浙江理工大學

Zhejiang Sci-Tech University

# 算法分析与设计


题目：图像分割典型算法：深度学习进展与前沿应用

| 姓名 | 学号 | 班级 |
| --- | --- | --- |
| 李凯涛 | 2023327100056 | 计算机科学与技术 23（4）班 |

# 图像分割典型算法：深度学习进展与前沿应用

**摘　要：**图像分割是计算机视觉领域的一项基础性任务，旨在根据像素的共享属性（如灰度、纹理或颜色）将图像划分为具有语义意义的区域。其核心目标在于简化图像表示，从而促进更高效、更深入的图像分析，并精确地定位图像中的目标对象或其边界。本文全面回顾了图像分割领域的发展历程，从经典的分割方法到前沿的深度学习模型。

**关键词:** 图像分割；分割算法；语义分割；实例分割；深度学习；像素分类；Otsu 算法；分水岭算法；FCN；U-Net；Mask R-CNN；DeepLab；SegNet；PSPNet

## 1. 绪论

### 1.1 研究背景与意义

图像分割是数字图像处理和计算机视觉领域中一个长期存在且极具挑战性的核心问题 [1]。其根本目的在于将数字图像细分为多个有意义的图像片段（即区域或对象），从而简化图像的复杂表示，使其更易于理解和分析 [3]。这一过程的本质是为图像中的每个像素分配一个标签，使得具有相同标签的像素共享某些视觉特性，例如颜色、强度或纹理 。

图像分割作为计算机视觉任务的基石，为更高层次的视觉理解提供了基础。它不仅用于定位图像中的对象和边界，更是场景理解、视觉常识推理以及感知社会可供性等高级目标实现的关键初始步骤 [2]。该领域的研究持续演进，从早期的 N-Cut、FCN 和 MaskFormer 等开创性算法，到近期以 CLIP、Stable Diffusion、DINO 和 SAM 等基础模型（Foundation Models, FMs）为代表的新纪元，其发展轨迹反映了从依赖人工特征工程向端到端学习的重大范式转变 。传统的图像处理方法通常依赖于手工设计的特征和统计模型，而现代深度学习和基础模型则能够直接从像素数据中进行端到端学习，自动提取特征，从而在性能上超越传统方法。这种从显式规则到隐式数据驱动学习的转变，得益于计算能力的提升和大规模数据集的可用性。

### 1.2 图像分割定义与分类体系

图像分割是将数字图像细分为多个互不重叠的区域或片段的过程 [1]。这些区域内的像素通常共享相似的属性（例如灰度、纹理或颜色），而不同区域间的像素属性则存在显著差异 [1]。其根本目标是为图像中的每个像素分配一个特定的标签，以实现对图像内容的精细化理解 。

一个有效的图像分割结果必须满足以下几个基本条件：所有像素都必须被分配到某个子集中；各个子集之间必须互不重叠，并且每个子集内的像素应具有连通性；在每个子集内部，像素的属性应保持一致性，而不同子集间的像素属性则应表现出显著的差异 。

图像分割方法可以根据其处理像素信息的方式进行分类：

1. **基于灰度不连续性:** 这类方法主要通过检测图像中局部亮度的急剧变化来识别区域边界，例如边缘检测技术，常用的算子包括 Sobel 和 Canny 。
2. **基于像素相似性:** 这类方法通过聚合具有相似属性的像素来形成区域，常见的技术包括阈值处理、区域生长和聚类分析，如 Otsu 算法和 K-means 聚类 。

3. **混合方法**: 混合方法结合了上述两种或更多策略，以利用各自的优势。例如，图论方法（如 Graph Cuts）和能量泛函方法（如水平集）以及现代的深度学习方法（如 U-Net）。

下表 1 总结了图像分割的典型分类体系，展示了不同类别算法的代表性实例及其主要特点。

**表 1: 图像分割典型分类体系**

| 类别 | 代表算法 | 特点 |
| --- | --- | --- |
| 阈值分割 | Otsu、迭代法 | 简单高效，适合背景单一图像 |
| 边缘检测 | Canny、LoG 算子 | 依赖梯度变化，对噪声敏感 |
| 区域分割 | 分水岭、区域生长 | 需种子点，易过分割或欠分割 |
| 基于图论 | FH 算法、Graph Cuts | 全局优化，计算复杂度高 |
| 深度学习 | FCN、U-Net、DeepLab | 端到端学习，高精度但需大量标注数据 |

从表 1 中可以看出，图像分割算法存在一个性能与资源需求之间的权衡谱系。例如，阈值分割方法以其“简单高效”而著称，但其适用范围通常限于“背景单一图像”。这种对比揭示了在算法选择过程中，研究人员和工程师必须在效率、精度和所需资源之间进行多目标优化，而非简单追求单一指标的最优。

这种分类体系也反映了图像分割技术从启发式方法向数据驱动学习的演进。早期的经典方法，如 Canny 边缘检测和 Otsu 阈值分割，是基于预定义的数学规则（如梯度变化和方差最大化）来执行操作的 。区域生长算法则依赖于明确的像素相似性准则 。然而，随着图像复杂度的增加，手动设计这些规则变得越来越困难且效率低下。深度学习的出现标志着一个根本性的转变，它使得模型能够直接从大规模数据中学习复杂的模式，从而在处理复杂、多变的真实场景时展现出更高的精度和泛化能力。

