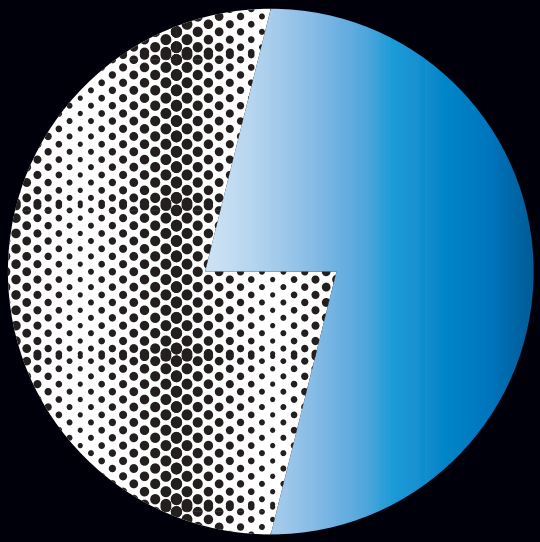
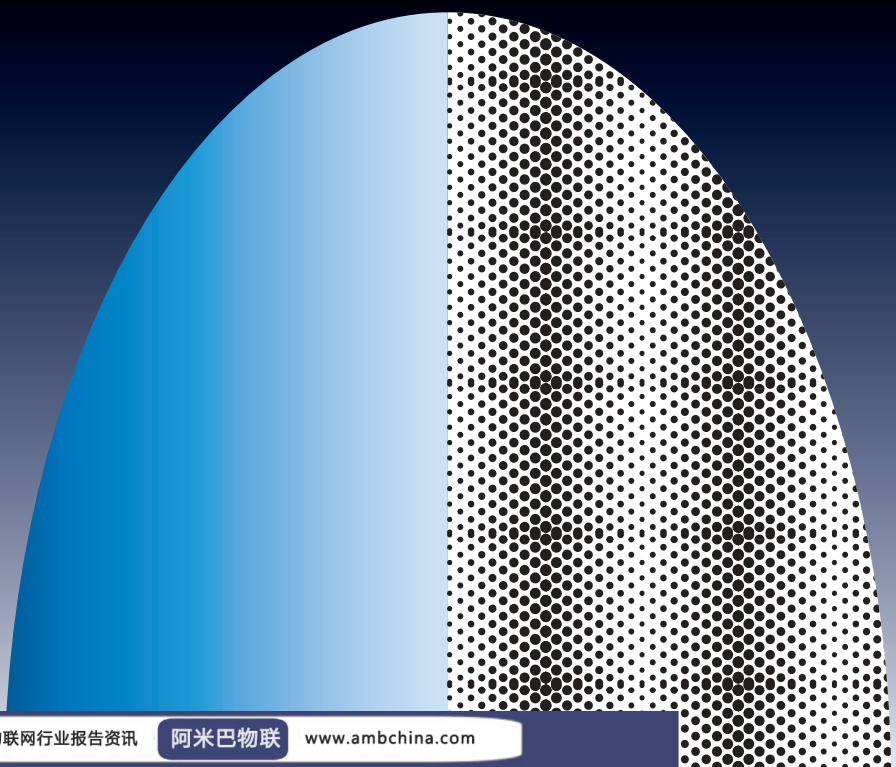
腾讯研究院 腾讯优图

AI生成内容发展报告 2020

“深度合成”(deep synthesis)商业化元年





腾讯研究院 X 腾讯优图实验室

目录

前言：

随着人工智能时代的到来，下一代媒体将由人工智能驱动，人工智能可能给数字内容领域带来重塑。近几年，人工智能合成内容（AI-generated Media）正在快速兴起，利用 AI 算法生产、修改数据和信息内容，从而让文字、音乐、图像、语音、视频等都可由 AI 自动生成。尤其是随着生成对抗网络（GAN）这一 AI 算法的诞生，AI 生成内容中的“深度合成”（deep synthesis）技术可以实现换脸、人脸合成、语音合成、视频生成甚至数字虚拟人等诸多应用，[1] 获得了各界的广泛关注。2017 年是深度合成技术进入大众视野的第一年，随后的两年里，随着 GAN 算法的发展应用和开源项目的增多，依托深度合成技术的诸多商业化应用问世，技术应用潜力逐渐显现。与此同时，深度合成技术强大的仿真能力也引发对技术作恶和技术滥用的担忧，例如金融诈骗、色情复仇、隐私侵权、商业诋毁，乃至威胁国家安全、公共安全。商业环境中，一方面，deepnude 等应用程序因为涉嫌侵权被下架或阉割，充斥着色情合成内容等黑灰产的 Reddit 论坛 deepfake 被关闭；另一方面，开源工具受到欢迎，诸多 AI 初创公司试水深度合成工具的开发，相继寻找下一个能引发爆点的应用，在通讯、社交、电影、网络购物等领域的尝试令人惊喜。回顾技术的发展历程可以发现，经历过技术初问世时的狂热追捧以及随之而来的“威胁论”和“恐慌论”，社会对深度合成技术及相关应用的认识逐渐趋于理性，而 2020 年也有望成为深度合成技术走向大规模商业化应用的元年。总之，深度合成技术是人工智能发展到一定阶段的产物，它不会让社会真相失守，更不是世界秩序的威胁者。面对新技术的挑战，政府和监管者应当包容审慎，避免阻碍深度合成技术的有益的、创新性的应用，通过法律、技术、行业、用户的多重治理将其纳入可控的发展轨道。

1. https://en.wikipedia.org/wiki/Synthetic_media

1.“深度合成”的概念

1.1“深度合成”的起源

“深度合成”作为一种 AI 合成内容（AI-generated media）技术，最早引起关注是在 2017 年，当时，美国新闻网站 Reddit 的一个名为“deepfakes”的用户上传了经过数字化篡改的色情视频，即这些视频中的成人演员的脸被替换成了电影明星的脸。此后，Reddit 网站成为了分享虚假色情视频的一个阵地。从那时起，新闻媒体开始使用“deepfake”一词来描述这种基于人工智能技术的合成视频内容。尽管后来 Reddit 网站上的 deepfake 论坛因为充斥着大量合成的色情视频而被关闭，但 deepfake 背后的人工智能技术却引起了技术社区的广泛兴趣，开源方法和工具性的应用不断涌现，例如 Faceswap、FakeAPP、face2face 等。[2] 而后来一些涉及普京、特朗普等政治人物的 deepfake 视频，更是将 deepfake 及其背后的技术推到了社会舆论的风口浪尖，欧美国家的政府机构开始积极跟进。

1.2 为技术正名：“深度合成”（deep synthesis）而非“深度伪造”（deep fake）

deepfake 在 Reddit 网站上兴起之后，国外媒体开始使用 deepfake 来泛指这类新型的合成内容及其背后的 AI 技术，国内媒体则一般根据 deepfake 这一合成词将这种技术翻译为“深度伪造”，但实际上，deepfake 这一用语并未得到技术社区的广泛认可，也未被正式纳入词典之中。相反，使用“深度合成”（deepfake synthesis）这一术语来描述可以实现换脸、脸部表情修改、人脸和语音合成等活动的 AI 合成内容技术，更加科学合理。

原因有二。一方面，“深度伪造”（deepfake）一词以偏概全，不足以涵盖所有的“深度合成”技术和相应的合成内容。如前所述，考察 deepfake 一词的起源可以发现，其本来含义是指向一种人工智能换脸技术，Reddit 网站的用户最早利用该技术来合成虚假的色情视频。而目前所存在的 AI 换脸技术除了 deepfake，还有 face2face 等。因此，利用特指“一种人工智能换脸技术”的“深度伪造”来泛指借助人工智能算法合成和自动生成语音、音乐、图像、人脸、视频等内容的所有“深度合成”技术，是以偏概全，既不专业，也不科学。

另一方面，“深度伪造”这一用语容易给相应的 AI 技术造成污名化影响，不利于技术发展应用。追根溯源，deepfake 一词最初只用于描述 AI 换脸的色情视频。虽然 deepfake 的出现让背后的 AI 技术获得了广泛的关注，但基于技术使用的意图（即 deepfake）去定义技术，强调技术的潜在欺骗性或可能带来的负面影响，这一做法并不公正。[3] 更进一步而言，deepfake 背后的技术具有很大的

2. https://en.wikipedia.org/wiki/Deepfake

3. https://www.google.com.hk/amp/s/www.theverge.com/platform/amp/2018/5/22/17380306/deepfake-definition-ai-manipulation-fake-news

正向应用价值，已被广泛应用于新闻传媒、电影制作、娱乐、电商、教育等诸多领域，如新华社的AI合成主播、网络上的虚拟歌手、社交媒体中的换脸应用等。2020年作为“深度合成”的商业化元年，将涌现出更多的应用场景。

因此，为了更加科学合理地认识deepfake背后的AI合成内容技术，本报告将deepfake翻译为“深度合成”，并使用“深度合成”（deep synthesis）这一术语来泛指相关的AI技术和合成内容。

