道高一尺魔高一丈?

科研打假新对手：以假乱真的人工智能 (AI) 生成图像

裴磊

华中科技大学同济医学院

3 月中旬，一段乌克兰总统沃洛季米尔 • 泽连斯基的一分钟的视频首先出现在社交媒体上，随后出现在一个乌克兰新闻网站上。在视频中，泽连斯基告诉乌克兰士兵放下武器，向俄罗斯军队投降。但这段视频之后被证明是深度伪造（deepfake）的，是一个通过人工智能（AI）机器学习技术合成的一段假视频。

一些科学家现在担心，类似的技术可能被人用来制造虚假的光谱或生物标本图像进行科研欺诈行为。

致力于学术打假的微生物学家和科学诚信专家 Elisabeth Bik 说：“我一直非常担心这类技术。我认为深度伪造的研究图像用于学术论文并被发表这种情况正在发生”。她怀疑她此前曾揭露的 600 多份貌似来自同一家论文工厂的完全捏造的研究图像（图 1-4）可能是人工智能生成的。

与人工操作的图像不同，AI 生成的图像几乎不可能被肉眼分辨出来。在一项由中国厦门大学计算机科学家 Rongshan Yu 领导的团队所作的研究中，研究人员利用 AI 技术创造了一系列深度伪造的 Western blot 和肿瘤图像（图 5）。这些伪造的图像与真实图像根本无法用肉眼进行区分。

深度造假通常基于生成式对抗网络（GAN），其中一个生成器和一个鉴别器试图相互竞争。假设一个网络试图从白噪声中生成一个假图像，比方说一张脸，它最初是不知道如何生成人脸的，所以它需要鉴别器的帮助，鉴别器是另一个网络，学习如何区分图像的真假。最终，生成器将欺骗鉴别器，使其认为图像是真实的。

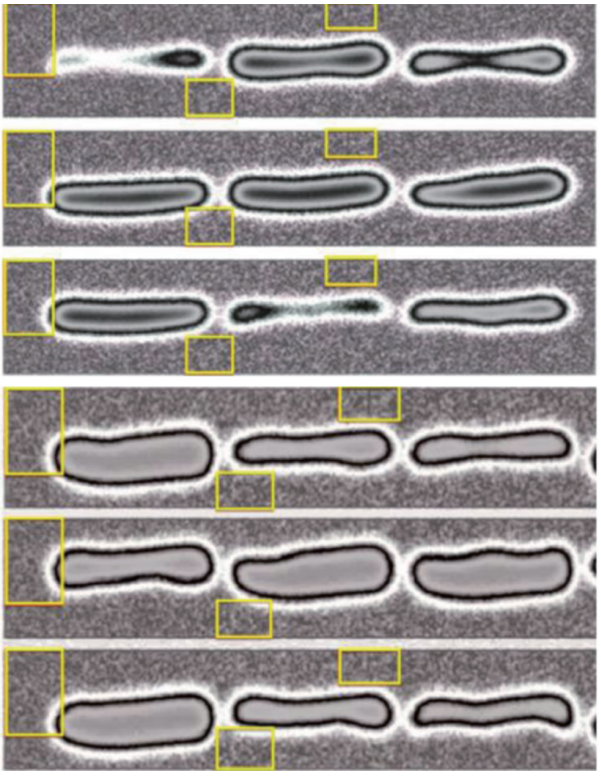

图1 注意带状形状和背景图案的相似性。上图和下图来自不同的手稿，有不同的作者和所属单位。Figure 2 from Christopher, FEBS Letters 592 (2018) 3027–3029, DOI: 10.1002/1873-3468.13201.