## 2. 经典图像分割算法原理

### 2.1 基于阈值分割：Otsu 算法

Otsu 算法（大津法）是一种自动确定图像二值化阈值的常用方法，尤其适用于灰度直方图呈现双峰或

多峰分布的图像[1]。该算法的核心思想是自动将图像像素分为前景和背景两个类别，并通过选择一个最佳阈值来最大化这两类像素之间的可分离性[5]。

在数学上，Otsu 算法利用灰度直方图的零阶和一阶累积矩来推导最佳阈值[6]。假设 p(i) 是灰度级 i 处像素的归一化计数（即概率）。对于给定的阈值 T，Otsu 算法旨在找到使类间方差（Between-Class Variance, BCV）最大化的 T 值 。BCV 的一种常见公式为：

$$\sigma B^2(k) = w0(k) * (\mu0(k) - \mu T)^2 + w1(k) * (\mu1(k) - \mu T)^2,$$

其中 w0 和 w1 分别是阈值 k 分割出的两个类别的概率，μ0 和 μ1 是它们各自的灰度均值，而 μT 是图像的全局灰度均值[7]。更简洁的表示形式是 BCV = Na * Nb * (Ma - Mb)^2，其中 Na 和 Nb 分别是阈值上方和下方像素的归一化计数（概率），Ma 和 Mb 是它们各自的平均灰度值 。

Otsu 算法的优势在于其统计学上的优化特性。通过最大化类间方差，它能够自动寻找最具统计学意义的分离点，从而在一定程度上对图像整体亮度的变化具有鲁棒性，只要图像中前景和背景的相对强度差异保持一致即可[5]。它最适用于“背景单一”的图像 ，即前景和背景具有明显且均匀的灰度差异。在真实世界中，图像往往包含不均匀的光照、阴影、或多个具有不同灰度分布的对象，这些复杂因素会导致直方图不再呈现清晰的双峰模式。

### 2.2 基于区域分割：分水岭算法

分水岭算法是一种基于区域的图像分割方法，其灵感来源于水文学中的分水岭概念[1]。该算法将图像视为一个地形景观，其中像素的亮度值对应于“海拔”高度（通常较暗的像素被视为较低的“最小值”或盆地） 。算法模拟水流从最低点开始填充盆地的过程，当来自不同盆地的水流即将汇合时，便会建立“堤坝”，这些堤坝最终构成了图像的“分水岭”线 。

分水岭算法的典型处理步骤如下：

1. **计算分割函数**: 通常，图像的梯度图或强度图被用作分割函数。对于多光谱图像，需要将其转换为单波段图像。这可以通过“边缘方法”（使用 Sobel 边缘检测来突出高对比度区域）或“强度方法”（对选定波段的像素值进行平均）来实现 。
2. **根据尺度级别修改地图**: 这一步通过累积分布函数确定一个“尺度级别”值，用于调整分割函数。例如，在“边缘方法”中，可以丢弃低梯度值以保留最显著的边缘；在“强度方法”中，则设定一个起始海拔高度进行“注水” 。
3. **执行分水岭变换**: 算法将像素按灰度值从小到大排序，从最低点（最小值）开始“注水”，并将图像划分为“盆地”（像素强度相似的区域），这些盆地由“分水岭”线分隔开 。

标记控制分水岭算法通过引入显式的前景和背景标记，有效地将先验知识融入到分割过程中。这意味着，虽然纯粹基于数据（或强度）的方法具有一定能力，但结合这些人工或算法生成的先验信息，能够显著克服算法固有的局限性（如过分割），从而获得更具实用价值的分割结果 。这种混合方法在实际应用中往往表现出更优越的性能。

即使在经典的图像处理范畴内，复杂的特征工程（不仅仅是简单的梯度计算）也是生成鲁棒输入、从而实现有效分割的关键。这些预处理步骤，通过一系列精心设计的操作，使得图像更适合分水岭算法

的处理，这在某种程度上预示了深度学习中多层特征提取的复杂性。

## 3. 深度学习图像分割模型

### 3.1 全卷积网络 (FCN)

全卷积网络（Fully Convolutional Networks, FCN）是语义图像分割领域的开创性模型，其核心创新在于用卷积层取代了传统分类网络末端的全连接层 [1]。这一改变使得 FCN 能够接受任意尺寸的输入图像，并生成相应尺寸的像素级预测输出 。这种端到端（pixels-to-pixels）的学习方式，极大地简化了图像分割的流程 [16]。

FCN 的架构通常基于预训练的图像分类网络（如 AlexNet、VGG 或 GoogLeNet），通过将它们的全部或部分层转换为卷积层，并进行微调以适应分割任务，从而继承了这些网络强大的特征提取能力 。

FCN 的关键创新包括：

1. **网络内上采样（反卷积）**: 由于分类网络通常会通过池化操作降低特征图的分辨率，FCN 通过引入“反卷积层”（也称为转置卷积或学习上采样）来恢复原始输入图像的分辨率 。这些层能够学习非线性的上采样，从而生成更精细的分割结果。
2. **跳跃连接（Skip Architecture）**: 这是 FCN 最核心的创新之一，旨在解决深度卷积网络在下采样过程中丢失空间细节的问题 [1]。跳跃连接将浅层（包含丰富外观信息和高分辨率细节）的特征图与深层（包含高层次语义信息但分辨率较低）的特征图进行融合。例如，FCN-16s 融合了最终层和 pool4 层（步长 16）的预测，而 FCN-8s 则进一步加入了 pool3 层（步长 8）的预测，从而逐步生成更精确和细致的分割结果 。