2.“深度合成”的技术发展状况

2.1 技术原理

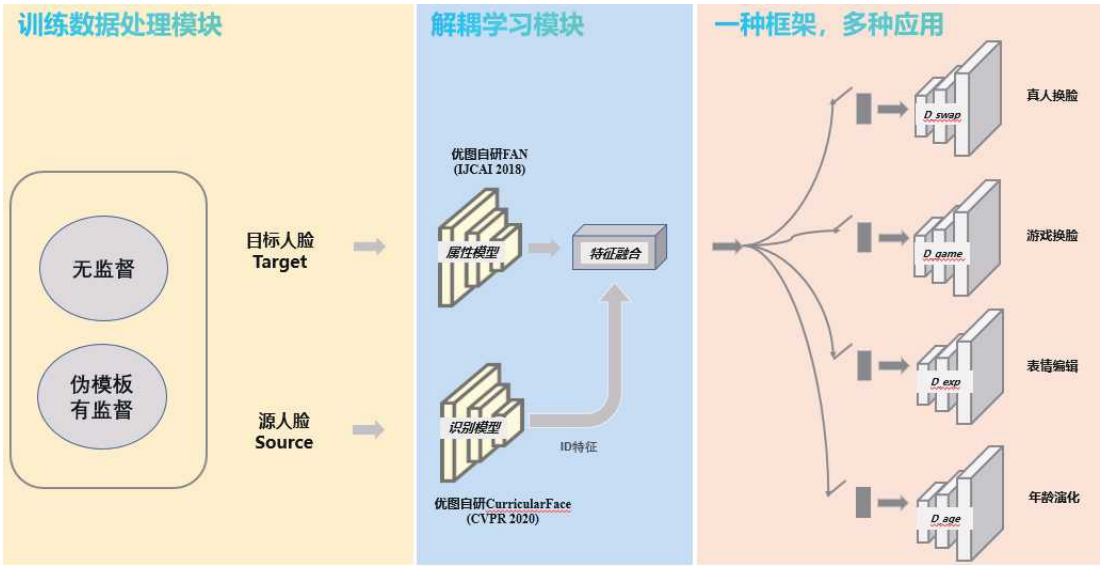
“深度合成”依赖于人工智能技术，尤其是可以从大量数据中自主学习的深度学习算法模型。[4]“深度合成”背后的AI技术主要包括自动编码机（autoencoders）和生成对抗网络（generative adversarial networks，简称GAN）。自动编码器是一个人工神经网络，被训练来对输入数据进行重建（reconstruct）以实现数据合成。GAN由两组相互对抗的人工神经网络组成，其中一个网络负责生成数据，另一个网络负责甄别。具体而言，在GAN的两个机器学习系统中，生成网络（generative network）或者说生成器（generator）负责制作复制了原始数据集特征的合成数据如图片、音频记录、视频等，鉴别网络（discriminative network）或者说鉴别器（discriminator）则负责识别合成的数据。基于每次迭代的结果，生成网络不断进行调整以创造越来越逼真、越来越接近于原始数据的新数据。生成器与鉴别器的竞争往往需要数千次或数百万次迭代，这让生成器不断改善其性能，直至让鉴别器不再能够区分真实数据和合成数据。[5]最终，GAN可以对视频中的人脸进行高度逼真的渲染，从而合成高保真的信息内容。虽然GAN通常可以生成高度逼真的合成视频，但使用起来却更加困难且复杂。

就“深度合成”的实现过程而言，总的来看可以分为三个步骤——数据提取、数据训练和转换，过程中需要用到图片融合等技术。以人脸替换为例，第一步是数据提取，这步需要收集足够的人脸图像，以便来训练算法模型。主流的方法是借助软件从视频中提取源人物（the source person）和目标人物（the target person）的多角度图像并裁剪出脸部肖像，形成脸部结构和头部尺寸相似的人脸图像。第二步是训练，即利用收集到的图像对人脸替换模型进行训练。模型训练通常用到两种AI技术，即包含编码器和解码器的自动编码机（autoencoder）这一神经网络，以及更复杂的生成对

4. 值得说明的是，并非所有的内容合成或者说内容生成应用，都用到了深度神经网络；但在内容合成或者说内容生成应用中使用深度神经网络等AI技术是发展趋势，因此本报告使用深度合成来指代类似的内容合成或者说内容生成应用。

5. Deep Fakes and National Security, https://crsreports.congress.gov/product/pdf/IF/IF11333

抗网络（GAN）。最后一步是合成，这是技术上最具挑战性的任务，需要将合成的图像插入视频中。这意味着要确保视频中的每帧合成图像的自然度和真实性，让合成人脸的角度与目标人物的头部角度完全一致。根据一些 AI 专家的看法，这是“深度合成”过程中唯一需要依靠手写代码而非端到端机器学习算法的阶段。[6] 正因如此，目前视频的深度合成难度远大于图像。下图为腾讯优图实验室的人脸深度合成模型训练框架。



图一：腾讯优图人脸深度合成模型训练框架

2.2 分类

就目前而言，典型的“深度合成”主要包括以下四种形式。

（1）人脸替换（face replacement）：也被称为换脸（face swapping），主要是指将某一个人的脸部图像（源人物）“缝合”到另外一个人的脸上（目标人物），从而覆盖目标人物的面部。

（2）人脸再现（face re-enactment）：主要是指利用深度合成技术改变人的面部特征，包括目标对象的嘴部、眉毛、眼睛和头部的倾斜，从而操纵目标对象的脸部表情。人脸再现不同于 AI 换脸，不是为了替换身份，而是改变某个人的脸部表情，从而让其看起来在说他们从未说过的话。

（3）人脸合成（face generation）：“深度合成”技术还可被用来创建全新的人脸图像。这些随机生成的人脸图像很多都可以媲美真实的人脸图像，有一部分可以代替一些真实肖像的使用，比如广告

6. CDEI, Snapshot Paper - Deepfakes and Audiovisual Disinformation

宣传、用户头像等。Generated Photos 就是一个 AI 自动生成人脸的网站，该资源库包含有 10 万张由人工智能生成的免版税人脸，该公司的免费图片可被用于网络及移动应用程序、教育、讲义、电子邮件与时事通讯、登陆页面以及用户头像等方面，只要求用户在使用时标明来源即可。

（4）语音合成（speech synthesis）：语音合成涉及创建特定的声音模型，不仅可以将文字转化成声音，而且可以转化为接近真人语调和节奏的声音。例如，加拿大的语音合成系统 RealTalk，与以往基于语音输入学习人声的系统不同，它可以仅基于文本输入生成完美逼近真人的声音。此外，Modulate.ai 的语音合成产品，允许用户自主选择任何年龄段和性别的语音模型，而不是模仿特定目标的声音。

以上四种深度合成形式在目前都存在一定局限，例如人脸替换需要对源图像和目标图像进行多角度的面部特征数据训练；面部表情操控实现的前提是目标人物面部直对镜头，并且能保持一段时间；而将语音合成技术自然融入运行的视频之中也较为困难。

2.3 技术发展趋势：技术演进加快，商业化应用开始流行

自出现在公众视野以来，深度合成无论是从技术成熟度还是社会影响力来看都在迅速发展。

技术方面，“深度合成”背后的核心 AI 技术 GAN 从 2017 年以来持续受到高度关注，发布在 arXiv 上的论文数量连年保持高速增长，为“深度合成”的进一步发展和规模化应用奠定技术基础。[7]

在单一的音频、图像合成之外，深度合成技术朝着综合性的方向发展。去年，来自斯坦福大学、德国马克斯普朗克信息学院、普林斯顿大学和 Adobe 研究院的研究人员融合了多种深度学习方法，包括语音识别、唇形搜索、人脸识别和重建，以及语音合成，实现了根据输入的文字即可改变目标人物口型的技术效果，让视频的修改更加自然。[8] 而为了完成更加复杂的视频合成，卡内基梅隆大学的研究人员在其最新的论文中介绍了一种新的视听合成技术——“多对多音视频转换”，这种技术不仅能让给定的音频与目标对象的唇动相匹配，还能识别同一媒介中多主体的声音特征，同时处理不同身份的人的声音变化，视频综合处理技术有了显著进步。[9]

其次，面部合成之外，全身合成将成为新热点。在 AI 换脸一度引发体验热潮的同时，合成媒体的分支——针对全身的“深度合成”也悄然而至。2018 年加州大学伯克利分校的研究团队在其论文中介绍了一种新的 AI 算法，这种算法可用于学习源人物的舞蹈动作并映射到目标人物之上，让一

7. Deeptrace, the State of Deepfakes: Landscape, Threats, and Impact
8. https://www.huxiu.com/article/303626.html
9. https://medium.com/deeptrace/tracer-41-20-01-20-carnegie-mellon-researchers-publish-a-new-technique-for-transforming-a-live-de05e02523a1

个不会跳舞的人也能呈现出近乎专业的舞蹈动作。[10]

另外，在 2D 合成之外，3D 合成技术尤其是虚拟人（数字人）技术将是下一阶段的重点。以往的数字虚拟形象，需要运用三维动画、红外感应、大屏显示等技术，通过拍摄真实人物，结合三维后期技术制作再投射在屏幕上。而现在，数字王国独立研发的虚拟替身——Digi Doug，让这一技术迈向实时 3D 渲染。Digi Doug 是一个高度仿真的立体虚拟形象，能够实现对真人动作的实时捕捉和渲染，实时反应原人物的表情、动作甚至睫毛的抖动，与真人的动作只有 6 分之一秒的延迟。[11] 虚拟人技术不仅降低了制作成本，更重要的是在保证仿真效果的同时能实现同步映射。

