# 通过基于包装类神经网络的光流的深度伪造视频检测

艾琳·阿梅里尼（Irene Amerini），莱昂纳多·加特里（Leonardo Galteri），罗伯托·卡尔德利（Roberto Caldelli），阿尔贝托·德尔本比（Alberto Del Bimbo）

意大利佛罗伦萨大学媒体整合与传播中心（MICC）

国家大学间电信联盟（CNIT），意大利帕尔马

## 摘要

视觉媒体技术的最新进展已导致进行处理的新工具，最重要的是生成多媒体内容。特别是基于现代AI的技术提供了易于使用的工具，可创建极为逼真的可操纵的视频。这样的合成视频，名为*Deep Fakes*，可能对公开主题的声誉或解决一般在某个事件上的观点构成严重威胁。 据此，能够个体化这种虚假信息变得至关重要。在这项工作中，一种新的法医技术能够辨别给出的假视频和原始视频的序列；不像其他依靠单个视频帧的最新方法，我们建议采用光流场来探索可能的帧间差异。这样的线索将被用作作CNN分类器要学习的特性。初步的结果在FaceForensics ++数据集上获得了非常有前途的表现。

## 1.简介

深度学习技术正在使有关创建和处理多媒体内容的技术不断升级。被称为*Deep Fakes*（DF）的一种新现象，最近出现了：它允许非常简单地在人们面部的地方或有时仅在嘴唇和眼睛的动作的地方创建真实的视频，它们被修改为可能在一定条件下模拟另一个特定文字主题的存在或使某人与其他人连贯说话，并且可能会损害演讲。效果是可以想象的，当这些伪造的信息故意用来伤害诸如公众人物或一个政治家，甚至是一个政党之类的组织。深度假冒的影响也可以通过快速传递信息的社交网络到全世界被放大。据此，机器学习社区为此现象付出了特殊而双重的关注。一方面，我们花费了很多精力来开发新型有效的合成视频生成技术，例如Face2Face，Deep Video Portraits(深度视频人像)，StarGAN [ 5 ]和Deep Fake 1。从另一面看，各种最新的研究集中在发现仿冒的影片问题上；他们大多数通过分析视频 RGB的帧中的不一致可能。通常，成熟的和经过预训练的CNN技术直接应用于从每个序列的单个帧学习独特的功能。在[ 11 ]中，复发性卷积传统策略用于人脸检测，那里一组框架被视为一个整体。其他方法考虑物理特征，例如工作在作者建议检测眨眼暴露生成的假脸视频，并在[ 2 ]中显示对表情进行建模以从自然模式区分假说话模式。

在此扩展摘要中，介绍了能够从原始视频中检测类似Deepfake的视频的一种新技术。特别是，与通常基于帧的方式的最新方法不同，我们展示了基于序列的，致力于调查在视频的时间结构可能存在差异的方法。具体来说，光流已被提取，来开发被用作CNN分类器输入的帧间关联。

论文结构如下：第2节，通过讨论向量字段运动的用法，介绍了提出的方法，而第3节讨论了一些初步的实验结果；最后，第4节得出结论。

## 2.提出的方法

在本节中，提出的方法的基本架构是如图1所示。这样的结构已经建立，以了解optical flow fields(光学流场)实际的效果，以区分深层伪造和原始视频。光流是一个矢量场，是在两个连续帧f（t）和f（t +1）上计算得出，提取观察者和场景本身之间的视在运动。特别地，我们的假设是光流是能够发现，跨帧（通过摄像机与自然产生的那些一起产生的帧）同步的运动差异。它应该可感知：光流矩阵，引入假冒和异常的嘴唇、眼睛以及整个脸部的运动。因此，由于这个原因，对于每个帧f（t），在特定时间t处，使用关于光流的CNN模型称为PWC-Net [ 13 ]，提取前向流量OF（f（t），f（t + 1））。该技术基于金字塔处理和扭曲，还基于通过CNN本身来估计光流量来处理的成本量的使用。依次（见图1），计算得出的前向流量OF（f（t），f（t + 1））作为半流动CNN（称为Flow-CNN，基于一些预先训练的网络）的输入。在我们的实验中，我们已经测试了VGG16 [ 12 ]和ResNet50 [ 6 ]作为主干。

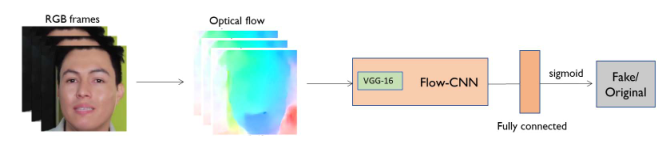


Figure 1. The proposed architecture.