FCN 的这些创新，特别是其“全卷积”特性和“跳跃连接”，奠定了现代语义分割架构的基础 。后续的深度学习分割模型，如 U-Net、DeepLab 和 SegNet 等，无一例外地都借鉴了这些核心思想，并通过改进上采样机制、增强跳跃连接或优化上下文聚合方式来进一步提升性能 。FCN 作为这些模型的概念先驱，其重要性不言而喻。

### 3.2 U-Net 及其变体

U-Net 是一种为生物医学图像分割而设计的卷积网络架构，以其独特的对称编码器-解码器结构而闻名 。该网络能够通过强大的数据增强策略，仅使用少量标注图像进行端到端训练，并在医学图像分割任务中取得了显著成功 。

U-Net 的关键创新在于：

- **对称跳跃连接**: 这是 U-Net 最显著的特征。收缩路径中的高分辨率特征图被复制并直接拼接（通过裁剪以匹配尺寸）到扩展路径中对应的上采样输出。这一机制对于精确的像素级定位至关重要，因为它弥补了下采样过程中丢失的精细空间细节，并能将深层抽象特征中的上下文信息有效地传播到高分辨率的解码层 。
- **强大的数据增强**: U-Net 通过广泛应用数据增强技术，有效地利用了有限的标注样本 。这对于医学图像等难以获取大量像素级标注数据的领域尤为重要，显著降低了数据标注的成本和难度。

U-Net 能够仅用“极少量标注图像”进行有效训练的能力 ，是其最重要的创新之一。这直接解决了像素

级标注数据成本高昂和稀缺的问题，尤其在医学影像等专业领域，标注通常需要专家知识且耗时费力。U-Net 的设计哲学在于最大化有限数据的利用效率，这与 FCN 最初依赖大规模预训练分类网络的策略形成对比。

### 3.3 Mask R-CNN

Mask R-CNN 是一个概念简洁、灵活且通用的框架，专为**目标实例分割**而设计[1]。它不仅能够高效地检测图像中的对象，还能同时为每个检测到的对象实例生成高质量的分割掩膜。Mask R-CNN 是 Faster R-CNN 的扩展。

Mask R-CNN 的架构沿用了 Faster R-CNN 的两阶段处理流程：

1. **区域提议网络（RPN）**: 第一阶段，RPN 负责生成候选的目标边界框 。
2. **第二阶段**: 在这一阶段，Mask R-CNN 除了并行预测每个候选区域（Region of Interest, RoI）的类别标签和边界框回归偏移量之外，还额外添加了一个分支，用于预测该 RoI 的二值掩膜。

Mask R-CNN 的关键创新包括：

1. **掩膜预测与类别预测并行**: Mask R-CNN 将掩膜预测与类别分类任务解耦 。在训练过程中，它使用多任务损失函数 L = Lcls + Lbox + Lmask。其中，Lcls（分类损失）和 Lbox（边界框损失）与 Fast R-CNN 相同。掩膜分支输出 K 个 m × m 的二值掩膜（对应 K 个类别），并对输出应用逐像素的 Sigmoid 激活函数。Lmask 被定义为平均二值交叉熵损失，且仅对真实类别 k 对应的掩膜计算损失 。
2. **全卷积网络（FCN）用于掩膜预测**: Mask R-CNN 使用 FCN 从每个 RoI 预测一个 m × m 的掩膜 。这种设计使得掩膜分支的每一层都能保持明确的 m × m 对象空间布局，而不是将其折叠成缺乏空间维度的向量表示。
3. **RoIAlign**: 这是 Mask R-CNN 最核心的创新，显著提升了像素级精确掩膜的预测能力[1]。RoIAlign 通过以下方式解决了这一错位问题：避免对 RoI 边界或 bin 进行任何量化；使用双线性插值计算每个 RoI bin 内四个规则采样位置的精确输入特征值；聚合这些结果（通过最大池化或平均池化） 。 RoIAlign 对掩膜精度有显著影响，尤其是在高 IoU（交并比）标准下，可将掩膜 AP（平均精度）提高约 3 个百分点 。

Mask R-CNN 的主要优势在于其在实例级分割任务中实现了高精度（约 85% IoU），尤其在 COCO 等数据集上表现出色 。它还可以推广到人体姿态估计等其他任务 。然而，其缺点是计算资源需求极高 。

Mask R-CNN 的核心创新 RoIAlign 直接解决了实例分割中对**像素级精度**的迫切需求。这突显出，对于需要精确对象形状（而非仅边界框）的任务，即使是池化操作引入的微小错位也可能显著降低性能。这表明分割不仅是识别“是什么”和“在哪里”，更是要精确到“到底在哪里”和“是什么形状”。

