属性知识引导的自适应视觉感知与结构理解研究进展

张知诚[1] 杨巨峰[1] 程明明[1] 林巍峣[2] 汤 进[3] 李成龙[3] 刘成林[4]

摘 要 机器通过自适应感知从环境中提取人类可理解的信息,从而在开放场景中构建类人智能.因属性知识具有类别无关的特性,以其为基础构建的感知模型与算法引起广泛关注.文中首先介绍属性知识引导的自适应视觉感知与结构理解的相关任务,分析其适用场景.然后,总结四个关键方面的代表性工作.1)视觉基元属性知识提取方法,涵盖底层几何属性和高层认知属性;2)属性知识引导的弱监督视觉感知,包括数据标签受限情况下的弱监督学习与无监督学习;3)图像无监督自主学习,包括自监督对比学习和无监督共性学习;4)场景图像结构化表示和理解及其应用.最后,讨论目前研究存在的不足,分析有价值的潜在研究方向,如大规模多属性基准数据集构建、多模态属性知识提取、属性知识感知模型场景泛化、轻量级属性知识引导的模型开发、场景图像表示的实际应用等.

关键词 自适应感知, 结构理解, 属性知识, 弱监督学习, 无监督学习

引用格式 张知诚,杨巨峰,程明明,林巍峣,汤 进,李成龙,刘成林.属性知识引导的自适应视觉感知与结构理解研究进展.模式识别与人工智能, 2023, 36(12): 1104-1126.

DOI 10.16451/j.cnki.issn1003-6059.202312004 **中图法分类号** TP 37

Progress in Attribution-Guided Adaptive Visual Perception and Structure Understanding

ZHANG Zhicheng[1], YANG Jufeng[1], CHENG Mingming[1], LIN Weiyao[2], TANG Jin[3], LI Chenglong[3], LIU Chenglin[4]

ABSTRACT Machines extract human-understandable information from the environment via adaptive perception to build intelligent system in open-world scenarios. Derived from the class-agnostic characteristics of attribute knowledge, attribution-guided perception methods and models are established and widely studied. In this paper, the tasks involved in attribution-guided adaptive visual perception and structure understanding are firstly introduced, and their applicable scenarios are analyzed. The representative

收稿日期:2023-10-07;录用日期:2023-12-25
Manuscript received October 7, 2023;
accepted December 25, 2023
科技创新 2030-"新一代人工智能"重大项目(No. 2018AAA0100400)、天津市自然科学基金杰出青年基金项目(No. 20JCJQJC00020)、国家自然科学基金项目(No. 62325109, U21B2013)、中央高校基本科研业务费资助
Supported by National Key Research and Development Program of China(No. 2018AAA0100400), Natural Science Foundation for Distinguished Young Scholars of Tianjin(No. 20JCJQJC00020), National Natural Science Foundation of China(No. 62325109, U21B2013), Fundamental Research Funds for the Central Universities
本文责任编委 桑 农
Recommended by Associate Editor SANG Nong

1. 南开大学 计算机学院 天津 300350
2. 上海交通大学 电子信息与电气工程学院 上海 200240
3. 安徽大学 计算机科学与技术学院 合肥 230601
4. 中国科学院自动化研究所 多模态人工智能系统全国重点实验室 北京 100190
1. College of Computer Science, Nankai University, Tianjin 300350
2. School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240
3. School of Computer Science and Technology, Anhui University, Hefei 230601
4. State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190

research on four key aspects is summarized. Basic visual attribute knowledge extraction methods cover low-level geometric attributes and high-level cognitive attributes. Attribute knowledge-guided weakly-supervised visual perception includes weakly supervised learning and unsupervised learning under data label restrictions. Image self-supervised learning covers self-supervise contrastive learning and unsupervised commonality learning. Structured representation and understanding of scene images and their applications are introduced as well. Finally, challenges and potential research directions are discussed, such as the construction of large-scale benchmark datasets with multiple attributes, multi-modal attribute knowledge extraction, scene generalization of attribute knowledge perception models, the development of lightweight attribute knowledge-guided models and the practical applications of scene image representation.

Key Words Adaptive Perception, Structure Understanding, Attribution Knowledge, Weakly-Supervised Learning, Unsupervised Learning

Citation ZHANG Z C, YANG J F, CHENG M M, LIN W Y, TANG J, LI C L, LIU C L. Progress in Attribution-Guided Adaptive Visual Perception and Structure Understanding. Pattern Recognition and Artificial Intelligence, 2023, 36(12): 1104-1126.

机器感知的本质是从传感器捕获的数据中提取智能系统可理解的信息,包括场景[1]、物体[2]、行为[3]、关系[4]、时间[5]等,为智能系统做出决策提供判据,是智能化的重要技术支撑[6-7]. 深度学习的发展推动模式识别和智能感知算法取得长足进步[8]. 然而,开放环境下的自适应感知仍面临新的挑战,包括感知对象类别集变化、数据分布变化与数据噪声、模态异质性的干扰等,因而需要感知模型和算法具有对领域变化的自适应性、对噪声与异常数据的鲁棒性、对感知结果的可解释性[9-12].

构建鲁棒、可解释、自适应的感知模型面临如下困难.

1)开放场景中的某些类别样本量较少,依赖大量数据的训练框架难以获得小样本泛化能力[13]. 不同类别数据的样本量差距可达百倍以上,且开放环境中不仅有数据集所含的常见类别,还有面向特定应用场景的少见类别,如工业车间中的元器件、教学场景中的工具等.

2)数据质量不一,存在噪声和异常,造成现有模型的稳定性与鲁棒性较差[14]. 开放环境中获取的图像视频数据通常包含大量噪声,既有视觉内容因低质、干扰等因素自己携带的噪声,也有因数据来源繁杂或搜集过程困难造成的很多标签噪声.

3)数据分布不受限制,模型难以适配数据分布与标注类别变化的情况[15]. 数据集上的训练样本难以覆盖所有属性的样本,当同一类别的数据样本存在较大的分布差异时,算法难以准确感知. 同时,当类别分布发生变化,如新增类别时,算法无法进行相应的迁移调整.

4)现有深度模型大部分属于黑盒模型,难以解释开放场景下的感知过程和结果[16]. 深度神经网络由数千个神经元组成,这些神经元以一种分散的方式共同工作以解决问题. 每个神经元可能对提取的某些特征进行编码,但这些编码往往缺乏物理含义,因此难以被人类理解.

针对上述挑战,学者们分别从感知模型结构设计、特征提取、学习算法、知识表示和推理等不同角度提出解决办法. 相关的研究工作包括可解释深度神经网络[17]、小样本学习[13]、场景图生成[18]等. 神经科学研究表明,采用对不同场景和目标具有通用性且符合人类视觉机理的基元属性作为特征,有利于提升感知模型的泛化性、可解释性、自适应性和鲁棒性[19-22]. 场景结构理解也依赖具有通用性和鲁棒性的基元属性知识. 因此,本文聚焦基元属性知识引导的自适应感知和结构理解,汇聚相关的研究问题,评述最新研究进展.

属性知识引导的自适应感知与结构理解任务主要包括如下方面.

1)任务类别无关的视觉基元属性感知. 通过模拟人类感知机制,算法从图像中提取任务类别无关的视觉基元属性,包括视觉结构的几何基元(如角点、线段、边缘、平面、显著区域等)和视觉图像的认知属性(如情感、复杂度、记忆性、图像质量、图像美学等). 这些基元属性可作为后续步骤的输入,用于场景理解和识别.

2)属性知识引导的弱监督视觉感知. 基于提取

的属性知识,指导模型从弱标注数据中学习模型结构和参数,完成视觉对象的识别、检测、分割和定位.例如:可以使用语义标签、场景描述等信息辅助图像分类、目标检测和分割模型的学习.

3)图像无监督自主学习.利用海量的无标签数据,挖掘数据中的潜在结构和规律,让模型借助提取的知识完成自主学习.然而开放环境中的数据通常包含大量噪声且数据质量不同,由此引入新的挑战.

4)场景图像的结构化表示和理解.基于从图像中提取的基元属性,得以训练并辅助模型提取更高级别的结构化表示,如场景图、语义网格等.这些表示帮助算法更好地理解图像中的语义信息和空间关系,从而实现更准确和全面的场景理解.

本文回顾属性知识引导的自适应感知与结构理解的研究进展,总结 4 个关键任务的发展历程及其在开放场景自适应感知领域的代表性工作,分析不同方法的特点和适用条件.最后,总结全文并展望未来值得探索的方向.

目前针对视觉感知与结构理解中的各个感知任务已有充分的综述工作,这些综述在弱监督感知、无监督学习与自监督学习等领域中进行大量的调研整理工作.尽管研究任务相同,但本文基于属性引导的研究工作与它们具有显著差异,尤其是本文提供一种新的视角(属性知识)实现自适应的感知方法,涵盖从底层到高层属性知识的提取、标签不足与缺失时的模型、算法与学习范式的设计以及面向开放环境场景图像的算法应用.并且现有综述多针对单一自适应视觉感知与结构理解任务,而本文回顾属性知识应用的多个感知任务,提供更全面的视角.

1 视觉基元属性感知

视觉基元属性感知可追溯到 20 世纪 80 年代初,Fukushima[23]提出神经认知机(Neocognitron),模拟人类视觉感知系统的一些基本特性.

视觉基元属性感知领域早期的研究主要集中在边缘、角点等基本图像特征的检测和描述上.随着计算机技术和算法的发展,人们开始研究更复杂的几何基元,如线段、曲线、平面等,进而关注图像的高层属性,如物体类别、场景类型、情感等.这些属性通常是由多个基本特征组合而成,因此需要进行更高级别的特征提取和分析.

基元属性广泛存在于视觉数据中,有助于完成很多不同的图像分析和理解任务.

现有各种基元属性的定义如表 1 所示.各种基元属性的示例样本如图 1 所示.

表 1 视觉基元属性的定义

Table 1 Definition of visual primitive attribution

基元属性		知识定义
几何属性	关键点	图像特征的局部表达,如斑点和角点
	直线线段	直线上两点和它们之间的部分
	曲线边缘	图像中发生急剧变化的区域边界
	平面区域	以特定角度倾斜的平面
	显著区域	图像中引起人眼注意的区域或物体
认知属性	视觉情感	人从图像中感受的生理和心理反应
	复杂度	图像中包含的信息量与细节数量
	记忆性	图像被人记住和识别的能力
	图像质量	图像信号的准确程度和视觉质量
	图像美学	图像的表现力和视觉吸引力

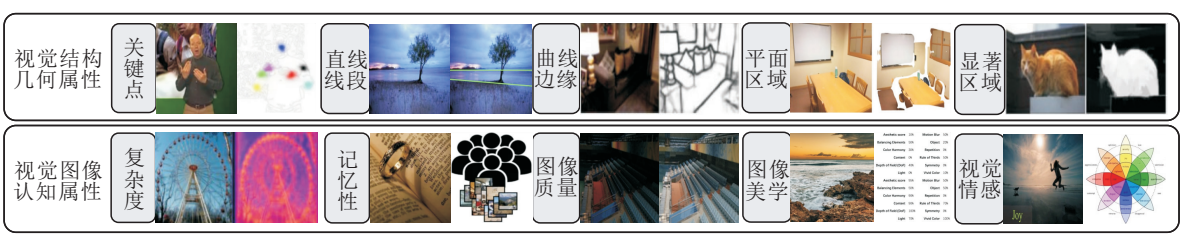


图 1 视觉基元属性示例

Fig. 1 Examples of visual primitive attribution

1.1 视觉结构的几何属性

关键点作为视觉基元属性,在计算机视觉和图像处理领域得到广泛应用,包括图像拼接、目标跟踪、三维重建等.Lowe 等[24-25]提出 SIFT(Scale Invariant Feature Transform),可在不同尺度下检测到图像中的特征点,这些特征点具有旋转、尺度和光照不变性.

随着深度学习的发展,大量工作采用数据驱动的方式提取关键点,大幅提升关键点提取的精确性.DeTone 等[26]提出 SuperPoint(Self-Supervised Interest

Point Detection),在合成图像数据集与真实图像数据集上进行联合自监督训练,具有跨域检测上的泛化性. Liu 等[27]提出 GIFT(Group Invariant Feature Transform),考虑特征点在多个视角上的一致性,设计组卷积神经网络提取特征,并进一步通过群线性池化融合这些特征.

在现实场景中,关键点涉及的任务类别广泛,如人的身体部件、物体的角点等. 近年来,一些工作利用提取得到的关键点,指导完成特定任务. 基于提取到的特征点,Lin 等[28]提出用于无损关键点视频序列压缩的方法,消除视频中关键点的时空冗余以实现压缩,关键点序列的范围涵盖 2D 边界框、人体骨架、3D 边界框和面部特征点. 该无损压缩方法可保持数据的完整性,确保在后续的视频分析任务中准确可靠.

在图像中,直线线段是一种有用的中间层表示,在人类视觉由底层到高层进行语义信息转换时发挥重要作用. Hough 变换[29]是具有代表性的方法之一. 根据截距和斜率进行参数化,图像中的直线被转化为 Hough 空间中横坐标为截距、纵坐标为斜率的参数点,此时直线检测等价变换为点检测的问题.

近年来,一些方法通过训练卷积神经网络以检测图像中的直线. PPGNet[30]提取直线的关键点信息,采用图结构描述连接点、直线段以及它们的关联关系. 不同于参数空间变换,PPGNet 利用 CNN(Convolutional Neural Network)直接从图像推理得到图结构,从而获得直线和端点的结构化信息. Xue 等[31]进一步利用图像的区域划分图指导直线的检测和提取,首先将每个像素分配给对应的一个线段,构建区域划分图,再通过其相对于线段的二维投影向量对区域中的每个像素进行编码. 最终,压缩编码后的区域,得到预测的线段图.

