基于飞桨的《数字取证》课程建设与实践教学

方艳梅[1,2] 骆伟祺[1,2] 赵慧民[3]

1.中山大学计算机学院，广州，510006
2.广东省信息安全技术重点实验室，广州
3.广东技术师范大学计算机科学学院，广州，510665

摘 要 随着深度伪造（Deepfake）等相关技术的发展迅猛，对数字媒体取证提出了更高的挑战，同时也对信息安全人才培养也提出了更高要求。“数字取证”（Digital Forensics）作为网络空间安全方向的专业选修课程，在教学模式和实践教学环节需与时俱进。本文对 “数字取证” 课程的培养方案和实践教学模式进行了探索与思考，开展了基于百度AI Studio平台和飞桨（PaddlePaddle）深度学习框架的 “数字取证“在线课程，对课程建设方案以及实践教学模式进行了总结，并分析了目前仍存在的问题及改进方案。

关键字 数字取证，飞桨，课程建设

Curriculum Construction and Practical Teaching of Digital Forensics on PaddlePaddle

Yanmei Fang Weiqi Luo

School of Computer Science,
Sun Yat-sen University, Guangzhou 510006, China
fangym@mail.sysu.edu.cn;
luoweiqi@mail.sysu.edu.cn

Huimin Zhao

School of Computer Science,
Guangdong Polytechnic Normal University,
Guangzhou, 510665, China
zhaohuimin@gpnu.edu.cn

Abstract—With the rushly development of Deepfake related technology, there is a more challenge for the digital media forensics and higher requirement for the talent cultivation on information security. As a professional elective course, Digital Forensics needs to be catch up with both in the model of course teaching and practice training. This paper explores the course program and practical teaching model, and offers an online course of Digital Forensics on the platform of AI Studio, and then, presents the summarizes and prospect.

Key words—Digital Forensics, PaddlePaddle, Curriculum Construction

1 引 言

近些年，以深度伪造（Deepfake）技术[1]为代表的数字媒体内容生成和迁移技术得到快速发展，如今各类伪造的图像和视频在网络上泛滥成灾，给人们生活、司法取证、社会稳定等带来了极大的负面影响。为了尽量减少深度伪造技术带来的负面影响，亟须提高对真假图像/视频的检测和鉴别技术，这对数字媒体取证提出了更高挑战，同时也对信息安全人才培养也提出了更高要求。

由于深度伪造技术更新速度快，“数字取证“（Digital Forensics）作为网络空间安全方向的专业选修课程，在教学模式和实践教学环节需与时俱进。本文在此背景下，经过中山大学广东省信息安全重点实验室团队近几年的教学实践积累经验，对 “数字取证” 课程的培养方案和实践教学模式进行了探索与思考，并开展基于百度 AI Studio 平台和飞桨（PaddlePaddle）深度学习框架的 “数字取证“在线课程。

2 “数字取证”课程建设的思考

数字取证相关术语有计算机取证、电子取证、网络取证等，这些术语之间的界定不是很清晰，各有侧重，但取证的结果都是电子证据。电子证据既具有脆弱易失性、不可靠性等特点，其表现形式又具有灵活多样。若将其作为法庭证据，必须采用科学的方法和严格的程序保证电子证据具有客观性、关联性、合法性以及证据连续性。

不同于先前的计算机取证或网络取证，本文谈到“数字取证” 一般指 “数字媒体取证”(Digital Media Forensics)，并聚焦到 “数字图像取证”(Digital Image Forensics)[2]。

在课程内容方面，“数字取证”综合了人工智能、机器学习与深度学习、模式识别、多媒体信号处理、统计分析方法、最优化方法以及司法鉴定技术等多个领域知识，是一门多学科交叉融合的新型学科。在教学过程中，应该重视对相关领域的相关基础理论、基本概念、典型模型及算法等进行分析与讲解，以及对当前人工智能、机器学习、深度学习新技术的实战训练及应用。

在教学模式方面，“数字取证”属于专业技术应用课程，课程建设过程应强调理论与实践并重，学生在掌握课程基础理论的同时，应具备较强的动手实践技能。与传统的物证技术相比，数字取证无论是在法律依据上，还是在技术上都需要解决大量的问题，因此需要在实验教学中采取从数据集准备、算法模型搭建、模型优化、推理测试、取证报告等一系列适合本课程特点的方法。