### 3.4 DeepLab 系列

DeepLab 系列模型（包括 v1、v2、v3 和 v3+）专注于语义图像分割，特别是为了有效处理多尺度上下文信息和精确恢复目标边界而设计[1]。

DeepLab 系列的关键创新包括：

1. **空洞空间金字塔池化（Atrous Spatial Pyramid Pooling, ASPP）**: ASPP 是 DeepLabv3 和 DeepLabv3+中的关键模块，旨在捕捉多尺度上下文信息。它通过并行应用多个具有不同空洞率的空洞卷积来探测卷积特征，从而获取不同尺度的上下文信息。DeepLabv3 进一步将 ASPP 模块与图像级特征相结合，以编码全局上下文信息 [4]。
2. **DeepLabv3+（编码器-解码器结构）**: DeepLabv3+是 DeepLabv3 的扩展，通过添加一个简单而有效的解码器模块来精细化分割结果，尤其是在对象边界处。其编码器（DeepLabv3 的输出）首先进行上采样，然后与骨干网络中的低级特征进行拼接，并通过后续的卷积操作进行精炼 。
3. **Xception 模型和深度可分离卷积的集成**: DeepLabv3+进一步探索了 Xception 模型，并将深度可分离卷积应用于 ASPP 和解码器模块 [8]。这种方法通过将标准卷积分解为深度卷积和逐点卷积，显著降低了计算成本和参数数量，同时保持或略微提升了性能，从而构建了一个更快、更强的编码器-解码器网络 。

DeepLab 系列模型的优势在于其在复杂场景语义分割中实现了极高的精度（约 89% IoU）。

DeepLab 的演进，特别是空洞卷积和 ASPP 与 DeepLabv3+中编码器-解码器结构的结合 ，代表了一种精巧的方法，能够同时捕捉广泛的上下文信息（通过空洞卷积/ASPP 的不同感受野）和细粒度的空间细节（通过解码器和低级特征融合）。这比 FCN 的跳跃连接更进一步地解决了“是什么”和“在哪里”的问题。其设计旨在克服早期模型的局限性，通过在一个强大框架内明确处理全局上下文和局部细节，超越了简单的特征拼接。

### 3.5 SegNet

SegNet 是一种新颖且实用的深度全卷积编码器-解码器架构，专为像素级语义分割而设计，其主要动机是应用于场景理解任务。该模型在推理过程中，尤其注重内存效率和计算时间效率 [9]。

SegNet 的关键创新在于：**利用池化索引进行非线性上采样**: 解码器利用在对应编码器最大池化步骤中计算得到的**池化索引**来执行非线性上采样。这种方法避免了模型需要学习上采样过程，从而减少了参数量。通过这种方式上采样得到的特征图是稀疏的，随后再通过可训练的卷积滤波器进行卷积，生成密集的特征图 。

SegNet 的优势包括：改善了边界勾勒的精确性，减少了参数数量（使得端到端训练更容易），在推理过程中具有内存和计算效率，适用于移动端部署。它还具有较高的处理速度（16 FPS）。然而，与更先进的模型相比，其精度相对中等（约 72% IoU）。

SegNet 利用池化索引 是一种巧妙且资源高效的设计选择。它通过重用下采样路径中的信息，而非学习复杂的反卷积滤波器或依赖简单的双线性上采样，显著减少了参数数量和计算开销。这表明了 SegNet 在早期就注重轻量化和实时性能，为后续轻量化架构的发展奠定了基础。

### 3.6 金字塔场景解析网络 (PSPNet)

金字塔场景解析网络（Pyramid Scene Parsing Network, PSPNet）引入了金字塔池化模块（Pyramid Pooling Module, PPM）来聚合多尺度上下文信息，为场景解析提供全局先验表示。这对于复杂场景中的像素级预测至关重要。

PSPNet 的关键创新在于：

- **金字塔池化模块（PPM）**: PPM 旨在克服卷积神经网络经验感受野的局限性，这些感受野通常小于其理论值，导致全局场景先验信息的整合不足 。PPM 通过融合来自不同子区域、具有不同感受野的信息来解决这一问题，从而提供更强大的表示。 PPM 通过在四个不同的金字塔尺度下融合特征来构建分层全局先验 ：①**最粗糙级别（全局池化）**: 第一个级别执行全局池化，生成一个单一的 bin 输出，捕捉整个图像最普遍的上下文。②**子区域池化**: 随后的金字塔级别将特征图划分为不同的子区域（例如，2x2、3x3、6x6 的 bin），形成不同粒度的池化表示。③**降维与上采样**: 为了保持全局特征的权重，在每个金字塔级别之后应用 1x1 卷积层进行降维，然后通过双线性插值上采样到与原始特征图相同的尺寸。④**拼接**: 最后，将来自不同金字塔级别的这些上采样特征拼接起来，形成全面的全局特征 。

PSPNet 的优势在于其在复杂场景（如街景）中表现出色，能够有效地捕捉局部和全局上下文信息。

PSPNet 中的 PPM 是对场景理解中上下文歧义问题的一种直接的架构响应。简单的全局池化可能丢失空间关系 ，导致对包含大量对象的复杂场景产生误解。PPM 的多级池化明确地在不同尺度上捕捉上下文，从而实现对场景更细致的理解。这凸显了不仅要拥有上下文，更要有效地“组织”上下文的重要性。

