文章编号：1003-0530(2021)12-2278-24

基于深度学习的数字图像篡改定位方法综述

李昊东[1,2,3] 庄培裕[1] 李 斌[1,2,3]

（1. 深圳大学电子与信息工程学院，广东深圳 518060；2. 广东省智能信息处理重点实验室，广东深圳 518060；
3. 深圳市媒体信息内容安全重点实验室，广东深圳 518060）

摘 要：日益进步的图像处理技术让数字图像编辑的门槛变得越来越低。利用触手可及的图像处理软件，人们可以方便地改动图像内容，而篡改后的图像往往十分逼真，以至于肉眼难以辨认。这些篡改图像已对个人隐私、社会秩序、国家安全造成了严重的威胁。因此，检测及定位图像中的篡改区域具有重要现实意义，并已成为多媒体信息安全领域中的重要研究课题。近年来，深度学习技术在图像篡改定位中得到了广泛的应用，所取得的性能已显著超越了传统的篡改取证方法。本文对基于深度学习的图像篡改定位方法进行了梳理。介绍了图像篡改定位中常用的数据集及评价标准，以在篡改定位中应用的不同网络架构为依据分析了现有方法的技术特点和定位性能，并讨论了图像篡改定位面临的挑战和未来的研究方向。

关键词：数字图像取证；图像篡改检测；篡改区域定位；深度学习

中图分类号：TP309.2 **文献标识码**：A **DOI**：10.16798/j.issn.1003-0530.2021.12.004

引用格式：李昊东，庄培裕，李斌. 基于深度学习的数字图像篡改定位方法综述[J]. 信号处理，2021，37(12)：2278-2301. DOI：10.16798/j.issn.1003-0530.2021.12.004.

Reference format：LI Haodong，ZHUANG Peiyu，LI Bin. A survey on deep learning based digital image tampering localization methods[J]. Journal of Signal Processing，2021，37(12)：2278-2301. DOI：10.16798/j.issn.1003-0530.2021.12.004.

A Survey on Deep Learning Based Digital Image Tampering Localization Methods

LI Haodong[1,2,3] ZHUANG Peiyu[1] LI Bin[1,2,3]

（1. College of Electronics and Information Engineering，Shenzhen University，Shenzhen，Guangdong 518060，China；
2. Guangdong Key Laboratory of Intelligent Information Processing，Shenzhen，Guangdong 518060，China；
3. Shenzhen Key Laboratory of Media Security，Shenzhen，Guangdong 518060，China）

Abstract：The sustained advancement of image processing technology makes digital image editing more and more easy. With popular image processing software，people can easily manipulate image contents. The manipulated images are becoming so realistic that they are difficult to be identified with the naked eyes，which have posed serious threats to personal privacy，social order，and even national security. Therefore，it is of great importance to detect and localize the tampered regions in digital images，which has attracted much attention in the field of multimedia information security. In recent years，deep learning technology has been widely adopted in image tampering localization and has significantly outperformed traditional forensic methods. This paper reviews the image tampering localization methods based on deep learning. It introduces the commonly used datasets and evaluation criteria for image tampering localization. Based on the applications of different

收稿日期：2021-07-12；修回日期：2021-09-27

基金项目：广东省重点领域研发计划(2019B010139003)；国家自然科学基金(61802262，U19B2022，61872244)；广东省自然科学基金(2019B151502001)；深圳市基础研究项目(JCYJ20200109105008228)

network architectures, the technical features and localization performance of existing methods are presented. In addition, the challenges of image tampering localization and future research directions are discussed.

Key words: digital image forensics; image tampering detection; tampering localization; deep learning

1　引言

如今,不断进步的图像处理技术及功能强大的图像处理软件已经高度融入了我们的生活。即便是不具备专业图像处理知识的用户,也能够随心所欲地利用各种图像处理工具对数字图像中的内容进行修改。诚然,这给人们带来了相当多的方便和乐趣,但也不可避免地导致一些安全问题。随着篡改图像的逼真程度越来越高,仅凭肉眼已经难以识别篡改图像,这使得“眼见为实”、“有图有真相”等传统观念受到了极大的冲击,人们对于数字图像真实性的置信程度显著降低。同时,不乏一些别有用心之人刻意篡改图像内容,并将篡改图像上传到网络及社交媒体中,企图传播虚假消息以扰乱或操纵公众舆论,这更是会造成十分恶劣的影响。例如,在新冠肺炎疫情期间,某网民修改了一张照片中疫情防控标牌上的内容,篡改图像在网络平台发布后被迅速扩散,严重扰乱了防疫公共秩序,而该网民最终也受到了相应的处罚*。可见,如果不对篡改图像加以监管,则它们必将对个人隐私、社会秩序、国家安全等方面造成严重威胁。为了应对篡改图像引发的安全隐患,亟需研究能够鉴定数字图像的真实性和完整性的技术,即数字图像取证技术[1]。

数字图像取证的目的是解决一系列有关图像认证的问题,如:图像通过何种成像设备获取,其生成方式是否可信,是否经过某些操作的篡改,篡改的区域处于图像的哪些位置等等。数字图像取证依据的基本原理是:在数字图像产生的过程中,实际场景内容、相机软硬件处理的特性等因素都会在图像中留下某些固有特征,当图像被篡改后,原始图像中的固有特征会不同程度地受到破坏或改变,通过提取与检测图像中的固有特征,就可以解决相关的图像取证问题。经过近二十年的发展,数字图像取证技术对图像进行分析的细化程度在不断提高。早期的取证技术仅能判断给定图像是否经过篡改,而不能预测图像中被篡改的区域或像素。显然,与图像级预测相比,像素级预测能提供更加细致的关于篡改的有效信息,也更符合现实应用场景的需求。因此近年来图像篡改定位问题越来越受到研究人员的重视,新的图像篡改定位方法不断涌现。另一方面,取证技术所依赖的基本工具也发生了更迭。随着深度学习在计算机视觉等应用中取得令人瞩目的优异性能,各种深度学习模型被引入到图像取证之中。深度学习技术和图像取证的领域知识相结合,提升了取证方法的性能,这使得传统的基于手工设计特征的取证方法逐渐淡出主流。

考虑到数字图像取证技术的研究现状,本文主要对基于深度学习的图像篡改定位方法进行梳理,同时介绍了图像篡改定位中常用的数据集及评价标准,并对该领域面临的问题和未来的研究方向进行了讨论。我们注意到,此前已有一些关于数字图像取证技术的综述工作[2-6],但本文关注的范围和叙述的角度与这些工作存在区别:本文主要关注图像篡改定位问题,因此对包括图像源识别、图像处理操作辨识、图像篡改检测等在内许多取证方法并没有赘述;同时,考虑到传统篡改定位方法的性能已落后于新提出的方法,本文集中讨论基于深度学习的图像篡改定位方法,而对传统篡改定位方法没有投入过多笔墨;与文献[6]相比,本文更加关注不同深度网络架构在篡改定位中的应用,因此采用了不同的方式对相关文献进行归类。

本文的后续部分安排如下:第 2 节简要介绍图像篡改的类型及篡改定位的发展过程;第 3 节介绍篡改定位中常用的数据集和性能评价指标;第 4 节对已存在的基于深度学习的图像篡改定位方法进行归纳和分析;第 5 节讨论图像篡改定位面临的挑战和未来的研究方向;第 6 节对全文进行总结。

* P 图篡改标牌扰乱公共秩序? 警方:拘留 4 日,http://m.news.cctv.com/2021/01/05/ARTIr67sHFgOWTnXdbwoZcqC210105.shtml,2021-01-05.

2 背景知识

2.1 图像篡改的类型

在数字图像取证中,一般将经过图像处理软件(或算法)编辑过的图像认为是篡改图像(非原始图像)。实际上,人们更加关心图像的语义是否经过篡改。通常,作用于图像全局的操作不会显著改变图像的语义,而图像语义的改变往往由局部篡改导致,故本文主要关注图像的局部篡改。按图像语义内容改变的结果可将图像篡改分为两种基本类型:

1)内容添加(Content addition):即在图像中增加本不存在的物体。

2)内容移除(Content removal):即把图像中实际存在的物体删除并将相应部位替换为与图像背景相吻合的内容。

在实际的篡改情形中,上述两种篡改结果可以单独出现,也可以相互组合。例如:当移动图像中物体的位置时,可认为是先对位于某区域的物体进行移除,再在图像的另一区域添加相同的物体;对图像中的内容进行替换时,可认为是先对位于某区域的原内容进行移除,再在同一区域中添加新内容。此外,在内容添加或移除的过程中,物体的形状、角度、尺寸等属性均可能发生变化,即发生形变,从而形成了各种各样的图像篡改实例。

另一方面,按在实施图像篡改时使用的具体操作来看,图像篡改主要包含三种类型:

1)拼接(Splicing):此类操作将来源于宿主图像之外的内容移接到宿主图像中的某个区域,移接的内容可以是从某一供体图像中抠图得到的物体,也可以是通过某些方式生成的物体。

2)复制移动(Copy-move):此类操作首先复制图像中的某些区域(源区域),并将它们移动覆盖到其他区域(目标区域)。

3)内容填充(Region filling):此类操作首先在图像中选择目标区域,然后在这些区域中填充与区域之外图像语义相匹配的内容。可有效实现内容填充的主流技术是图像修复技术(Inpainting)。

不难看出,以上篡改操作可产生不同的篡改结果。其中,利用拼接和复制移动都可以实现内容添加及移除,但篡改部位的来源存在差异,前者来自图像外部,而后者来自图像本身;内容填充在绝大多数情况下被用于移除图像中的物体,因此一些已有文献[7-8]也将第三类篡改操作称为“移除”。为了保证篡改图像的视觉逼真性,这些篡改操作往往还伴随着旋转、缩放、亮度调整、边界平滑、滤波等处理操作,而且篡改图像在保存的过程中通常还会受到有损压缩。这一系列操作都会在图像上引入相关的痕迹,而这些痕迹则是可以进行图像篡改检测及定位的基础。

2.2 图像篡改定位技术发展历程

所谓图像篡改定位,指的是识别并标记出图像中存在的篡改区域。如图 1 所示,在图像篡改定位任务中,需要设计开发有效的算法和模型,以待测图像作为输入,相应地输出一张与输入图像尺寸相同的概率图,其中各元素取值的大小表示相应位置像素被篡改的可能性的大小,对此概率图进行阈值化后,得到一张二值图像(掩模)作为输出结果。由此可见,图像篡改定位实质上是一个像素级别的二分类问题。篡改定位与篡改检测有着密切的联系,它们都需要判断图像是否遭到篡改;但两者又存在显著的区别:后者只是判断给定图像是原始的还是经过篡改的,利用图像的全局信息进行决策,决策精细度要求低;前者则需利用图像的局部信息进行决策,决策精细度要求高。因此,篡改定位的技术水平要求比篡改检测更高。

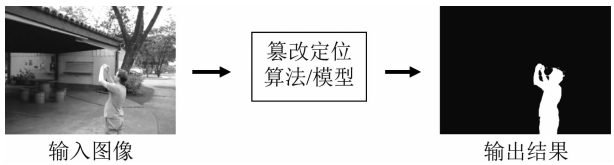


图 1 图像篡改定位任务的示意图

Fig. 1 Illustration of image tampering localization task

在图像篡改取证的早期研究中,受限于技术水平,相关工作主要集中在篡改检测方面。其中,一部分工作着眼于检测图像是否经历过某些典型图像操作的处理[9],如重采样[10-11]、滤波[12-13]、JPEG 压缩[14-15]等,另一些工作则尝试对图像整体是否经受了拼接[16-17]、复制移动[18-20]等操作进行判断。尽管这些方法在它们各自所关注的篡改检测任务上取得了良好的性能,但大多数并不适用于篡改定位任务。这可从 2013 年 IEEE 信息取证与安全技术委员会(IFS-TC)组织的首届图像取证竞赛中得到

验证。该竞赛包含了篡改检测与篡改定位两个任务。从比赛结果看，篡改检测在一定条件下能达到相当高的精度[21]（ACC=0.9421），而篡改定位的性能则还有很大的提升空间[22]（F1-score=0.4072）。

从可行性上分析，大多数针对整张图像的篡改检测方法可以通过滑动窗口的方式来实现篡改定位，但它们都避免不了窗口尺寸与定位性能相互制约的问题——提高定位的细致度需要减小滑动窗口的尺寸，然而窗口尺寸减小往往就会导致算法准确率显著降低。为了提高篡改定位的性能，通常需要结合小尺寸滑动窗口的特性，设计合适的特征提取算法或决策算法。例如：Bianchi 和 Piva[23] 提出在 8×8 的图像块中分析 JPEG 重压缩痕迹来定位篡改区域；Ferrara 等人[24] 尝试在最小为 2×2 的图像块中提取去马赛克痕迹来判断是否发生了篡改；Chierchia 等人[25] 利用贝叶斯框架来估计传感器模式噪声，并结合马尔可夫随机场（MRF）对像素间关系建模，得到关于篡改区域的预测；Fan 等人[26] 提出利用高斯混合模型（GMM）对图像小块的统计特性进行建模；Korus 和 Huang[27] 通过融合多种尺寸滑动窗口的输出结果来提高篡改定位的性能。虽然这些方法在某些篡改定位任务上取得了不错的效果，但仍有一些因素影响了它们进一步的发展。首先，它们所采用的基于手工设计的特征受到一定先验条件的限制，往往并不能较好地应对现实情况中纷繁复杂的图像来源和篡改手段。其次，基于滑动窗口的定位策略在使用时不易确定较优的窗口尺寸参数，而且计算效率不高，尤其是当图像尺寸较大时处理时间显著增加。