技术发展的同时，“深度合成”正在通过越来越多的计算机应用程序和服务走向商业化、工具化应用。**商业化首先得益于源代码的开放。** Faceswap 的源代码很早就被创建者匿名捐赠给开源平台，并上传到 GitHub，开源代码工具驱动了诸多基于合成技术的分支研究，针对非技术受众的深度合成创作工具日益增多。对于开发人员而言，开源代码的发布使其能够借鉴其他人提出的想法，或者与其他技术人员合作，从而不断克服来自新技术的挑战。其次，**合法的“深度合成技术”论坛又为使用这些工具提供了交流平台。**大多数开放源代码工具都需要一定的编程知识和强大的图形处理器才能有效运行，但是即使是非专业技术人员，也可求助于一些线上技术交流论坛和讨论组获得详细的使用教程。根据荷兰一家网络安全公司 deeptrace 统计，仅统计中 13 个公开的 deepfake 论坛，其总会员数就有大约 10 万人。[12]

在这样的背景下，2018 年以来“深度合成”技术被许多公司争相商业化应用，包括诸多 AI 初创公司，也包括一些发展成熟的互联网企业。有些利用 AI 换脸体验吸引用户，进行产业布局，例如 ZAO；有些则在网上出售在线生成的“深度合成”图片或音视频。可以预见，随着 AI 换脸、语音合成、数字虚拟人等“深度合成”的不断成熟和技术门槛的持续降低，未来几年“深度合成”将迎来商业化时代，创新性的应用形式将持续涌现。

3.“深度合成”的商业化应用

短期内，“深度合成”技术已经作用于影视、娱乐和社交等诸多领域，它们或是被用于升级传统的音视频处理或后期技术，带来更好的影音体验；或是被用来进一步打破语言障碍，优化社交体验。中长期来看，“深度合成”技术既可以基于其深度仿真的特征，超越时空限制，加深我们与虚拟世

10. Everybody Dance Now, https://arxiv.org/abs/1808.07371
11. https://vrroom.buzz/vr-news/people/meet-digi-doug-doug-robles-digital-twin
12. Deeptrace, the State of Deepfakes: Landscape, Threats, and Impact

界的交互，也可以基于其合成性，创造一些超越真实世界的“素材”，比如合成数据，为科研或生产所用。具体而言，深度合成技术已开始在如下领域中使用，2020 年更多商业化应用也将从这些领域涌现出来。

3.1 影视：升级后期制作与扩展角色创作空间

深度合成之于电影，一方面可用来升级音视频剪辑技术，为影视制作中的特效、配音呈现更好的效果，减轻视频编辑人员的工作压力；另一方面还可以减少因为演员、拍摄场景的局限，拓展电影的创作空间，衍生出更多改编作品。例如，它可以创建演员的声音模型，帮助为因疾病而失去自己声音的演员使用数字声音继续表演，或因为剧情需要改变角色的台词，还可以自动执行各种语言的逼真配音，从而使不同的受众群体能够更好地欣赏电影。在 2017 年《星球大战 8：最后的绝地武士》中，莱雅公主的饰演者凯莉 · 费舍尔因心脏病突发逝世，在电影中，制片方利用以往的真实录音合成了更多台词，结合未公开使用的素材，延续了这一角色的“生命”。深度合成技术还可用于“数字复活”已故演员，重现电影中的经典场景或角色。2016 年，《星球大战外传：侠盗一号》就通过合成技术重现了 1977 年《星球大战 4》中总督威尔霍夫 · 塔金的形象。而在《速度与激情 7》中，结合 CGI 动作捕捉技术，保罗 · 沃克被“数字复活”，在电影之中完成了角色的谢幕。[13] 此外，借助深度合成技术，影视制作方也可以针对既有内容拍摄续集、进行衍生和适应不同文化的改编等。

3.2 娱乐：深度合成技术带来全新娱乐体验

2017 年至今，包括 FaceAPP、Snapchat、Face2Face、ZAO 等图像、视频合成应用在国内外市场上反响热烈。例如 FaceApp 推出的老年滤镜、婴儿滤镜，会自动添加皱纹、白发和皮肤松弛等衰老迹象，也能基于现有面部特征合成幼化的面孔，同时保证照片的真实感。ZAO 则主打视频换脸，利用特定的影视化素材进行表情迁移和头部姿势迁移，让用户与明星同框，体会“表演”的乐趣并能够展示自我。社交平台 Snap 此前和 AI Factory 合作增强 Cameos 功能，让用户可以将自拍嵌入 GIF 图像，从而创造出合成的动图，而且越来越多的网络平台开始推出类似功能。[14] 腾讯旗下网络游戏“和平精英”，也引入了深度合成应用，玩家可化身游戏中的“和平精英”与火箭少女 101 同框合影，背后所依托的正是腾讯优图实验室的人像融合技术 DittoGAN：通过自研深度解耦学习（Deep Disentangle Learning）框架实现对人脸身份特征和属性特征的精细化建模；伪模板图像生成技术（Pseudo-Template Image Generation）制作带真实标签的训练数据对，实现有监督学习，进一步提升图像的恢复质量；真实人像超分辨率 (Real-world Face Super-Resolution) 框架实现高清人像编辑与生成。这一技术还能用于 H5 等活动营销传播、游戏影视制作、以及各类相机 APP

13. Bobby Chesney and Danielle Citron, Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*
14. CB Insights, Artificial Intelligence Trends To Watch in 2020

和小程序，带来更多富有创造性的体验。[15] 这类应用玩法简单、体验流畅，还能将换脸后的照片或视频分享到交友平台，满足了用户的好奇心与新鲜感。这些虽然是深度合成技术较为基础的应用方式，但为大众了解这一技术提供了很好的机会。此外，深度合成技术还可以创造虚拟主播来播报新闻、天气预报等内容，创造出虚拟歌手来演唱歌曲（如虚拟歌手洛天依等）；也可以给医疗、零售、娱乐等领域提供更具亲和力的人形问诊机器人、虚拟客服、虚拟偶像等。总之，未来的应用场景十分广泛，值得期待。

图二：腾讯优图的人脸融合技术在游戏中的应用

3.3 教育：更生动的教育工具

深度合成之于教育的意义，类似于早期的视频等多媒体教学带来的教育创新浪潮。相对于阅读和讲座等传统方式，深度合成技术为教育工作者提供了新的工具，能以更加生动、更加令人信服的方式向学生传递知识，例如制作历史人物直接与学生对话的视频，给一场毫无吸引力的演讲注入新的活力；或者在多媒体视频的播放途中，改动某部电影或节目中的一个场景，插入知识点的讲解，强化学生印象。此外，深度合成技术可以合成逼真的虚拟教师，让数字教学更具互动性和趣味性。深度合成技术的教育价值还会扩展到课堂之外，发挥其科普和启发民众的功能，比如作用于公益广告。2019 年大卫 · 贝克汉姆拍摄的疟疾宣传广告中，通过视觉和语音合成打破了语言障碍，让其在广告中看起来能自如地使用九种语言进行公益宣传，提升了宣传效果。在教育领域运用深度合成技术，可能涉及到关于知识产权保护和合理使用豁免的问题，但不可否认这一技术将激发更大的教育效益。

15. https://mp.weixin.qq.com/s/lvBV31XuZE6K6sp2Pv788A

3.4 通讯和社交：更真实的虚拟互动

深度合成技术能用于深度模拟个体的生物特征，从而可为个人创造虚拟化身（avatar），并能通过虚拟化身参与媒体中各类社交活动，创造出超越现实的个性化体验和化身体验。在这个意义上，深度合成技术可以带来更真实的虚拟体验，产生更强大的化身效果和代入感，以实现各种目的的自我表达和体验。例如在线上会议中，结合语音识别与机器翻译技术，同时改变参会者呈现出来的脸部表情和嘴巴动作，可以进一步打破语言交流障碍，使每个人看起来像在自然地使用相同的语言进行交流，发展更好的人际关系和线上互动。更进一步，伴随着人体深度合成的发展，这一技术还将参与到虚拟现实技术中去。虚拟现实需要以人工智能为基础，而深度合成所具备的仿真能力完全符合虚拟现实的需要。使用人像数据创建 3D 模型，并实时更新眼神、表情及肢体语言，进一步增强虚拟与现实的交互，这一技术已经在腾讯旗下的腾讯会议中使用。未来，随着数字虚拟人技术的发展成熟，将其与 VR 等技术结合，将带来前所未有的社交体验。

3.5 电子商务和内容营销：优化服务和升级商业模式

商业领域中，深度合成技术还可以改变广告宣传方式和商品交易过程，可针对消费者动态的决策过程提供更精准的营销。比如为广告进行自动、逼真的配音，让不同地方的观众获得更好的观看体验和宣传效果，扩大宣传受众。[16] 而在未来，合成视频制作门槛的降低，还可能使广告投放更加个性化，比如根据大数据需求分析，结合时令季节与流行趋势定向合成有针对性内容的视频，从而有效吸引用户，拓展市场。[17] 声音合成技术也能在商品宣传中发挥作用，在商标资源有限的情形下，声音商标已经得到了多国法律的承认。深度语音合成可以创造独特的人工合成声音，发挥指引消费者认牌购货的功能，扩大声音商标的适用。另外，深度合成技术可以提升线上购物的体验，品牌可以使用肤色，身高和肤色各异的虚拟模特展示服装效果，而不用再雇佣摄影师和专业模特，甚至可以鼓励消费者自己进行数字建模，预览服装的上身效果，实现快速的数字试衣、试穿，进一步增强网购的个性化体验。[18] 例如，AI 初创公司 Superpersonal 已经在探索将用户的脸换到短的视频片段中，从而在购买前实现虚拟试穿。此外，在广告宣传、内容营销等领域，AI 合成的人脸和虚拟形象可以替代网红、模特等，既能带来新鲜感，也免去了传统上使用他人肖像的授权。

