充满争议的“双刃剑”——AI DeepFake 技术浅谈

关文聪　2016060601008

摘要：

二十一世纪，我们所处的时代，是一个飞速发展的时代，科学技术的发展日新月异。其中，人工智能迅速崛起，成为了时代的风头浪尖。其中，计算机视觉，作为人工智能领域的重要分支，也在飞快地发展着。近几年来，一种名为“AI DeepFake”的技术在互联网网站上出现，这种技术可以利用人工智能算法，将一个人的脸合成到另一个人的视频中，并且只要训练量足够大，视频合成就能达到非常逼真的效果！此技术一经发布，就在互联网乃至全社会引起轩然大波，尽管网站随后对其采取进行了封锁和删除，但由于传播范围广、社会影响大，它带来的风波却从未停息。本文将对 AI DeepFake 技术作简要的介绍与分析，并且对其在社会的影响和争议进行一些相关讨论。

关键词：人工智能、计算机视觉、AI DeepFake 技术、换脸、社会影响、争议

正文：

DeepFake 技术引起互联网广泛讨论是在 2017 年底。2017 年 12 月，一个名为“DeepFake”的用户在 Reddit 网站上发布了一个“假视频”，它展示了通过 AI 合成的明星色情片，视频中的艺人其实是后期加上的，但是看起来几乎毫无破绽。他利用了深度学习和 AI 新技术，在成人电影中把演员的脸替换成某个艺人的脸，从而制作成了这个看上去以假乱真的视频。一石激起千层浪，该技术由于其用途的争议性而迅速引爆互联网，在社会上引起广泛而激烈的讨论。虽然网站随后采取了相关措施封锁删除，但由于作者将代码开源发布，且传播甚广，至今依然一直处于互联网和社会的讨论和争议中。

那么，什么是 DeepFake 技术？Deepfake，是英文“deep learning”（深度学习）和“fake”（伪造）的混成词，专指用基于人工智能的人体图像合成技术。此技术可将已有的图像和影片叠加至目标图像或影片上。简单来说就是脸部替换，可以将 B 的脸换到 A 的脸上。和 PS 不同的是，这项技术不仅可以生成图片，还可以生成视频的。事实上，人脸交换技术在电影制作领域已经不是个新鲜词了，

但是之前电影视频中的人脸交换非常复杂，专业的视频剪辑师和 CGI 专家需要花费大量时间和精力才能完成视频中的人脸交换。DeepFake 的出现可以说是人脸交换技术的一个突破。利用 DeepFake 技术，你只需要一个 GPU 和一些训练数据，就能够制作出以假乱真的换脸视频。并且，用于训练的数据量越大，其类型越丰富，得到视频往往就越逼真，越让人难以分辨。

其实，DeepFake 技术看似很神奇，但背后的相关技术原理，对于了解机器学习、深度学习和神经网络相关知识的人士而言并不算很复杂。DeepFake 的原理用一句话可以概括：用监督学习训练一个神经网络将某人的扭曲处理过的脸还原成原始脸，并且期望这个网络具备将任意人脸还原成其的脸的能力。最原始版本的 DeepFake 实际上就是一个自编码模型。为此，此处首先要简要介绍一下自编码器：自编码器（AutoEncoder）是一种能够通过无监督学习，学到输入数据高效表示的人工神经网络。输入数据的这一高效表示称为编码（codings），其维度一般远小于输入数据，使得自编码器可用于降维。更重要的是，自编码器可作为强大的特征检测器（feature detectors），应用于深度神经网络的预训练。此外，自编码器还可以随机生成与训练数据类似的数据，这被称作生成模型（generative model）。比如，可以用人脸图片训练一个自编码器，它可以生成新的图片。神经网络中的权重矩阵可看作是对输入的数据进行特征转换，即先将数据编码为另一种形式，然后在此基础上进行一系列学习。然而，在对权重初始化时，我们并不知道初始的权重值在训练时会起到怎样的作用，也不知道在训练过程中权重会怎样的变化。因此一种较好的思路是，利用初始化生成的权重矩阵进行编码时，我们希望编码后的数据能够较好的保留原始数据的主要特征。如果编码后的数据能够较为容易地通过解码恢复成原始数据，我们则认为较好的保留了数据信息。