近年来，深度学习在计算机视觉、语音识别、自然语言处理等领域引领了技术进步的潮流。在此过程中，深度学习技术也被逐渐引入到图像篡改定位领域。首先，借助深度学习对复杂数据的强大表征能力，可以通过深度网络自动地从图像中提取有效的图像篡改特征[8]，从而可在一定程度上摆脱传统方法对手工设计特征的依赖。其次，利用合适的深度网络架构，可以构造端到端的图像篡改定位模型，直接输出篡改定位结果[28]，从而可以摒弃低效的滑动窗口策略。再者，使用深度网络还可以更灵活地结合不同特性的图像表征作为输入，并设置不同的学习目标，通过多任务联合优化来提升定位的性能[29]。由于具有这些优势，基于深度学习的图像篡改定位方法在技术路线上比传统方法更加合理。另外，近年来篡改图像数据的规模不断增大，这为充分发挥深度学习模型的潜能提供了必要条件。技术方案和数据质量的共同进步，促使基于深度学习的方法在图像篡改定位中取得良好的性能，占据了主流地位。

3　数据集与性能评价指标

数据对深度学习有着至关重要的作用，数据集的质量直接影响着所构建模型的性能，因此只有使用符合篡改定位任务要求的数据集，才能够获得可靠的篡改定位模型，进而才能够合理地评估篡改定位方法的优劣。另一方面，篡改定位任务中通常存在严重的类别不均衡问题，即原始像素多、篡改像素少，因而也需要使用合理的性能评价指标，才可以客观地反映篡改定位方法的有效性。本节将总结和介绍图像篡改定位中常用的数据集和性能评价指标。

3.1　数据集

收集和构造适用于篡改定位任务的图像数据集并非易事。在图像处理操作检测等任务中，通常可以利用程序对图像进行批量处理来生成大量数据，但使用类似的方法难以得到高质量的篡改图像数据集。这是因为数据集中的图像理应能够客观反映实际的篡改情形，这就要求在原始图像中所进行的修改应确实扭曲了其语义，且所得图像不应含有明显的视觉异常痕迹。然而，目前还无法通过计算机程序来较好地实现这样的功能。同时，为了辅助分类器的训练，需要为每张篡改图像提供相应的像素级标签。这又使得很难直接收集网络中潜在的大量篡改图像作为数据集，因为它们并不携带所需的像素级标签。综上而言，一种较理想的构造篡改图像数据集的方法是在可控条件下通过人工产生篡改图像，但这显然需要耗费大量人力物力。目前，已存在一些公开的篡改图像数据集，相关的信息整理总结如表 1 所示。

表 1 数字图像篡改定位中常用的数据集

Tab. 1 Commonly used datasets for image tampering localization

名称	发布时间	图像格式	图像尺寸	图像数量（真/假）	篡改方式	像素级标签
Columbia gray [30]	2004	BMP	128×128	933 / 912	拼接	无
Columbia color [31]	2006	BMP、TIF	757×568-1152×768	183 / 180	拼接	有
CASIA v1 [32]	2013	JPG	384×256	800 / 921	拼接、复制移动	无
CASIA v2 [32]	2013	JPG、BMP、TIF	240×160-900×600	7200 / 5123	拼接、复制移动	无
IEEE IFS-TC [34]	2013	PNG	1024×768-3000×2500	1050 / 1150	拼接、复制移动	有
DSO-1 (Carvalho) [35]	2013	PNG	2048×1536	100 / 100	拼接	有
CoMoFoD [36]	2013	PNG、JPG	512×512-3000×2000	260 / 260	复制移动	有
Coverage [37]	2016	TIF	400×486	100 / 100	复制移动	有
Wild Web [38]	2015	PNG、BMP、JPG、GIF	72×45-3000×2222	90 / 9657	真实案例	有
RFD-Korus [44]	2016	TIF	1920×1080	220 / 220	拼接、复制移动	有
NIST NC 16 [45]	2016	JPG	500×500-5616×3744	560 / 564	拼接、复制移动、移除	有
NIST NC 17 [45]	2017	RAW、PNG、BMP、JPG	160×120-8000×5320	2667 / 1410	多种操作	有
MFC 18 [45]	2018	RAW、PNG、BMP、JPG、TIF	128×104-7952×5304	14156 / 3265	多种操作	有
MFC 19 [46]	2019	RAW、PNG、BMP、JPG、TIF	160×120-2624×19680	10279 / 5750	多种操作	有
PS Battle [47]	2018	PNG、JPG	130×60-10000×8558	11142 / 102028	多种操作	无
DEFACTO [48]	2019	TIF	240×320-640×640	-/ 229000	多种操作	有
FantasticReality [50]	2019	JPG	280×800-6000×4000	16592 / 19423	拼接	有
IMD 2020 [51]	2020	PNG、JPG	193×260-4437×2958	37010 / 37010	多种操作	有

据我们所知，第一个公开的篡改图像数据集是由哥伦比亚大学的研究团队在 2004 年发布的 Columbia gray 数据集[30]。该数据集共有 1845 张大小为 128×128 的灰度图像，其中包含 933 张真实图像、912 张拼接篡改图像。随后，该团队发布了 Columbia color 数据集[31]，其中包含 183 张真实图像和 180 张拼接篡改图像。这些图像都是彩色图像，且尺寸不尽相同。这两个拼接图像数据集都采用了随机拼接的办法来得到篡改图像，篡改区域没有经过任何后处理，且篡改后的图像均保存为非压缩格式，因此篡改图像具有明显的视觉异常，远不能和实际的篡改图像相比拟。

由中科院自动化所发布的 CASIA 数据集[32]在图像篡改定位的相关工作中被广泛使用。CASIA 数据集包括 v1 和 v2 两个版本，后者图像数量更多且包含不同格式的图像。CASIA v1 中图像的篡改区域没有经过后处理，存在肉眼可见的异常痕迹；CASIA v2 中的图像在进行拼接或复制移动后，篡改区域会经过合适的后处理，故看起来更加逼真。CASIA 数据集中并未提供关于篡改区域的像素级标签，但可以通过将篡改图像和相应的真实图像进行比较来产生近似的像素级标签。文献[33]指出该数据集中的图像使用了若干不同的质量因子进行 JPEG 压缩，因此在使用该数据集时应充分考虑到 JPEG 压缩因素对性能的影响，譬如适当地通过数据增强来提高算法的性能。

IEEE 信息取证与安全技术委员会组织的图像取证竞赛中也发布了一个篡改图像数据集[34]，其中包含 1050 张真实图像和 1150 张篡改图像。这些篡改图像是通过人工篡改产生的，主要使用拼接、复制移动等篡改操作，且应用了适当的后处理操作，大部分图像的篡改效果较好。Carvalho 等人[35]在以上数据集中挑选部分篡改效果逼真拼接图像构成了 DSO-1 数据集（也称为 Carvalho 数据集），其中包含 100 张真实图像和 100 张拼接图像。IEEE IFS-TC 数据集和 DSO-1 数据集中图像均为 PNG 格式，但其中的大部分图像事先都经历过 JPEG 压缩。

复制移动作为一种典型的篡改操作，一直是许多取证方法的检测目标，因而也出现了不少专门为复制移动检测构造的数据集。最常用的数据集之一是 Tralic 等人公开的 CoMoFoD 数据集[36]，其中包含 200 张低分辨率图像（512×512）和 60 张高分辨率图像（3000×2000）。为更好地评估复制移动检测算法的鲁棒性，数据集作者在对这些图像进行复制移动操作时，还伴随使用了旋转、缩放、加噪、模糊、压缩等处理。另一常用的复制移动数据集是 Coverage 数据集[37]。该数据集除了包括常规的复制移动操作和相应的多种后处理外，还引入“相似但真实

的物体”(Similar but Genuine Objects)，这给复制移动检测带来了新的挑战。由于现有的复制移动数据集所包含的图像数量并不多，它们通常被用于测试阶段，而较少被用于训练深度网络模型。

为了给图像篡改定位提供更实际的数据，Zampoglou 等人[38]从网络上收集一系列真实的图像篡改案例构成了 Wild Web 数据集。该数据集包含 80 个实际图像篡改案例(共 90 个子案例)，每个案例中有若干不同版本的篡改图像。通过对比属于同一案例的图像，作者给出了标记篡改区域的像素级标签。他们还利用此数据集测评了一些传统的篡改定位方法的性能[39]，包括基于 JPEG 压缩误差[40]、DCT 系数分布[23,41-42]、去马赛克痕迹[24]、局部噪声不一致[43]等特征的方法。Korus 也提出了一个篡改效果较好的数据集 Realistic Tampering Dataset(RTD)[44]，其中包含 220 组原始图像和篡改图像。这些图像来自于 4 台不同的相机，涉及的篡改操作包括拼接、复制移动等。除了像素级标签外，该数据集中还加入了图像的模式噪声信息，这有助于设计基于相机模式噪声的篡改取证方法。

自 2016 年起，美国国家标准与技术研究院(NIST)发布了一系列篡改图像数据集[45]。首个数据集 NC16 包含 564 张篡改图像，其中包括内容相同的篡改图像的不同版本：篡改区域的边界经过或不经过后处理，图像被以高低不同的质量因子进行 JPEG 压缩。这有助于研究篡改定位方法受这些因素影响的程度。在后续的数据集 NC17、MFC18、MFC19[46]中，内容相同的篡改图像不再提供多个版本，但图像的数量显著增加，而且图像的分辨率、格式、以及涵盖的篡改操作更加丰富多样。这给图像篡改定位方法提供了更具挑战性的评测数据。

随着深度学习技术在图像篡改定位中的流行，篡改图像数据匮乏的问题越显严重。这也促使人们构造容量更大的篡改图像数据集。Heller 等人[47]从图像处理爱好者活跃的网络社区中收集了超过 10000 个图像篡改实例，每个实例都包含真实图像及若干内容和篡改方式各异的篡改图像，形成包含超过 10 万张篡改图像的 PS-Battles 数据集。遗憾的是，该数据集并不包含像素级标签。Mahfoudi 等人[48]在 MS COCO 数据集[49]的基础上通过构造了 DEFACTO 数据集。该数据集包含超过 22 万张篡改图像，涵盖了拼接、复制移动、物体移除、人脸变形等多种方式的篡改。Kniaz 等人发布了 FantasticReality 数据集[50]，其中包含超过 16000 张真实图像和 19000 张篡改图像。该数据集的篡改图像都是拼接得到的，其中约一半的拼接图像为“粗糙”样本，即没有经过细致的修改，包含明显的篡改痕迹；而另一半则经过人为润饰，肉眼不易发现其异常。除了篡改区域的像素级标签外，该数据集还为 10 类不同的物体提供了标签。Novozamsky 等人也构建了一个较大规模的数据集，称为 IMD2020[51]。其中包含 35000 张由 2322 台不同相机拍摄的原始图像，以及 35000 张经由拼接、复制移动、修复、形变等多种操作得到的篡改图像。此外，该数据集还含有 2010 张从网络中收集的实际篡改图像以及它们相应的真实图像。

值得注意的是，基于深度学习的方法对数据规模的要求较高，万级甚至十万级的数据量可能还不足以训练得到可靠的模型。因此，在一些基于深度学习的图像篡改定位工作中还会利用其他数据集来自动化地生成大量篡改图像作为训练数据[7,29]，但这种自动生成的篡改图像质量并不太高。就已有文献来看，MS COCO[49]、Deresden[52]、ImageNet[53]、MIT Places[54]、SUN[55]等几个数据集常用作自动生成篡改图像的原始素材。

3.2　性能评价指标

如前所述，图像篡改定位实际上是一个像素级别的二分类问题。因此，篡改定位模型的性能可以通过常用的分类评价指标衡量。常用的评价指标主要有：精度(Accuracy，ACC)、F1-分数(F1-score)、ROC 曲线下面积(Area Under the Curve，AUC)、马修斯相关系数(Matthews Correlation Coefficient，MCC)以及交并比(Intersection over Union，IoU)等。下面对这些性能指标的定义及适用性进行简要说明。

3.2.1　ACC

篡改定位模型所得预测结果中的元素可分为四类：篡改像素被判断正确，原始像素被判断正确，原始像素被判断错误，篡改像素被判断错误。将以上元素的数量分别记为 TP，TN，FP，FN 那么模型的预测精度由下式给出：

$$\mathrm{ACC}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{1}$$

精度对于正负两类样本给予相同的权重，而在

实际的篡改场景中,篡改像素与原始像素通常存在严重的比例不均衡现象。例如,在 IMD2020 和 CASIA v1 数据集中,篡改像素占全部像素比例分别只有 0.076 和 0.087。这意味着,即使一个模型把所有像素都预测为原始的,其精度也能超过 90%。可见将精度作为篡改定位性能的评价指标并不能有效地衡量一个模型的好坏。

3.2.2 F1-score

在样本类别不均衡的情况下,精度不能客观地反映模型的性能,另一更合适的性能指标是 F1-score。F1-score 是查准率(Precision)和查全率(Recall)的调和平均,即:

$$\mathrm{F1}=2\times\frac{\mathrm{Precision}\times\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}} \tag{2}$$

其中,查准率和查全率分别定义为:Precision = TP/(TP+FP),Recall = TP/(TP+FN)。由于 F1-score 综合考虑了查准率和查全率,因此与精度相比其能更好地衡量篡改定位模型的性能。值得注意的是,一些现有工作在计算平均 F1-score 时,会对模型输出的每张篡改概率图都用其各自的最优阈值进行二值化,这在实际取证场合中并不适用,因为最优阈值无法确定。

