# Deepfake 的探究与防御

## 1 Deepfake 的来源

2017 年 12 月，一位名为“Deepfakes”的用户在全球流量排名第四的国际互联网社区“Reddit”上发布了一段好莱坞女星盖尔·加朵的伪造人脸视频，掀起了一阵轰动，这一事件作为开端，标志着人脸深度伪造技术的兴起，而该用户的用户名也被引用成为了这一类技术的代名词“Deepfake[1]”。

因此，Deepfake 指代人脸的深度伪造，即将目标视频人物的脸替换成指定的原始视频人脸，或让目标人脸重演、模仿原始人脸的动作、表情等，从而制作出目标人脸的伪造视频。

## 2 Deepfake 的理论与方法

Deepfake 技术在总体上可以分成两类[2]：基于图像域特征编码的方法和基于隐变量编辑的方法，其中基于图像域特征编码的方法中又可分为面部替换和属性编辑两大类，面部替换旨在用原始人脸面部替换目标人脸的面部区域，涉及目标图像身份属性的变化，属性编辑主要针对目标人脸身份信息外的各类属性进行编辑篡改，又包括表情迁移、面部重演、唇形篡改等具体篡改形式。

### 2.1 基于图像域特征编码的方法

#### 2.1.1 面部替换

面部替换是人脸视频深度伪造技术里最典型的一类算法，其主体结构基于自动编码器实现。

在训练阶段，对于原始人脸 A 目标人脸 B，训练一个权值共享的编码器，用于编码人脸特征，而在解码阶段，A 和 B 自训练一个独立的解码器用于重构人脸。在测试阶段，为了实现 A 和 B 之间的人脸替换，先用训好的编码器对目标人脸 B 进行编码，此时，用训好的 A 的解码器来解码由 B 编码得到的特征，由此可以实现将 A 的面部解码到 B 的人脸肖像上，基本原理如下图所示：

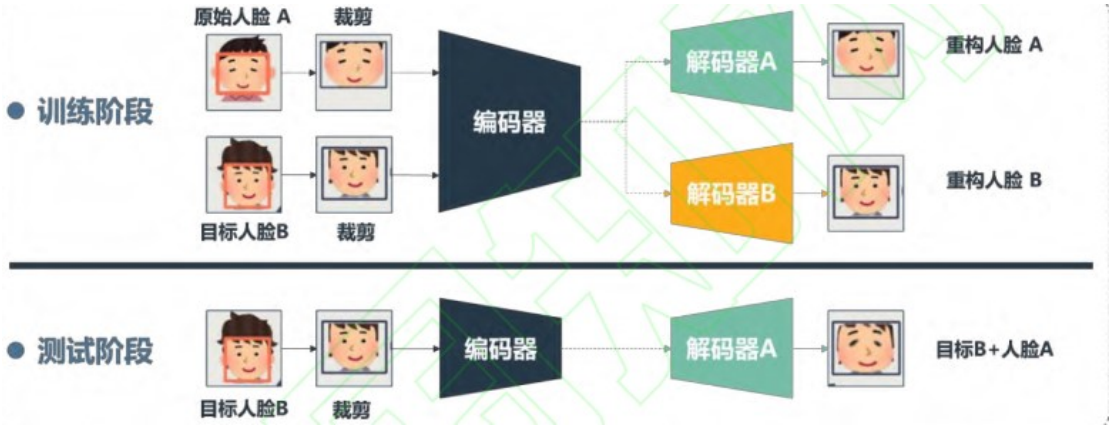


图 1 Deepfake 自动编码器原理框图[1]

[1] 图片来源与参考文献[2]

这种自动编码器的结构大大提升了面部替换的可操作性，同时也大大降低了换脸的技术门槛。但由于自动编码器结构上的限制，其对人脸特征的表征能力有限，因此在生成质量和生成的可控性方面都有待进一步提升。

随着对人脸深度伪造理解的不断深入，更多提升面部替换逼真度的方法被提出。如微软亚洲研究院的学者提出的 FaceShifter[3]，能解决交换的人像中可能存在的遮挡问题，中科院自动化所的研究者从最优化传输的角度提出的 AOT 算法[4]，该方法能够有效改善现有面部替换方法光照、肤色不协调的情况，并且取得了良好的视觉质量。