对自编码有了初步的了解和认识后，我们回到 DeepFake，它可以简单的用公式表示为：

$$X' = \mathrm{Decoder}\big(\mathrm{Encoder}(XW)\big)$$

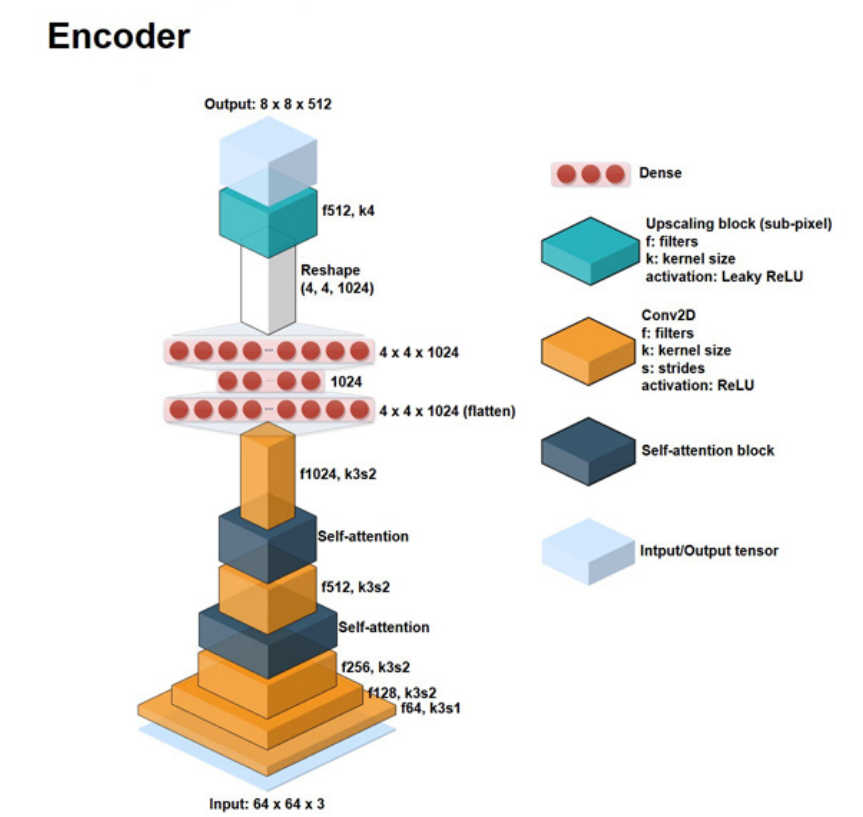
$$\mathrm{Loss} = \mathrm{L1Loss}(X' - X)$$

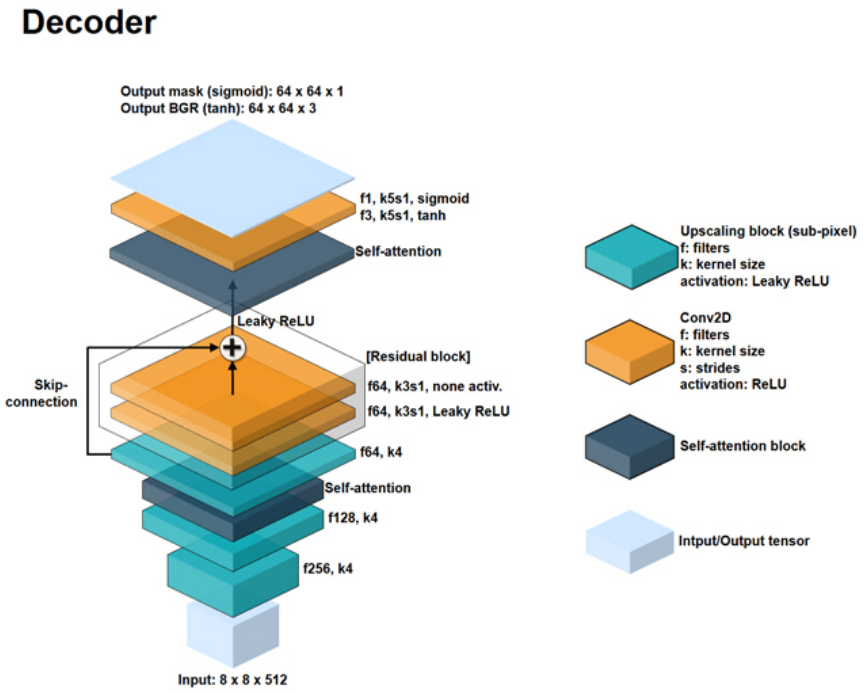
这里的 XW 是经过扭曲处理过的图片。要运行 DeepFake 算法，需要提供两个人的人脸数据：