不同于物体中存在的线结构,自然图像中的“语义线”是指能勾勒图像内容结构的直线,如不同区域的分割线、建筑物的中轴线等. 语义线检测在摄影构图、图像处理等下游任务中具有广泛应用. 例如:将图像中的语义线置于照片的黄金分割位置,有助于拍出视觉效果更好的图像.

Lee 等[32]首次使用 CNN 提取线条的候选集合,将语义线识别视作目标检测问题的特例,沿整条直线对特征进行双线性插值. 然后类似于 Faster-R-CNN(Faster Region-CNN),通过分类器和回归器验证线条表示. Han 等[33]从 Hough 变换的角度将语义线提取视作线条提取问题,遍历图像中所有可能的直线,沿直线将特征聚合到参数空间中对应的点上,语义线估计因此转换为参数域中的点检测问题,获得更高的检测效率. Zhao 等[34]进一步提出边缘信息指导的修正模块,将空间域与参数域的直线进行对齐,获取精确的直线估计.

边缘提取算法的基本思想是通过对图像进行滤波、梯度计算等操作,找到像素值变化较大的区域,从而得到物体的边缘轮廓. 最具代表性的 Sobel 类离散差分算子[35]计算图像亮度函数的梯度近似值. 在图像的任何一点使用此算子,将会产生对应的梯度向量或是其范数.

Soria 等[36]提出 DexiNed(Dense Extreme Inception Network for Edge Detection),基于视觉特征解耦的假说,即边缘、轮廓与边界是三个不同的视觉特征,分开进行基准数据集的评估. 为此,该工作提出一个新的边缘数据集,并展示 DexiNed 在该数据集上的表现. DexiNed 不需要预先训练的权重,也无需基于目标识别的预训练.

场景中的平面区域为基于视觉的很多应用提供重要信息,包括立体视觉与机器人视觉等. 平面区域检测旨在使机器可以像人类一样具有理解高层场景结构的感知能力. 在从单个图像中提取所有平面后,人们可选择他们感兴趣的平面,并基于这些平面区域设计有效且有吸引力的应用程序. 例如:可使用喜爱的纹理装饰墙壁;广告商可在宣传视频中充分利用信息稀疏的区域(如桌子、墙壁和木板),更有效地营销他们的产品. 此外,平面特征也是自主机器人感知周围环境和通过相机视图构建地图的关键线索.

随着深度神经网络的兴起,平面区域分析逐渐演化为平面分割、平面重建、平面跟踪等若干子任务. Liu 等[37]提出 PlaneNet,基于 CNN 的端到端架构检测平面区域. 作者将这项任务考虑为像素分割问题,因此只能检测固定数量的平面.

为了解决平面数量受限的问题,Liu 等[38]利用 Mask R-CNN 生成任意数量的平面,设计精炼模块,同时集成所有平面的特征,进一步对预测结果进行细化处理.

为了提升图像对间的平面搜索能力,Wang 等[39]提出 Gracker(Graph-Based Tracker),将问题建模为图像对之间的图匹配问题,在参考帧与搜索帧之间分别构建平面区域特征图,并进行跨图匹配的搜索优化,在室内场景和室外场景均获得良好的跟踪性能.

尽管大多数工作关注于跟踪视频中的单个平面,事实上,真实世界中通常同时存在多个平面. Zhang 等[40]提出 PRTrack(Tracking Framework Comprised of Procedures for Appearance Perception and Occlusion Reasoning). 为了应对引入多个平面后的相互遮挡问题,PRTrack 统一外观感知和遮挡推理,使用双分支网络跟踪平面对象的可见部分,包括顶点和掩码. 他们还开发一个遮挡区域定位策略,用于推断不可见部分,即被遮挡的区域,最终通过双流注意力网络实现精细化预测.

近来,为了整合平面分析任务,Zhang 等[41]提出 PlaneSeg,利用物体的边缘信息,给模型提供具有判别性的平面特征,以即插即用的形式嵌入到下游任意的平面区域分析深度网络中. 该框架分别提取边缘特征与上下文特征,并对这两种特征在多个层级进行成对融合. 之后提供给神经网络,完成众多平面分析任务,包括平面分割、平面重建与深度估计等.

显著性是指图像中引起人眼注意的区域或物体. 这些区域可能与周围环境不同,或者在图像中有特殊的位置、颜色、纹理或形状. 显著性对于计算机视觉和图像处理非常重要,因为它可以帮助计算机自动识别和理解图像中的重要信息. 常见的应用包括图像检索、自动驾驶、视频监控等.

为了从多种角度综合考虑显著性,大量工作都选择设计深度神经网络架构. Liu 等[42]提出重新设计的 U 形结构,用于显著目标检测中的多尺度特征提取,该结构集中底部向上和顶部向下的通路,实现跨尺度信息交互,并提取语义上更强、位置更精确的特征. Wu 等[43]提出 EDN(Extremely Downsampled Network),通过极端下采样技术有效学习整个图像的全局视图,提高高级特征的表现,从而实现精确的显著目标定位. 同时,作者构建 SCPC(Scale-Correlated Pyramid Convolution),用于从上述极端下采样中恢复目标细节. Wu 等[44]提出用于 SIS(Salient Instance Segmentation)的网络架构,网络预测每个显著实例的类别无关掩码. 经过正则化的密集连接,从所有特征金字塔中注意到促进信息性特征并抑制非信息性特征.

为了联合提取多种知识基元,近年来,研究者设计通用网络框架,同时整合多种基元属性的提取. Liu 等[45]提出 PoolNet+,应用于边缘检测、RGB-D 显著性对象检测和伪装对象检测任务. PoolNet+可逐步提炼高层语义特征,获得细节丰富的显著性地图,适合于移动应用. 现有显著物体检测模型存在的问题主要是针对理想条件下的数据集进行的训练难以应对真实场景中的复杂情况.

为了更好地反映真实场景中的复杂性,Fan 等[2]分析现有显著物体检测(Salient Object Detection, SOD)数据集的设计偏差问题,并提出一个高质量数据集——SOC,该数据集包含多个常见物体类别的显著图像和非显著图像. 基于真实场景的 SOC 数据集,作者提出的数据增强策略包括标签平滑、随机图像增强和自监督学习,可提高现有 SOD 模型的性能. SOC 数据集对 SOD 领域的研究意义在于提供一个更综合和平衡的基准,可从不同角度客观评估模型性能.

为了解决显著性主观定义的问题,Fan 等[46]提出评估前景映射的结构度量指标 S-measure,同时评估前景映射和基准映射之间的区域感知和对象感知结构相似性. 该指标能有效区分模型的优点和缺点,是之前评估方法的有益补充.

在显著性检测的实际应用中,深度图由于其便捷性广泛应用于移动边端设备[47]. 为了有效整合 RGB 和深度信息以提高检测性能,Fan 等[48]提出 BBS-Net(Bifurcated Backbone Strategy Network),采用分叉的骨干策略,将多级特征分成教师特征和学生特征,并使用深度增强模块,从通道和空间视图中挖掘深度线索的信息. BBS-Net+[49]更进一步使用深度适配器模块以压缩模型参数量.

此外,受应用场景驱动,轻量化的显著性检测成为重要研究方向之一. HVPNet[6]模拟灵长类的视觉皮层进行分层感知学习,设计 HVP(Hierarchical Visual Perception)模块. MobileSal[50]采用移动网络进行深度特征提取,实现高效 RGB-D 显著目标检测. Gao 等[51]提出 CSNet,利用灵活的广义卷积模块 Oct-Conv 提取多尺度特征,同时通过动态权重衰减方案减少特征冗余. 该模型仅需 100 K 的参数量,并在常用的显著性目标检测基准测试中取得与大型模型相当的性能. Cheng 等[52]进一步研究显著性目标检测模型的语义信息编码方式,以及它们是否是类别不可知的.

研究表明,显著性检测和分类方法基于不同的机制,因此 SOD 对类别不敏感且不必要使用 ImageNet 预训练对 SOD 进行训练,同时,SOD 所需的参数量少于分类模型.

1.2 视觉图像的认知属性

情感是重要的认知属性. Minsky 指出,“问题不

在于智能机器是否会有情感,而是没有情感的机器能否智能”[53]. 基于图像载体,情感计算技术可帮助智能机器更好地理解人类的观点和意图,从而更好地与人类进行交互. 例如:在人机对话中,情感识别帮助机器更好地理解用户的情感状态,从而更好地回应用户的需求. 随着视觉情感领域的快速发展,其广泛的应用场景受到更多关注,如社交助手[54-55]、商务智能[56]、意见挖掘[57-58]等.

与图像感知层面的研究不同,情感图像内容分析的目标是理解认知层次的语义信息. 为了建模情感的刺激诱因,一些工作研究如何定位图像中传递情感的物体[59-60]. She 等[61]提出 WSCNet(Weakly Supervised Coupled Network),考虑非受限于事物性的更普遍的情感区域,采用弱监督学习的方式在图像级情感标签的帮助下进行定位. 最近,Feng 等[7]将情感产生的诱因归结为三个阶段:刺激获取、整体组织以及高层感知. 为每个阶段设计专门的情感预训练任务以发掘具有区分性的特征表示.

情感的另一个特点来自于人的认知差异,现有工作通过刻画这种差异以获取更准确的情感预测. Yang 等[62]探讨社交媒体中用户情绪的表达方式,揭示朋友互动的作用. Yang 等[63]提出 SAMNet(Subjectivity Appraise-and-Match Network),为了描述群众投票过程中的多样性,进行多支路的主观评估,其中每条支路模拟一个特定个体的情感唤醒过程. 考虑到情感的特性,研究者们发现标签共现存在规律性,特定情感标签往往成对出现[64-65]. 针对标签距离情感相关性,Yang 等[66]设计隐含和排斥两种生成策略,构建标签分布.

由于心理学专家定义的情感模型没有确定的统一形式,现实世界对不同情感类别的需求也多种多样. Yang 等[67]提出三元对比损失,统一情感模型与极性模型的关系. Yao 等[68]提出 APSE(Attention-Aware Polarity Sensitive Embedding),进一步提出来自极性-情感两个尺度的有监督注意力模块,增强特征表示能力.

此外,标签稀缺是情感识别长久以来的问题,大量工作[69-71]探索是否可只使用少量标签或获取便捷的情感线索训练模型. 最近,视频因其包含多种情感刺激物而备受关注[72],其中,情感主要由视频的某些关键帧和相应的判别区域引起. Zhao 等[73]提出 VAANet(Visual-Audio Attention Network),在帧间实施空间维度、通道维度与时间维度上的注意力检测,提取到具有关键信息的关键帧. Zhang 等[74]提出视频情感定位任务,同时定位情感段落并识别相应的情感类别,提出弱监督视频情感定位框架,基于对比一致性的多模态特征融合模块,利用反向映射策略融合多模态特征,并利用情感分布建模定位所需的伪标签. Zhang 等[75]设计跨模态时域擦除网络,捕获关键帧和非关键帧的互补信息. 除了分类以外,情感的时域定位在实际应用中也具有重要意义,尤其是长视频中传达情感的片段位置与时长不定.

视觉复杂度即图像中包含的信息量和视觉元素的数量. 对于人类来说,图像复杂度是理解图像的基本视觉线索之一,因此评估图像复杂度可更好地模拟人类感知并提高计算机视觉任务的性能. 其应用场景包括图像分割、图像隐写、网页设计、文本检测和图像增强等. 早期的复杂度研究关注于复杂度的表示方式,如图像熵[76]、边缘信息[77]或颜色[78]. 此外,数据驱动的深度学习在复杂度评估领域展现令人印象深刻的性能. Chen 等[79]研究基于纹理、边缘和区域的神经网络,评估图像的复杂度,Saraee 等[80]使用深度网络中间层特征进行视觉复杂度分析.

为了进一步促进图像复杂度领域的发展,Feng 等[81]设计由 9 600 幅图像组成的、实例级的、经过精心注释的图像复杂度基准数据集 IC9600,涵盖抽象、广告、建筑、物体、绘画、人物、场景和交通等多种类别. 每幅图像都经过 17 个人的精细标注,目的是降低主观性. 作者还提出 ICNet,可在弱监督的情况下预测图像的复杂度得分,生成复杂度密度图. 此外,ICNet 可协助提升计算机视觉任务的性能,如图像美学评估、人群计数和显著性目标检测等.

相比非记忆性图像,记忆性图像往往具有更明显、更容易识别的模式,这是因为具有强烈模式的图像更容易在人的脑海中留下深刻印象. 因此,模式识别在决定图像可记忆性方面起着重要作用. Khosla 等[82]构建大型数据集 LaMem,包含 60 000 幅带注释的图像. 通过使用 CNN,他们发现经过微调的深度特征是预测可记性的最佳指标. 此外,他们还发现哪些物体和区域与可记性呈正相关和负相关关系,能为每幅图像创建可记性的分布图. 从文献[83]开始,已有多项研究证明,训练神经网络的输出可用于成功执行成员推理攻击,即高精度地推断给定数据点是否属于训练集的一部分. 该领域的一个重要问题是找到更能抵御此类隐私攻击的学习算法. Arpit 等[84]研究随机标签的记忆与网络在真实标签上的性能之间的关系. 研究表明,使用各种正则化技术可

降低算法拟合随机标签的能力,而不会显著影响其在真实标签上的测试准确性.

