像素级语义理解: 从分类到回归

李学龙*, 赵致远

西北工业大学计算机学院与光电与智能研究院, 西安 710072

* 通信作者. E-mail: li@nwpu.edu.cn

收稿日期: 2020-10-31; 接受日期: 2021-01-14; 网络出版日期: 2021-03-18

科技部重点研发计划 (批准号: 2018YFB1107400) 和国家自然科学基金 (批准号: 61871470) 资助项目

摘要 近年来, 随着科学技术的快速发展和硬件设备的不断迭代, 人工智能在各种领域 (如安防监控、医疗辅助、健康诊断、智能推荐、遥感监测、目标定位等) 都得到了广泛的应用. 随着人们对智能处理任务的要求逐步提高, 人工智能算法所需要理解的语义信息层次和输出数据精准度要求也步步攀升. 因此, 像素级语义理解任务也因其精准度要求远高于图像级理解而越来越受到重视. 与图像级理解相比, 像素级语义理解具有输出数据量大、逐像素输出精度高的优点, 相应地其难度也更大, 内部成因更值得关注与研究. 为此本文从信息度量的角度出发, 结合像素级语义理解任务的特有属性, 给出了像素级语义理解任务的定义与优化目标, 进一步依据实际任务的特性从初始定义衍生出像素级语义分类和像素级语义回归两类任务; 随后分别讨论了在这两类任务中优化目标的退化和演变, 并通过详尽的调研总结了常见像素级语义理解任务的发展现状; 紧接着探究了当前像素级语义理解的难点和未来发展方向, 针对亟待解决的问题给出了深入的分析思考以及可行的解决方案; 最后重点反思了后深度学习时代像素级语义理解乃至人工智能领域所面对的机遇与挑战, 提出知识的方向指导和数据的优化驱动是未来人工智能发展的重点关注目标. 本文意图从像素级语义理解的定义与发展现状出发, 延伸出对当前工作的思考以及对整个领域的反思, 强调整个领域面临的风险; 在介绍像素级语义理解基础认知的同时对相关技术的发展方向和路径进行深入的思考与深度的展望.

关键词 像素级语义理解, 人工智能, 深度学习, 分类, 回归

1 引言

近年来, 人工智能成为一个热点话题, 关于它的各种讨论和研究层出不穷. 在全球范围, 人工智能都被视为一项能够引领未来的战略技术, 这主要是基于以下几点原因. 首先人工智能催生的新技术、新产品能够对传统行业起到极其有力的赋能作用, 迅速地提升社会的整体生产力; 相关技术的合理应用能够在简单、重复或者具有危险性的场景中替代人类的工作, 减少人力成本的投入, 同时还提升了

引用格式: 李学龙, 赵致远. 像素级语义理解: 从分类到回归. 中国科学: 信息科学, 2021, 51: 521–564, doi: 10.1360/SSI-2020-0340
Li X L, Zhao Z Y. Pixel level semantic understanding: from classification to regression (in Chinese). Sci Sin Inform, 2021, 51: 521–564, doi: 10.1360/SSI-2020-0340

工作效率和工作质量;除了传统行业,人工智能还在教育、医疗、环境保护、城市管理等多个领域获得了广泛的应用,能够极大提高公共服务水平,有效提升群众生活质量;人工智能还能够辅助预测感知预警系统主动作出决策应对,提升社会治安水平,保障人民生命财产安全。当今人们的生活也已经进入一个泛智能化的时代,人工智能技术被广泛应用于各种领域,从海量数据分析^[1]、智慧城市^[2,3]等国家战略级应用,到全自动智能工厂^[4]、自动驾驶^[5]、智能家居等服务于社会的应用,再到手机中的植物识别软件等面向个人消费者的应用,人工智能已经渗透入每个人生活的点点滴滴。毫无疑问,未来是人工智能掌控的未来。

视觉作为人类最重要的感官之一,为人类提供了几大感官中几乎最多的信息,对于人工智能来说亦是如此。目前来说除了部分语音领域的应用,绝大多数的人工智能技术或系统都是基于视觉信息的。首先整个系统使用自己的眼睛,即光学传感器,采集目标区域的图像信息,然后采集到的图像被送入系统,使用特定的机器学习算法从中提取到所需要的信息,然后再进一步依据提取的信息作出相应的决策并给出应对方案。在最开始的应用中,系统只需要依据输入图像作出一些简单的判断,比如在生产线上依据产品的外观初步筛选产品合格与否;随后人们尝试使用算法进行稍复杂一些的任务,比如闭路监控系统中进行行人的跟踪和再识别;随着社会和技术的发展,人们所希望依赖人工智能完成的任务越来越多,越来越复杂,比如自动驾驶任务中就需要对可视场景中几乎所有物体有相应的理解和判断。因此,由于人类需求的不断发展,系统对算法提取、识别和利用信息的要求越来越高,机器学习算法也跟随着人类需求一路发展,从初期的图像级信息理解到图像局部信息理解再到像素级语义理解,算法要处理的任务难度逐步增大,要输出的信息也越来越多,对高级语义信息的理解能力要求也逐渐增强。

传统机器学习算法基于人工手动构建的特征模式从输入中提取和匹配相应的特征类型,这样的方法在应对初期不那么复杂的小规模机器学习任务时表现良好,但是受限于手动构建特征的有限表征能力,传统机器学习方法泛化能力差,只能用来处理手动构建特征时就纳入考虑的情况。可以通过构建更有力的特征提升传统机器学习方法的表征能力,但是高效特征的构建依赖于大量专业知识对特定解决问题的建模。因此对专业知识构建模型不那么依赖的深度学习方法在近年来大受欢迎。依赖于大量数据的堆砌,深度神经网络自动地完成特征提取方法以及特征利用方法的学习。无论是基于传统的机器学习方法还是基于深度学习的方法,像素级语义理解任务始终是机器学习乃至人工智能研究范畴中不可或缺的一部分。

本文对像素级语义理解进行系统性的介绍,给出了相应的定义以及优化目标,讨论了不同情况下从像素级语义理解到像素级语义分类和像素级语义回归的任务演变和优化目标差异,并分别介绍了每一阶段常见任务的发展脉络及研究现状。紧接着分析了当前亟待解决的数据和算力问题,最后分析了深度学习方法给当前像素级语义理解乃至人工智能领域带来的利与弊,反思了后深度学习时代所面临的机遇与挑战。

2 像素级语义理解基础

本节首先进行像素级语义理解相关概念的定义,随后介绍发展历史及国内外现状,最后讨论像素级语义理解的应用方向。

2.1 定义

在人工智能的应用领域中最基础而常见的图像识别以及分类任务通常被认为是图像级理解任务。例如推动深度学习发展的 ImageNet^[6],耳熟能详的 MNIST^[7], Fashion MNIST^[8] 和 CIFAR^[9] 等都

图 1 不同机器学习任务输出信息量的比较与发展趋势

Figure 1 Comparison and trend of output information for different machine learning tasks

属于图像级理解任务的范畴. 这类任务的特点对于任意一幅给定的图像, 相应的模型或者方法只需要给出该图像对应类别 (描述) 的预测即可. 输出的预测数据量通常很少, 往往只有几百、几十甚至几个, 仅仅是对整个图像的一个或几个数值的描述. 图像级理解只关注图像内少数特定目标的属性与整个图像间的内在关联.

与图像级理解任务不同, 像素级语义理解则致力于探索图像中所包含的每一个像素互相之间的作用机理, 是一种逐个像素理解图像语义信息的机制. 对于一个给定的目标图像, 像素级语义理解任务需要对图像中的每一个像素进行预测 (无论是分类还是回归), 这就意味着输出数据量通常是和输入数据量相同的, 甚至在某些特定任务中输出数据量会大于输入数据量. 图 1 展示了一些常见机器学习任务以及它们对应的输出信息量比较. 自然, 对于相同的输入数据, 随着需要预测的数据量增大, 任务的难度也会随之加大. 所以通常情况下像素级语义理解任务难度大于图像级理解任务, 但与之对应的像素级语义理解任务的输出结果所包含的信息量与价值也远大于图像级理解任务. 表 1 比较了通常情况下的像素级语义理解任务和图像级理解任务以及语音理解任务所具有的特点, 可以看到像素级语义理解任务具有输入输出数据量大、运算量大、计算设备依赖性高等特点, 这也从侧面说明了其应用和发展难度大的原因.

从图像级理解任务到像素级语义理解任务, 算法所关注的范围从整个图像到逐个像素, 最小单元逐渐缩小. 也有研究者尝试再进一步越过像素的尺寸限制探寻像素的内部元素之间的关系, 这样范畴的研究被称作亚像素理解, 在这里我们不作过多的介绍.

像素级语义理解任务可以更进一步被划分成为像素级语义分类任务和像素级语义回归任务, 也就是机器学习中分类和回归任务分别与像素级语义理解任务的结合. 分类和回归都是通过对成对已有数据进行建模拟合, 构建从输入空间到输出空间的映射关系. 区别在于分类任务的输出空间是一个定性的离散空间, 对输出结果来说只有错误与正确之分, 无法衡量各个类别之间的距离; 而回归任务的输

表 1 像素级语义理解和图像级理解、语音理解任务特点比较

Table 1 Comparison of pixel level semantic understanding, image level understanding and audio understanding			
	Pixel level semantic understanding	Image level understanding	Audio understanding
Input data size	$>10^6$	$<10^4$	$<10^4$
Output data size	$>10^6$	$1 \sim 10^3$	$<10^2$
Operation time	s ~ min	ms	ms
Power dependence	Extremely strong	Strong	Slightly weaker

出空间是一个定量的度量空间, 这个输出空间带有一个能衡量预测值和实际值差异的度量. 本文后面的章节将从分类和回归两个角度出发, 以数据输出量的从少到多和数据输入量的从多到少为主线分别介绍和阐述像素级语义分类任务和像素级语义回归任务.

从信息度量^[10]的角度出发, 对于给定的输入空间 A 和期望得到的输出空间 B , 使用如下公式:

$$\max \mathcal{I}_{\mathcal{P}}[Y; \mathcal{F}(X)], \quad X \in A, Y \in B \quad (1)$$

来描述本文所讨论的像素级语义理解任务, 其中 X 代表从 A 中随机抽样的一个元素样本, Y 代表与之对应的空间 B 中的样本. 像素级语义理解任务的目标就在于寻找一种映射关系 $\mathcal{F}(\cdot)$ 来最大化 Y 与经映射后的 $\mathcal{F}(X)$ 间的逐像素互信息 $\mathcal{I}_{\mathcal{P}}(\cdot)$. 依据互信息的定义可以将式 (1) 变换到如下形式:

$$\max[\mathcal{H}_{\mathcal{P}}(Y) - \mathcal{H}_{\mathcal{P}}(Y|\mathcal{F}(X))], \quad X \in A, Y \in B, \quad (2)$$

其中 $\mathcal{H}_{\mathcal{P}}(\cdot)$ 代表逐像素尺度下的熵, 这样一来优化目标就转换为最大化给定 $\mathcal{F}(X)$ 的条件下 Y 的不确定度的缩减量. 由于目标空间 B 中元素 Y 的熵是客观存在不可改变的未知固定值, 式 (2) 就可以简化为

$$\min \mathcal{H}_{\mathcal{P}}(Y|\mathcal{F}(X)), \quad X \in A, Y \in B, \quad (3)$$

即寻找 $\mathcal{F}(\cdot)$ 来最小化给定 $\mathcal{F}(X)$ 条件下 Y 的熵. 该公式给出了最终优化目标的描述, 但是在实际的像素级语义理解过程中, 这种定义下的熵通常难以计算, 所以在不同任务中会使用不同的衡量指标来替代该指标进行优化. 在随后的章节中我们将继续围绕该定义来分析不同情况下的像素级语义理解任务.

2.2 历史

人工智能是一个研究计算机的视觉、听觉、感觉甚至是情感的科学, 或者从另一个角度来说, 是通过硬件层面的传感器和软件层面的智能算法相配合, 使得智能设备能够自主地探测分析目标区域环境, 并对环境中的指令和信息作出自动化的反馈与交互的学科. 各种交互中最重要的计算机视觉涉及到对数字图像的理解与评估. 英国机器视觉协会 (The British Machine Vision Association, BMVA) 将机器视觉定义为“对单张图像或一系列图像的有用信息进行自动提取、分析和理解”. 计算机视觉任务的最终目的是使用视觉传感器和计算机相配合, 替代人眼、人脑对视觉目标进行识别、跟踪、测量和概括理解等任务. 计算机视觉领域涉及的研究范围十分广泛, 有大量一直在被研究的任务, 也有大量不断涌现的新问题. 图 2 展示了计算机视觉任务从图像级理解到像素级语义理解的演化过程.

早期的发展中由于算法和各种软硬件的限制, 计算机视觉领域的任务主要集中在图像级理解任务, 例如经典的人脸识别 (face recognition)、图像分类^[3](image classification) 等任务仅仅预测输出当前输入人脸 (图片) 的对应人物信息 (类别信息) 即可. 随着社会发展, 人们对计算机视觉获取信息的

图 2 (网络版彩图) 计算机视觉任务从图像级理解到像素级语义理解的演化

Figure 2 (Color online) The evolution of computer vision tasks from image level understanding to pixel level semantic understanding

精度要求进一步提高, 任务需求从原本的“图中有什么”提升到了“图中物体在哪儿”, 随之出现了诸如图像定位 (image location)^[11,12]、目标检测 (object detection)^[13]、目标跟踪 (object tracking)^[14~16]、群组检测 (group detection)^[2] 和异常检测 (abnormality detection)^[17,18] 的视觉任务. 这类问题想要得到的预测输出往往包含多个实例, 同时每个实例又由对应的类别信息和边界框 (bounding box) 共同组成. 这样一来, 对于相同的输入, 需要预测得到的信息大量增多. 这在一定程度上为整个视觉系统提供了更加丰富的视觉信息理解能力. 初期, 传统的机器学习方法诸如支持向量机 (support vector machine)^[19,20]、决策树 (decision tree)^[21,22]、逻辑回归 (logistic regression)^[23,24]、贝叶斯分类器 (Bayesian classifier)^[25,26]、随机森林 (random forest)^[27,28] 等在上述所提到的图像级理解任务中得到了广泛的应用.

自 2012 年 AlexNet^[29] 在 ImageNet 上展现出了出色的性能以来, 伴随着相关软硬件的逐步完善和性能提升, 深度学习的热度随之水涨船高. VGGNet^[30] 和 MSRA Net^[31] 先后尝试拓展网络深度, GoogLeNet^[32~35] 则尝试在横向上拓展网络的宽度. ResNet^[36] 和 DenseNet^[37] 则改进了网络优化的目标, 极大增加了网络的深度. 网络深度增加带来的问题是深度神经网络参数量的爆炸性增长, 过大规模的参数在提升网络性能的同时也极大降低了训练和推理速度. 所以一些研究者从优化网络运行速度的角度出发提出了诸如 SqueezeNet^[38], MobileNet^[39~41], ShuffleNet^[42,43] 等一系列改进的轻量化网络结构, 从而在网络性能和运算量之间达到一定的平衡.

随着深度神经网络在各个领域大展身手^[44], 再加上相关硬件设施的飞速发展, 科研人员意识到神经网络的强大拟合能力能做到的远不止简单的图像级理解工作, 同时工业界以及应用领域也对计算机视觉信息的理解精度要求逐步提升. 有学者尝试从像素级语义理解的角度解决原有的问题 (如边缘检测^[45] 和目标跟踪^[46,47]), 此外各方面需求的提升和条件的完善催生了多种新的计算机像素级语义理解任务.

全卷积网络 (fully convolutional networks, FCN)^[48] 开创性地抛开了常见神经网络结构中常用的全连接层, 使用端到端 (end to end) 的方式直接使用深度神经网络输出对图像中所有像素的类别预测结果. 对 $H \times W \times D$ (高度 $\times$ 宽度 $\times$ 通道数) 的输入, FCN 预测输出的最终结果为 $H \times W \times 1$, 即对于任意尺寸的图像输入, FCN 最终输出一个单通道的同尺寸数组, 这个数组中每个元素包含的值为预测出的原图像中相应位置像素点的所属类别. 这样的任务被称作语义分割任务. FCN 的意义在于引入反卷积 (deconvolution) 实现卷积网络的上采样过程, 开拓了卷积神经网络的全新运行思路, 推开了端到端神经网络在像素级语义理解中应用的大门. 随后各种语义分割方法大量涌现. 在语义分割迅猛发展的同时, 实例分割和新兴的全景分割任务也得到了不错的发展, 这些任务都属于像素级语义分类任务, 具体的细节请参考第 3 节.

除了上述提到的分类任务, 在某些特定的应用场景中, 简单的分类还无法满足实际的需求. 比如

超分辨、去模糊和图像生成任务中, 我们希望得到的是每一个像素点在指定范围内的一个精确值, 而不是所属类别这样一个简单而不精确的输出. 诸如此类的任务我们称为像素级语义回归任务. 深度神经网络也自然而然地被拓展应用到了这样的回归任务中.

2.3 国内外现状

虽然人工智能的概念诞生至今已经有半个多世纪, 但是它的发展经历几起几落, 磕磕绊绊并不顺利. 自 21 世纪以来, 以深度神经网络为代表的各类机器学习算法在广泛领域展现了优异的性能, 人工智能及其背后的算法再次受到全球学者和工业从业者的广泛关注. 人工智能的进展与计算机视觉任务的发展密不可分, 其中像素级语义理解在计算机视觉任务中扮演了越来越重要的角色.

国外研究现状. 早在 2013 年的美国国情咨文中, 奥巴马政府就提出人脑绘图计划, 预期投资 30 亿美元在预期时间点内完成人类大脑图谱的绘制, 借以探寻人脑内部的工作原理, 模拟人脑构建智能体系, 推动人工智能的发展. 欧洲联盟也在同年初公布了一系列“未来新兴旗舰技术项目”, 人脑工程和石墨烯位列其中, 每个项目都配备有专项的研究计划, 每项计划预计在未来十年内获得 10 亿欧元的研究经费. 为了维护其在全球科技霸主的地位, 美国将人工智能置于科技版图的核心位置. 从奥巴马政府时期以来, 就一直积极大力支持相关研究, 并有将政策态度从扶持和引导转向要求领先的趋势. 2019 年美国陆续颁布《维护美国在人工智能领域领导地位》《国家人工智能研发战略计划》和《美国人工智能时代: 行动蓝图》三项重要政策文件, 在相关政策文件中指出并强调像素级语义理解任务在人工智能领域的重要性. 欧洲联盟则于 2018 年颁布了《人工智能合作宣言》, 重点关注人工智能在工业、制造业的创新赋能, 针对像素级语义理解任务与传统工业的有机结合, 依托于“数字欧洲”计划和“地平线 2020”计划, 投资数十亿欧元提供扶持资金建立重点实验室. 《欧盟 2030 自动驾驶战略》更是展示了欧洲联盟希望凭借基于像素级图像解析的人工智能技术升级强化汽车产业、制造业等自身优势产业的计划. 根据不完全统计结果, 全球有数十个国家和地区发布了与人工智能相关的战略规划并进行了政策部署.

