# 使用不连续的头部姿势暴露深层造假

辛阳，李月尊和吕思薇

美国纽约州立大学，奥尔巴尼大学

## 摘要

在本文中，我们提出了一种新的方法公开AI 生成伪造的人脸图像或视频（众所周知的Deep Fakes）。我们的方法基于观察：假冒伪劣作品是通过将合成人脸重新拼接成为原始图像，并在此过程中介绍，从面部图像中，当3D头部姿势被估计时，可以显示的错误。我们进行实验演示以论证这种现象并进一步发展基于此提示的分类方法。使用基于此提示的功能，使用一组真实面部图像和深造假评估SVM分类器。

索引字词-媒体取证，DeepFake检测，头姿势估计

## 1.引言

由于机器学习的最新发展，用于处理和伪造图像和视频的技术已达到新的复杂高度[1、2、3，4，5，6，7]。这种趋势的前沿是所谓的深度造假，就是通过插入使用深度神经网络合成的面孔到原始图像/视频而创建。与其他形式的通过数字社交网络共享的错误信息一起，Deep Fakes创建的数字模仿已经成为一个严重的有消极社会影响[8]的问题。因此，有效的公开伪造品的方法是迫切需要的。

迄今为止，Deep Fakes的检测方法已经依赖关于合成算法伪影或固有的不一致性，例如，算法缺乏真实的眨眼[9]，颜色配置不匹配[10]和说话时的的视觉嘴唇[11]。基于分类方法的神经网络也已被用来从 “深造假”[12、13、14]直接识别真实图像。在这项工作中，我们提出了一种新的方法来检测深造假。我们的方法基于深度神经网络人脸综合模型的内在局限性，这是Deep Fake产品管道的核心组成部分。具体来说，这些算法会创建不同人的面孔，但保持原始人的表情。但是，这两张脸的面部区域标记（即人脸上与重要的结构相一致的地方，例如眼和嘴尖端，因为神经网络综合算法不能保证原始脸和合成脸具有一致的脸部地标，如图1所示。）不匹配。

地标位置的错误可能在人眼不能直接看到，但可以从头部姿势中看出来（即头部的方向和位置），它是根据在脸部的真实和伪造部位上的2D地标估算的。具体来说，我们对 使用所有面部标志估计的头部姿势 和 仅使用中心区域估计的那些 进行比较，如图1所示。原理是两个估计的头部姿势接近于原始脸部，图1（k）。但是对于深度造假，由于中央脸部区域来自合成脸部，由原始地标位置与上述生成的图像位置不匹配造成的误差，在两个估计的头部姿势之间，会导致更大的差异，图1（n）。我们实验上证实了在深度造假中的头部姿势估算的显着差异。然后我们用在估计的头部姿势中的差异，作为特征向量来训练简单的基于分类器的SVM，以区分原始和深度造假。逼真的Deep Fake视频实验表明了我们算法的有效性。

## 2. 深度造假生产线

制作“伪造品”的总体过程在图1（a）-（h）。要生成“深造假”，我们给算法一幅包含源面将被替换的图像（或视频中的帧）。此脸的边界框可由面部检测器得到，然后检测面部地标。脸部区域变形为标准配置，通过仿射变换M（通过最小化面部中心标志的对准误差（图1（c）中的红点）到一组标准地标位置得到），这是一个称为人脸对齐的过程。然后将该图像裁剪为64×64像素，并馈入深度生成神经网络来创建一张合成的脸。合成的脸用M-1向后转换以匹配原始面部。最后，通过边界平滑等后处理，Deep Fake图像/视频帧就会被创建。

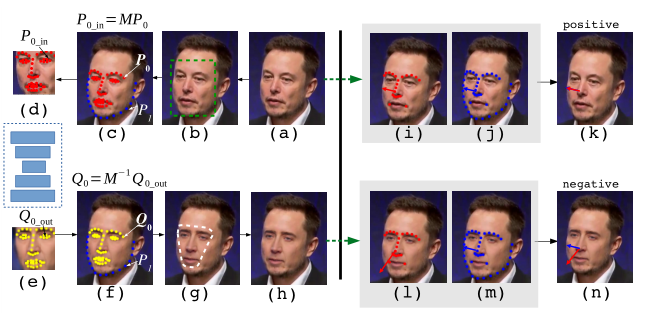


图1. Deep Fake工作流程概述（左）和我们的方法（右）。在（深层假冒工作流程）中：（a）是原始图像。（b）图像中检测到的脸。（c）检测到的2D面部地标。（d）使用仿射变换M将（a）中的裁剪面变形为标准面。（e）由深度神经网络合成的Deep Fake脸。（f）使用M-1将深假面转换回。（g）变换后的脸部面具会根据地标进行优化。（g）合成的脸部合并为原始图像。（h）最终的伪造图像。对于（我们的方法）：顶行对应于真实图像，而底端对应于Deep Fake。我们比较，使用来自全脸（j），（m）或仅中央脸部区域（i），（l）的面部标志进行评估，的头部姿势。对准误差显示为头部姿势的差异，显示为它们在像平面上的投影。然后将头部姿势的差异馈送到SVM分类器，以区分仿冒（n）和原始图像（k）。