$$A' = \text{Decoder}_A\big(\text{Encoder}(AW)\big)$$

$$B' = Decoder_B\big(Encoder(BW)\big)$$

Encoder 与 Decoder 的示意图如下所示：

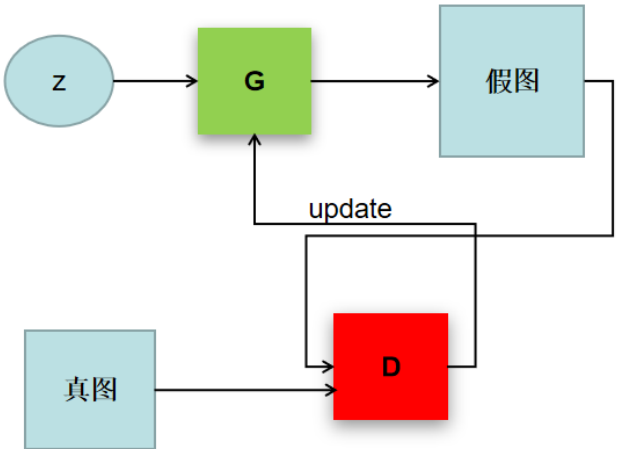


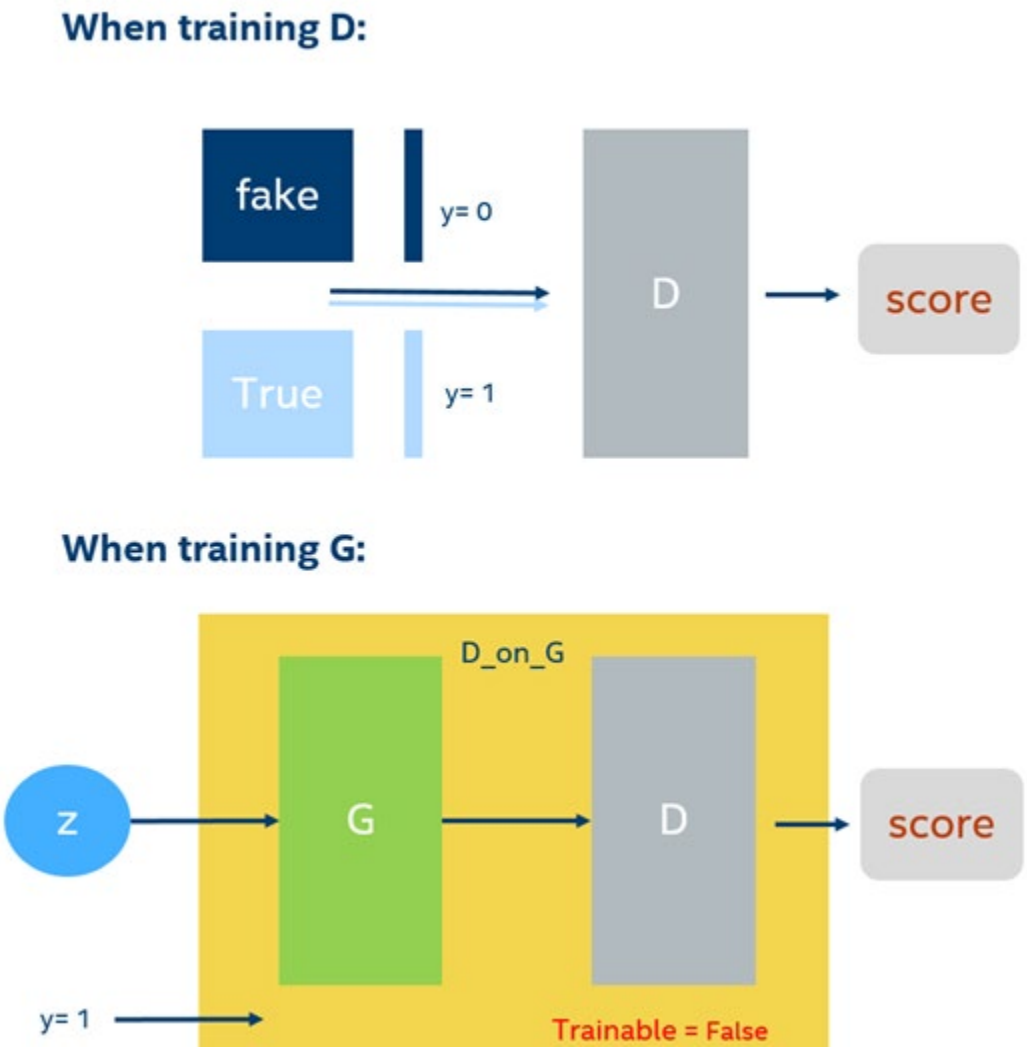


DeepFake 技术一经开源发布，在互联网得到广泛流传，随后，在最原始版本的基础上，又有很多人做了改进和创新，其中，最广为人知、应用最广泛的，应该莫过于应用了 GAN（Generative Adversarial Networks，生成式对抗网络）的改进 DeepFake 了。