在学术界, 早在 1959 年美国麻省理工大学 (Massachusetts Institute of Technology, MIT) 就创办了人工智能实验室, 后来在 2003 年同计算机实验室合并成为计算机科学与人工智能实验室 (Computer Science and Artificial Intelligence Laboratory, CSAIL), 该实验室的研究范围十分广泛, 涉及人工智能、系统和理论的研究, 通过像素级语义理解算法推动机器人及系统的研究进展. 斯坦福大学人工智能实验室 (Stanford Artificial Intelligence Laboratory) 成立于 1962 年, 长期以来和工业界紧密合作, 紧抓实际产业和工业需求, 其发布的 ImageNet 计算机视觉识别数据库与挑战赛极大推动了计算机视觉乃至整个人工智能领域的发展进程, 通过图像级的理解任务为后续像素级语义理解的发展打下坚实的基础. 卡耐基梅隆大学 (Carnegie Mellon University) 在 1976 年成立了机器人学院 (Robotics Institute), 将图像像素级语义理解从算法层面和实际机器应用牢牢结合, 推进机器自动化的发展. 牛津大学 (University of Oxford) 虽然没有设立专门的人工智能实验室, 但是派出了大量相关领域的专家同谷歌人工智能研究组通力合作共同克服学术难题, 研制出了名噪一时的 AlphaGo. 微软亚洲研究院也和 20 多个国家的 300 余所高校建立合作关系, 拉近学术界同工业界前沿科技的关系, 组建多个团队致力于不同的像素级语义理解任务的探索.

在工业界, 国外大量知名企业也早早部署了各自在人工智能领域的产业升级方案, 通过像素级语义理解的实际落地应用与传统业务相结合. 诸如谷歌、微软、IBM, Amazon 等国际知名企业都给出了各自的人工智能产业赋能方案和应用. 早在 2013 年谷歌就囤积相关领域的人才资源, 先后收购 DeepMind 和 Andrew Zisserman 等领域领军企业. 安排公司内部多名基础算法平台开发者转战人工智

能领域,多个团队分别致力于图像级任务和像素级语义理解任务的研究.谷歌与相关高校的积极合作和大量投入更是催生了在某些特定领域超越人类表现的智能算法. Apple 多年来在智能自动化生产流程上的支出超过百亿美元,像素级的智能理解算法提供了极其精细的语义信息,使得超高精度的全自动/半自动工业化生产流程成为可能.在某些项目上,机器人的表现甚至超过了同成本的人类技术员.得益于像素级语义理解的极高解析精度,各大汽车企业针对像素级语义理解与自动驾驶和辅助驾驶的有机结合进行了大量科研上和工业化落地的投入,有力推动了像素级语义理解任务的发展.

在像素级语义理解研究发展的初期阶段,欧美等国家的相关探索更早一步,全球各个国家地区都已经认识到高精度人工智能算法对产业升级、未来建设的重要意义,并将其发展提升到战略高度,依据自身科研实力、人才储备、基础设施和国情需求调整侧重点,大力促进相关领域的研究发展.

国内研究现状.我国政府和相关的学术界、工业界也早已重视高精度人工智能任务,即像素级语义理解任务的相关研究和应用工作,相关的政府、地区政策和研究计划也纷纷落地启动.鉴于笔者有限的了解,本文仅简要介绍其中一部分.

在政府政策层面,在 2015 年国务院发布的《中国制造 2025》就提出“智能制造”的概念,同年发布的《关于积极推进“互联网+”行动的指导意见》将人工智能列入 11 项重点行动之一. 2016 年 3 月,人工智能被列入国家“十三五”规划纲要,同年 5 月发布的《“互联网+”人工智能 3 年行动实施方案》给出了到 2018 年人工智能基础资源与发展平台的创新目标,为基于像素级语义理解的人工智能研究打下了硬件和平台基础. 2017 年人工智能被写入全国政府工作报告. 同年 7 月国务院正式印发了《新一代人工智能发展规划》,确立了新一代人工智能发展的三步走战略,将其发展的重要意义提升至国家战略层面. 十九大报告强调了互联网、大数据、人工智能和实体经济深度融合的重要性. 2018 年 11 月《新一代人工智能产业创新重点任务揭榜工作方案》提出在人工智能主要细分领域选拔领头羊标杆企业培育创新发展主力军. 2019 年中央全面深化改革委员会审议通过的《关于促进人工智能和实体经济深度融合的指导意见》指出要把握新一代人工智能的特点,与产业化、区域化的特色深度融合,探索基于像素级语义理解的相关创新和落地应用成果转化的路径与方法. 国家层面的种种政策逐步将基于像素级语义理解的人工智能及其相关创新发展应用提升至国家战略和国家意志层面,这展现了国家大力发展该领域的决心,也进一步反映了像素级语义理解级别的人工智能已经成为引领科技发展、产业升级的重要驱动力.

在学术层面,早在 1981 年国家就注册成立了中国人工智能学会 (Chinese Association for Artificial Intelligence, CAAI),目前已经发展了四十余个相关的分支机构,成立了专门的分支团队致力于像素级语义理解的研究. 截至 2019 年国内已有 35 所高校获批人工智能专业建设资格,共有超过三百所高校申请了人工智能相关专业的建设. 在各个学校人工智能学院完成建设之前,相关专业的研究团队也早已大力推动了像素级语义理解任务在本专业领域的创新和研究进展. 同时近年来大量的国内科研机构和学会也开展了诸如中国人工智能大会、中国模式识别与计算机视觉大会等众多的学术研讨活动,像素级语义理解相关的研究也在各大研讨会上频频现身.

在工业界层面,当前已有超 30 个地区发布多项智能制造与人工智能专项扶持政策与计划,其中多项政策直接面向像素级语义理解在产业化中的落地应用. 每年的相关领域核心产业规模已经达到近千亿元. 国家地区政策的大力扶持也极大促进了相关产业的发展,孵化了大量的新生创业公司. 各个新生企业努力创新,借人工智能之势大力发展. 老牌企业也不甘示弱,纷纷成立自己的研究团队,深入创新发展融合产业和像素级语义理解技术,进行大量人力和资金投入,力争在新一代人工智能时代占据领先地位.

像素级语义理解作为计算机视觉乃至人工智能领域中最重要研究方向之一,国内外的相关实

表 2 常见像素级语义理解任务的研究方向及其应用
Table 2 Application of common pixel level semantic understanding tasks

	Research direction	Application
Pixel level semantic classification	Edge detection	Medical image segmentation
	Semantic segmentation	Auxiliary driving
	Instance segmentation	Autopilot
Pixel level semantic regression	Crowd counting	Public area security
	Image super resolution	Image restoration
	Image generation	Virtual data generation

验室和研究团队都把其相关研究作为首要关注点, 并且紧跟时代发展在像素级语义理解领域投入大量精力. 值得一提的是得益于网络的快速发展和开源精神的普及, 无论国内外的工业界还是学术界, 有越来越多的研究者们选择在发布成果的同时公开自己的算法源代码, 算法源代码的公开极大地方便了后续学习研究者复现以及改进工作, 有力地推动了整个人工智能领域的快速发展迭代进程. 有研究者制作了专门的网站¹⁾, 统计收集了带有算法的论文, 这里专门指出这一点供广大读者参考学习.

2.4 应用

图像级理解已经在人脸识别^[49]、车道线定位^[50]等多个行业领域得到了广泛的应用. 相比之下, 像素级语义理解提供了更加精细的语义信息输出, 自然也会在多个行业领域得到大面积的应用. 如表 2 所示, 在诸如道路检测^[51]、辅助驾驶^[52~54]、医疗图像分割^[55~57]、姿态检测^[58,59]等多个研究方向下的不同落地领域, 像素级语义理解都有大面积的应用, 随着社会发展的需求, 越来越多的落地应用方向将被逐渐开发.

3 像素级语义分类

依据 2.1 小节给出的定义, 本文将像素级语义理解分为像素级语义分类和像素级语义回归两个细节分支, 本节与第 4 节将分别从这两个方面具体讨论.

对目标区域进行逐像素的类别预测的所有任务统称为像素级语义分类任务. 对于式 (3) 给出的描述性定义, 当目标空间 B 是有限取值的离散空间 B_α 时, 针对抽样出的大小为 $M \times N$ 的特定向量 Y_α , 逐像素的条件熵退化为如下形式:

$$\min \left[- \sum_{i=1}^M \sum_{j=1}^N \sum_{c=1}^C p_\alpha(x_{ij}) \log(q_\alpha(x_{ij})) \right], \quad x_{ij} \in Y_\alpha, Y_\alpha \in B_\alpha, \tag{4}$$

其中 x_{ij} 代表 Y_α 在位置 $[i, j]$ 处的值, $p_\alpha(x_{ij})$ 代表抽样向量 Y_α 的真实离散概率分布, $q_\alpha(x_{ij})$ 则代表了使用映射 $\mathcal{F}(\cdot)$ 预测出结果的离散概率分布, $\sum^C$ 表示在所有可能的类别中逐类别统计叠加. 仔细对比可以发现, 这个时候的优化目标具有和交叉熵损失函数一致的形式, 而交叉熵损失函数便是像素级语义分类任务中最常见的损失函数和优化目标.

从像素级二分类 (pixel level binary classification) 的边缘检测到像素级多分类 (pixel level multi-class classification) 的语义分割, 对输入图像中的任意像素点, 我们只关注它们所属的类别信息; 而残

1) <https://www.paperswithcode.com/>.

图 3 (网络版彩图) 常见的像素级语义分类任务

Figure 3 (Color online) Common pixel level semantic classification tasks

实例像素级多分类 (partial instance pixel level multi-class classification) 中的实例分割与全实例像素级多分类 (entire instance pixel level multi-class classification) 中的全景分割除类别信息之外, 还在输出中包含了每个像素点对应的所属实例信息, 提供了更加丰富的语义信息. 这里依照算法输出信息逐步增多, 也就是目标域信息熵逐渐增大的主线对常见的像素级语义分类任务作简单的介绍. 为了方便理解, 图 3 简要展示了下文中提及的任务.

3.1 像素级二分类

像素级二分类任务的特点是目标空间可取的类别值只有两个, 导致累加熵的最小化变得相对容易, 因此从一定程度上来说分类难度低于多分类任务. 有限的类别取值使得式 (4) 进一步退化为

$$\min \left[- \sum_{i=1}^M \sum_{j=1}^N [p_{\alpha}(x_{ij}) \log(q_{\alpha}(x_{ij})) + (1 - p_{\alpha}(x_{ij})) \log(1 - q_{\alpha}(x_{ij}))] \right], \quad x_{ij} \in Y_{\alpha}, Y_{\alpha} \in B_{\alpha}, \quad (5)$$

其中使用的字母含义与式 (4) 中一致. 这里以边缘检测为例, 介绍典型的像素级二分类任务与方法. 边缘检测是在视觉领域长久以来一直被研究的任务, 是一种常见的图像预处理和特征转换方法, 也被看作当前分割技术的一种初期形态. 早期的算法将其转化为检测图像中连续点幅值变化明显的位置, 并从该角度出发, 尝试使用图像信息中的不连续性来实现边缘检测. 这个时候的方法集中于使用各种不同的算子来对图像进行滤波, 得到图像的梯度信息从而检测出图像的边缘信息. 常见的检测算子包括 Roberts 算子^[60]、Sobel 算子^[61]、Canny 算子^[62]等.

在基于深度学习的机器学习方法兴起之后, 研究者意识到对于边缘检测这个任务也可以从另一个角度进行理解. 即, 可以把边缘检测任务看作是一个分类任务, 更具体地说, 是一个逐像素的二分类任务. 对于图像中的每个像素, 它拥有边界像素和非边界像素两种类别. 通过这样的定义, 边缘检测任务就被转化为像素级分类任务. 这样能更容易地应用现有的计算机视觉方法.

2015 年在 IEEE 国际计算机视觉与模式识别会议 (IEEE Conference on Computer Vision and Pattern Recognition, CVPR) 上面世的 Deepcontour^[63] 尝试将神经网络应用到边缘检测中, 他们首先使用给定的参数 (包含用于训练神经网络的图像切片的大小和数量) 和标注信息将输入图像切片为大量指定大小的小块, 然后依据指定的数量进行聚类, 接着使用聚类得到的带类别信息的图像小块训练神经网络提取深度特征, 最后可视化每个图像小块的深度特征, 再使用结构化随机森林将可视化的深度特征进一步转化为相应图像小块的轮廓图. 该方法的贡献在于有效利用了神经网络强有力的表征能力来提取强有力的深度特征, 最后使用该特征可视化出相应的边缘信息. 但是依旧存在大量的问题, 例如, 方法不是端到端的, 数据的准备、网络训练以及最终结果的获取这些过程之间是割裂而又相对独立的, 不利于流程化的执行. 此外对于每个输入的图像, 无法直接得到想要的输出, 需要将输入图像切割为指定大小的小块, 每块单独处理, 最后再将所有小块的结果合并成整个图像的边缘图, 这

图 4 (网络版彩图) 不同思路下的边缘检测方法

Figure 4 (Color online) Edge detection methods under different directions

样就导致了整个流程的效率较低, 而且只能处理指定尺寸图像, 无法自由处理任意大小图像. 同年被提出的整体嵌套边缘检测方法 (holistically-nested edge detection) [45] 由 FCN [48] 得到启发, 直接使用一整张图像作为网络的输入, 最终网络直接输出完整的边缘检测结果, 整个方法是一个端到端的边缘检测系统. 该方法有效解决了之前方法的重复小图计算问题, 提升了速度, 同时能够接受任意尺寸的图像作为输入. 但是上述方法只利用了骨干网络中每个卷积阶段的最后一层特征输出, 导致最终得到的结果对卷积提取的特征利用率不够. 为了提升对每一阶段卷积特征的利用率, 富集卷积特征 (richer convolutional features, RCF) [64] 使用新设计的网络结构从骨干网络的每个阶段中抽出每个卷积层提取出的特征, 最终结果的输出利用到了所有的特征, 有效提升了网络输出在细节和边缘上的表现.

图 4 以 Sobel 滤波和一个简单全卷积网络为例展示了两种不同思路下的边缘检测方法. 滤波方法的优势在于运算量小、速度快, 但是在复杂图像上的检测精度不如深度学习方法. 得益于深度卷积神经网络的发展, 神经网络拥有强有力的特征提取表征能力, 这使得基于神经网络的边缘检测方法取得了远超传统方法的表现, 但是优异的检测结果依赖于大量的运算, 拖累了检测速度.

3.2 像素级多分类

人们对分类的要求不会仅限于两个类别, 随之衍生而来的是像素级多分类任务. 式 (4) 中所描述的就是最基本的像素级多分类任务的目标, 但是在实际的应用过程中, 目标域对应的标签是精确而又固定的, 即 $p_{\alpha}(x_{ij})$ 只会在当前像素点对应的正确类别上有值, 其他所有类别上值为零. 因此逐类别的累加过程可以在计算中被省略, 使得式 (4) 简化为

$$\min \left[- \sum_{i=1}^M \sum_{j=1}^N p_{\alpha}(x_{ij}) \log(q_{\alpha}(x_{ij})) \right], \quad x_{ij} \in Y_{\alpha}, Y_{\alpha} \in B_{\alpha}, \quad (6)$$

其中的字母含义与之前的定义保持一致. 边缘检测可以被理解为一个像素级的二分类任务, 那么语义分割就是与之形式类似但是难度加大的像素级多分类任务. 语义分割的过程通过为图像中所有像素逐一赋予预设的类别编号实现全图像的像素级多分类. 预设类别的增加也直接增大了分类的难度. 语义分割这一过程是由图像处理到图像分析的关键步骤.

在神经网络被广泛应用之前, 也有一些诸如基于阈值 [65]、区域 [66]、边缘 [67] 和直方图分割 [68] 的传统方法, 但是总体上效果都不是很理想. 当时传统方法的表征能力不足、硬件设备的计算能力偏弱、

没有大规模的语义分割数据集等诸多限制约束了初期语义分割工作的发展. 这里主要关注基于深度学习的语义分割方法, 不对传统方法作过多讨论.

在语义分割的开山之作全卷积神经网络出现之前, 卷积网络通常被用于大型数据集上的图像任务, 这些骨干网络通常都同神经网络的鼻祖 LeNet^[7] 一样, 在网络的最后接上全连接层作为分类器输出对应每一个类别的概率. 这样的结构设计可以说完全是为了图像分类任务而存在. 全连接层给网络带来的问题是它限定了网络输入和输出的尺寸, 这在很大程度上限制了网络在其他领域中的拓展应用. 全卷积网络摆脱了全连接层的限制, 完全使用卷积进行编码提取特征, 反卷积 (实质上还是一种卷积) 解码特征并完成分类. 这样, 在一个固定的网络结构下, 最后的输出大小是由输入的大小所决定的. 同时解除输入和输出的大小限制极大拓展了神经网络可能的应用面, 此后很多的方法可以说都是基于全卷积网络的这个思路而来的.

全卷积神经网络奠定了深度语义分割的基础, 但是也存在很多可以改进的地方. 它的卷积层感受野是固定大小的, 输入上采样 (反卷积) 阶段的特征空间尺度太小, 而且解码模块过于简单, 这些都导致了其对小物体的分割效果较差. 为了缓解这个问题, Noh 等^[69] 提出 DeconvNet, 使用全对称的编码器和解码器结构, 反卷积的过程从小尺度到大尺度逐步恢复图像的尺寸, 这样层层递进的解码过程有效提升了小尺寸物体的分割效果. 美中不足的是它在编码器的尾部和解码器的头部重新应用了全连接层, 给结构又套上了尺寸的限制, 同时导致了运算量大, 训练运行速度慢的问题. SegNet^[70] 对此提出了改进, 抛弃了 DeconvNet 中间的全连接层部分, 使用反池化层 (UpPooling) 对特征图进行上采样, 特征编码过程中的每一个最大池化操作中对应的最大值位置都被记录下来用于后续的反池化操作. 虽然这样的操作在一定程度上缓解了高频特征的丢失, 但是池化的过程依然造成了大量细节信息的损失. 同时该方法还引入了带权重的 SoftMax 函数应对分割样本不均匀的现象. U-Net^[71] 采用了和 SegNet 类似的结构, 但是与之不同的是它使用转置卷积 (UpConvolution) 恢复特征图尺寸, 在编码器和解码器的对称阶段中加入了跳跃连接 (skip connection), 将编码器阶段的特征跨层提供给解码器, 通过为解码器提供相应阶段池化层丢失的大量信息辅助提升分割细节.

与前面提到的这些对称网络结构不同, 有些方法从改进卷积的角度提升网络效果. DeepLab 系列在这个方向上进行了较为丰富的探索. 针对现有的大量池化操作和降采样过程带来的空间分辨率降低、位置和细节信息丢失等问题, DeepLabV1^[72] 引入了空洞卷积 (atrous convolution), 在不增加运算量的前提下扩大了卷积核的感受野, 提升网络的空间建模理解能力. 最终使用全连接条件随机场输出分割结果. DeepLabV2^[73] 中借鉴了 PSP Net^[74] 中提出的金字塔池化模块 (pyramid pooling module), 提出空洞空间金字塔池化 (atrous spatial pyramid pooling, ASPP). ASPP 中使用基于不同膨胀系数的密集空洞卷积替代金字塔池化中的不同尺度的池化操作, 辅以常规尺度的池化累积融合多尺度下的特征使得模型能够对不同尺度的物体同时拥有较好的建模效果, 进一步提升了模型在不同尺度上的细节表现. DeepLabV3^[75] 中进一步调整优化了 ASPP 的结构, 具体来说改变了其中空洞卷积的膨胀系数, 同时增加了池化特征, 骨干网络升级为 ResNet101, 并使用卷积网络替换了原本条件随机场直接卷积得到分割结果. DeepLabV3+^[76] 则更进一步完全将结构改成了编码器-解码器结构, 使用了更加复杂的解码器结构, 同时从编码器中引入底层卷积特征辅助解码器解码. DeepLab 系列的最新进展是基于网络架构搜索的 AutoDeepLab^[77], 尝试使用神经网络在给定的基于 DeepLab 结构的预设选择中自动搜索性能最好的网络结构.

也有研究者尝试将建模序列信息使用的循环神经网络用在语义分割中^[78], 这种尝试的出发点是将图像的行和列视为多个连续的序列信息, 然后使用基于序列的连续信息建模手段提取图像像素间的空间关系信息. 近期的注意力机制^[79,80] 和图卷积^[81] 也在语义分割中得到了应用^[82,83].