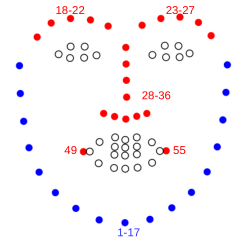


图2. 68个面部地标。红点用作中心脸部区域。蓝色和红色地标被用作整个面部。那些表示为空圆圈的标记在头部姿势估计中不使用。

## 3. 3D头部姿势估计

3D头部姿势对应于旋转和平移世界坐标来对应相机坐标。具体来说，将[U，V，W] T 表示为一个面部标志的全局坐标，[X，Y，Z] T是其相机坐标，并且（x，y）T是其图像坐标。世界与相机坐标系之间的转换可以表示为



 (1)

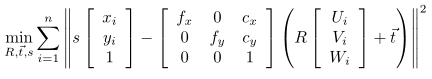


其中R是3×3旋转矩阵， 是3×1平移向量。相机和图像坐标之间的转换系统定义为

 （2）

其中f x和f y是x-和y-方向的焦距，（c x，c y）是光学中心，而s是未知比例因子。

在3D头部姿势估计中，我们需要解决反向问题，即，使用从标准模型（例如3D平均面部模型，假设我们知道相机参数）中获得的相同脸部标志集合的2D图像坐标和3D世界坐标，估计s，R和。具体来说，对于一组n个面部标志点，这可以用公式表示为优化问题，因为



使用Levenberg-Marquardt算法[15]可以有效解决。估计的R是相机姿势，即相机相对于全局坐标的旋转，并通过将其反转为R T（其中R是正交矩阵）获得头部姿势。

## 4.在深度伪造中头部姿势的不连贯性

由于图1中的Deep Fake过程交换了中央面部区域中的面部，伪造面部的标志性位置通常会偏离原始面孔。如图1（c）所示，在中心面部区域P 0中的标志首先仿射变换为P 0 in = MP 0。经过生成神经网络之后，在伪造面部中，其对应的伪造地标是Q 0 out。

因为在Deep Fake中生成神经网络的配置，不保证地标匹配，以及人们有不同的面部结构，这一生成的人脸的标志Q 0out可能对P 0in有不同的位置。基于795对像素为64×64的图像的51个中心区域地标的比较，生成神经网络从输入（图1（d））到输出（图1（e））的标记平均转换为1.540像素，其标准偏差为0.921像素。经过逆变换Q 0 = M -1 Q 0 out，在伪造的面孔的地标位置Q 0，将与相应的在原来的脸上的地标P 0不同。但是，由于Deep Fake仅交换中央面部区域中的面部的事实，脸部外部轮廓上的标志位置（图1（c）和（f）中的蓝色标记P 1）将保持不变。这个假脸的中心和外部轮廓之间的地标不匹配，被当做评估中央和整个面部地标的3D头部姿势不一致性。特别的，中心脸和全脸区域之间的头部姿势差异在真实图像中会很小，但在伪图像中会很大。

我们进行实验以确认我们的假设。为简单起见，我们仅查看头部方向向量。作为使用来自使用第3节描述的方法的全脸（图2中的红色和蓝色地标）的面部标志估计的旋转矩阵， 作为仅使用中部地区的地标（在图2中红色地标）的估计值。我们获得3D单位向量v a和v c相应于以此方式估算的头部朝向，v a = w和v c =  w，其中w = [0，0，1]T是世界坐标系中w轴的方向。我们然后比较两个单位向量 v c和v a之间的余弦距离，1 − v a ·v c /（||v a|| ||v c||），其取值为[0，2]，0表示两个向量彼此一致。该值越小，两个向量越接近彼此。图3所示为：用于一组原始的和Deep Fake生成的图像的，v c和v a之间的余弦距离的直方图。如这些结果所示，实像的两个估计的头部姿势向量中的余弦距离专注于较小的值范围，直至0.02，而对于“深造假”，大多数值范围在介于0.02和0.08之间。区别在于，对于真实和深造假的，两个头部方位向量的余弦距离的分布，根据此提示，表明它们可以是不同的。