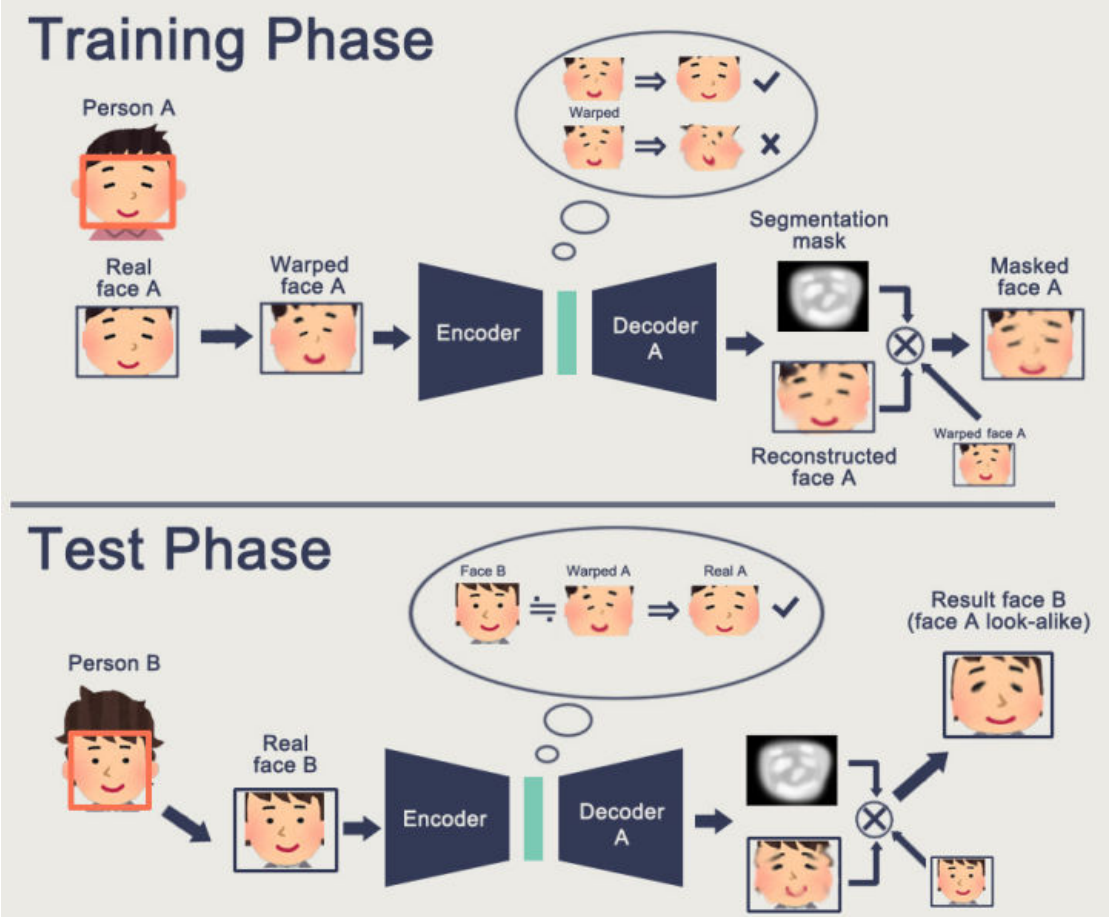
GAN 是一种深度学习模型，由 Ian J. Goodfellow 等人于 2014 年提出，是近年来复杂分布上无监督学习最具前景的方法之一。框架中同时训练两个模型：捕获数据分布的生成模型 G（generator，生成器：负责凭空捏造数据出来），和估计样本来自训练数据的概率的判别模型 D（discriminator，判别器：负责判断数据是不是真数据）。G 的训练程序是将 D 错误的概率最大化。GAN 的初衷就是生成不存在于真实世界的数据，类似于使得 AI 具有创造力或者想象力。目前 GAN 最常使用的地方就是图像生成。接下来就以图像生成为例，进一步阐述 GAN 的思想与原理。