图像质量描述从物理世界拍摄的数据的失真程度.感知技术深受图像失真的影响,一幅具有清晰边界的高质量图像可容易地被感知算法识别.为了评估图像质量,早期工作聚焦在对比理想场景下收集的图像与拍摄图像之间的差异,出现大量指标,如SSIM(Structure Similarity Index Measure)[85]、PSNR(Peak Signal-to-Noise Ratio)[86].这些指标根据图像的内容差异和边缘的结构信息进行估计.而考虑到人类对于局部结构信息的认知,LPIPS(Learned Perceptual Image Patch Similarity)[87]利用深度卷积神经网络分别提取参考图像和拍摄图像的局部特征,再计算两种深度特征之间的余弦距离.更进一步,Ding 等[88]基于深度特征的空间相似性和结构相似性评估图像质量,以研究影响人类感知的关键因素.研究表明人类视觉系统不同于现有的逐像素比较图像差异的指标,能从视觉纹理中进行度量,如空间上同质的区域.与此同时,研究者提出盲图像质量评估,直接估计输入图像的质量分数[89-92].Kang 等[93]提出首个端到端的基于卷积神经网络的盲图像质量评估方法,直接优化回归损失以预测图像的质量得分.更进一步,Pan 等[94]利用全卷积神经网络以及池化网络,同时预测图像质量的得分以及对应的逐像素质量图.根据 FCNN(Fully Connected Neural Network)生成的中间质量图,池化网络通过压缩空间维度以预测位置无关的图像质量得分.Su 等[91]提出自适应的超网架构,适应于通用场景下的图像质量评估.作者设计内容理解超网和质量得分预测网络.内容理解超网根据提取的语义特征自适应预测质量得分网络的权重,再将骨架网络提取的多层特征输入质量得分预测网络,完成最终的预测.更进一步,Roy 等[89]提出基于批次和样本层级的测试阶段增强方法,在批次级别的多个样本中进行分组对比学习以及在样本的不同退化实例之间进行排序学习,完成测试阶段的自适应质量评估.

图像美学是一种描述视觉吸引力的图像属性,是一个与图像质量、图像复杂度类似的主观概念.图像美学的评估受主客观因素影响:客观因素有图像的色调、纹理、色彩丰富度等因素;主观因素包括观察者的情感反应、个人偏好、文化背景等因素.图像美学的评估可模拟人类的视觉感知和情感反应,进而提高计算机视觉系统的性能.

图像美学的应用非常广泛,包括摄影和图像设计、网页和用户界面设计、市场营销和商业广告等.Murray 等[95]建立 AVA(Aesthetic Visual Analysis)数据集,是目前最大的图像美学数据集,包括大约 250 000 幅图像,涵盖 66 个语义类别,每幅图像都标注美学分数和摄影风格类别.该数据集已广泛应用于图像美学评估之中.对于图像美学评估的模型,Zhang 等[96]提出一个门控视网膜卷积神经网络,用于模仿人类的审美感知机制,还可以同时编码整体信息和细粒度特征.

1.3 现有工作回顾与挑战

受下游感知任务的需求引导,研究者从数据中提取视觉结构的几何属性和图像对应的认知属性作为先验知识.现有工作通过数据驱动的方式,从开放场景中收集数据并构建大规模数据集.基于构建好的数据集,知识提取算法可在使用大量标注情况下,基于监督信号学习到图像的属性,提取数据中蕴含的知识.一些研究基于此探索如何在多个属性之间进行关系建模,并将其应用于目标检测和图像分割任务中.此外,由于属性知识与感知任务的类别无关,通常可覆盖多个感知任务,如关键点、边缘可用于目标检测、图像分割、视频理解等多个感知任务,因此受到广泛的关注.

本节回顾现有的几何基元属性感知与认知基元属性感知的代表性工作.作为计算机视觉的经典问题,基元属性感知具有悠久的研究历史.基元属性感知可分为视觉结构的几何属性和视觉图像的认知属性.视觉结构的几何属性包括角点、线段、边缘、平面和显著区域等,研究的感知任务覆盖检测、分割、重建、关联、描述及跟踪.而视觉图像的认知属性包括情感、复杂度、记忆性、图像质量和图像美学等,研究的感知任务覆盖分类、检测、回归、分割、排序.这些属性可作为自适应视觉感知与结构理解后续步骤的输入,辅助感知模型的训练和学习过程.

总之,从视觉数据中提取几何属性和认知属性作为先验知识是一种有效方法,可提高下游感知任务的准确性.然而,这种方法仍然存在一些挑战.如何准确提取多模态的属性信息、如何处理开放环境下数据场景变化等都是值得探索的问题.

2 属性知识引导的弱监督视觉感知

属性知识引导的弱监督视觉感知旨在利用大规

模未标记或粗糙标记的图像配合少量良好标记数据以训练深度神经网络. 其中,属性知识可来自领域专家、从其它相关任务中获得、从视觉对象中自动提取. 通过使用属性作为先验知识引导弱监督学习,可在不需要大量标记数据的情况下有效训练深度神经网络,从而提高计算机视觉任务性能. 如图 2 所示,本节围绕弱监督识别、弱监督检测与弱监督分割展开,介绍针对不同场景的现有解决方案.

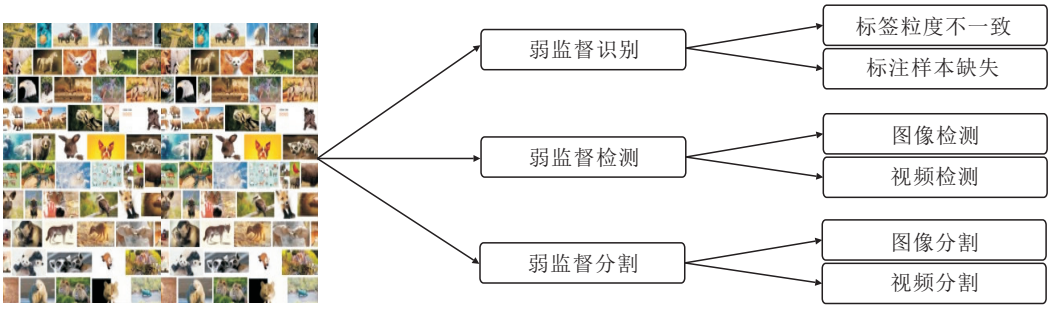


图 2　属性知识引导的弱监督视觉感知

Fig. 2　Attribution knowledge-guided weakly-supervised visual perception

2.1　弱监督识别

模型训练作为模式识别的关键技术,研究者关注如何有效利用数据的标签. 然而在开放场景中,某些数据集因其规模较大而面临难以标注的挑战,无法获得足量完全的数据标签. 为此,弱监督识别使用不完全准确的标签数据训练模型,涉及到两个关键问题:1)如何使用不正确的数据标签;2)如何使用不完全的数据标签.

在通常情况下,互联网图像数据集是通过爬虫软件从多个不同的社交平台等网络站点获取的. 这些爬取的图像数据受限于用户提供的关键词,标签信息往往具有不完备和不准确的特点,导致数据集上存在不理想标注,如仅有粗粒度的标注或仅有部分数据被标注.

当数据标注粒度不足[97]时,通常只有原始的网络 tag 标签可用. Nayak 等[98]提出一种弱监督识别方法,使用组级别的二进制标签作为弱监督信号,训练实例级别的二进制分类模型. 他们将组级别标签建模为适用于单个实例的类条件噪声(Class-Conditional Noisy, CCN)标签,并使用噪声标签规范在强标记实例上训练模型的预测.

更进一步针对细粒度识别,Xu 等[99]讨论当只有粗略分类标签时,学习目标任务的细粒度模式的挑战,利用粗略类别的信息,减弱该挑战带来的负面影响,学习适当的表示并用于目标任务.

当仅有部分数据被标注时,半监督学习利用数据标签的密度作为指导,引导模型构建伪标签进行训练. Jiang 等[100]提出 GPENs(Graph Propagation-Embedding Networks),将特征传播和低维嵌入同时集成到一个网络中,实现图结构数据的紧凑表示. Jia 等[69]设计 S^2-VER(Semi-Supervised Visual Emotion Recognition),算法包含可信情感标签学习和模糊感知自适应阈值策略. 可信情感标签学习策略计算维护的情感原型与样本嵌入之间的相似度,生成平滑标签,提高伪标签的准确性;模糊感知自适应使用信息熵衡量平滑标签的模糊程度,然后自适应调整阈值,选择高置信度的未标记样本.

2.2　弱监督检测

传统检测任务[101-103]需要训练数据标注每个目标的位置和类别,与之不同,弱监督检测[104-105]只需使用不完全的标注,如图像的标签或图像中是否存在目标,就可训练模型. 属性引导的弱监督检测采用不同属性知识作为先验,以此指出或矫正物体可能出现的位置,补全给定的不完全标注,指导模型的学习和训练.

图像中的弱监督检测通过图像级别的标注(如图像中存在的物体类别)训练模型. 在弱监督检测的早期工作中[106],研究者将显著性作为属性知识[107-108],协助估计物体可能出现的位置,并进一步训练检测模型.

Shi 等[109]提出基于几何与外观知识的贝叶斯联合主题建模框架,联合编码多个目标的共现特征,并通过弱标记和未标记的互联网图像数据实现混合学习. 不同于直接指定物体位置,Cinbis 等[110]通过

多实例学习,将图像划分成多个区域,根据每个区域中的特征点迭代检测,矫正物体可能出现的位置并完成训练. Deselaers 等[111]借助类间的语义信息挖掘目标位置,并通过网络训练循环更新可能的候选区域.

近年来,越来越多的工作[112-113]开始利用其它领域的预训练模型作为知识的提取器,辅助检测模型的学习. Shi 等[114]利用物体的先验知识,辅助弱监督学习过程,其思想是基于源域数据集训练一个语义分割模型,在目标域数据集进行弱监督的学习和推理. Zhang 等[115]基于预训练深度特征的协作式课程学习网络,定位感兴趣的对象.

在视频弱监督检测任务中,一段几秒钟的短视频就可包含上百幅图像. 因其增加额外的时间维度,使获取视频数据的完整标注信息变得更困难. 因此,近期的研究工作开始关注利用视频级别的标签,提取时间一致性作为先验知识,引导模型在学习过程中更好地理解视频数据. 这种方法可帮助模型更准确地定位和检测对象,提高检测性能.

Zhang 等[74]提出 TSL-Net,每个片段只需标记一帧,使用贪婪搜索策略为未标记帧生成伪标签,再融合视觉和音频模态的特征,预测时间标签分布. Sang 等[116]提出从单帧图像中估计物体 6D 姿态的网络框架 INVNet,结合不可见信息与可见的 2D-3D 对应关系,建模物体的几何特征. INVNet 通过跟踪生成密集的可见对应关系以及几何的路径图,以此构建在视频所有帧中的物体姿态伪标签并用于训练.

2.3 弱监督分割

分割的目的是在像素级别进行精细分类,因此需要更细节的标注信息. 现有的弱监督分割工作除了使用图像级别的标签以外,还利用大量属性知识,如物体类别、关键前景点、线条,提供有效且容易标注的弱标签,辅助模型训练.

由于弱监督分割[117]需要在像素层级定位和区分物体,因此带来新的挑战. Liu 等[118]提出弱监督实例分割的方法,将所有训练数据的粗糙标注信息汇集成一个大型知识图谱,再从该图谱中利用语义关系辅助后续处理. 作者提出多实例学习框架,同时计算每个候选区域的概率分布和类别感知的语义特征,并使用这些特征构建一个大型无向图. 该图的最优多路切割可为每个候选分配可靠的类别标签. KnifeCut[119]通过用户在误分割的细小部分上绘制一条线条,作为分割边界,从而有效实现细小部分的分割. 相比传统的交互式图像分割方法,KnifeCut 更直观易懂,用户只需进行低强度的交互操作即可完成任务,不再需要进行复杂的点击、涂鸦或多边形绘制等操作.

针对类别噪声问题,侯淇彬等[120]设计一种噪声擦除模型,以跨样本的方式,从小批次样本的置信区域学习语义信息,实现对图像中与类别无关的区域的擦除.

在下游应用中,医疗图像分割由于其数据的私密性,难以获得准确的标注. SANet(Slice-Aware Network)[121]是一种肺结节区域分割的切片感知网络. 作者建立大规模的 PN9 数据集,包含 8 798 幅 CT 扫描图像和 40 439 个标注结节. SANet 利用切片分组非局部模块,捕获特征图中任意位置和任意通道之间的长程依赖性. 引入 3D 区域提议网络,生成具有高灵敏度的肺结节候选项,而检测阶段通常伴随着许多假阳性. 为此,作者使用多尺度特征图生成一个假阳性消除模块.

相比图像,在视频中分割物体更难,这是因为视频的语义信息很难被直接用于指导分割过程. 大量弱监督视频分割研究工作关注于仅使用第一帧图像标注的情况下如何有效训练模型[122-124]. Wang 等[125]提出一种时间环一致性框架,将第一帧的标注传播到后续帧,并回传到第一帧,以此计算时间一致性损失. Li 等[126]进一步利用视频内的重要区域和关键点,分别从目标和像素级别构建区域跟踪和像素传播的自监督任务,并构建渐进式训练策略,耦合两个自监督任务,完成弱监督学习.

相比使用掩码,坐标框因其更容易标注而在检测领域被广泛使用,但其只能提供有限的目标位置信息,并且引入额外的背景像素. 为了解决背景混淆,Yan 等[127]设计 STC-Seg(Spatio-Temporal Collaboration for Instance Segmentation in Videos),利用深度和光流信息,从目标框中去除背景像素并构建伪标签. 在弱监督训练过程中,STC-Seg 考虑伪标签的置信度,将掩码切分为多个区域,分块计算拼图损失.