PSPNet 将预训练的 ResNet 骨干网络（带有空洞网络策略）与 PPM 相结合 ，展示了一种强大的组合：一个强大的特征提取器提供局部细节，并通过专门的模块增强全局上下文。这种模块化设计，即通用骨干网络通过任务特定的上下文聚合模块得到增强，是高性能分割模型中反复出现的主题。

| 算法 | 网络结构 | 核心机制 | 优点 | 缺点 | 适应场景 |
| --- | --- | --- | --- | --- | --- |
| FCN | 将 CNN 网络的全连接层都替换为卷积层 | 使用反卷积进行上采样；跳跃连接 | 实现了端到端的语义分割；能接受任意大小图片输入 | 损失了空间信息 | 像素级分割场景 |
| U-Net | 对称的编码器解码器架构 | U 型结构；overlap-tile 等 | 叠操作拼接，恢复部分语义信息，提高精度；更好处理图像的边缘细节 | 上采样的拼接方法过于粗糙 | 医学图像分割 |
| SegNet | VGG 编码器，完全对称的解码器 | 上采样操作只存在于解码器部分，且和编码器的池化层一一对应，其上采样仅利用池化索引 | 在分割性能较好的同时，使得参数量和推理时间都有显著的降低 | 在精度上没有提升 | 适合移动设备上的推理 |

## 4. 性能评估与比较分析

### 4.1 常用评估指标

图像分割的评估相比简单的图像分类更为复杂，因为它不仅需要判断像素的类别，还需要精确量化预测分割与真实标注掩膜之间的重叠程度 。

1. **像素精度（Pixel Accuracy）**: 这是最简单的评估指标，衡量像素分类的精确度 。然而，当图像中目标像素占比较小（例如，图像大部分是背景）时，像素精度可能会被高估，因为即使模型没有有效分割目标，背景的正确分类也会导致高分，使其在某些情况下变得无用 。
2. **交并比（Intersection over Union, IoU）/ 杰卡德指数（Jaccard Index）**: IoU 是图像分割中最常用的评估指标之一，它量化了预测分割（Predicted）与真实标注（Ground Truth, GT）之间的重叠程度。其定义为预测区域与真实区域交集面积除以两者并集面积的比值：IoU = |GT ∩ Predicted| / |GT ∪ Predicted| 。IoU 值越接近 1，表示预测结果与真实标注越相似 。
3. **Dice 系数（Dice Score）**: 另一个广泛使用的评估指标，它表示预测区域与真实区域重叠面积的两倍，再除以两者面积之和。公式为：Dice = 2 * |GT ∩ Predicted| / (|GT| + |Predicted|) 。与 IoU 类似，Dice 系数越接近 1 表示预测结果越完美。
4. **计算时间（Computation Time / FPS）**: 衡量算法生成初始分割掩膜所需的时间，这对于自动驾驶等需要实时或近实时性能的应用至关重要。
5. **计算资源需求（Computational Resource Requirements）**: 评估算法所需的 GPU/CPU 内存和处理能力。
6. **用户交互时间（User Interaction Time）**: 对于交互式分割方法，此指标衡量用户为精炼初始分割结果所花费的时间。

除了 IoU 和 Dice 等精度指标外，计算时间、计算资源需求和用户交互时间等指标的纳入，表明对图像分割算法的评估，尤其是在实际部署场景中，远不止于像素的“正确性”。实用性、效率和用户体验（对于交互式系统）同样关键，这反映了将研究原型转化为可部署解决方案所面临的工程挑战。

### 4.2 主流算法性能对比

下表 2 总结了主流深度学习图像分割算法在精度、速度、计算资源需求和适用场景方面的性能对比。

**表 2: 主流深度学习图像分割算法性能对比**

| 算法 | 精度（IoU/Dice） | 速度（FPS） | 计算资源需求 | 适用场景 |
|---|---|---|---|---|
| FCN | 中等（~75% IoU） | 低（0.5 FPS） | 高 | 通用语义分割 |
| U-Net | 高（~90% Dice） | 中（10 FPS） | 中等 | 医学图像、小样本数据 |
| Mask R-CNN | 高（~85% | 低（5 FPS） | 极高 | 实例级分割（如 |

| | | | | |
|---|---|---|---|---|
| | IoU） | | | COCO 数据集） |
| DeepLab v3+ | 极高（~89% IoU） | 中（8 FPS） | 高 | 复杂场景语义分割 |
| SegNet | 中等（~72% IoU） | 高（16 FPS） | 低 | 实时应用、移动端部署 |
| PSPNet | 极高 | 中 | 高 | 复杂场景（如街景）中表现优异 |

从表 2 中可以清晰地看出，没有哪一种算法能在所有指标上都达到最优。相反，每种模型都有其独特的优势和劣势，使其适用于特定的应用场景。例如，U-Net 在医学图像和处理小样本数据方面表现出色，SegNet 则更适合实时应用和移动端部署，而 Mask R-CNN 专注于实例级分割，DeepLab 和 PSPNet 则在复杂场景的语义分割中展现出卓越性能 。这强调了算法选择是一个高度依赖上下文的工程决策，而非简单的性能排名。