DOI: 10.14218/MRP.2022.507
通讯作者：裴磊 Email: 154948836@qq.com

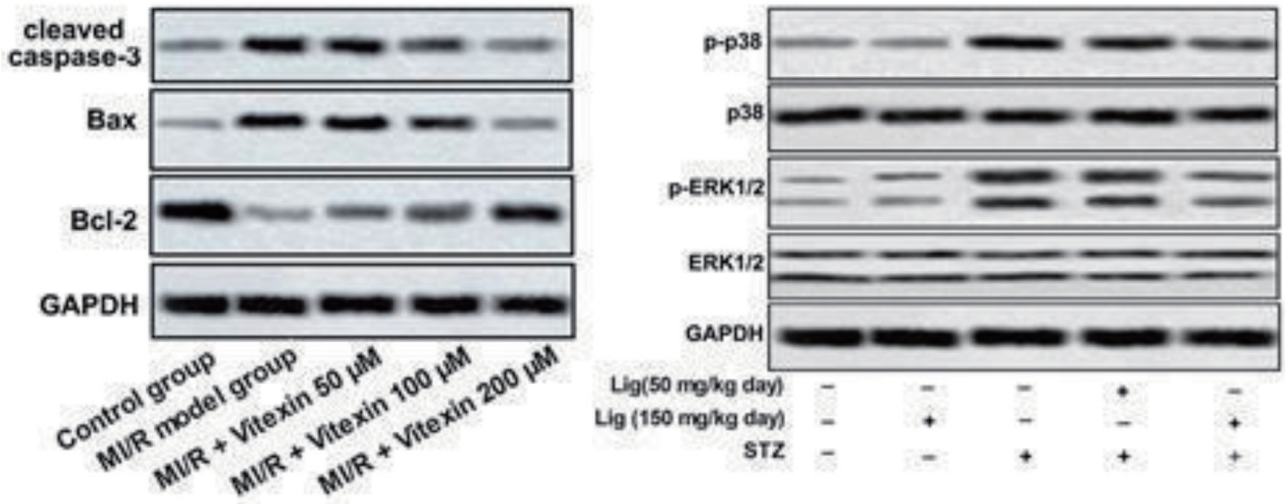

图2 这里的两个Western blots来自不同作者的不相关的论文，但底部的GADPH却是完全相同的。
来源：Source: Retracted 2017 & 2018 Royal Society of Chemistry papers.

美国加州大学伯克利分校专门研究数字取证和错误信息的哈尼 - 法里德（Hany Farid）说，鉴于 GAN 可以合成与真实面孔无法区分的假面孔，它完全可以生成以假乱真的生物图像。但是，尽管深度造假是一个需要认真对待的威胁，“我更关心的是‘可重复性、P 值黑客、Photoshop 操纵’这些传统作假方式。它们仍将在相当长的一段时间内占据主导地位”。