图 5 (网络版彩图) 基于轮廓表示和基于检测的实例分割方法

Figure 5 (Color online) Instance segmentation methods based on contour representation and detection

3.3 残实例像素级多分类

仔细考虑前面介绍到的像素级分类任务, 会发现它存在一个问题, 即虽然提供了足够精细的像素级类别信息, 但却无法区分图像中的同一类别像素. 例如, 它可以告诉你图中哪些像素是属于人这个类别, 但是通过它提供的信息无法区分图中不同的人. 为了应对这个问题, 在像素级多分类任务发展的同时, 也一直存在带实例信息的像素级多分类任务. 在目标任务中附加上实例信息后, 式 (4) 转变为式 (7):

$$\min \left[- \sum_{k=1}^C \sum_{l=1}^{I^k} p_{\alpha}(x_{ij}; I_{ij}^k) \log(q_{\alpha}(x_{ij}; I_{ij}^k)) \right], \quad x_{ij} \in Y_{\alpha}, Y_{\alpha} \in B_{\alpha}, I_{ij}^k \in I^k, \quad (7)$$

其中 C 为类别数目, I^k 表示在样本 Y_{α} 中包含的属于类别 k 的实例总数量, I_{ij}^k 是位置在 (i, j) 处的点所属的类别 k 的实例序号. 相应的 $p_{\alpha}(x_{ij}; I_{ij}^k), q_{\alpha}(x_{ij}; I_{ij}^k)$ 分别为实际和预测的目标类别与实例编号服从的概率分布, 其余字母与之前出现过的定义一致. 实例分割便是一种带实例信息的像素级多分类任务, 但是它不要求整个图像的每个像素的类别信息, 只关注感兴趣类别的像素以及它所属的实例类别, 因此我们将其归为残实例像素级多分类, 从分割的角度它也被称为基于区域的分割任务. 现在主流的实例分割算法有不同思路的解决方案, 下面依照解决问题的思路进行介绍.

基于检测的实例分割方法解决问题的思路是先使用检测算法检测得到关注类别物体的边界框, 然后在每一个边界框中问题就简化成为一个二分类问题, 也就是对框中每个像素, 它只可能有两个类别, 即属于框中主体类别的像素和不属于框中主体类别的像素. 图 5 上半部分展示了基于这种思路解决问题的过程. 这种检测思路下的实例分割方法与目标检测技术的发展不可分割. 从 R-CNN^[84], Fast RCNN^[85], Faster RCNN^[86] 这一系列目标检测方法发展而来的 Mask RCNN^[87] 沿用了 Faster RCNN 检测网络的设计思想, 采用更加有力的 ResNet-FPN^[88] 进行特征提取, 然后在原本的目标检测结构中添加一个用于预测二值化分割结果的网络分支. 最终在得到预测框的同时得到相应的分割结果. Mask RCNN 相对来说精度较高, 但是因为遵从的是 Faster RCNN 的检测思路, 算法中使用的诸如 ROI (region of interest) Pooling 操作推理速度较慢. FCIS^[89] 则是生成每个单类别物体的二值化语义分割结果, 然后使用目标检测的边界框生成该类别每个实例的位置敏感图, 最终结合语义分割结果和

位置敏感图得到实例分割的结果. 此外诸如 TensorMask^[90] 和 Yolact^[91] 都属于和检测算法紧密绑定的实例分割方法.

除了基于检测结果边界框做实例分割的思路, 还有一种思路是基于轮廓进行实例分割, 这种思路通常是对每个实例定义一种轮廓表示方法, 然后将实例分割任务转化为相应轮廓的预测任务. CurveGCN^[92] 使用图结构建模分割轮廓或分割曲面, 然后采用 GCN 预测轮廓多边形每个顶点的偏移量对曲线或曲面进行迭代, 从而实现分割. DeepSnack^[93] 基于传统的 Snake^[94] 方法提出进一步的深度化方法, 使用环形卷积来调整每个输入轮廓的顶点位置, 利用模型学习到的深度特征提取每个顶点相对于初始位置的偏移量, 从而使得所有顶点与实例本身更为贴近. PolorMask^[95] 则选择使用一个中心点加多个轮廓点的方式描述实例, 将实例分割转化为预测每个实例的轮廓点基于相应中心点的极坐标任务. 图 5 的下半部分可视化了这种极坐标表示下的实例分割方法. 这些方法通过转换问题, 提出了相当巧妙的实例分割解决方案, 让人耳目一新.

3.4 全实例像素级多分类

残实例像素级多分类任务所关注的像素类别有限, 往往只涉及少数前景元素, 为了弥补这一缺陷, 新兴的全实例像素级多分类任务要求获得全部元素的实例信息以及类别信息. 因此可以用如下的全实例逐像素优化目标来表示全实例像素级多分类任务:

$$\min \left[- \sum_{i=1}^M \sum_{j=1}^N \sum_{l=1}^{I^k} p_{\alpha}(x_{ij}; I_{ij}^k) \log(q_{\alpha}(x_{ij}; I_{ij}^k)) \right], \quad x_{ij} \in Y_{\alpha}, Y_{\alpha} \in B_{\alpha}, I_{ij}^k \in I^k, \quad (8)$$

其中所有变量含义都与前文定义保持一致. 全景分割 (panoptic segmentation)^[96] 便是这样一个新兴的全实例像素级多分类任务, 它相当于前面提到的语义分割和实例分割的有机结合体. 全景分割任务为图像中的每个像素点赋予一个包含类别信息的语义输出和一个包含实例信息的编号输出. 该任务的提出团队 FAIR 也随任务定义给出了一种名为 Panoptic FPN 的网络结构^[97], 通过设计同时包含语义分割子网络和实例分割子网络的多路结构实现一个最终直接用于全景分割的统一网络. MIT 联合谷歌提出 DeeperLab^[98], 基于 DeepNet 的结构, 将其拓展为同时完成物体实例分割和背景分割两项子任务. 前面两种方法虽然最终都实现了全景分割的任务, 但是从网络设计上来说, 都还是将全景分割视为语义分割和实例分割两个割裂的子任务来处理的, 只在两个分支网络利用了相同的特征提取器. 为了进一步探究子任务间的关联, 中国科学院自动化研究所提出的 AUNet^[99]、丰田研究院设计的 TASCNet^[100] 和 Uber 提出的 UPSNet^[101] 都尝试在实例分割子网络和语义分割子网络间构建联系、交换信息以探究互相之间的辅助指导作用. 图 6 展示了全景分割任务的不同解决思路.

3.5 像素级语义分类小结

本节从条件熵优化目标的演变过程出发, 以边缘检测为例介绍了像素级二分类任务的发展, 基于语义分割方法分析了像素级多分类方法的现状, 接着分别讨论了残实例像素级多分类任务和全实例像素级多分类任务的优化目标和当前研究进展. 对于这 4 类子任务来说, 在输入相同信息量级别的图像情况下, 输出的信息量却逐步提升, 输出所属的目标域信息熵也逐渐增大, 任务难度也逐步增大. 图 7 展示了本节提到的 4 种像素级语义分类任务的发展历程, 选取了一些关键的算法作为例子展现在发展时间线上, 方便读者配合本文查阅资料快速了解相关研究工作.

篇幅所限, 本文没有进行详细的代码实现讨论, 有需要的读者可以参考相关的代码开源网站, 例如

图 6 (网络版彩图) 全景分割方法的发展
Figure 6 (Color online) Development of panoramic segmentation methods

图 7 像素级语义分类的发展
Figure 7 Development of pixel level semantic classification

边缘检测²⁾、语义分割³⁾、实例分割⁴⁾、全景分割⁵⁾。

2) <https://github.com/MarkMoHR/Awesome-Edge-Detection-Papers>.
3) <https://github.com/mrgloom/awesome-semantic-segmentation>.
4) <https://github.com/wutianyiRosun/Segmentation.X>.
5) <https://github.com/Angzz/awesome-panoptic-segmentation>.

图 8 (网络版彩图) 常见的像素级语义回归任务

Figure 8 (Color online) Common pixel level semantic regression tasks

4 像素级语义回归

与像素级语义分类类似, 像素级语义回归任务期望输出的也是逐像素的预测结果, 其区别在于分类的结果是不连续的类别信息, 而回归结果是在指定范围内的连续的具体数值. 不同于式 (4) 展示的离散空间中的优化目标, 如果目标空间为无限取值的连续空间 B_β , 那么其中抽样向量 Y_β 在位置 $[i, j]$ 处的取值 $x_{i,j}$ 所服从的概率分布 $p_\beta(x_{i,j})$ 也随之转变成了连续的概率分布函数. 相应地, 以式 (3) 作为优化目标的逐像素条件熵可以使用如下公式衡量:

$$\min \left[- \sum_{i=1}^M \sum_{j=1}^N \int \int p_\beta(x_{ij}) \log(q_\beta(x_{ij})) dp_\beta dq_\beta \right], \quad x_{ij} \in Y_\beta, Y_\beta \in B_\beta, \quad (9)$$

其中 M 和 N 的定义与式 (4) 中一致, $q_\beta(x_{ij})$ 依然代表 $\mathcal{F}(\cdot)$ 预测出的结果在位置 $[i, j]$ 处服从的分布, 但是在像素级回归问题中其服从的是连续的分布. 由于 $p_\beta(x_{ij})$ 服从未知的连续分布, 所以此时的优化目标是无法准确计算的. 但是目标空间 B_β 拥有自己空间内的度量方法, 所以在像素级语义回归任务中研究者常常使用该度量来替代式 (9) 衡量预测结果与期望值之间的距离, 最终通过最小化该度量下的距离优化映射 $\mathcal{F}(\cdot)$.

依据不同的回归任务所拥有的输入信息量和输出信息量比例的不同, 像素级语义回归任务可划分为: 输出小于输入的单通道像素级回归 (single channel pixel level regression), 输出等于输入的抽象语义指导的多通道像素级回归 (multi-channel pixel level regression with guidance) 和无额外指导的多通道像素级回归 (multi-channel pixel level regression without guidance), 输出大于输入的倍增多通道像素级回归 (multi-channel pixel level regression with multiple output), 以及不依赖于输入信息的零输入多通道像素级回归 (multi-channel pixel level regression without input). 为便于直观理解, 图 8 简要展示了介绍到的像素级语义回归任务实例.

4.1 单通道像素级回归

单通道像素级回归的特点是从输入图像回归得到同等尺寸的单通道输出. 与前文中所讨论的多分类任务类似, 对于抽样样本 Y_β 来说, 它拥有以精准数值形式存在的标注信息, 因此像素点 (i, j) 处真实的分布函数具有冲激函数的形式, 即只在标签值处有值为 1, 其余值处处为 0. 此时, 式 (9) 便简

化为

$$\min \left[- \sum_{i=1}^M \sum_{j=1}^N \int p_{\beta}(x_{ij}) \log(q_{\beta}(x_{ij})) dq_{\beta} \right], \quad x_{ij} \in Y_{\beta}, Y_{\beta} \in B_{\beta}, \quad (10)$$

其中所有的变量含义与前文定义保持一致. 这里以人群计数为例简要介绍单通道像素级回归任务的特点和相应的算法方案.

面向大型活动安保需求, 需要对人群场景进行实时监控分析, 以便及时安排相应的安保措施, 从而有效避免人群踩踏、伤亡等事件的发生, 因此, 需要利用图像采集设备进行自动化的人群场景采集、理解和监管, 而人群计数便是密集人群场景理解中的一个基础任务. 早期阶段, 很多研究者尝试使用基于检测的方法解决人群计数的问题. 这类算法的流程是使用一个预设的滑动窗口^[102]逐步检测图像中每个小块场景中是否包含行人, 最后累加在一起得到总的人数. 至于具体的检测方法, 有的解决方案是基于人体整体的检测方法, 这类方法首先基于诸如小波变换、HOG^[103]特征、边缘检测^[104]的方法提取生成图像特征信息, 然后使用 SVM^[105]、随机森林^[106]等分类方法判断当前小窗中的特征是否与行人特征相匹配. 基于整体检测的方法比较适合检测稀疏场景中的人群, 随着人群密度的增加, 人和人之间的遮挡影响越来越严重, 这样就会极大地阻碍特征的匹配. 为了应对这个问题, 基于身体部分结构的检测也被应用到人群计数任务中. 例如文献^[107, 108]中就通过检测人头、肩膀统计人群的数量. 基于人体部分结构的检测方法在稠密一些的场景中表现优于基于整体的方法, 但是此类方法都只能处理不太稠密的场景, 因为当人群密度过高时很容易出现行人间的大面积遮挡, 这样就会导致此类方法失效 (无论是基于人体整体的检测还是基于部分人体的检测). 因此人们尝试从回归的角度出发, 直接预测人数而不是检测后进行累加统计.

从回归的角度去理解, 人群计数可以认为是一个图像级理解任务, 对于一个给定的输入图像, 它只要求一项对应的输出, 即图中包含的人群数量. 初期的基于回归的人群计数方法解决问题的思路是学习一种特征到人群数量的映射关系. 这类方法拥有两大主要步骤, 首先使用合适的特征提取方法 (如前景特征^[109]、边缘特征^[110]、纹理^[111]和梯度特征) 提取低级图像特征; 然后基于提取的特征构建回归模型 (如线性回归、分段线性回归、岭回归和高斯 (Gauss) 过程回归^[110]等), 建立从低级特征到人群数量的映射关系. 但是在研究的过程中发现, 这样直接预测一个数字容易导致预测偏差变大, 不利于任务的实现. 于是人们转而将单个数字的预测转化为对人群密度图的预测. 对于一张给定人头位置标注的输入图像, 使用高斯核对图像进行滤波就可以将孤立的人头点分散到其附近的区域中, 也就是说最终得到的密度图中每个像素都包含着它所含有的人头个数, 当然这个时候人头个数变成了连续的小数, 而不是离散的整数. 这样的话问题从对一个数的预测转化成了对很多个数的预测, 虽然在一种理解上加大了任务的难度, 但是却通过增加预测数量减少了误差波动的影响, 最终提升了预测精度.

当人群计数问题被转化为密度图估计问题后, 人群计数也自然而然成为了一个像素级语义回归任务. 这样的话, 之前提到的多种像素级语义分类任务经过简单的转换就可以应用到人群计数问题中来. 基于深度学习的人群计数任务也紧跟着各种网络结构设计的步伐不断发展, 多分支网络^[112]、注意力机制^[113, 114]、可变形卷积^[115]等网络结构通过改进都被引入到人群计数任务中. 也有研究者尝试将人群计数和定位、目标检测或者分割任务结合起来做多任务 (multi-task) 的工作^[116~118], 并探索不同任务之间的互相关联和辅助效果. 图 9 展示了常见的基于检测和基于神经网络回归方法解决人群计数问题的思路.

4.2 抽象语义指导的多通道像素级回归

当回归任务要求的输出和输入同尺寸同通道数时, 我们称之为多通道像素级回归. 相比于单通道

图 9 (网络版彩图) 基于检测和基于回归的人群计数方法

Figure 9 (Color online) Crowd counting methods based on detection or regression

像素级回归, 多通道回归的任务量直接翻倍. 因为直接回归大量的数据难度很大, 在很多的任务中, 会使用额外的信息指导回归任务. 通常把这种带有额外指导信息的多通道像素级回归称为抽象语义指导的多通道像素级回归. 为了保持回归输出结果与给定信息的抽象语义层面上的一致性, 语义提取变换 $\mathcal{F}_S(\cdot)$ 被引入到优化目标中:

$$\min \left[- \sum_{i=1}^D \sum_{j=1}^M \sum_{k=1}^N \iint \mathcal{F}_S(p_{\beta}(x_{ij})) \log(\mathcal{F}_S(q_{\beta}(x_{ij}))) dp_{\beta} dq_{\beta} \right], \quad x_{ij} \in Y_{\beta}, Y_{\beta} \in B_{\beta}, \quad (11)$$

其中 $\mathcal{F}_S(p_{\beta}(x_{ij}))$, $\mathcal{F}_S(q_{\beta}(x_{ij}))$ 分别表示语义提取变换后的真实分布和预测分布情况, D 为回归目标的通道数. 这里以图像风格迁移为例介绍此类任务的特点.

图像风格迁移是指通过一定的技术和算法将指定图像的内容和其他的一个或多个图像的风格融合在一起, 生成一个兼具前者内容和后者风格的新图像. 这也可以说是一个图像生成任务, 但是与后续 4.5 小节介绍的图像生成任务不同的是, 这里的生成过程必须使用所提供的图像内容信息以及图像纹理信息, 即生成的是“非全新”的图像, 而后者生成的是全新的图像. 图像风格迁移的任务通常来说会被拆解为 3 个步骤: 首先是目标图像风格的提取, 这一步也经常被称为图像纹理信息的建模过程; 然后是原始图像的内容信息提取, 确定需要融合目标风格的对象主体; 最后是基于前面提取到的风格特征和内容特征重建全新的图像.

图像的风格是一个很难量化的特征, 没有直观的衡量方法. 在研究的初期, 图像的纹理特征也被视为风格特征的一种. 纹理合成分析的方法也是多种多样的, 其中基于样图的纹理合成也就是图像风格迁移的前半部分, 即风格 (纹理) 特征的构建. 它要解决的问题是提取给定的一小块纹理样本的纹理信息, 然后生成更大的拥有同样纹理信息的图像. 基于样图的纹理合成主要面临以下问题: 纹元的采样方法、纹元之间关系模型的构建、高效率地合成高质量纹理. 文献 [119] 首先生成预设尺寸的初始图像, 然后对图像中的每一个像素点逐个进行合成. 在每个像素点的合成过程中, 以当前待合成像素点为中心选取不同大小的方形窗口, 在纹理样本图像中滑动遍历匹配与选取出的窗口最相似的一个, 然后使用这个最相似窗口的中心像素点替代待合成的像素点. 这里衡量两个窗口相似度的方法是计算两个窗口中除去需要合成的点之外所有像素点的均方误差, 均方误差越大也就意味着两个滑动窗口之间

的差距越大, 越不相似. 此外为了区分一个窗口中与中心点距离不同的像素点的影响, 在计算均方误差时还加入了基于距离考量服从高斯分布的加权系数. 系数的加入使得距离中心点越近的像素权重越大, 增加了模型的局部结构建模能力. 这样的方法在算法实现上比较简单清晰, 较大的窗口设置下可以较好地保持纹理结构特性, 但是逐像素点进行预测的思路导致算法效率低, 计算量大. 文献 [120] 紧接着改进了上述方法, 整体的思路是一致的, 最大的区别在于进行窗口滑动匹配时选取的窗口不再是原本的简单方形窗口, 而是使用了改进的 L 形窗口, 并且提出使用带随机噪声的纹理作为初始结果以辅助边缘处的像素合成. 针对合成速度过慢的问题, 作者也提出多尺度金字塔方法对样本图和合成图进行下采样, 再利用树结构向量量化算法加速窗口的匹配查找过程, 最终提升了算法的效率. 为了进一步提升像素合成的速度, 文献 [121] 延续同样的思路, 但是每次匹配完成之后一次性合成所有未合成区域的一块像素点而不是一个像素点, 这样极大加快了像素的合成速度, 但是由于每次生成的区域面积过大, 在每次生成的区域之间会存在比较明显的接缝. 为了解决这个问题, 作者还提出使用动态规划的方法找到能使得相邻两个窗口之间差异最小的接缝. 此外, 如果将算法使用的初始化图像从噪声图像替换为特定图像经过处理计算得到的亮度特征图像, 然后这个方法就可以实现纹理迁移的效果. 而后文献 [122] 则从改进滑动窗口的角度出发, 不使用预设形状和大小的窗口, 而是先通过图切割方法计算出图像的最佳接缝, 然后依据得到的最佳接缝分割出合适的窗口用于纹理匹配和生成. 这个阶段的纹理合成和迁移工作都只是基于局部信息和人工构造的低级特征的工作, 对图像的语义信息表征能力不充分, 因此进步缓慢.