3 课程教学内容建设

数字图像取证主要研究对象是数字图像与视频信息，而不是计算机文件、磁盘格式等。本课程重点以数字图像内容安全取证为核心，主要是通过对图像统计特性的分析来判断数字图像内容的真实性、完整性和原始性，即判断数字图像在数码相机拍摄之后，是否经过某些篡改操作。它属于计算机取证的一个分支，是对源于数字图像资源的数字证据进行确定、收集、识别、分析及出示法庭的过程。

数字图像取证的基本教学内容涵盖以下几个方面：

（1） **图像篡改历史**. 讲述从早期的图像篡改至今150年来发生的典型篡改案例；概述从90年代至今出现的图像篡改技术[3,4]。

（2） **数字取证基本概念**. 讲述数字图像取证的基本概念、数字图像取证方法和技术。主动取证的定义、分类及技术。盲取证技术的理论框架，现有的盲取证技术，分析目前盲取证技术存在的主要问题[6,7]，并介绍国际知名研究团队及主要研究成果。

（3） **数字图像篡改技术**. 从数字图像的真实性篡改、完整性篡改、原始性篡改、版权篡改等几个方面，分别讲述近年来流行的数字图像篡改技术。

（4） **相机模式噪声检测方法**. 针对数码相机内部产生的传感器噪声（模式噪声）进行图像源识别及篡改检测。目前，数码相机使用的成像传感器有CCD、CMOS等都是采用半导体工艺制造的，因此它们产生的噪声有相似的特性。本课程介绍一种典型的模式噪声取证方法PRNU (Photo-Response Non-Uniformity)[7]。

（5） **复制-粘贴篡改取证**. 讲述同幅图像复制-粘贴取证(Copy-Move Forgery Detection)、不同幅图复制-粘贴拼接取证(Splicing Detection)、以及基于JPEG数字照片块效应的复制-粘贴取证方法[8,9]。

（6） **双重JPEG压缩及重采样操作取证**. JPEG重压缩是对JPEG图像进行篡改的必要过程。重JPEG压缩检测可以作为一种辅助手段与其他检测手段一起检测图像是否经过了篡改[10]。

（7） **模糊润饰取证**. 模糊作为一种图像篡改辅助手段，其目的是消除或减少由于局部篡改而留下的痕迹，可利用同态滤波等方法对图像拼接边缘的模糊润饰取证[11]

（8） **数字图像源辨识**. 又称图像设备取证. 分析数码相机、扫描仪、手机等图像获取设备的统计特征及区分方法[12-15]；探讨基于色彩滤波阵列插值的数码相机类型取证[16]，以及基于支持向量机等小样本分类器进行训练分类的图像来源取证。

（9） **自然图像统计规律取证**. 分析图像的固有统计相关性以及篡改对相关性的影响；图像样本数据库建库准则、拓扑结构、建库规模，以及分类器选择等基于机器学习的辨识过程[17]。

（10）**图像隐写分析取证**. 图像隐写(Steganography)可以看作是一种特殊的篡改形式：它将秘密信息以一种无感知的方式嵌入到数字图像中，并通过公开信道进行传输含秘图像，以掩盖秘密通信的行为。图像隐写分析 (Steganalysis) 则通过发掘图像在隐写前后统计特征差异，以检测含有秘密信息的图像。经典的隐写分析方法有：早期针对最低有效位替代的检测算法[18]、基于图像残差统计特性的方法[19]、以及目前基于深度学习的检测算法[20]。

（11）**深度伪造检测**（Deepfake Detection）. 目前学术界主要从深度学习检测算法、数字来源取证、生命日志记录三个主要方向探索应对深度伪造技术威胁。主要方法有基于生理信号特征的方法、基于图像篡改痕迹的方法和基于GAN(Generative Adversarial Network）[21,22]图像特征的方法。本课程内容主要集中在后两者，即基于篡改痕迹的检测方法，以及基于生成对抗网络GAN图像特征的方法。