3.2.3 AUC

受试者工作特征曲线(Receiver Operating Characteristic curve,ROC 曲线)通常被用于展示分类模型的性能。ROC 曲线分别以假阳性率和真阳性率作为横轴和纵轴,综合反映了两者之间的关系。ROC 曲线下面积,即 AUC,是评价分类模型性能优劣的一个重要指标。AUC 的取值在 0 到 1 之间,值越大则模型的分类性能越好。在评价具体的篡改定位结果时,AUC 所反映出来的差距有时并不如 F1-score 明显。如图 2 所示,A 和 B 两种方法所得的 AUC 值十分接近(1.00 和 0.98),但对两种方法输出的概率图以阈值 0.5 进行二值化后,计算所得的 F1-score 有相当大的差距(0.89 和 0.03)。即使分别以最优阈值对两者输出的概率图二值化,所得 F1-score 依然有显著差异,而从图 2(d)和(f)也可看出方法 A 的定位结果确实更准确。由此可见,F1-score 对篡改定位的性能更加敏感。

3.2.4 MCC

马修斯相关系数(Matthews Correlation Coefficient,MCC)也是一种被常用于对类别不均衡分类问题进行性能评价的指标,其计算公式为:

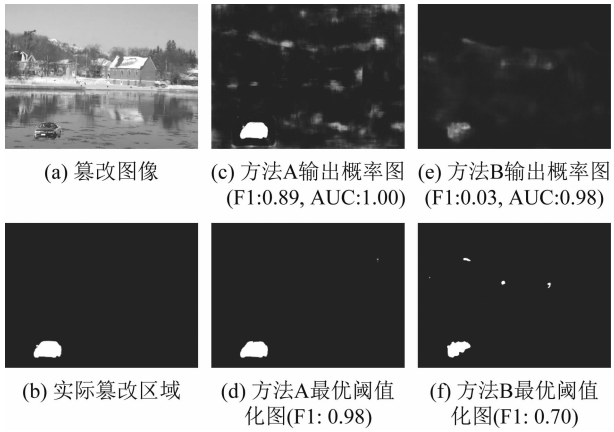


图 2 两种不同方法得到 AUC 和 F1-score

Fig. 2 AUC and F1-score obtained by two different methods

$$\mathrm{MCC}=\frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \tag{3}$$

MCC 的值介于-1 到 1 之间,-1 说明分类器将正负样本完全分错,0 表示分类器性能与随机猜测相当,1 则表示分类器能够实现完美分类。

3.2.5 IoU

在篡改定位中,有时还会使用图像语义分割领域常用的交并比(IoU)来评估性能的好坏。此时,IoU 记为模型预测的篡改区域和实际篡改区域的交集的面积与两者并集的面积之比,可按下式计算:

$$\mathrm{IoU}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}} \tag{4}$$

IoU 的取值介于 0 到 1 之间,值越大则意味着模型的性能越好。而且,对于固定的测试集,IoU 与 F1-score 总是呈现出正相关的关系。

4 基于深度网络的图像篡改定位方法

迄今为止,学者们已提出了许多基于深度学习的图像篡改定位方法。从这些方法的取证目标来看,部分方法仅限于检测某种特定的篡改操作,如拼接、复制移动、内容修复等,而部分方法则更注重通用性,可用于应对多种图像篡改操作。文献[6]从上述角度对部分已有方法进行了归纳总结。我们注意到:这些方法都不可忽视的一个关键问题是如何使用合适的深度网络架构来实现篡改定位;而

某些专用于特定篡改操作的方法，如果给予合适的数据，也能训练得到具有通用性的篡改定位模型。因此，本文尝试按所使用的深度网络架构来对现有的相关方法进行归类，以期提供一个不同的角度来帮助读者了解图像篡改定位的研究现状。

回顾已有的基于深度学习的图像篡改定位方法，自编码器[56]（Autoencoder，AE）在部分工作中被作为基础框架，而卷积神经网络[57]（Convolutional Neural Network，CNN）则被普遍地使用。特别地，在图像语义分割和目标检测中常用的全卷积网络[58]（Fully Convolutional Network，FCN）和 R-CNN（Region proposals with CNN features）系列网络[59-60]得到大量应用，而孪生网络[61]（Siamese Network）在部分问题中也占据一席之地。此外，长短期记忆网络[62]（Long short-term memory，LSTM）和生成对抗网络[63]（Generative Adversarial Network，GAN）也有相关的应用。下面将对基于不同网络架构的各类方法进行详细介绍。

4.1　基于自编码器的方法

自编码器是一种无监督学习框架，它包含编码器与解码器两部分，其中编码器将输入图像映射成维数较低的特征，解码器则利用该特征重构输入图像。在训练过程中，通常将网络输出与输入的均方误差（MSE）作为损失函数来对网络参数加以优化，使得输出和输入逐渐接近，这样就能使编码器提取到关于图像的有效特征表达。

Zhang 等人[64]较早地将自编码器应用于图像篡改定位。该方法构造了一个 3 层堆栈自编码器，并在其后加上一个全连接层来进行分类。其中，自编码器的输入并非图像本身，而是从图像块中提取的 450 维小波特征。堆栈自编码器中的参数由贪婪的逐层无监督训练确定，而全连接层参数则通过图像块的类标进行监督训练得到。由于以图像块作为输入，该方法依然需要使用滑动窗口的方式来得到篡改定位结果。为了提高性能，该方法利用相邻块的输出结果来提高预测的准确性，其对 32×32 图像块的检测精度可达到 91.09%。在后续的工作中，作者还对输入图像进行语义分割，结合图像语义分割的结果来进一步提高篡改定位的准确度[65]，在 CASIA v2 数据集中获得了 F1-score 为 0.5839 的结果。

Cozzolino 等人[66]提出了一个基于异常检测的无监督的篡改定位方法，可直接使用单张待测图像训练模型来得到篡改定位结果。他们在图像块中提取 108 维的空域富模型[67]（Spatial Rich Model，SRM）特征，然后将特征送入仅包含一个隐藏层的自编码器进行重构。得到图像中所有块的重构特征后，通过大津算法对重构误差选择阈值，将图像块标记成两部分，然后利用重构误差和标记结果形成的损失函数更新自编码器参数，如此迭代训练。当输出的标记结果趋于稳定后即停止迭代，这时将数量较少的一部分像素认为是篡改的。在该工作中，作者发现隐藏层的神经元数量小于输入特征的维数时，会产生大量的虚警错误，而只有将隐藏层的神经元数量增加才能得到较好的结果。进一步地，通过将自编码器与 LSTM 结合，上述方法被扩展用于识别视频中的篡改区域[68]。

以上基于自编码器的篡改定位方法均使用图像的某种特征作为网络输入，故而它们的性能也会受到所用特征的影响。与之不同的是，Yarlagadda 等人[69]直接将原始图像块的像素送入自编码器网络，并加入一个判别器来进行对抗性训练，使得自编码器学习到关于原始图像的良好的特征表达。接着，使用自编码器得到的特征训练一个基于支持向量机（Support Vector Machine，SVM）的单分类器，对篡改图像块和原始图像块进行辨别。该方法在训练过程中仅需使用原始图像，并在一个卫星图像数据集中验证了其定位篡改区域的能力。

4.2　基于卷积神经网络的方法

卷积神经网络在多种图像处理任务中有着出色的表现，因而它在图像篡改定位中也受到了广泛的使用。Rao 等人[70]提出一个基于 CNN 的拼接和复制移动检测方法。与计算机视觉中的常规做法不同的是，该方法结合了取证领域的先验知识，将 CNN 第一层的卷积核设置为 SRM 中的高通滤波器，用于提取图像的残差信息。该 CNN 使用图像块进行预训练。训练完成后，融合图像中各图像块通过 CNN 得到的特征，进一步训练一个 SVM 分类器来实现分类。此方法在 Columbia gray 和 CASIA 数据集中获得了超过 96% 的图像级别检测精度，但它并不能给出篡改定位的结果。文献[71]对此方法进行了改进：首先，对用 SRM 滤波器初始化 CNN 首层卷积核的策略进行优化，使得到的残差更多样，

同时通过受约束的学习策略来微调这些卷积核以保持它们的高通滤波特性;其次,在训练 CNN 时加入了对比损失函数,这可在特征空间中增大类间差异而减小类内差异,有助于提高模型的泛化性;再者,改进了融合图像块特征的方法,提高了算法对于 JPEG 压缩的鲁棒性;最后,在 CNN 网络之后加入条件随机场(Conditional Random Field,CRF),实现了像素级别的篡改定位。在使用 CASIA v2 数据集训练,DSO-1 数据集测试时,该方法得到的 F1-score 为 0.5813。与直接使用 SRM 中的滤波器对卷积核初始化不同的是,Cozzolino 等人[72]使用只有两个卷积层的 CNN 来近似 SRM 的提取过程。CNN 网络首先在约束条件下被初始化为可输出近似 SRM 特征的状态。随后可放松约束并对网络进行微调,使其性能进一步提升。

除了应用基于 SRM 的初始化策略,其他一些传统取证方法的经验也有与 CNN 结合的例子。考虑到不同相机拍摄的图像会表现出不同的特性,Bondi 等人[73]设计了一个可用于区分相机模型的 CNN 网络,通过该网络提取图像块的特征。同时,利用图像块的统计特性来评估特征的置信度,然后采用聚类的方法推断图像中的拼接篡改区域。在此方法中,需要使用已知拍摄相机的图像对 CNN 进行训练。重采样检测是一类重要的图像处理操作取证方法,Bunk 等人[74]将重采样特征与 CNN 结合来实现篡改检测和定位。具体地,他们使用文献[11]提出的重采样检测方法提取图像块的特征作为 CNN 的输入,然后使用 CNN 进行分类来得到预测结果。作者还发现使用 CNN 来近似文献[11]中的特征提取过程(即:依次进行拉普拉斯滤波、Radon 变换、FFT 等操作)并不能获得更好的性能。由于拼接会产生照度信息的不一致,Pomari 等人[75]将图像的照度图(Illuminant Maps)输入 ResNet-50[76] 网络提取特征,然后使用 SVM 对所得特征分类产生拼接检测结果。当图像被检测为拼接时,使用 Grad-CAM[77] 技术生成类激活热力图,并将它转换到 HSV 颜色空间,最后根据 H 通道的信息得到篡改定位结果。

在篡改定位中,使用单一特征往往不易取得令人满意的结果。传统做法是利用不同特征得到分类结果,然后再融合这些结果进行决策[78-79]。这在基于深度学习的篡改定位中也得到借鉴。Zhou 等人[80]提出一个双分支网络来检测图像中被篡改的人脸:一个分支使用 InceptionNet[81] 来判断提取的人脸是否真实,另一分支则基于三元组损失函数[82](Triplet Loss)使用网络对图像块中提取的隐写分析特征[83]进行优化,并送入 SVM 进行分类,最终结果由两个分支输出的预测分数加权求和得到。该工作构造了 FaceSwap 和 SwapMe 两个人脸篡改数据集,使用前者训练模型并在后者上测试,所得 AUC 为 0.927。除了直接对输出结果进行融合,使用深度网络还能更方便地在特征层面上进行融合。Shi 等人[84]利用两个 CNN 子网络分别对输入图像的空域和小波域进行处理。其中,工作于空域的 CNN 对图像滤波残差进行二维卷积,工作于小波域的 CNN 对基于小波的统计特征向量进行一维卷积,各自得到 256 维特征。两个域的特征随后被连接在一起,再经过两层全连接层得到分类结果。该工作使用 CASIA v2 数据集训练网络,在 Columbia 和 DSO-1 数据集中分别得到了 0.69 和 0.58 的 F1-score。Xiao 等人[85]提出一种粗网络和精网络相结合的拼接检测与定位方法。该方法首先将图像输入基于 VGG16[86] 的粗网络,大致地预测出可疑的篡改区域,然后再利用基于 VGG19[86] 的精网络对粗网络的预测结果进行改善。文中还设计了一种自适应聚类算法,通过异常点过滤和凸包填充来得到更准确的篡改定位结果。该方法在 Columbia 和 CASIA v2 数据集分别获得了 0.6950 和 0.6758 的 F1-score。然而,受限于所用的聚类算法的特性,该方法仅适用于包含单个篡改区域的图像。

基于 CNN 也能构建针对复制移动篡改的检测和定位模型。Ouyang 等人[87]对在 ImageNet 中预训练的 AlexNet[57] 进行微调得到一个复制移动检测模型。该模型在仿真产生的复制移动篡改图像上表现较好,但对实际的复制移动篡改图像表现不佳,且该模型也不具备定位篡改区域的能力。Wu 等人[88]提出一个端到端的框架来执行传统复制移动检测中涉及的特征提取、匹配、后处理三个步骤。该方法首先将图像输入不包含全连接层的 VGG16 网络得到相应的特征,然后对特征图计算自相关来表示不同位置的相似程度,并根据自相关值的大小通过逐点卷积生成关于复制移动区域的匹配信息,最后通过解码网络得到定位结果。为充分地对网