在卷积神经网络获得长足的进步之后, 人们发觉了它极强的特征提取与表征能力. 于是 Deep Textures^[123] 提出使用经过预训练的神经网络来提取输入图像的特征向量, 然后在特征上计算 Gram 矩阵来表示这个图像所拥有的纹理信息. 在图像的初步重建过程中, 该方法使用基于梯度下降的慢速重建方法逐个更新重建图像的所有像素点, 最终优化的目标是使得重建的图像经过卷积网络提取出的特征和对应的 Gram 矩阵接近于风格图像的特征向量以及 Gram 矩阵. 在完成重建之后再使用一个平滑函数来处理图像边缘. 受到前者工作的启发, 文献 [124] 通过严密的计算证明了最小化 Gram 矩阵差异等同于最小化两个图像内部统计变量分布间的二阶核函数下的最大均值差异, 并由此推测出其他的最大均值差异核函数诸如一阶线性核函数、高斯核函数, 也都可以用于衡量图像的风格或纹理信息. 随后他们设计实验证明了这一点. 这种基于高层卷积特征的建模方法可以有效提取图像的特征纹理信息, 但是高层特征下的分辨率损失导致了图像重构过程中不可避免的边缘、细节信息缺失, 最终使得生成的图像带有不和谐的形变等问题. 为了提升重构过程中的细节表现, 文献 [125] 在重建图像的过程中加入新的损失函数, 首先使用基于拉普拉斯 (Laplace) 算子的滤波方法处理输入的内容图像和重建之后的图像, 然后将这两个经过预处理的图像间的差异纳入重建过程的考量, 最终提升了重建图像的细节表现. 文献 [126] 则将逐窗口匹配生成的思路应用到卷积特征空间中, 同时引入内容损失来降低高层语义信息的丢失.

上面提到的算法重建过程需要不断地更新迭代才能完成重建图像的像素值更新, 最终得到重建输出, 这种算法属于慢速重建算法 (也被称为在线算法). 为了提升迁移重建图像的速度, 研究者们提出使用基于离线模型优化的快速图像重建算法进行风格迁移. 快速图像重建算法通过事先训练好的模型或者设定的方法直接使用输入图像即可输出迁移后的图像. 文献 [127] 提出一种使用基于输入图像的主色调进行特征级重着色的颜色量化方法, 然后基于这种方法实现了一种油画纹理着色方案用于将输入图像转化为油画风格. 文献 [128] 则是通过分析对比度、亮度等差异, 实现了一种将普通头像图片转化为特定头像摄影风格的离线算法. 在工业化的实际应用中也有一些基于多层不同参数滤波器叠加实现图像风格迁移的算法. Johnson 等^[129] 提出使用感知损失函数训练前馈网络, 用于图像变换任务, 测

试或者应用阶段只有简单明了的前向传播过程,没有迭代和反向传播.文献[130]则是使用基于多尺度的网络结构进行风格迁移.文献[131]在团队之前提出的慢速重建算法的基础上进行了改进,使用对抗生成网络替代原本算法中的窗口匹配过程,从而优化算法的速度.但是诸如这样的方法都是基于单风格的解决方案,也就是一个训练好的模型只能生成一种风格的图像.与之相比,如果训练一个模型能够学习到多种不同的风格,这就是单模型多分割快速图像风格迁移方法.文献[132]通过实验发现在一个已经训练好的风格迁移网络上只需要再加一个实例正则化层上的仿射变换就可以实现完全不同的风格效果,基于这个设定他们使用一个小型的网络学习不同的条件实例正则化层的参数,从而使同一个网络中包含能生成不同风格的参数.文献[133]和[134]则分别将使用风格图像计算得到的风格特征和与噪声结合的风格图像作为输入信息指导网络进行风格的学习,从而实现一个网络学习到多种不同风格的迁移方法.更进一步有学者研究单模型的任意风格快速重建算法,文献[135]在卷积特征空间中分别找到内容图像和风格图像的内容特征以及对应的风格特征,然后进行风格特征的交换,最后使用交换后的特征重建出兼具内容图像的内容和风格图像的风格的新图像.但是因为风格特征的交换耗时较大,这个方法的速度还是比较慢.文献[136]改进了文献[132]中的实例正则化方法,提出了自适应实例正则化方法将内容特征与风格特征的统计参数自适应对齐,从而在使用过程中直接将输入图像自适应正则化为想要的风格.文献[137]使用给定的风格图像通过一个风格预测网络预测每个风格变换对应的仿射变换的参数,然后再使用基于文献[132]的方法生成重建的图像.这些单模型多风格的方法都是基于数据驱动训练的网络,对于数据的依赖性很大,这就会造成应用过程中生成图像偏向于训练数据中的图像风格的问题.为了应对这类基于数据驱动算法的缺陷,文献[138]提出一种不基于数据训练学习的风格迁移方法,该方法首先使用神经网络抽取内容图像的卷积特征然后使用风格擦除变换抹去特征中包含的风格信息,只保留包含内容信息的高级语义特征,然后使用色彩变换将风格图像的颜色和风格信息迁移到语义特征上,最后重建出风格化结果.

4.3 无额外指导的多通道像素级回归

在实际的应用中,有额外语义信息指导的任务始终还是少量,虽然说没有额外指导的多通道像素级回归任务难度变大,但是其研究价值相对也更高.从数学和信息熵的层面理解,无额外指导的多通道像素级回归任务直接最小化多通道回归值和目标值的熵,而不是式(10)中经过语义信息变换提取的分布之间的熵.具体公式表示如下:

$$\min \left[- \sum_{i=1}^D \sum_{j=1}^M \sum_{k=1}^N \iint p_{\beta}(x_{ij}) \log(q_{\beta}(x_{ij})) dp_{\beta} dq_{\beta} \right], \quad x_{ij} \in Y_{\beta}, Y_{\beta} \in B_{\beta}, \quad (12)$$

其中所有的字母与前文定义一致.这里以图像去模糊为例,介绍典型的无额外指导多通道像素级回归任务特点与相关算法思路.值得注意的是早期的部分图像去模糊任务解决思路只涉及一些诸如简单滤波器等底层图像处理方法,这样的方法不考虑具体的图像语义信息.但是随着研究的进一步深入,越来越多的方法使用精细的语义信息指导图像去模糊任务的具体实现,像素级语义回归方法的出现为图像去模糊任务开辟了新的道路.

计算机视觉领域的所有工作都离不开图像信息的获取,但是受到抖动、灰尘、失焦、采集装置性能限制等不可控因素的影响,在图像的采集过程中难免会出现图像模糊,最终导致图像质量的下降,在很多领域这种图像质量损失带来的影响不可估量.因此,图像去模糊技术应运而生.该任务的目标是研究如何从模糊的图像出发获取清晰的图像.当前的研究过程中,研究者将图像产生模糊、分辨率降低等质量下降的过程称为图像的退化.对于图像模糊这一退化过程,一般被建模为通过模糊核与原图

像进行卷积生成模糊图像的过程. 通常使用如下公式表示图像的模糊过程:

$$\mathcal{B} = \mathcal{I} \otimes \mathcal{K} + \mathcal{N}, \quad (13)$$

其中 $\mathcal{B}$ 代表模糊图像, $\mathcal{I}$ 指代原本的初始图像 (也就是我们想要得到的图像). $\mathcal{K}$ 是导致图像模糊的模糊核 (卷积核), $\mathcal{N}$ 是累计附加的噪声信息, $\otimes$ 代表着卷积操作. 基于这样的建模思路, 去模糊方法解决问题的关键点集中在模糊核的构建、噪声干扰的去除, 然后是基于模糊图像、模糊核和噪声建模反向推理原始清晰图像的过程.

根据 $\mathcal{K}$ 是否已知, 去模糊任务通常被分为非盲去模糊和盲去模糊两个分支. 顾名思义, 非盲去模糊指的是在已知模糊核的情况下尝试解决图像去模糊任务. 在这种设定下, 任务被简化为对噪声信息的去除和反向卷积推理的拟合. 由于模糊核提供了模糊构成的重要信息, 反向推理的过程相对来说被简化, 研究者的注意力更多地集中在图像修复过程中如何在提升细节的同时抑制噪声. 当然这里的模糊核可能是准确的真实模糊核, 也有可能是预测得到的模糊核. 从理论的角度来说, 绝对真实准确的模糊核是不存在的, 这里所提到的“真实模糊核”只是指在一些特定的实验环境和设备下, 在采集图像的同时记录采集设备、环境甚至被采集对象状态等难以获取的数据, 然后使用被采集的数据推算出相应的模糊核用于图像修复.

根据处理数据的域不同, 可以将非盲去模糊方法分为基于频域的去模糊方法和基于时域的去模糊方法. 基于频域的方法首先将图像通过诸如傅里叶变换 (Fourier transformation)^[139]、小波变换^[140]的方法转换到频域, 然后在频域数据的基础上分析图像的模糊状况, 完成去模糊之后再转换为人眼可以直接理解的时域图像. 最初, 简单的逆滤波被引入到去模糊中^[141], 也就是只在频域进行卷积的逆过程. 基于逆滤波的方法完全没有应对噪声的能力, 为了更好地处理图像中的噪声以生成更高质量的清晰图像, 研究者们引入维纳 (Wiener) 滤波^[142]、切比雪夫 (Chebyshev) 滤波器等方法在频域去除模糊和噪声信息. 基于时域的方法则通过直接在原本的图像建模实现去模糊. 比如 Lucy-Richardson 去模糊算法^[143]通过贝叶斯 (Bayes) 理论、泊松 (Poisson) 分布和最大似然估计这类基于概率的方法修复模糊图像. 文献 [144] 则提出一种动态结合全变差和分数阶全变差的方法建立去模糊的凸函数模型, 再用变量分裂和交替方向法求解模型, 这样的求解方式为模糊图像提供了更多的细节信息.

相比较来说, 非盲去模糊任务更加容易解决, 但是现实情况中, 盲去模糊任务的应用却更加广泛. 因为通常的图像采集过程中没有充分的条件记录用于生成模糊核的辅助信息. 比如最常见的运动模糊, 一般至少掺杂了采集设备的运动和被采集设备的运动, 复杂的运动过程以及环境因素、设备条件的限制导致我们通常能获取到的只有单独的一张模糊图像. 最初, 盲去模糊的解决方案关注于先使用一定的方法模拟计算出相应的模糊核, 然后再使用上面提到的非盲去模糊方法推理出清晰图像. 这种思路中模糊核的拟合对最终的结果起到了决定性的作用. 代表性的方法如文献 [145] 中首先使用双边滤波器和冲击滤波器组合成为滤波器组, 通过调整后的滤波器组预处理图像突出图像细节、削弱噪声影响, 然后使用基于正则化稀疏先验的方法拟合模糊核, 最后依赖于近似的模糊核结合全变差正则化方法恢复清晰图像. 此外还有使用 Radon 变换^[146]、运动密度方程^[147]等途径推算模糊核的研究. 在神经网络兴起后, 很多研究者尝试使用卷积网络拟合模糊核. 文献 [148] 使用卷积神经网络学习输入图像的运动模糊场, 然后基于学习到的运动模糊场完成去模糊. 文献 [149] 则尝试直接使用神经网络拟合频域下的卷积核, 跳过从频域到时域转换的步骤以减少转换过程中造成的信息损失. 随着神经网络展现出越来越强的能力, 研究者意识到与其使用神经网络拟合卷积核, 然后再使用模糊图像和卷积核作逆卷积生成清晰图像, 不如直接使用神经网络预测输出清晰图像. 随后去模糊的发展便逐步走向基于神经网络的端到端学习过程. 文献 [150] 提出了一种多尺度卷积神经网络, 以模糊图像作为输入直接输出生成的清晰图像, 同

时采集生成了一个包含成对的模糊和清晰图像的大型数据集,有力推动了端到端思路下的神经网络去模糊方法发展.文献[151]则针对大动作尺度差异下的去模糊提出共享参数的多尺度递归网络,有效拓展了网络的感受野范围.如果换个角度来看,也可以将去模糊问题理解成一个图像生成任务,所以也有研究者提出基于生成对抗网络(generative adversarial networks, GAN)的去模糊方法. DeblurGAN^[152]使用从模糊图像生成清晰图像的生成器和判别图像是模糊还是清晰的判别器对抗训练,最终得到可以生成高质量清晰图像的生成器.文献[153]则提出了同时包含使用清晰图像生成模糊图像和使用模糊图像生成清晰图像的两套对抗网络的结构,第一个对抗网络通过未配对的清晰图像、模糊图像学习生成更加逼真的模糊图像,再使用它指导后者如何从模糊图像生成更好的清晰图像.光流^[154]以及序列信息^[155]也被通过不同的方式融入到了去模糊任务中.

4.4 倍增多通道像素级回归

进一步地,当任务的输出在通道数上与输入一致,但是尺寸上超出了输入信息的尺寸时,我们将这样的任务归为倍增多通道像素级回归.用 m 和 n 分别表示输出图像与输入图像相比在宽度和高度上的倍增比例,那么可以用如下公式来表述倍增多通道像素级回归任务的优化目标:

$$\min \left[- \sum_{i=1}^D \sum_{m=1}^M \sum_{n=1}^N \iint p_{\beta}(x_{ij}) \log(q_{\beta}(x_{ij})) dp_{\beta} dq_{\beta} \right], \quad x_{ij} \in Y_{\beta}, Y_{\beta} \in B_{\beta}. \quad (14)$$

最典型的倍增多通道像素级回归任务就是图像超分辨任务.这里将以图像超分辨任务为例,对倍增多通道像素级回归任务作简要的介绍.

通常来说,我们使用一张图像中所包含的总的像素点数量来衡量其分辨率,它反映了相应图像中包含的数据量.在不考虑其他因素的前提下,一张图像的分辨率越高就意味着它的像素点数越多、数据量越大,进而表示图像能够展现出的细节信息更丰富、内容信息更有价值.但是在实际的图像采集过程中,由于硬件设备条件与成像环境的限制,往往会影响到最终采集到的图像质量,导致只能获取到分辨率较低的图像数据.想要解决这个问题获取到高质量的图像,最直接的方式是升级成像硬件设备,改善成像环境.但是硬件层面传感器技术的发展限制了成像设备的发展,而且硬件与成像环境的改善成本较高,普适性差.受限与硬件层面提升图像分辨率的高成本,有研究者提出基于软件的技术来提升现有图像的分辨率.此类基于现有低分辨率图像利用软件技术生成相应的高分辨率图像的算法被称为超分辨(super resolution, SR)算法.

为了从软件层面从低分辨率图像获取高分辨率图像,早在 1984 年就有研究者提出图像超分辨算法,尝试从算法层面配合已有的信号处理方法重构图像采集过程中的图像劣化过程,从而实现复原增强图像高频段细节信息,提升图像分辨率.经过多年的发展,图像超分辨技术随后演化出基于插值、基于重建和基于学习这 3 种不同的解决问题的路线.其中基于插值的方法在早期就得到了较大范围的应用.插值方法实现超分辨的思路是将其视为一个新像素构建问题,在原有低分辨率图像的像素间填充新的像素而得到高分辨率图像(如果使用插值得到的像素替代原本图像相应区域的像素,则可以实现基于插值的图像缩小).填充使用的新像素点是通过将相应像素插入位置附近的离散像素值拟合插值为连续值,然后在拟合出的连续值上重新采样得到新像素点的离散像素值.依据生成新像素值时所使用的原图像素区域和离散数据连续化时的算法不同,插值方法可以分为最近邻插值^[156]、双线性插值^[157]和双立方插值^[158]等.这 3 种方法在插值过程中使用的原图像素量逐渐增多,相应的算法复杂度也逐渐增大.但是使用的像素点越多,意味着用于建模的上下文信息越多,也使得超分辨结果拥有更好的细节、更平滑的边缘纹理和更优异的视觉效果.基于重构的图像超分辨算法采用特定条件约束

的方法, 通过观测低分辨率图像的生成过程来建模图像的劣化过程, 接着使用新建的模型通过反向演变构建重建过程中的约束条件. 常见的基于重构的超分辨算法有凸集投影法^[159]、迭代反投影法^[160]和最大后验概率法^[161, 162]. 基于插值和基于重构的传统图像超分辨方法更多地关注图像底层分布信息和局部关联信息, 不涉及高层的语义信息, 而出现较晚的基于学习的方法, 尤其是基于深度学习的方法, 将图像的高层次语义信息引入到图像超分辨任务中, 提供从全局到像素级的语义指导, 为该领域未来的发展埋下了铺垫. 基于学习的方法通过大量数据的训练拟合低分辨率与高分辨率图片内部隐含的数据关联. 这种关联的构建通常需要依赖于成组的低 - 高分辨率图像对 (LR-HR pairs), 也就是我们常说的数据集. 常见的方法包含基于流形学习^[163]、稀疏表示^[164, 165]以及基于深度神经网络的学习方法.

传统机器学习方法在图像超分辨领域的应用推动研究进程向前迈进了一小步. 然而, 基于传统机器学习方法的图像超分辨算法需要人工手动构建诸如特征模型、特征选择、训练参数等诸多和学习效果紧密相关的技术要点, 这些关键参数的设置很多都是基于经验的设定, 很大程度上限制了方法的拓展性和泛化能力. 随着近年来硬件设备水平的飞速发展, 深度学习方法在各个领域都得到了广泛的应用, 图像超分辨领域自然也从中受益. 得益于深度学习的特性, 人们不需要再过度关注特征的结构设计和过多的参数设定, 卷积网络丰富的参数集合可以自动地学习到理想的特征提取以及利用方式. 早期的尝试结合了深度学习方法和传统方法, 例如 SRCNN^[166] 首先基于传统的插值方法直接生成符合预期尺寸的粗糙中间图像, 紧接着通过简单的卷积网络学习从粗糙中间图像到精细的高分辨率图像之间的映射关系. 作为神经网络在超分辨领域的初步尝试, SRCNN 迈出了重要的一步, 但是依然存在很多有待改进的地方. 比如神经网络的输入输出都是尺寸较大的图像, 导致计算量急剧增大; LR 图像和 HR 图像的相似性过高, 重复信息太多不利于映射的学习. VDSR^[167] 则致力于解决上面提到的第 2 个问题, 不同于之前学习 LR 到 HR 映射的思路, VDSR 尝试学习从 LR 到残差图像之间的映射. 这里的残差是指输入图像和高分辨率图像之间的高频细节差异. 同时它还加深了网络的深度, 提升了网络的感受野. DnCNN^[168] 更进一步直接只学习 LR 图像到高频残差的映射, 网络最终的输出就是高频残差. IRCNN^[169] 则重点关注小尺寸训练样本的学习以及卷积过程中的边界填充问题, 以减弱由卷积运算引起的边界伪影, 同时它还是一个多任务 (multi-task) 解决方案, 作者同时将其应用于去噪和去模糊任务. FSRCNN^[170] 作为 SRCNN 的改进工作, 去除了之前工作中的插值步骤, 直接使用低分辨率图像输入到神经网络中, 然后在网络中依靠反卷积提升图像的尺寸, 在网络深度增加的情况下依然提升了模型的运算速度. 前面的工作在使用卷积网络时都是直接使用现有的网络, EDSR^[171] 则尝试从网络底层结构来探究网络各个成分在图像超分辨任务中的有效性, 作者重点设计实验验证了批量正则化层 (batch normalization, BN) 带来的不好的效果, 并指出 BN 对一个批次内的图像特征进行正则化破坏了图像原本的色彩、对比度、亮度分布, 而超分辨任务要求输出的高分辨图像和输入图像保持一致的分布, 仅仅提升分辨率和图像细节, 所以去除网络中的 BN 层可以大幅增加性能表现, 同时减少计算资源的消耗. 近期也有工作通过对抗生成网络进行超分辨的尝试, SRGAN^[172] 结合对抗生成网络中的对抗损失、像素级别相似度的 MSE 损失和自行定义的基于高层图像表征的感知相似性度量, 使用多任务聚合损失进行网络参数的迭代优化. EnhanceNet^[173] 则集成了基于网络中间层特征表示上定义的感知损失、匹配 LR 和 HR 图像纹理相似度的纹理匹配损失和对抗生成网络的对抗损失联合优化网络参数.