罗切斯特全球网络安全研究所所长马修 - 赖特 (Matthew Wright) 表示同意上述观点。他认为 AI 造假目前还不具有特别的威胁性，尽管这在技术上很

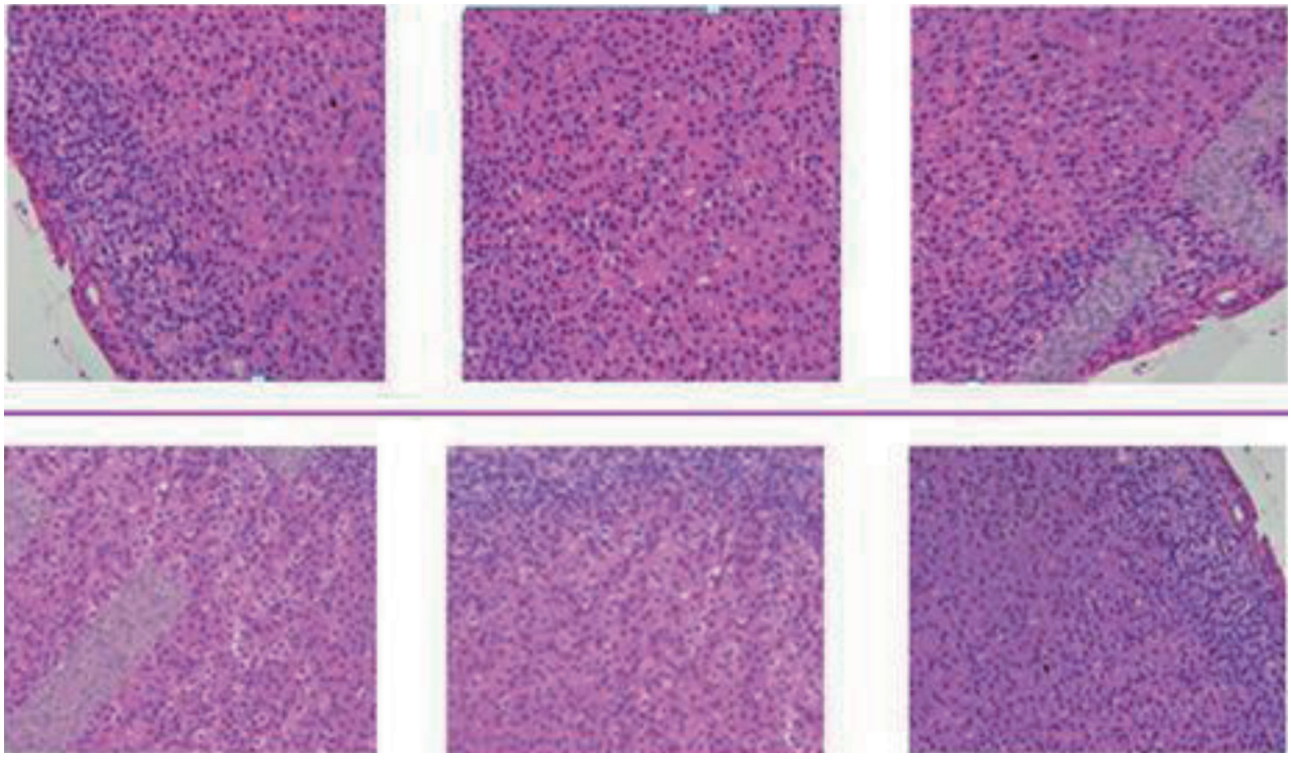

图3 两篇不相关的论文的肾脏组织切片（上/下），没有共同的作者。其中至少三张切片的染色是重复的，已经被重新定位和旋转了。
来源: Retracted 2018 Royal Society of Chemistry papers.