络进行训练，该工作利用 SUN、MS COCO、ImageNet 中的图像仿真生成了大量训练数据，且在训练过程中使用了压缩、模糊、加噪等处理进行数据增强。所训练的模型在 CASIA v2 中得到了 F1-score 为 0.7572的检测性能，比传统的复制移动篡改检测算法更好，且运行时间也更短。为提高针对复制移动取证的性能，Zhong 和 Pun[89] 设计了一个包含密集连接[90]的 InceptionNet 来进行多尺度特征分析和分层特征匹配，这对训练中未出现的物体类型有更好的检测效果。Zhu 等人[91]则在网络中引入关于位置和通道的注意力机制来提高检测性能，并加入一个精细化模块来优化输出的定位结果。以上方法尽管得到了较好的性能，但并没有解决复制移动篡改定位中一个棘手的问题，即定位结果存在歧义，不能区分复制移动的源区域与目标区域。实际上，通过灵活配置深度网络的结构，可以消除这种歧义性。Wu 等人[92]设计了一个包含两个并行分支的 BusterNet 网络，其中一个分支用于检测由复制移动及其后处理操作引入的痕迹，另一个分支用于分析图像中相似的区域，将两个分支的结果融合就可以给出不存在歧义的复制移动篡改定位结果。实验结果表明，该方法在 CASIA v2 的复制移动篡改图像和 CoMoFoD 数据集中能分别得到 0.4556 和 0.4926 的 F1-score；在检测到复制移动篡改时，该方法能以接近 78% 的成功率区分出源区域和目标区域。BusterNet 需要在两个分支均能识别到篡改区域时才能得到良好的结果，为克服此不足，Chen 等人[93]提出一个串行的解决方案：首先使用一个子网络（CMSDNet）来进行相似性检测，得到配对的复制移动区域，然后使用另一个子网络（STRDNet）来进一步区分配对区域中的源区域和篡改区域。与 BusterNet 相比，该方法定位的准确度以及区分源区域和目标区域的成功率均有显著提升。

4.3　基于全卷积网络的方法

上一小节提到的大多数方法，是将图像划分为小块输入网络提取特征的，然后利用网络或额外的分类器来判断图像块是否篡改。这意味着它们依然需要通过滑动窗口来给出篡改定位结果，也就无法避免滑动窗口策略存在的缺陷。即使有些工作尝试使用多个 CNN 提取不同尺寸图像块的特征，然后通过融合多尺度的预测结果来进行篡改定位[94]，也无法从根本上解决这一问题。全卷积网络，即 FCN，为解决这一问题提供了良好的框架。全卷积网络由 Long 等人[58]提出，最初被应用于图像语义分割。如图 3 所示，FCN 和用于分类的普通 CNN 最主要的区别是使用卷积层代替了全连接层，因而可以处理任意尺寸的输入图像。一般地，FCN 中包含一个用于提取特征的编码网络，以及一个将提取的分辨率较小的特征图放大到与输入图像尺寸一致的解码网络，这样就能对图像进行逐像素点的“稠密”预测。由于具有上述特点，全卷积网络适用于图像篡改定位任务。

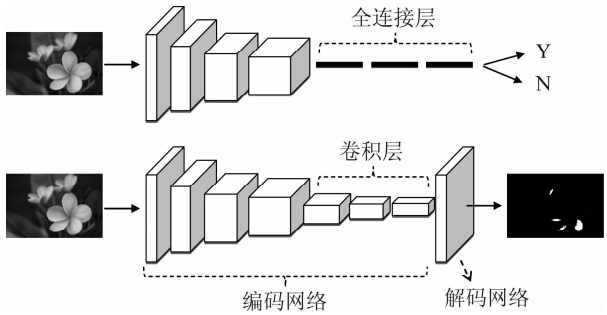


图 3　普通卷积网络（上）和全卷积网络（下）的对比

Fig. 3　Comparison between vanilla CNN (top) and FCN (bottom)

Salloum 等人[28]较早地将 FCN 用于图像拼接的篡改定位。他们借鉴了文献[58]提出的全卷积网络结构，并在输出端设计了两个对应于不同任务的分支—— 一个预测拼接区域的内部，一个预测拼接区域的边界，在训练时联合优化两个分支。为提高定位性能，该方法中还提出一种边界增强的决策手段，即将边界预测分支所得结果阈值化后进行填充，然后与另一分支的阈值化结果取交集，形成最终的篡改定位结果。该方法使用 CASIA v2 数据集训练，所得模型在 Columbia、CAISA v1、NIST NC 16、DSO-1 等数据集中的定位性能均优于文献[39]中测评的方法。Liu 和 Pun[95]综合利用了文献[58]中提出的三个不同尺度的 FCN，使用条件随机场来对它们的输出结果进行融合和优化。Kwon 等人[96]提出了一个包含 RGB 和 DCT 双分支的全卷积网络来进行拼接定位，其中的 RGB 分支以图像的 RGB 像素值作为输入，而 DCT 分支则以图像 Y 通道的 DCT 系数和相应的 JPEG 量化表作为输入，用于学习 JPEG 压缩痕迹。两个分支都使用了基于 HRNet[97]的网络结构，通过并联结构来保持高分辨率表征，

并形成多分辨率的特征表示,从而能更好地提取不同形状和尺度的拼接物体的特征。实验结果表明该方法对JPEG图像和非JPEG图像均有良好的篡改定位性能。Rao等人[98]设计了一个包含8个残差单元[76]的全卷积网络,在每两个残差单元之后加入基于条件随机场的注意力模块对学习到的特征进行加权,并在后四个残差单元中使用空洞卷积来扩大感受野,最后使用空洞卷积空间金字塔池化[99](Atrous Spatial Pyramid Pooling, ASPP)来得到篡改定位的结果。该方法使用IEEE IFS-TC数据集训练网络,所得模型在多个数据集中测试得到的F1-score、MCC等指标都好于已有方法。Zhuang等人[100]利用密集连接单元[90]构造了一个用于篡改定位的全卷积网络,其中密集连接单元在编码和解码网络中均使用,并也加入了空洞卷积以提高定位性能。该方法通过Photoshop脚本程序产生大量篡改图像用于训练模型,对书籍封面和自然场景这两类不同内容的图像都有较好的篡改定位效果。

作为一种典型且性能优良的全卷积网络,U型网络[101](U-Net)在不少图像篡改定位工作中得到了应用。在U-Net中,编码网络和解码网络有着几乎对称的结构,编码网络产生的不同层级的特征图通过跳跃连接与解码网络得到的具有相同分辨率的特征图相结合,因此网络能更有效地利用不同尺度的细节信息来形成定位结果。Bi等人[102]提出一个用于拼接篡改检测的U型网络,其编码网络和解码网络都由若干个环形残差单元构成。在环形残差单元中,包含了正向残差传播和反向残差反馈两种连接,前者类似人脑的召回机制,有助于解决深度网络的梯度退化问题;后者则类似人脑的巩固机制,能够增强原始和篡改区域在特征表达上的差异。该网络能更好地利用图像的上下文信息以减少预测错误,在使用同一数据集训练和测试的条件下,能在Columbia和CASIA v2中分别得到0.915和0.841的F1-score。为设计一个不局限于应对拼接篡改的方法,Chen等人[103]也在网络中使用了和文献[102]中类似的双向残差结构来增强提取的特征,同时在编码网络部分加入一个基于LSTM的结构来对重采样特征建模,将两部分特征一起送入解码网络得到篡改定位结果。通过6.5万张合成篡改图像训练该网络并用NIST NC 16中70%的数据进行微调,该方法在NIST NC 16剩余的30%数据中取得了超过0.91的F1-score。Yin等人[104]设计了一个与U-Net类似的编解码网络框架来进行篡改定位,除了通过解码网络预测篡改区域外,还使用块预测模块对编码网络不同层级的特征图进行多尺度的逐块分类,并设计混合的损失函数来指导整个网络的优化。Zhang和Ni[105]提出了一个基于密集连接U型网络的篡改定位方法。所设计的网络不仅在编码和解码部分使用了密集卷积块,而且把U-Net中编码和解码网络之间的跳跃连接也密集化,加入了将编码部分较低分辨率的特征图与解码部分较高分辨率的特征图相结合的连接路径。该方法在Columbia数据集中得到的F1-score达到0.9307。Shi等人[106]认为语义的差异导致在网络中利用相加或串联等简单的方式进行特征融合效果不佳。为此,作者在编码网络中使用格拉姆矩阵处理每一层的特征图,以使得编码特征中包含更丰富的全局纹理信息;在解码阶段,使用双向卷积的LSTM来将编码网络不同层级的特征与解码网络相连,而不使用U-Net中的简单连接模式,从而使得相同层级的特征图具有语义一致性。该方法在NIST NC 16数据集中得到的AUC和F1-score分别为0.917和0.837。为了避免网络中的池化操作造成的信息丢失,Bi等人[107]提出使用小波分解来代替传统的池化操作。该方法根据篡改区域内部预测、篡改区域边界预测、图像内容重构等目标设计了多任务学习框架,对不同任务使用不同的解码网络,并相应地使用不同小波子带进行反池化,有效地改善了拼接篡改定位的性能。

在一些工作中,全卷积网络也被用于识别和定位经过图像修复(Inpainting)篡改的区域。Zhu等人[108]提出了一个15层的全卷积网络来定位被基于块的图像修复算法[109]篡改的区域。Li和Huang[110]设计了一个高通全卷积网络以定位被基于深度学习的修复算法[111]篡改的区域,其中包括预滤波、特征提取、上采样三个模块。网络的预滤波模块以一阶差分滤波器进行初始化,用于增强修复留下的痕迹;特征提取网络模块主要通过残差单元构造;上采样模块则通过转置卷积来获得逐像素的篡改定位结果。由于图像中的被修复区域通常较小,以上两个工作在训练网络时都特别使用了带

权重的损失函数，其中文献[108]根据图像中篡改像素和原始像素的比例对交叉熵损失进行加权，而文献[110]则应用 Focal 损失函数[112]来减少数量多且易分类的类别在总体损失中的权重。上述针对图像修复的篡改定位方法对修复算法比较敏感，当遇到不包含在训练数据中的修复算法时，性能明显降低。Wu 和 Zhou[113]对此进行了改进：同时使用普通卷积层、受约束的卷积层[114]、以高通滤波器初始化的卷积层对修复痕迹进行增强，利用网络架构搜索(Neural Architecture Search，NAS)自适应地确定特征提取模块的结构，将注意力机制加入决策模块以提高预测的准确性。该方法对多种图像修复算法都取得了良好的效果，表现出较好的泛化能力。

4.4　基于 R-CNN 的方法

R-CNN 系列网络是目标检测领域的主流技术之一。图 4 给出了典型的 Mask R-CNN[60]的基本框架。其中使用一个卷积网络作为主干网络以提取输入图像的特征，然后通过区域生成网络(Region Proposal Network，RPN)产生候选区域(proposals)，接着利用 ROI Align 来提取各个 proposal 的特征，将特征送入后续的网络以对 proposal 的类型(class)和边界框(bbox)进行预测。以上过程和 Faster R-CNN[59]是基本一致的，在此基础上，Mask R-CNN 还将 proposal 的特征送入一个全卷积网络，得到像素级别的预测(mask)。可见基于 R-CNN 的处理框架也可以应用于篡改定位任务。而且，由于 R-CNN 具有多任务输出的特性，其在定位篡改区域的同时，可以利用 class 输出端来预测篡改区域所属的篡改类型。

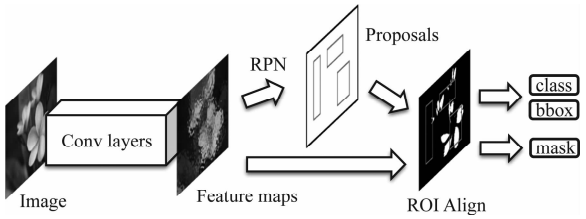


图 4　Mask R-CNN 的基本框架

Fig. 4　Basic framework of Mask R-CNN

Zhou 等人[29]基于 Faster R-CNN 提出了一个双输入流的篡改检测方法。该方法将图像的 RGB 像素值和经过 SRM 滤波器得到的残差噪声分别输入两个特征提取网络，使用 RPN 从 RGB 分支得到的特征中生成候选区域，并根据这些候选区域分别在 RGB 分支和噪声分支的特征图中提取特征，利用双线性池化融合两种特征，送入分类网络以判断候选区域的篡改类型，同时仅使用候选区域的 RGB 特征来预测篡改区域的边界框。为了得到充足的训练样本，作者利用 MS COCO[49]数据集生成了 42000 张篡改图像。由于该方法基于 Faster R-CNN 网络，其只能给出篡改区域的边框，而不能输出更精确的像素级预测结果，其在 NIST NC 16 和 Coverage 数据集中获得的 F1-score 分别为 0.722 和 0.437。Zhang 等人[115]将上述方法与基于 CNN 的篡改检测方法[84,94]相结合，利用基于 CNN 方法的预测结果来生成额外的候选框，加入基于 Faster R-CNN 的检测框架中以得到改进的检测边框，然后把基于 Faster R-CNN 方法的输出和基于 CNN 方法的预测结果相融合以减少虚警，最后通过稠密条件随机场来优化定位结果。由于结合了两类方法的优势，其定位性能相比文献[29]有所提高。为利用 R-CNN 框架得到像素级的预测结果，Ahmed 等人[116]将 Mask R-CNN 应用于拼接检测，其中对基于 ResNet-FPN[117]的主干网络进行了简化，采用单独的卷积层代替了相对复杂的特征金字塔网络(Feature Pyramid Network，FPN)。Yang 等人[118]利用以 ResNet-101 为主干网络的 Mask R-CNN 框架设计了一个通用的篡改检测与定位方法，其中图像先经过约束卷积层后才被送入特征提取网络。这与文献[29]采用的双流输入相比，减少了网络的参数量和复杂程度。为提高篡改检测和定位的能力，该方法在 RPN 网络中加入 CBAM 注意力模块[119]，通过跳跃连接融合低层和高层特征，并利用关于边界框的预测结果来指导像素级预测分支的学习。与文献[29]相比，该方法显著提升了定位的准确性，在 NIST NC 16 和 Coverage 数据集中获得的 F1-score 分别为 0.927 和 0.757。和文献[29]及[118]中的做法均不同的是，文献[120]中依然使用 RGB 和噪声两个分支，但并不对两个分支产生的特征进行融合，而是利用注意力机制通过噪声分支来指导 RGB 分支的训练。该方法在 NIST NC 16 和 Coverage 数据集中分别得到 0.930 和 0.764 的 F1-score，性能相较文献[118]有一定提高。

