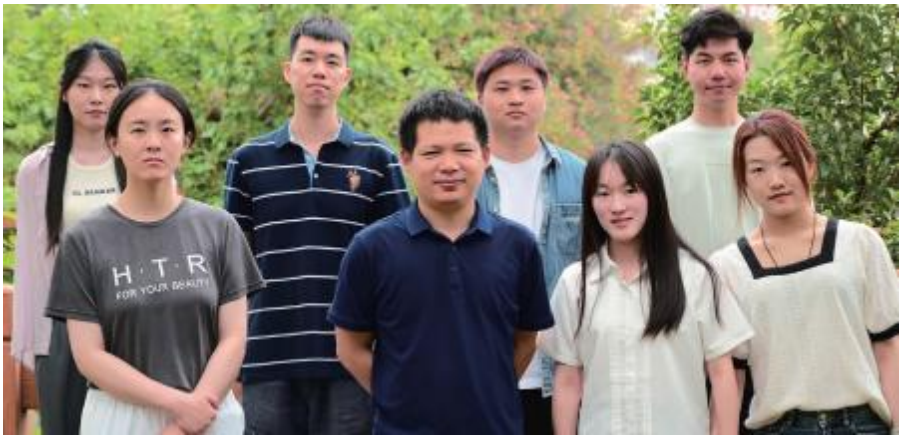
为数据世界“签名”

——记上海大学通信与信息工程学院副教授吴汉舟

郑 心　张锦玉

在数字洪流奔涌的当下，一张照片、一段音频、一份电子合同，都可能在转瞬之间被复制、篡改、传播，而原作者却无从追溯。而为这些无形的信息赋予“身份”的秘密，正潜藏在一种名为“数字水印防伪技术”的精密机制中——它不是浮于表面的标签，而是将一段加密信息如基因般嵌入多媒体数据的纹理、频率或像素结构之中，并且这种嵌入必须满足双重标准：对人类感官“透明”，对攻击手段“坚韧”——既不破坏原始内容的观感，又要在图像被裁剪、压缩、加噪后，水印仍可被可靠提取。简单而言，就像在一幅油画的笔触深处藏入作者的签名，肉眼不可见，紫外线一照就格外清晰。



吴汉舟（前排中）与学生合影

要实现如此“黑科技”，科学原理要横跨信号处理、信息论与密码学等不同领域的交叉融合：通过在离散余弦变换（DCT）或小波域中调整特定系数，嵌入鲁棒性强、抗压缩与噪声干扰的水印信息，构筑抗攻击模型。所以，在上海大学通信与信息工程学院数据安全处理实验室，聚集着一批心怀理想的学者，他们致力于调节水印的“不可感知性”与“鲁棒性”之间的微妙平衡，在由0和1构建的数字迷宫中，找寻从“能嵌入”向“难去除、易验证”进化的算法出口，吴汉舟正是其中一员。在他看来，网络空间没有“安全孤岛”，维护网络安全也不应该有旁观者，研究者尤其应该“冲锋在前”，“近年来，随着生成式AI伪造内容（Deepfake）的泛滥，数字水印更被进一步赋予了‘内容真实性验证’的新使命，这也同时意味着，我们应该用更多的努力面对愈加纷繁复杂的算法世界，做好网络净土的卫士”，他如此说道。

“山凿一尺宽一尺”

“2007年，我考入西南交通大学信息科学与技术学院；2011年本科毕业后保送本校直接攻读博士学位，后获国家留学基金委资助，2014年前往美国新泽西理工学院联合培养两年；2016年回国，2017年博士毕业……”吴汉舟用简单几个时间点就描述完自己数年的求学生涯，但任谁的往昔岁月中都有许多让人印象深刻的记忆坐标。对吴汉舟而言，那些俯首书案的日夜和标注密密麻麻的笔记本便藏着“岁月神偷”，翻开任何一页，都镌刻着其攀峰之路的生动注脚。

当走入“计算机科学与技术茅以升班”的那一天起，吴汉舟就预料到追寻梦想的远途，既会看到和风细雨，也会遭遇疾风骤雨。身边的同窗和师长已经足够优秀，但仍在用极高的标准要求自己。“那时候上课，老师对我们的要求高，很多课程都是双语教学，再加上学校还鼓励我们参加学科竞赛，所以每一天都是忙碌而充实的。”吴汉舟回忆。逐渐地，他的本科生涯便在忙碌的常态与ACM-ICPC国际大学生程序设计亚洲区预选赛和全国邀请赛银奖等荣誉的笼罩下逐渐走到了终点。可以说，在进入博士的研习之前，他就已经在计算机科学领域中打下了坚实的地基。这也为他之后顺利进入数字水印研究领域埋下了伏笔。

2014年，经导师推荐，吴汉舟有机会前往美国新泽西理工学院数字水印领域国际知名学者施云庆教授带领的团队开展为期两年的学习。“到了国外，王老师（导师）很关心我的生活和学习状态，给予了很多帮助。也就是在这两年，在导师和施老师的共同指导下，我完成了博士阶段的主要研究工作，可以说这一段经历塑造了我今天的科研风格和理念。”多年之后，他如此评价这段时光。