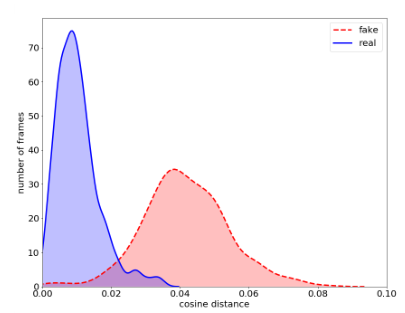


图3虚假和真实的人脸图像 v c and v a之间的余弦距离的分布。

## 5.基于头部姿势的分类

我们根据使用完整的面部标记和中央面部区域的标记估计的头部姿势之间的差异进一步训练了SVM分类器以区分真实图像还是视频的伪造品。这些特性是遵循以下步骤提取的：（1）对于每个图像或视频帧，我们使用软件包DLib [16]，运行面部检测器，并提取68个面部标志。（2）然后，用来自OpenFace2 [17]，相同的68点的，标准3D面部界标模型，来自中央面部区域（R c和t c）和整个脸（R a和t a）的头部姿势，分别估计为标记18 − 36、49、55（图2中的红色）和1 − 36、49、55（图2中的红色和蓝色）。在这里，我们近似相机焦距为图像宽度，相机中心为图像中心，而忽略镜头变形的影响。（3）将所获得的旋转矩阵（R a -R c）和平移向量（t a -t c）之间的差展平为一个向量，它是通过减去其平均值然后除以分类的标准偏差来标准化的。

SVM分类器的训练和测试数据是基于真实和Deep Fake图片视频的两个数据集的。第一个，UADFV，是一组Deep Fake视频，它们对应的是我们之前工作[9]中使用的真实视频。该数据集包含49个真实视频，它们被用于创建49个Deep Fake视频。这些视频的平均长度大约11.14秒，普遍分辨率为294×500像素。第二个数据集是一个来自DARPA MediFor GAN图像/视频挑战赛[18]的子集，其中包含241个真实图像和252个Deep Fake图像。为了SVM分类器的训练，我们使用来自UADFV数据集中的35个真实的和35个Deep Fake视频，总计21694张图像。其余14部真实视频的帧（共11，058帧）和14部来自UADFV数据集的Deep Fake视频和DARPA GAN集中的所有图像用于测试SVM分类器。我们在训练数据上使用RBF内核训练SVM分类器，并使用5倍交叉验证的超参数进行网格搜索。

使用单帧作为分析单元，AUROC作为演示量度，评估的性能如图4的两个数据集所示。如结果所示，在UADFV数据集，SVM分类器达到0.89的AUROC。这表明中心区域和全脸评估的头部姿势的不同之处是识别生成的Deep Fake图片的好功能。同样，在DARPA GAN挑战数据集上，SVM分类器的AUROC为0.843。这是由于DARPA GAN挑战中合成面孔的事实是通常模糊不清，导致难以准确预测面部标志位置，并因此估计头部姿势。我们还使用将单个视频作为UADFV数据集的分析单位方法估算效果。

这是通过对各个视频的帧的分类预测取平均值实现的。表现如表1的最后一行。

我们还进行了消融研究，以比较SVM分类器中使用的各种不同类型功能的表现。具体来说，我们比较了五种不同类型的功能基于估计的相机坐标中的3D头部姿势的旋转和平移，也如表1所示。（1）如第4节所述，我们将头部姿势简化为头部方向，va和v c。使用v a -v c作为特征进行分类，在Deep Fake Dataset上达到0.738 AUROC。这是所期待的，是因为这种简化忽略了在其他轴上的平移和旋转。（2）因为有3个自由度旋转，将头部姿势旋转矩阵表示为Rodrigues旋转向量（r a -r c），可以将AUROC增加到0.798。（3）代替罗德里格斯的向量r∈R 3，展平3 x 3旋转矩阵R a -R c的差异作为特征，进一步将AUROC提高到0.840。（4）介绍对应（1）和（2）的平移向量t a − t c的差异结果为AUROC分别为0.866和0.890，这是因为在平移中增加了头部姿势。

