,deepfakes 检测的大部分工作都花在了基于单帧的方法上。其中一些技术基于手动选择的简单特征。例如,Matern 等人。 [21]专注于简单的视觉伪影,例如虹膜颜色、面部有线阴影以及眼睛和牙齿的缺失细节。其他人转向 DC NN 提取的深层特征。阿夫查尔等。 [1]提出了基于图像细观特性的MesoNet。罗斯勒等人。 [25]成功地将 Xception [5]转移到 deepfake 检测任务中。李等。 [16]利用一种名为 HRNet [28]的高级架构来检测 Deepfakes 操纵图像的混合边界。得益于强大的 CNN,这些方法以高成本换取了良好的性能。尽管如此,它们缺乏稳健性或难以重现。

视频级检测。最近,视频包含比图像更多信息的直觉极大地刺激了基于时间特征的 Deepfakes 检测。

一些基于几何特征的方案已经做出了有价值的尝试。李等。 [17]捕获了假视频中眨眼频率的异常。杨等。 [33]使用外部地标和中心地标分别组成头部方向和面部方向,并检测它们之间的不一致。手动选择的特征区分度较低,这限制了它们的性能,而我们转向利用富有表现力的深度时间特征。

基于外观的解决方案相对更为普遍。
G¨uera 等人。 [10]提出了一个框架,利用 CNN 从帧中提取特征,并使用 LSTM 处理特征序列。萨比尔等人。 [26]采用了类似的架构,但用双向门控电流单元 (GRU) 代替了 LSTM。这些方法也非常依赖 CNN,因此它们遇到与那些帧级检测器类似的问题。

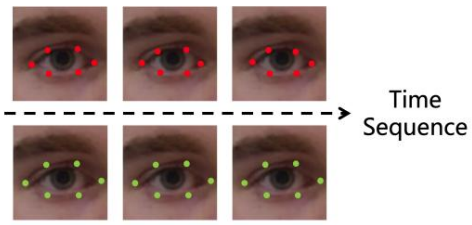

图 2. 准确度和精密度的比较。红色地标（上)准确但不精确。尽管它们都依附于轮廓,但它们抖动很大。绿色点（较低)不太准确但精确,可以更好地描述动态属性。

2.2.地标检测

面部标志是具有代表性和重要的几何特征。它的检测方法已被广泛研究多年。一开始,研究人员提出了主动外观模型（AAM） [6]和约束局部模型（CLM） [7]。之后,基于级联形状回归（CSR） [32、13]的检测方法取得了突出的进展。这些方法对地标位置进行初步估计,然后通过回归模型（例如,回归树)的集合迭代地改进它们。它们被广泛使用的开源图像处理存储库采用,如 Dlib [14],易于使用且检测速度快。最近,设计了丰富的基于深度学习的模型,例如 Cascade CNN [36]、卷积姿态机 (CPM) [31]、卷积专家 CLM [35]等。有些还通过开源工具包如 Openface [4] 实现].它们具有更好的性能,但速度较慢。此外,还引入了复杂的架构来解决面部遮挡、极端头部姿势等问题。

确保检测到的地标的准确性和精确度至关重要,因为它们是基于几何的 Deepfakes 检测中的决定性特征。具体来说,术语“准确”是指检测结果具有低偏差,而“精确”是指低方差（如图2所示）。精度相对更重要,因为抖动噪声会严重干扰时间建模。然而,目前的地标检测器大多在帧级运行,无法达到高精度。为此,我们设计了一个校准模块来提高地标检测结果的精度。

3. 方法论

我们提出的 Deepfakes 检测框架 LRNet 由四个部分组成（如图3所示）:人脸预处理模块、校准模块、特征嵌入程序和 RNN 分类程序。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。与大多数需要端到端训练的建议方法不同,我们的框架

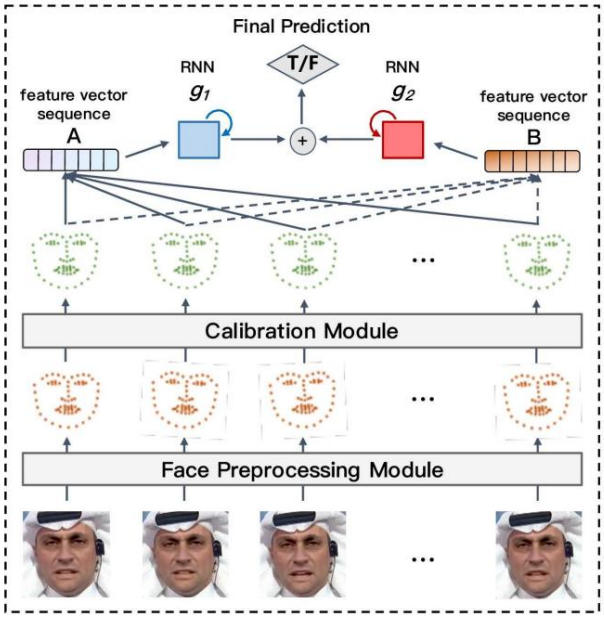

图 3. LRNet 检测框架概述。待检测的视频被分割成帧,并通过预处理程序以及精心设计的校准模块以获得一系列更精确的面部标志。随后的嵌入过程将地标嵌入到两种类型的特征向量中,并使用双流 RNN 来挖掘时间信息并判断其真实性。