### 2.1.2 属性编辑

属性编辑是人脸深度伪造技术中另一类重要算法。该类算法以人脸属性为对象进行篡改，不涉及到目标人物身份信息的改变。通常，属性编辑可以改变视频人物的外观或动作表情特征，这类方法的输入可以是成对人脸视频，来实现目标人脸对原始人脸表情的模仿，也可以是单一的目标人脸加上某一指定的条件，将目标人脸的某种属性改变为指定的条件，如给定指定的语音内容，对目标人物的唇形进行篡改使得其口型配合上给定的语音内容等。

表情迁移算法 Face2Face[5]是人脸属性编辑的代表性算法，该算法使用成对的原始人脸和目标人脸作为输入，以五官的关键点作为表征，刻画并驱动不同表情的生成。在训练阶段，首先对目标人物的五官关键点进行提取，并以此作为图像翻译网络的输入，驱动网络训练学习并重构出具有相应表情的目标人脸。为了实现表情迁移，在模型的测试阶段，同样对待迁移的原始人脸提取五官关键点，以该关键点图形作为驱动输入到训练好的图像翻译模型中，重构出具有驱动表情的目标人脸，再根据需求合成视频。基本原理如下图所示：

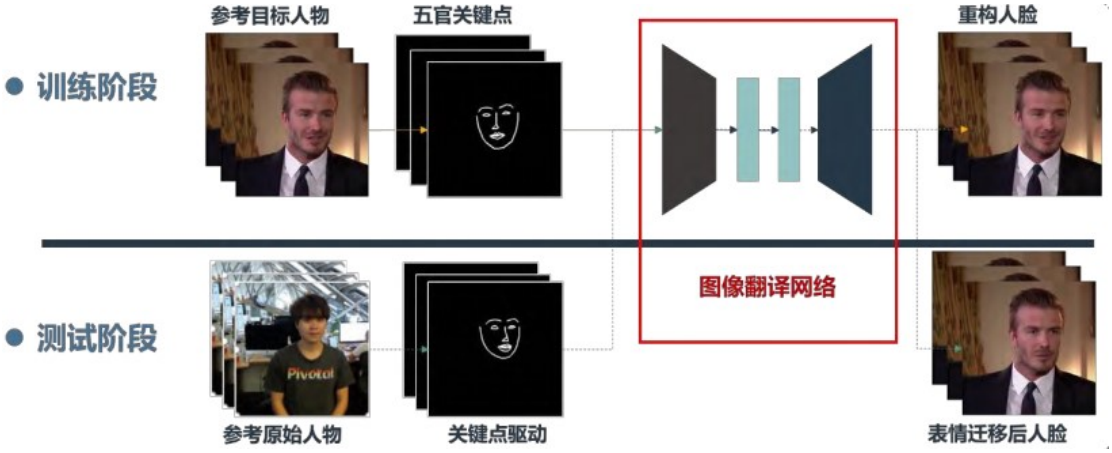


图 2 Face2Face 基本原理框图[2]

除了这类特定人物之间的表情迁移，一项由视频驱动图片进行表情迁移方法 First Order Motion[6]也获得了学术界的大量关注， 在其基础上也发展出了一系列娱乐应用。该方法大体上来说分成两个模块，一个是运动估计模块，另一个是图像生成模块。其中运动估计模块的输出有两个，一个是密集运动场，它表征了驱动视频人物图像中的每个关键点到目标图像的映射关系；一个是混淆遮罩，该参数约束了驱动图像的不同区域，表明了哪部分可以通过扭曲得到，哪部分需要通过图像填充得到。而在图像生成模块中，输入目标人物图像，可以在编码器得到的特征层中进行形变，再解码回去，得到最终的输出。以该方法作为基础，已

[2] 图片来源与参考文献[2]