3.6 艺术：内容创意升级

戏仿与讽刺是通过对原艺术作品的模仿和变形达到喜剧的、讽刺目的的艺术手段或类型。将深度合成技术用于创造新形式的戏仿与讽刺具有很强的冲击力，包括但不限于合成明显与演讲者的语言、

16. 例如，初创公司 Synthesia 的“深度合成”技术让大卫 • 贝克汉姆看起来可以用 9 种语言来宣传抵抗疟疾的信息，不仅降低了后期配音成本，还能增强广告的宣传效果。这还进一步增加了广告商的市场。

17. https://mp.weixin.qq.com/s/oHfG3m_LtscubWUI2IELsg

18. U.S. CAO, SCIENCE & TECH SPOTLIGHT: Deepfakes, https://www.gao.gov/products/GAO-20-379SP

身份不协调的作品，讽刺、模仿、批评公众人物和社会事件，这是以往的作品无法做到的。[19] 以色列公司 Canny.ai 曾经制作了一个关于扎克伯格的合成视频，在视频中，扎克伯格十分冷静的说道，“想象一下，有一个人，完全控制着数十亿人被盗的数据，他们所有的秘密，他们的生活，他们的未来……谁能掌控数据，谁就能掌控未来。”[20] 这一基于深度合成的讽刺作品无疑进一步提升了公众对平台数据隐私管控能力的关注度。更具有创意的是，深度合成技术还曾经被用来提示技术本身的风险。Buzzfeed 网站制作的“Deepfake Obama PSA”是一个早期的深度合成作品，这段视频曾半天内在推特上收获了 200 多万次播放，1300 多条评论和 5 万多个赞。[21] 视频中奥巴马受合成技术驱动说到，“我们已经进入了这样一个时代，我们的敌人可以做出看起来像任何人在任何时候说任何话的东西。”[22] 这个视频的目标是警示社会深度合成技术可能带来的危险。此外，利用深度合成技术还可以将平面艺术作品加工成富有趣味的动态视频，俄罗斯 Skolkovo 研究所曾利用人工智能算法学习多名志愿者人脸特征数据，为达芬奇著名画作主角——蒙娜丽莎制作动态表情，在保留志愿者不同面部特征的基础上，融合蒙娜丽莎的外貌形象，使蒙娜丽莎具备不同表情，呈现出不同的“个性”。

3.7 医疗：辅助诊断治疗康复全过程

AI 合成技术也可介入治疗和康复。去年，语音合成软件制造商 Lyrebird 为渐冻症患者设计一套新的语音合成系统，用患者自身的语音数据替换以往机器合成的语音，这意味着今后对于可能有失声风险的患者，都有可能通过提前采集自己的语音，获得能继续用“自己的声音”交流的机会。[23] 此外，该技术还可以用于帮助老年痴呆症患者与他们可能记得的年轻面孔互动或数字化地再现截肢者的肢体。对于心理治疗而言，通过合成面部而非自己的真实面部去接受心理咨询能减轻咨询者的心理负担，而医生仍然可以从他们的面部表情中捕捉患者的情绪，降低对治疗质量的影响。同时，据《MIT 科技评论》报道，吕贝克大学的研究人员已利用人工智能自动合成技术合成了与真实影像无异的医学图像，解决了没有足够训练数据的问题，而这些图像将可以用于训练 AI，优化针对罕见疾病的检测算法。

3.8 科研：合成数据打破数据壁垒，助力 AI 模型训练

深度学习的发展越来越依赖高质量的数据“投喂”，然而很多场景下获取真实数据成本太高。基于深度合成技术的合成数据就是很好的 AI 训练资源，很有可能打破数据壁垒，让更多基础研究和商业应用获得想要的数据样本。2019 年，科学家们在模拟环境下已经可以完成一些数据构建和高质

19. Bobby Chesney and Danielle Citron, Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*
20. https://www.huxiu.com/article/303626.html
21. https://tech.sina.com.cn/csj/2018-04-19/doc-ifzihnep7688648.shtml
22. 视频见 https://www.bilibili.com/video/av66254375
23. http://dy.163.com/v2/article/detail/EFVQG8C6051280SH.html

量的数据训练工作。比如 Amazon 旗下的虚拟助手 Alexa，送货无人机和机器人都在使用模拟数据进行训练。[24] 合成数据还可用于一些存在风险的作业，如自动驾驶领域，国内外都在尝试开发自动驾驶仿真系统（AADS）：根据使用激光雷达（LiDAR）和相机扫描街景获得的轨迹数据，生成较为合理的汽车和行人交通流，为自动驾驶系统的训练和测试创造虚拟环境。[25] 例如腾讯自动驾驶团队就开发了自动驾驶仿真系统 TAD Sim。此外，在医疗领域，利用深度合成技术可以合成 MRI 图像，为训练 AI 诊断系统提供可供投入学习的数据库。医学影像 AI 诊断系统的训练需要投入一定量的影像数据，而实践中数据主体的隐私、经济状况、罕见病例数据等原因都会干扰甚至阻碍 AI 诊断系统的开发。在 NVIDIA，MGH&BWH 和 Mayo Clinic 联合发表的论文中，则展示了利用 GAN 算法合成带有肿瘤的脑部 MRI 图像的方法，在算法训练生成过程中，仅需投入 10% 的真实数据，经训练后的 AI 诊断系统可以准确检测出真实影像中的肿瘤。[26]

“深度合成”作为人工智能技术发展到一定阶段的产物，自 2017 年 11 月首次进入大众视野以来，经过过去两年的喧嚣和争议，随着技术的成熟，在 2020 年新的应用开始大量涌现，可谓是商业化应用元年。可以确定的是，深度合成技术在内容创意、营销、社交、娱乐、电商、通讯等诸多领域应用前景广阔，其未来应用令人期待。这也说明，深度合成并非关于“伪造”和“欺骗”的技术，而是极富创造力和突破性的技术，虽然它像其他任何技术一样，也催生了一系列必须面对的难题，但这并不会磨灭这一技术给社会带来的进步。

4.“深度合成”的治理

4.1 深度合成的风险

风险客观存在，需要得到正视。深度合成领域已经出现的一些滥用现象如明星换脸的色情视频、虚假的政治人物视频等，引发了人们对深度合成技术作恶的担忧。一般而言，深度合成技术带来的新挑战主要表现为，利用深度合成技术伪造或合成真假难辨的图片、音频、视频等影像资料来进行欺骗和欺诈等非法活动，将网络攻击场景和信息安全问题带到一个全新的层面，如色情报复、敲诈勒索[27]、假冒身份、商业诋毁、散布虚假信息、非法获取个人信息[28]、虚假情报、选举干扰、外交和国际秩序扰乱等，给个人和企业利益以及国家安全、公共安全带来威胁。而且深度合成开源方法和软件的增多，极大地降低了操纵、伪造音视频的门槛。以合成视频为例，报告显示，到 2019 年 12

24. https://www.guokr.com/article/442864/
25. https://tech.sina.com.cn/csj/2020-01-02/doc-iihnzahk1419898.shtml
26. https://arxiv.org/abs/1807.10225
27. 例如，2019 年，一位英国能源公司的 CEO，因为相信了这种根据其上司的声音合成的音频，被骗 22 万欧元（折合 24.3 万美元）。
28. 例如，通过 3D 合成“假脸”认证账号注册或登录后，不法分子可在受害人毫不知情的情况下，用于黑卡虚假注册、刷单、薅羊毛、诈骗等不法行为。

月的时候，网上的合成视频的总数比 2018 年 12 月翻了一番，达到近 15000 个，其中合成的色情视频占比高达 96%，深度合成已经成为了色情复仇的重要工具。[29] 而早在 2016 年，欧洲刑警组织《网络有组织犯罪威胁评估报告》中，就已经点明了 deepfake 犯罪将成为趋势。

长远来看，由于这些负面影响，这一技术有可能演变为“不信任”的符号，所产生的文化影响可能远大于技术影响。任何人都可以使用深度合成技术来质疑他们不喜欢的事情，诋毁原本真实的证据，这意味着在将来我们不仅需要证明什么是假的，还需要证明什么是真的。[30] 技术滥用的风险已经摆在我们面前，然而技术本身没有善恶，开发人员也一直遵循严格的道德标准，许多研究者甚至从未使用这项技术制作带有负面影响的合成内容。[31] 由此，深度合成的未来不仅取决于技术的发展水平，也取决于在实践中的规范和引导，更依赖于公众信息分辨能力的进一步提升。对深度合成技术的治理，体现在法律、技术、市场、行业自律等层面。

4.2 法律方案

域外立法方面，美国最为积极。美国情报界发布的研究报告《2019 年全球威胁评估》认为，deepfake 技术已经对美国国家安全构成威胁，敌对势力和战略竞争对手很有可能企图利用深度合成技术或类似的机器学习技术，创造出高度可信但却完全虚假的图片、音频和视频资料，以加强针对美国及其盟友和合作伙伴的影响渗透运动。美国一些政治人物的虚假合成视频在 Facebook 等社交媒体上的广泛流传更是加剧了美国政府对该技术之滥用的担忧。在此背景下，美国开始提出监管措施，尤其是在美国 2020 年大选来临之际，美国议会先后提出了《Deepfakes 责任法案》(Deepfakes Accountability Act，法案全称为 Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability Act of 2019）[32] 和《2019 年 Deepfake 报告法案》(Deepfake Report Act of 2019）[33]，旨在限制深度合成技术的不当利用，防范外国竞争对手利用该技术散布虚假信息，干涉选举活动。主要包括以下措施：