4.5 零输入多通道像素级回归

不管是分割、人群计数还是去模糊, 都是人们在尝试理解现有的图像, 尝试从现有的图像中提取

图 10 常见的图像生成方法
Figure 10 Common image generation methods

所需的信息. 自然, 我们对世界的探究欲不会只满足于利用现有的信息, 我们还在尝试创造. 数字化的世界中, 声音、图像或者说任何数据都是以离散的点的状态被存储、计算和应用. 以我们这里最关注的图像为例, 一副图像中每个像素点就是一个离散的采样点, 多个像素以一定的概率组合到一起就构成了最终的图像. 对于一系列图像, 我们假设每张图像中包含的所有像素点构成一个随机向量 X , 那么可以设定自然场景下的图像 (随机变量) 都服从于一个复杂的位置分布 $p(x)$. 如果我们能够通过观测现有的数据对这个未知的分布进行建模, 在估计出具体的分布之后我们便能够基于这个分布来生成 (采样) 一些符合这个分布的随机变量, 也就完成了图像的生成. 在图像生成过程中, 使用随机噪声信息或者简单的条件约束信息作为输入, 而期望的输出则是大尺寸的多通道图像. 这样的任务我们称为零输入多通道像素级回归任务. 由于零输入多通道像素级回归任务中没有真实的目标空间抽样样本作为标签, 我们也无从获取目标域的真实分布情况, 因此研究过程中使用基于学习到的预估目标空间 B 的密度分布 p_β^B 作为优化时参考的目标值. 此时式 (4) 所定义的优化目标转变为

$$\min \left[- \sum_{i=1}^D \sum_{j=1}^M \sum_{k=1}^N \iint p_\beta^B(x_{ij}) \log(q_\beta(x_{ij})) dp_\beta^B dq_\beta \right], \quad x_{ij} \in \hat{Y}_\beta, \quad Y_\beta \in \hat{B}_\beta, \quad (15)$$

相应的, $\hat{B}_\beta$ 和 $\hat{Y}_\beta$ 分别表示预估的目标域和在预估目标域中抽样出来的样本.

至此, 图像生成任务被解析为对现有数据分布的估计. 基于对未知分布建模时的方法, 图像生成任务可以分为基于显式概率密度估计的方法和基于隐式概率密度估计的方法. 图 10 展示了常见的图像生成方法, 随后的段落会对展示的方法进行详细的介绍.

显式概率密度估计是指在概率分布拟合过程中明确地定义出需要优化的概率模型, 然后基于现有的数据和构建的模型优化对未知分布的最大似然. PixelRNN^[174] 在学习阶段通过链式法则将最大似然函数分解为一维分布的乘积, 然后使用循环神经网络拟合定义分布. 在生成阶段则是从角落处的像素开始, 基于学习到的分布序列化地生成未知的每一个像素, 这样的生成过程速度非常慢. 为了提升图像的生成速度, 作者又提出了 PixelCNN^[175], 使用相同的分布建模方法, 但是在生成阶段改为使用卷积神经网络生成某个像素周围块状范围内其他像素的值, 这极大地加快了生成的速度. 这种直接对当前分布建模的方法带来的最大好处是最后得到的概率分布可以被用作一个很好的评估度量, 通过计算生成数据的似然衡量生成数据的质量. 但是由于真实数据极具多样复杂的特点, 想要直接构建出真实的分布模型难度极大. 于是基于变分自编码器 (variational auto-encoder, VAE)^[176] 的方法提出引入一个隐空间, 并假定隐空间服从一些常见的分布 (如正态分布和泊松分布), 然后通过大量数据训练学习从隐空间分布到真实数据分布的映射, 在数据生成阶段就在隐空间进行随机的抽样, 然后通过学习到

的映射将抽样到的数据转换到想要的分布下. 条件变分自编码器 (conditional variational auto-encoder, CVAE) [177] 则通过条件对数似然 (conditional log-likelihood) 完成在给定条件下生成符合学习到的分布的方法.

隐式的概率密度估计方法则不关注具体的分布衡量方式的构建, 这种方法解决问题的思路是, 在没有合适的度量方法的条件下, 通过学习的方式直接拟合出想要的度量方法. 生成对抗网络 (GAN) [178] 便是这一思路的典型方法. 它的核心思路是基于博弈论的思想, 首先设计用于生成数据的生成器和一个用来判别数据来源的判别器, 然后通过博弈的过程不断优化提升两个对抗模块的能力, 使得生成器可以生成更加接近真实分布的图片, 判别器对生成图片和真实图片的识别也越来越敏锐. 初始对抗生成网络的思路清晰, 效果也略好于 VAE 方法, 但是存在训练难度大、生成器和判别器训练不同步、对抗式训练容易出现崩溃状况等问题. WGAN [179, 180] 首先在数学的层面上通过严谨的推导证明分析原始的 GAN 存在的问题, 然后通过生成样本和真实样本中手动加入噪声使得两个分布产生重叠区域, 用来缓解训练不稳定问题, 同时在算法流程上提出了损失函数不取对数、判别器参数截断和不使用基于动量的优化器算法等改进. WGAN-GP [181] 则在 WGAN 基础上进一步引入梯度惩罚使得训练过程更加稳定和更快收敛. LSGAN [182] 则使用最小二乘损失替代了初始的交叉熵损失函数, 增加了模型对分布边缘样本的关注, 提升了训练过程的稳定性和生成器质量. DCGAN [183] 将更强有力的深度神经网络引入对抗生成模型中, 替代原本的多层感知机, 凭借卷积神经网络更强的拟合能力, 有效提升了图像生成的能力. 初始的对抗生成网络只能通过随机的噪声输入生成图像, 也就是说输入输出都是随机的, 不能够被人为地控制, 在最终结果被生成出来之前没有人知道会生成出怎样的图像. 这样的生成方式自然不能满足人们的需求, 所以条件对抗生成网络 (conditional generative adversarial nets, CGAN) [184] 提出在给定条件下生成数据的方法, 能够对生成数据的过程进行干预, 从而实现按照自己的想法生成特定类型数据的目标. SAGAN [185] 基于注意力机制放开卷积网络的局部感受野的限制, 利用全局信息来辅助提升生成图像的细节信息. BigGan [186] 基于 SAGAN 的结构使用超大规模的集成 TPU 和高达 2048 的批大小 (batch size), 训练出大型的生成网络, 通过实际操作证明大规模的训练数据为生成网络的性能带来了有力的提升.

也有人尝试将显式和隐式的方法相结合, 探究更加有力的图像生成方法. 如 VAE-GAN [187] 使用完整的变分自编码器作为对抗生成网络的生成器部分, 然后同时使用 VAE 和 GAN 的方法优化更新生成器的参数以提升方法的性能. BiGan [188] 则是将变分自编码器的编码器、解码器的输入输出分别打包为成对的图像-特征组合送入判别器中, 判别器的工作则从原来的判断图像真假转换为判断输入的图像-特征组合的来源是编码器还是解码器. BigBiGan [189] 则集成了 BiGan 和 BigGan 的思路, 基于双向的生成过程训练大规模对抗网络, 改进后的联合判别器性能更稳定, 最终得到了超越 BigGan 的性能表现.

4.6 像素级语义回归小结

本节延续第 3 节的规划思路, 在拥有距离度量定义的连续目标空间中讨论了逐像素的语义理解任务, 即像素级语义回归任务. 不同的是本节介绍的任务输入信息量不再保持一致, 输出信息量也不尽相同, 因此本节依据输入数据与输出数据信息量之比逐渐减小的线索展开探讨. 该比例的减小反映了目标域相对于输入域的熵增, 这也意味着任务难度逐步提升. 图 11 展示了本节提到的 5 种像素级语义回归任务的发展历程, 选取了一些关键的算法作为例子展现在发展时间线上, 方便读者配合本文查阅资料快速了解相关研究工作.

篇幅所限, 本文没有进行详细的代码实现讨论, 有需要的读者可以参考相关的代码开源网站, 例如

图 11 像素级语义回归的发展
Figure 11 Development of pixel level semantic regression

人群计数⁶⁾、去模糊⁷⁾、风格迁移⁸⁾、超分辨率⁹⁾、图像生成¹⁰⁾。

5 像素级语义理解的难点和未来发展方向

从更宏观的角度看, 上述所有任务都是在建立从一个域到另一个域的映射. 我们称输入信息来自源域 $\mathbb{D}_S$, 输出信息所在的域为目标域 $\mathbb{D}_T$. 像素级语义理解的特别之处在于, 源域和目标域在尺寸上是一致的, 甚至目标域的尺寸会大于源域. 不同的子任务其源域不同 (其实本文涉及到的任务也可以说是来自相同的源域, 只是分布不同), 想要求得的目标域也不同. 不同的方法和技术也都是在尝试寻求更加准确有力的映射关系. 从同尺寸域到域的映射来看这两大类任务可以共享很多方法, 同时也共同面临着相似的挑战, 本节将重点从数据和算力两方面进行分析.

5.1 数据挑战

当前人工智能领域中使用较多的深度学习方法本身是一个严重依赖数据的研究方法. 超大型数据集的出现往往可以推动某个甚至整个机器学习领域的进展, ImageNet 数据集就是其中之一. 研究者们基于它的图像分类任务设计了众多的网络结构, 并且 ImageNet 上预训练的模型参数也被应用于各个领域的应用网络中. 像素级语义分类^[190~196]和回归^[197]的发展也离不开相关领域数据集的发布.

通常情况下数据集的发布都需要带有精准的标注结果以供研究者训练以及测试, 表 3 展示了一些常见计算机视觉数据集中任务标注一张图像所需要的大致时间. 数据集标注和构建本身就是一个费时费力的工作, 而像素级语义理解任务的标注成本又是远超图像级视觉任务的.

这里以人群计数问题为例展开分析, 指出当前像素级语义理解面临的挑战与机遇. 随着硬件、算法和技术的不断发展, 算法效果在现有数据集上的表现被快速地刷新. 以 ShanghaiTech Part B 这个

6) <https://github.com/gjy3035/Awesome-Crowd-Counting>.

7) <https://github.com/subeeshvasu/Awesome-Deblurring>.

8) <https://github.com/ycjing/Neural-Style-Transfer-Papers>.

9) <https://github.com/ChaofWang/Awesome-Super-Resolution>.

10) <https://github.com/nightrome/really-awesome-gan>.

表 3 不同任务标注一张图像大致所需平均时间

Table 3 The average time it takes to label an image for different tasks

	Image classification	Object detection	Instance segmentation	Crowd counting	Semantic segmentation
Labeling time	Few seconds	Few minutes	Dozens of minutes	Few hours	Dozens of hours

图 12 (网络版彩图) ShanghaiTech Part B 数据集上的衡量指标发展

Figure 12 (Color online) Test results on ShanghaiTech Part B dataset

数据集为例, 从图 12 可以看到, 短短几年内它的衡量指标平均绝对误差 (mean average error, MAE) 和均方误差 (mean square error, MSE) 被分别从 26.4 和 41.3 刷新到了 6.4 和 11.3 (这两个误差越低代表算法表现越好). 因此单个数据集往往无法满足科研发展的需求, 需要不断更新有差异化的、有挑战性的数据集, 才能继续推动算法的不断进步.

表 4 [111, 112, 197~205] 展示了一些常用的人群计数数据集以及它们对应的统计信息. 可以看到随着时间的推进, 新的数据集不断出现, 数据集使用图片的分辨率、数据集包含的人数都呈现一个稳步提升的趋势. 新型大型数据集的构建是推动领域发展必不可少的要素, 但是越来越大的数据集规模也急剧增加了数据标注的成本, 消耗大量的人力物力. 除去成本的考量, 人工标注数据的方法还存在一个很严重的问题, 即无法保证标注的完全准确性, 同一个人不同时间标注同一张图可能都会产生 10% 以上的差异, 尤其是在像素级语义理解数据标注中, 差异可能更大. 这个问题很容易被忽视, 其实它从一定程度上限制了一个数据集能达到的效果上限, 也变相地限制了算法的发展上限.

5.1.1 虚拟数据

为了克服上面提到的人工数据集带来的限制和挑战, 很多研究者也从各种不同的层面尝试去缓解相应的问题. 部分研究者尝试使用逼真的虚拟软件去生成类似于现实场景下的场景图像, 并且使用虚拟场景渲染时的信息直接生成对应的像素级标注信息. 例如 Ros 等 [206] 就使用 Unity 开发的专门软件生成场景和相应的语义分割标注结果. Richter 等 [207] 则选择使用现有的游戏软件生成数据场景和标注信息. 表 4 中的 GCC [200] 数据集其实是一个基于 GTAV 游戏生成的人群计数数据集, 通过简单的对比就可以发现, 它在图像数量、总人数上是远超现有的真实数据集的. 得益于计算机生成的全自动以及完美精准的特性, 这样的数据可以较容易地被大量生成、自动标注, 在数据量和标注质量上远超人工数据集.

表 4 一些常见人群计数数据集的统计数据
Table 4 Statistics of common crowd counting datasets

Dataset	Average resolution	Sample quantity	Total crowd number	Average crowd number	Release year
NWPU-Crowd ^[197]	2191×3209	5109	2133375	418	2020
JHU-CROWD++ ^[198]	1430×910	4372	1515005	346	2020
Fudan-ShanghaiTech ^[199]	1080×1920	15000	394081	27	2019
GCC ^[200, 201]	1080×1920	15211	7625843	501	2019
UCF-QNRF ^[202]	2013×2902	1535	1251642	815	2018
CityUHK-X ^[203]	384×512	3191	106783	33	2017
ShanghaiTech Part B ^[112]	768×1024	716	88488	123	2016
WorldExpo'10 ^[204]	576×720	3980	199923	50	2015
Mall ^[111]	480×640	2000	62325	31	2012
UCSD ^[205]	158×238	2000	49885	25	2008

自然, 虚拟数据是不能直接用到真实数据的预测上的, 而人们的最终目的还是解决真实场景中的视觉任务. 于是大量研究者针对从虚拟域到真实域的转换进行研究工作. 虚拟数据集的输入我们称为虚拟源域 $\mathbb{D}_S^{\text{sync}}$, 对应的语义级标注信息我们称为虚拟目标域 $\mathbb{D}_T^{\text{sync}}$. 从 $\mathbb{D}_S^{\text{sync}}$ 到 $\mathbb{D}_T^{\text{sync}}$ 的映射与从 $\mathbb{D}_S$ 到 $\mathbb{D}_T$ 的映射有相似的地方也有区别. 部分研究者尝试找出两个映射间共通的关系, 提取源域中最适合计算目标域的特征, 使用虚拟数据辅助训练特征提取器^[200, 208~210]. 也有研究者尝试将 $\mathbb{D}_S^{\text{sync}}$ 通过风格迁移技术转化为类似 $\mathbb{D}_S$ 的数据然后和真实数据混合在一起作为训练数据^[201, 211].

5.1.2 算法优化

除了对数据本身的拓展, 另一个缓解数据紧张问题的思路就是通过升级现有算法, 对现有的数据进行更高效的利用和发掘, 对于这个思路的未来发展方向归纳总结如下.

(1) **多源数据融合.** 在深度神经网络的加持下, 之前我们讨论过的所有任务基本都获得了有效的进步与提升, 但是各种方法性能表现越好, 越接近数据集本身的上限, 算法的性能提升难度也越大, 越容易陷于局部最优解. 这是由于算法学习的是数据集的特异特征 (即只在当前数据集数据内部存在的特征), 学习到的特异特征越多, 越容易在当前数据集上过拟合.

为了缓解单个数据集上的过拟合问题, 提升数据集的性能上限, 很多研究者尝试使用来自多源的数据信息融合思路来解决现存的视觉任务. 比如在图像去模糊任务中, 由于很多的模糊图像是由运动造成的, 而现有的图像采集终端自带加速度传感器、运动传感器以及方位传感器, 所以就可以在去模糊的过程中参考图像采集设备在采集图像一瞬间的运动状态和姿态信息. 这样的运动信息可以有效辅助运动模糊的模糊过程建模, 从而达到辅助去模糊的效果. 此外, 如表 4 中的 NWPU-Crowd 数据集不仅标注了人头点位置, 还标注了包含人头的边界框. Fudan-ShanghaiTech 数据集则是以视频的形式给出了序列化的帧间信息. 也有用于图像分割的数据集连同分割数据给出了相应的深度信息、红外信息, 这些额外的多传感器源数据包含了图像传感器不能采集获取到的大量数据, 可以为相应的计算机视觉任务提供有力的帮助, 就如同人体本身也不单单依赖于视觉本体, 触觉、嗅觉、听觉也是感知世界的重要方式.

(2) **目标函数和度量方法的改进.** 目标函数是机器学习算法中优化参数以及结构的依据, 以深度学习为例, 它的目标函数是一个人为设定的损失函数 (loss function), 算法的直接优化依据就是通过参

表 5 常见的像素级语义理解任务及其常用的目标函数和度量方法

Table 5 Common pixel level semantic understanding tasks with their target functions and measurements			
	Semantic segmentation	Crowd counting	Super resolution
Target function (common)	Cross entropy loss	Mean square error	Mean square error
Measurements (common)	Mean intersection over union	Mean square error	Peak signal-to-noise ratio

数的迭代学习最小化给定的损失函数. 度量方法则是用来衡量程序算法输出结果质量的一个指标.

由于机器学习算法的核心在于优化指定的目标函数, 所以目标函数的设计在各种任务中都至关重要. 例如常见的像素级语义分类任务的损失函数是交叉熵损失, 而像素级语义回归任务常用的损失函数是均方误差函数, 这类基本损失函数在理想情况下可以起到不错的效果. 但是真实世界中的数据总是会出现不理想的情况, 例如数据类别的不平衡问题就会导致模型在某些特定类别上的性能不理想. 为了解决这样的问题, 很多研究者都会尝试设计新的目标函数或者增加用于辅助的目标函数以提升模型的性能.

就像人们衡量美的标准并不统一, 对于很多像素级语义理解任务来说, 衡量结果表现的标准也不是一成不变的. 例如, 边缘检测需要界定边缘限度, 并非无论大小粗细地输出所有边缘就是最好的; 语义分割可以通过计算平均像素准确率衡量分割结果, 但是这个结果并不能完全代表算法分割效果的好坏, 某些算法虽然在图中面积较大的区域 (比如背景) 取得了很高的准确率 (导致平均像素准确率较高), 但是可能在一些小面积的重要图像元素上表现一团糟. 所以人们提出了新的语义分割指标平均交并比 (mean intersection over union, mIoU) 来替代平均像素准确率. 不同的度量方法自然也会引领着机器学习算法向不同方向优化. 随着任务需求的发展, 度量方法也在逐步地进步与改变.