Lin 等[128]讨论跨帧的目标框表征一致性,将两帧的目标框表征通过双边聚合模块进行特征对齐. 为了获取更精确的物体位置信息,Liu 等[129]利用关键点作为弱标签,设计记忆蒸馏策略对齐跨帧的关键点表征.

2.4 现有工作回顾与挑战

本节回顾属性知识引导的弱监督视觉感知的代

表性工作. 弱监督学习是开放场景下数据标签缺失的代表性学习方法,通常受制于标签质量,模型难以取得令人满意的准确率. 属性知识指导的弱监督学习通过提取任务无关的先验,指导模型学习. 基于提取的基元属性,一些研究通过属性知识将弱监督标签转换为具有全监督标签形式的伪标签,以此指导模型学习,如通过显著性图将分类标签转换为分割任务所需的掩码标签. 另一些研究则聚焦在学习范式上,以属性知识为评价指标,设计损失函数、学习策略,提高模型性能. 因此,通过属性知识引导的弱监督学习可在不需要大量标注数据的情况下,学习到图像中物体的属性信息,进而提高模型的准确性.

然而,在现有的弱监督视觉感知研究工作中,常用的模型架构与弱监督学习范式由于引入额外的结构以提取属性知识,虽然可以提高模型性能,但属性知识的提取产生额外的计算开销,延长训练所需的时长.

3　图像无监督自主学习

开放场景平台作为一个巨大的图像数据源,为各种复杂任务,如动作分析、场景图生成等,提供丰富的数据支持. 然而,传统的图像数据标注和监督学习方法却伴随着昂贵的时间成本以及庞大的人力资源投入,这成为训练模型的制约因素之一. 为了应对这些挑战,无需标签的无监督/自监督学习引起学者的高度关注.

本节聚焦于图像无监督自主学习方法,旨在克服传统方法面临的种种限制. 如图 3 所示,深入探讨图像无监督自主学习中自监督对比学习及无监督共性学习这两个关键子领域,它们在推动图像自主学习方面具有重要作用.

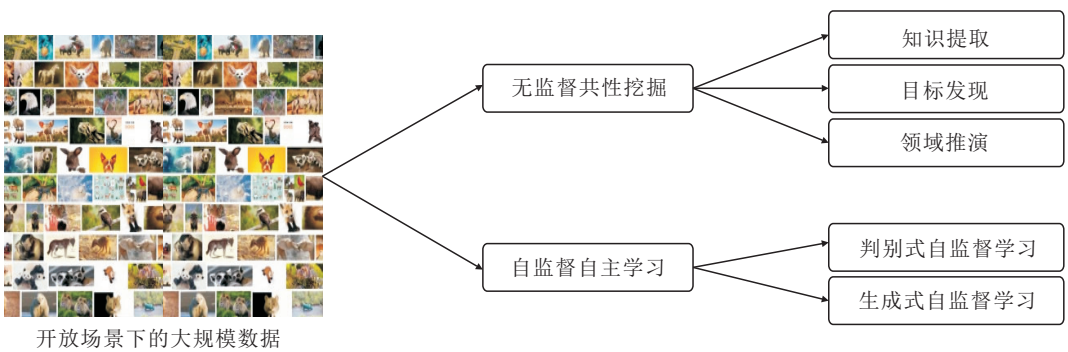


图 3　图像无监督自主学习

Fig. 3　Image unsupervised autonomous learning

3.1　无监督共性挖掘

共性学习强调在大规模图像数据集上挖掘共同特征的重要性. 通过发现图像集合中的普适模式,建立跨样本的联系,为跨任务的学习提供更丰富的信息支持. 这种共性体现在物体形状、轨迹等语义特征上的共性、不同数据域的共性等方面.

通常使用聚类或相似度等方法挖掘数据之中共现的属性知识,如关键点、显著区域. 自监督关键点提取以多视角图像数据之间的关键点共现规律为线索,提取图像中不变的关键点.

SuperPoint[26]构建单应变换引导的多视角图像对,在图像对之间提取关键点匹配关系作为伪标签,预训练关键点检测网络和关键点表征网络以提取关键点. 协同显著性检测利用图像中显著物体的共性特征,在相关图像组中分割共同显著的前景物体. Zhang 等[130]借鉴人类行为,提出基于梯度诱导的共同显著性检测方法,将单个图像和一组图像的一致表示进行对比,利用反馈的梯度信息,将更多注意力放在引起协同显著性的区域. Li 等[131]结合多目标跟踪方法与基于轨迹的聚类方法,构建视频摘要系统,提升对监控视频分析的速度和准确性.

在没有标签的情况下,研究者试图从图像或视频中自动发现物体,对其进行分类和定位. Slot Attention Module[132]作为一种与感知表示对应的架构组件,可产生一组任务相关的抽象表示,称为 slot. 这些 slot 是可交换的,并且可通过多轮注意力竞争过程专门绑定到输入中的任何对象. 研究结果表明,Slot Attention Module 可提取以对象为中心的表示,在无监督对象发现和监督属性预测任务上进行训练,并实现对未见组合的泛化.

在大规模的开放数据上,以往无监督的聚类方法或概率模型的尺度泛化性有限. 为了解决这一问题,Vo 等[133]提出无监督目标发现的方法,将目标发现转化为排序问题,并使用自监督特征进行处理. 该方法在单个物体和多个物体的发现设置中都表现出色,并适用于大型数据集.

为了更精确地理解图像中的语义,无监督语义分割也获得更多关注. Gao 等[134]不仅提出大规模语义分割的数据集,还构建 PASS 作为基准模型. 模型自动学习形状和类别表示,应用基于像素注意力的聚类方案,获得伪类别,并将生成的类别分配给每个图像像素.

利用无标记数据减少数据跨域影响,可提高模型的泛化性能. Yao 等[135]提出包含多个源域网络和一个伪目标域网络的架构. 在这种架构中,不同源域网络的参数以加权的方式搜索目标子集的最优参数,并提出候选区域网络的一致性正则化,促使不同域的子网学习更多抽象的域共性.

3.2 自监督自主学习

自监督学习旨在从无标注的数据中构造伪标签以训练有效网络,以此完成网络的自主学习. 通过引入额外的数据结构知识,如对不同图像样本之间的关系、数据增强对的相似和不同情感极性数据的差异进行建模,从而学习更具有判别性的特征表示.

通过对比正负样本,可挖掘图像数据中用于感知任务的模式,进而提升学习算法的性能,在缺乏标签信息的情况下,能有效捕捉数据的内在结构. 这种对比体现在语义的差异、情感的两极性等方面. Qian 等[136]在视频表示学习中,利用朴素对比学习和原型对比学习构建分布图,指导低中级特征的学习过程. Chen 等[137]提出 SSL++,利用下游任务学习的低级通用特性与生成语义伪标签的高级语义特性之间的互补性,缓解学习表征在不同下游代理任务中产生的特异性和局限性. Yao 等[138]构建一个极性敏感的嵌入网络,使用强推力、弱推力、拉力三种损失类型,充分考虑标签的极性内关系和极性间关系. Huang 等[9]提出类特定语义重建策略,集成自编码器和原型学习的能力.

生成式自监督学习旨在重建数据本身的特征和信息,引入数据的底层结构知识,使用数据集本身的信息构造伪标签. 早期的自回归式模型可视作贝叶斯网络结构,通过最大化前向自回归分解的似然函数,逐像素建模图像的概率分布[139-140]. 考虑到直接优化似然函数(即概率密度)是一件困难的事情,Flow 类的模型通过一系列的几何变换函数描述不同的数据点[141-142].

不同于上述方法,自编码器类模型[143]通过训练前向传播网络,在输出层预测其输入,目标由最大化似然函数转变为最小化输入与编解码的结果,最近,掩码自编码器(Masked Autoencoders, MAE)通过恢复被隐去的部分数据结构,如图像的关键区域、视频的部分关键帧等,有效完成自监督学习[144]. 掩码学习在自然语言中早有应用,经典的 BERT(Bidirectional Encoder Representations from Transformers)通过预测下一句文本建模语句之间的顺序关系[145]. 受其启发,MAE 将图像切分成块并随机进行掩码,使模型学习不同区域之间的视觉底层结构. 更进一步,MaskFea(Masked Feature Prediction)[146]研究中层的视觉知识表征,要求模型恢复具备语义信息的结构,包括像素、边缘 HoG(Histogram of Oriented Gradients)、网络提取的特征等. 不同于视觉可见的特征,A^2MIM(Architecture-Agnostic Masked Image Modeling)[147]引入频率域约束,对齐 Transformer 自监督学习与 CNN 自监督学习的性能,缓解 MAE 难以拓展到 CNN 架构的局限. 除了视觉知识以外,MILAN(Masked Image Pretraining on Language Assisted Representation)[148]更进一步设计视觉-语言跨模态模型 CLIP(Contrastive Language-Image Pre-training)引入语言知识,要求模型恢复 CLIP 输出的语言特征,并在特征层次进行对齐.

3.3 现有工作回顾与挑战

本节回顾图像无监督自主学习的代表性工作. 计算机视觉的自主学习方法主要包括无监督共性学习和自监督自主学习两个关键子领域. 其中,无监督共性学习方法基于大规模的开放场景数据,挖掘无监督数据中的属性知识、共现规律,并进一步泛化到多个场景. 自监督自主学习根据从数据中提取的属性知识,构造伪标签以训练模型. 两者均在无可用标签的情况下,利用数据的属性知识进行特征学习和模型训练. 现有无监督自主学习方法仍然依赖长耗时的训练过程,比监督式学习方法更难以收敛,因此有待后续工作研究解决.

4 场景图像的结构化表示和理解

场景图像的结构化表示和理解一直是计算机视觉领域的重要研究方向. 如图 4 所示,本节重点关注

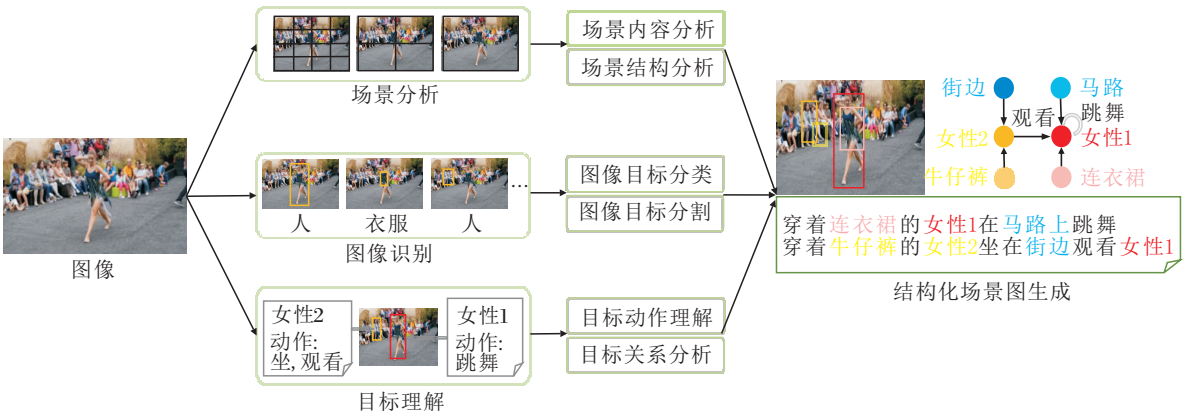


图 4　场景图像的结构化表示和理解

Fig. 4　Structured representation and understanding of scene images

场景分析、图像识别和目标理解等关键问题,通过更完备的结构化表示,提高模型对场景图像的理解能力.

4.1　场景分析

场景内容分析旨在理解当前图像的上下文信息,提取图像拍摄的地点等属性. 根据图像所处的环境、图像中包含的对象类别以及各对象的布局关系,将场景图像分类为预定义的场景类别之一(如图书馆、海滩、商场等)[149]. 作为场景分析的重要环节,场景分类引导智能系统对视觉场景中出现的物体、动作或事件等要素进行重新思考,构建物体与物体以及物体与环境之间的联系. 通过掌握不同物体的语义类别以及物体间的组织关系,智能系统构建对场景的全面理解,能服务于自动驾驶[150]、智能机器人[151]、智慧城市[152]等下游应用.

场景内容分析任务根据数据来源及分类目标不同被分为室内/外场景分类[153-154]、遥感场景分类[155-156]、声学场景分类[151,157]、地点分类[158-159]等子任务. 这些任务各自面临着领域内特有的挑战. 例如:室内场景受物体摆放位置和拍摄角度影响较大、遥感场景的图像空间分辨率较低、声学场景分类需要考虑声音的时变性和时序性等,但仍在存在如下共性问题.

1)类内变化较大. 同一类别的场景可能呈现出完全不一致的特征.

2)语义模糊. 即使包含相同的视觉特征,但姿态、布局的差异可能造成不同场景下的同一物体语义上的差异,影响对场景类别的判断. 此外,类别标签本身依赖于主观注释,不同的场景标签并非完全互斥,同一场景可能与多个标签相关.

3)计算效率. 下游应用一般服务于移动场景,对计算资源限制较大.

传统场景分类方法常被分为特征检测、特征描述和特征分类三个阶段,在特征检测阶段,学者们提出大量的特征检测器,包括:SIFT[160]、FAST(Features from Accelerated Segment Test)[161]、SURF(Speeded-Up Robust Features)[162]等. 考虑到手工特征在面对复杂场景时的性能退化,传统方法已逐渐被具有强大特征提取和学习能力的 CNN 取代. 即使在深度学习方法中,研究人员对特征提取的关注也远大于分类阶段,最新的工作[163-164]通过构建更强大的特征提取网络,得到视觉特征的更优表示,服务于场景理解的下游任务. 未来的工作将更关注于场景理解领域的共有特性,构建更通用和可扩展的网络框架,统一各个子任务.