利用 R-CNN 框架也可以设计针对图像修复的检测和定位方法。Wang 等人[121]首先使用以 Res-

Net-101 为主干网络的 Faster R-CNN 模型对基于深度学习的两种图像修复算法[122-123]进行了检测。他们的另一个工作[124]改用 Mask R-CNN 模型,将图像像素和 LBP 特征图同时作为网络的输入,并对 RPN 网络改进以提取多尺度的特征金字塔。该方法被用于对更多种图像修复算法进行检测,结果验证了其有效性。

4.5 基于孪生网络的方法

孪生网络是由两个结构相同且权重共享的网络组合而成的一种网络结构,它通常用于比较两个样本的相似程度。具体地,孪生网络接收一对样本作为输入,其中的两个子网络各自处理一个样本,输出样本在高维空间的嵌入表征。通过计算两个表征的距离,就可以得到两个样本的相似程度。孪生网络具有的这种"相似性度量"的特点,使之在一些篡改取证问题中得以派上用场。

Huh 等人[125]利用孪生网络来预测两个图像块是否对应于一致的元数据属性。该网络使用带有 EXIF 头文件的原始图像通过自监督的方式进行训练。训练完成后,模型能对给定的图像输出一个"一致性"图,篡改区域和原始区域在该图中会产生不同的响应。该方法在 Columbia 数据集中表现较好,所得 F1-score 为 0.88,而对于篡改情形更复杂的 DSO-1 数据集,其得到的 F1-score 为 0.52。与之类似的是,Mayer 和 Stamm[126]提出利用孪生网络来比较两个图像块的"取证相似性",即两图像块是否呈现出相同或相异的取证痕迹。该方法首先用两个权值共享的约束卷积网络[114]提取一对图像块的特征,然后将特征送入另一个网络来判断两者是否相似。当图像经过拼接等操作的篡改时,篡改区域和原始区域的特征将表现出较低的相似性,故此方法能够定位出篡改区域。该方法的优点是,即使某些取证痕迹在训练集中没有出现,训练好的模型也能正确地做出处置,因此更适应于开放的取证场合。文献[127]通过上述的"取证相似性"对图像中的各小块构建一个关系图。由于取证痕迹的差异,篡改区域和原始区域会在关系图中形成不同的群体,通过判断是否存在多个群体并对群体进行分割,就可以实现篡改检测和定位。该方法在 Columbia 和 DSO-1 数据集中取得的 F1-score 达到 0.89 和 0.82。

在传统的篡改检测和定位方法中,一类重要的方法是利用相机在图像中留下的传感器模式噪声[128]来进行决策的。不同相机的模式噪声各不相同,故可认为模式噪声是相机的一种指纹信息。由相同相机拍摄的图像具有相同的指纹,而不同相机拍摄的图像则具有不同的指纹;当图像经过篡改后,篡改区域的指纹信息也会受到破坏。Cozzolino 和 Verdoliva[129]提出了一种利用孪生网络来提取图像的噪声指纹(noiseprint)的方法。该方法将成对的图像块输入一个孪生网络,网络提取出图像的噪声信息,通过计算两噪声的距离来预测两图像块是否来自于同一相机模型,根据预测结果与实际结果之间的差异损失来更新孪生网络中的参数。训练完成后,网络就具备了从图像中提取与相机模型相关的噪声指纹的能力。文中还提出了一种不需额外辅助信息的篡改定位方法:对于给定的图像,首先通过训练好的网络提取噪声指纹,然后利用 EM 算法来将像素分成两类,得到篡改定位结果。该方法在 9 个数据集中进行了测试并与 15 种取证方法进行了比较,结果表明其定位性能优于多种传统方法,且在大多数情况下也比文献[125]提出的方法要更好。此外,文献[130]提出了一种在有参考信息条件下基于噪声指纹的篡改定位方法。与文献[129]不同的是,在此情况下可利用网络提取一批参考图像的噪声指纹,对它们求平均得到相机的噪声指纹模板。对于测试图像,将其噪声指纹与相应的参考模板进行匹配,相似程度低的区域即被认为是篡改区域。

Barni 等人[131]认为复制移动篡改中的源区域和目标区域在插值痕迹和边界处的连贯性上具有差异。基于此,他们设计了一个双分支的网络框架,分别对插值痕迹和边界处的痕迹进行分析,其中两个网络分支都采用了孪生网络的结构。通过融合两个分支的结果,可以区分源区域和目标区域。该方法可以应用于任何复制移动检测算法产生的结果,以消除歧义性。

在文献[132]中,Wu 等人尝试解决约束条件下的拼接检测问题。与普通拼接检测不同的是,此问题中除了待测图像外还给定一张疑似的供体图像,需要判断待测图像中是否有从疑似的供体图像拼接而来的区域。所提出的方法使用基于 VGG16 的权值共享的网络分别提取两张图像的特征,然后计

算特征之间的相关性并提取相似图像区域，然后将结果送入具有孪生网络结构的视觉一致性检测器，进一步判断是否确实发生了拼接操作，从而确定篡改的位置。

4.6　结合 LSTM 的方法

长短期记忆（LSTM）网络是循环神经网络的一种，它擅长于解决序列数据中的长期依赖问题，在长序列数据中往往有更好的表现。在图像篡改定位中，LSTM 通常与 CNN 结合起来使用。此时，图像或其特征图会被分割成小块，形成序列数据，从而可通过 LSTM 来对小块之间的关系建模，这有助于获得可区分原始区域和篡改区域的有效特征。

Bunk 等人[74]首先将 LSTM 用于图像的篡改取证。他们把图像块的重采样特征输入一个 3 层 LSTM，得到 256 维的特征向量，使用该特征来判断图像块是否经过篡改。Bappy 等人[133]则先利用两层卷积网络对图像块提取特征，将所得特征输入 LSTM 来得到块级别的分类结果；同时，LSTM 最后一层输出的特征被重组成二维的特征图，再经过 3 层全卷积网络得到像素级别的定位结果。该方法使用 NIST NC 16 中 75% 的图像来训练模型，并将剩余的 25% 图像用于测试，得到块级别分类的精度为 89.38%，像素级别定位的 AUC 为 0.7641。在进一步的工作中，Bappy 等人[7]设计了一个混合 LSTM 和 CNN 编解码网络的框架。该框架包含两个特征提取分支，其一使用一个 2 层 LSTM 对图像块的重采样特征进行建模，另一使用由 4 个残差网络块构成的 CNN 提取图像的空域特征。随后，两个分支得到的特征被串联起来，送入解码网络得到篡改定位结果。为了构建充足的训练数据，作者利用 MS COCO[49]、Deresden[52]、NIST NC16[45]等数据集自动生成了超过 6 万张篡改图像。该方法对 NIST NC 16 中的图像进行篡改定位的 AUC 值为 0.7936。为了更好地捕捉篡改区域边界附近的痕迹，Mazaheri 等人[134]在文献[7]的网络框架中引入跳跃连接，将编码网络中浅层的特征与编码网络最后一层的输出相融合，促使网络更准确地定位篡改区域，使得在 NIST NC 16 数据集中得到的 AUC 提高至 0.8570。

与上述方法不同的是，Wu 等人[8]将篡改定位视为一个局部异常检测问题，利用 LSTM 构建局部异常检测网络得到篡改定位的结果。具体地，该方法首先使用基于 VGG16 的网络提取特征，该特征提取网络被训练以区分 385 种不同方式和参数的图像操作，图像经特征提取网络得到的输出被转换为可以表征局部异常的特征，这些特征经过拼接组合形成具有 4 个维度的特征张量，再输入 LSTM 单元以得到关于异常区域（即篡改区域）的预测结果。在 Columbia、CAISA、NIST NC 16、Coverage 等常用的篡改图像数据集中，该方法不需要使用相应数据集中的图像对预训练模型微调，就能获得可与文献[29, 133]相当的结果。Hu 等人[135]也利用了上述的特征提取网络来获得图像特征，但将基于 LSTM 的局部异常检测网络替换为多层的自注意力结构[136]，这能更好地对图像块之间的关系建模，因此性能相比文献[8]有所提升。

4.7　结合 GAN 的方法

生成对抗网络（GAN）由 Goodfellow 等人[63]提出，其中包含两个网络：一个是生成器（Generator），一个是判别器（Discriminator）。GAN 的训练可以视为生成器和判别器之间的“零和博弈”：生成器尝试产生看起来自然真实的、和原始数据尽量相似的样本，而判别器则需要判断给定样本是真实的还是生成的。经过交替的对抗性学习的优化，生成器和判别器的性能都可得到提升。在理想情况下，生成器最终能再现真实数据的分布，而此时判别器则无法区分真实样本和生成样本。GAN 在图像篡改定位中的应用主要有两方面：一是将产生篡改定位结果看作生成问题，通过加入判别器进行对抗训练，优化生成的结果；二是利用 GAN 生成“困难样本”，进行数据增强，从而强化模型的训练效果。

Bartusiak 等人[137]提出利用条件 GAN（Conditional GAN）来检测和定位卫星图像中的拼接区域。在该 GAN 模型中，生成器由一个 16 层的 U-Net 构成，其作用是预测给定图像的篡改区域掩模；判别器由一个 5 层的 CNN 分类网络构成，它接收一对由图像和掩模组成的输入，并判断掩模是该图像的实际掩模还是由生成网络预测的掩模。通过对抗性学习使预测掩模能够欺骗判别器，并通过损失函数对预测掩模与实际掩模的相似度进行约束，从而使生成器能正确地预测图像中的篡改区域。Li 等人[138]和 Islam 等人[139]将条件 GAN 用于复制移动篡改检测和定位中。除 GAN 网络的对抗损失外，Li

等人[138]在训练时引入了基于 L_1 距离的损失和衡量篡改区域检测误差的损失,并使用适量的非篡改图像作为弱监督样本来提升性能。Islam 等人[139]在生成器中引入了通过特征图的亲和矩阵产生的一阶和二阶注意力,其中一阶注意力用于强化复制移动篡改的位置信息,二阶注意力则表征了图像块之间的共生关系。通过这两种注意力对特征进行融合,能够提升模型的检测和定位性能。Liu 等人[140]将对抗性学习的思想应用到约束拼接检测问题。该方法使用一个匹配网络来预测拼接区域在两图中的位置。具体地,使用带空洞卷积的网络提取待测图像和疑似供体图像的特征,通过比较特征相关性来生成预测掩模。为了提高性能,在训练阶段加入了检测网络和判别网络来对预测结果进行约束:检测网络用于判断预测的拼接区域内容是否相关,判别网络则用于区分输入的掩模是由匹配网络生成的还是真实的。这两个网络都通过对抗性训练来辅助更新匹配网络中的参数,使之得到更准确的预测结果。实验结果表明,该方法获得了比文献[132]中的方法更好的性能。

Kniaz 等人[50]提出了一个用于拼接检测及定位的混合对抗生成框架。该框架包含两个生成器和两个判别器。其中一个生成器称为修图生成器,其作用是对输入的拼接图像进行润饰,以尽量消除拼接产生的可被检测的痕迹;另一个生成器称为标注生成器,用于对输入的图像预测像素级的篡改与否的标签;两个判别器则分别就生成的润饰图像和预测的篡改标签是否与实际一致进行判断。这四个网络在对抗性学习的框架下同时训练,最终目标是使得标注生成器具有良好的篡改定位性能。该方法使用 FantasticReality 中的"粗糙"拼接图像训练模型,在 CASIA v2.0、Columbia、DSO-1、FantasticReality 等数据集中进行测试,所得结果在 IoU 等性能指标上要好于[8]、[28]等文献中提出的方法。Zhou 等人[141]也将 GAN 用于生成篡改图像以辅助篡改定位模型的训练,所提出的方法包括生成、分割、优化三个阶段。在生成阶段,先将一张篡改图像中的篡改区域简单地复制到一张原始图像中,得到一个简单样本,然后使用 GAN 网络对简单样本的篡改区域和背景进行混合,生成困难样本。在分割阶段,使用基于 DeepLab-VGG16[99]的网络模型预测给定图像的篡改区域和篡改区域边界,上一阶段中的篡改图像及由其得到的两种样本都作为此阶段的输入。在优化阶段,把在困难样本中被预测为篡改区域边界的像素用原始图像中的相应像素代替,形成新的样本,并再次送入分割网络进行训练。通过交替地执行各阶段的训练过程,可使得分割网络具备泛化性较强的篡改定位能力。该方法使用 CASIA v2 数据集对在 ImageNet 中预训练的模型微调,在 CASIA v1、DSO-1、Coverage 等数据集中都取得了较好的结果。

4.8 篡改定位方法性能总结

为了综合比较上述方法的篡改定位性能,我们将部分代表性方法在若干公开数据集中的性能表现总结如表 2 所示。该表中所有数据均由各文献报道的实验数据整理而得,限于版面篇幅,我们仅给出了 AUC 和 F1-score 两个性能指标的结果。表中的方法根据训练策略的不同进行排序,相同训练策略的方法大致按照在上文出现的先后顺序列出。表中还加入了在现有文献中常用于比较的部分传统方法的实验结果,列在表格前 6 行。此外,由于复制移动篡改检测方法的训练和测试数据和其他方法有较大差异,它们被统一列在最后 5 行。通过对表 2 的分析可得到以下结论:

1)基于深度学习的方法的性能通常显著优于传统方法。可以看到,大多数基于深度学习的方法在各个数据集中取得的性能指标数值都要明显高于传统方法,这很大程度上得益于深度学习为篡改定位任务提供了更优异的基础技术框架。