在教授学生完成数字取证课程基础内容的学习，并完成基础算法的实战训练之后，针对几种典型的图像篡改技术，从以下几个方面专题进行算法实现及探讨分析：

- 专题一 PRNU相机模式噪声检测技术；
- 专题二 基于直方图统计特性的双重JPEG压缩检测技术；

- 专题三 基于傅立叶变换域的图像篡改检测技术；
- 专题四 数字图像源辨识技术，包括 CG 图像/扫描图像/翻拍图像的源辨识；
- 专题五 基于 CNN 卷积神经网络的深度伪造检测技术；
- 专题六 基于深度学习和 GAN 图像特征的深度伪造检测技术。

4 基于飞桨的实践教学模式

“数字取证”课程属于信息安全或网络空间安全专业应用型选修课，需侧重培养学生解决实际取证问题的专业知识和基本技能。建议一学期安排六次实验，以大作业的形式，每次评分考察学生的学习效果，并进行反馈和改进。在教学方法方面，突出以下几个方面特色：

✧ 强调理论和实践相结合，安排有大小结合的算法作业练习，以及结合实际应用的项目实践，以提高教学的质量和效率；课堂讲授包括数字取证基本算法原理、信号处理及数据处理方法、机器学习与深度学习模型算法，借助 AI Studio 平台提供的免费算力和数据集，完成数据处理、模型搭建、优化策略方法、数据可视化，完成项目总结。并结合课堂提问研讨，课后习题和答疑等环节激发学生的学习兴趣，增加团队学习、小组大作业、实验课和理论课的结合、使用信息技术方法、由教师和知识为中心转化为以学生和学习为中心。

图 1 基于飞桨 AI Studio 平台的数字取证课程管理界面

✧ 进一步完善考核学生学习质量的评估机制，培养学习兴趣，强调学生必须注重平时自觉主动的学习、必须注重自身能力的培养与素质的提高，而不能以应试作为学习目标。

✧ 引导学生查阅最新学术成果文献，跟踪信息安全专业领域的最新技术和研究动向，阅读国内外有影响力的方法，跟踪 AI 与计算机视觉领域的顶会论文算法，如 CVPR、AAAI、ECCV、ICML 等，并通过动手实践能够进行复现论文，鼓励学生尽可能开拓思路创新，为后续的学习或科研工作打下扎实基础。

百度基于飞桨的 AI Studio 一站式开发平台，给广大开发者提供了更加完善自由的编程环境，可帮助学生更快捷简便地完成深度学习项目，并持续提供更多的增值服务。该平台集合了 AI 相关教程、代码环境、算法算力和数据集，并为大家提供了免费的在线云计算编程环境，学生不需要再进行环境配置和依赖包等繁琐步骤，随时随地可以上线 AI Studio 开展深度学习项目。AI Studio 具有如下四大特性，能够为数字取证课程带来多方位的服务。

✧ 云端集成. 在云端集成了视频教程、样例模型、代码、计算资源、比赛平台等多种能力，学生可以一站式达到学、练、用的目的，免除环境配置的困扰。

✧ 简单易用. 为初学者准备了多个领域的不同模型范例，以及数十个经典数据集，供学习练习使用。

✧ 运行高效. 自动根据模型大小来分配计算资源，确保模型训练高效执行。

✧ 免费资源. 视频教程、云空间、计算资源全部免费，特别是在 AI 最新前沿技术方面，针对高校课程给予丰富的教学资源支持。

“数字取证“课程基于该平台建设并在线使用，其管理界面[25]如图 1 所示。在 AI Studio 平台可以直接创建项目，进行深度学习项目训练。同时，百度提供丰富的样例工程，每个案例从历史背景到代码实现都有非常详尽的内容，可以直接分叉（fork）现有项目。还提供了丰富数据集和 PaddleGAN 等套件，数据集包括从人脸检测、人像生成到人体特征、交通信息等各个方面的典型数据，以及针对深度伪造检测方面的数据集，学生可以在数据集标签栏中查看管理这些数据集，也可以在创建项目的时候直接调用。AI Studio 平台为学生的课程学习提供了相当宝贵的资源和算力，借助这些数据、算法与算力，可以为学生