此外, 由于目标函数直接参与优化的过程, 所以算法要求它具有一些必备的特性 (如深度学习算法中要求目标函数必须是可导的, 这样才能通过梯度回传的方式更新网络参数). 然而很多任务的度量方法是不具有这些特性的, 任务中可以使用的目标函数又不能很好地衡量输出结果的表现. 表 5 就展示了常见的像素级语义理解任务及其常用的目标函数和度量方法, 由于篇幅限制这里只展示了 3 个任务, 可以看到只有一个任务在通常情况下是具有统一的目标函数和度量方法的. 这样的不一致性就导致了一个很严重的问题, 那就是我们算法实际优化的目标不是我们想要得到的结果! 虽然说大多数情况下目标函数得到优化也会提升算法在最终度量上的表现, 但是如果要是能够从算法初始设计的层面统一目标函数与度量方法, 一定能够使得算法的表现提升一个台阶.

(3) 弱监督/无监督解决方案. 数据的量始终是有限的, 然而算法需要面对的场景以及任务是无穷的, 所以除了从数据的角度出发增加数据量的思路以外, 有很多研究者致力于基于小样本的学习方法, 通过使用少量标注的数据, 或者配合部分不带标注的数据进行学习完成相应的计算机视觉任务, 尽量在减少数据依赖、人工消耗的情况下带来比较好的性能表现.

弱监督的方法有多种多样的解决方案, 只要不是完全的全监督方法就可以被称为弱监督的方法. 有的方法使用低等级标注信息进行高等级视觉任务, 比如只使用图像类别信息辅助学习完成分割任务. 有的方法基于少样本学习 (few-shot learning), 训练过程中对于每个类别只使用一部分的标注数据, 而将剩余的未标注数据通过聚类的方法聚合到相应的类别中进行学习. 有的方法基于元学习, 不需要在学习过程中涉及到所有的类别和场景, 学习到的不是直接的结果映射, 而是如何在新的类别或场景中快速学习. 还有方法使用大量虚拟数据完成对网络的训练, 然后在不同的真实数据集上只使用寥寥几对带标注的数据, 以较小的代价将网络参数迁移到适用于真实数据集的状态.

有更加激进的学者不满足于弱监督的解决方案, 甚至提出了无监督的解决方案 [212~214]. 当然现

有的大多数无监督解决方案无法解决像素级语义理解任务,研究领域还主要处在分类任务阶段.唯一一种像素级语义理解任务无监督的解决方案,就是只使用自动生成的虚拟数据指导模型学习,然后将该模型应用在真实场景下的数据中(这一过程也被称为迁移的过程).

当然,弱监督和无监督的解决方案目前在性能上远远落后于全监督的解决方案,所以这个方向还有很大的发展空间.

(4) 网络结构优化.除了数据和整体任务目标级别的算法优化思路,对于当前最常见的基于深度学习的像素级语义理解任务解决思路来说,网络结构设计也是提升算法性能的重要因素.最早期的神经网络专注于图像分类任务,并不能方便地直接拓展到其他任务中.得益于全卷积网络在结构层面抛弃了全连接层,后续像素级语义理解方法百花齐放,这也反映了网络结构设计的重要意义.依照网络结构设计发展的时间节点来看,基于深度学习的像素级语义理解方法在前期算法整体表现严重依赖于骨干网络性能,使用新设计的强力骨干网络简单地替换最终的输出结构即可大幅度提升算法表现.进而在骨干网络的发展陷入平台期后,研究者们将研究重心调到面向不同具体任务的特异化网络结构设计,将特定的像素级语义理解算法与最终的目标任务结合,从网络初始设计层面就考虑最终任务需求,针对不同的目标特征设计专用的网络模块与结构.更进一步,有学者疲于特异化网络结构的搭建与设计工作,提出从前期设计的多种特异化网络结构中构建网络结构搜索空间,然后基于网络架构搜索的思想在预设的网络结构空间中通过学习自动化地搜索最适合当前任务的网络结构设计.这种自动化的网络结构生成方法可以有效提升像素级语义理解网络结构设计的效率,但是就当前的实现来说,工程量巨大、实现难度高、搜索空间构建不合理等多种因素阻碍了其进一步的发展.在未来,合理的搜索空间构建以及廉价的架构搜索实现想必能够极大地提升研究者们探索高性能网络结构的效率.

5.2 算力挑战

除了数据,当代深度学习的飞速发展离不开硬件算力的大幅提升.但是目前高算力的计算设备都是功耗体积都很大的 GPU 设备,还都是在大型计算平台上才能使用的设备.图 13^[215,216]中的绿色三角形展示了一些常见的语义分割方法在 Cityscapes 数据集的测试集上的测试结果以及相应的推理速度.这里的推理速度是基于一个桌面端 NVIDIA TITAN X GPU (功耗高达 250W) 计算得到的.可以看到即便是在如此强劲的设备上,很多精度较高的神经网络的推理速度都无法达到 2 帧每秒,甚至远低于 1 帧每秒,这样的速度完全无法匹配通常图像采集设备的速度 (30 帧每秒).因此,语义级像素理解任务的实时应用遥遥无期.计算机视觉的实际应用中,大多数情况下也完全无法支持这样的大型设备进行计算.

解决这个问题思路的一个思路是移动终端采集数据,通过网络传输到计算中心完成计算,最后再将结果返回到移动终端,但是这样的方案受限于高成本的计算中心以及网络的约束.另一个思路是使当前算法能够运行在低功耗移动平台中,这个思路又分为软件和硬件两个层面的解决方案,后续段落从这两个层面分别探讨.

5.2.1 软件层面

软件层面的优化主要从两个角度出发,一个是在解决问题之初,就将优化网络运行速度纳入考虑,从网络结构和卷积层内部使用轻量化的结构提升网络的运行速度.另一个角度是利用通用的压缩技术压缩已有的模型,尽量在不影响模型性能的情况下压缩模型体积,减少计算消耗.

(1) 网络结构优化.图 13 中的红色圆点展示了当前常见的骨干网络的计算量以及它们在 ImageNet 分类数据集上的 Top-1 准确率,可以清晰地看到网络准确率和相应的计算量呈现一个显

图 13 (网络版彩图) 常见骨干网络的计算量和 ImageNet 上的 Top-1 准确率^[215] (红色圆点表示) 以及常见语义分割方法运算速度和准确度^[216] (绿色三角形表示)

Figure 13 (Color online) Amount of calculation and Top-1 accuracy on ImageNet of common backbone nets (in red dots); speed and accuracy of common semantic segmentation methods (in green triangles)

著的正相关趋势. 为了在保持准确率的同时降低网络的计算量, 很多学者在网络结构优化层面作出努力, 致力于研究轻量化的网络结构^[42, 43, 217]. 网络结构层级的优化是尝试去除网络结构设计中带来的冗余信息, 但是神经网络中携带的冗余信息远不止在网络结构设计上. 也有研究者尝试优化卷积内部结构, 例如 GhostNet^[218] 通过设定特征包含信息差别的衡量方法发现通常的卷积特征在通道维度上会存在很多包含相似信息的特征块, 这些相似的特征块可以互相之间通过简单的线性变换得到, 而不需要动用包含大量运算的卷积, 他们从这个发现出发提出通过去除卷积特征冗余降低神经网络的运算量. 而 AdderNet^[219] 则提出基于加法的网络和参数学习算法, 构建完整的只包含加法运算的网络后向传播以及更新迭代参数的算法. 当然参数上的减少在带来运算速度提升的同时势必也会带来性能表现上的损失, 依据不同的任务需求, 当前很多轻量化的网络结构设计在移动终端设备上得到了广泛的应用. 而诸如 GhostNet 和 AdderNet 这样的设计目前因为性能上的差距还无法取得实际的应用, 但是它们的存在是对未来网络运算优化的一种探索, 提供了一些思路上的可行方向.

(2) 网络压缩. 网络结构层面的优化通常面对的是特定的任务, 相应的结构迁移到不同的任务中不一定会取得同样理想的结果. 于是研究者将目光转向了现有的表现优秀的模型, 尝试通过压缩已有的训练好的模型获得同时拥有不错的性能表现和速度表现的模型. 一种努力的方向是非通用化的模型压缩方法, 针对一个已经训练达到预期结果的模型, 通过逐步拆除已有模型的不同组成部分, 然后在特定的验证集合上测试改动后网络的性能表现, 最终在可接受范围内抛弃一定的网络结构来达到减少网络参数、提升运算速度的目的. 这种压缩方法通常压缩效果较好, 但是测试验证过程繁杂, 而且完全不具备通用性, 对于任何的网路都需要重复复杂的探索测试过程. 所以这类方法通常被用于工业界中特定任务的最终优化过程.

另一种思路致力于通用化的网络压缩方法, 常见的方法是神经网络量化. 通常的网络参数的存储使用的是 32 位或 64 位的浮点型数据类型, 这样的数据类型存储小数精度较高, 但是十分占用存储空

间而且运算时的内存和计算量开销也很大. 一些研究验证发现少量的数据位数减少不会给模型的性能表现带来过大的影响, 所以可以通过降低现有模型的参数位数来对已经训练好的模型进行压缩. 这种压缩的过程被称为神经网络参数的量化^[220, 221], 在尽可能少损失网络性能的情况下降低参数精度从而降低计算量, 减少内存开销. 在极端情况下甚至出现了二值化神经网络^[222], 即只使用两个值、一位存储来表示网络的参数, 可想而知, 这样程度的量化会给模型精度带来极大的损失, 但是它的压缩效果也是其他方法无可比拟的.

5.2.2 硬件层面

使用低功耗移动终端设备在靠近物或数据源头的一侧完成计算任务的计算方式被称为边缘计算. 由于受到设备体积以及功耗的限制, 移动终端的算力普遍偏弱, 而且通常情况下, 设备内置的都是基于 CPU 的计算核心, 这类核心的重要特征是计算单元数量少, 但是核心主频高, 适用于常见的计算问题, 但是在深度学习中最重要举证计算中它的推理速度远赶不上主频低一些, 但是核心数多很多的 GPU 设备. 基于现行算法对矩阵计算的极度依赖, 很多硬件常在设计计算核心时单独将矩阵计算能力纳入考虑. 比如谷歌推出专为神经网络中常用的乘积累加运算执行优化的 TPU 设备, 通过多种最低层级别的优化有效提升了低功耗平台上的矩阵计算能力. 华为也在麒麟处理器中加入神经网络处理单元 (neural-network processing units, NPU) 专用于矩阵计算. 这一类的解决方案属于通用算力提升的解决方案, 提供的是一个更有力的矩阵计算平台, 其他任何算法都可以在该平台上运行来提升速度. 此外还有诸如基于 FPGA^[223] 或者光神经网络^[224] 的专用解决方案. 基于 FPGA 的方法是使用底层逻辑运算电路来实现已有的特定运算, 从底层电路出发的实现极大加速了推理运算的运行速度. 光神经网络则尝试通过光学滤波器的形式在光学硬件层面上达到模拟卷积核卷积计算的效果, 这样的基于光学器件的推理方式运算成本低, 且运算速度极快. 这类解决方案的思路是将现有的算法方案通过专用的硬件设备实现, 进而优化算法的推理速度. 但是因为这类方法的不可通用性, 相关的研究目前进展不充分.

5.3 后深度学习时代的思考

深度学习如火如荼的发展, 给人工智能领域带来诸多振奋人心的进步. 在进步带来欣喜之余, 我们也应该冷静下来思考这一切背后的因果, 思考浪潮之后的世界, 思考机器学习乃至人工智能的未来. 前文探讨了像素级语义理解未来面对的机遇与挑战, 当前的像素级语义理解任务发展得益于深度学习的发展, 但同时也受限于深度学习的发展. 数据和算力两大难点恰恰也是近年来有力推动人工智能以及深度学习发展的两大重要因素. 大数据时代的海量数据提供机遇的同时也带来了巨大的挑战. 在人工智能发展的路途中, 仍然面对的是崎岖不平的道路.

5.3.1 深度方法的不可解释性

可解释性是指在了解或者解决问题的过程中, 使用清晰的解决思路处理问题, 以人类能够理解的方式给出问题的解决办法和过程. 传统机器学习方法构建解决方案时一般会对问题进行完整的建模, 将待解决的问题转换为一个能够使用数学方法求出最优解的问题. 而这个建模的过程是基于研究者对问题的深入了解和详尽分析才能完成的. 不同的建模方法会直接影响到问题的最终求解, 大部分时候建模的方法决定了算法能够达到的最优水平. 与传统机器学习方法不同, 深度学习的方法通过搭建带有超大量参数的网络模型, 然后使用训练数据和梯度反向传播算法更新模型参数, 从而尝试得到能使损失函数最小化的模型参数值. 在这样的一个解决问题的流程下, 特征提取的工作和对特征利用的

图 14 (网络版彩图) 传统机器学习方法与深度学习方法解决问题思路上的差别

Figure 14 (Color online) Differences in problem-solving ideas between traditional machine learning methods and deep learning methods

工作被一体化, 统一由深度神经网络完成. 这时没有人为的干预和知识的指导, 模型进行自主的建模学习. 最终所有学习到的特征提取方法和特征使用方法被以参数的形式保存在模型内部. 我们无法了解到深度学习作出某个判断或决策的缘故, 只能给模型一个输入然后观测它的输出. 这样的一个过程被称作黑盒模型, 显然, 黑盒模型是不具有可解释性的. 如图 14 所示, 我们可以清晰地了解到传统机器学习方法作出判断的依据, 而无法从深度学习方法中获取这样直观的判断依据. 黑盒模型在特定的环境下可能会取得不错的结果, 但是源于模型的不可解释性, 缺乏详尽的科学原理支撑, 这一类型的方法在应用过程中始终受到限制.

也有研究者尝试打开模型, 对模型学习到的东西进行探究, 这其中最常见的方法是尝试通过可视化的方式展现模型学习到的参数和特征, 如图 15 所示^[225]. 作者通过编写特定的程序对已经训练好的神经网络参数及其提取的特征向量进行可视化得到这样的结果, 乍一看可视化结果中若隐若现的物体形状仿佛有力地增强了神经网络特征的可理解性和说服力. 但是冷静下来仔细一想, 首先我们使用可视化程序的目的是将特定的数据转换成我们能看得懂的样子, 再加上大脑的一点想象和泛化, 无法证明目前看到的就是原本的特征; 其次这样的可视化方法是以人类理解特征的方式去强制理解模型学习到的特征以及衡量方式, 然而人类衡量特征的方式并不一定能够有效地代表机器学习算法理解的特征. 所以诸如此类的可视化方法是对可解释深度学习的有效探索, 但是并不能从根本上解决深度学习方法的不可解释问题.

与深度学习不同, 利用知识图谱和基于贝叶斯、决策树等方法构建的机器学习算法往往在构建的过程中融入了大量的专业知识和数学原理, 这种基于知识构建的特征选择过程可以有效地抑制非目标语义特征的生成. 未来的机器学习方法应该兼具知识的方向指导和数据的优化驱动, 即以先验知识构建模型为导向, 利用大量数据完善模型构成, 协同并进地引领人工智能走上新的台阶.

5.3.2 人工智能的数据依赖

深度学习作为一种典型的依赖于数据的学习方法, 虽然在现存挑战上展现出了优异的成绩, 但是它自身依然存在许多缺陷. 其优异的性能表现在极大程度上依赖于高质量大数据, 以及在训练决策

图 15 (网络版彩图) 深度网络参数及特征的可视化 [225]

Figure 15 (Color online) Visualization of network parameters and features

层面的数据高度一致性. 在基于拥有独立同分布假设的封闭环境中, 深度学习方法可以通过大量数据堆砌自动地学习到封闭数据集内部有利于完成特定任务的特征. 这种自动学习特征构建的特性是它的优点同时也导致了它的缺点. 优点在于免去了人工构建特征的复杂过程, 自动生成具有较强表征能力的特征; 缺点在于全自主依赖于数据学习的方法缺乏背后专业知识和数学知识的指导, 可能会纯粹地学习到一些数据集内部与实际的语义特征不符、但却能够帮助完成特定任务的特征, 这样的特征只存在于训练所使用的数据集内部, 而不能真正表征特定的语义信息. 神经网络作为一个黑盒模型, 其中包含了大量的参数, 其中的参数更新完全基于每次数据输入计算损失函数得到的梯度反向传播迭代更新. 而在现如今大量的应用中, 特征的提取、分类利用过程被混为一谈, 迭代获取的参数集合既被用来当作特征提取器, 也被用来当作分类器和回归器. 这导致我们无法直观地判断特征提取器以及特征的有效性. 黑盒模型的可解释性将影响到在很多应用中模型的可依赖性. 深度模型对于训练时使用数据的过度依赖问题同时也会导致模型泛化能力差, 无法在同目标但是差异较大的场景中取得较好的效果. 同时一些研究结果也表明, 输入信息中叠加的极少量随机噪声就可能很轻易地破坏脆弱的神经网络, 这也就是神经网络攻击所研究的内容.

神经网络在设计初期致力于模仿人类大脑中密集连接的神经元结构, 但是训练的过程对数据和算力的依赖已经远超过正常人类学习过程所需的数据, 可以说现在已经处于一种和人类学习完全不同的状态之中. 不同于神经网络机械化的学习过程, 人类认知世界的过程可以通过一个新获得的例子自行构建完整的学习思路, 以举一反三的方式极大地拓展大脑的泛化能力. 同时面对未知的事物, 人类还具有极强的探索和自我学习能力. 现有深度学习算法乃至人工智能算法对这些方面的探索十分不充足, 总的来说对任务场景变换的适应性极其脆弱, 无法很好地在复杂环境中协调处理多个任务, 不能应对不断演变的具有较强博弈性质的任务. 诸如自监督学习、元学习、无监督学习等领域仍然存在大量待研究和改进的工作.

5.3.3 不可忽视的风险

无论是本文中提及的像素级语义理解算法, 或是其他的机器学习模式识别算法, 都不应该是一个独立于实际应用最终目标而孤立存在的“算法”. 我们在自己所在的领域努力耕耘的同时也更应当高瞻远瞩, 从问题产生的源头出发到最终需要解决的问题, 贯穿式地深入了解自己所面对的任务. 以解决最终问题为目标驱动科研, 而非以算法的迭代更新驱动科研. 以本文提及的语义分割和实例分割任务为例, 这两种任务当前多被应用于城市交通场景理解中, 最终贡献于车辆自动驾驶应用. 这些方法在当前的自动驾驶场景下是有效的, 但是设想一下在遥远的未来, 倘若交通工具不再是独立的存在, 新型传感器的出现使得道路和交通工具之间存在直接的交互, 基于整个路网直接指导全体交通工具的自动协调与驾驶. 这样一来当前的这些方法在自动驾驶领域的意义便大幅降低. 在科技高速发展、各个

领域新技术日新月异的当下, 这样的未来也许并不遥远, 因此科研人员应当对自己所在领域的任务保持极强的风险敏感性.

5.3.4 一些大胆设想

当前所讨论的所有方法和问题都是基于计算机信号处理, 即数字信号处理范畴下的任务. 我们身处一个依赖模拟信号而存在的世界, 但是计算机科学探索世界是通过基于数字传感器采集到的离散化数字信号实现的. 在数字信号数据的采集和量化过程中一定会造成数据的损失, 这其中包含了大量频域信息的损耗, 说不定就包含了人类探索智能化世界所需的重要信息. 而且当前基于数字信号处理的计算方式严重依赖于强有力的计算设备, 也就进而依赖于大量而稳定的能量供给. 那么, 有没有跨过数字信号直接在模拟信号域进行数据处理乃至“机器学习”的可能性? 在笔者的知识范畴内, 生物计算和光学计算是最接近于这一设想的两种非数字信号处理方法. 生物计算是指基于现有生物系统信息处理方式研发的用于特定目标的计算模式与计算设备. 它往往可以精细到分子尺度下的有序计算体系的构建, 生物计算系统的复杂度和对于能量的依赖性均远优于现有的数字信号处理体系. 但是通常的生物计算系统只能接受特定的生物信号作为输入, 很难直接处理现实世界的模拟信号数据. 而光学计算则是通过特定的光学元件搭建复杂的计算体系, 进一步直接接受模拟信号作为输入. 以光学神经网络 (optical neural network, ONN) 为例, 在这一开创性探索中, 研究者尝试开辟一种利用光学衍射、干涉等来加速或直接实现人工神经网络计算功能的光学计算方法, 从而实现将光学与人工智能两大领域紧密地融合. 相比于传统的神经网络, 光学神经网络具有高带宽、低能耗、近光速计算、高互联、内在并行处理等特性, 可显著减轻软硬件环境的计算负载, 为改进甚至替代传统神经网络提供了一种极具前景的思路. 这是一个极具新意而又大胆的尝试, 有力地展现了光学计算在实际应用中的丰富可能性. 基于光信号本身所展现的极强的信息负载能力与信息处理速度, 如果今后能够取得更多的进展, 甚至可以使用光信号替代电信号来处理包括图像、声音、文本等一切数据, 实现比如模拟世界这种在数字电信号时代看来遥不可及的设想. 摆脱算力限制、以光速完成计算甚至不依赖于能量供给的计算方式带来的智能世界将是怎样的一种存在, 让我们拭目以待.

