计算机视觉中的图像特征提取技术

黄梓淳

Jun 09, 2026

计算机视觉任务通常包含分类、检测、分割与检索等核心环节，这些任务都需要把原始像素逐步映射为可被算法理解的高层语义。特征提取正处于这一映射过程的中心位置，它决定了后续模块能否高效地利用视觉信息。传统手工设计特征依赖于先验知识与人工调参，深度学习则通过端到端训练自动学习层次化表示，而多模态大模型进一步把视觉特征与语言、音频对齐。理解这一演进脉络，既能帮助读者在精度、速度与数据约束之间做出权衡，也能为实际工程选型提供清晰思路。后续内容将依次梳理手工特征、深度特征、任务适配、可视化、评估、工程实践以及未来方向。

1　传统手工特征提取方法

在深度学习普及之前，研究者依靠梯度、颜色与纹理等低层统计量来描述图像。方向梯度直方图简称 HOG，通过在局部单元内统计梯度幅值与方向的分布来捕捉物体形状；其计算流程先对图像做 Gamma 校正，再求取水平与垂直方向梯度，然后把梯度方向量化为若干个 bin 并统计直方图，最后把相邻单元组合成块并做 L2 归一化以获得对光照的鲁棒性。尺度不变特征变换简称 SIFT，先在高斯尺度空间中检测 DoG 极值以定位关键点，再计算关键点周围 16×16 区域的梯度直方图并生成 128 维描述子，从而实现对尺度与旋转的鲁棒。SURF、ORB 与 BRIEF 则通过积分图加速或二进制测试来降低计算量，体现了工程化权衡。颜色直方图与颜色矩直接统计像素在各通道的分布，LBP 通过比较中心像素与邻域像素的大小关系生成二值编码，Gabor 滤波器组则用多尺度多方向的复正弦波来模拟人类视觉的纹理感知。全局描述子如 GIST 对整幅图像做 Gabor 响应统计，Spatial Pyramid Matching 把图像划分为多层网格并分别提取局部特征再加权融合。手工特征虽然在特定场景下仍有价值，但对光照、视角与语义变化的敏感性使其难以泛化到复杂任务。

2　深度学习驱动的特征提取

卷积神经网络通过堆叠卷积、池化与全连接层，逐步把像素级信息抽象为边缘、纹理、部件与物体级语义。经典骨干网络的演进体现了深度与效率的平衡：AlexNet 首次在 ImageNet 上证明深层卷积网络的威力，VGG 通过重复使用 3×3 卷积堆叠出更深的结构，ResNet 引入残差连接缓解梯度消失，EfficientNet 则用复合缩放系数同时调节深度、宽度与分辨率，ConvNeXt 把 Vision Transformer 的设计思想反哺卷积架构。特征金字塔网络 FPN 在不同分辨率特征图之间自顶向下融合，空洞卷积 ASPP 通过不同膨胀率捕捉多尺度上下文。Vision Transformer 把图像切分为固定大小的 patch 并用自注意力建模全局依赖，证明了纯注意力机制在视觉任务上的可行性。自监督学习进一步降低对标注的依赖，SimCLR 通过最大化同一图像不同增强视图之间的一致性来学习不变特征，MoCo 用队列维护负样本，DINO 把自注意力蒸馏到学生网络，MAE 则通过掩码自编码重建被遮挡区域。轻量化网络如 MobileNet 用深度可分离卷积降低计算量，ShuffleNet 通过通道混洗实现跨组信息流