对“数字取证”课程的学习与实践提供了很大帮助，会使学生在实践学习过程中收益匪浅，并为将来就业打下坚实的技术基础。

5 实践教学环节的探索与思考

实践教学环节的设计内容主要围绕两个方面，一方面是基于传统的统计信号处理技术的取证方法实践，另一方面是基于深度学习算法以及 GAN 图像特征的对抗式取证算法。

传统取证技术主要是建立在概率论与数理统计的基础之上，利用图像信号的一阶或高阶统计特征，通过检测数字媒体内容在篡改前后的差异进行取证，部分算法的实践项目列表如表 1 所示，其中部分数据集以作业形式可由学生自建完成。

基于深度学习的数字的媒体取证内容主要包括伪造生成图像、语音、视频的检测。目前比较流行的基于 GAN 的图像生成技术有 DCGAN[23]、LSGAN[24]、StyleGAN[25] 、 PRGAN[26] 、 Pixel2Pixel[27] 、ESRGAN[28]、RCAN[29]等模型算法。我们部分采用百度 AI Studio 平台上提供的飞桨开源实战项目作为教学案例，例如 DCGAN 实践项目，PGGAN 及 Pixel2Pixel 实践项目，并针对课程内容设计相应的实战检测项目如表 1 所示。

根据目前比较典型的 Deepfake 检测算法几个流派，不但有传统的基于图像篡改痕迹的方法，还有基于生理特征的检测方法，以及基于 GAN 图像指纹特征的检测方法。后一类方法将作为该课程在 Deepfake 检测部分的实践教学案例，例如，Two-stream 模型[31]使用了一个双流 CNN 在生成目的的图像伪造中达到最好性能；Xception 网络[32]是一个基于 XceptionNet 模型的检测算法，并在 FaceForensic++[32]数据集上训练；Capsule 网络[33]使用了基于 VGG-19 的胶囊结构作为 Deepfake 分类的骨干网络，该方法在 FaceForensic++数据集上完成训练并测试；DSP-FWA 模型[34]在 FWA 算法基础上，使用空间金字塔池化模块以更好地处理原始图像分辨率的变化，该算法在自建数据集上完成训练并得到很大性能提升。这些算法可以在课程平台上借助开源数据集完成训练和测试，让学生对基于深度学习的数字取证前沿技术有一个直观感受和深刻体会，实践项目内容设计如表 1 所示。

表 1 实践教学项目列表

编号	实践项目名称	数据集
实践 1	基于 PRNU 的图像来源取证算法	自建数据集
实践 2	基于相关性匹配的图像复制-粘贴检测	ImageForgery dataset
实践 3	基于直方图统计特性的图像双重 JPEG 压缩检测	自建数据集
实践 4	基于高阶统计特性的图像源辨识算法	自建数据集
实践 5	基于 DCGAN、PGGAN 的人脸图像生成实战	无需
实践 6	基于 GAN 特征的人脸生成图像检测	自建数据集
实践 7	基于 ESRGAN 的图像超分模型应用	卡通化图像数据集
实践 8	基于 GAN 特征的图像超分辨率检测	卡通化图像数据集
实践 9 大作业	基于深度学习的开放性 Deepfake 检测算法（参考最近两年顶会论文，自选完成一篇论文复现，并呈现检测效果，完成实验报告）	FaceForensic++ Celeb-DF, DFD DFDC preview dataset

另外一个重要的环节就是深度伪造开源数据集的选取和应用，目前学术界和工业界均已开源了一些用于伪造图像、视频检测的数据集，以促进该领域检测算法的研究，在大量数据集中，选用其中几个比较有代表性的典型数据集，如 FaceForensic++[32]、Celeb-DF[35]、DFDC[36]、DFD[37]等数据集如表 2 所示。