只需要训练RNNs部分。我们的框架的细节将在下面展示。

3.1.人脸预处理

人脸预处理模块从人脸图像中提取几何信息。它由人脸检测、面部标志检测和标志对齐组成。

一开始,对视频的每一帧进行人脸检测,我们保留人脸的感兴趣区域（ROI）。裁剪出人脸图像后,我们在其上检测到 68 个面部特征点,这些特征点勾勒出面部的标志性轮廓。最后,我们将地标点对齐到通过仿射变换实现的预设位置。

请注意,我们的框架足够灵活,可以解耦预处理模块（更具体地说,地标检测器）。首先,地标检测器可以通过预训练模型实现,无需任何额外训练。其次,我们整个框架的性能并没有严重依赖地标检测器。此属性由我们提出的校准模块保证,这将在下面讨论。我们在4.3.1 节中通过实验对其进行了详细证明。

3.2.地标校准

通过预处理模块提取的地标基本可以满足精度要求。然而,他们

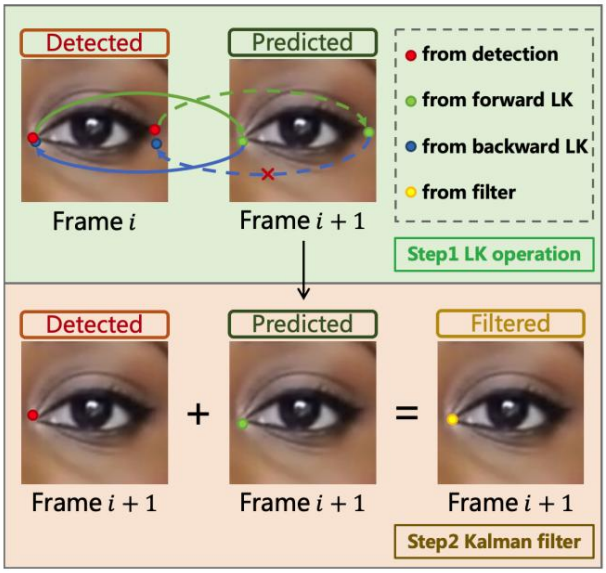

图 4. 校准模块的详细步骤。第一步使用 LK 操作来跟踪地标点。还执行前向后向检查以消除不精确的预测。

第二步使用卡尔曼滤波器合并检测和预测结果。

远非精确,因为它们是逐帧检测的。
从我们的观察来看,即使人脸几乎不动,检测到的地标也会有明显的抖动。因此,我们提出了一种新的校准模块来解决这个问题（如图4 所示）。我们使用连续的帧来预测基于 Lucas-Kanade 光流计算算法[20, 3] 的地标的后期位置。

并且通过定制的卡尔曼滤波器[12]将有效预测与其相应的检测结果合并以去噪并获得具有更高精度的校准地标。

3.2.1 跟踪

对于校准地标,我们的直觉在于我们可以通过匹配它们周围的小图像块来调整它们的位置。出于这个原因,Lucas-Kanade 算法是合适的,因为它计算光流,帧之间几个特征点的移动,本质上以与这种直觉相同的方式。受到 Baker 等人的作品的启发。 [3]和董等人。 [8],我们提出了一个金字塔 Lucas Kanade 操作（以下称为 LK 操作)来预测地标位置,换句话说,跟踪地标。

给定framei中以xi = [x, y] T 为中心的小图像块Pi ,其中来自 framei+1 的另一个相同大小的块Pi+ 1,我们尝试找到位移矢量d = [dx, dy]以最小化 P和Pi+1之差,则可以得到跟踪预测xi+1 = xi + d。
吨

因此,我们可以通过min计算出位移矢量d

缩小

$$\sum_{x\in\Omega} \alpha x \, Pi(x + \Delta d) - Pi+1(x + d) \quad , \qquad (1)$$

其中d首先被初始化为 [0, 0]T。从等式。 (1)我们可以求解Δd并通过以下方式迭代更新d

$$d \leftarrow d + \Delta d \qquad (2)$$

直到收敛。在等式中。 (1), Ω指patch中所有以xi为中心的位置的集合,
αx = exp(－远离中心的阳离子做软预测。
$\frac{x-xi}{2\sigma 2}$), 用于减少 lo 的权重