场景结构感知对场景中的物体和结构进行感知,以获取丰富和准确的场景信息. 然而直接从图像中理解场景结构需要大量的先验知识,如物体的类别、形状、尺寸以及场景类型、朝向等. 现有方法利用深度信息,也就是图像中物体到相机的距离,作为辅助信息感知场景的结构. 在传统的计算机视觉中,深度信息通常通过使用双目或多目相机等硬件设备获取.

最近,深度估计方法可通过识别物体的轮廓和特征点自动确定物体的位置和方向,从而估计图像的深度信息,这在计算机视觉领域、机器人、增强现实等领域具有广泛的应用前景. 张羽丰等[165]提出基于双目视差回归的目标距离估计方法,基于区域卷积神经网络,同时进行目标检测和目标距离估计. 通过双目图像输入网络,提取区域特征,利用双目视

差回归算法计算目标距离,并将结果通过双目包围框进行输出. 该方法可有效解决传统双目目标距离估计方法精度较低或数据准备困难等问题. Zhang 等[166]提出基于 R-CNN 结构的区域回归网络,用于实现单目物体距离估计. 网络通过添加浅层网络处理相机外参参数,并优化特征处理结构,改善估计结果.

4.2 图像识别

图像识别是计算机视觉的基础任务,旨在确定图像中物体的类别和位置. 在场景图像中,准确识别不同物体的存在对于进一步理解和分析图像至关重要.

判别区域定位和特征学习对细粒度视觉识别至关重要. Liu 等[167]集成网格门注意力单元、尺度一致的注意力部件选择策略和部件关系建模模块,充分利用注意力机制与局部部件之间的一致性以及部件之间丰富的关系信息,实现细粒度识别. 开放环境下的图像种类多样,高分辨率的图像类型也在监控等领域发挥重要作用. 以往针对高分辨率图像识别任务是简单地将图像切割成小块,Fan 等[168]提出小块排列网络,通过确定哪些小块可以打包成一个紧凑的画布以加速检测. 随后,从剩余的小块中,决定如何将这些小块打包成更小数量的画布. 这些画布被分别送入检测器,得到最终结果.

目标分割是将图像中的每个像素分配给特定的物体类别,从而获得物体的精确边界. 这对于场景理解和图像编辑等任务具有重要意义. 传统的图像分割方法通常考虑分割结果在像素级上的覆盖率或交并比等. Guo 等[169]证明卷积注意是比自注意机制更有效的编码上下文信息的方法,并据此提出 SegNeXt. Wu 等[1]将金字塔池应用于视觉转换器的多头自注意中,同时减少序列长度并捕获强大的上下文特征,构建一个通用的视觉变压器骨干,称为金字塔池变压器. 它在图像分类、语义分割、目标检测和实例分割等各种视觉任务中表现出实质性的优势. 在下游应用中,语义分割面临更多的挑战,如避免遥感图像中碎片化的分割结果[170]或是针对细胞完成精细的分割[171]. 为了提升道路抽取的拓扑正确性,Mei 等[170]从道路形状角度出发,提出条带卷积模块,并从连通性角度出发,考虑到建筑物和树木的遮挡,提出连通性注意模块,探索相邻像素之间的关系.

4.3 目标分析与理解

视频动作分类可直接理解物体动作,将视频中目标发生的动作按照类别进行分类. 由于目标动作的复杂性,通常需要使用深度学习等技术,对视频进行特征提取和分类. Li 等[172]将事件边界作为指导时间对齐的先验,设计基于时间边界的帧采样策略,减小类内方差,此外,引入一种边界选择模块,将视频特征与其动作持续时间进行对齐. Li 等[3]还构建两阶段动作识别模型,先定位动作的时间起始范围,降低动作持续时间估计偏差,再学习运动变化特征,降低动作演化偏差. 对于动作识别问题,Liu 等[173]注重帧选择策略,通过时间选择器选择更重要的视频帧,通过空间放大器寻找关键帧中最显著的部分.

更进一步,视频定位沿时间轴定位动作发生的起止时间,可提供更丰富的信息. Li 等[174]提出用于有效的时空动作定位的检测器,首先从视频流中估计粗略的时空动作变换体积,再根据抽取的关键时间戳进一步精细化该体积. Li 等[175]提出 LSTC(Long-Short Term Context),为了更准确地进行原子动作检测,将动作识别线索独立分解为信息丰富的短期依赖和用于交互推理的长期依赖.

除了仅通过视觉模态进行定位以外,多模态视觉理解与定位提供丰富的判别信息以提高动作理解的准确性,通过使用包括文本、音频、温度在内的额外模态作为辅助信息进行动作理解和定位.

Qian 等[176]利用视觉和音频模态解决在非受限场景下一次性定位多个声源位置的问题,从复杂场景中分离不同类别的视听表示,在不同粗细粒度的尺度上进行模态间特征对齐. Guo 等[4]根据视觉问答中可能包含对图像的描述,提出一种重新注意机制框架,对答案提供的信息进行注意力修正,将视觉注意力图重新定位到正确的位置. Li 等[177]利用 RGB 视觉和温度模态,解决热成像下的目标跟踪问题,提出的挑战感知神经网络通过参数共享分支解决模态共享的问题,通过参数独立分支处理各模态的特定问题.

4.4 结构化场景图生成

场景图是一种用于描述场景中的对象、属性和对象关系的结构化表示方法. 作为图像的一种语义化表示,图像中的目标对应场景图中的节点,目标间的关系对应场景图中的边[178]. 为了尽可能涵盖图像更多的细节,有时图中还会表示目标的属性,如颜色、状态、大小等. 尽管场景图将图像语义化表示,在形式上与知识图谱十分相似,但场景图强调对象主体,而知识图谱强调语义标签[179].

场景图自从被提出后,得益于其直观、高效、模

态无关的特性,广泛应用于推理、检索和推荐等领域,在多模态信息融合领域表现出巨大的潜力. 深度学习领域的飞速进展造成对大规模数据集的迫切需求,基于目标检测数据集人工标注场景图的方式逐渐被抛弃,如何自动生成高质量的场景图成为热门的研究任务.

场景图生成旨在对图像或视频进行解析,生成一个结构化表示,弥合视觉特征和人类感知之间的差距,帮助智能系统理解现实视觉场景. 场景图的生成需要检测到视觉特征中的所有实体并挖掘实体间的视觉关系,这是一个自下而上的过程,从图像的基本元素出发,逐渐建立对其的感知[180].

现有场景图的生成有两种主流方案[181].

1)将场景图生成进行拆分,视作一个两阶段任务. 首先检测图像中的对象,得到场景图的节点及每个对象的属性,再推断对象之间的关系.

2)先定位目标,再构建一个未标记的图结构,从图的角度分别对节点和边进行类别预测.

进一步,根据采用的技术手段,场景图生成又被分为 4 类:基于 CRF(Conditional Random Field)[182]的方法、基于 TransE(Translating Embeddings)[183]的方法、基于深度学习的方案、引入外部先验的场景图生成方法.

CRF 是一种将统计关系纳入判别任务的工具. 研究人员基于场景图中的关系谓词和对象之间存在的统计相关性推理对象间相关性并给出有效性证明. 受此启发,DRNet(Deep Relational Network)[184]、SG-CRF(Scene Graph Generation via CRF)[185]等模型都借用这种基于频率的方式进行关系预测. 基于 CRF 的方法虽然有效,但忽略图像的上下文信息,无法准确区分同一类别的多个实例.

基于 TransE 的方法是从场景图和知识图谱的相似性层面考虑的. TransE 已经被证明在将知识图谱中的三元组嵌入到低维向量空间中的有效性,场景图拥有与知识图谱类似的定义和属性,在视觉关系嵌入层面应共享有效性. 因此,研究人员尝试使用 TransE 将场景图的相关元素进行低维嵌入,并建模不同元素间的视觉关系,相关的研究结果表明这种方案的有效性[186-187].

深度学习的快速发展为场景图生成提供更有效和可靠的方案. 一个通用的框架是先使用常规的目标检测器,分别获得图像中的不同物体,再对不同物体进行特征提取,包括视觉特征、空间关系、语义信息等,得到多个中间表示. 然后考虑多个特征表示之间的差异与关联,应用注意力机制调整权重,最终对关系进行预测.

除了常规架构以外,一些特殊架构也被认为对于建模物体间关系具有优势. 例如:RNN(Recurrent Neural Network)的多次迭代能融合全局信息,这种考虑上下文的方案可减少关系预测的模糊性[188];基于 GNN 的方法借助图论改进关系图的生成[181]. 更进一步,纳入更多人为先验或自然常识被认为有助于减轻模型推理物体间关系时的负担,并能有效提高结果的准确性. 语言先验[189]、统计信息先验[183]都被证实能指导模型进行更准确的关系推理.

4.5　现有工作回顾与挑战

场景图像的结构化表示和理解是计算机视觉领域的一个重要研究方向,它将图像中的物体、关系和属性呈现为结构化的形式. 传统的场景图像表示方法主要基于手工设计的特征和规则,在复杂场景下往往难以扩展和适应. 近年来,随着深度学习技术的发展,基于深度神经网络的场景图像表示方法逐渐成为主流.

虽然场景图生成已被广泛研究并构建一系列解决方案,但仍面临一些挑战. 从场景图本身来说,有限的关系种类、浅层语义的关系表示、难以对节点进行实例级别的区分等问题,都限制场景图的实用性. 另外,视觉图像中包含的信息难以支撑复杂的关系推理,这促使研究者进一步探索更多种类的额外先验以协助模型生成场景图. 最后,和大多数任务一样,数据规模和细粒度都限制深度学习方法性能的进一步提高,探索更高效的数据标注方案或有限数据的深度学习方法值得进一步探索.

5　未来研究方向

1)大规模多属性知识数据集. 在开放场景中,同幅图像涵盖的属性知识多样,而现有基准数据集均针对单一属性知识构建,难以研究属性之间的相关性. 因此大规模的多属性知识数据集具有巨大的研究价值. 在构建大规模属性知识数据集时,需要考虑如何获取包含丰富属性信息的数据,并对数据进行标注和质量控制. 同时,还需要考虑如何将这些数据转化为结构化的形式,以便指导后续感知任务的分析和应用. 与普通单一属性知识数据集不同,多属性数据集在构建过程中需要把多种属性知识之间的

影响考虑在内,避免多种属性捆绑出现的偏置情况.

2)多模态属性知识提取. 现有工作的属性知识提取聚焦在单一模态(主要是视觉模态),然而开放世界中可大量获取的数据通常涵盖多种模态,包括图像、文本、语音等. 针对这种场景,需要从实际应用,如模态对齐、图文匹配中,总结所需的模态知识,如视频多模态数据的时间一致性、图文数据的语义性. 常见的提取方法涵盖跨模态关联学习、多模态生成模型及强化学习等. 因此,多模态属性知识提取是一个复杂而又具有实际意义的研究问题,需要结合不同领域的技术和方法进行研究和实践.

3)属性知识感知模型的场景泛化性. 在开放场景的图像中,属性知识的提取和表示需要考虑到不同场景和环境的变化,以保证模型具有良好的泛化性能. 设计和优化属性知识感知模型,提高其在不同场景下的泛化性能成为当务之急. 这个问题涉及 3 个方面:基准数据集的数据多样性、评估指标的设计合理性以及感知算法的鲁棒性. 此外,对于开放场景下的感知问题,还可从感知模型结构设计、特征提取、学习算法、知识表示和推理等不同角度提出解决办法. 相关研究工作包括可解释性深度神经网络、小样本学习等.

4)轻量级属性知识引导的视觉感知模型. 在设计视觉感知模型的过程中,由于引入属性知识,模型通过中间结果取得更好的可解释性和预测表现,也同时引入额外的结构进行属性知识的提取,因而需要进行轻量化模型的设计,并考虑如何保证模型具有足够的感知能力. 此外,属性知识的学习与物体语义的理解不同,其类别无关且更具有一般性,通常所需的模型参数量大幅小于常规的感知模型,如显著区域提取可仅用 100 K 参数的模型完成.

6 结束语

本文从属性知识的角度切入,综述开放场景自适应感知领域的研究进展. 本文并未囊括领域内的所有方法,而是关注一些有代表性的方法. 总结和对比广泛使用的属性知识提取方法、弱监督与无监督视觉感知模型、图像结构化表示和理解方法. 最后,讨论属性知识引导的自适应感知与结构理解的一些开放性问题和潜在的研究方向.

尽管近年来属性知识的应用取得快速发展,但是仍未出现一个完全解决开放场景下有效、高效、鲁棒的多模态多属性知识引导的算法框架. 随着领域专家知识的不断扩充、深度感知模型的快速发展,属性知识引导的自适应感知会在之后很长时间内保持活跃,作为计算机视觉与人工智能领域的前沿方向与研究热点之一.

参 考 文 献

[1] WU Y H, LIU Y, ZHAN X, *et al*. P2T: Pyramid Pooling Transformer for Scene Understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(11): 12760-12771.

[2] FAN D P, ZHANG J, XU G, *et al*. Salient Objects in Clutter. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(2): 2344-2366.

[3] LI S Y, LIU H B, QIAN R, *et al*. TA^{2}N: Two-Stage Action Alignment Network for Few-Shot Action Recognition. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(2): 1404-1411.

[4] GUO W Y, ZHANG Y, YANG J F, *et al*. Re-Attention for Visual Question Answering. IEEE Transactions on Image Processing, 2021, 30: 6730-6743.

[5] WANG T, XU N, CHEN K A, *et al*. End-to-End Video Instance Segmentation via Spatial-Temporal Graph Neural Networks // Proc of the IEEE/CVF International Conference on Computer Vision. Washington, USA: IEEE, 2021: 10777-10786.