2)算法跨数据集测试的性能有待提高。通过对比各方法在不同训练策略下得到的结果可以发现,在同一数据集中训练测试,以及先用外部数据训练再在同一数据集中微调并测试,在多数情况下均能得到比仅用外部数据训练更好的结果。这表明现有方法的鲁棒性、泛化性仍不够理想,需要进一步对算法进行优化。同时,这也意味着在搜集构造用于训练的外部数据时,需充分考虑实际测试场景,使外部数据中的篡改形式尽可能和实际相符。

3)评价指标可能会导致性能比较上的偏见。通过比较 HFSRNet、Yin2021Hybrid、GSCNet、RGB-N、Constrained R-CNN、MM-Net 等方法在 NIST NC 16 中得到的结果可以发现,它们的 AUC 都在 0.9 以

表 2　部分代表性方法的篡改定位性能*

Tab. 2　Tampering localization performance of some typical methods

方法名称	发表年份	网络架构	目标操作	训练策略	Columbia		CASIA v1		CASIA v2	NIST NC 16		DSO-1	Coverage		CoMoFoD
					AUC	F1	AUC	F1	F1	AUC	F1	F1	AUC	F1	F1
DCT[41]	2007	–	Mul	–	–	0.488	–	0.301	0.516	–	0.262	0.328	–	–	–
ADQ1[42]	2009	–	Mul	–	–	0.502	–	0.295	0.502	0.568	0.235	0.332	–	–	–
NADQ[23]	2012	–	Mul	–	–	0.450	–	0.176	0.175	0.613	0.197	0.247	–	–	–
ELA[40]	2007	–	Mul	–	0.581	0.471	0.613	0.214	–	0.429	0.231	0.280	0.583	0.222	–
CFA1[24]	2012	–	Mul	–	0.720	0.477	0.522	0.201	0.135	0.501	0.179	0.294	0.485	0.197	–
NOI1[43]	2009	–	Mul	–	0.546	0.574	0.612	0.263	–	0.487	0.282	0.366	0.587	0.269	–
Zhang2018Semi[65]	2018	AE	Mul	T	–	–	–	–	0.584	–	–	–	–	–	–
C2RNet[85]	2020	CNN	SP	T	–	0.695	–	–	0.676	–	–	–	–	–	–
RRU-Net[102]	2019	FCN	SP	T	–	0.915	–	–	0.841	–	–	–	–	–	–
DU-DC-EC Net[105]	2020	FCN	Mul	T	–	0.931	–	0.572	–	–	0.524	–	–	–	–
MWC-Net[107]	2021	FCN	SP	T	–	0.828	–	–	0.834	–	0.638	–	–	–	–
J-Conv-LSTM-Conv[133]	2017	CNN+LSTM	Mul	T	–	–	–	–	–	0.764	–	–	0.614	–	–
HFSRNet[103]	2021	FCN	Mul	P+F	–	–	–	0.467	–	0.954	0.918	–	0.782	0.624	–
Yin2021Hybrid[104]	2021	FCN	Mul	P+F	0.942	0.902	0.858	0.548	–	0.927	0.756	–	0.813	0.435	–
GSCNet[106]	2020	FCN	Mul	P+F	–	–	0.833	0.471	–	0.917	0.837	–	–	–	–
RGB-N[29]	2018	Faster R-CNN	Mul	P+F	0.858	0.697	0.795	0.408	–	0.937	0.722	–	0.817	0.437	–
HybridArch * {1+3}[115]	2021	R-CNN+CNN	Mul	P+F	–	0.755	–	–	0.525	–	0.725	–	–	0.589	–
Constrained R-CNN[118]	2020	Mask R-CNN	Mul	P+F	0.861	0.790	0.789	0.475	–	0.992	0.927	–	0.939	0.757	–
MM-Net[120]	2021	Mask R-CNN	Mul	P+F	0.897	0.823	–	–	–	0.983	0.930	–	0.913	0.764	–
LSTM-EnDec[7]	2019	CNN+LSTM	Mul	P+F	–	–	–	–	–	0.794	–	–	0.712	–	–
LSTM-EnDec-Skip[134]	2019	CNN+LSTM	Mul	P+F	–	–	0.814	0.432	–	0.857	–	–	–	–	–
C_ISRM_C-CNN[71]	2020	CNN	SP	P	–	–	–	–	–	–	–	0.581	–	–	–
D-CNNs-bi[84]	2018	CNN	Mul	P	–	0.690	–	–	–	–	–	0.580	–	–	–
MFCN[28]	2018	FCN	SP	P	–	0.612	–	0.541	–	–	0.571	0.480	–	–	–
Rao2021Multi[98]	2021	FCN	Mul	P	–	–	–	0.259	0.315	–	0.348	0.804	–	0.636	–
Dense-FCN[100]	2021	FCN	Mul	P	–	–	–	–	–	0.750	0.310	–	–	–	–
EXIF-Consistency[125]	2018	Siamese Net	Mul	P	–	0.880	–	–	–	–	–	0.520	–	–	–
Forensic simi graph[127]	2020	Siamese Net	Mul	P	0.930	0.890	–	–	–	–	–	0.820	–	–	–
Noiseprint[129]	2020	Siamese Net	Mul	P	–	–	–	–	–	–	0.395	0.780	–	–	–
ManTra-Net[8]	2019	CNN+LSTM	Mul	P	0.824	–	0.817	–	–	0.795	–	–	0.819	–	–
SPAN[135]	2020	CNN+SelfAttn	Mul	P	0.936	0.815	0.814	0.336	–	0.836	0.290	–	0.912	0.535	–
GSR-Net[141]	2020	GAN	Mul	P	–	–	–	0.574	–	–	–	0.525	–	0.489	–
BusterNet[92]	2018	CNN	CM	P	–	–	–	–	0.456†	–	–	–	–	–	0.493
Dense-InceptionNet[89]	2020	CNN	CM	P	–	–	–	–	0.643†	–	–	–	–	–	0.441
AR-Net[91]	2020	CNN	CM	P	–	–	–	–	0.455†	–	–	–	0.849	–	0.501
CMSDNet+STRDNet[93]	2020	CNN	CM	P	–	–	–	–	0.538†	–	–	–	–	0.677	0.511
DOA-GAN[139]	2020	GAN	CM	P	–	–	–	–	0.414†	–	–	–	–	–	0.369

注：* 表中“目标操作”列中的 SP 表示拼接，CM 表示复制移动，Mul 表示多种操作；“训练策略”列中的 P 表示使用不包含在数据集中的外部数据训练，T 表示使用同一数据集的图像训练，F 表示使用同一数据集的图像微调；文献中未报道的测试结果以“–”表示。

† 针对复制移动的篡改定位方法的结果是在 CASIA v2 数据集中的部分复制移动篡改图像上得到的。

上，差距不大，而相应的 F1-score 则在 0.72 至 0.93 之间波动，差异比较明显。可见选择不同性能评价指标会对这些方法的性能差异产生不同的结论，如何公平合理地评估算法性能是值得进一步探讨的问题。

4）不同网络架构并未引起性能上的显著差异。总体来看，上述介绍的方法都在一定程度上借鉴了计算机视觉领域中为语义分割、目标检测所设计的网络架构。综合比较这些方法的定位性能，我们并未发现有某种网络架构表现出显著超越其他网络架构的性能。这一方面表明将计算机视觉中的各类网络架构应用于篡改定位是可行的，随着计算机视觉的发展，将有更多性能出色的技术方案能为我们所用。但另一方面也意味着纯粹借鉴计算机视觉中的解决方案可能会遇到性能瓶颈，只有深入分析篡改定位的需求和特点，更合理地优化改进网络架构，才能形成新的突破。

5　机遇与挑战

深度学习范式已推动图像篡改定位技术迈上

一个新的台阶,但这并不意味着取证者已在这场持续的“猫和老鼠”的博弈中处于上风。事实上,随着人工智能技术的迅速发展,篡改伪造的手段也在日益更新,这给取证带来了更多新的挑战。另一方面,我们也注意到数字媒体取证的重要性已得到学术界甚至工业界越来越多的认同,不断有新的力量投入到图像篡改取证之中。可以预见,数字图像取证在未来很长一段时间内依然会处于挑战和机遇并存的环境中。在图像篡改定位这一重要的应用方向上,以下一些问题值得继续深入研究:

1)基准数据集的构建:在深度学习技术被大量地应用于图像篡改定位任务的环境下,篡改图像数据集的缺乏已成为制约篡改定位技术发展的瓶颈之一。在现有的篡改定位工作中,依然大量使用Columbia、CASIA、NIST NC 16等相对陈旧的数据集来进行性能验证。由于这些数据集中图像的采集、篡改、后处理过程并非完全透明,且可能和实际的篡改情形有较大偏差,这会导致对算法的性能产生过于乐观的估计。因此,如何构建满足实际取证需求的基准数据集必须引起重视。我们认为一个良好的数据集应能客观地反映现实中的篡改情形。这意味着数据集中需涉及足够数量的拍摄设备,涵盖常见的图像篡改方式,引入多种类的后处理操作,且这些相关信息能在一定程度上追溯以便分析不同因素对算法性能的影响。显然,构建这样的数据集并非一日之功。

2)性能评价体系的改进:3.2节介绍了篡改定位任务中常用的性能评价指标,但这些指标并不能全面地评价篡改定位方法的优劣。特别地,在一些实际场合中(如司法鉴定),篡改定位算法给出的结果是用于辅助人工进行最终决策的。此时,如果产生的虚警较多,将会大大降低工作效率,而即使对于篡改区域的预测标注不是十分准确,也不会对最终决策产生本质的影响。在这种情况下,上述性能指标的数值大小并不能够客观反映某个篡改定位算法是否足够实用。因此,如何根据实际取证场合的需求设计合理的性能评价体系,也是值得研究的问题。

3)算法泛化性和鲁棒性的提升:目前,基于深度学习的篡改定位模型的性能严重依赖于训练数据集。对于来自不同数据集的测试样本,其性能通常明显下降。这意味着需要更深入地分析不同来源的图像之间的内在联系,相应地改进网络架构以学习更有效的特征,从而提高方法的泛化能力。另一方面,当图像经受了JPEG压缩或尺寸缩放等处理后,模型性能也会显著降低。尽管可以通过在训练数据中引入后处理来进行数据增强,以减轻性能下降的程度,但这没有从根本上解决问题。如何提高算法对这些后处理操作的鲁棒性,是将算法推向实际应用的过程中必须考虑的。同时,可以预计这些后处理达到一定强度时,篡改区域和原始区域的差异已被完全破坏,导致无法定位篡改区域。是否存在这样的性能下界以及如何确定它们,也是有意义的研究方向。此外,深度神经网络本身还容易遭到对抗样本的攻击,这意味着在设计算法时还需顾及到如何对此种攻击进行抵抗。

4)模型可解释性的突破:与深度学习在其他领域中遇到的问题类似,深度学习模型可解释性弱的特点限制了它在篡改定位中的实用与推广。在实际的篡改取证场合,仅给出关于篡改区域的预测结果是不够的,往往还需要得知模型是根据什么线索而给出这样的结果。但现有的基于深度学习的篡改定位方法都无法很好地回答这一问题,这就导致人们无法给予模型足够的信任。因此,为了促进篡改定位模型在实际中的部署,在关注提高模型的篡改定位性能的同时,还必须考虑如何在模型的可解释性方面产生突破。目前,深度模型可解释性研究在图像分类领域已形成一定规模,借鉴相关的研究成果是一种可行的办法。另外,结合取证分析运作模式的自身特点,设计具有可解释性的深度网络结构,也是一种值得探索的途径。

6 结论

本文归纳总结了基于深度学习的图像篡改定位方法。特别地,我们按照这些方法所使用的网络架构对它们进行梳理。可以看到,不同的网络架构有着各自的特点和优势,为面向不同具体问题的篡改定位方法的设计提供了多样化的选择。深度学习技术仍在持续发展,这给图像篡改定位带来了大量的挑战和机遇:我们既需要对付不断更新升级的图像篡改技术,也可以利用更加有效的网络架构或学习策略来提升算法的性能。本文还介绍了图像篡改定位中常用的数据集和性能评价指标,并讨论了该领域目前

存在的问题和一些可能的研究方向。这有助于读者全面把握图像篡改定位领域的研究动向。

从更宏观的角度来看，遏制图像篡改伪造并非纯粹的技术问题，通过法律法规来对篡改伪造行为进行约束和震慑，同样是防止篡改图像导致不必要损失的有效途径，但这已超出本文的讨论范围。展望未来，我们仍将在与图像篡改伪造相互斗争的漫漫长路中求索，不断充实数字图像取证的技术装备库，为多媒体信息安全保驾护航。

参考文献

[1] FARID H. Image forgery detection[J]. IEEE Signal Processing Magazine, 2009, 26(2): 16-25.

[2] STAMM M C, WU Min, LIU K J R. Information forensics: An overview of the first decade[J]. IEEE Access, 2013, 1: 167-200.

[3] 杨锐, 骆伟祺, 黄继武. 多媒体取证[J]. 中国科学: 信息科学, 2013, 43(12): 1654-1672.
YANG Rui, LUO Weiqi, HUANG Jiwu. Multimedia forensics[J]. Scientia Sinica (Informationis), 2013, 43(12): 1654-1672. (in Chinese)

[4] KORUS P. Digital image integrity-a survey of protection and verification techniques[J]. Digital Signal Processing, 2017, 71: 1-26.

[5] VERDOLIVA L. Media forensics and DeepFakes: An overview[J]. IEEE Journal of Selected Topics in Signal Processing, 2020, 14(5): 910-932.