其一，不得误导，要求披露人工智能身份。美国加州 2018 年的一项法案要求，AI 机器人以商业或政治目的与人交流或互动时，必须披露其是人工智能，不能误导人。

其二，划定红线，禁止政治干扰、色情报复、假冒身份等目的的深度合成。例如，2019 年 10 月生效的美国加州 AB-730 法案禁止在选举前 60 天内制作、散布任何经过篡改的针对竞选公职人员的

29. https://edition.cnn.com/2019/10/07/tech/deepfake-videos-increase/index.html
30. https://www.google.com.hk/amp/s/www.theverge.com/platform/amp/2018/5/22/17380306/deepfake-definition-ai-manipulation-fake-news
31. https://blog.csdn.net/ConferenceEye/article/details/104059899
32. https://www.congress.gov/bill/116th-congress/house-bill/3230/text
33. https://www.congress.gov/bill/116th-congress/senate-bill/2065

候选人的恶意音频或视频。[34] 美国得克萨斯州 2019 年 9 月 1 日生效的一项法案将在选举前 30 日内制作、散布 deepfake 视频意图伤害候选人或影响选举结果的行为，认定为刑事犯罪。美国弗吉尼亚州 2019 年 7 月 1 日生效的反色情复仇修正法案，将“制作、传播虚假的裸体或性视频、图像”意图胁迫、骚扰或恐吓他人的行为，认定为刑事犯罪，把发布和传播 deepfake 视频视为实施色情报复的手段之一，违法者将面临最高 12 个月的监禁和 2500 美元的罚款。美国加州 AB-602 法案禁止未经同意制作 deepfake 色情内容的行为。[35] 2018 年美国纽约州的 A08155 法案将未经同意使用他人肖像制作合成色情视听内容的行为定义为侵权，行为人有责任向受害者支付赔偿金。[36] 该法案招致了以迪士尼，漫威为代表的影视公司以及行业协会的强烈反对，他们认为这可能违反保护言论自由的第一修正案，影响作品的创作力和想象力。[37] 此外，《Deepfakes 责任法案》规定，利用深度合成技术实施数字冒名顶替行为也应视为假冒身份行为。

其三，设置披露义务，要求制作者以适当方式披露、标记合成内容。《Deepfakes 责任法案》规定，利用深度合成技术合成虚假内容放置于网络上传播的，制作者应当采用嵌入数字水印、文字、语音标识等方式披露合成信息。违反披露义务的制作者或者恶意删除披露信息的行为人需承担民事责任，行为恶劣、造成严重后果的还会面临罚金、人身监禁等刑事处罚。

其四，加强技术攻防，呼吁开发检测识别技术。《Deepfakes 责任法案》要求成立 deepfake 特别小组，其职责包括研究开发针对包括深度合成在内的图像、音视频操纵技术的检测识别和反制技术、为研究此类技术的其他政府部门提供行政和科学支持、与私营企业或学术机构合作开发检测识别工具等。此外，《2019 年 Deepfake 报告法案》要求美国国土安全部（DHS）定期发布关于 deepfake 技术的评估报告。

美国以外，其他一些国家也在积极探索深度合成技术的治理之道，尤其是对音视频形式的虚假信息（disinformation）的规制。2018 年 4 月 26 日欧盟委员会发布《应对线上虚假信息：欧洲方案》,《方案》集中阐释了欧盟委员会面对线上虚假信息挑战的基本观点，提出改进信息来源及其生产、传播、定向投放和获得赞助方式的透明度，改善信息的多样性，提高信息的可信度，制定包容性的解决方案等原则，以实现全面防范视频、图像和文字等虚假信息，避免信息发布者违法操纵舆论等状况。[38] 此外，欧盟正式实施的《通用数据保护条例》(GDPR）通过对个人信息的保护，也能规范深度合成技术，即 GDPR 框架适用于可能被用于制作深度合成内容的公民图片等数据，并适用于社交媒体平台和软件公司发布的换脸软件产品等可能引发的个人隐私泄露问题。在英国，恶意制作深度合成视频者可能会被起诉，但是现在还没有一个特定的罪名，部分议员提出扩大《偷拍裙底

34. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB730
35. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB602
36. https://nyassembly.gov/leg/?default_fld=%0D%0A&leg_video=&bn=A08155&term=2017&Summary=Y&Actions=Y&Text=Y
37. https://www.theregister.co.uk/2018/06/12/new_york_state_is_trying_to_ban_deepfakes_and_hollywood_isnt_happy/
38. https://www.secrss.com/articles/17221

法案》(Upskirting Bill)的调整范围，将利用深度合成技术制作合成色情视频（即 deepfake 色情内容）列入性犯罪行为，[39] 但最终通过的法案并未包含针对 deepfake 的条款；[40] 此外，英国 2019 年的《在线危害白皮书》(Online Harms White Paper）旨在鼓励社交媒体平台对其平台上发布的内容负担更大责任，提议给科技公司施加新的注意义务来应对一系列的在线危害，其中就可能包括音视频形式的虚假信息的传播。对于深度合成技术，一些国家选择优先通过解释现行法律加以规范，例如通过既有的著作权法、隐私法、侵权法甚至刑法等予以规制；加拿大现行的法律框架下，未经许可的合成行为可能落入著作权侵权的范畴，而一些深度合成的虚假内容还可能涉及诽谤、诈骗等；此外，根据加拿大《魁北克民法典》，将一个人的姓名、肖像、声音等公民的合法信息以外的目的属于侵犯隐私的行为。

就我国而言，深度合成技术及其应用也已经引起我国立法的高度重视。中国国家互联网信息办公室《网络信息内容生态治理规定》已经于 2020 年 3 月正式施行，其中第 23 条规定，网络信息内容服务使用者和网络信息内容生产者、网络信息内容服务平台不得利用深度学习、虚拟现实等新技术新应用从事法律、行政法规禁止的活动。这一规定表明了我国对深度合成技术的态度——给“深度合成”技术划定了应用边界，同时为行业探索有益应用场景留出了发展空间。除了此次《规定》提出的概括性禁止要求，《民法典人格权编（草案）》第 799 条、《数据安全管理（征求意见稿）》第 24 条，以及已发布的《网络音视频信息服务管理规定》第 10-13 条等都作出了规定。此类规定主要包括：禁止利用信息技术手段伪造的方式侵犯他人的肖像权和声音；针对自动合成的信息内容标明“合成”字样；上线相关功能或服务需开展安全评估，对非真实音视频信息予以标识，禁止深度合成的虚假新闻消息，部署鉴别技术并在对非真实音视频信息进行标识后方可继续传播，建立辟谣机制，等等。[41]

4.3 技术方案

法律措施并非对深度合成技术进行治理的唯一方式，且单靠法律的规制难免会滞后于技术的发展与演变，还可能矫枉过正，扼杀技术的潜在社会福利，最终不利于技术的社会经济价值的发挥。因此，用于实现技术治理的技术手段显得同样重要。

首先是鉴别技术。随着深度合成技术的进化，相应的鉴别技术也在同步发展，以期能够迅速鉴别并可靠标记 AI 合成内容，从而从传播媒介上阻止负面的虚假内容的扩散和蔓延。在这方面，司法实践中对证据真伪进行判断的司法鉴定可以提供一定帮助。但是，学术和商业界的防伪开发项目多针对特定产品而非通用的音视频，即需要针对每一种新兴的视频内容篡改技术都训练一个对应的鉴别

39. https://www.theguardian.com/world/2018/jun/21/call-for-upskirting-bill-to-include-deepfake-pornography-ban
40. https://www.parliament.uk/business/news/2019/february/royal-assent-upskirting-bill-signed-into-law/
41. 张钦坤、曹建峰：《从 < 网络信息内容生态治理规定 > 看互联网新技术的治理》，载《信息安全与通信保密》2020 年第 2 期。

网络，还没有通用性的视频鉴别网络。正如 Photo DNA（识别和屏蔽儿童色情图片的技术）技术开发者、达特茅斯学院教授 Hany Farid 所说，我们距离能够确切地识别深度合成内容的鉴别技术还有几十年的路要走。这意味着需要加大对通用且高效鉴伪技术的研发投入和支持力度。例如，美国国防部高级研究计划局（DARPA）目前有两个项目致力于深度合成内容的检测鉴别——媒体鉴定(Media Forensics) 和语义鉴定（Semantic Forensics）。媒体鉴定项目计划开发一个算法平台，自动评估照片和视频的完整性，并向分析师提供有关假冒内容是如何生成的信息。据介绍，媒体鉴定平台将使用三种重要的指标去识别不一致的视觉媒体：像素不一致（数字完整性）；物理不一致（物理完整性）；以及与其他信息源不一致（语义完整性）。媒体鉴定项目在 2019 财年获得了 1750 万美元的扶持，随着该计划开始过渡到作战指挥和情报界，在 2020 财年还将获得 530 万美元的扶持。与此类似，语义鉴定平台试图开发一种媒体信息的自动识别方法，检测并识别不寻常的信息或面部特征。无论是媒体鉴定还是语义鉴定，两者都是为了提高识别和对抗虚假信息的能力。[42] 在国内，腾讯优图实验室也在研发人脸合成检测技术，该技术主要从图像结构完整性、整体分布合理性、局部纹理一致性三个层面进行递进式鉴定，结合数字图像处理和深度学习技术，完成对照片、视频合成与编辑的检测。