[6] LIU Y, GU Y C, ZHANG X Y, *et al*. Lightweight Salient Object Detection via Hierarchical Visual Perception Learning. IEEE Transactions on Cybernetics, 2021, 51(9): 4439-4449.

[7] FENG T L, LIU J X, YANG J F. Probing Sentiment-Oriented PreTraining Inspired by Human Sentiment Perception Mechanism // Proc of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2023: 2850-2860.

[8] LECUN Y, BENGIO Y, HINTON G. Deep Learning. Nature, 2015, 521(7553): 436-444.

[9] HUANG H Z, WANG Y, HU Q H, *et al*. Class Specific Semantic Reconstruction for Open Set Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(4): 4214-4228.

[10] GAN R T, FAN J S, WANG Y X, *et al*. Interact with Open Scenes: A Life-Long Evolution Framework for Interactive Segmentation Models // Proc of the 30th ACM International Conference on Multimedia. New York, USA: ACM, 2022: 5688-5697.

[11] ZHU F, CHENG Z, ZHANG X Y, *et al*. OpenMix: Exploring Outlier Samples for Misclassification Detection // Proc of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2023: 12074-12083.

[12] LI J C, XIE C Y, WU X Y, *et al*. What Makes Good Open-Vocabulary Detector: A Disassembling Perspective[C/OL]. [2023-

09-20]. https://arxiv.org/pdf/2309.00227.pdf.

[13] MAO B J, ZHANG X B, WANG L F, *et al*. Learning from the Target: Dual Prototype Network for Few Shot Semantic Segmentation. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(2): 1953-1961.

[14] LI X C, XIA X B, ZHU F, *et al*. Dynamics-Aware Loss for Learning with Label Noise. Pattern Recognition, 2023, 144. DOI: 10.1016/j.patcog.2023.109835.

[15] CHENG Z, ZHU F, ZHANG X Y, *et al*. Adversarial Training with Distribution Normalization and Margin Balance. Pattern Recognition, 2023, 136. DOI: 10.1016/j.patcog.2022.109182.

[16] ZHANG Y, TIŇO P, LEONARDIS A, *et al*. A Survey on Neural Network Interpretability. IEEE Transactions on Emerging Topics in Computational Intelligence, 2021, 5(5): 726-742.

[17] WU Y H, GAO S H, MEI J, *et al*. JCS: An Explainable COVID-19 Diagnosis System by Joint Classification and Segmentation. IEEE Transactions on Image Processing, 2021, 30: 3113-3126.

[18] TANG K H, ZHANG H W, WU B Y, *et al*. Learning to Compose Dynamic Tree Structures for Visual Contexts // Proc of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2019: 6612-6621.

[19] FAN J S, ZHANG Z X. Memory-Based Cross-Image Contexts for Weakly Supervised Semantic Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(5): 6006-6020.

[20] HUANG Y, WANG Y M, ZENG Y N, *et al*. MACK: Multimodal Aligned Conceptual Knowledge for Unpaired Image-Text Matching [C/OL]. [2023-09-20]. https://papers.nips.cc/paper_files/paper/2022/file/3379ce104189b72d5f7baaa03ae81329-Paper-Conference.pdf.

[21] TIAN K, ZHANG C H, WANG Y, *et al*. Knowledge Mining and Transferring for Domain Adaptive Object Detection // Proc of the IEEE/CVF International Conference on Computer Vision. Washington, USA: IEEE, 2021: 9113-9122.

[22] YU H Y, LI T, YU W C, *et al*. Regularized Graph Structure Learning with Semantic Knowledge for Multi-variates Time-Series Forecasting // Proc of the 31st International Joint Conference on Artificial Intelligence. San Francisco, USA: IJCAI, 2022: 2362-2368.

[23] FUKUSHIMA K. Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position-Neocognitron. IEICE Technical Report A, 1979, 62(10): 658-665.

[24] LOWE D G. Object Recognition from Local Scale-Invariant Features // Proc of the 7th IEEE International Conference on Computer Vision. Washington, USA: IEEE, 1999. DOI: 10.1109/ICCV.1999.790410

[25] LOWE D G. Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision, 2004, 60(2): 91-110.

[26] DETONE D, MALISIEWICZ T, RABINOVICH A. SuperPoint: Self-Supervised Interest Point Detection and Description // Proc of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Washington, USA: IEEE, 2018: 337-349.

[27] LIU Y, SHEN Z H, LIN Z X, *et al*. GIFT: Learning Transformation-Invariant Dense Visual Descriptors via Group CNNs[C/OL]. [2023-09-20]. https://arxiv.org/pdf/1911.05932.pdf.

[28] LIN W Y, HE X Y, DAI W R, *et al*. Key-Point Sequence Lossless Compression for Intelligent Video Analysis. IEEE MultiMedia, 2020, 27(3): 12-22.

[29] DUDA R O, HART P E. Use of Hough Transformation to Detect Lines and Curves in Pictures. Communications of the ACM, 1972, 15(1): 11-15.

[30] ZHANG Z H, LI Z X, BI N, *et al*. PPGNet: Learning Point-Pair Graph for Line Segment Detection // Proc of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2019: 7098-7107.

[31] XUE N, BAI S, WANG F D, *et al*. Learning Attraction Field Map for Robust Line Segment Detection // Proc of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2019: 1595-1603.

[32] LEE J T, KIM H U, LEE C, *et al*. Semantic Line Detection and its Applications // Proc of the IEEE International Conference on Computer Vision. Washington, USA: IEEE, 2017: 3249-3257.

[33] HAN Q, ZHAO K, XU J, *et al*. Deep Hough Transform for Semantic Line Detection // Proc of the European Conference on Computer Vision. Berlin, Germany: Springer, 2020: 249-265.

[34] ZHAO K, HAN Q, ZHANG C B, *et al*. Deep Hough Transform for Semantic Line Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(9): 4793-4806.

[35] KANOPOULOS N, VASANTHAVADA N, BAKER R L. Design of an Image Edge Detection Filter Using the Sobel Operator. IEEE Journal of Solid-State Circuits, 1988, 23(2): 358-367.

[36] SORIA X, RIBA E, SAPPA A. Dense Extreme Inception Network: Towards a Robust CNN Model for Edge Detection // Proc of the IEEE Winter Conference on Applications of Computer Vision. Washington, USA: IEEE, 2020: 1912-1921.

[37] LIU C, YANG J M, CEYLAN D, *et al*. PlaneNet: Piece-Wise Planar Reconstruction From a Single RGB Image // Proc of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2018: 2579-2588.

[38] LIU C, KIM K, GU J W, *et al*. PlaneRCNN: 3D Plane Detection and Reconstruction from a Single Image // Proc of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2019: 4445-4454.

[39] WANG T, LING H B. Gracker: A Graph-Based Planar Object Tracker. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(6): 1494-1501.

[40] ZHANG Z C, LIU S Z, YANG J F. Multiple Planar Object Trac-

king // Proc of the IEEE/CVF International Conference on Computer Vision. Washington, USA: IEEE, 2023: 23460-23470.

[41] ZHANG Z C, CHEN S, WANG Z C, *et al*. PlaneSeg: Building a Plug-In for Boosting Planar Region Segmentation. IEEE Transactions on Neural Networks and Learning Systems, 2023. DOI: 10.1109/TNNLS.2023.3262544

[42] LIU J J, LIU Z A, PENG P, *et al*. Rethinking the U-Shape Structure for Salient Object Detection. IEEE Transactions on Image Processing, 2021, 30: 9030-9042.

[43] WU Y H, LIU Y, ZHANG L, *et al*. EDN: Salient Object Detection via Extremely-Downsampled Network. IEEE Transactions on Image Processing, 2022, 31: 3125-3136.

[44] WU Y H, LIU Y, ZHANG L, *et al*. Regularized Densely-Connected Pyramid Network for Salient Instance Segmentation. IEEE Transactions on Image Processing, 2021, 30: 3897-3907.

[45] LIU J J, HOU Q B, LIU Z A, *et al*. PoolNet+: Exploring the Potential of Pooling for Salient Object Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(1): 887-904.

[46] FAN D P, CHENG M M, LIU Y, *et al*. Structure-Measure: A New Way to Evaluate Foreground Maps// Proc of the IEEE International Conference on Computer Vision. Washington, USA: IEEE, 2017: 4558-4567.

[47] ZHOU T, FAN D P, CHENG M M, *et al*. RGB-D Salient Object Detection: A Survey. Computational Visual Media, 2021, 7: 37-69.

[48] FAN D P, ZHAI Y J, BORJI A, *et al*. BBS-Net: RGB-D Salient Object Detection with a Bifurcated Backbone Strategy Network // Proc of the 16th European Conference on Computer Vision. Berlin, Germany: Springer, 2020: 275-292.

[49] ZHAI Y J, FAN D P, YANG J F, *et al*. Bifurcated Backbone Strategy for RGB-D Salient Object Detection. IEEE Transactions on Image Processing, 2021, 30: 8727-8742.

[50] WU Y H, LIU Y, XU J, *et al*. MobileSal: Extremely Efficient RGB-D Salient Object Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(12): 10261-10269.

[51] GAO S H, TAN Y Q, CHENG M M, *et al*. Highly Efficient Salient Object Detection with 100K Parameters // Proc of the European Conference on Computer Vision. Berlin, Germany: Springer, 2020: 702-721.

[52] CHENG M M, GAO S H, BORJI A, *et al*. A Highly Efficient Model to Study the Semantics of Salient Object Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(11): 8006-8021.

[53] MINSKY M. The Society of Mind. New York, USA: Simon & Schuster, 1988.

[54] LIU S Z, ZHANG X, YANG J F. SER30K: A Large-Scale Dataset for Sticker Emotion Recognition // Proc of the 30th ACM International Conference on Multimedia. New York, USA: ACM, 2022: 33-41.

[55] ZHAO S J, GE Y X, QI Z A, *et al*. Sticker820K: Empowering Interactive Retrieval with Stickers[C/OL]. [2023-09-20]. https://arxiv.org/pdf/2306.06870.pdf.

[56] WANG L J, GUO W Y, YAO X X, *et al*. Multimodal Event-Aware Network for Sentiment Analysis in Tourism. IEEE MultiMedia, 2021, 28(2): 49-58.

[57] WEN C S, JIA G L, YANG J F. DIP: Dual Incongruity Perceiving Network for Sarcasm Detection // Proc of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2023: 2540-2550.

[58] ESTRADA M L B, CABADA R Z, BUSTILLOS R O, *et al*. Opinion Mining and Emotion Recognition Applied to Learning Environments. Expert Systems with Applications, 2020, 150. DOI: 10.1016/j.eswa.2020.113265.

[59] BORTH D, CHEN T, JI R T, *et al*. SentiBank: Large-Scale Ontology and Classifiers for Detecting Sentiment and Emotions in Visual Content // Proc of the 21st ACM International Conference on Multimedia. New York, USA: ACM, 2013: 459-460.

[60] SUN M, YANG J F, WANG K, *et al*. Discovering Affective Regions in Deep Convolutional Neural Networks for Visual Sentiment Prediction // Proc of the IEEE International Conference on Multimedia and Expo. Washington, USA: IEEE, 2016. DOI: 10.1109/ICME.2016.7552961.

[61] SHE D Y, YANG J F, CHENG M M, *et al*. WSCNet: Weakly Supervised Coupled Networks for Visual Sentiment Classification and Detection. IEEE Transactions on Multimedia, 2020, 22(5): 1358-1371.

[62] YANG Y, JIA J, ZHANG S M, *et al*. How Do Your Friends on Social Media Disclose Your Emotions? Proceedings of the AAAI Conference on Artificial Intelligence, 2014, 28(1): 306-312.

[63] YANG J Y, LI J, LI L D, *et al*. Seeking Subjectivity in Visual Emotion Distribution Learning. IEEE Transactions on Image Processing, 2022, 31: 5189-5202.

[64] WANG L J, ZHANG X, JIANG N, *et al*. D^2S: Dynamic Distribution Supervision for Multi-label Facial Expression Recognition // Proc of the IEEE International Conference on Multimedia and Expo. Washington, USA: IEEE, 2022. DOI: 10.1109/ICME52920.2022.9859687.

[65] YANG J F, SUN M, SUN X X. Learning Visual Sentiment Distributions via Augmented Conditional Probability Neural Network. Proceedings of the AAAI Conference on Artificial Intelligence, 2017, 31(1): 224-230.

[66] YANG J F, SHE D Y, SUN M. Joint Image Emotion Classification and Distribution Learning via Deep Convolutional Neural Network // Proc of the 26th International Joint Conference on Artificial Intelligence. San Francisco, USA: IJCAI, 2017: 3266-3272.

[67] YANG J F, SHE D Y, LAI Y K, *et al*. Retrieving and Classifying Affective Images via Deep Metric Learning. Proceedings of the

AAAI Conference on Artificial Intelligence, 2018, 32(1): 491-498.

[68] YAO X X, SHE D Y, ZHAO S C, *et al*. Attention-Aware Polarity Sensitive Embedding for Affective Image Retrieval // Proc of the IEEE/CVF International Conference on Computer Vision. Washington, USA: IEEE, 2019: 1140-1150.

[69] JIA G L, YANG J F. S^2-VER: Semi-Supervised Visual Emotion Recognition // Proc of the 17th European Conference on Computer Vision. Berlin, Germany: Springer, 2022: 493-509.

[70] YOU Q Z, LUO J B, JIN H L, *et al*. Robust Image Sentiment Analysis Using Progressively Trained and Domain Transferred Deep Networks. Proceedings of the AAAI conference on Artificial Intelligence, 2015, 29(1): 381-388.