表 2 深度伪造开源数据集

数据集	篡改类型	描述	假真比	大小
FaceForensic++	DeepFakes FaceSwap Face2Face NeuralTexture	每一类篡改视频均被 C0,C23,C40 这 3 种参数压缩	1 : 1	5000 视频
Celeb-DF	Deepfakes	针对以往伪造视频质量差、不稳定等缺点进行改进后	1 : 0.11	6229 视频
DFDC preview dataset	Unkown	Deepfakes 竞赛的预赛数据	1 : 0.28	5244 视频
Deepfake-Detection (DFD)	DeepFakes	不同场景原视频，进行换脸，篡改视频均被压缩，压缩参数 C0,C23,C40	1 : 0.12	3431 视频

6 结束语

数字取证科学是近年来应运而生的新领域，“数字取证”课程的综合性和实践性都比较强，需要结合各种新问题、新技术不断地建设和完善其教学体系，特别是针对飞速发展的深度伪造技术。一方面，新的伪造技术和篡改形式还在不断出现；另一方面，反取证技术也对目前的取证技术产生一些干扰。此外除了图像造假，还不断涌现各种各样的音视频伪造技术，这些对我们数字取证研究发起了更严峻的挑战，并且已经引起国际业界的高度重视。如 2019 年 Facebook 与微软、MIT、亚马逊等宣布合作，共同打击深度伪造。该计划被称为 Deepfake 检测挑战赛（DeepFake Detection Challenge，DFDC），旨在创建能用于“打假”模型训练的开源工具。

在进行“数字取证”课程教学探索研究过程中，我们深刻体会到做好实践项目的设计与实施是课程建设的关键环节。因此，课程建设过程需要借助产业界资源，开展产学合作协同育人工程。本课程是借助百度公司 AI Studio 课程平台打造的在线课程，旨在培养一批在数字取证技术方面的新型人才，以便更好地应对这一来势凶猛的造假技术，能够检测声音、图像或视频等媒体内容是否被篡改，维护媒体信息传播的良好生态环境。

本文从课程内容的设置与实验环节设计的角度阐述了数字取证课程建设的一些探索与思考，每年 AI 顶会大量论文不断刷新媒体生成技术以及深度伪造检测算法，因此对课程的及时更新还存在一定难度，本文所讲有关课程内容建设还很粗浅，还望在后续的课程中不断地加以改进和完善。

参 考 文 献

[1] 李旭嵘,纪守领,吴春明,刘振广,邓水光,程鹏,杨珉,孔祥维.深度伪造与检测技术综述[J].软件学报,2021,32(02):496-518

[2] Sencar H T, Memon N. Digital Image Forensics[M]. Springer New York, 2013.

[3] Farid H. Digital Image Forensics.[J]. Scientific American, 2008, 298(6): 66-71.

[4] Farid H, Photo Forensic[M]. MIT press, 2016.3

[5] 周琳娜, 王东明. 数字图像取证技术[M]. 北京邮电大学出版社, 2008.

[6] 周琳娜, 张茹, 郭云彪. 数字图像内容取证[M]. 高等教育出版社, 2011.

[7] Lukas J, Fridrich J, Goljan M. Digital Camera Identification from Sensor Pattern Noise[J]. IEEE Transactions on Information Forensics & Security, 2006, 1(2):205-214.

[8] Fridrich A J, Soukal B D, Lukáš A J. Detection of Copy-move Forgery in Digital Images[C]//in Proceedings of Digital Forensic Research Workshop. 2003.

[9] Farid H. Detecting Digital Forgeries using Bispectral Analysis[J]. Technical Report, AIM-1657, MIT AI Memo, 1999.

[10] Popescu A C, Farid H. Exposing Digital Forgeries by Detecting Traces of Resampling[J]. IEEE Transactions on Signal Processing, 2005, 53(2): 758-767.

[11] Zhou L, Wang D, Guo Y, et al. Blur Detection of Digital Forgery using Mathematical Morphology[C]// KES International Symposium on Agent and Mult

i-Agent Systems: Technologies and Applications. Springer, 2007: 990-998.

[12] Fang Y, Emir A, Sun X, et al. Source Class Identification for DSLR and Compact Cameras[C]// In IEEE International Workshop on Multimedia Signal Processing (MMSP), 2009.

[13] Yin J, Fang Y. Markov-based Image Forensics for Photographic Copying from Printed Picture[C]// In the 20th ACM International Conference on Multimedia. 2012: 1113-1116.