其次是溯源技术。深度合成技术的检测与反检测逐渐演变成一场猫捉老鼠的技术竞赛，深度合成技术正在快速更新以逃避检测工具的识别。因此，还需要从内容源头上区分真实内容与合成内容，确认内容的来源包括制作者、制作的地点等。有一种设想是提供一种标记方法，要求用户在特定位置标识内容的原始来源或注明内容是否已经过编辑。但是标记和分类的前提是用户或平台能够确定特定内容是否真实，而运用区块链技术进行内容识别被认为是一种有效的解决方案。目前市场上一些语音合成和视频合成设备的制造商已经使用各种时间戳工具，在由特定设备创建的图像和视频上添加数字水印或数字签名，用来记录图像或视频是在何时何地拍摄的，用的什么设备等信息，这些可以用来检测某个文件是否随着时间的推移而被修改。例如 Signed at Source 是一家提供区块链服务的初创公司，目标是保护物联网数据的完整性，它支持为用户改装音视频硬件设备，在音视频设备中加入加密签名数据控制器来证明特定的内容是由该设备所拍摄的，以此来对抗后期的篡改。[43] 要让这些数字水印或者数据被认可，还需要通过架构行业公有区块链或私有区块链来共享可信的数据记录，并利用区块链的不可篡改性来确保数据的完整性。当然，开发这种复杂的技术解决方案需要大量的投资和测试。

4.4 行业自律

深度合成技术的治理也离不开行业自律做法的支持。例如，谷歌（Google）、脸书（Facebook）等美国主流科技公司也开始担忧深度合成的违法内容在其社交媒体平台上广泛传播的可能性，及其潜

42. Deep Fakes and National Security, https://crsreports.congress.gov/product/pdf/IF/IF11333
43. https://iot-sas.tech/

在的不利影响。并通过多种举措，发展甄别 AI 合成内容、对抗深度合成技术滥用的方法和工具。

一是构建并开放深度合成数据集，为研究、开发检测识别技术提供基准。随着深度合成技术的不断进化和日益复杂，相应的检测识别技术的开发也在加紧步伐，但这离不开大量的高质量数据集的支持。为此，美国主流科技科技公司已在积极构建并开放深度合成数据集。2019 年 9 月，谷歌推出了自主开发的大型深度合成视频数据集，该数据集现已被纳入谷歌赞助的 Face Forensics 视频基准测试数据集中，供研究社区免费试用，研究者可基于开源的数据集开发检测合成视频的方法。随着深度合成技术的进化，谷歌后续将增加新视频到这个数据库中，并与合作伙伴持续加深这一领域的工作。此外，在去年推出 AI 助手 Duplex 之后，谷歌即着手构建合成语音数据集，于 2019 年 1 月发布了虚假语音（fake speech）数据集并开放给社会，以促进虚假音频检测技术的研究与开发。

二是支持、发起深度合成检测挑战赛，与行业携手推动检测技术的研究与开发。2019 年 1 月，谷歌 AI 团队和谷歌新闻计划（GNI）合作创建了一个深度合成数据集，其中包括大量利用深度学习 TTS（text-to-speech）模型合成的语音。谷歌通过支持 2019 AVSspoof 挑战赛（即自动语音验证模仿和对策挑战赛），将这一数据集开放给挑战赛的参与者来开发鉴别真实语音和虚假语音的系统。脸书也在积极推进类似举措。2019 年 9 月，脸书宣布投入约 1000 万美元，联合微软、美国 AI 联盟（Partnership on AI）、MIT 等九家机构发起深度合成检测挑战赛（Deepfake Detection Challenge，DFDC），以促进这一领域更多的研究和发展，确保开发出更好的开源工具来检测识别深度合成内容。DFDC 包括一个数据集、排行榜，以及资助和奖励；Facebook 负责为挑战赛开发深度合成数据集，在 2019 年 10 月的国际计算机视觉大会（ICCV）上测试这些数据，之后在 2019 年 12 月的神经信息处理系统大会（NeurIPS）上发布整个数据集并正式启动 DFDC 挑战赛。

三是积极开发深度合成检测识别和标注工具。当前，深度合成内容的检测识别和标注技术，日益得到行业的重视。例如谷歌积极与新闻机构展开合作，采取训练检测系统等措施消除错误及虚假新闻；去年，谷歌的工程师与初创公司 AI Foundation 合作开发了名为“Reality Defender”的浏览器插件，该插件可扫描用户浏览的图像、视频或其他数字媒介，标记并报告可疑的伪造内容，检测经窜改的或人工合成的内容。AI Foundation 还与内容创作者进行合作，在其 Human-AI 协作平台上采用“AI 水印（Honest AI watermark）”标注 AI 生成的文本、图像、视频和音频。微软也在研发针对不同 AI 换脸算法的鉴别算法，以及通用的鉴别模型。此外，越来越多的创业公司也在加入到研发 deepfake 对抗技术的行列，例如 2019 年 9 月，初创公司 Quantum Integrity 与瑞士联邦理工学院（EPFL）合作开发 deepfake 检测工具并获得了瑞士创新机构 Innosuisse 的资助。在国内，腾讯信息安全团队已自研了 GFN 网络算法鉴别 AI 换脸，达到了很高的准确率；腾讯优图实验室也在构建人脸合成检测平台——“FaceIn 人脸防伪”，支持对多种换脸方法进行检测，达到了很高的准确率。

四是培训专门的合成内容审查人员，加强对视频内容的真实性审核。除了研发、部署检测识别技术

外，网络平台也开始考虑加强其内容审查队伍，培训专门的合成内容审查人员来鉴别深度内容。例如，随着网络上合成内容的增多，在过去的两年中，路透社负责验证视频内容真实性的审查人员翻了一番。如今，该团队在全球范围内每周需要验证约 80 个视频。审查人员在验证的过程中需要借助某些技术，例如在 Google 地图上交叉参照位置或反向图像搜索。在 2019 年 2 月印度和巴基斯坦的冲突事件中，路透社一共发现了大约 30 个伪造的视频。此外，对于 deepfake 内容，不同平台的内容审核政策存在差异，例如 Facebook 就曾公开拒绝移除一些被篡改的视频片段，如美国政客 Nancy Pelosi 和 Facebook 的 CEO 马克 · 扎克伯格的视频片段，认为并未违反平台的内容政策，而其他平台则可能对类似内容采取下架措施。

4.5 公众教育

立法和技术检测之外，进一步提高网络用户的信息分辨能力也是深度合成技术治理的必要手段。[44] 合成内容往往能够满足用户的好奇心和猎奇心理，并促使用户与他人分享这些内容。而平台算法倾向于突出显示流行的信息，尤其是被相对同质的群体分享的信息，内容的反对声或者澄清声明可能就被过滤掉了。更进一步，受到信息级联的影响，人们更倾向于相信别人传递的信息，而忽略自己的判断，这强化了社交媒体用户对他们先前所接触信息的信任，加速深度合成的虚假内容的传播速度。一项研究表明，恶作剧和虚假谣言传播的速度是真实故事传播速度的十倍。即使研究人员控制了谣言源头之间的差异，谎言被转发的可能性依然比准确的新闻高出 70%。一言以蔽之，深度合成的虚假内容利用大众的认知偏见，更容易捕获公众的注意力，并经由传播增强对这种虚假信息的信任度，让真实内容和合成内容的区分难度加大。

因此，应强化教育宣传，提高用户的数字素养（digital literacy），培养批判性思维，逐步提高用户的警惕防范意识和鉴别能力，以批判性的眼光看待音视频资料并进行多方检索印证，在接收信息和传播信息的过程中保持警醒，不要轻易相信眼见为“实”，也不要因为眼见之“虚”就过快否定其真实性。[45] 提高数字素养有助于加强用户发现假新闻的能力，并且在网上互动时更加尊重彼此，这对于年纪尚小的网络用户和年长、不太理解技术的群体同样重要，用户需要能够批判性地评估他们可能希望观看的视频的真实性和社会背景，以及其来源的可信度，以便理解视频的真实意图。[46] 更进一步而言，还需要通过教育宣传等措施提升社会各界对于深度合成潜在滥用和风险问题的意识，让决策者、科技行业、学术界和用户能够清醒地意识到深度合成技术的滥用风险及其危害，逐步提高社会大众的防范意识和鉴别能力；同时发挥新闻舆论的监督作用，营造正向氛围，促进深度合成技术的妥善利用，在防范风险的同时为对抗性技术和鉴别技术的发展提供机会。

44. https://www.google.com.hk/amp/s/sciencemediahub.eu/2019/12/04/deepfakes-shallowfakes-and-speech-synthesis-tackling-audiovisual-manipulation/amp/
45. 曹建峰、方龄曼：《“深度伪造”的风险及对策研究》，载《信息安全与通信保密》2020 年第 2 期。
46. Mika Westerlund, the Emergence of Deepfake Technology： A Review

结论：

我们正在走向一个人工智能更加大众化的世界，机器学习、神经网络和人工智能系统将变得更加强大，深度合成技术就是在这种背景下出现的。从技术实质来看，深度合成以深度学习等 AI 技术为基础，包括语音、图像、视频等多种合成类型，并朝着综合性的方向发展，最终可以实现数字虚拟人，带来前所未有的数字体验。就目前而言，深度合成技术在电影、娱乐、教育、社交、电子商务、医疗、艺术创造、科研等诸多领域的应用潜力已经显现，2020 年也有望成为这一技术走向大规模商业化应用的元年；国内外诸多商业化应用印证了这一技术的价值，它不仅能降低创作门槛，激发新形式的创造，还能以多种形式造福社会。但是就像多年前索尼 Betamax 录像机，现在的虚拟现实技术、自动驾驶、虚拟货币等一样，新技术总是具有两面性，技术的应用一方面可以消除一些风险，推动社会的进步；另一方面也会带来新的风险。但这不应成为限制或禁止深度合成技术发展应用的理由。相反，需要在理解这一技术及其影响的基础上，通过法律、技术、行业、用户的多重治理，最大程度地减少其滥用的可能性，同时最大程度地发挥其作为学习、创作、商用、科研合法工具的潜力。