[71] PAN J C, WANG S F, FANG L. Representation Learning through Multimodal Attention and Time-Sync Comments for Affective Video Content Analysis // Proc of the 30th ACM International Conference on Multimedia. New York, USA: ACM, 2022: 42-50.

[72] ZHAO S C, JIA G L, YANG J F, *et al*. Emotion Recognition from Multiple Modalities: Fundamentals and Methodologies. IEEE Signal Processing Magazine, 2021, 38(6): 59-73.

[73] ZHAO S C, MA Y S, GU Y, *et al*. An End-to-End Visual-Audio Attention Network for Emotion Recognition in User-Generated Videos. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(1): 303-311.

[74] ZHANG Z C, YANG J F. Temporal Sentiment Localization: Listen and Look in Untrimmed Videos // Proc of the 30th ACM International Conference on Multimedia. New York, USA: ACM, 2022: 199-208.

[75] ZHANG Z C, WANG L J, YANG J F. Weakly Supervised Video Emotion Detection and Prediction via Cross-Modal Temporal Erasing Network // Proc of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2023: 18888-18897.

[76] LI P, YANG Y, ZHAO W D, *et al*. Evaluation of Image Fire Detection Algorithms Based on Image Complexity. Fire Safety Journal, 2021, 121. DOI: 10.1016/j.firesaf.2021.103306.

[77] DAI L C, ZHANG K, ZHENG X S, *et al*. Visual Complexity of Shapes: A Hierarchical Perceptual Learning Model. The Visual Computer, 2022, 38: 419-432.

[78] OLIVIA A, MACK M L, SHRESTHA M, *et al*. Identifying the Perceptual Dimensions of Visual Complexity of Scenes // Proc of the Annual Meeting of the Cognitive Science Society. New York, USA: ACM, 2004: 1041-1046.

[79] CHEN Y Q, DUAN J, ZHU Y, *et al*. Research on the Image Complexity Based on Neural Network // Proc of the International Conference on Machine Learning and Cybernetics. Washington, USA: IEEE, 2015: 295-300.

[80] SARAEE E, JALAL M, BETKE M. Visual Complexity Analysis Using Deep Intermediate-Layer Features. Computer Vision and Image Understanding, 2020, 195. DOI: 10.1016/j.cviu.2020.102 949.

[81] FENG T L, ZHAI Y J, YANG J F, *et al*. IC9600: A Benchmark Dataset for Automatic Image Complexity Assessment. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(7): 8577-8593.

[82] KHOSLA A, RAJU A S, TORRALBA A, *et al*. Understanding and Predicting Image Memorability at a Large Scale // Proc of the IEEE International Conference on Computer Vision. Washington, USA: IEEE, 2015: 2390-2398.

[83] SHOKRI R, STRONATI M, SONG C Z, *et al*. Membership Inference Attacks against Machine Learning Models // Proc of the IEEE Symposium on Security and Privacy. Washington, USA: IEEE, 2017: 3-18.

[84] ARPIT D, JASTRZĘBSKI S, BALLAS N, *et al*. A Closer Look at Memorization in Deep Networks // Proc of the 34th International Conference on Machine Learning. San Diego, USA: JMLR, 2017: 233-242.

[85] WANG Z, BOVIK A C, SHEIKH H R, *et al*. Image Quality Assessment: From Error Visibility to Structural Similarity. IEEE Transactions on Image Processing, 2004, 13(4): 600-612.

[86] GIROD B. What's Wrong with Mean Squared Error // WASTON A B, ed. Digital Images and Human Vision. Cambridge, USA: MIT Press, 1993: 207-220.

[87] ZHANG R, ISOLA P, EFROS A A, *et al*. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric // Proc of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2018: 586-595.

[88] DING K Y, MA K D, WANG S Q, *et al*. Image Quality Assessment: Unifying Structure and Texture Similarity. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(5): 2567-2581.

[89] ROY S, MITRA S, BISWAS S, *et al*. Test Time Adaptation for Blind Image Quality Assessment // Proc of the IEEE/CVF International Conference on Computer Vision. Washington, USA: IEEE, 2023: 16742-16751.

[90] LIU X L, VAN DE WEIJER J, BAGDANOV A D. RankIQA: Learning from Rankings for No-Reference Image Quality Assessment // Proc of the IEEE International Conference on Computer Vision. Washington, USA: IEEE, 2017: 1040-1049.

[91] SU S L, YAN Q S, ZHU Y, *et al*. Blindly Assess Image Quality in the Wild Guided by a Self-Adaptive Hyper Network // Proc of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2020: 3664-3673.

[92] ZHANG W X, LI D Q, MA C, *et al*. Continual Learning for Blind Image Quality Assessment. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(3): 2864-2878.

[93] KANG L, YE P, LI Y, *et al*. Convolutional Neural Networks for No-Reference Image Quality Assessment // Proc of the IEEE Con-

ference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2014: 1733-1740.

[94] PAN D, SHI P, HOU M, *et al*. Blind Predicting Similar Quality Map for Image Quality Assessment // Proc of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2018: 6373-6382.

[95] MURRAY N, MARCHESOTTI L, PERRONNIN F. AVA: A Large-Scale Database for Aesthetic Visual Analysis // Proc of the IEEE Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2012: 2408-2415.

[96] ZHANG X D, GAO X B, LU W, *et al*. A Gated Peripheral-Foveal Convolutional Neural Network for Unified Image Aesthetic Prediction. IEEE Transactions on Multimedia, 2019, 21(11): 2815-2826.

[97] ZHUANG B H, LIU L Q, LI Y, *et al*. Attend in Groups: A Weakly-Supervised Deep Learning Framework for Learning from Web Data // Proc of the IEEE Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2017: 2915-2924.

[98] NAYAK G, GHOSH R, JIA X W, *et al*. Weakly Supervised Classification Using Group-Level Labels[C/OL]. [2023-09-20]. https://arxiv.org/pdf/2108.07330v1.pdf.

[99] XU Y H, QIAN Q, LI H, *et al*. Weakly Supervised Representation Learning with Coarse Labels // Proc of the IEEE/CVF International Conference on Computer Vision. Washington, USA: IEEE, 2021: 10573-10581.

[100] JIANG B, WANG L L, CHENG J, *et al*. GPENs: Graph Data Learning with Graph Propagation-Embedding Networks. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(8): 3925-3938.

[101] ZHENG Z H, YE R G, WANG P, *et al*. Localization Distillation for Dense Object Detection // Proc of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2022: 9397-9406.

[102] ZHANG S Q, LI C L, JIA Z, *et al*. Diag-IoU Loss for Object Detection. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(12): 7671-7683.

[103] CHEN Z M, CHEN K, LIN W Y, *et al*. PIoU Loss: Towards Accurate Oriented Object Detection in Complex Environments // Proc of the European Conference on Computer Vision. Berlin, Germany: Springer, 2020: 195-211.

[104] ZHANG D W, HAN J W, CHENG G, *et al*. Weakly Supervised Object Localization and Detection: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(9): 5866-5885.

[105] SHAO F F, CHEN L, SHAO J, *et al*. Deep Learning for Weakly-Supervised Object Detection and Object Localization: A Survey. Neurocomputing, 2022, 496: 192-207.

[106] ZHANG Y M, CHEN T. Weakly Supervised Object Recognition and Localization with Invariant High Order Features[C/OL]. [2023-09-20]. https://bmvc10.dcs.aber.ac.uk/proc/conference/paper47/paper47.pdf.

[107] TANG Y X, WANG X F, DELLANDREA E, *et al*. Fusing Generic Objectness and Deformable Part-Based Models for Weakly Supervised Object Detection // Proc of the IEEE International Conference on Image Processing. Washington, USA: IEEE, 2014: 4072-4076.

[108] SIVA P, RUSSELL C, XIANG T, *et al*. Looking Beyond the Image: Unsupervised Learning for Object Saliency and Detection // Proc of the IEEE Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2013: 3238-3245.

[109] SHI Z Y, HOSPEDALES T M, XIANG T. Bayesian Joint Modelling for Object Localisation in Weakly Labelled Images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(10): 1959-1972.

[110] CINBIS R G, VERBEEK J, SCHMID C. Multi-fold MIL Training for Weakly Supervised Object Localization // Proc of the IEEE Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2014: 2409-2416.

[111] DESELAERS T, ALEXE B, FERRARI V. Localizing Objects While Learning Their Appearance // Proc of the European Conference on Computer Vision. Berlin, Germany: Springer, 2010: 452-466.

[112] SINGH K K, XIAO F Y, LEE Y J. Track and Transfer: Watching Videos to Simulate Strong Human Supervision for Weakly-Supervised Object Detection // Proc of the IEEE Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2016: 3548-3556.

[113] LI D, HUANG J B, LI Y L, *et al*. Weakly Supervised Object Localization with Progressive Domain Adaptation // Proc of the IEEE Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2016: 3512-3520.

[114] SHI M J, CAESAR H, FERRARI V. Weakly Supervised Object Localization Using Things and Stuff Transfer // Proc of the IEEE International Conference on Computer Vision. Washington, USA: IEEE, 2017: 3401-3410.

[115] ZHANG D W, HAN J W, ZHAO L, *et al*. Leveraging Prior-Knowledge for Weakly Supervised Object Detection under a Collaborative Self-Paced Curriculum Learning Framework. International Journal of Computer Vision, 2019, 127: 363-380.

[116] SANG H B, NI Z L, HE H Y, *et al*. Trace-Level Invisible Enhanced Network for 6D Pose Estimation // Proc of the IEEE International Conference on Multimedia and Expo. Washington, USA: IEEE, 2022. DOI: 10.1109/ICME52920.2022.9859613.

[117] JIANG P T, HAN L H, HOU Q B, *et al*. Online Attention Accumulation for Weakly Supervised Semantic Segmentation. IEEE

Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(10): 7062-7077.

[118] LIU Y, WU Y H, WEN P S, *et al*. Leveraging Instance-, Image- and Dataset-Level Information for Weakly Supervised Instance Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(3): 1415-1428.

[119] LIN Z, DUAN Z P, ZHANG Z, *et al*. KnifeCut: Refining Thin Part Segmentation with Cutting Lines // Proc of the 30th ACM International Conference on Multimedia. New York, USA: ACM, 2022: 809-817.

[120] 侯淇彬,韩凌昊,刘姜江,等. 互联网图像驱动的语义分割自主学习. 中国科学(信息科学), 2021, 51(7): 1084-1099.
(HOU Q B, HAN L H, LIU J J, *et al*. Autonomous Learning of Semantic Segmentation from Internet Images. Scientia Sinica Informationis, 2021, 51(7): 1084-1099.)

[121] MEI J, CHENG M M, XU G, *et al*. SANet: A Slice-Aware Network for Pulmonary Nodule Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(8): 4374-4387.

[122] CHEN J, LI Z H, LUO J B, *et al*. Learning a Weakly-Supervised Video Actor-Action Segmentation Model with a Wise Selection // Proc of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2020: 9898-9908.

[123] LIU Q, RAMANATHAN V, MAHAJAN D, *et al*. Weakly Supervised Instance Segmentation for Videos with Temporal Mask Consistency // Proc of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2021: 13963-13973.

[124] LI Y X, XU N, YANG W J, *et al*. Exploring the Semi-Supervised Video Object Segmentation Problem from a Cyclic Perspective. International Journal of Computer Vision, 2022, 130(10): 2408-2424.

[125] WANG X L, JABRI A, EFROS A A. Learning Correspondence from the Cycle-Consistency of Time // Proc of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2019: 2566-2576.

[126] LI X T, LIU S F, DE MELLO S, *et al*. Joint-Task Self-Supervised Learning for Temporal Correspondence // Proc of the 33rd International Conference on Neural Information Processing Systems. Cambridge, USA: MIT Press, 2019: 318-328.

[127] YAN L Q, WANG Q F, MA S Q, *et al*. Solve the Puzzle of Instance Segmentation in Videos: A Weakly Supervised Framework with Spatio-Temporal Collaboration. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(1): 393-406.

[128] LIN F C, XIE H T, LIU C B, *et al*. Bilateral Temporal Re-Aggregation for Weakly-Supervised Video Object Segmentation. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(7): 4498-4512.

[129] LIU P D, HE Z B, YAN X Y, *et al*. WeClick: Weakly-Supervised Video Semantic Segmentation with Click Annotations // Proc of the 29th ACM International Conference on Multimedia. New York, USA: ACM, 2021: 2995-3004.

[130] ZHANG Z, JIN W D, XU J, *et al*. Gradient-Induced Co-Saliency Detection // Proc of the European Conference on Computer Vision. Berlin, Germany: Springer, 2020: 455-472.

[131] LI Y X, LIN W Y, WANG T, *et al*. Video Summarization via Cluster-Based Object Tracking and Type-Based Synopsis // Proc of the IEEE Conference on Multimedia Information Processing and Retrieval. Washington, USA: IEEE, 2020: 113-116.

[132] LOCATELLO F, WEISSENBORN D, UNTERTHINER T, *et al*. Object-Centric Learning with Slot Attention[C/OL]. [2023-09-20]. https://arxiv.org/pdf/2006.15055.pdf.

[133] VO V H, SIZIKOVA E, SCHMID C, *et al*. Large-Scale Unsupervised Object Discovery[C/OL]. [2023-09-20]. https://arxiv.org/abs/2106.06650.

[134] GAO S H, LI Z Y, YANG M H, *et al*. Large-Scale Unsupervised Semantic Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(6): 7457-7476.

[135] YAO X X, ZHAO S C, XU P F, *et al*. Multi-source Domain Adaptation for Object Detection // Proc of the IEEE/CVF International Conference on Computer Vision. Washington, USA: IEEE, 2021: 3253- 3262.