计算资源需求与适用场景之间的关联性也揭示了硬件能力和部署环境（例如边缘设备与云端 GPU）是驱动架构设计的关键因素。例如，Mask R-CNN 和 DeepLab 需要 GPU 支持，而 SegNet 可以部署在边缘设备上 。这种对轻量级模型的需求（如 SegNet）与高精度模型（如 Mask R-CNN 和 DeepLab）的高资源需求形成对比，表明了性能与可部署性之间持续存在的权衡。这种权衡推动了研究人员在精度和效率之间寻找更优的平衡点，以满足不同实际应用的需求。

| Methods | Backbones | Times (ms) | | Memory (MB) | | Model Size (MB) | Batch Size |
|---|---|---|---|---|---|---|---|
| | | Train | Infer | Train | Infer | | |
| Mask R-CNN [12] | R-101 | 251.2 | 74.7 | 5407 | 1876 | 480 | 2 |
| Mask R-CNN [12] | RX-101 | 316.5 | 87.5 | 6398 | 1872 | 477 | 2 |
| Mask R-CNN [12] | R2-101 | 359.4 | 83.1 | 6648 | 1923 | 485 | 2 |
| Mask R-CNN [12] | S-50 | 214.1 | 79.5 | 7414 | 1657 | 260 | 4 |
| Mask R-CNN [12] | RegNet | 203.7 | 59.6 | 7414 | 1657 | 260 | 4 |
| Mask R-CNN [12] | R-50 | 361.7 | 60.1 | 6499 | 1769 | 334 | 4 |
| MS R-CNN [17] | R-50 | 246.6 | 62.7 | 7249 | 1794 | 428 | 4 |
| CARAFE [21] | R-50 | 515.7 | 65.2 | 7373 | 1873 | 376 | 4 |
| Cascade M-R-CNN [47] | R-50 | 820.9 | 77.6 | 5280 | 2005 | 587 | 2 |
| HTC [18] | R-50 | 893.4 | 78.9 | 6165 | 2314 | 588 | 2 |
| GRoIE [23] | R-50 | 414.3 | 110.7 | 5407 | 2030 | 363 | 2 |
| SCNet [22] | R-50 | 311.8 | 82.4 | 8273 | 2406 | 699 | 4 |
| YOLACT [13] | R-50 | 64.0 | 32.8 | 10,513 | 7610 | 265 | 8 |
| YOLACT [13] | R-101 | 82.7 | 40.9 | 10,001 | 7758 | 410 | 8 |
| BlendMask [24] | R-50 | 276.9 | 58.9 | 4669 | 1641 | 274 | 4 |
| CondInst [26] | R-50 | 312.8 | 56.7 | 5385 | 1713 | 259 | 4 |
| SOLOV2 [27] | R-50 | 392.7 | 62.4 | 5025 | 3123 | 354 | 4 |
| BoxInst [29] | R-50 | 454.6 | 55.0 | 9276 | 1681 | 261 | 4 |
| LosNet | R-101 | 704.2 | 35.4 | 5397 | 1525 | 338 | 12 |
| LosNet | R-50 | 539.9 | 32.3 | 3759 | 1467 | 193 | 12 |
| LosNet | M-V3 | 448.7 | 25.2 | 2838 | 1359 | 26 | 12 |

## 5. 典型应用案例

### 5.1 医学影像分析

图像分割在医疗健康领域扮演着举足轻重的角色，是疾病诊断、治疗规划和病情监测的关键工具 [10]。

1. **肿瘤分割**: U-Net 广泛应用于 MRI/CT 图像中的脑肿瘤精确分割，为手术规划提供辅助 [7]。Mask R-CNN 则用于肺部结节检测，结合三维重建技术可提升诊断效率 。
2. **器官定位与组织体积测量**: 这对于疾病诊断、解剖结构研究以及手术规划和模拟至关重要。例如，对脾脏、肝脏、胰腺、肾脏和胆囊等器官的分割 。
3. **放射治疗**: 图像分割用于精确确定放射治疗的靶区和剂量规划 。

在医疗领域，对精度和可靠性的要求极高，因为诊断和治疗的失误可能带来严重后果。U-Net 在医学图像处理中的广泛应用，以及对“精确分割”的持续强调 ，都凸显了在医疗这类高风险应用中，算法的准确性和鲁棒性是首要考量。这促使研究人员和临床医生选择那些能够提供高度可靠结果的算法，即使这意味着更高的计算需求或需要特定的训练策略（如 U-Net 的数据效率）。

然而，医学图像的像素级标注工作通常是“劳动密集型、耗时且需要专家知识”的 [18]。这种高昂的标注成本是深度学习在医学领域规模化应用的一个显著瓶颈，直接推动了弱监督学习和少样本学习成为重要的研究方向 [1]。这些方法旨在通过更少或更不精确的标注来训练模型，从而降低标注负担，加速医疗 AI 解决方案的开发和部署。

### 5.2 自动驾驶与机器人视觉

图像分割是实现自动驾驶车辆和机器人智能感知与运动控制不可或缺的技术。

1. **道路识别**: DeepLab v3+能够实时分割车道线和交通标志，显著增强车辆的环境感知能力 。
2. **障碍物检测**: PSPNet 用于识别行人、车辆等障碍物，从而提升路径规划的安全性 。
3. **导航**: 图像分割为机器人导航提供关键的视觉信息 。
4. **智慧城市**: 图像分割技术也为智慧城市中的实时交通监控和视频监控系统提供支持 。