经发展出了 Avatarify 等互联网上非常流行的换脸软件。

近期，一些跨模态的唇形篡改方法如 Wave2Lip[7]也逐渐成为研究热点。这类方法通常以驱动音频以及目标人物视频作为输入，首先对音频和目标人脸分别进行特征提取，再通过不同类型的联合共享特征空间的表征，配合不同视角的损失函数进行约束，最后进行音频和篡改后视频的同步，得到能够配合驱动音频的篡改唇形视频。

### 2.2 基于隐变量编辑的方法

在人脸伪造相关技术中，有一类方法通过编辑人脸图像的隐空间变量实现篡改，这类方法大多基于生成对抗网络（GAN）来实现。与基于图像域特征编码的方法不同，基于 GAN 隐空间实现人脸语义篡改的方法依赖于已训练好的 GAN 网络，探索人脸图像在隐空间中对应的隐变量，找到待篡改的语义方向，再利用预训练好的 GAN 生成器来生成编辑后的人脸。

#### 2.2.1 GAN 网络结构

GAN 的网络结构由生成网络和判别网络组成，模型结构如图 3 所示。生成器 G 接收随机变量 z，生成假样本数据 G(z)。生成器的目的是尽量使得生成的样本和真实样本一样。判别器 D 的输入由两部分组成，分别是真实数据 x 和生成器生成的数据 G(x)，其输出通常是一个概率值，表示 D 认定输入是真实分布的概率，若输入来自真实数据，则输出 1，否则输出 0。同时判别器的输出会反馈给 G，用于指导 G 的训练。理想情况下 D 无法判别输入数据是来自真实数据 x 还是生成数据 G(z)，即 D 每次的输出概率值都为 1/2，此时模型达到最优。在实际应用中，生成网络和判别网络通常用深层神经网络来实现。

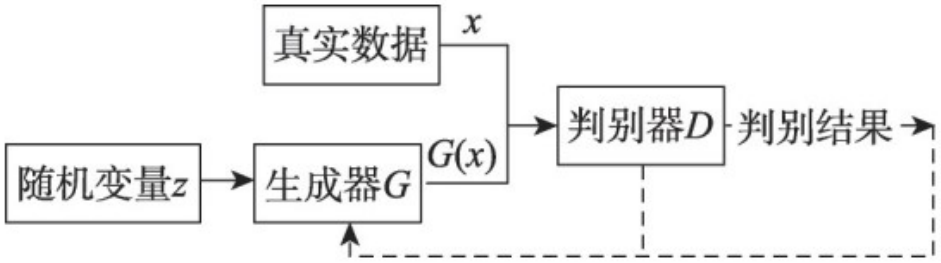


图 3 GAN 网络模型结构示意图[3]

GAN 的思想来自于博弈论中的二人零和博弈[8]，生成器和判别器可以看成是博弈中的两个玩家。在模型训练的过程中生成器和判别器会各自更新自身的参数使得损失最小，通过不断迭代优化，最终达到一个纳什均衡状态，此时模型达到最优。

#### 2.2.2 原始 GAN 模型存在的问题与解决

原始的 GAN 并不成熟，存在着诸多问题，主要问题是梯度消失和模式崩溃。梯度消失即利用误差反向传播算法对深度神经网络进行训练时，梯度后向传播到浅层网络时基本不能引起数值的扰动，最终导致神经网络收敛很慢甚至不能收敛。GAN 在判别器训练得越好的时候，生成器梯度消失得越严重。GAN 模式崩溃是指 GAN 生成不了多样性的样本，而是生成了与真实样本相同的样本。

---

[3] 图片来源与参考文献[8]

研究者们针对 GAN 存在的问题，提出了很多基于 GAN 的变体。比如 CGAN[9]在原始 GAN 的基础上增加了约束条件，控制了 GAN 过于自由的问题，使网络朝着既定的方向生成样本，但其存在模型训练不稳定的缺点。LAPGAN[10]基于 CGAN 进行改进，它能够生成高品质的图片，但它必须在有监督下训练。

