 DOI: https://doi.org/10.12345/xdjyjz.v4i3.36724

Knowledge Distillation-Based Instance Segmentation System Design

Qiangda Shan Xianfei Zhou Hongfang Cheng

Wuhu Vocational Technical University, Wuhu, Anhui, 241000, China

Abstract

With the rapid development of edge computing and mobile vision applications, there is a growing demand for efficient and low-latency instance segmentation systems. This paper designs and implements a complete real-time instance segmentation system. The system is centered on a lightweight instance segmentation model based on knowledge distillation and adopts a modular, pipelined engineering architecture, achieving end-to-end processing from multi-source data input to instance segmentation result output. This paper focuses on elaborating the engineering design and implementation details of the system, including system architecture, processing flow, module design, and optimization strategies. The system supports various input methods such as cameras, video streams, and image files, and through key technologies such as adaptive preprocessing, multi-threaded parallel inference, and intelligent post-processing, it significantly improves processing efficiency while ensuring segmentation accuracy. Experimental results show that the system achieves an accuracy of 31.0 AP on the COCO dataset and a processing speed of 32 FPS, successfully balancing accuracy and real-time performance. It is suitable for various practical scenarios such as intelligent security, autonomous driving, and industrial inspection.

Keywords

instance segmentation; system design; knowledge distillation; process optimization

基于知识蒸馏的实例分割系统设计

单强达 周先飞 程鸿芳

芜湖职业技术大学，中国 · 安徽 芜湖 241000

摘 要

随着边缘计算与移动视觉应用的快速发展，对高效、低延迟的实例分割系统需求日益增长。本文设计并实现了一套完整的实时实例分割系统，系统以基于知识蒸馏的轻量级实例分割模型为核心，采用模块化、流水线化的工程架构，实现了从多源数据输入到实例分割结果输出的全流程处理。本文重点阐述系统的工程设计与实现细节，包括系统架构、处理流程、模块设计与优化策略。系统支持摄像头、视频流、图像文件等多种输入方式，并通过自适应预处理、多线程并行推理、智能后处理等关键技术，在保证分割精度的同时显著提升处理效率。实验结果表明，本系统在COCO数据集上达到31.0 AP的精度，处理速度达32 FPS，成功实现了精度与实时性的平衡，适用于智能安防、自动驾驶、工业检测等多种实际场景。

关键词

实例分割；系统设计；知识蒸馏；流程优化

1 引言

实例分割 [1] 作为计算机视觉领域的重要研究方向，旨在同时完成目标检测与像素级语义分割，为智能系统提供精细化、实例级的场景理解能力。近年来，随着深度学习技术的快速发展，实例分割算法在精度上取得了显著突破，然而大多数高性能模型如 Mask R-CNN[2]、SOLOv2[3] 等存在参数量大、计算复杂度高的问题，难以在资源受限的边缘设备上实现实时推理。

在实际应用场景中，如自动驾驶 [4]、智能监控、移动机器人等，系统不仅需要高精度的分割结果，更对处理延迟和实时性有着严格要求。传统的实例分割系统往往注重算法精度而忽略工程实现细节，导致在实际部署中难以满足实时性

【基金项目】芜湖职业技术大学校级科学研究项目“基于知识蒸馏的实例分割系统设计”(wzdzr202635)；安徽省特色专业教学资源库项目“计算机应用技术专业教学资源库”(2024jxzyk003)；芜湖职业技术大学教育教学改革研究（重点）项目“AIGC 技术在高职物联网应用技术专业教学中的应用研究”(2025jyzd01)。

【作者简介】单强达（1996–），男，中国湖北黄冈人，硕士，助教，从事计算机视觉研究。

要求 [5]。因此，设计一套兼顾精度与效率、具备良好工程化特性的实时实例分割系统具有重要的理论价值与实践意义。

本文基于前期研究的轻量级知识蒸馏实例分割模型 [6]，从工程实现角度出发，设计并实现了一套完整的实时实例分割系统。与以往研究不同，本文重点聚焦于系统的工程架构设计、处理流程优化、模块化实现等工程性问题，旨在为实例分割技术的实际应用提供一套可参考的工程解决方案。

2 系统总体设计

2.1 系统设计目标

本系统的设计围绕以下核心目标展开：

1. 实时性：系统需在标准硬件配置下实现 30 FPS 以上的稳定处理速度。

2. 精度平衡：在保持较高分割精度的前提下，通过算法优化和工程技巧提高处理效率。

3. 可扩展可维护：采用模块化设计，各功能组件松耦合，便于功能扩展和系统维护。

2.2 系统整体架构

系统采用经典的分层架构设计，如图 1 所示，整体分为硬件层、支持层、处理层和应用层四个层次。

各层功能明确，职责清晰：硬件层提供计算、存储和输入输出能力；支持层负责资源管理、进程通信和系统监控；处理层是系统的核心，包含完整的实例分割处理流水线；应用层提供用户界面和结果管理功能。这种分层设计确保了系统的可扩展性和可维护性。