在使用 GAN 生成图像过程中，同样的，同时训练两个模型 G 和 D。G 是一个生成图片的网络，它接收一个随机的噪声 z，通过这个噪声生成图片，记做 G(z)。D 是一个判别网络，判别一张图片是不是“真实的”。它的输入参数是 x，x 代表一张图片，输出 D（x）代表 x 为真实图片的概率，如果为 1，就代表 100%是真实的图片，而输出为 0，就代表不可能是真实的图片。在训练过程中，生成网络 G 的目标就是尽量生成真实的图片去欺骗判别网络 D。而 D 的目标就是尽量把 G 生成的图片和真实的图片分别开来。这样，G 和 D 构成了一个动态的“博弈过程”。GAN 的主要思想就体现在“对抗”以及“欺骗”上，生成模型和判别模型通过互相博弈学习，可以产生相当好的输出。下图展示了 GAN 模型的整体模型结构以及模型 G 和模型 D 各自的训练学习过程：

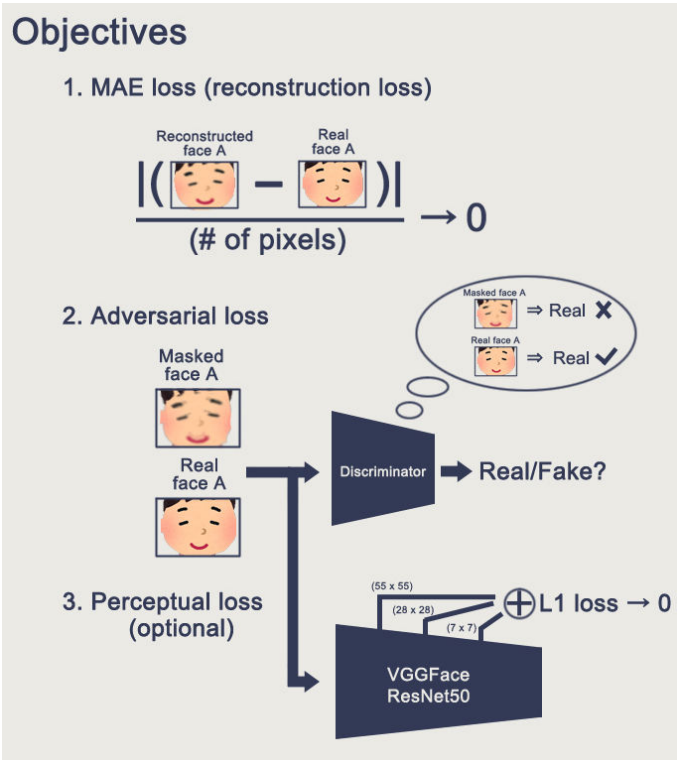




同理可以将 GAN 应用在 DeepFake 技术上。在制作 DeepFake 的过程中，GAN 会让两个神经网络彼此对抗，第一个神经网络被称为生成网络，它负责制作尽可能逼真的作品，而第二个网络为鉴别器，它将前者生成的作品和原始数据库中大量的“真迹”进行对比，来鉴别哪些是真的哪些是假的。基于每一次的“对抗”结果，生成网络会调整它制作时使用到的参数，直到鉴别器无法辨别它生成的作品和真迹，就算“成功”。这时候，最终合成的视频基本可以以假乱真了。对于使用 GAN 进行改进的 DeepFake 技术，其示意图如下：



简而言之就是增加了 Adversarial Loss 和 Perceptual Loss，后者是用训练好的 VGGFace 网络（该网络不做训练）的参数做一个语义的比对。