根据 Baker 等人的工作。 [3],我们获得了
方程式的最终解。 (1)即：

$$\Delta d = H - 1 \sum_{x\in\Omega} J(x)\alpha x(Pi+1(x + d) - Pi(x + 0))。 \quad (3)$$

在等式中。 (3), H = J TAJ ∈ R trix。 J $^{2\times 2}$ 是 Hessian ma
∈ R C|Ω|×2是通过垂直拼接生成的
J(x) ∈ R Pi(x$^{C\times 2}$ (x ∈ Ω),它是的雅可比矩阵
+ 0)。 C是Pi的通道数（RGB 图像为 3）。 A是一个对角矩阵,其元素由αx组成,对J中x对应的雅可比矩阵进行加权。

该方案的优点在于迭代过程中Pi(x)是常数,复杂的J和H-1只计算一次。

考虑到 LK 操作对 patch 的大小敏感,我们引入金字塔 LK 操作（在算法1 中有详细描述）,首先对图像进行多次下采样（通常将其大小减半)以构建它的金字塔表示,并执行简单的 LK 操作在具有相同补丁大小的不同大小的图像上。

注意到 LK 操作并不总是成功的,因此引入了前向后向检查,如图 3 所示。
4. 我们对前一帧执行前向 LK 操作（绿色箭头和点）,对从后一帧返回到前一帧的预测点执行后向 LK 操作（蓝色箭头和点）。其原始点与后向 LK 点差异较大的预测点将被丢弃（虚线箭头）。

3.2.2 去噪

我们从实验结果中发现,LK 操作也会带来噪声,从而扰乱地标的稳定性。因此,我们设计了一个定制的卡尔曼滤波器来整合来自检测和预测的信息,而不仅仅是 LK 运算结果。

卡尔曼滤波器估计最优地标点x 选择
在 framei+1 中通过相应的加权平均 我+1

算法 1:金字塔 LK 运算

输入:前一帧Fi ,后一帧Fi+1,前一帧待跟踪点xi

输出:字母框xi+1中的预测点

1构建Fi和Fi+1 的金字塔表示：

F i L L=0,...,Lm Lm ， 低频; _ i+1 L=0,...,Lm

2 g 3对于 L = Lm;L ⩾ 0;L ← [0, 0]T; d ← xi/2 L;

提取补丁P L 6计算P i的雅可比矩阵J和H；

号4 × 大 L ← [0, 0]T; 7 d for i = 1 :最大

5个 做 大号 来自F的L以x i为中心;一世

8个

9 从以i+1 i+1 + d L为中心的F中提取 大号 patch P；

号_大

10 计算Δ d 大号 通过方程式。 (3);

11 更新d 12结 大号 通过方程式。 (2);

束

13 g L−1 14结 ← 2(克 大号 +分升）；

束

15得到最终预测： d ← g 0 + d 0;

16返回： xi+1 ← xi + d;

算法 2:地标校准

输入： Li , Li+1, Fi , Fi+1

输出:校准的地标L^ 我+1

1对于xi ∈ Li do

2个 xi+1 ←算法1 (Fi , Fi+1, xi); x i ←算法1 (Fi+1,

3个 Fi , xi+1);

4个 用x i和执行前向后向检查

x~i ;

5个 如果xi有效,则/* 卡尔曼滤波器 */ 通过等式。 (7);

6个 计算博士 i+1

7 通过等式计算Ki+1 。 (5); opt估计

8个 x i+1 通过方程式。 (4);

9 通过等式更新Pi+2 。 (6);

10 xi^+1 ← x 选择 我+1;

11 else xi^+1 ←对应

12 的 xi+1 ∈ Li+1 ;结尾

13

14结束

15返回： L^ 我+1 = [1 68 xx i+1, ..., i+1] T;

地标检测结果x i+1 pred预测x det 和LK操作跟踪

我+1。 这个过程可以表示为：

$$x_{\text{选择 我+1}} = x_{\text{前 我+1}} + K_{i+1}\,(x_{\text{这 我+1}} - x_{\text{前 我+1}}), \tag{4}$$

其中Ki+1是估计x时的卡尔曼增益,可以通过以下方式计算： opt 我+1。 它

$$K_{i+1} = \frac{\pi_{+1}}{P_{i+1} + D_{i+1}}, \tag{5}$$

其中Pi+1是 LK pred操作预测x时的方差（表示不稳定性）， Di+1类似地标检测时检测x病房时的方差,我们通过以下方式更新下一个 LK 操作的方差Pi+2 我 + 1, 这之后。在i+1

$$P_{i+2} = (1 - K_{i+1}) \cdot P_{i+1} + Q, \tag{6}$$

其中Q是 LK 操作的固有方差。

然而,很难计算P和D ,因为 LK 操作和地标检测器都不是一个简单可表示的数学模型。因此,我们提出了近似相对方差Dr的概念：

$$\text{我+1博士} = \frac{x - x_{i+1}^{\text{det}} - x_{i+1}^{\text{pred}}}{x_{i+1}^{\text{det}}}. \tag{7}$$

由于连续帧中的每个ground-truth landmark点不会发生很大的偏移,如果检测结果有明显的振动， Dr会大于 1。然后我们在实验中根据视觉效果凭经验设置Q = 0.3 ,并在计算 Eq 时将D替换为Dr。 (5).