动，GhostNet 则用廉价线性变换生成冗余特征图，配合知识蒸馏与结构剪枝可在移动端保持较高精度。

3 特征表达的 downstream 任务适配

不同视觉任务对特征的聚合方式与损失函数有不同要求。在图像分类中，全局平均池化把空间维度压缩为通道向量，再接线性分类头即可得到类别概率。目标检测则需要在特征图上生成候选框，R-CNN 系列先用选择性搜索生成区域，再用 CNN 提取特征并分类；YOLO 把检测转化为单阶段回归，直接在特征图上预测边界框与类别；DETR 用 Transformer 的编码-解码结构把检测视为集合预测问题，省去 NMS 后处理。语义分割常用全卷积网络 FCN 把分类网络改写为上采样结构，Mask R-CNN 在检测分支之外增加掩码分支，Segment Anything 模型 SAM 则通过提示编码器与掩码解码器实现零样本分割。图像检索与重识别常用度量学习损失，Triplet Loss 拉近正样本对并推远负样本对，ArcFace 在角度空间引入加性边界以增强类间可分性，Proxy-NCA 用代理向量近似类中心。向量索引库 Faiss 提供倒排文件与乘积量化，可在十亿级向量上实现毫秒级近似最近邻搜索。三维视觉中，PointNet++ 通过最远点采样与局部邻域聚合把点云映射为置换不变特征，多视图方法则把同一物体的不同视角特征融合后再做三维推理。

4 特征可视化与可解释性

为了理解网络学到了什么，研究者提出了多种可视化手段。激活最大化通过梯度上升在输入空间搜索能最大化特定神经元响应的图像，Grad-CAM 用目标类别对最后一个卷积层的梯度加权得到类激活热力图，Score-CAM 则用前向传播得分代替梯度来避免噪声。t-SNE 与 UMAP 把高维特征投影到二维或三维，便于观察类间可分性与聚类结构。特征图可视化可直接显示各通道的响应模式，注意力热力图则高亮模型在推理时关注的区域。这些工具不仅帮助调试模型，也常用于论文中直观展示特征表达的有效性。

5 性能评估与基准

公平比较需要统一的数据集、指标与硬件。ImageNet 提供一千类百万张图像，用于衡量分类能力；COCO 同时标注检测与分割，可计算平均精度 mAP；Places 数据集聚焦场景识别；Oxford5k 与 Paris6k 用于图像检索评测。常用指标包括 Top-1 与 Top-5 准确率、平均精度 mAP、Recall@K 以及每秒帧数 FPS，同时记录模型参数量与浮点运算量 FLOPs 以评估效率。实验时需固定预训练权重、输入分辨率与推理硬件，避免因配置差异导致误判。

6 工程实践与工具链

从原型到部署需要完整工具链支持。传统方法可直接调用 OpenCV 实现 SIFT、HOG 等算子，VLFeat 提供更底层的 C 实现。深度学习框架中，PyTorch 与 TensorFlow 提供自动微分与分布式训练，timm 仓库收集了大量预训练分类模型，mmpretrain 与 Detectron2 分别针对分类与检测提供统一接口。模型部署时，TensorRT 可对 NVIDIA GPU 做层融合与精度校准，ONNX Runtime 支持跨平台推理，OpenVINO 针对 Intel 硬件优化，Core ML 则面向 Apple 生态。量化与混合精度训练能在精度损失可控的前提下显著降低延迟。MLOps 层面，特征版本管理、持续训练与 A/B 测试确保模型在生产环境中持续迭代。

7 挑战与未来方向

尽管深度特征已取得显著进展，仍存在若干开放问题。数据稀缺场景下，领域自适应与少样本学习成为关键技术。多模态统一表示如 CLIP 通过对比图像-文本对学习跨模态对齐，LLaVA 把视觉编码器接入大语言模型实现视觉问答，ImageBind 进一步把图像、文本、音频与深度图映射到同一空间。鲁棒性研究聚焦对抗样本防御，旨在保持模型在微小扰动下的稳定性。神经架构搜索 NAS 尝试自动化发现最优网络结构，减少人工设计成本。边缘智能与存算一体芯片则把特征提取直接嵌入传感器，降低数据搬运开销。

8 结论

特征提取已从手工设计演进为端到端学习，再迈向多模态对齐的新范式。实践中需根据任务对精度、速度与数据的约束选择合适技术栈。展望未来，统一的多模态大模型有望进一步简化视觉 pipeline，但针对特定领域的高效特征仍将长期存在。

9 参考文献与延伸阅读

必读论文包括《ImageNet Classification with Deep Convolutional Neural Networks》《Deep Residual Learning for Image Recognition》《Attention Is All You Need》与《Learning Transferable Visual Models From Natural Language Supervision》。开源代码可访问 timm、Detectron2 与 Segment-Anything 仓库。进一步学习资源推荐斯坦福 CS231n 课程、清华计算机视觉公开课以及顶会论文速递平台。