在自动驾驶等涉及生命安全的实时系统中，速度和低延迟是至关重要的。一个准确但运行缓慢的分割算法将无法及时提供决策所需的信息。这直接推动了对轻量级架构和高效推理算法的研究，以满足实时处理的需求 [1]。

此外，自动驾驶环境本质上是一个"开放世界"，这意味着系统必须能够处理各种未预见的物体、光照和天气条件。这促使图像分割算法必须具备强大的泛化能力，能够鲁棒地处理训练数据之外的新场景，从而推动了开放世界学习和提示式基础模型等研究方向的发展 。

### 5.3 遥感与工业检测

图像分割对于分析大规模空间数据和确保产品质量至关重要 [1]。

1. **土地分类**: FCN 被用于分割卫星图像，以支持城市规划或灾害评估 。语义分割和实例分割技术能够自动化识别不同的地形和地貌特征 。
2. **缺陷检测**: 分水岭算法 和深度学习模型 在工业质量检测中用于定位产品表面瑕疵。
3. **农业**: 图像分割帮助农民估算作物产量和检测杂草，从而优化农业实践 。

遥感应用（如土地分类）涉及处理海量高分辨率图像数据，这要求算法不仅要精确，还要具备高度的可扩展性和效率。这促使了能够处理大规模输入或进行分布式数据处理的算法的发展 。

尽管深度学习在许多领域占据主导地位，但分水岭算法在"缺陷检测"中的持续应用 表明，对于某些特定且定义明确的工业任务，计算量较小的经典算法仍然有效且更受青睐。特别是在深度学习数据稀缺或问题复杂度较低的情况下，经典方法可能比开发和训练复杂的深度学习模型更具成本效益。这表明"最佳"解决方案并非总是最复杂的。

## 6. 挑战与未来发展趋势

### 6.1 小样本与弱监督学习

**挑战**:深度学习模型对大规模像素级标注数据集的依赖性是其主要挑战之一 [1]。这些标注工作成本高昂、耗时且需要专业知识，尤其在三维医学图像领域更为突出 [23]。此外，传统模型在遇到训练集中未出现的新类别时，往往难以有效识别 。

像素级标注的高成本和高难度是弱监督学习和少样本学习迅速发展和兴起的根本原因 [1]。这种标注瓶颈构成了深度学习解决方案规模化应用的一个基本经济和实践限制。

**弱监督学习（Weakly Supervised Learning, WSL）**: 旨在通过使用稀疏标注（如图像级标签、边界框、涂鸦或点）而非完整的像素级掩膜来降低标注成本 [8]。通常涉及一个图像分类模型来预测类别，然后利用类别激活图（CAMs）来检测最具判别力的区域，将其作为监督模型的种子标注 。近期研究

提出了针对三维医学图像的概率感知型弱监督学习流程，利用点标注生成密集的伪标签，并结合基于概率的损失函数和注意力机制 [3]。迭代训练和精炼方案也能有效提升性能 。

**少样本学习（Few-shot Learning, FSL）**: 使模型能够从有限的示例中学习通用规律，并将其应用于新问题，从而预测新的图像类别 [5]。通常包括一个特征提取网络（共享权重）用于支持图像和查询图像，一个关联映射学习模块，以及一个用于目标网络分割的引导模块 。基于迁移学习的 FSL 通过在较小数据集上微调预训练模型来实现 。元学习方法利用先验知识从少量示例中进行泛化 。基础模型如 SAM 通过用户提供的提示实现可泛化的零样本分割 。

从仅仅减少标注负担（弱监督学习）到预测全新的、未曾见过的类别（少样本学习，以及 SAM 等基础模型的零样本能力 [2]），这代表了图像分割领域的一个重大飞跃。这表明研究重点已从降低现有任务成本转向使人工智能系统能够在真正的"开放世界"场景中运行，即能够处理以前未建模的类别。这种演进不仅提升了效率，更拓展了人工智能感知的边界。

### 6.2 实时性优化与轻量化模型

**挑战**: 许多应用场景（如自动驾驶、移动设备部署、即时医疗影像分析）对实时或近实时性能有严格要求 。然而，复杂的深度学习模型往往伴随着高昂的推理时间和内存消耗，这限制了它们在资源受限环境中的部署。

对"轻量化"模型和"实时性优化"的明确关注 [1] 是对工程现实的直接回应：虽然原始精度很重要，但对于实际部署而言，它往往不足够。这表明研究重点已从单纯最大化基准性能，转向优化一个包含速度、内存和功耗在内的多目标函数。

**轻量化模型**: 旨在以更少的参数和计算量实现高精度，同时保持快速推理时间 [6]。实时性优化技术: ①空洞卷积: 在不增加参数的情况下扩大感受野（改进型 ENet ）。②模型压缩: 移除冗余操作（改进型 ENet ）。③优化推理: 高效的前向传播，视频中的时间分摊 。④专用硬件/部署: 为特定硬件（如边缘设备）设计模型 。

轻量级架构（如编码器-解码器、多分辨率、双路径、基于注意力、基于金字塔、基于 Transformer ）和优化技术（如分解卷积、池化索引、重参数化）的多样性，表明计算约束是推动架构创新的强大催化剂。研究人员正在探索所有可能的途径，以更少的资源实现更高的性能。