我们的校准模块依赖于 frame1 的地标来校准 frame2 的地标。然后frame2的这些优化的地标将用于校准frame3等。给定从帧Fi和Fi+ 1 中提取的lanmarks Li = [x和Li+1 ,地标校准的整体过程在算法 2 中有详细描述。为简单起见,我们只在其中表示一个校准步骤。 我，...,我 × 68] 吨

3.3.假视频分类

将上述步骤中提取和校准的地标序列嵌入到两种类型的特征向量序列中，然后输入到双流 RNN 中进行假视频分类。

每个地标点x , ya] T , 从而可以 可以用x表示 =

生成第一类特征向量αi embed [x a

来自地标Li = [x by: 我，...,我 × 68] 吨

$$\alpha_i = x_{i,y1}\text{我}, x_{2y2}\text{我}, \ldots, x_{68,y68}\text{我}, \tag{}$$

这可以看作是直接从Li扁平化。 那么第二类特征向量βi可以通过以下方式计算：

$$\beta_i = \alpha_{i+1} - \alpha_i$$

$$i+1\,i, \quad 1\,1 = x - x \quad \ldots, y68 \quad - y_{i+1}^{68}, \tag{}$$

表示连续帧之间地标位置的差异。

5个

通过嵌入我们得到两个特征向量 se 和B = [β1, ..., βn－1] T。根据 A = [α1, ..., αn－1] T，RNN 之间的另一个模型地标基于模型或上的面部形状运动模式，以及 B 上的唇部运动模式（或者可以看作是速度模式，用于捕获时间不连续性）。全连接层附加到每个 RNN 的输出以进行自己的预测，并且两个流被平均作为最终预测。我们将这些预测操作总结在一个函数f(·, ·) 中。因此，最终的预测，即视频剪辑的真实或虚假可能性，可以记为：

$$f(g_1(A), g_2(B))。 \tag{8}$$

为了执行视频级检测，每个视频样本被分割成固定长度的片段。剪辑的预测标签被聚合以用于视频的预测。

4. 实验

在本节中，我们首先声明实验设置。
然后我们在几个基准上评估我们提出的 LRNet 框架的效率。此外，我们分析了 LRNet 的影响因素。

4.1.实验设置

4.1.1 数据集

在 Deepfakes 的研究进展中，已经提出了几个具有挑战性的数据集。为了使评估具有代表性和全面性，我们选择了 4 个典型数据集。

UADFV [17]包含 49 个原始视频和 49 个经过处理的视频。它代表了早期的数据集，并被许多经典作品所采用。

FaceForensic++ (FF++) [25]包含 1000 个视频及其处理版本。每个视频都有原始版本 (raw)、轻微压缩版本 (c23) 和高度压缩版本 (c40)。它是最典型的近期数据集，已被广泛采用。

Celeb-DF [19]和 DeeperForensics-1.0 (DF1.0)[11]是两个新提出的具有高视觉质量的数据集。 Celeb-DF包含5639个假视频和540个真实视频，DF1.0包含1000个真实和对应的类似于FF++的假视频。每项工作还提供了一个有助于我们评估的基准。

4.1.2 参数及实现细节

在预处理步骤中，我们采用 Dlib [14]进行人脸和地标检测（另一个检测器 Openface [4]在消融研究中采用）。在分类过程中，我们的双流网络中的每个 RNN 都是双向的，由输出单元数设置为k = 64 的 GRU（门控循环单元)和两个全连接网络组成

方法	配置测试数据集大小八月训练	UADFV	FF++	Celeb-DF
中观 4 [1]	0.03 M ×未发布。	84.3	84.7	54.8
时间[18]	26 M √未发布。	97.4	80.1	56.9
DSP-FWA [18]	28 M √未公开。	93.0	97.7 80.4	64.6
Xception [25]	20.8 M × FF++ 胶囊[23]	99.7	61.3	48.2
	15 M × FF++ CNN+RNN	70.9	96.6 98.3	57.5
[26]	24.3 M × FF++ LRNet（我们的） 0.18 M ×			61.5
FF++		98.5	99.9	56.9

表 1. AUC socres (%) 在不同测试数据集上的一般性能评估。 “八月。”指该方法是否采用数据增强。我们提出的 LRNet 在模型大小上相对较轻，不需要数据增强，同时在 FF++ 上表现最佳。