首先，立法和监管应当包容审慎，避免矫枉过正，挫伤深度合成技术的发展应用，影响技术的社会经济价值的发挥。实际上，对于任何新技术，无论是人工智能还是自动驾驶，立法和监管都需要把握合理的限度，遵循包容审慎的理念，通过多方参与、风险评估、成本效益分析等机制，确保立法和监管的科学化、精细化、灵活化，并可考虑设立“安全港”规则或者监管例外来鼓励 AI 应用。[47] 就深度合成技术而言，技术是中立的，技术的价值观取决于使用技术的人。深度合成技术并非在所有情况下都会给个人、社会和国家带来重大损害和威胁，相反这项技术具有广泛的正向应用价值，能够以多种形式造福社会。因此笼统地禁止该项技术，禁止制作或传播深度合成内容并不科学，也无法真正打击不法分子，对于该技术需要细分应用情形分类管理，采取分类分场景监管。具体而言，一方面，应明确违法有害的深度合成内容的范围，应限于危害国家安全、散布淫秽色情等“九不准”内容；对于禁止范围之外的深度合成应用形式和内容不应施加额外的法律和监管限制，因为即使带来法律风险，也可以通过既有的民事、行政和刑事规则得到较好的解决，如著作权、个人信息保护、隐私、肖像、声音等方面的法律规定，未必需要新的监管措施。

其次，加强源头治理。从源头要求制作者对深度合成内容行进行披露和标记，向公众提示内容的合成性质，避免造成误导，并确保可以溯源。当前美欧的做法主要是要求制作者对深度合成内容进行标注，否则可能承担

47. https://new.qq.com/omn/20200302/20200302A0LSK900.html

民事责任，严重时还将承担刑事责任，从而从源头上对深度合成进行规范。这一源头治理的做法具有合理性，因为没有标注的深度合成内容一旦传播出去之后，第三方就很难鉴别，检测技术的开发、成熟也面临着诸多困难，而且难以跟上深度技术进化的步伐，所以加强源头治理是最有效的措施。

再次，鼓励行业开发鉴别技术和溯源技术。当前深度合成技术快速进发，相应的鉴别技术和溯源技术尚在初步阶段，需要大量的数据来发展鉴别技术。未来，平台可通过部署鉴别、溯源等技术来实现对深度合成内容的治理。但不宜强制要求平台对用户上传的或第三方的视频是否属于“深度合成”或“AI 合成”进行准确的检测识别，因为这将给企业带来不成比例的管理负担和成本。正因如此，美欧没有强硬要求平台部署鉴别技术。技术攻防本来就是“魔高一尺道高一丈”，处于不断发展之中，故不宜将其转变为硬性要求并与法律责任挂钩。

最后，行业应当加强自律做法。随着深度合成技术的广泛应用，相关负面事件和问题时有出现，呼吁行业和企业加强自我监管。例如，为深度合成技术的合法合规应用和健康有序发展制定行业公约、标准、最佳实践、伦理指南等，营造良好的行业发展氛围。

附录：

关于“深度合成”技术的十个误解

随着人工智能时代的到来，下一代媒体将由人工智能驱动，人工智能可能给数字内容领域带来重塑。其中，可以实现换脸、人脸合成、语音合成、视频生成甚至数字虚拟人等诸多应用形式的“深度合成”技术，作为人工智能发展到一定阶段的产物，逐步从 deepfake、deepnude 等色情性的换脸视频的阴影中走了出来，迎来了商业化时代。AI 虚拟主播、电商平台上的“数字试穿”、电影后期制作、社交产品中的人脸融合、合成人脸和合成虚拟形象用于在线营销、合成声音用于失声患者发声，以及数字虚拟人等创新性的应用持续涌现，“深度合成”技术的社会福祉日益彰显。但由于对技术的不了解，人们对“深度合成”技术还存在诸多偏见和误解，例如认为“深度合成”就是“深度伪造”，认为“深度合成”会彻底冲击社会信任，等等。为此，我们总结出了人们对该技术的十个误解，希望通过澄清这些误解，帮助人们更全面地了解深度合成技术的发展和应用情况。

误解 1：

深度合成技术仅包括 AI 换脸一种形式。

实际上，现阶段的深度合成技术，除了广为人知的“AI 换脸”以外，还包括人脸再现、人脸生成、语音合成等技术，并朝着全身合成、数字虚拟人等方向发展。AI 换脸是最早进入公众视野，也是目前应用较多的深度合成形式，可以借助人工智能技术对视频中的人脸进行替换，在一些 AI 换脸应用中，用户只需上传一张面部照片，就可实现化身电影中的演员、游戏中的角色等效果。除此之外，“人脸再现”涉及对目标人物的脸部表情进行驱动；“人脸合成”涉及创建媲美真实人脸的全新人脸图像；“语音合成”涉及创建特定的声音模型，可以将文字转化成接近真人语调和节奏的声音。同时，深度合成正从局部合成转向全身合成，从二维合成转向 3D 合成；前者例如对目标人物的全身动作进行操控，后者则以数字虚拟人技术为代表。目前，国内外互联网公司纷纷试水数字虚拟人技术，例如，2018 年腾讯携手 Epic 等企业启动“Siren”虚拟人项目，2019 年腾讯 AI Lab 正式发布首个电竞虚拟人“T.E.G”（天鹅静），整合 3D 人脸和人体重建、文本 / 语音 / 口型驱动和神经网络渲染等技术，特别是利用生成对抗网络完成人体动作的迁移。随着 5G 时代的到来，这种捕捉和渲染将会更加灵敏生动，数字虚拟人在游戏、社交、影视、医疗等领域将大有可为。

误解 2：

任何人都可以制作高质量、高仿真的深度合成内容。

深度合成内容的制作门槛已大为降低，但是高质量、高仿真的深度合成内容的制作还未普遍实现，仍需专业技能和专业工具。相比于 PS 等传统的图像处理软件，得益于源代码的开放和易用性工具的开发，深度合成技术的使用门槛已大为降低，普通用户在智能手机、电脑等终端设备上，借助深度合成应用程序，即可轻易制作、获取 AI 换脸、人脸合成、语音合成等娱乐性的深度合成内容。这类合成内容往往较为容易辨别，且存在来源标记，不至以假乱真。因此就目前而言，虽然像 FakeApp、ZAO 这样的软件已经开始让更多的人接触到深度合成技术，但高质量、高仿真的深度合成内容仍然难以创建，需要掌握专业技能和专业工具的专业人员的大量投入。

误解 3：

深度合成技术已被大量滥用，用于在社交媒体平台上制作、传播虚假信息。

实际上，无论是在国内还是在国外，社交媒体平台上涉及政治和政治人物的深度合成视频都是很少见的，深度合成性质的虚假信息也很少。此前在国内外引发广泛关注的奥巴马、普京等政治人物的深度合成视频，更多是警示性的和教育性的，意在表明深度合成技术可能出现此类滥用，而非为了传播政治谣言和虚假信息。而且主流社交平台已采取了针对深度合成内容的审核政策，因此深度合成内容并未在社交媒体平台中失控，也并未给公众话语权与社会舆论造成扭曲。但色情性的深度合成视频，是深度合成技术滥用的重灾区，应予以重视，报告显示，2019 年 12 月全网共有 14678 个深度合成视频，其中 96% 属于色情性的深度合成视频，主要存在于色情网站。

误解 4：

快速立法是应对深度合成技术滥用风险的唯一有效方式。

在新技术的治理与风险防范方面，法律规制一直是必不可少的手段，但由于很难识别深度合成内容的来源，立法可能起不到应有的效果，还可能阻碍技术的有益应用与正向发展。因此，立法和监管应当包容审慎，把握合理的限度，避免因矫枉过正而挫伤技术的发展应用从而影响技术的社会经济价值的发挥。更进一步而言，可通过多方参与、风险评估、成本效益分析等机制，确保立法和监管的科学化、精细化、灵活化，并可考虑设立“安全港”规则或者监管例外来鼓励 AI 应用。当然，立法并非唯一有效的方式，而且具有滞后性，难以跟上技术发展演变的步伐，尤其是对于仍在快速发展的深度合成技术而言；更为合理的路径是，借助鉴别技术、溯源技术等技术措施，要求制作者对深度合成内容进行标记的源头治理，行业公约、标准、最佳实践、伦理指南等行业自律措施，以及公众教育和数字素养的培养等更为敏捷灵活的治理措施，来实现多元治理。

误解 5：

深度合成内容无法通过技术工具鉴别，只能通过生物特征测试（例如“眨眼测试”）。

实际上，眨眼测试等根据生物特征进行鉴别的方式，是非常低效、不可靠的，只能阶段性地起作用，而且随着深度合成技术的发展进化，生物特征测试越来越难以发挥作用。相反，深度合成内容的检测识别，需要基于 AI 的鉴别技术，来实现对深度合成内容的自动化检测。目前，随着深度合成技术的进化，学界和业界已在大量投入和支持鉴别技术的开发，但目前的鉴别网络多针对特定的深度合成方法，尚没有通用的鉴别网络，因此 AI 检测工具需要随时更新。在国内，腾讯优图实验室也在构建人脸合成检测平台——“FaceIn 人脸防伪”，支持对多种换脸方法进行检测，达到了很高的准确率。