### 2.2.3 GAN 反演

GAN 反演的目的是将真实图像反转到预先训练好的 GAN 模型的隐空间中，以便生成器准确地从被反演的隐编码中重建图像。它不仅提供了一个灵活的图像编辑框架，而且还有助于揭示深度生成模型的内在机制。通过 GAN 反演方法获得了真实图像在隐空间中对应的隐编码后，便可以利用有监督或无监督的方法找到某个具体面部语义在隐空间中的方向，沿该方向调整隐编码，即可实现面部相应语义的篡改。

其中，典型的有监督的代表方法是由香港中文大学汤晓鸥团队提出的 InterfaceGAN[11]。其基本原理如图 4 所示。针对某个待篡改属性，例如：眼镜、年龄、性别等，首先通过预训练的相应属性的预测模型对于从隐空间中随机采样的隐编码所对应的合成图像进行判决，经过判决后便可以对该隐编码给予确定的标签，即是否包含目标属性。接着，利用这些被给予标签的隐编码即可训练 SVM 分类器。对于给定的二分类属性，在隐空间中存在一个线性超平面可以很好地将隐编码分为两类。因此，我们便获得了该属性对应的语义方向。

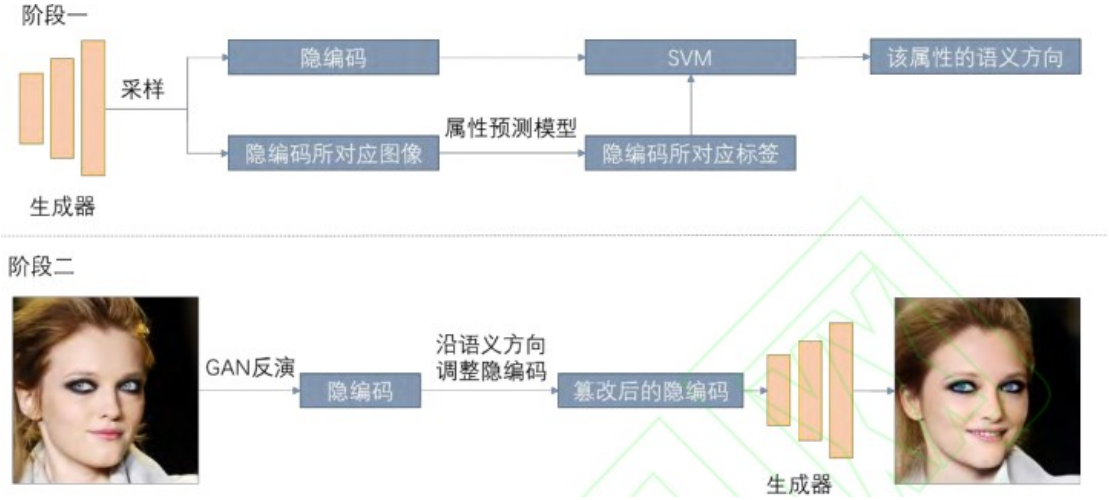


图 4 InterfaceGAN 基本原理图[4]

有了相应属性的语义方向和通过 GAN 反演方法获得的真实图像对应的隐编码，我们可以通过令隐编码在该语义方向进行线性插值的方式篡改相应语义，之后将篡改后的隐编码送入生成对抗网络生成篡改后的图像。

## 3 Anti-deepfake 的理论与方法

Anti-deepfake 指代 Deepfake 人脸视频防御技术，根据防御策略的不同，现有的防御技术可以大体划分为被动式检测和主动式防御两大类，其中被动式检测技术侧重于事后取证，即针对已经制作并传播的视频进行检测，判别其是否属于伪造人脸视频；另一类主动式防御技术侧重于事前防御，即在人脸数据发布传播前添加隐藏信息，如水印、对抗噪声等，进行主动溯源或使得恶意使用者无法利

[4] 图片来源与参考文献[2]