图1.建议的体系结构。

由于尺寸（以样本数量计），对于Deepfake识别任务的常见数据集是不足以从头训练网络，我们仅在一部分网络上采用了迁移学习技术，网络的其余部分在deepfake数据集上进行了微调。所以在微调期间，网络的第一层已经被修复，当最后的卷积层与密集层训练有素。之后，为了二进制分类，为了每一帧（假的或原始的），最后的带有一个输出单元（跟随着一个S型激活作用）的全连接层， 放在网络的末尾。为了发现有措施和培训原始RGB图像的预训练网络，光流转换为使用固定的颜色编码方法的3通道图像。像素颜色由流矢量之间的角度和水平轴确定，而运动的强度是由颜色的饱和度编码。

## 3.初步实验结果

在本节中，介绍一些初步的实验结果，来评估提议的概念的好处。在图2中，一个光流场在原始视频（左侧）和对应Deepfake（右侧）的两个连续帧上进行计算的分别拍照以提供视觉检查的示例。如预期，可以注意到，与更改后的看起来更加流畅的视频相比，实际序列的下巴周围的运动向量噪声较多。在此基础上，我们试图验证这种线索是否可以由神经网络正确学习。为此，经过随机左右翻转，在更大的包含面部的300×300 像素的平方块随机选取的，大小为224×224像素的平方块，通用网络已被训练；为了训练，我们使用10-4的 Adam优化器，

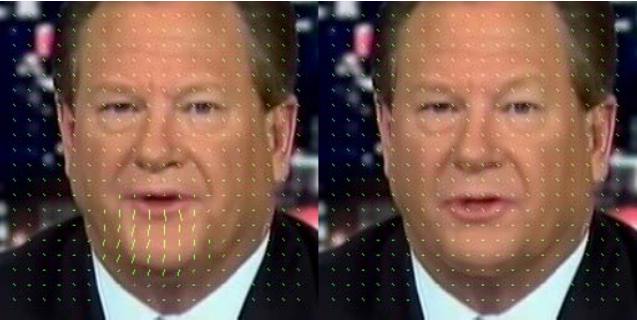


Figure 2. Optical flow for original (left) and deepfake (right) videos.

图2.原始（左）和深度作假（右）的光流

默认动量值，批量大小为256。我们在[ 10 ]中提出的FaceForensics ++数据集上运行实验；该集合由1000个原始视频序列组成，序列通过三个自动面部操作方法(Deepfake，Face2Face和FaceSwap)进行操作。720个视频用于培训，120个用于验证，其它120个进行测试。

|  |  |  |
| --- | --- | --- |
|  | VGG16 | ResNet50 |
| Face2Face | 81.61％ | 75.46％ |

表1.我们的Face2Face操作体系结构的二进制检测精度（％）

在两个已实施的网络（VGG16和ResNet50）的初步结果在表1中已经列出。他们是在FaceForensics ++的整个测试集上获得的，为了操纵Face2Face并见证该方法是能够区分两种视频。

## 4.结论

在这项工作中，介绍和研究了利用光流场不同作为作为区分Deepfake视频和原始视频的线索的想法。这是一种非常创新的尝试，它考虑到了序列在时间维度上的明显异常。在此初始实验中，要解决问题要使用预训练网络，运动矢量已表示为3通道图像，然后作为新神经网络的输入。在具有不同类型的网络的FaceForensics ++数据集上获得的初步结果，很有前途，并证明这种功能似乎是能够指出在这两个分析案例之间存在的一些不同质性。该证据为许多可能的未来工作做了铺垫：首先，它通过针对更多数据集和其与他神经网络的测试，评估流场对伪造视频的识别；其次，研究如何使，这种开发关于暂时的轴不一致的方法，能够与众所周知的最新基于帧的方法结合使用，以改善其性能，应该是有趣的。

## 参考文献

[1] D.Afchar, V.Nozick, J.Yamagishi, andI.Echizen. Mesonet: a compact facial video forgery detection network. pages 1–7, 12 2018. 1

[2] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li. Protecting world leaders against deep fakes. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2019. 1

[3] L. Alparone, M. Barni, F. Bartolini, and R. Caldelli. Regularization of optic flow estimates by means of weighted vector median filtering. IEEE Transactions on Image Processing, 8(10):1462–1467, Oct 1999. 1

[4] S. S. Beauchemin and J. L. Barron. The computation of optical flow. ACM Comput. Surv., 27(3):433–466, Sept. 1995. 1

[5] Y. Choi, M. Choi, M. Kim, J. Ha, S. Kim, and J. Choo. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. CoRR, abs/1711.09020, 2017. 1

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016. 2

[7] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Niessner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt. Deep video portraits. ACM Trans. Graph., 37(4):163:1–163:14, July 2018. 1

[8] Y. Li, M. Chang, and S. Lyu. In ictu oculi: Exposing AI generated fake face videos by detecting eye blinking. CoRR, abs/1806.02877, 2018. 1

[9] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. Faceforensics: A large-scale video

dataset for forgery detection in human faces. CoRR, abs/1803.09179, 2018. 1

[10] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. Faceforensics++: Learning to detect manipulated facial images. CoRR, abs/1901.08971, 2019. 1, 2

[11] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan. Recurrent convolutional strategies for face manipulation detection in videos, 05 2019. 1

[12] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv 1409.1556, 09 2014. 2

[13] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018. 2

[14] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Niessner. Demo of face2face: Real-time face capture and reenactment of RGB videos. In ACM SIGGRAPH 2016 Emerging Technologies, SIGGRAPH ’16, pages 5:1–5:2, New York, NY, USA, 2016. ACM.