单元数为 64 和 2 的层连接到 RNN 层的输出。在输入和 RNN 之间插入一个丢弃率dr1 = 0.25 的丢弃层，另外 3 个丢弃率dr2 = 0.5的丢弃层用于分离其余层。这些设置部分基于现有的研究结果[26]。此外，我们采用 8:2 数据集分割，即 80% 用于训练，20% 用于测试。

每个视频被分割成固定长度为 60 的片段，当 fps 为 30 时总计为 2 秒。我们采用lr = 0.001 的 Adam 优化器，批大小设置为 1024。将训练此分类模型到 500 个纪元。

4.2.绩效评估

4.2.1 总体评价

我们首先基于 Celeb-DF 基准[19] 对 LRNet 进行一般评估。由于目前提出的检测方法很大一部分没有开源，难以复现，我们遵循Celeb-DF benchmark的评估设置，在一个数据集（主要是FF++)上训练我们的模型，并在不同的数据集上进行测试.

评估指标是 AUC 分数（ROC 曲线下的面积），结果如表1 所示。我们只展示了部分最佳性能方法的结果。我们的方法在其训练数据集 FF++ (99.9) 上获得了几乎完整的 AUC 分数，表明它可以有效地捕获异常运动和不连续性。此外，它还可以推广到其他数据集（未见过的样本）。

4.2.2 视频压缩的鲁棒性

我们进一步测试了我们的方法对视频压缩的稳健性，这被大多数当前工作所忽视。
在 FF++ 上，我们比较了其基准测试中最好的检测器 Xception，以及新提出的和先进的 X-Ray [16]。每个检测器都在原始视频（raw)上进行训练，并在具有不同压缩率的三个版本的视频上进行测试。在 Celeb-DF 上，我们使用了它的基准设置

Xcpetion 用 FF++(c23) 训练,FWA(DSP-FWA) 用数据增强训练。而我们的方法直接在 FF++(raw) 上训练。结果如表2 所示。我们可以从结果中得出,我们的方法的性能对视频压缩相对更具不变性。

方法	FF++			衰退
	生的	c23	c40	
异常[25]	99.7	93.3	86.5	6.4/13.2
X 射线[16]	99.1	87.3	61.6	11.8/37.5
LRNet（我们的）	99.9	97.3	95.7	2.6/4.2
方法	名人-DF			衰退
	生的	c23	c40	
Xception-c23 [25]	65.3	65.5	52.5	-0.2/12.8
时间[18]	56.9	54.6	52.2	2.3/4.7
DSP-FWA [18]	64.6	57.7	47.2	6.9/17.4
LRNet（我们的）	56.9	56.3	55.4	0.6/1.5

表 2. 遇到视频压缩时不同方法的 AUC 分数 (%)。

4.2.3 对视频噪声的鲁棒性

我们还在被不同噪声破坏的视频上挑战我们提出的 LRNet。我们选择适合本次评测的DF1.0 benchmark。因为它提供了此设置并测试了其他基准测试忽略的各种高级视频级检测方法。结果如表3 所示。我们可以看到所有方法在相同的训练和测试数据集上都表现良好（包括我们的 LRNet,它达到了 97.74% 的准确率和 99.2% 的 AUC）。而我们的方法在面对噪音时性能下降最少。

方法	训练/测试		衰退
	标准/标准	标准/噪音	
C3D [29]	98.50	87.63	10.87
TSN [30]	99.25	91.50	7.75
I3D [29]	100.00	90.75	9.25
CNN+RNN [26]	100.00	90.63	9.37
异常[25]	100.00	88.38	11.62
LRNet（我们的）	97.74	96.83	0.91

表 3. 不同方法对视频噪声的鲁棒性比较,可以通过二分类精度 (%) 来评估。 “std”是指干净的样本，“噪声”是指具有基准[11]中描述的几种强噪声的样本。

4.2.4 培训效率

为了更好地展示我们框架的效率,我们对几种具有代表性的基线外部方法进行了评估,并将结果显示在表4 中。所有模型都需要

模型	预处理	训练
	操作时间#Param GPU 磁盘时间	
异常[5]	F+L+A 6h 20.8M 12G 64G 21h	
X射线[16] F+L+A+D 24h 37.7M >12G >180G >30h CNN+RNN[26] F+L+A 6h 24.3M 9G 64G 22.5h F+L+A+O 10h 22.5M >12G >120G >30h		
天网[30]		
LRNet（我们的)F+L+A+C 8h 0.18M 3G 1.1G 0.2h		