用添加了噪声的人脸视频进行伪造，从而达到保护人脸，实现主动防御的目的。

### 3.1 被动式检测方法

被动式检测技术指仅从人脸视频自身获取信息或提取特征，对伪造人脸视频进行鉴别的技术，这个任务本质上是一个二分类任务。

#### 3.1.1 有伪造样本学习方法

这类方法的核心特点是利用真假成对数据作为训练的数据驱动，分类模型的学习过程需要有伪造人脸样本的参与。根据信息提取视角的不同，有伪造样本学习方法又可细分为基于空域信息的方法、基于时域信息的方法、基于频域信息的方法、基于通用伪造痕迹的方法、基于注意力机制的方法、跨模态检测方法等。

#### 3.1.2 无伪造样本学习方法

无伪造样本学习方法的模型训练过程不需要使用伪造人脸的负样本作为数据驱动，而是抓住了人脸这一特殊信息载体的某些特性，或抓住了深度伪造过程中某一固有的流程漏洞实现检测与鉴别。

#### 3.1.3 基于多任务迁移的方法

多任务迁移的方法本质上是利用其它取证或视觉任务中已有的方法进行迁移改造，应用到 Deepfake 人脸伪造视频的检测任务中。

例如，Haliassos 等人提出的 Lip Forensics[12]，将唇读任务中的预训练模型迁移到了 Deepfake 伪造人脸视频的检测任务中。利用了针对嘴部动作中的高级语义不规则性，以及现有伪造技术在唇形生成方面的弱点，区分真假唇形，进而鉴别伪造人脸视频。该方法由于在大规模的唇读任务数据集上进行了预训练，因此在 Deepfake 检测任务中，网络不会过拟合到少量的伪造人脸数据唇形上，所以在库内和跨库迁移性方面都表现出了非常优秀的性能。

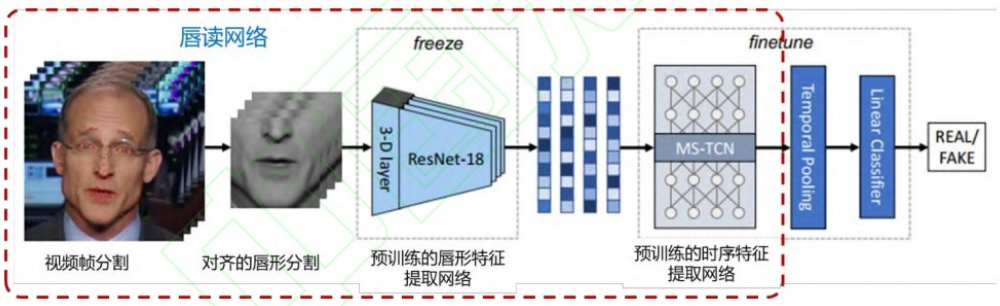


图 5 基于唇读任务的 Deepfake 人脸伪造检测算法框图[5]

### 3.2 主动式防御方法

主动式防御的核心思路是在人脸视频发布前便添加一定程度的信号干扰，使得恶意使用者无法利用添加了干扰后的人脸素材进行伪造，或即便伪造也能够顺利溯源，找到伪造者，以实现“事前防御”的目标。

主动防御技术也可归纳为主动干扰和主动取证两类，主动干扰指的是对发布的人脸素材添加噪声干扰，如对抗攻击或数据毒化，使恶意使用者无法对该人脸进行伪造。主动取证指的是将某种标识嵌入训练数据中，从而能够在生成的伪造结果中检测到这种标识。

[5] 图片来源与参考文献[2]

# 4 我的应对思路

## 4.1 Deepfake 的两面性

虽然 Deepfake 的出现就伴随着技术的滥用，在其发展过程中，一度严重威胁国家政治、外交安全，甚至引发战争，但我觉得仍不能将其完全否定。在我看来，技术只是工具，人的使用才使其成为双刃剑。

诚然，Deepfake 在娱乐、影视行业有着天然的用武之地，比如 Deepfake 目前已经逐渐开始取代一些传统影视的 CG 动作捕捉技术，能够降低行业成本，推动技术创新。

但是，若无管制，一些虚假的换脸视频大肆传播，不仅会构成当事人名誉的损害，甚至造成社会动荡。因此，Anti-deepfake 也必不可缺。在当下，我觉得更应当发展 Anti-deepfake 技术，从而遏制 Deepfake 的滥用。对此，我有两个角度的应对思路，分别是主动式 Anti-deepfake 和被动式 Anti-deepfake。