## 6.结论

在本文中，我们提出了一种公开AI-生成伪造的人脸图像或视频（众所周知的Deep Fakes）的新方法。我们的方法基于以下观察，这样的“深造假”是通过拼接合成人脸区域放入原始图像中来创建的，然后介绍了从面部图像估计3D头部姿势时可能涉及的错误。我们进行实验演示来论证这种现象并进一步发展基于此提示的分类方法。我们还报告了我们的方法对一组真实面部图像和深度造假图像的实验评估。

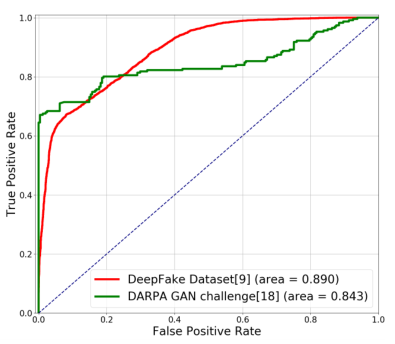
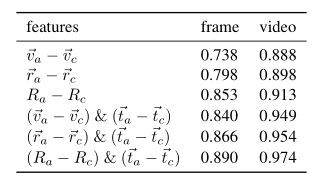


图4. SVM分类结果的ROC曲线，详见文本。

表1.基于来自UADFV数据集[9]的视频和帧的AUROC



## 7.参考文献

[1] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros,“Image-to-image translation with conditional adversarial networks,” arxiv, 2016.

[2] IrynaKorshunova, WenzheShi, JoniDambre, andLucasTheis,“Fast face-swap using convolutional neural networks,” in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3677–3685.

[3] Yuval Nirkin, Iacopo Masi, Anh Tran Tuan, Tal Hassner, and Gerard Medioni, “On face segmentation, face swapping, and face perception,” in Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on.IEEE, 2018, pp. 98–105.

[4] Hai X Pham, Yuting Wang, and Vladimir Pavlovic, “Generative adversarial talking head: Bringing portraits to life with a weakly supervised neural network,” arXiv preprint arXiv:1803.07716, 2018.

[5] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, N. Nießner,P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt, “Deep Video Portraits,” ACM Transactions on Graphics 2018 (TOG), 2018.

[6] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner, “Face2face: Real-time face capture and reenactment of rgb videos,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2387–2395.

[7] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen, “Progressive growing of gans for improved quality, stability, and variation,” arXiv preprint arXiv:1710.10196, 2017.

[8] Robert Chesney and Danielle Keats Citron, “Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security,” 107 California Law Review (2019, Forthcoming); U of Texas Law, Public Law Research Paper No. 692; U of Maryland Legal Studies Research Paper No. 2018-21.

[9] Yuezun Li, Ming-Ching Chang, and Siwei Lyu, “In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking,” in IEEE International Workshop on Information Forensics and Security (WIFS), 2018.

[10] Haodong Li, Bin Li, Shunquan Tan, and Jiwu Huang, “Detection of deep network generated images using disparities in color components,” arXiv preprint arXiv:1808.07276, 2018.

[11] Pavel Korshunov and Sébastien Marcel, “Speaker inconsistency detection in tampered video,” in 2018 26th European Signal Processing Conference (EUSIPCO). IEEE, 2018, pp. 2375–2379.

[12] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen, “Mesonet: a compact facial video forgery detection network,” in IEEE International Workshop on Information Forensics and Security (WIFS), 2018.

[13] Pavel Korshunov and Sébastien Marcel, “Deepfakes: a new threat to face recognition? assessment and detection,” arXiv preprint arXiv:1812.08685, 2018.

[14] David Güera and Edward J Delp, “Deepfake video detection using recurrent neural networks,” in IEEE International Conference on Advanced Video and Signal-based Surveillance (AVSS), 2018.

[15] G. Bradski, “The OpenCV Library,” Dr. Dobb’s Journal of Software Tools, 2000.

[16] Davis E. King, “Dlib-ml: A machine learning toolkit,” Journal of Machine Learning Research, vol. 10, pp. 1755–1758, 2009.

[17] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency, “Openface 2.0: Facial behavior analysis toolkit,” in Automatic Face & Gesture Recognition (FG 2018),

2018 13th IEEE International Conference on. IEEE, 2018, pp. 59–66.

[18] Haiying Guan, Mark Kozak, Eric Robertson, Yooyoung Lee, Amy N. Yates, Andrew Delgado, Daniel Zhou, Timothee Kheyrkhah, Jeff Smith, and Jonathan Fiscus, “MFC datasets:

Large-scale benchmark datasets for media forensic challenge evaluation,” in IEEE Winter Conf. on Applications of ComputerVision(WACV),WorkshoponImageandVideoForensics, 2019.