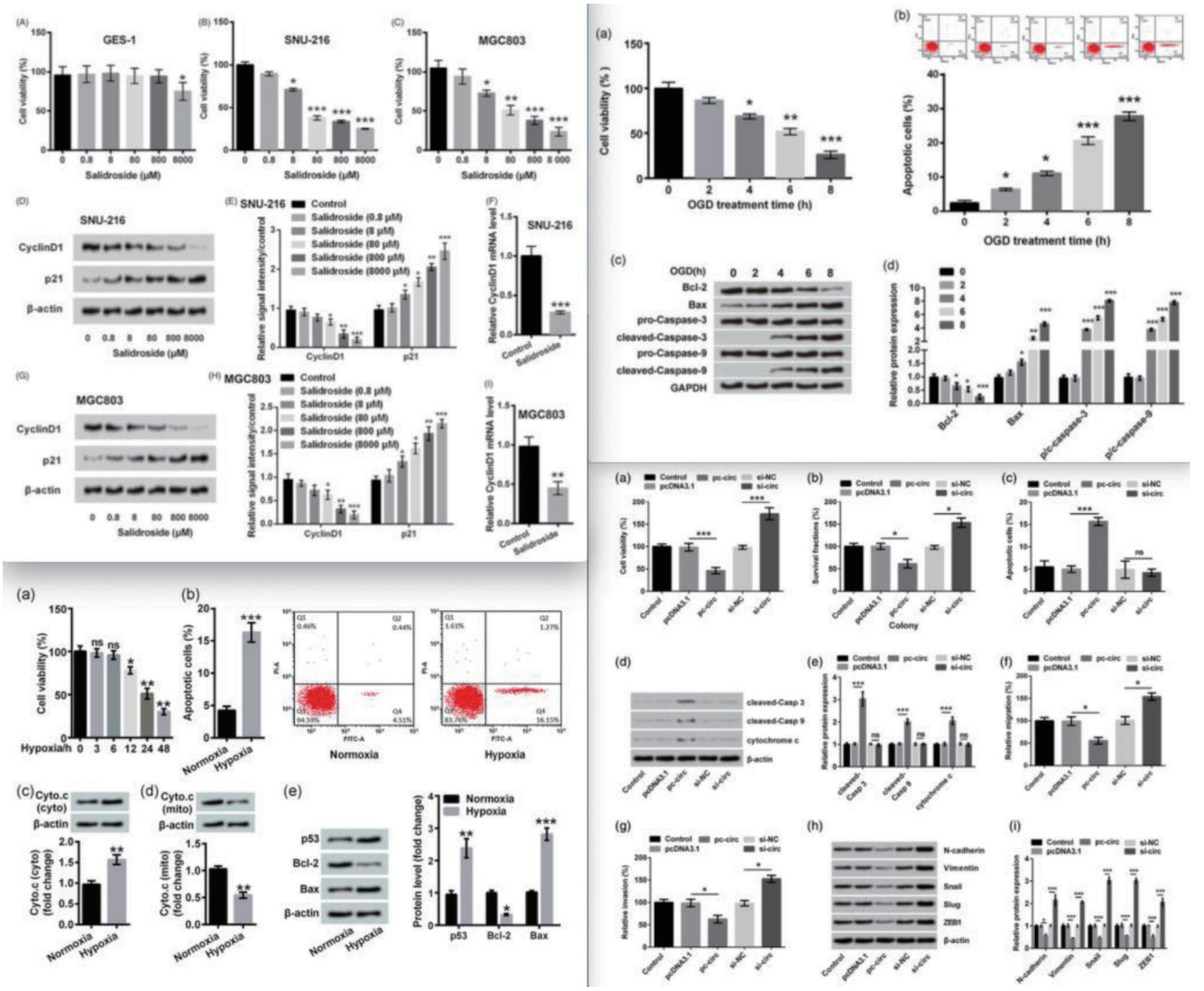

图4 来自不同研究小组的四篇不同论文的结果，它们都有类似的布局。
注意柱形图布局、字体和Western blots之间的相似性 (图片来自The Tadpole Paper Mill – Science Integrity Digest).

难被发现。机器学习留下的数据库可以用来识别假图像，尽管欺诈者通常在几个月后就能找到绕过这种方法的办法。不过他相信，科学的自我纠正机制最终会处理掉虚假研究。

俞容山教授说，目前还不清楚文献中是否已经包含 AI 生成的图像。Bik 认为我们已经到了无法分辨论文真假的地步。我们需要更努力地与政府和研究机构合作，促使它们负起责任来，为研究人员减压，否则他们整个职业生涯都可能取决于在国际期刊上发表论文。

参考文献

[1] WangL, ZhouL, YangW, YuR. Deepfakes: A new threat to image fabrication in scientific publications? Patterns (N Y) 2022;3(5):100509.

[2] AI-generated images could make it almost impossible to detect fake papers | News | Chemistry World. Available from: https://www.chemistryworld.com/news/ai-generated-images-could-make-it-almost-impossible-to-detect-fake-papers/4015708.article.

[3] Deepfakes: A new threat to image fabrication in scientific publications? (www.cell.com). Available from: https://www.cell.com/action/showPdf?pii=S2666-3899%2822%2900101-5.

[4] The Tadpole Paper Mill – Science Integrity Digest. Available from: https://scienceintegritydigest.com/2020/02/21/the-tadpole-paper-mill/.