[6] CASTILLO CAMACHO I, WANG Kai. A comprehensive review of deep-learning-based methods for image forensics[J]. Journal of Imaging, 2021, 7(4): 69.

[7] BAPPY J H, SIMONS C, NATARAJ L, et al. Hybrid LSTM and encoder-decoder architecture for detection of image forgeries[J]. IEEE Transactions on Image Processing, 2019, 28(7): 3286-3300.

[8] WU Yue, ABDALMAGEED W, NATARAJAN P. Man-Tra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA. IEEE, 2019: 9535-9544.

[9] LI Haodong, LUO Weiqi, QIU Xiaoqing, et al. Identification of various image operations using residual-based features[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 28(1): 31-45.

[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] MAHDIAN B, SAIC S. Blind authentication using periodic properties of interpolation[J]. IEEE Transactions on Information Forensics and Security, 2008, 3(3): 529-538.

[12] KIRCHNER M, FRIDRICH J. On detection of median filtering in digital images[C]// Proceedings of the SPIE 7541, Media Forensics and Security II, 2010, 7541: 754110.

[13] CHEN Chenglong, NI Jiangqun, HUANG Jiwu. Blind detection of median filtering in digital images: A difference domain based approach[J]. IEEE Transactions on Image Processing, 2013, 22(12): 4699-4710.

[14] FAN Zhigang, DE QUEIROZ R L. Identification of bitmap compression history: JPEG detection and quantizer estimation[J]. IEEE Transactions on Image Processing, 2003, 12(2): 230-235.

[15] LI Bin, SHI Y Q, HUANG Jiwu. Detecting doubly compressed JPEG images by using mode based first digit features[C]// Proceedings of the 10th IEEE Workshop on Multimedia Signal Processing. Cairns, QLD, Australia. IEEE, 2008: 730-735.

[16] SHI Y Q, CHEN Chunhua, CHEN Wen. A natural image model approach to splicing detection[C]// Proceedings of the 9th Workshop on Multimedia and Security. 2007: 51-62.

[17] WEI Wang, DONG Jing, TAN Tieniu. Effective image splicing detection based on image chroma[C]// Proceedings of the 16th IEEE International Conference on Image Processing. Cairo, Egypt. IEEE, 2009: 1257-1260.

[18] FRIDRICH A J, SOUKAL B D, LUKÁŠ A J. Detection of copy-move forgery in digital images[C]// Proceedings of the Digital Forensic Research Workshop, 2003.

[19] LI Jian, LI Xiaolong, YANG Bin, et al. Segmentation-based image copy-move forgery detection scheme[J]. IEEE Transactions on Information Forensics and Security, 2015, 10(3): 507-518.

[20] 李子健, 阮秋琦. 基于 LPP 和改进 SIFT 的 copy-move 篡改检测[J]. 信号处理, 2017, 33(4): 589-594.
LI Zijian, RUAN Qiuqi. Copy-move forgery detection based on LPP and improved SIFT algorithm[J]. Journal of Signal Processing, 2017, 33(4): 589-594. (in Chinese)

[21] COZZOLINO D, GRAGNANIELLO D, VERDOLIVA L. Image forgery detection through residual-based local de-

scriptors and block-matching[C]// Proceedings of the IEEE International Conference on Image Processing. Paris, France. IEEE, 2014: 5297-5301.

[22] VERDOLIVA L, COZZOLINO D, POGGI G. A feature-based approach for image tampering detection and localization[C]// Proceedings of the IEEE International Workshop on Information Forensics and Security. Atlanta, GA, USA. IEEE, 2014: 149-154.

[23] BIANCHI T, PIVA A. Image forgery localization via block-grained analysis of JPEG artifacts[J]. IEEE Transactions on Information Forensics and Security, 2012, 7(3): 1003-1017.

[24] FERRARA P, BIANCHI T, DE ROSA A, et al. Image forgery localization via fine-grained analysis of CFA artifacts[J]. IEEE Transactions on Information Forensics and Security, 2012, 7(5): 1566-1577.

[25] CHIERCHIA G, POGGI G, SANSONE C, et al. A Bayesian-MRF approach for PRNU-based image forgery detection[J]. IEEE Transactions on Information Forensics and Security, 2014, 9(4): 554-567.

[26] FAN Wei, WANG Kai, CAYRE F. General-purpose image forensics using patch likelihood under image statistical models[C]// Proceedings of the IEEE International Workshop on Information Forensics and Security. Rome, Italy. IEEE, 2015: 1-6.

[27] KORUS P, HUANG Jiwu. Multi-scale fusion for improved localization of malicious tampering in digital images[J]. IEEE Transactions on Image Processing, 2016, 25(3): 1312-1326.

[28] SALLOUM R, REN Yuzhuo, KUO C C J. Image splicing localization using a multi-task fully convolutional network (MFCN)[J]. Journal of Visual Communication and Image Representation, 2018, 51: 201-209.

[29] ZHOU Peng, HAN Xintong, MORARIU V I, et al. Learning rich features for image manipulation detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. IEEE, 2018: 1053-1061.

[30] NG T, HSU J, CHANG S F. A data set of authentic and spliced image blocks[EB/OL]. http://www.ee.columbia.edu/ln/dvmm/downloads/AuthSplicedDataSet/AuthSplicedDataSet.htm. 2004.

[31] HSU Y F, CHANG S F. Detecting image splicing using geometry invariants and camera characteristics consistency[C]// Proceedings of the IEEE International Conference on Multimedia and Expo. Toronto, ON, Canada. IEEE, 2006: 549-552.

[32] DONG Jing, WANG Wei, TAN Tieniu. CASIA image tampering detection evaluation database[C]// Proceedings of the IEEE China Summit and International Conference on Signal and Information Processing. Beijing, China. IEEE, 2013: 422-426.

[33] CATTANEO G, ROSCIGNO G. A possible pitfall in the experimental analysis of tampering detection algorithms[C]// Proceedings of the 17th International Conference on Network-Based Information Systems. Salerno, Italy. IEEE, 2014: 279-286.

[34] IEEE IFS-TC Image Forensics Challenge Dataset[EB/OL]. http://ifc.recod.ic.unicamp.br/fc.website/index.py. 2014.

[35] DE CARVALHO T J, RIESS C, ANGELOPOULOU E, et al. Exposing digital image forgeries by illumination color classification[J]. IEEE Transactions on Information Forensics and Security, 2013, 8(7): 1182-1194.

[36] TRALIC D, ZUPANCIC I, GRGIC S, et al. CoMoFoD—New database for copy-move forgery detection[C]// Proceedings ELMAR-2013. Zadar, Croatia. IEEE, 2013: 49-54.

[37] WEN Bihan, ZHU Ye, SUBRAMANIAN R, et al. COVERAGE—A novel database for copy-move forgery detection[C]// Proceedings of the IEEE International Conference on Image Processing. Phoenix, AZ, USA. IEEE, 2016: 161-165.

[38] ZAMPOGLOU M, PAPADOPOULOS S, KOMPATSIARIS Y. Detecting image splicing in the wild (WEB)[C]// Proceedings of the IEEE International Conference on Multimedia and Expo Workshops. Turin, Italy. IEEE, 2015: 1-6.

[39] ZAMPOGLOU M, PAPADOPOULOS S, KOMPATSIARIS Y. Large-scale evaluation of splicing localization algorithms for web images[J]. Multimedia Tools and Applications, 2017, 76(4): 4801-4834.

[40] KRAWETZ N. A picture's worth[EB/OL]. http://www.hackerfactor.com/papers/bh-usa-07-krawetz-wp.pdf. 2007.

[41] YE Shuiming, SUN Qibin, CHANG E C. Detecting digital image forgeries by measuring inconsistencies of blocking artifact[C]// Proceedings of the IEEE International Conference on Multimedia and Expo. Beijing, China. IEEE, 2007: 12-15.

[42] LIN Zhouchen, HE Junfeng, TANG Xiaoou, et al. Fast, automatic and fine-grained tampered JPEG image detection via DCT coefficient analysis[J]. Pattern Recognition, 2009, 42(11): 2492-2501.

[43] MAHDIAN B, SAIC S. Using noise inconsistencies for blind image forensics[J]. Image and Vision Computing, 2009, 27(10): 1497-1503.

[44] KORUS P, HUANG Jiwu. Evaluation of random field models in multi-modal unsupervised tampering localization [C]// Proceedings of the IEEE International Workshop on Information Forensics and Security. Abu Dhabi, United Arab Emirates. IEEE, 2016: 1-6.

[45] GUAN Haiying, KOZAK M, ROBERTSON E, et al. MFC datasets: Large-scale benchmark datasets for media forensic challenge evaluation[C]// Proceedings of the IEEE Winter Conference on Applications of Computer Vision Workshops. Waikoloa, HI, USA. IEEE, 2019: 63-72.

[46] MFC2019[EB/OL]. https://www.nist.gov/itl/iad/mig/media-forensics-challenge-2019-0. 2019.

[47] HELLER S, ROSSETTO L, SCHULDT H. The PS-battles dataset -an image collection for image manipulation detection[EB/OL]. arXiv preprint arXiv:1804.04866, 2018.

[48] MAHFOUDI G, TAJINI B, RETRAINT F, et al. DEFACTO: image and face manipulation dataset[C]// Proceedings of the 27th European Signal Processing Conference. A Coruna, Spain. IEEE, 2019: 1-5.

[49] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context[M]// Computer Vision-ECCV 2014. Cham: Springer International Publishing, 2014: 740-755.

[50] KNIAZ V V, KNYAZ V, REMONDINO F. The point where reality meets fantasy: Mixed adversarial generators for image splice detection[C]// Advances in Neural Information Processing Systems. 2019, 32: 215-226.

[51] NOVOZÁMSKÝ A, MAHDIAN B, SAIC S. IMD2020: A large-scale annotated dataset tailored for detecting manipulated images[C]// Proceedings of the IEEE Winter Conference on Applications of Computer Vision Workshops. Snowmass, CO, USA. IEEE, 2020: 71-80.

[52] GLOE T, BÖHME R. The 'Dresden Image Database' for benchmarking digital image forensics[C]// Proceedings of the ACM Symposium on Applied Computing. Sierre, Switzerland. New York: ACM Press, 2010: 1584-1590.

[53] DENG Jia, DONG Wei, SOCHER R, et al. ImageNet: A large-scale hierarchical image database[C]// Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL, USA. IEEE, 2009: 248-255.

[54] ZHOU Bolei, LAPEDRIZA A, XIAO Jianxiong, et al. Learning deep features for scene recognition using places database[C]// Advances in Neural Information Processing Systems, 2014, 27: 487-495.

[55] XIAO Jianxiong, HAYS J, EHINGER K A, et al. SUN database: Large-scale scene recognition from abbey to zoo [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. San Francisco, CA, USA. IEEE, 2010: 3485-3492.

[56] KRAMER M A. Nonlinear principal component analysis using autoassociative neural networks[J]. AIChE Journal, 1991, 37(2): 233-243.

[57] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[C]// Advances in Neural Information Processing Systems, 2012, 25: 1097-1105.

[58] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. IEEE, 2015: 3431-3440.

[59] REN Shaoqing, HE Kaiming, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 39(6): 1137-1149.

[60] HE Kaiming, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]// Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy. IEEE, 2017: 2980-2988.

[61] CHOPRA S, HADSELL R, LECUN Y. Learning a similarity metric discriminatively, with application to face verification[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. San Diego, CA, USA. IEEE, 2005: 539-546.

[62] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[63] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]// Advances in Neural Information Processing Systems. 2014, 27: 2672-2680.

[64] ZHANG Ying, GOH J, WIN Leilei, et al. Image region forgery detection: A deep learning approach[C]// Pro-

ceedings of the Singapore Cyber-Security Conference, 2016: 1-11.

[65] ZHANG Ying, THING V L L. A semi-feature learning approach for tampered region localization across multi-format images [J]. Multimedia Tools and Applications, 2018, 77(19): 25027-25052.

[66] COZZOLINO D, VERDOLIVA L. Single-image splicing localization through autoencoder-based anomaly detection [C] // Proceedings of the IEEE International Workshop on Information Forensics and Security. Abu Dhabi, United Arab Emirates. IEEE, 2016: 1-6.

[67] FRIDRICH J, KODOVSKY J. Rich models for steganalysis of digital images[J]. IEEE Transactions on Information Forensics and Security, 2012, 7(3): 868-882.

[68] D'AVINO D, COZZOLINO D, POGGI G, et al. Autoencoder with recurrent neural networks for video forgery detection[J]. Electronic Imaging, 2017, 2017(7): 92-99.

[69] YARLAGADDA S K, GÜERA D, BESTAGINI P, et al. Satellite image forgery detection and localization using GAN and one-class classifier[J]. Electronic Imaging, 2018, 2018(7): 214-1.

[70] RAO Yuan, NI Jiangqun. A deep learning approach to detection of splicing and copy-move forgeries in images[C] // Proceedings of the IEEE International Workshop on Information Forensics and Security. Abu Dhabi, United Arab Emirates. IEEE, 2016: 1-6.

[71] RAO Yuan, NI Jiangqun, ZHAO Huimin. Deep learning local descriptor for image splicing detection and localization[J]. IEEE Access, 2020, 8: 25611-25625.

[72] COZZOLINO D, POGGI G, VERDOLIVA L. Recasting residual-based local descriptors as convolutional neural networks: An application to image forgery detection[C] // Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security. Philadelphia Pennsylvania USA. New York, NY, USA: ACM, 2017: 159-164.

[73] BONDI L, LAMERI S, GÜERA D, et al. Tampering detection and localization through clustering of camera-based CNN features[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu, HI, USA. IEEE, 2017: 1855-1864.