表 4. 训练成本的定量比较（在 FF++ 上）。这些操作包括人脸检测（F）、地标检测（L）、对齐（A）、数据增强（D）、光流（O)和建议的地标校准（C）。 #Param是指模型的可训练参数大小。我们还评估了 GPU 内存占用 (GPU) 和训练数据占用的磁盘内存 (DISK)。

	train	Dlib	
test	calib	Y	N
OpenFace	Y	98.5	85.5
	N	81.6	90.2

	train	OpenFace	
test	calib	Y	N
Dlib	Y	98.2	90.6
	N	71.6	78.4

图 5. 在一种地标上训练并用另一种地标进行测试的混淆矩阵。我们在 FF++(raw) 上评估每个设置的准确性 (%)。 “Y”表示使用我们的校准模块检测地标,而“N”表示不检测。

$\times 10^{-2}$	jaw	eyebrow (left)	eyebrow (right)	nose	eye (left)	eye (right)	lips
raw	5.87	3.26	3.45	2.49	1.16	1.15	3.67
raw+calib	5.34	3.04	3.25	2.21	0.89	0.89	3.08

图 6. 不同检测器（此处为 Dlib 和 OpenFace)检测到的地标之间的平均距离。 “+calib”表示使用我们的校准模块。我们将 68 个地标按它们所属的器官合并为 7 组。

类似的预处理操作和我们提出的 LRNet 消耗了可接受的时间,即仅比基本要求（6 小时)多 2 小时。虽然我们的模型在训练中速度明显更快,内存占用更少。

4.3.框架分析

4.3.1 校准模块的作用

地标标定是 LRNet 的核心组成部分。我们对其进行消融研究（如表5所示）。我们可以看到校准模块提高了检测框架的整体性能并保持了其鲁棒性。

我们在评估中更进一步。首先,我们尝试使用性能更好的基于 CNN 的检测器 Openface [4] 来检测地标,并重新训练 LRNet 的 RNN 部分。

性能与使用结果非常相似

方法	FF++（原始）		FF++(c40)	
	累计 AUC	累计 AUC		
网络	99.7	99.9	91.2	95.7
不带卡尔曼滤波器	98.5	99.4	87.2	94.3
无校准	92.8	97.4	84.3	92.6

表 5. LRNet 中校准模块的消融研究,通过评估二进制检测精度 (%) 和 AUC 分数 (%)。所有模型仅在 FF++(raw) 上训练。

方法	FF++（原始）		FF++(c40)	
	累计 AUC	累计 AUC		
LRNet (g1 + g2)	99.7	99.9	91.2	95.7
g1	83.4	89.3	80.4	86.1
g2	98.3	99.2	85.2	93.9

表 6. 通过评估二进制检测精度 (%) 和 AUC 分数 (%) 对网络架构进行消融研究。 g1指的是模拟异常面部运动的 RNN， g2指的是另一个捕捉时间不连续性的 RNN。

Dlib,我们这里就不一一列举了。然后我们探索不重新训练模型的影响,即,将 Openface 地标馈送到由 Dlib 地标训练的 LRNet 中,反之亦然。如图5所示，只有借助校准模块才能避免性能下降。

原因在于通过校准的地标训练的模型可以更好地捕获异常的面部运动,而不是地标检测器带来的噪声。我们通过计算来自不同检测器的地标之间的差异来进一步证明这一点,如图 6 所示。校准模块有助于缩短不同地标检测的差距,使它们更接近地面真实位置。

4.3.2 网络架构的影响

我们通过将其与仅使用其一个流进行比较来证明我们的双流 RNN 架构的有效性（如表6 所示）。结果表明该结构具有优越的能力。来自两条道路的信息相互促进,其中g2检测到的时间不连续性线索对最终准确性的贡献更大， g1识别的异常运动消息为预测提供了更强的鲁棒性。

4.3.3 输入长度的影响

输入长度是指我们将其作为单个样本输入模型的连续帧数。我们在控制其他条件的情况下评估不同的输入长度（如图7 所示）。尽管在同一数据集上训练和测试时输入长度几乎不影响性能,但合适的输入长度（我们采用的 60)可以提高两者

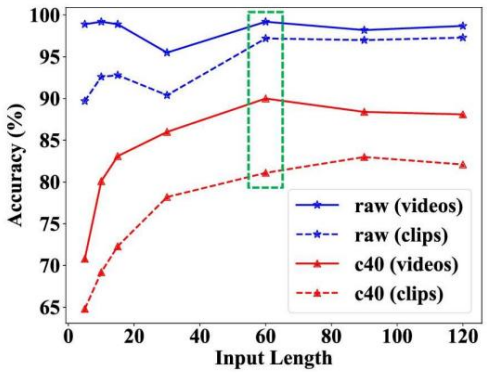

图 7. 输入长度变化时 LRNet 的精度 (%)。
所有模型都在 FF++(raw) 上训练。 “（剪辑）”指的是样本检测精度，“（视频）”是视频级分类的结果。