6 结论

本文介绍了人工智能中的一大类任务, 即像素级语义理解的基本概念和发展历史, 强调了像素级语义理解任务的特性. 分别从任务输出数据量逐步增多和输入数据量逐渐减少的主线出发介绍了常见的像素级语义分类和像素级语义回归任务的发展状况和解决方案, 讨论了一些代表性的算法. 此外从数据和算力两个角度分析了当前像素级语义理解任务面临的巨大挑战, 并分别从虚拟数据构建、算法优化、软件层面和硬件层面给出了相应的解决方案以及后续值得研究和探索的方向. 最后针对深度学习对当前像素级语义理解乃至人工智能带来的机遇和挑战进行了探讨和总结.

参考文献

- 1 Li X L, Gong H G. A survey on big data systems. *Sci Sin Inform*, 2015, 45: 1–44 [李学龙, 龚海刚. 大数据系统综述. *中国科学: 信息科学*, 2015, 45: 1–44]
- 2 Li X L, Chen M L, Wang Q. Multiview-based group behavior analysis in optical image sequence. *Sci Sin Inform*, 2018, 48: 1227–1241 [李学龙, 陈穆林, 王琦. 光学影像序列中基于多视角聚类的群组行为分析. *中国科学: 信息科学*, 2018, 48: 1227–1241]

- 3 Li X L, Shi J H, Dong Y S, et al. A survey on scene image classification. *Sci Sin Inform*, 2015, 45: 827–848 [李学龙, 史建华, 董永生, 等. 场景图像分类技术综述. *中国科学: 信息科学*, 2015, 45: 827–848]
- 4 Wan J F, Yang J, Wang Z R, et al. Artificial intelligence for cloud-assisted smart factory. *IEEE Access*, 2018, 6: 55419–55430
- 5 Yuan Y, Lu Y W, Wang Q. Tracking as a whole: multi-target tracking by modeling group behavior with sequential detection. *IEEE Trans Intell Transp Syst*, 2017, 18: 3339–3349
- 6 Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 248–255
- 7 Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. *Proc IEEE*, 1998, 86: 2278–2324
- 8 Xiao H, Rasul K, Vollgraf R. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. 2017. *ArXiv:1708.07747*
- 9 Krizhevsky A. Learning Multiple Layers of Features From Tiny Images. Technical Report TR-2009. Toronto: University of Toronto, 2009
- 10 Kullback S. *Information Theory and Statistics*. New York: Dover Publications, 1997
- 11 Yan J Q, Tao D C, Tian C N, et al. Chinese text detection and location for images in multimedia messaging service. In: *Proceedings of IEEE International Conference on Systems, Man and Cybernetics, Istanbul*, 2010. 3896–3901
- 12 Liu G C, Li L L, Jiao L C, et al. Stacked fisher autoencoder for SAR change detection. *Pattern Recogn*, 2019, 96: 106971
- 13 Shen Y H, Ji R R, Wang C H, et al. Weakly supervised object detection via object-specific pixel gradient. *IEEE Trans Neural Netw Learn Syst*, 2018, 29: 5960–5970
- 14 Zhang S P, Lan X Y, Yao H X, et al. A biologically inspired appearance model for robust visual tracking. *IEEE Trans Neural Netw Learn Syst*, 2017, 28: 2357–2370
- 15 Aziz M A A, Niu J W, Zhao X K, et al. Efficient and robust learning for sustainable and reacquisition-enabled hand tracking. *IEEE Trans Cybern*, 2016, 46: 945–958
- 16 Li X L, Chen M L, Wang Q. Quantifying and detecting collective motion in crowd scenes. *IEEE Trans Image Process*, 2020, 29: 5571–5583
- 17 Huang W, Xiong Z T, Wang Q, et al. Kalm: key area localization mechanism for abnormality detection in musculoskeletal radiographs. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020. 1399–1403
- 18 Li X L, Yuan Z H, Wang Q. Unsupervised deep noise modeling for hyperspectral image change detection. *Remote Sens*, 2019, 11: 258
- 19 Cortes C, Vapnik V. Support-vector networks. *Mach Learn*, 1995, 20: 273–297
- 20 Lin Y, Lv F J, Zhu S H, et al. Large-scale image classification: fast feature extraction and svm training. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Providence*, 2011. 1689–1696
- 21 Breima L, Friedman J, Stone C, et al. *Classification and Regression Trees*. Boca Raton: CRC Press, 1984
- 22 Liu H, Cocea M, Ding W L. Decision tree learning based feature evaluation and selection for image classification. In: *Proceedings of International Conference on Machine Learning and Cybernetic, Singapore*, 2017. 2: 569–574
- 23 Wright R. *Logistic Regression*. Washington: American Psychological Association, 1995
- 24 Cheng Q, Varshney P K, Arora M K. Logistic regression for feature selection and soft classification of remote sensing data. *IEEE Geosci Remote Sens Lett*, 2006, 3: 491–494
- 25 Kononenko I. Semi-naive Bayesian classifier. In: *Proceedings of European Working Session on Learning*. Berlin: Springer, 1991. 206–219
- 26 Lee S, Crawford M M. Unsupervised multistage image classification using hierarchical clustering with a Bayesian similarity measure. *IEEE Trans Image Process*, 2005, 14: 312–320
- 27 Breiman L. Random forests. *Mach Learn*, 2001, 45: 5–32
- 28 Bosch A, Zisserman A, Munoz X. Image classification using random forests and ferns. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2007. 1–8
- 29 Krizhevsky A, Sutskever I, Hinton G. ImageNet classification with deep convolutional neural networks. In: *Proceedings of Advances in Neural Information Processing Systems*, 2012. 1097–1105

- 30 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. ArXiv:1409.1556
- 31 He K M, Zhang X Y, Ren S Q, et al. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Santiago, 2015. 1026–1034
- 32 Szegedy C, Liu W, Jia Y Q, et al. Going deeper with convolutions. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Santiago, 2015. 1–9
- 33 Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. 2015. ArXiv:1502.03167
- 34 Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016. 2818–2826
- 35 Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning. 2016. ArXiv:1602.07261
- 36 He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016. 770–778
- 37 Huang G, Liu Z, Der-Maaten L V, et al. Densely connected convolutional networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017. 4700–4708
- 38 Iandola F-N, Han S, Moskewicz M-W, et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and < 0.5 MB model size. 2016. ArXiv:1602.07360
- 39 Howard A-G, Zhu M L, Chen B, et al. MobileNets: efficient convolutional neural networks for mobile vision applications. 2017. ArXiv:1704.04861
- 40 Sandler M, Howard A, Zhu M L, et al. MobileNetV2: inverted residuals and linear bottlenecks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018. 4510–4520
- 41 Howard A, Sandler M, Chu G, et al. Searching for MobileNetV3. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 1314–1324
- 42 Zhang X Y, Zhou X Y, Lin M X, et al. ShuffleNet: an extremely efficient convolutional neural network for mobile devices. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018. 6848–6856
- 43 Ma N N, Zhang X Y, Zheng H T, et al. ShuffleNet V2: practical guidelines for efficient CNN architecture design. In: Proceedings of European Conference on Computer Vision, Munich, 2018. 116–131
- 44 Li X L, Yuan Y, Wang Q. Hyperspectral and multispectral image fusion via nonlocal low-rank tensor approximation and sparse representation. IEEE Trans Geosci Remote Sens, 2021, 59: 550–562
- 45 Xie S N, Tu Z W. Holistically-nested edge detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Santiago, 2015. 1395–1403
- 46 Jiang M X, Deng C, Shan J S, et al. Hierarchical multi-modal fusion FCN with attention model for RGB-D tracking. Inf Fusion, 2019, 50: 1–8
- 47 Cao J L, Pang Y W, Li X L. Triply supervised decoder networks for joint detection and segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 7392–7401
- 48 Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Santiago, 2015. 3431–3440
- 49 Jin X, Tan X. Face alignment in-the-wild: a survey. Comput Vision Image Underst, 2017, 162: 1–22
- 50 Zhao Z Y, Wang Q, Li X L. Deep reinforcement learning based lane detection and localization. Neurocomputing, 2020, 413: 328–338
- 51 Oliveira G-L, Burgard W, Brox T. Efficient deep models for monocular road segmentation. In: Proceedings of IEEE International Conference on Intelligent Robots and Systems, Daejeon, 2016. 4885–4891
- 52 Pan X G, Shi J P, Luo P, et al. Spatial as deep: spatial CNN for traffic scene understanding. 2017. ArXiv:1712.06080
- 53 Franke U, Joos A. Real-time stereo vision for urban traffic scene understanding. In: Proceedings of IEEE Intelligent Vehicles Symposium, Dearborn, 2000. 273–278
- 54 Wang B, Yuan X Y, Gao X B, et al. A hybrid level set with semantic shape constraint for object segmentation. IEEE Trans Cybern, 2019, 49: 1558–1569
- 55 Milletari F, Navab N, Ahmadi S. V-Net: fully convolutional neural networks for volumetric medical image segmen-

- tation. In: Proceedings of International Conference on 3D Vision, 2016. 565–571
- 56 Simpson A-L, Antonelli M, Bakas S, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. 2019. ArXiv:1902.09063
 - 57 Huang Q H, Huang Y H, Luo Y Z, et al. Segmentation of breast ultrasound image with semantic classification of superpixels. *Med Image Anal*, 2020, 61: 101657
 - 58 Ge Y Y, Zhang R M, Wang X G, et al. DeepFashion2: a versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 5337–5345
 - 59 Zhang X H, Wong Y K, Kankanhalli M-S, et al. Unsupervised domain adaptation for 3D human pose estimation. In: Proceedings of 27th ACM International Conference on Multimedia, Nice, 2019. 926–934
 - 60 Al-Amri S S, Kalyankar N V, Khamitkar S D. Image segmentation by using edge detection. *Int J Comput Sci Eng*, 2010, 2: 804–807
 - 61 Gao W S, Zhang X G, Yang L, et al. An improved Sobel edge detection. In: Proceedings of International Conference on Computer Science and Information Technology, Chengdu, 2010. 67–71
 - 62 Canny J. A computational approach to edge detection. *IEEE Trans Pattern Anal Mach Intell*, 1986, PAMI-8: 679–698
 - 63 Shen W, Wang X G, Wang Y, et al. DeepContour: a deep convolutional feature learned by positive-sharing loss for contour detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Santiago, 2015. 3982–3991
 - 64 Liu Y, Cheng M M, Hu X W, et al. Richer convolutional features for edge detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017. 3000–3009
 - 65 Snyder W, Bilbro G, Logenthiran A, et al. Optimal thresholding—a new approach. *Pattern Recogn Lett*, 1990, 11: 803–809
 - 66 Ohta Y I, Kanade T, Sakai T. Color information for region segmentation. *Comput Graph Image Process*, 1980, 13: 222–241
 - 67 Fan J P, Yau D K Y, Elmagarmid A K, et al. Automatic image segmentation by integrating color-edge extraction and seeded region growing. *IEEE Trans Image Process*, 2001, 10: 1454–1466
 - 68 Delon J, Desolneux A, Lisani J-L, et al. A nonparametric approach for histogram segmentation. *IEEE Trans Image Process*, 2007, 16: 253–261
 - 69 Noh H, Hong S, Han B, et al. Learning deconvolution network for semantic segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Santiago, 2015. 1520–1528
 - 70 Badrinarayanan V, Kendall A, Cipolla R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans Pattern Anal Mach Intell*, 2017, 39: 2481–2495
 - 71 Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Proceedings of International Conference on Medical Image Computing and Computer Assisted Intervention, Munich, 2015. 234–241
 - 72 Chen L-C, Papandreou G, Kokkinos I, et al. Semantic image segmentation with deep convolutional nets and fully connected CRFs. 2014. ArXiv:1412.7062
 - 73 Chen L-C, Papandreou G, Kokkinos I, et al. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans Pattern Anal Mach Intell*, 2018, 40: 834–848
 - 74 Zhao H S, Shi J P, Qi X J, et al. Pyramid scene parsing network. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017. 2881–2890
 - 75 Chen L-C, Papandreou G, Schroff F, et al. Rethinking atrous convolution for semantic image segmentation. 2017. ArXiv:1706.05587
 - 76 Chen L-C, Zhu Y K, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of European Conference on Computer Vision, Munich, 2018. 801–818
 - 77 Liu C X, Chen L-C, Schroff F, et al. Auto-DeepLab: hierarchical neural architecture search for semantic image segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 82–92
 - 78 Hu R H, Rohrbach M, Darrell T. Segmentation from natural language expressions. In: Proceedings of European Conference on Computer Vision, Amsterdam, 2016. 108–124

- 79 Woo S, Park J, Lee J-Y, et al. CBAM: convolutional block attention module. In: Proceedings of European Conference on Computer Vision, Munich, 2018. 3–19
- 80 Wang X L, Girshick R, Gupta A, et al. Non-local neural networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018. 7794–7803
- 81 Zhang C X, Song D J, Huang C, et al. Heterogeneous graph neural network. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, New York, 2019. 793–803
- 82 Fu J, Liu J, Tian H J, et al. Dual attention network for scene segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 3146–3154
- 83 Wang W G, Lu X K, Shen J B, et al. Zero-shot video object segmentation via attentive graph neural networks. In: Proceedings of IEEE International Conference on Computer Vision, Seoul, 2019. 9236–9245
- 84 Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Columbus, 2014. 580–587
- 85 Girshick R. Fast R-CNN. In: Proceedings of IEEE International Conference on Computer Vision, Santiago, 2015. 1440–1448
- 86 Ren S Q, He K M, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks. In: Proceedings of Advances in Neural Information Processing Systems, Quebec, 2015. 91–99
- 87 He K M, Gkioxari G, Dollár P, et al. Mask R-CNN. In: Proceedings of IEEE International Conference on Computer Vision, Venice, 2017. 2961–2969
- 88 Lin T-Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017. 2117–2125
- 89 Li Y, Qi H Z, Dai J F, et al. Fully convolutional instance-aware semantic segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017. 2359–2367
- 90 Chen X L, Girshick R, He K M, et al. TensorMask: a foundation for dense object segmentation. In: Proceedings of IEEE International Conference on Computer Vision, Seoul, 2019. 2061–2069
- 91 Bolya D, Zhou C, Xiao F Y, et al. YOLACT: real-time instance segmentation. In: Proceedings of IEEE International Conference on Computer Vision, Seoul, 2019. 9157–9166
- 92 Ling H, Gao J, Kar A, et al. Fast interactive object annotation with curve-GCN. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 5257–5266
- 93 Peng S, Jiang W, Pi H J, et al. Deep snake for real-time instance segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2020. 8533–8542
- 94 Kass M, Witkin A, Terzopoulos D. Snakes: active contour models. *Int J Comput Vision*, 1988, 1: 321–331
- 95 Xie E, Sun P Z, Song X G, et al. PolarMask: single shot instance segmentation with polar representation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2020. 12193–12202
- 96 Kirillov A, He K M, Girshick R, et al. Panoptic segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 9404–9413
- 97 Kirillov A, Girshick R, He K M, et al. Panoptic feature pyramid networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 6399–6408
- 98 Yang T J, Collins M-D, Zhu Y K, et al. DeeperLab: single-shot image parser. 2019. ArXiv:1902.05093
- 99 Li Y W, Chen X Z, Zhu Z, et al. Attention-guided unified network for panoptic segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 7026–7035
- 100 Li J, Raventos A, Bhargava A, et al. Learning to fuse things and stuff. 2018. ArXiv:1812.01192
- 101 Xiong Y W, Liao R J, Zhao H S, et al. UPSNet: a unified panoptic segmentation network. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 8818–8826
- 102 Dollar P, Wojek C, Schiele B, et al. Pedestrian detection: an evaluation of the state of the art. *IEEE Trans Pattern Anal Mach Intell*, 2011, 34: 743–761
- 103 Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, San Diego, 2005. 886–893
- 104 Leibe B, Seemann E, Schiele B. Pedestrian detection in crowded scenes. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, San Diego, 2005. 878–885

- 105 Enzweiler M, Gavrilu D M. Monocular pedestrian detection: survey and experiments. *IEEE Trans Pattern Anal Mach Intell*, 2009, 31: 2179–2195
- 106 Tuzel O, Porikli F, Meer P. Pedestrian detection via classification on riemannian manifolds. *IEEE Trans Pattern Anal Mach Intell*, 2008, 30: 1713–1727
- 107 Felzenszwalb P F, Girshick R B, McAllester D, et al. Object detection with discriminatively trained part-based models. *IEEE Trans Pattern Anal Mach Intell*, 2010, 32: 1627–1645
- 108 Wu B, Nevatia R. Detection and tracking of multiple, partially occluded humans by Bayesian combination of edgelet based part detectors. *Int J Comput Vis*, 2007, 75: 247–266
- 109 Ryan D, Denman S, Fookes C, et al. Crowd counting using multiple local features. In: *Proceedings of Digital Image Computing: Techniques and Applications*, Melbourne, 2009. 81–88
- 110 Chan A-B, Vasconcelos N. Bayesian poisson regression for crowd counting. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Kyoto, 2009. 545–551
- 111 Chen K, Loy C C, Gong S G, et al. Feature mining for localised crowd counting. In: *Proceedings of British Machine Vision Conference*, Guildford, 2012. 1–11
- 112 Zhang Y Y, Zhou D S, Chen S Q, et al. Single-image crowd counting via multi-column convolutional neural network. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 2016. 589–597
- 113 Zhang A R, Shen J Y, Xiao Z H, et al. Relational attention network for crowd counting. In: *Proceedings of IEEE International Conference on Computer Vision*, Seoul, 2019. 6788–6797
- 114 Chen J W, Su W, Wang Z F. Crowd counting with crowd attention convolutional neural network. *Neurocomputing*, 2020, 382: 210–220
- 115 Guo D, Li K, Zha Z J, et al. DADNet: dilated-attention-deformable ConvNet for crowd counting. In: *Proceedings of the 27th ACM International Conference on Multimedia*, Nice, 2019. 1823–1832
- 116 Liu J, Gao C Q, Meng D Y, et al. DecideNet: counting varying density crowds through attention guided detection and density estimation. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 2018. 5197–5206
- 117 Marsden M, McGuinness K, Little S, et al. ResnetCrowd: a residual deep learning architecture for crowd counting, violent behaviour detection and crowd density level classification. In: *Proceedings of IEEE International Conference on Advanced Video and Signal Based Surveillance*, Lecce, 2017. 1–7
- 118 Gao J Y, Wang Q, Li X L. PCC Net: perspective crowd counting via spatial convolutional network. *IEEE Trans Circ Syst Video Technol*, 2020, 30: 3486–3498
- 119 Efros A A, Leung T K. Texture synthesis by non-parametric sampling. In: *Proceedings of IEEE International Conference on Computer Vision*, Corfu, 1999. 1033–1038
- 120 Wei L Y, Levoy M. Fast texture synthesis using tree-structured vector quantization. In: *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, New Orleans, 2000. 479–488
- 121 Efros A A, Freeman W T. Image quilting for texture synthesis and transfer. In: *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, Los Angeles, 2001. 341–346
- 122 Kwatra V, Schödl A, Essa I, et al. Graphcut textures. *ACM Trans Graph*, 2003, 22: 277–286
- 123 Gatys L, Ecker A S, Bethge M. Texture synthesis using convolutional neural networks. In: *Proceedings of Advances in Neural Information Processing Systems*, Quebec, 2015. 262–270
- 124 Li Y H, Wang N Y, Liu J Y, et al. Demystifying neural style transfer. In: *Proceedings of International Joint Conference on Artificial Intelligence*, Melbourne, 2017. 2230–2236
- 125 Li S H, Xu X X, Nie L Q, et al. Laplacian-steered neural style transfer. In: *Proceedings of the 27th ACM International Conference on Multimedia*, Mountain View, 2017. 1716–1724
- 126 Li C, Wand M. Combining Markov random fields and convolutional neural networks for image synthesis. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 2016. 2479–2486
- 127 Semmo A, Limberger D, Kyprianidis J E, et al. Image stylization by oil paint filtering using color palettes. In: *Proceedings of International Symposium on Computational Aesthetics in Graphics, Visualization, and Imaging*, Istanbul, 2015. 149–158
- 128 Shih Y C, Paris S, Paris C, et al. Style transfer for headshot portraits. *ACM Trans Graph*, 2014, 148: 1–14
- 129 Johnson J, Alahi A, Li F F. Perceptual losses for real-time style transfer and super-resolution. In: *Proceedings of*