至此，可以总结一下 DeepFake 技术的主要流程，与其它的图像处理技术非常类似，无非就是 1. 图像预处理，从大图（或视频）中识别，并抠出人脸图像。2.深度神经网络训练（例如应用上文提到的 GAN）3.图像融合。另外，随着作者开源发布源码，很多人在此基础上封装、制作了使用更方便的脚本或软件，这使得越来越多的人可以更方便地接触和使用 DeepFake 技术，甚至不需要了解其具体的细节与训练过程，只需要准备充足数量的图像数据集，以及足够的 GPU 计算资源，普通人也可利用此技术制造出各种各样的合成伪造视频。

正是由于这种技术变得越来越方便，越来越“黑盒化”“平民化”，在科学技术飞速发展，互联网高度发达的今天，它得到了极为广泛的传播，而伴随而来的社会争议也一直未曾平息。DeepFake 技术的横空出世如同打开了“潘多拉魔盒”，带来了很多极具争议的影响。有人将 DeepFake 技术用于合成一些明星的色情影片，并发布在成人网站上，这给明星的名誉和社会形象带来了巨大的影响，甚至已经触犯了法律；也有人将此技术用于伪造公众人物、社会名流的公开演讲等等，以发布假信息、谣言，让人难辨真假……不仅是名人，普通人也可能因此遭到骚扰、攻击、敲诈勒索等威胁，甚至以 DeepFake 合成的视频可以通过面部识别“刷脸”机器，以此进行盗刷和其他非法用途……

针对以上种种出现的争议和问题，人们开始研究识破这种伪装的方法。同样的，利用人工智能技术，目前出现了例如“眨眼检测”的检测算法。该算法利用的是训练数据集本身的缺陷，即很多训练数据往往缺少闭上眼睛的图片，这使得使用人工智能算法训练出来的视频往往具有缺陷，例如人物很少甚至不眨眼睛，或者在眨眼睛的瞬间会有变形、抖动等异常，据此可以进行判别。但是，目前而言，这种方法的成功率并不太高。其它的方法例如肉眼鉴别、用加密散列标记视频来确定视频是否被篡改等方法同样面临种种困难。对于业余的 DeepFake 伪造作品，往往很容易就能被辨别出来，但是对于经过很专业训练诞生的伪造视频，如何有效识别和检测出来还有待研究。这将是一场科学技术的顶尖攻防较量，在未来可能会持续相当长的一段时间。

我们应该意识到，科学技术是一把双刃剑，DeepFake 技术也不例外，如何正确地使用它并为人类带来便利，取决于我们人类本身。我们更希望将 DeepFake 技术用于正途，例如影视拍摄和后期特效，效果可能是传统的拍摄和

制作难以企及的；再例如用于纪录片、历史资料影像的还原，真实地还原历史人物的原貌等等……我们应该相信，在科学技术不断发展的时代，在不远的将来，DeepFake 技术能够继续发展，并被人们正确利用，给人类创造更大的价值，带来更光明美好的未来！

参考资料：

[1] Ian J. Goodfellow, Jean Pouget-Abadie————————————————, Mehdi Mirza, Bing Xu, David Warde-Farley,Sherjil Ozairy, Aaron Courville, Yoshua Bengioz. Generative Adversarial Nets. https://arxiv.org/pdf/1406.2661.pdf

[2] Tero Karras, Samuli Laine, Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. https://arxiv.org/pdf/1812.04948.pdf

[3] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B. Tenenbaum, William T. Freeman, Antonio Torralba. GAN DISSECTION:VISUALIZING AND UNDERSTANDING GENERATIVE ADVERSARIAL NETWORKS. https://arxiv.org/pdf/1811.10597.pdf

[4] Korshunova, I., W. Shi, J. Dambre, et al. Fast Face-Swap Using Convolutional Neural Networks. in 2017 IEEE International Conference on Computer Vision (ICCV). 2017.

[5] Gatys, L.A., A.S. Ecker, and M. Bethge. Image Style Transfer Using Convolutional Neural Networks. in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016.

[6] Pavel Korshunov, Sebastien Marcel. DeepFakes: a New Threat to Face Recogniton? Assessment and Detection. https://arxiv.org/pdf/1812.08685.pdf