误解 6：

深度合成就是“深度伪造”（deepfake）。

国内媒体一般根据“deepfake”这一合成词，将其背后的技术翻译为“深度伪造”，但“深度伪造”是以偏概全，不足以涵盖所有的深度合成技术和相应的合成内容。追根溯源，deepfake 最初只用于描述 AI 换脸的色情视频，是一种特定的 AI 换脸技术，后来被媒体拿来泛指所有的深度合成技术，是以偏概全，既不专业，也不科学。因为“深度合成”的内涵更为广泛，意指借助人工智能算法实现语音、音乐、图像、人脸、视频等内容的合成和自动生成，而以“深度伪造”为代表的 AI 换脸只是其中的一种应用形式而已。此外，“深度伪造”这一不甚科学的术语容易给相应的 AI 技术造成污名化影响，可能扼杀技术的潜在社会福利，不利于技术发展应用，因为 deepfake 背后的 AI 技术具有很大的正向应用价值，如新华社的 AI 合成主播、网络上的虚拟歌手、社交媒体中的换脸应用等。因此，虽然 deepfake 的出现让背后的 AI 技术获得了广泛的关注，但基于技术使用的意图（即 deepfake）去定义技术，强调技术的潜在欺骗性或可能带来的负面影响，这一做法并不科学。基于以上考虑，“深度伪造”（deepfake）这一用语实际上并未得到技术社区的广泛认可；相反，使用“深度合成”（deep synthesis）来描述相关的 AI 技术和合成内容，更为科学合理。

误解 7:

深度合成是人工智能技术作恶，只会给社会带来负面影响没有正向价值。

具备高度仿真能力的深度合成技术，虽然也存在被滥用的风险，但其巨大的正向应用价值将持续带来社会福利，正被广泛应用于影视、娱乐、教育、医疗、社交、电商、内容营销、艺术创作、科研等诸多领域。随着过去几年的发展成熟，深度合成技术在 2020 年迎来商业化元年，大规模商用成为可能，未来几年将持续涌现创新性的应用形式。例如，在影视作品的后期制作方面，深度合成技术已被用于“数字复活”演员或演员的声音，或者实现多种语言的“数字配音”。亦开始大量涌现 AI 主播、虚拟歌手、AI 换脸、数字虚拟人等社交与内容类应用。在电商领域，深度合成技术可以将用户的脸部换到短的视频片段中，从而让用户在购买前可以实现“数字试穿”。在广告宣传、内容营销等领域，AI 合成的人脸和虚拟形象可以替代网红、模特等，既能带来新鲜感，也免去了传统上使用他人肖像的授权。在医疗领域，深度合成技术可以让有失声风险的患者重新获得“自己的声音”，也可以生成与真实影像无异的医学图像来训练 AI 系统，解决数据不足、病人隐私保护等问题。在语音合成方面，腾讯云上线的语音合成以及实时语音合成技术，可以将任意文本转化为语音，用于新闻、车载导航等个性化语音播报、有声读物制作、机器人发声等。总之，深度合成并非关于“伪造”和“欺骗”的技术，而是极富创造力和突破性的技术，虽然它像其他任何技术一样，也催生了一系列必须面对的难题，但这并不会磨灭这一技术给社会带来的进步。

误解 8:

深度合成就是“深度伪造”(deepfake)。

实际上，互联网行业内的主流网络平台已经着手采取自律措施应对深度合成技术的潜在滥用。谷歌、Facebook 等美国主流科技公司已经采取了应对方案，积极开发甄别 AI 合成内容、对抗深度合成技术滥用的方法和工具，如谷歌开发的“Reality Defender”工具，可扫描用户浏览的图像、视频或其他数字媒介，标记并报告可疑的伪造内容，检测经窜改的人工合成内容；在此基础上降低合成内容的权重，让算法不再为用户推荐被认定为深度合成并可能造成负面影响的内容。利用平台优势，这些科技公司已经在积极构建深度合成数据集，并开放给研究人员免费使用，以此来促进检测技术的研究与开发。同时，各平台之间还携手开展深度合成检测挑战赛，为检测技术的开发提供资金和深度合成数据集，以促进更多检测识别技术的开发。仅 2019 年，谷歌、Facebook 等相继投资此类竞赛，例如 Facebook 联合微软、美国 AI 联盟（Partnership on AI）、MIT 等九家机构发起的深度合成检测挑战赛（Deepfake Detection Challenge），已取得一定效果。在

技术赛道之外，平台也在培训专门的合成内容审查人员，主要目的是增加审核的准确性，特别是在深度合成与戏仿讽刺的界限还难以把握的情况下，需要人工审核的参与，确保内容符合平台的政策要求。在国内，腾讯信息安全团队自研的 GFN 网络算法鉴别 AI 换脸，及腾讯优图实验室研发的人脸合成检测技术，对相关深度合成内容的检测都达到了很高的准确率。

误解 9：

深度合成已经被国外立法禁止。

实际上，被禁止的不是深度合成技术本身，而是利用此项技术从事色情视频合成、虚假新闻、干扰选举等非法行为。Reddit 网站上 deepfake 论坛关闭、一键裸照应用 deepnude 下架等事件似乎表明国外对这项技术很不友好，但事实上，国外立法都承认深度合成技术的有益应用和正向价值，没有“一刀切”禁止使用深度合成技术，而是根据使用意图和使用效果进行划分，主要对利用深度合成技术从事的违法行为进行打击，而没有对正常的深度合成技术应用施加过多的限制。例如，美国国会“Deepfakes 责任法案”及美国德州、加州、弗吉尼亚州、纽约州的相关法案等只禁止政治干扰、色情报复、假冒身份等目的的深度合成，但没有强制要求平台部署检测识别措施，而是加强源头治理，要求制作者、上传者对深度合成内容添加水印、文字、语音等标记。欧盟则对深度合成技术可能引致的假新闻以及个人信息保护等问题关注度颇高，在考虑用 GDPR 进行规制的合理性。回到我国，《网络信息内容生态治理规定》第 23 条、《网络音视频信息服务管理规定》第 10-13 条、《民法典人格权编（草案）》第 799 条、《数据安全管理（征求意见稿）》第 24 条等规定给“深度合成”技术划定了应用边界，同时为行业探索有益应用场景留出了发展空间。

误解 10：

深度合成会彻底冲击媒体信任。

深度合成技术将如何影响大众的行为和认知，目前还没有足够的研究支持，但是它提示我们，进入人工智能大众化时期，对大众信息分辨能力的培养也是治理的重要一环。以往 PS 等编辑技术也能进行一定程度的内容合成，但是并未冲击社会的信任，相反社会能很好地适应并使用这一技术。深度合成媒体将比 PS 等技术更容易操作和使用，随着开源工具的出现，深度合成内容的应用规模和使用范围也将更大，内容的说服力更强。这为识别真实信息与合成内容带来了挑战，在一些报道中，深度合成技术被形容成社会真相的破坏者，认为深度合成技术的存在会导致对媒体信息天然的不信任，公众可以用“deepfake”去怀疑一切他们想怀疑的事物。问题是，在这一技术出现之前，使用传统的音视频剪辑技术，甚至不使用技术手段，通过断章取义等简单方式就可以炮制虚假信息。媒体信任的塑造绝对不仅仅是封杀某一技术可以达到的，而需要从内容的生产、传播、接收等多方面进行规范。深度合成技术的出现已经让我们意识到了眼见不一定为“实”，这是加强公众信息辨别能力的一个重要契机。

腾讯研究院
Tencent
Research Institute

腾讯研究院是腾讯公司设立的社会科学研究机构，旨在依托腾讯公司多元的产品、丰富的案例和海量的数据，围绕产业发展的焦点问题，通过开放合作的研究平台，汇集各界智慧，共同推动互联网产业健康、有序的发展。围绕互联网法律、公共政策、互联网经济、大数据等研究方向，与国内外研究机构、智库开展多元化的合作，不断推出面向互联网产业的数据和报告，为学术研究、产业发展和政策制定提供有力的研究支持。我们坚守开放、包容、前瞻的研究视野，致力于成为现代科技与社会人文交叉汇聚的研究平台。

欢迎扫描二维码关注
腾讯研究院官方公号

腾讯优图

腾讯优图实验室成立于 2012 年，是腾讯公司旗下顶级人工智能实验室之一。优图聚焦计算机视觉，专注人脸识别、图像识别、OCR、机器学习、数据挖掘等领域开展技术研发和行业落地，在推动产业数字化升级过程中，优图始终专注基础研究、产业落地两条腿走路的发展战略，与腾讯云与智慧产业深度融合，挖掘客户痛点，切实为行业降本增效。与此同时，优图关注科技的社会价值，践行科技向善理念，致力于通过视觉 AI 技术解决社会问题，帮助弱势群体。

顾问团队

司晓

腾讯研究院院长

张钦坤

腾讯研究院秘书长

黄飞跃

腾讯优图实验室总监

研究团队

曹建峰

腾讯研究院高级研究员

丁守鸿

腾讯优图实验室高级研究员

熊辰

腾讯研究院助理研究员

研究联系

曹建峰

邮箱：jeffcao@tencent.com 微信：xinzhelibrary

腾讯研究院 腾讯优图

AI生成内容发展报告2020——“深度合成”（deep synthesis）商业化元年