## 4.2 主动式 Anti-deepfake

### 4.2.1 加强监管

法律能从根本上约束人的行为，同样能从根本上遏制 deepfake 的滥用。纵观全球，美国现行的《诽谤法》《欺诈法》以及其他刑事和民事法律规定，创作、分享深度伪造会被列为违法行为，法国通过了《反假新闻法》，用来遏制在选举中出现和流传的虚假新闻，德国通过了《改进社交网络中的法律执行的法案》，强化了社交网络平台删除或屏蔽某些刑事违法内容的义务。

在我国，国家互联网办公室发布的《网络音视频信息服务管理规定》对网络音视频信息服务提供者和使用者的行为进行了规范。规定中专门要求，网络音视频信息服务提供者、使用者利用基于 Deepfake 等深度学习技术制作、发布、传播的非真实音视频信息，应当以显著方式予以标识，禁止利用 Deepfake 等深度学习技术制作、发布、传播虚假新闻信息．网络音视频信息服务提供者要加强对音视频信息的管理，一旦发现网络音视频信息服务使用者利用 Deepfake 等深度学习技术制作、发布、传播违法违规内容，应立即依法停止传输、保存记录，并向网信等部门备案。

然而，在现实中，仍然面临着挑战。比如，Deepfake 虚假信息在互联网上发布后，很难追踪溯源。深度伪造的创作者在社交媒体上发布虚假信息，或以其他可访问的方式将其公布在互联网上时会努力做到匿名。在这种情况下，当个人或实体受到 Deepfake 虚假信息的伤害，根本无法对散布者进行追诉。网络虚拟性、匿名性的特征导致追踪发布者的现实世界身份困难重重，受害者只能根据相关法律法规，要求发布平台进行补救，而这些举措反而会促使涉及自身的虚假信息更进一步传播[13]。其次，现实世界存在执法辖区。网络连接在全球范围无处不在，但一般情况下，国家的法律法规执行范围只限于国境之内，任何的刑事起诉、民事诉讼和监管行动的效用都会受到地域限制，不能超出本国法规政策的有效辖区。另外，诉讼深度伪造相关责任人成本较大，甚至会引发二次伤害，且 Deepfake 取证困难。

为了应对上述挑战，我认为监管方可从下面 3 个层面进行提升：

1、立法机构要加快制定与深度伪造相关的民事与刑事责任确认、隐私和产权保护、信息安全评估等法律法规，建立追溯和问责制度，明确深度伪造的创作者、传播者、社交媒体等角色的相关权利、义务和责任。严厉打击利用 Deepfake

技术实施的各种违法犯罪活动，确保 Deepfake 技术整个生命周期的安全可控。

2、监管机构要全方位监测风险、引导发展和应对危机，构建动态的 Deepfake 技术评估评价机制，对衍生应用建立安全审查制度和定期检测制度，开展相关科学和伦理问题研究，建立多层次的社会伦理框架，引导多方主体参与治理。制定因深度伪造而引发政治、舆论、公共安全等突发事件应急方案。

3、网信部门要对企业、公众进行针对性的宣传教育，使其认识到传播 Deepfake 虚假信息造成的影响和危害，提高社会辨别不良信息的能力。引导公众形成批判性思维，不要轻易相信“眼见为实”，培养网民谨慎的网络信息获取习惯。

#### 4.2.2 积极采取主动式防御方法

在 Anti-deep 中曾提到，可以采取“事前防御”的主动式防御方法，即在人脸视频的发布源中加入干扰或标识，从而能更容易地将修改过的人脸视频识别出来。

目前的相关部门可以督促各大视频平台，采用类似技术，从而从源头上杜绝了 Deepfake 的滥用。当然，此项措施的范围无法涵盖全网，但对大部分的人脸视频能够起到保护作用。