2014年以前，学者们主要利用传统的机器学习方法检测数字图像中隐藏的机密信息，2014年有文章开始阐述利用深度学习技术检测隐藏信息，但效果还较差，这与深度学习在其他领域取得的成功形成巨大反差，而这也引起了施云庆教授的关注和重视，带领正在攻读博士学位的徐冠朔及访学的吴汉舟一起开展研究。“我们与导师在2014年进入这一问题的研究，虽然经历了大量失败实验，但最终还是很幸运，在2015年底设计了一种有效的神经网络结构，取得了超越传统方法的检测性能。”次年，《IEEE信号处理快报》（IEEE Signal Processing Letters）期刊记录了这一成果，而文章一经发表也迅速得到了国内外同行的关注和正面评价。其中就包括吴汉舟钦佩的杰西卡·弗雷德里奇（Jessica Fridrich）教授。

“弗雷德里奇教授是美国纽约州立大学汉姆顿分校杰出教授，她在我们领域内取得了很多有影响力的研究成果，是国内外公认的信息隐藏领域权威学者，其论文引用量在我们领域一直排在第一位。难能可贵的是，弗雷德里奇教授非常专注，仅依靠数名研究生便能做出有重要影响力的成果，我想这源于她多年沉淀的真知灼见，源源不竭的自驱力，以及发于心底的热爱。”所以，当研究成果得到弗雷德里奇教授“the first architecture with a competitive performance（第一个具有竞争力性能的架构）”的评价之后，吴汉舟感受到了前所未有的动力。“我想我应该感谢这种认可，这或许是我坚持科研的源泉之一”。

回国之后，吴汉舟言出必践，抱着从前辈身上汲取的精神与哲理进入了中国科学院自动化研究所（以下简称“自动化所”）工作。“这是我的第一份工作，求职的时候没有想太多，有对首都的向往，也同时了解到自动化所是国内人工智能领域首屈一指的研究机构，拥有很多模式识别、计算机视觉及自然语言处理等领域的知名学者，我想我又有机会学习榜样、进益自己了。”而现实也正如他所想，在自动所度过的几年岁月让他深刻认识到“山外有山，人外有人”，于是吴汉舟的心愿也慢慢从“做科研”变成了“做好科研”。“希望在本领域做出更多同行认可的学术贡献吧。”他说。

“路修一丈长一丈”

2019年，由于家庭原因，吴汉舟最终决定离开北方，重投南国怀抱。而彼时，上海大学恰好以其悠久的数字水印研究历史走入了他的视野。众多杰出的研究成果让他即便面对着一切几乎从零开始的短暂困境，也充满信心与干劲地加入了人工智能安全团队，从项目申报、研究生招生到教学等工作一项项抓起。

光阴如梭，转眼吴汉舟已在岗位坚守五年有余。如果将其职业生涯发展比喻为一首庄严肃穆的交响乐，那么多媒体安全、人工智能安全就始终是其中的主旋律。2020年，吴汉舟与合作者共同提出了“无盒”模型水印的新概念。这项成果发表在重要学术期刊《IEEE电路与系统汇刊》（IEEE Transactions on Circuits and Systems for Video Technology），曾连续10个月入选此期刊高关注度文章榜单（Popular Article List），在最新的“谷歌学术评价指标”中，入选该期刊近五年h5指数Top 30名单，得到了国内外学者的关注和引述。文章中详细论述了仅根据人工智能模型的输出结果就可以完成水印提取的过程与原理，实现了人工智能模型的知识产权保护和追踪溯源。可以说，围绕深度伪造的检测、溯源与防御目标，此研究探索出了一条切实可行的深度伪造主动识别与对抗技术路线，为维护社会公信力与促进数字经济安全发展提供了有力的技术支撑。

“我认为，未来的研究重点之一会是人工智能创作内容（Artificial Intelligence Generated Contents，AIGC）的鉴别与溯源。”方向既定，吴汉舟对于未来的思考便从未停止，“社交网络已经成为了当下人们获取信息的主要渠道，而信息源的真实性是受众最大的诉求之一。数字水印技术能够发挥重要作用，国家也出台了相关管理办法。例如，国家网信办等七部门近期联合发布了《生成式人工智能服务管理暂行办法》，明确要求对人工智能生成的内容进行主动标识，这其实就是说要利用数字水印技术对人工智能进行主动监管。”捕捉大势，紧联当下，在科技的发展道路上，前进的每一步都值得称赞；在时代的星河里，每位研究者的汗水都熠熠生辉。虽然吴汉舟清楚，这样不遗余力向前奔跑的日子还有很多，攀登科技之峰的道路也不会一帆风顺，但他已经笃定：“我们经常讲的‘好好学习，天天向上’几个简单而朴素的字眼其实正是破题之法。何以拓展认知的疆域、铺平发展的道路？唯有持续学习而已。”科

（本文获国家自然科学基金项目资助，项目编号：U23B2023）