[136] QIAN R, LI Y X, LIU H B, *et al*. Enhancing Self-Supervised Video Representation Learning via Multi-level Feature Optimization // Proc of the IEEE/CVF International Conference on Computer Vision. Washington, USA: IEEE, 2021: 7970-7981.

[137] CHEN S, XUE J H, CHANG J L, *et al*. SSL++: Improving Self-supervised Learning by Mitigating the Proxy Task-Specificity Problem. IEEE Transactions on Image Processing, 2021, 31: 1134-1148.

[138] YAO X X, ZHAO S C, LAI Y K, *et al*. APSE: Attention-Aware Polarity-Sensitive Embedding for Emotion-Based Image Retrieval. IEEE Transactions on Multimedia, 2020, 23: 4469-4482.

[139] VAN DEN OORD A, KALCHBRENNER N, ESPEHOLT L, *et al*. Conditional Image Generation with PixelCNN Decoders // Proc of the 32nd International Conference on Neural Information Processing Systems. Cambridge, USA: MIT Press, 2016: 4797-4805.

[140] VAN DEN OORD A, KALCHBRENNER N, KAVUKCUOGLU K. Pixel Recurrent Neural Networks // Proc of the 33rd International Conference on Machine Learning. San Diego, USA: JMLR, 2016: 1747-1756.

[141] KINGMA D P, DHARIWAL P. Glow: Generative Flow with Invertible 1×1 Convolutions // Proc of the 32nd International Conference on Neural Information Processing Systems. Cambridge,

USA: MIT Press, 2016: 10236-10245.

[142] DINH L, SOHL-DICKSTEIN J, BENGIO S. Density Estimation Using Real NVP[C/OL]. [2023-09-20]. https://arxiv.org/pdf/1605.08803.pdf.

[143] VAN DEN OORD A, VINYALS O, KAVUKCUOGLU K, *et al*. Neural Discrete Representation Learning // Proc of the 31st International Conference on Neural Information Processing Systems. Cambridge, USA: MIT Press, 2017: 6309-6318.

[144] HE K M, CHEN X L, XIE S N, *et al*. Masked Autoencoders Are Scalable Vision Learners // Proc of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2022: 15979-15988.

[145] DEVLIN J, CHANG M W, LEE K, *et al*. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proc of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies(Long and Short Papers). Stroudsburg, USA: ACL, 2019: 4171-4186.

[146] WEI C, FAN H Q, XIE S N, *et al*. Masked Feature Prediction for Self-Supervised Visual Pre-Training // Proc of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2022: 14648-14658.

[147] LI S Y, WU D, WU F, *et al*. Architecture-Agnostic Masked Image Modeling-From ViT Back to CNN // Proc of the 40th International Conference on Machine Learning. San Diego, USA: JMLR, 2023: 20149-20167.

[148] HOU Z J, SUN F, CHEN Y K, *et al*. MILAN: Masked Image Pretraining on Language Assisted Representation[C/OL]. [2023-09-20]. https://arxiv.org/pdf/2208.06049.pdf.

[149] ZENG D L, LIAO M Y, TAVAKOLIAN M, *et al*. Deep Learning for Scene Classification: A Survey[C/OL]. [2023-09-20]. https://arxiv.org/abs/2101.10531.

[150] SIAGIAN C, ITTI L. Rapid Biologically-Inspired Scene Classification Using Features Shared with Visual Attention. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(2): 300-312.

[151] REN Z, QIAN K, ZHANG Z X, *et al*. Deep Scalogram Representations for Acoustic Scene Classification. IEEE/CAA Journal of Automatica Sinica, 2018, 5(3): 662-669.

[152] WANG L, SNG D. Deep Learning Algorithms with Applications to Video Analytics for A Smart City: A Survey[C/OL]. [2023-09-20]. https://arxiv.org/abs/1512.03131.

[153] LÓPEZ-CIFUENTES A, ESCUDERO-VINOLO M, BESCÓS J, *et al*. Semantic-Aware Scene Recognition. Pattern Recognition, 2020, 102. DOI: 10.1016/j.patcog.2020.107256

[154] TONG Z H, SHI D X, YAN B Z, *et al*. A Review of Indoor-Outdoor Scene Classification // Proc of the 2nd International Conference on Control, Automation and Artificial Intelligence. New York, USA: ACM, 2017: 469-474.

[155] CHENG G, HAN J W, LU X Q. Remote Sensing Image Scene Classification: Benchmark and State of the Art. Proceedings of the IEEE, 2017, 105(10): 1865-1883.

[156] XIA G S, HU J W, HU F, *et al*. AID: A Benchmark Dataset for Performance Evaluation of Aerial Scene Classification. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(7): 3965-3981.

[157] MESAROS A, HEITTOLA T, VIRTANEN T. TUT Database for Acoustic Scene Classification and Sound Event Detection // Proc of the 24th European Signal Processing Conference. Washington, USA: IEEE, 2016: 1128-1132.

[158] LOWRY S, SÜNDERHAUF N, NEWMAN P, *et al*. Visual Place Recognition: A Survey. IEEE Transactions on Robotics, 2016, 32(1): 1-19.

[159] ARANDJELOVIC R, GRONAT P, TORII A, *et al*. NetVLAD: CNN Architecture for Weakly Supervised Place Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(6): 1437-1451.

[160] BROWN M, SÜSSTRUNK S. Multi-spectral SIFT for Scene Category Recognition // Proc of the IEEE Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2011: 177-184.

[161] VISWANATHAN D G. Features from Accelerated Segment Test (FAST)[C/OL]. [2023-09-20]. https://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/AV1011/AV1FeaturefromAcceleratedSegmentTest.pdf.

[162] BAY H, ESS A, TUYTELAARS T, *et al*. Speeded-Up Robust Features(SURF). Computer Vision and Image Understanding, 2008, 110(3): 346-359.

[163] JEEVAN P P, VISWANATHAN K, ANANDU A S, *et al*. WaveMix: A Resource-Efficient Neural Network for Image Analysis[C/OL]. [2023-09-20]. https://arxiv.org/abs/2205.14375.

[164] WANG Q L, XIE J T, ZUO W M, *et al*. Deep CNNs Meet Global Covariance Pooling: Better Representation and Generalization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(8): 2582-2597.

[165] 张羽丰,李昱希,赵明璧,等.局部双目视差回归的目标距离估计.中国图象图形学报, 2021, 26(7): 1604-1613.
(ZHANG Y F, LI Y X, ZHAO M B, *et al*. Object Distance Estimation Based on Stereo Regional Disparity Regression. Journal of Image and Graphics, 2021, 26(7): 1604-1613.)

[166] ZHANG Y F, LI Y X, ZHAO M B, *et al*. A Regional Regression Network for Monocular Object Distance Estimation // Proc of the IEEE International Conference on Multimedia and Expo Workshops. Washington, USA: IEEE, 2020. DOI: 10.1109/ICMEW46912.2020.9106012.

[167] LIU H B, LI J G, LI D, *et al*. Learning Scale-Consistent Atten-

tion Part Network for Fine-Grained Image Recognition. IEEE Transactions on Multimedia, 2021, 24: 2902-2913.

[168] FAN J H, LIU H B, YANG W J, *et al*. Speed Up Object Detection on Gigapixel-Level Images With Patch Arrangement // Proc of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2022: 4643-4651.

[169] GUO M H, LU C Z, HOU Q B, *et al*. SegNeXt: Rethinking Convolutional Attention Design for Semantic Segmentation[C/OL]. [2023-09-20]. https://arxiv.org/pdf/2209.08575v1.pdf.

[170] MEI J, LI R J, GAO W, *et al*. CoANet: Connectivity Attention Network for Road Extraction From Satellite Imagery. IEEE Transactions on Image Processing, 2021, 30: 8540-8552.

[171] PRANGEMEIER T, REICH C, KOEPPL H. Attention-Based Transformers for Instance Segmentation of Cells in Microstructures // Proc of the IEEE International Conference on Bioinformatics and Biomedicine. Washington, USA: IEEE, 2020: 700-707.

[172] LI S Y, LIU H B, FEI M J, *et al*. Temporal Alignment via Event Boundary for Few-shot Action Recognition[C/OL]. [2023-09-20]. https://www.bmvc2021-virtualconference.com/assets/papers/0878.pdf.

[173] LIU H B, LÜ W U X, SEE J, *et al*. Task-adaptive Spatial-Temporal Video Sampler for Few-Shot Action Recognition // Proc of the 30th ACM International Conference on Multimedia. New York, USA: ACM, 2022: 6230-6240.

[174] LI Y X, LIN W Y, SEE J, *et al*. CFAD: Coarse-to-Fine Action Detector for Spatiotemporal Action Localization // Proc of the European Conference on Computer Vision. Berlin, Germany: Springer, 2020: 510-527.

[175] LI Y X, ZHANG B S, LI J, *et al*. LSTC: Boosting Atomic Action Detection with Long-Short-Term Context // Proc of the 29th ACM International Conference on Multimedia. New York, USA: ACM, 2021: 2158-2166.

[176] QIAN R, HU D, DINKEL H, *et al*. Multiple Sound Sources Localization from Coarse to Fine // Proc of the European Conference on Computer Vision. Berlin, Germany: Springer, 2020: 292-308.

[177] LI C L, LIU L, LU A D, *et al*. Challenge-Aware RGBT Tracking // Proc of the European Conference on Computer Vision. Berlin, Germany: Springer, 2020: 222-237.

[178] CHANG X J, REN P Z, XU P F, *et al*. A Comprehensive Survey of Scene Graphs: Generation and Application. IEEE Transactions on Neural Networks and Learning Systems, 2023, 45(1): 1-26.

[179] ZAREIAN A, KARAMAN S, CHANG S F. Bridging Knowledge Graphs to Generate Scene Graphs // Proc of the European Conference on Computer Vision. Berlin, Germany: Springer, 2020: 606-623.

[180] LI H S, ZHU G M, ZHANG L, *et al*. Scene Graph Generation: A Comprehensive Survey. Neurocomputing, 2023. DOI: 10.1016/j.neucom.2023.127052.

[181] LI Y K, OUYANG W L, ZHOU B L, *et al*. Factorizable Net: An Efficient Subgraph-Based Framework for Scene Graph Generation // Proc of the European Conference on Computer Vision. Berlin, Germany: Springer, 2018: 346- 363.

[182] LAFFERTY J, MCCALLUM A, PEREIRA F C N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data // Proc of the 18th International Conference on Machine Learning. San Diego, USA: JMLR, 2001: 282-289.

[183] BORDES A, USUNIER N, GARCIA-DURAN A, *et al*. Translating Embeddings for Modeling Multi-relational Data // Proc of the 26th International Conference on Neural Information Processing Systems. Cambridge, USA: MIT Press, 2013: 2787-2795.

[184] DAI B, ZHANG Y Q, LIN D H. Detecting Visual Relationships with Deep Relational Networks // Proc of the IEEE Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2017: 3298-3308.

[185] CONG W L, WANG W, LEE W C. Scene Graph Generation via Conditional Random Fields[C/OL]. [2023-09-20]. https://arxiv.org/pdf/1811.08075.pdf.

[186] ZHANG H W, KYAW Z, CHANG S F, *et al*. Visual Translation Embedding Network for Visual Relation Detection // Proc of the IEEE Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2017: 3107-3115.

[187] HUNG Z S, MALLYA A, LAZEBNIK S. Contextual Translation Embedding for Visual Relationship Detection and Scene Graph Generation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 43(11): 3820-3832.

[188] XU D F, ZHU Y K, CHOY C B, *et al*. Scene Graph Generation by Iterative Message Passing // Proc of the IEEE Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2017: 3097-3106.

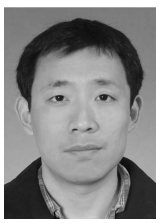
[189] PLUMMER B A, MALLYA A, CERVANTES C M, *et al*. Phrase Localization and Visual Relationship Detection with Comprehensive Image-Language Cues // Proc of the IEEE International Conference on Computer Vision. Washington, USA: IEEE, 2017: 1946-1955.

作者简介

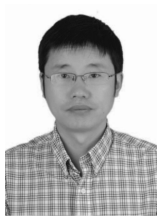
张知诚,博士研究生,主要研究方向为计算机视觉. E-mail:gloryzzc6@sina.com.
(**ZHANG Zhicheng**, Ph. D. candidate. His research interests include computer vision.)

杨巨峰(通信作者),博士,教授,主要研究方向为计算机视觉. E-mail:yangjufeng@nankai.edu.cn.
(**YANG Jufeng** (Corresponding author), Ph. D., professor. His research interests include computer vision.)

程明明,博士,教授,主要研究方向为计算机视觉. E-mail:cmm@nankai.edu.cn.
(**CHENG Mingming**, Ph. D., professor. His research interests include computer vision.)



林巍峣,博士,教授,主要研究方向为计算机视觉. E-mail:wylin@sjtu.edu.cn.
(**LIN Weiyao**, Ph. D., professor. His research interests include computer vision.)



汤 进,博士,教授,主要研究方向为计算机视觉. E-mail:tangjin@ahu.edu.cn.
(**TANG Jin**, Ph. D., professor. His research interests include computer vision.)

李成龙,博士,教授,主要研究方向为计算机视觉. E-mail:lcl1314@foxmail.com.
(**LI Chenglong**, Ph. D., professor. His research interests include computer vision.)

刘成林,博士,研究员,主要研究方向为模式识别、机器学习、文档分析与识别等. E-mail:liucl@nlpr.ia.ac.cn.
(**LIU Chenglin**, Ph. D., professor. His research interests include pattern recognition, machine learning, and document analysis and recognition.)