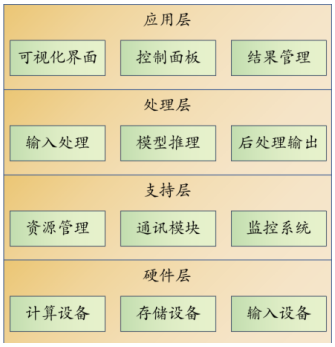

图 1 系统分层架构

2.3 核心处理流程

采用多阶段流水线处理机制，系统启动后，首先完成初始化工作。随后进入如图 2 所示核心处理流程，每帧数据依次经过以下六个关键阶段：

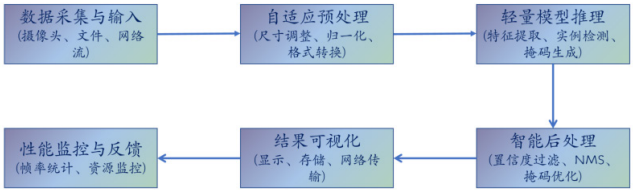

图 2 系统核心处理流程

数据采集。系统根据配置自动选择输入源，支持 USB 摄像头、视频流、图像文件等多种输入方式。采集模块实现智能帧率控制和缓冲管理，确保数据连续稳定供应。

自适应预处理。预处理模块根据当前系统负载和设备性能，动态选择最优处理参数。主要处理包括：图像尺寸调整、色彩空间转换、数据归一化。

轻量模型推理。采用基于知识蒸馏的轻量实例分割模型进行前向计算。推理过程分为三个子阶段：特征提取、实例检测、掩码生成。

智能后处理。后处理模块对模型输出的原始结果进行优化，包括：置信度过滤、改进的 Soft-NMS[10] 处理、掩码二值化与形态学优化、结果整合与格式转换。

结果输出与可视化。输出模块提供多种结果展示和保存方式，包括实时可视化显示、文件存储、网络传输等。用户可根据需要选择合适的输出方式。

性能监控与调优。监控模块实时收集各阶段处理时延、资源使用率等指标。基于这些数据，自适应调优模块动态调整系统参数，如输入分辨率、批处理大小、后处理强度等，确保系统在不同场景下都能保持最佳性能。

3 核心模块详细设计

3.1 数据输入与预处理模块

数据输入模块负责管理系统与外部数据源的连接，支持多种输入方式。模块采用统一的接口设计，对外提供一致的读取接口，对内根据输入类型选择相应的处理策略。预处理模块是保证系统实时性的关键环节，采用自适应处理策略。核心预处理操作包括：

1. 智能尺寸调整：根据目标设备的计算能力和当前系统负载，动态选择输入分辨率。系统维护一个分辨率 - 性能映射表，根据实时监控数据选择最合适的分辨率。

2. 色彩空间转换：将输入图像从 BGR 格式转换为 RGB 格式，并进行标准化处理，使像素值分布符合模型要求。

3. 在线数据增强：仅在训练阶段启用，包括随机翻转、色彩抖动、尺度变换等增强技术，提升模型的泛化能力。

3.2 轻量实例分割模型推理模块

模型推理模块是本系统的核心，负责实例分割的前向计算。模块基于前期研究的轻量级知识蒸馏模型构建，采用教师 - 学生网络架构。在实际部署中，仅使用学生网络进行推理，教师网络仅用于训练阶段的知识迁移。

模型加载与初始化。系统启动时，模型推理模块从指定路径加载预训练权重，初始化网络参数。为减少启动时间，模块采用惰性加载策略，仅在实际需要时分配 GPU 内存。

推理流水线设计。将推理过程分解为数据准备、特征提取、实例检测与掩码生成四个子阶段，形成处理流水线，充分利用硬件并行能力，提升整体吞吐量。

内存与计算优化。使用内存池减少动态分配开销；支

持模型分片加载，适应内存受限设备；在 GPU 环境中启用混合精度计算，在 CPU 环境中使用 OpenMP 并行。

3.3 后处理与结果优化模块

后处理模块负责对模型输出的原始结果进行精炼和优化，提高最终分割结果的质量和一致性。模块采用多阶段处理流程，如图 3 所示，包括置信度过滤、非极大值抑制、掩码优化和结果整合四个主要阶段。