不同数据分布（例如,压缩样本)的有效性和鲁棒性。

5. 讨论

在本节中,我们首先讨论我们的局限性和未来的工作。从一般的评估结果来看,我们的框架的通用性仍有改进的空间。此外,很难解释我们的模型捕捉到的时间特征,也很难将真人脸和假人脸之间的运动模式差异可视化。未来我们将对不同人脸操作技术的这些异常动态模式进行更深入的研究和分析,进而提升模型的普适性。

我们还阐述了外观和几何特征的影响。从目前的工作结果来看,外观特征是高维的,更具有泛化的表现力,但鲁棒性差且成本高。虽然几何特征更健壮、成本更低,但更难泛化。所以这是性能（尤其是泛化能力)和成本的权衡。虽然我们努力提高仅依靠几何特征的Deepfakes检测效率,但我们是否可以将外观和几何特征结合在一起同时避免高成本来提高效率值得探索。

六,结论

Deepfakes 是对人类社会的巨大威胁,其快速发展需要有效的解决方案。我们的工作表明,面部标志和时间特征的整合可以成为对 Deepfakes 的快速而稳健的测试。我们还探索了如何增强地标检测结果并充分利用时间特征。我们发现面部几何信息及其动态特征是必不可少的线索,值得在未来的工作中探索,以实现更高效、更稳健的野外 Deepfakes 检测。

参考

[1] Darius Afchar、Vincent Nozick、Junichi Yamagishi 和 Isao Echizen。 Mesonet:一个紧凑的面部视频伪造检测网络。在 IEEE 国际信息取证与安全研讨会上,第 1-7 页,2018。1、2、6 [2] Shruti Agarwal、Hany Farid、Yuming Gu、Mingming He、Koki Nagano 和 Hao Li。保护世界领导人免受深度造假。在 IEEE/CVF 计算机视觉和模式识别研讨会会议上,第 38–45 页,2019年。1、2

[3] 西蒙·贝克和伊恩·马修斯。 Lucas-kanade 20 年:一个统一的框架。国际计算机视觉杂志,56(3):221–255, 2004. 4

[4] Tadas Baltrusaitis、Amir Zadeh、Yao Chong Lim 和 Louis Philippe Morency。 Openface 2.0:面部行为分析工具包。在 IEEE 自动人脸和手势识别国际会议上,第 59–66 页,2018年。3、6、7 [5] Frank ois Chollet。 Xception:具有深度可分离卷积的深度学习。在 IEEE/CVF 计算机视觉和模式识别会议上,第 1251–1258 页， 2017年。1、2、7

[6] Timothy F. Cootes、Gareth J. Edwards 和 Christopher J. 泰勒。活跃的外观模型。 IEEE 模式分析和机器智能汇刊,23(6):681–685, 2001. 3

[7] 大卫·克里斯蒂纳奇和蒂莫西·F·库特斯。使用受限局部模型进行特征检测和跟踪。 In British Machine Vision Conference, pages 929–938, 2006. 3 [8] Xuanyi Dong, Shoou-I Yu, Xinshuo Weng, Shih-En Wei, Yi Yang, and Yaser Sheikh. Supervision-by-registration:一种提高面部地标检测器精度的无监督方法。在 IEEE/CVF 计算机视觉和模式识别会议上,第 360–368 页,2018年。4 [9] Ian Goodfellow、Jean Pouget-Abadie、Mehdi Mirza、Bing Xu、David Warde-Farley、Sherjil Ozair、Aaron Courville 和约书亚·本吉奥。生成对抗网络。 In Advances in neural information processing systems, pages 2672–2680, 2014. 1

[10] 大卫·古埃拉和爱德华·J·德尔普。使用递归神经网络的 Deepfake 视频检测。在基于高级视频和信号的监视的 IEEE 国际会议上,第 1-6 页,2018年。2 [11] Liming Jiang、Ren Li、Wayne Wu、Chen Qian 和 Chen Change Loy。 Deeperforensics-1.0:用于真实世界人脸伪造检测的大规模数据集。在 IEEE/CVF 计算机视觉和模式识别会议上,第 2889–2898 页,2020年。6、7

[12] 鲁道夫·埃米尔·卡尔曼等人。线性滤波和预测问题的新方法[j]。基础工程学报, 82(1):35–45, 1960. 4

[13] 瓦希德·卡泽米和约瑟芬·沙利文。一毫秒面部对齐与一组回归树。在 IEEE/CVF 计算机视觉和模式识别会议上,第 1867–1874 页,2014 年。3 [14] Davis E King。 Dlib-ml:机器学习工具包。