### 6.3 多模态数据融合与 3D 分割

**3D 分割**:
**挑战**: 对于医学影像（如 CT/MRI 序列的三维渲染、肿瘤/器官分割）和自动驾驶/机器人技术至关重要 [7]。然而，三维体数据通常包含数百甚至数千个切片，使得标注工作极其耗时 。挑战还包括可扩展性、计算需求、纯 Vision Transformer（ViT）模型缺乏归纳偏置、数据稀缺以及标注一致性等 。

转向基于深度学习的技术，特别是 Vision Transformer（ViTs）和混合 Vision Transformer（HVTs）[10]。HVTs 结合了 CNN（用于局部特征）和 ViTs（用于全局关系）的优势，应用于编码器/解码器路径 。解决数据问题、改进训练范式、集成方法以及融入领域知识也是当前的重要趋势 。

**多模态数据融合**:
旨在整合来自多种模态（如 RGB、深度、热成像、音频、文本）的信息，以实现更准确的预测 [6]。挑战包括噪声数据（异构噪声、传感器误差、环境干扰）、不完整数据（模态缺失）、不平衡数据（模态间的偏差/差异）以及动态变化的质量 [5]。
利用 ViTs 处理序列数据的能力来处理多模态输入，捕捉不同数据源之间的依赖和交互 。开发灵活可靠的融合方法，以处理不完整数据 。根据动态变化的质量，动态整合数据 。

向三维分割和多模态融合的演进 标志着数据维度和异构性的显著提升。这种复杂性要求根本性的新架构范式（如三维 CNN、ViTs、HVTs）和复杂的融合策略，这些策略能够有效地结合不同的信息源，同时管理计算负载和数据质量问题。

三维分割和多模态融合的追求，最终是为了使人工智能系统能够对真实世界获得更全面、更鲁棒的理解，这与人类感知能力相仿。人类以三维方式感知，并整合来自多种感官的信息。这些研究方向旨在赋予人工智能类似的能力，从而在复杂环境中构建更智能、更可靠的系统。

## 7. 结论

图像分割作为计算机视觉领域的基础性任务，其重要性不言而喻，并持续经历着显著的演进。本报告回顾了从 Otsu 阈值分割和分水岭算法等经典方法，到 FCN、U-Net、Mask R-CNN、DeepLab、SegNet 和 PSPNet 等先进深度学习模型的演变历程，并详细阐述了它们的核心原理、架构创新和各自的优势。

研究表明，在选择图像分割算法时，需要根据具体的应用需求进行权衡，综合考虑精度、速度和计算资源消耗。例如，U-Net 在医学图像和处理小样本数据方面表现卓越，而 SegNet 则更适合实时应用和移动端部署。这种多维度的性能考量，使得没有单一的“最佳”算法，只有最适合特定场景的解决方案。

图像分割领域的发展呈现出动态性，其进步由一系列关键挑战所驱动：数据稀缺性（尤其在像素级标注成本高昂的领域）、对实时处理能力的迫切需求以及对多维度感知能力的追求。这些挑战促使了弱监督学习、少样本学习、轻量化模型、多模态数据融合和三维分割等前沿研究方向的兴起。这些看似独立的研究方向正在逐步融合，共同致力于构建更鲁棒、更具泛化能力且更易于部署的分割系统。这种研究的融合，反映了该领域正从单纯追求算法性能向实现系统级影响转变。

展望未来，图像分割将继续作为智能感知的基石，其不断发展不仅将提升现有应用的性能，还将催生更多创新性的应用。该领域正从证明算法能力转向展示系统价值，其成功将日益体现在对实际系统和各行各业的深远影响上。

## 6. 参考文献

[1]姜枫,顾庆,郝慧珍,等.基于内容的图像分割方法综述[J].软件学报,2017,28(1):160-183.
[2]LONG J,SHELHAMER E,DARRELL T. Fully convolution-al networks for semantic segmentation [J].IEEE Transac-tions on Pattern Analysis and Machine Intelligence, 2015,39
(4):640-651.
[3]王赛男,郑雄风.基于边缘计算的图像语义分割应用与研究[J].计算机科学,2020,47(S2):276-280.
[4]黄鹏,郑淇,梁超.图像分割方法综述[J].武汉大学学报:理学版,2020,66(6):519-531.

[5]OTSUN. A thrshold selection method from gray-level histo-grams[J]. IEEE Transactions on Systems Man & Cybernet-ics,2007,9(1):62-66.
[6]KAPUR J N,SAHOO P K, WONG A K C.A new methodfor gray-level picture thresholding using the entropy of thehistogram[J]. Computer Vision Graphics and Image Process-
ing,1980,29(3):273-285.
[7]YEN J C,CHANG FJ,CHANG S.A new criterion for auto-matic multilevel thresholding [J]. IEEE Transactions on Im-age Processing,1995,4(3):370-378.
[8]TREMEAU A, BOREL N. A region growing and merging al-gorithm to color segmentation [J]. Pattern Recognition,1997,30(7):1191-1203.
[9]FELZENSZWALB P F, HUTTENLOCHER D P. Efficientgraph-based image segmentation[J]. International Journal ofComputerVision,2004,59(2):167-181.
[10] BOYKOV Y, FUNKA-LEA G.Graph cuts and efficient N-D image segmentation[J]. International Journal of ComputerVision,2006,70(2):109-131.