[14] Yin J, Fang Y. Digital Image Forensics for Photographic Copying[C]// Media Watermarking, Security, and Forensics. International Society for Optics and Photonics, 2012, 8303: 83030F.

[15] 尹京, 方艳梅. 数码翻拍图像取证算法[J]. 中山大学学报(自然科学版), 2011, 050(006):48-52.

[16] Popescu A C, Farid H. Exposing Digital Forgeries in Color Filter Array Interpolated Images[J]. IEEE Transactions on Signal Processing, 2005, 53(10): 3948-3959.

[17] Popescu A C, Farid H. Statistical Tools for Digital Forensics[C]// International Workshop on Information Hiding. Springer, 2004: 128-147.

[18] Ker A D. A General Framework for Structural Steganalysis of LSB Replacement[C]// International Workshop on Information Hiding. 2005: 296-311.

[19] Fridrich J, Kodovsky J. Rich Models for Steganalysis of Digital Images[J]. IEEE Transactions on Information Forensics and Security, 2012, 7(3): 868-882.

[20] Deng X, Chen B, Luo W, et al. Fast and Effective Global Covariance Pooling Network for Image Steganalysis[C]// Proceedings of the ACM Workshop on Information Hiding and Multimedia Security. 2019: 230-234.

[21] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative Adversarial Nets. In Advances in Neural Information Processing Systems (NIPS), 2014, arXiv:1406.2661.

[22] Salimans T, Goodfellow I J, Zaremba W, et al. Improved Techniques for Training GANs[C]// Conference on Neural Information Processing Systems. 2016.

[23] Radford A, Metz L and Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks, 2016

[24] Mao X, Li Q, Xie H, et al. Least Squares Generative Adversarial Networks[C]// In IEEE International Conference on Computer Vision (CVPR), 2017: 2794-2802.

[25] Karras T, Laine S, Aila T. A Style-Based Generator Architecture for Generative Adversarial Networks[C]// In IEEE International Conference on Computer Vision (CVPR), 2019: 4401-4410.

[26] Karras T, Aila T, Laine S, et al. Progressive Growing of GANs for Improved Quality, Stability, and Variation[C]// In ACM International Conference on Learning Representations (ICLR), 2018.

[27] Isola P, Zhu J Y, Zhou T, et al. Image-to-Image Translation with Conditional Adversarial Networks[C]// In IEEE International Conference on Computer Vision (CVPR), 2017:1125-1134.

[28] Wang X, Yu K, Wu S, et al. ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks[C]// In European Conference on Computer Vision, Springer, 2018.

[29] Zhang Y, Li K, Li K, et al. Image Super-Resolution Using Very Deep Residual Channel Attention Networks[C]// In European Conference on Computer Vision, Springer, 2018: 286-301.

[30] Zhou P, Han X, Morariu V, et al. Two-stream Neural Networks for Tampered Face Detection[C]. In IEEE International Conference on Computer Vision (CVPR), 2017.

[31] Chollet F. Xception: Deep Learning with Depthwise Separable Convolutions[C]// In IEEE International Conference on Computer Vision (CVPR), 2017.

[32] Rssler A, Cozzolino D, Verdoliva L, et al. FaceForensics++: Learning to Detect Manipulated Facial Images[C]// In IEEE International Conference on Computer Vision (CVPR), 2017: 1831-1839.

[33] Nguyen H H, Yamagishi J, Echizen I. Capsule-forensics: Using Capsule Networks to Detect Forged Images and Videos[C]// In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019: 2307-2311.

[34] He K, Zhang X, Ren S, et al. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 37(9): 1904-1916.

[35] Li Y, Yang X, Sun P, et al. Celeb-DF A Large-scale Challenging Dataset for DeepFake Forensics[C]// In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020: 3207-3216.

[36] Dolhansky B, Howes R, Pflaum B, Baram N, and Ferrer C. The DeepFake Detection Challenge (DFDC) Preview Dataset. arXiv preprint arXiv:1910.08854, 2019.

[37] Dufour N, Gully A, Karlsson P, et al. Deepfakes Detection Dataset by Google & Jigsaw. 2019. https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html