Deepfakse 的原理是什么？

Deepfakes 2017 年在互联网上流传开来，引起轩然大波。

什么是 Deepfakes 呢？翻译过来是“深度造假”，直白一些的说法就是给图片或视频的主角换脸。比如一段奥巴马的演讲，换上自己的模样，是不是很有成就感？然而，如果是一段色情视频呢??

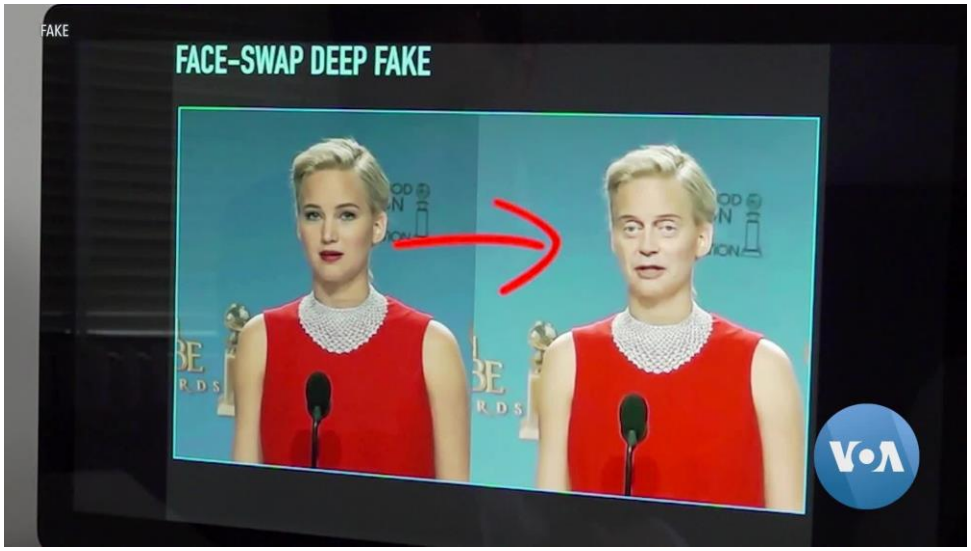

图 1：Deepfakes 实现换脸[1]

给图片或视频换脸其实并不新奇，人们已经研究了很多年[3,4]。传统换脸多基于图形学的 3D 模型重建追踪等技术。例如，首先捕捉到人脸，然后获取脸部的关键点，对这些关键点位置进行渲染，使之逐渐接近目标人脸。这些方法虽然可实现一定的换脸效果，但模型复杂，时间开销大，且生成的人脸有较明显的修改痕迹，表情不够自然。

Deepfake 采用的是深度学习方法，利用神经网络对人脸特征的学习能力，基于原人脸的表情特征合成目标人脸。Deepfakes 可采用多种网络结构和训练准则，图 2 给出一种基于编码-解码框架的实现方案。在训练阶段，将目标人 A 的脸部图片 X 经过扭曲变换之后得到图片 X’，将 X’输入编码器，得到隐藏层编码后再经过一个解码器恢复原人脸图片 X。训练完成后，这一编码-解码网络就学会了如何从扭曲的图片中恢复出人脸的方法。实际训练中，模型基于多个人的脸部图片同时训练，不同人共享一个编码器，但具有各自的解码器。经过这样的训练之后，共享的编码器将编码出所有人共有的表情特征，而特定人的解码器将依据这一共性特征恢复出对应人的脸部图片。

模型训练完成后，如果我们将 B 的图片送入共享的编码器，将得到 B 的表情特征，再通过 A

的解码器即可得到具有 B 的表情但是 A 的人脸的图片，从而实现由 B 到 A 的换脸。

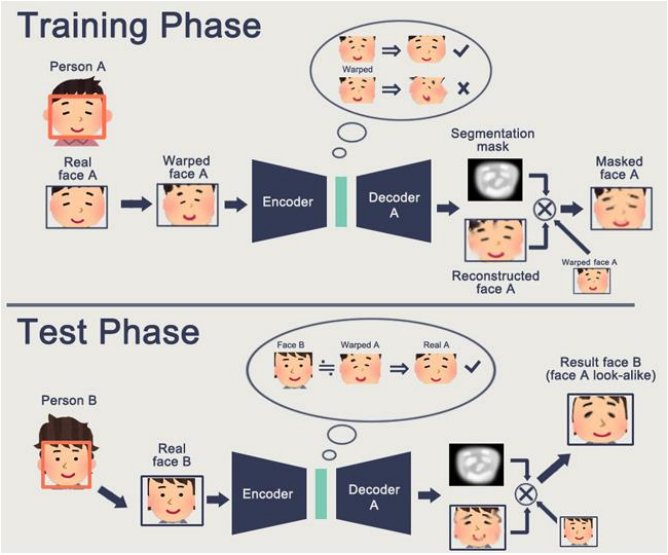

图 2：Deepfakes 的实现原理[5]

Deepfakes 不仅实现了逼真的换脸，而且公开了源码，稍微有点儿经验的人都能成为虚假图片和视频的制造者，当然也可能成为假货的受害者。目前，不仅换脸不成问题，换表情，换声音都已经是小儿科了，而且逼真程度绝对以假乱真。这是 AI 迄今为止给我们带来的最大的麻烦之一。

为了防止伪造视频带来的危害，研究者已经在研究各种方法对虚假视频进行检测，Facebook 和微软也发起了 Deepfakes 检测竞赛[6]。然而，道高一尺，魔高一丈，伪造和鉴伪之间战斗目前还在胶着中。对于普通老百姓的我们，靠肉眼分辨仿冒视频是不可能了，只能在遇到事情的时候多个心眼，保持警醒，多方求证。

[1] https://edtimes.in/how-is-the-rise-of-deepfake-a-threat-to-democracy/
[2] https://github.com/deepfakes/faceswap
[3] Zollhöfer M, Thies J, Garrido P, Bradley D, Beeler T, Perez P, Stamminger M, Niessner M, Theobalt C. State of the art on monocular 3D face reconstruction, tracking, and applications. Computer Graphics Forum, 2018,37(2):523-550.
[4] 李旭嵘等，深度伪造与检测技术综述，软件学报， 2021， 32（2）
[5] https://github.com/shaoanlu/faceswap-GAN
[6] https://www.kaggle.com/c/deepfake-detection-challenge