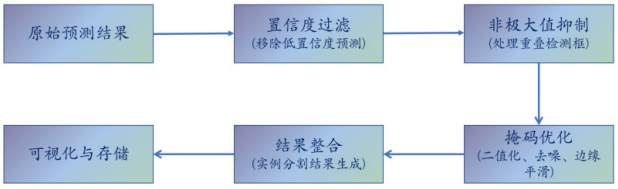

图 3 后处理流程图

4 系统实现

为实现实时处理，系统采用多线程并行架构。如图 4 所示，系统创建多个工作线程，分别负责不同的处理任务，通过精心设计的线程同步和数据共享机制协同工作。

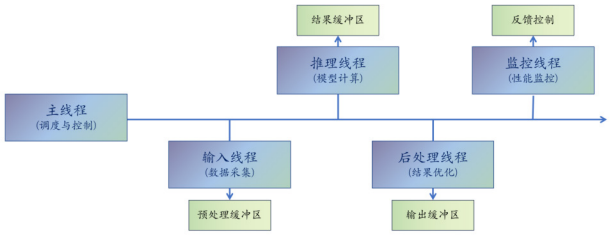

图 4 多线程处理模型

线程间通过共享缓冲区和同步原语进行通信。系统采用双重缓冲机制，一个缓冲区用于写入，另一个用于读取，避免读写冲突。线程同步采用条件变量和互斥锁结合的方式，确保数据的一致性和处理的正确性。线程调度采用优先级策略，推理线程具有最高优先级，确保模型计算的及时性。系统还实现了动态负载均衡，当某个线程成为瓶颈时，自动调整任务分配策略，平衡各线程的工作负载。

整体的 UI 设计如图 5 所示，分为输入区、结果显示区、控制面板、系统监控、统计面板五个区域。在输入区可以选择待处理的文件；控制面板用于系统基本行为控制和模型参数设置；统计面板可展示当前系统的处理性能。

5 结论与展望

本文设计并实现了一套完整的实时实例分割系统，系统以轻量级知识蒸馏模型为核心，通过创新的工程架构和优化策略，在保证分割精度的同时实现了实时处理能力。未来工作可从以下几个方面展开：首先，探索更高效的神经网络架构和知识蒸馏策略，进一步提升模型的精度和效率平衡。其次，研究更智能的自适应优化算法，使系统能够学习最优的处理策略。第三，扩展系统的输入模态支持，融合深度、红外等多模态信息，提升系统在复杂环境下的鲁棒性。最后，研究分布式部署和联邦学习技术，支持大规模分布式实例分割应用。

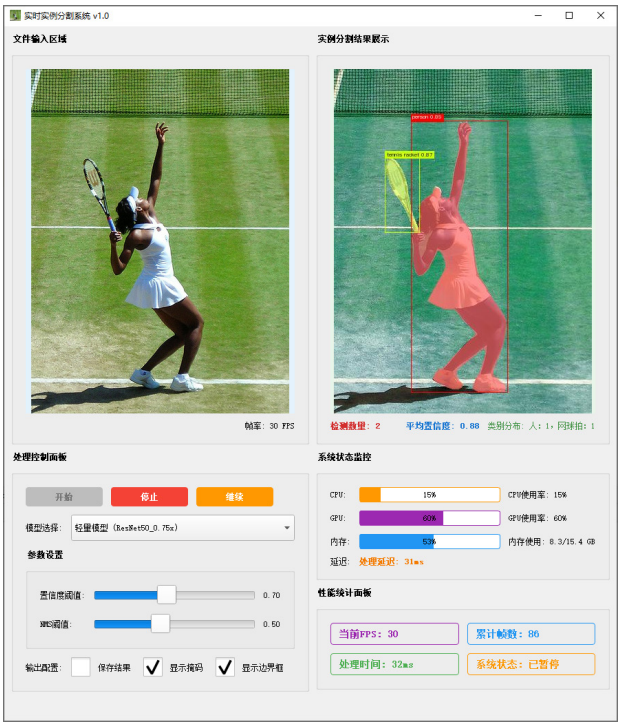

图 5 系统 UI 设计

参考文献

[1] Garcia-Garcia, Alberto, Orts-Escolano, Sergio, Oprea, Sergiu, et al. A survey on deep learning techniques for image and video semantic segmentation. Applied Soft Computing, 70: 41--65, 2018.

[2] He, Kaiming, Gkioxari, Georgia, Dollar, Piotr, et al. Mask r-cnn. //Proceedings of the IEEE international conference on computer vision, 2017.

[3] Wang, Xinlong, Zhang, Rufeng, Kong, Tao, et al. Solov2: Dynamic and fast instance segmentation. Advances in Neural information processing systems, 33: 17721--17732, 2020.

[4] 王中宇; 倪显扬; 尚振东. 利用卷积神经网络的自动驾驶场景语义分割[J]. 光学精密工程, 2019,27(11):2429-2438.

[5] 朱凌云,杨小洪.LiteRevNet:一种轻量级工业图像实例分割算法[J].重庆理工大学学报(自然科学),2024,38(12):133-140.

[6] Qiangda Shan, Hao Chen, Ziru Liu. Lightweight Instance Segmentation Algorithm Based on Knowledge Distillation. Academic Journal of Computing & Information Science (2025), Vol. 8, Issue 4: 40-48.

[7] Bodla N , Singh B , Chellappa R ,et al.Improving Object Detection With One Line of Code[J]. 2017.DOI:10.48550/arXiv.1704.04503.