### 4.3 被动式 Anti-deepfake

#### 4.3.1 模型融合

被动式 Anti-deepfake，即增强 deepfake 的检测技术。目前存在着各种各样的不同角度的检测思路，效果各有千秋。如果能借鉴集成学习的思路，将各式各样的检测器放置在一个模型中，通过 bagging 或 boosting 的方式，可能会提高检测效率。

#### 4.3.2 迁移学习

目前高质量的数据集较为稀缺，一定程度上制约了检测模型的效果。因此，可以使用迁移学习的思路，将其它任务的优秀模型作为预训练模型，从而弥补小样本的缺陷。例如，上文提到的唇语识别模型。类似的，还可以迁移表情识别、身份识别等具有良好研究基础的任务中的方法，从而获得更好的检测性能。

#### 4.3.3 多维度识别

目前大多数的 Deepfake 检测方法仅根据原始画面，然而，音频信息也可用来检测。比如唇语检测模型，就将视频信息和音频信息做了结合，提升了检测性能。此外，还可通过身份识别来构建另一维度的检测，通过人脸信息识别身份，再结合周围环境，判断出身份与场合的匹配度，从而做出更精准的判别。

对于公众人物来说，网络流传的信息更多，还可以根据公开的信息构建知识图谱，通过人物之间的关系，进一步检验模型判别的结果。

### 4.4 总结

Deepfake 和 Anti-deepfake 如同 GAN 模型中的生成器和判别器，只有两者的长期的互相博弈，才能推动技术持续进步。但 Anti-deepfake 同样需要外力的加持，确保 Deepfake 走持续健康的发展道路。

参考文献：
[1]Deepfakes. Deepfakes github. http://github.com/Deepfakes/faceswap [EB/OL], 2017. Accessed 2020-08-18.
[2]周文柏,张卫明,俞能海,赵汉卿,刘泓谷,韦天一.人脸视频深度伪造与防御技术综述[J/OL].信号处理:1-21[2021-12-07].http://kns.cnki.net/kcms/detail/11.2406.TN.20210930.1358.010.html.
[3]LI Lingzhi, BAO Jianmin, YANG Hao, et al. Advancing high fidelity identity swapping for forgery detection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition 1CVPR). Seattle, WA, USA. IEEE, 2020: 5073-5082.
[4]ZHU H, FU C, WU Q, et al. AOT: Appearance Optimal Transport Based Identity Swapping for Forgery Detection[EB/OL]. arXiv preprint arXiv:2011.02674, 2020.
[5]THIES J, ZOLLHÖFER M, STAMMINGER M, et al. Face2Face: real-time face capture and reenactment of RGB videos[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition 1CVPR). Las Vegas, NV, USA. IEEE, 2016:2387-2395.
[6]Siarohin A, Lathuilière S, Tulyakov S, et al. First order motion model for image animation[J]. Advances in Neural Information Processing Systems, 2019, 32: 7137-7147.
[7]Siarohin A, Lathuilière S, Tulyakov S, et al. First order motion model for image animation[J]. Advances in Neural Information Processing Systems, 2019, 32: 7137-7147.
[8]梁俊杰,韦舰晶,蒋正锋.生成对抗网络 GAN 综述[J].计算机科学与探索,2020,14(01):1-17.
[9]Mirza M, Osindero S. Conditional generative adversarial nets[J]. arXiv:1411.1784, 2014
[10]Denton E, Chintala S, Szlam A, et al. Deep generativ image models using a Laplacian pyramid of adversarial networks[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems, Montreal, Dec 7-12, 2015. Cambridge: MIT Press, 2015: 1486-1494.
[11]SHEN Yujun, GU Jinjin, TANG Xiaoou, et al. Interpreting the latent space of GANs for semantic face editing[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition 1CVPR). Seattle, WA, USA. IEEE, 2020: 9240-9249.
[12]HALIASSOS A, VOUGIOUKAS K, PETRIDIS S, et al. Lips don't lie: A generalisable and robust approach to face forgery detection [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 5039-5049.
[13]白国柱,王蓓蓓.Deepfake 技术监管政策现状和面临的挑战及建议[J].信息安全研究,2020,6(05):454-457.