[74] BUNK J, BAPPY J H, MOHAMMED T M, et al. Detection and localization of image forgeries using resampling features and deep learning[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu, HI, USA. IEEE, 2017: 1881-1889.

[75] POMARI T, RUPPERT G, REZENDE E, et al. Image splicing detection through illumination inconsistencies and deep learning[C] // Proceedings of the 25th IEEE International Conference on Image Processing. Athens, Greece. IEEE, 2018: 3788-3792.

[76] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. IEEE, 2016: 770-778.

[77] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization [J]. International Journal of Computer Vision, 2020, 128(2): 336-359.

[78] COZZOLINO D, GRAGNANIELLO D, VERDOLIVA L. Image forgery localization through the fusion of camera-based, feature-based and pixel-based techniques [C] // Proceedings of the IEEE International Conference on Image Processing. Paris, France. IEEE, 2014: 5302-5306.

[79] LI Haodong, LUO Weiqi, QIU Xiaoqing, et al. Image forgery localization via integrating tampering possibility maps[J]. IEEE Transactions on Information Forensics and Security, 2017, 12(5): 1240-1252.

[80] ZHOU Peng, HAN Xintong, MORARIU V I, et al. Two-stream neural networks for tampered face detection[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu, HI, USA. IEEE, 2017: 1831-1839.

[81] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. IEEE, 2016: 2818-2826.

[82] SCHROFF F, KALENICHENKO D, PHILBIN J. FaceNet: A unified embedding for face recognition and clustering [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. IEEE, 2015: 815-823.

[83] GOLJAN M, FRIDRICH J. CFA-aware features for steganalysis of color images [C] // Proceedings of the SPIE 9409, Media Watermarking, Security, and Forensics 2015, 2015, 9409: 94090V.

[84] SHI Zenan, SHEN Xuanjing, KANG Hui, et al. Image manipulation detection and localization based on the dual-

domain convolutional neural networks[J]. IEEE Access, 2018, 6: 76437-76453.

[85] XIAO Bin, WEI Yang, BI Xiuli, et al. Image splicing forgery detection combining coarse to refined convolutional neural network and adaptive clustering[J]. Information Sciences, 2020, 511: 172-191.

[86] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [EB/OL]. arXiv preprint arXiv:1409.1556, 2014.

[87] OUYANG Junlin, LIU Yizhi, LIAO Miao. Copy-move forgery detection based on deep learning[C]// Proceedings of the 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI). Shanghai, China. IEEE, 2017: 1-5.

[88] WU Yue, ABD-ALMAGEED W, NATARAJAN P. Image copy-move forgery detection via an end-to-end deep neural network[C]// Proceedings of the IEEE Winter Conference on Applications of Computer Vision. Lake Tahoe, NV, USA. IEEE, 2018: 1907-1915.

[89] ZHONG Junliu, PUN C M. An end-to-end dense-InceptionNet for image copy-move forgery detection[J]. IEEE Transactions on Information Forensics and Security, 2020, 15: 2134-2146.

[90] HUANG Gao, LIU Zhuang, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. IEEE, 2017: 2261-2269.

[91] ZHU Ye, CHEN Chaofan, YAN Gang, et al. AR-net: Adaptive attention and residual refinement network for copy-move forgery detection[J]. IEEE Transactions on Industrial Informatics, 2020, 16(10): 6714-6723.

[92] WU Yue, ABD-ALMAGEED W, NATARAJAN P. Busternet: Detecting copy-move image forgery with source/target localization[C]// Proceedings of the European Conference on Computer Vision, 2018: 168-184.

[93] CHEN Beijing, TAN Weijin, COATRIEUX G, et al. A serial image copy-move forgery localization scheme with source/target distinguishment[J]. IEEE Transactions on Multimedia, 2020, PP(99): 1.

[94] LIU Yaqi, GUAN Qingxiao, ZHAO Xianfeng, et al. Image forgery localization based on multi-scale convolutional neural networks[C]// Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security, 2018: 85-90.

[95] LIU Bo, PUN C M. Locating splicing forgery by fully convolutional networks and conditional random field[J]. Signal Processing: Image Communication, 2018, 66: 103-112.

[96] KWON M J, YU I J, NAM S H, et al. CAT-Net: Compression artifact tracing network for detection and localization of image splicing[C]// Proceedings of the IEEE Winter Conference on Applications of Computer Vision. Virtual. IEEE, 2021: 375-384.

[97] WANG Jingdong, SUN Ke, CHENG Tianheng, et al. Deep high-resolution representation learning for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, PP(99):1.

[98] RAO Yuan, NI Jiangqun, XIE Hao. Multi-semantic CRF-based attention model for image forgery detection and localization[J]. Signal Processing, 2021, 183: 108051.

[99] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834-848.

[100] ZHUANG Peiyu, LI Haodong, TAN Shunquan, et al. Image tampering localization using a dense fully convolutional network[J]. IEEE Transactions on Information Forensics and Security, 2021, 16: 2986-2999.

[101] RONNEBERGER O, FISCHER P, BROX T. U-Net: Convolutional networks for biomedical image segmentation[C]// Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, 2015: 234-241.

[102] BI Xiuli, WEI Yang, XIAO Bin, et al. RRU-net: The ringed residual U-net for image splicing forgery detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Long Beach, CA, USA. IEEE, 2019: 30-39.

[103] CHEN Haipeng, CHANG Chaoqun, SHI Zenan, et al. Hybrid features and semantic reinforcement network for image forgery detection[J]. Multimedia Systems, 2021: 1-12.

[104] YIN Qilin, WANG Jinwei, LUO Xiangyang. A hybrid loss network for localization of image manipulation[M]// Digital Forensics and Watermarking. Cham: Springer International Publishing, 2021: 237-247.

[105] ZHANG Rongyu, NI Jiangqun. A dense U-net with cross-layer intersection for detection and localization of

image forgery[C]// Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Barcelona, Spain. IEEE, 2020: 2982-2986.

[106] SHI Zenan, SHEN Xuanjing, CHEN Haipeng, et al. Global semantic consistency network for image manipulation detection[J]. IEEE Signal Processing Letters, 2020, 27: 1755-1759.

[107] BI Xiuli, ZHANG Zhipeng, LIU Yanbin, et al. Multi-task wavelet corrected network for image splicing forgery detection and localization[C]// Proceedings of the IEEE International Conference on Multimedia and Expo. Shenzhen, China. IEEE, 2021: 1-6.

[108] ZHU Xinshan, QIAN Yongjun, ZHAO Xianfeng, et al. A deep learning approach to patch-based image inpainting forensics[J]. Signal Processing: Image Communication, 2018, 67: 90-99.

[109] CRIMINISI A, PEREZ P, TOYAMA K. Region filling and object removal by exemplar-based image inpainting [J]. IEEE Transactions on Image Processing, 2004, 13 (9): 1200-1212.

[110] LI Haodong, HUANG Jiwu. Localization of deep inpainting using high-pass fully convolutional network[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South). IEEE, 2019: 8300-8309.

[111] IIZUKA S, SIMO-SERRA E, ISHIKAWA H. Globally and locally consistent image completion [J]. ACM Transactions on Graphics, 2017, 36(4): 1-14.

[112] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]// Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy. IEEE, 2017: 2999-3007.

[113] WU Haiwei, ZHOU Jiantao. IID-net: Image inpainting detection network via neural architecture search and attention[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, PP(99): 1.

[114] BAYAR B, STAMM M C. Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection[J]. IEEE Transactions on Information Forensics and Security, 2018, 13 (11): 2691-2706.

[115] ZHANG Yixuan, ZHANG Jiguang, XU Shibiao. A hybrid convolutional architecture for accurate image manipulation localization at the pixel-level[J]. Multimedia Tools and Applications, 2021, 80:23377-23392.

[116] AHMED B, GULLIVER T A, ALZAHIR S. Image splicing detection using mask-RCNN[J]. Signal, Image and Video Processing, 2020, 14: 1035-1042.

[117] LIN T Y, DOLLAR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. IEEE, 2017: 2117-2125.

[118] YANG Chao, LI Huizhou, LIN Fangting, et al. Constrained R-CNN: A general image manipulation detection model[C]// Proceedings of the IEEE International Conference on Multimedia and Expo. London, UK. IEEE, 2020: 1-6.

[119] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module[M]// Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2018: 3-19.

[120] YANG Chao, WANG Zhiyu, SHEN Huawei, et al. Multi-modality image manipulation detection[C]// Proceedings of the IEEE International Conference on Multimedia and Expo. Shenzhen, China. IEEE, 2021: 1-6.

[121] WANG Xinyi, WANG He, NIU Shaozhang. An image forensic method for AI inpainting using faster R-CNN [M]// Artificial Intelligence and Security. Cham: Springer International Publishing, 2019: 476-487.

[122] PATHAK D, KRÄHENBÜHL P, DONAHUE J, et al. Context encoders: Feature learning by inpainting[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. IEEE, 2016: 2536-2544.

[123] YU Jiahui, LIN Zhe, YANG Jimei, et al. Generative image inpainting with contextual attention[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. IEEE, 2018: 5505-5514.

[124] WANG Xinyi, NIU Shaozhang, WANG He. Image inpainting detection based on multi-task deep learning network [J]. IETE Technical Review, 2021, 38(1): 149-157.

[125] HUH M, LIU A, OWENS A, et al. Fighting fake news: Image splice detection via learned self-consistency[C]// Proceedings of the European Conference on Computer Vision, 2018: 101-117.

[126] MAYER O, STAMM M C. Forensic similarity for digital images[J]. IEEE Transactions on Information Forensics and Security, 2020, 15: 1331-1346.

[127] MAYER O, STAMM M C. Exposing fake images with

forensic similarity graphs[J]. IEEE Journal of Selected Topics in Signal Processing, 2020, 14(5): 1049-1064.

[128] LUKÁŠ J, FRIDRICH J, GOLJAN M. Detecting digital image forgeries using sensor pattern noise[C]// Proceedings of the SPIE 6072, Security, Steganography, and Watermarking of Multimedia Contents VIII, 2006, 6072: 60720Y.

[129] COZZOLINO D, VERDOLIVA L. Noiseprint: A CNN-based camera model fingerprint[J]. IEEE Transactions on Information Forensics and Security, 2020, 15: 144-159.

[130] COZZOLINO D, VERDOLIVA L. Camera-based image forgery localization using convolutional neural networks [C]// Proceedings of the 26th European Signal Processing Conference. Rome. IEEE, 2018: 1372-1376.

[131] BARNI M, PHAN Q T, TONDI B. Copy move source-target disambiguation through multi-branch CNNs[J]. IEEE Transactions on Information Forensics and Security, 2021, 16: 1825-1840.

[132] WU Yue, ABD-ALMAGEED W, NATARAJAN P. Deep matching and validation network: An end-to-end solution to constrained image splicing localization and detection [C]// Proceedings of the 25th ACM International Conference on Multimedia. Mountain View California USA. New York, NY, USA: ACM, 2017: 1480-1502.

[133] BAPPY J H, ROY-CHOWDHURY A K, BUNK J, et al. Exploiting spatial structure for localizing manipulated image regions[C]// Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy. IEEE, 2017: 4980-4989.

[134] MAZAHERI G, MITHUN N C, BAPPY J H, et al. A skip connection architecture for localization of image manipulations[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Long Beach, CA, USA. IEEE, 2019: 119-129.

[135] HU Xuefeng, ZHANG Zhihan, JIANG Zhenye, et al. SPAN: spatial pyramid attention network for image manipulation localization[M]// Computer Vision-ECCV 2020. Cham: Springer International Publishing, 2020: 312-328.

[136] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Advances in Neural Information Processing Systems, 2017, 30: 5998-6008.

[137] BARTUSIAK E R, YARLAGADDA S K, GÜERA D, et al. Splicing detection and localization in satellite imagery using conditional GANs[C]// Proceedings of the 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). San Jose, CA, USA. IEEE, 2019: 91-96.

[138] 李应灿, 杨建权, 丁峰, 等. 区分来源和目标区域的图像 copy-move 伪造检测方法[J]. 信号处理, 2020, 36(9): 1533-1543.
LI Yingcan, YANG Jianquan, DING Feng, et al. Copy-move detection method for distinguishing between source and target regions[J]. Journal of Signal Processing, 2020, 36(9): 1533-1543. (in Chinese)

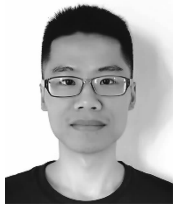
[139] ISLAM A, LONG Chengjiang, BASHARAT A, et al. DOA-GAN: Dual-order attentive generative adversarial network for image copy-move forgery detection and localization[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual. IEEE, 2020: 4676-4685.

[140] LIU Yaqi, ZHU Xiaobin, ZHAO Xianfeng, et al. Adversarial learning for constrained image splicing detection and localization based on atrous convolution[J]. IEEE Transactions on Information Forensics and Security, 2019, 14(10): 2551-2566.

[141] ZHOU Peng, CHEN B C, HAN Xintong, et al. Generate, segment, and refine: Towards generic manipulation segmentation[C]// Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 13058-13065.

作者简介

李昊东　男，1990 年生，广东湛江人。深圳大学电子与信息工程学院助理教授，工学博士，硕士生导师，主要研究方向为多媒体信息安全、智能信息处理等。
E-mail: lihaodong@szu.edu.cn



庄培裕　男，1995 年生，广东汕头人。深圳大学电子与信息工程学院博士研究生，主要研究方向为数字图像取证。
E-mail: 1800261051@email.szu.edu.cn

李　斌　男，1982 年生，广东五华人。深圳大学电子与信息工程学院教授，工学博士，博士生导师，主要研究方向为多媒体信息安全、智能信息处理等。
E-mail: libin@szu.edu.cn