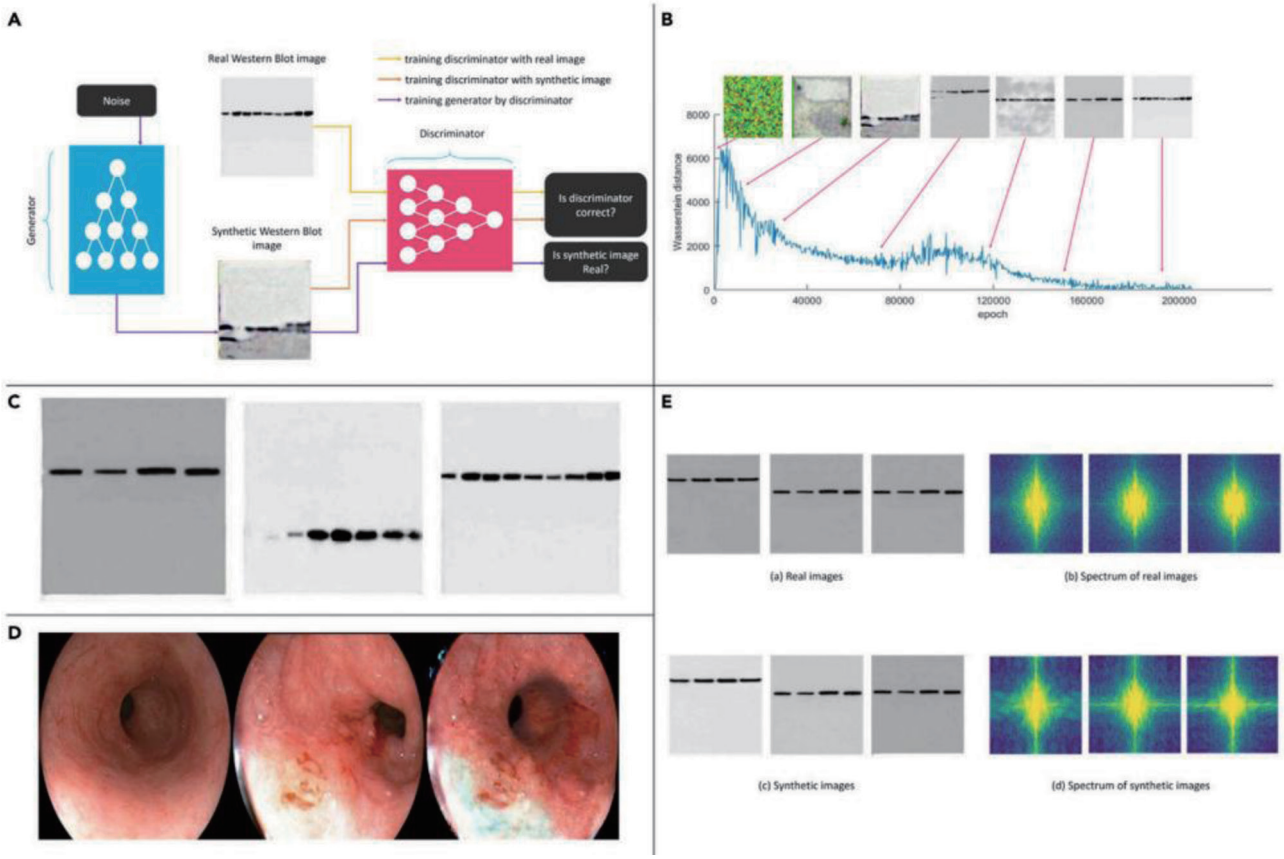

图5 AI深度伪造制图的工作流程和实例。

(A) generative adversarial network, GAN“生成对抗网络”流水线；(B) 当训练历时增加时，Wasserstein距离减少，在不同的训练历时产生的图像；(C) 生成的Western blot图像的例子; (D) 生成的食道癌图像的例子;(E) GAN的合成图像比真实图像有更多的高频部分。这些利用AI技术深度伪造的图像，足以以假乱真 (来源Source: © 2022 Liansheng Wang et al， https://doi.org/10.1016/j.patter.2022.100509).