机器学习研究杂志, 10:1755–1758, 2009. 3, 6

[15] 李浩东,李斌,谭顺全,黄继武.使用颜色分量的差异检测深度网络生成的图像。 arXiv 预印本 arXiv:1808.07276, 2018. 1

[16] 李令志,鲍建民,张婷,杨浩,陈东,文芳,郭百宁。用于更一般的面部伪造检测的面部 X 射线。 In IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5001– 5010, 2020. 1, 2, 6, 7 [17] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In ictu oculi:通过检测眨眼来暴露人工智能生成的假脸视频。 arXiv 预印本 arXiv:1806.02877, 2018. 1, 2, 6

[18] 李月尊, 吕四维.通过检测面部扭曲伪影来暴露 deepfake 视频。 In IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pages 46–52, 2019. 1, 6, 7 [19] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df:用于深度伪造取证的大规模挑战性数据集。在 IEEE/CVF 计算机视觉和模式识别会议上,第 3207–3216 页,2020 年。6 [20] Bruce D Lucas、Takeo Kanade 等。一种应用于立体视觉的迭代图像配准技术。

在人工智能国际联合会议上,第 674–679 页,1981 年。4 [21] Falko Matern、Christian Riess 和 Marc Stamminger。利用视觉伪像来揭露深度伪造和面部操纵。 In IEEE Winter Applications of Computer Vision Workshops, pages 83–92, 2019. 1, 2 [22] Scott McCloskey 和 Michael Albright。使用颜色提示检测 gan 生成的图像。 arXiv:1812.08247, 2018.1 [23] Huy H Nguyen、Junichi Yamagishi 和 Isao Echizen。

arXiv 预印本

胶囊取证:使用胶囊网络检测伪造的图像和视频。在 IEEE 声学、语音和信号处理国际会议上,第 2307–2311 页,2019.1、6 [24] Aja Romano 等人。乔丹·皮尔 (Jordan Peele) 模拟的奥巴马公益广告是对假新闻的双刃剑警告。澳大利亚警务,10(2): 44,2018。1 [25] Andreas R¨ossler、Davide Cozzolino、Luisa Verdoliva、Christian Riess、Justus Thies 和 Matthias Nießner。 Faceforen sics++:学习检测被操纵的面部图像。 arXiv 预印本 arXiv:1901.08971, 2019.1 , 2, 6, 7 [26] Ekraam Sabir、Jiaxin Cheng、Ayush Jaiswal、Wael AbdAl mageed、Iacopo Masi 和 Prem Natarajan。视频中人脸操纵检测的循环卷积策略。

在 IEEE/CVF 计算机视觉和模式识别研讨会会议上,第 80-87 页,2019 年。1、2、6、7 [27] Russell Spivak。 “deepfakes”:犯下最古老罪行之一的最新方法。乔治城法律技术评论,3(2):339–400, 2019. 1

[28] Ke Sun, Yang Zhao, Borui Jiang, Tianheng Cheng, Bin Xiao, Dong Liu, Yadong Mu, Xinggang Wang, Wenyu Liu, and Jingdong Wang. High-resolution representations for labeling pixels and regions. arXiv preprint arXiv:1904.04514, 2019. 2

[29] Du Tran、Lubomir Bourdev、Rob Fergus、Lorenzo Torresani 和 Manohar Paluri。使用 3d 卷积网络学习时空特征。 In IEEE International Conference on Computer Vision, pages 4489–4497, 2015. 7 [30] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool.时间段网络:走向深度动作识别的良好实践。在欧洲计算机视觉会议上,第 20-36 页,2016 年。7 [31] Shih-En Wei、Varun Ramakrishna、Takeo Kanade 和 Yaser Sheikh。卷积姿势机。在 IEEE/CVF 计算机视觉和模式识别会议上,第 4724–4732 页,2016 年。3

[32] 熊雪涵和费尔南多·德拉托雷。监督下降法及其在人脸对齐中的应用In IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 532–539, 2013. 3 [33] Xin Yang, Yuezun Li, and Siwei Lyu.使用不一致的头部姿势暴露深度假货。在 IEEE 声学、语音和信号处理国际会议上,第 8261–8265 页,2019。2 [34] 杨鑫、李悦尊、齐宏刚和吕思伟。使用地标位置暴露 gan 合成的人脸。 arXiv 预印本 arXiv:1904.00167, 2019.2 [35] Amir Zadeh、Yao Chong Lim、Tadas Baltrusaitis 和 Louis Philippe Morency。卷积专家限制了用于 3d 面部标志检测的局部模型。在 IEEE/CVF Con

关于计算机视觉和模式识别研讨会的参考文献,第 2519–2528 页,2017 年。3

[36] Erjin Zhou, Haoqiang Fan, Zhimin Cao, Yuning Jiang, and Qi Yin. Extensive facial landmark localization with coarse to-fine convolutional network cascade. In IEEE/CVF Con

计算机视觉和模式识别研讨会的参考文献,第 386–391 页,2013 年。3