- European Conference on Computer Vision, Amsterdam, 2016. 694–711
- 130 Ulyanov D, Lebedev V, Vedaldi A, et al. Texture networks: feed-forward synthesis of textures and stylized images. In: Proceedings of the International Conference on Machine Learning, New York City, 2016. 1349–1357
- 131 Li C, Wand M. Precomputed real-time texture synthesis with Markovian generative adversarial networks. In: Proceedings of European Conference on Computer Vision, Amsterdam, 2016. 702–716
- 132 Dumoulin V, Shlens J, Kudlur M. A learned representation for artistic style. In: Proceedings of International Conference on Learning Representations, Toulon, 2017
- 133 Zhang H, Dana K. Multi-style generative network for real-time transfer. In: Proceedings of European Conference on Computer Vision Workshops, Munich, 2018. 8–14
- 134 Li Y J, Fang C, Yang J M, et al. Diversified texture synthesis with feed-forward networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017. 3920–3928
- 135 Chen T Q, Schmidt M. Fast patch-based style transfer of arbitrary style. 2016. ArXiv:1612.04337
- 136 Huang X, Belongie S. Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of IEEE International Conference on Computer Vision, Venice, 2017. 1501–1510
- 137 Ghiasi G, Lee H, Kudlur M, et al. Exploring the structure of a real-time, arbitrary neural artistic stylization network. In: Proceedings of British Machine Vision Conference, London, 2017. 4–7
- 138 Li Y J, Fang C, Yang J M, et al. Universal style transfer via feature transforms. In: Proceedings of Advances in Neural Information Processing Systems, Long Beach, 2017. 386–396
- 139 Singleton R C. On computing the fast Fourier transform. Commun ACM, 1967, 10: 647–654
- 140 Chatfield C. Wavelet transforms and time-frequency signal analysis. Technometrics, 2002, 44: 87
- 141 Rafael C G, Richard E W. Digital Image Processing. 2nd ed. Berlin: Springer, 2002
- 142 Slepian D. Linear least-squares filtering of distorted images. J Opt Soc Am, 1967, 57: 918–922
- 143 Singh M K, Tiwary U S, Kim Y H. An adaptively accelerated Lucy-Richardson method for image deblurring. EURASIP J Adv Signal Process, 2007, 2008: 1–10
- 144 Chowdhury M R, Qin J, Lou Y. Non-blind and blind deconvolution under poisson noise using fractional-order total variation. J Math Imag Vis, 2020, 62: 1238–1255
- 145 Dong W S, Zhang L, Shi G M, et al. Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization. IEEE Trans Image Process, 2011, 20: 1838–1857
- 146 Chao T S, Paris S, Horn B K, et al. Blur kernel estimation using the Radon transform. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, 2011. 241–248
- 147 Gupta A, Joshi N, Zitnick C L, et al. Single image deblurring using motion density functions. In: Proceedings of European Conference on Computer Vision, Heraklion, 2010. 171–184
- 148 Sun J, Cao W F, Xu Z B, et al. Learning a convolutional neural network for non-uniform motion blur removal. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Santiago, 2015. 769–777
- 149 Chakrabarti A. A neural approach to blind motion deblurring. In: Proceedings of European Conference on Computer Vision, Amsterdam, 2016. 221–235
- 150 Nah S, Kim T H, Lee K M. Deep multi-scale convolutional neural network for dynamic scene deblurring. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017. 3883–3891
- 151 Tao X, Gao H Y, Shen X Y, et al. Scale-recurrent network for deep image deblurring. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018. 8174–8182
- 152 Kupyn O, Budzan V, Mykhailych M, et al. DeblurGAN: blind motion deblurring using conditional adversarial networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018. 8183–8192
- 153 Zhang K H, Luo W H, Zhong Y R, et al. Deblurring by realistic blurring. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2020. 2737–2746
- 154 Yuan Y, Su W, Ma D D. Efficient dynamic scene deblurring using spatially variant deconvolution network with optical flow guided training. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2020. 3555–3564
- 155 Jiang Z, Zhang Y, Zou D Q, et al. Learning event-based motion deblurring. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2020. 3320–3329

- 156 Schultz R R, Stevenson R L. A Bayesian approach to image expansion for improved definition. *IEEE Trans Image Process*, 1994, 3: 233–242
- 157 Hou H, Andrews H. Cubic splines for image interpolation and digital filtering. *IEEE Trans Acoust Speech Signal Process*, 1978, 26: 508–517
- 158 Li X, Orchard M T. New edge-directed interpolation. *IEEE Trans Image Process*, 2001, 10: 1521–1527
- 159 Stark H, Oskoui P. High-resolution image recovery from image-plane arrays, using convex projections. *J Opt Soc Am A*, 1989, 6: 1715–1726
- 160 Irani M, Peleg S. Improving resolution by image registration. *Graph Model Image Process*, 1991, 53: 231–239
- 161 Schultz R R, Stevenson R L. Improved definition video frame enhancement. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Detroit, 1995. 2169–2172
- 162 Schultz R R, Stevenson R L. Extraction of high-resolution frames from video sequences. *IEEE Trans Image Process*, 1996, 5: 996–1011
- 163 Chang H, Yeung D Y, Xiong Y M. Super-resolution through neighbor embedding. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Washington, 2004. 275–282
- 164 Yang J C, Wright J, Huang Y, et al. Image super-resolution as sparse representation of raw image patches. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, 2008. 1–8
- 165 Yang J C, Wright J, Huang T S, et al. Image super-resolution via sparse representation. *IEEE Trans Image Process*, 2010, 19: 2861–2873
- 166 Dong C, Loy C C, He K M, et al. Image super-resolution using deep convolutional networks. *IEEE Trans Pattern Anal Mach Intell*, 2016, 38: 295–307
- 167 Kim J, Lee J K, Lee K M. Accurate image super-resolution using very deep convolutional networks. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 2016. 1646–1654
- 168 Zhang K, Zuo W M, Chen Y J, et al. Beyond a Gaussian denoiser: residual learning of deep CNN for image denoising. *IEEE Trans Image Process*, 2017, 26: 3142–3155
- 169 Zhang K, Zuo W M, Gu S H, et al. Learning deep CNN denoiser prior for image restoration. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 2017. 3929–3938
- 170 Dong C, Loy C C, He K M, et al. Learning a deep convolutional network for image super-resolution. In: *Proceedings of European Conference on Computer Vision*, Zurich, 2014. 184–199
- 171 Lim B, Son S, Kim H, et al. Enhanced deep residual networks for single image super-resolution. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Honolulu, 2017. 136–144
- 172 Ledig C, Theis L, Huszár F, et al. Photo-realistic single image super-resolution using a generative adversarial network. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 2017. 4681–4690
- 173 Sajjadi S M, Scholkopf B, Hirsch M. EnhanceNet: single image super-resolution through automated texture. In: *Proceedings of IEEE International Conference on Computer Vision*, Venice, 2017. 4501–4510
- 174 Oord A V D, Kalchbrenner N, Kavukcuoglu K. Pixel recurrent neural networks. In: *Proceedings of International Conference on Machine Learning*, New York City, 2013. 1747–1756
- 175 Oord A V D, Kalchbrenner N, Espeholt L, et al. Conditional image generation with PixelCNN decoders. In: *Proceedings of Advances in Neural Information Processing Systems*, Barcelona, 2016. 4790–4798
- 176 Kingma D P, Welling M. Auto-encoding variational Bayes. In: *Proceedings of International Conference on Learning Representations*, Banff, 2014
- 177 Sohn K, Lee H, Yan X C. Learning structured output representation using deep conditional generative models. In: *Proceedings of Advances in Neural Information Processing Systems*, Montreal, 2015. 3483–3491
- 178 Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. In: *Proceedings of Advances in Neural Information Processing Systems*, Montreal, 2014. 2672–2680
- 179 Arjovsky M, Bottou L. Towards principled methods for training generative adversarial networks. In: *Proceedings of International Conference on Learning Representations*, Toulon, 2017
- 180 Arjovsky M, Chintala S, Bottou L. Wasserstein GAN. 2017. ArXiv:1701.07875
- 181 Gulrajani I, Ahmed F, Arjovsky M, et al. Improved training of wasserstein GANs. In: *Proceedings of Advances in Neural Information Processing Systems*, Long Beach, 2017. 5767–5777
- 182 Mao X D, Li Q, Xie H R, et al. Least squares generative adversarial networks. In: *Proceedings of IEEE International*

- Conference on Computer Vision, Venice, 2017. 2794–2802
- 183 Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. In: Proceedings of International Conference on Learning Representations, San Juan, 2016
- 184 Mirza M, Osindero S. Conditional generative adversarial nets. 2014. ArXiv:1411.1784
- 185 Zhang H, Goodfellow I, Metaxas D, et al. Self-attention generative adversarial networks. In: Proceedings of International Conference on Machine Learning, Long Beach, 2019. 7354–7363
- 186 Brock A, Donahue J, Simonyan K. Large scale GAN training for high fidelity natural image synthesis. In: Proceedings of International Conference on Learning Representations, New Orleans, 2019
- 187 Larsen A B L, Sønderby S K, Larochelle H, et al. Autoencoding beyond pixels using a learned similarity metric. In: Proceedings of International Conference on Machine Learning, Long Beach, 2016. 1558–1566
- 188 Donahue J, Krähenbühl P, Darrell T. Adversarial feature learning. 2016. ArXiv:1605.09782
- 189 Donahue J, Simonyan K. Large scale adversarial representation learning. In: Proceedings of Advances in Neural Information Processing Systems, Vancouver, 2019. 10542–10552
- 190 Liu B Y, Gould S, Koller D. Single image depth estimation from predicted semantic labels. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, 2010. 1253–1260
- 191 Tighe J, Lazebnik S. SuperParsing: scalable nonparametric image parsing with superpixels. In: Proceedings of European Conference on Computer Vision, Heraklion, 2010. 6315: 352–365
- 192 Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: common objects in context. In: Proceedings of European Conference on Computer Vision, Zurich, 2014. 740–755
- 193 Mottaghi R, Chen X J, Liu X B, et al. The role of context for object detection and semantic segmentation in the wild. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Columbus, 2014. 891–898
- 194 Fritsch J, Kuehnl T, Geiger A. A new performance measure and evaluation benchmark for road detection algorithms. In: Proceedings of IEEE International Conference on Intelligent Transportation Systems, Hague, 2013. 1693–1700
- 195 Zhou B L, Zhao H, Puig X, et al. Scene parsing through ADE20K dataset. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, 2017. 5122–5130
- 196 Huang X Y, Cheng X J, Geng Q C, et al. The apolloscape dataset for autonomous driving. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, 2018. 954–960
- 197 Wang Q, Gao J Y, Lin W, et al. NWPU-crowd: a large-scale benchmark for crowd counting and localization. IEEE Trans Pattern Anal Mach Intell, 2020. doi: 10.1109/TPAMI.2020.3013269
- 198 Sindagi V A, Yasarla R, Patel V M. JHU-crowd++: large-scale crowd counting dataset and a benchmark method. 2020. ArXiv:2004.03597
- 199 Fang Y Y, Zhan B Y, Cai W D, et al. Locality-constrained spatial transformer network for video crowd counting. In: Proceedings of IEEE International Conference on Multimedia and Expo, Shanghai, 2019. 814–819
- 200 Wang Q, Gao J Y, Lin W, et al. Learning from synthetic data for crowd counting in the wild. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 8198–8207
- 201 Wang Q, Gao J Y, Lin W, et al. Pixel-wise crowd understanding via synthetic data. Int J Comput Vis, 2021, 129: 225–245
- 202 Idrees H, Tayyab M, Athrey K, et al. Composition loss for counting, density map estimation and localization in dense crowds. In: Proceedings of European Conference on Computer Vision, Munich, 2018. 532–546
- 203 Zhang Q, Chan A B, Wide-area crowd counting via ground-plane density maps and multi-view fusion CNNs. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 8297–8306
- 204 Li J W, Song J X. Pedestrian counting via deep convolutional neural networks in crowded scene. In: Proceedings of International Conference on Advanced Materials and Information Technology Processing, Guilin, 2016
- 205 Chan A B, Liang Z S, Vasconcelos N. Privacy preserving crowd monitoring: counting people without people models or tracking. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, 2008. 1–7
- 206 Ros G, Sellart L, Materzynska J, et al. The synthia dataset: a large collection of synthetic images for semantic segmentation of urban scenes. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016. 3234–3243
- 207 Richter S R, Vineet V, Roth S, et al. Playing for data: ground truth from computer games. In: Proceedings of IEEE

- Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016. 102–118
- 208 Han T, Gao J Y, Yuan Y, et al. Focus on semantic consistency for cross-domain crowd understanding. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2020. 1848–1852
- 209 Wang Q, Gao J Y, Li X L. Weakly supervised adversarial domain adaptation for semantic segmentation in urban scenes. *IEEE Trans Image Process*, 2019, 28: 4376–4386
- 210 Zhao Z Y, Han T, Gao J Y, et al. A flow base bi-path network for cross-scene video crowd understanding in aerial view. In: Proceedings of European Conference on Computer Vision, Cham, 2020. 574–587
- 211 Zhu J Y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of International Conference on Computer Vision, Venice, 2017. 2223–2232
- 212 Li X L, Zhang H, Zhang R, et al. Discriminative and uncorrelated feature selection with constrained spectral analysis in unsupervised learning. *IEEE Trans Image Process*, 2020, 29: 2139–2149
- 213 Li X L, Chen M L, Wang Q. Adaptive consistency propagation method for graph clustering. *IEEE Trans Knowl Data Eng*, 2020, 32: 797–802
- 214 Li X L, Zhang R, Wang Q, et al. Autoencoder constrained clustering with adaptive neighbors. *IEEE Trans Neural Netw Learn Syst*, 2021, 32: 443–449
- 215 Canziani A, Paszke A, Culurciello E. An analysis of deep neural network models for practical applications. 2016. ArXiv:1605.07678
- 216 Zhao H S, Qi X J, Shen X Y, et al. ICNet for real-time semantic segmentation on high-resolution images. In: Proceedings of European Conference on Computer Vision, Munich, 2018. 405–420
- 217 Wang Y, Zhou Q, Liu J, et al. LEDNet: a lightweight encoder-decoder network for real-time semantic segmentation. In: Proceedings of IEEE International Conference on Image Processing, Taipei, 2019. 1860–1864
- 218 Han K, Wang Y H, Tian Q, et al. GhostNet: more features from cheap operations. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2020. 1580–1589
- 219 Chen H T, Wang Y H, Xu C J, et al. AdderNet: do we really need multiplications in deep learning? In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2020. 1468–1477
- 220 Gallo M L, Boybat I, Rajendran B, et al. Mixed-precision training of deep neural networks using computational memory. 2017. ArXiv:1712.01192
- 221 Li F F, Zhang B, Liu B. Ternary weight networks. 2016. ArXiv:1605.04711
- 222 Hubara I, Courbariaux M, Soudry D, et al. Binarized neural networks. In: Proceedings of Advances in Neural Information Processing Systems, Barcelona, 2016. 4107–4115
- 223 Zhao W L, Fu H H, Luk W, et al. F-CNN: an FPGA-based framework for training convolutional neural networks. In: Proceedings of IEEE International Conference on Application-specific Systems, Architectures and Processors, London, 2016. 107–114
- 224 Feldmann J, Youngblood N, Wright C D, et al. All-optical spiking neurosynaptic networks with self-learning capabilities. *Nature*, 2019, 569: 208–214
- 225 Zeiler M D, Fergus R. Visualizing and understanding convolutional networks. In: Proceedings of European Conference on Computer Vision, Zurich, 2014. 818–833

Pixel level semantic understanding: from classification to regression

Xuelong LI* & Zhiyuan ZHAO

School of Computer Science and School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China

* Corresponding author. E-mail: li@nwpu.edu.cn

Abstract With the rapid developments in science and technology and the continuous iteration of hardware equipment, artificial intelligence is being widely used in various fields (such as security monitoring, medical assistance, health diagnosis, intelligent recommendation, remote sensing monitoring, and target location). With the gradual expansion of user requirements for intelligent processing tasks, semantic understanding levels and accuracy requirements of intelligent algorithms are also gradually increasing. As such, the pixel-level task of semantic understanding, which requires much higher accuracy than image-level understanding, is attracting increasing attention. Compared to image-level understanding, in addition to the large amount and high precision of pixel-by-pixel output, the internal challenges in pixel-level semantic understanding merit greater attention and research. From the information measurement perspective, and given the unique attributes of pixel-level semantic understanding, in this paper, we present the definition and optimization goal of pixel-level semantic understanding, and derive pixel-level semantic classification and pixel-level semantic regression from the original definition based on the characteristics of actual tasks. We then discuss the degradation and evolution of optimization objectives in these two subcategories, respectively. Based on a detailed investigation, we summarize the development and current status of typical tasks in pixel-level semantic understanding. We consider the difficulties and future development direction of the current pixel-level semantic understanding, perform an in-depth analysis, and suggest feasible solutions to problems that require urgent solutions. Finally, we focus on the opportunities and challenges faced by pixel-level semantic understanding, and even artificial intelligence, in the post-deep-learning era and argue that knowledge guidance and data-driven optimization are fundamental objectives for the future development of artificial intelligence. From the definition and current development of pixel-level semantic understanding, in this work, we extend the current conceptual approach and reflect on the risks faced by the whole field, providing an in-depth discussion and consideration of the developmental direction of related technologies, while clarifying the knowledge base of pixel-level semantic understanding.

Keywords pixel level semantic understanding, artificial intelligence, deep learning, classification, regression

Xuelong LI is a full professor at the School of Computer Science and School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China.

Zhiyuan ZHAO received his B.E. degree in electronics and information engineering from Northwestern Polytechnical University, Xi'an, Shaanxi, China, in 2018. He is currently pursuing a Ph.D. degree at the School of Computer Science and School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision and pattern recognition.