Visual Recognition Homework 3: Cell Instance Segmentation

110550108

GitHub Repository: [Link]

1 Introduction

The goal of this task is **cell instance segmentation**, which requires detecting and segmenting each individual cell in microscopy images. Unlike semantic segmentation, which classifies each pixel without differentiating instances, instance segmentation distinguishes between separate objects of the same class. To solve this problem, I adopt the **Mask R-CNN** architecture proposed by He *et al.* [1], a two-stage framework that extends Faster R-CNN [2] by adding a branch for predicting object masks.

Mask R-CNN improves over previous object detectors by enabling simultaneous object detection and pixel-level instance segmentation. Its key contribution lies in its parallel mask prediction branch that operates on aligned RoI features using RoIAlign, significantly enhancing segmentation quality.

2 Method

2.1 Data Preprocessing

The dataset consists of microscopy images and corresponding cell annotations. Each image is padded and resized to a fixed input size for consistent processing. Ground truth annotations are converted into the COCO format [5], which is compatible with the MMDetection framework [7]. Binary masks are encoded in Run-Length Encoding (RLE) format for efficiency.

2.2 Model Architecture

The architecture used is the standard **Mask R-CNN** pipeline, implemented with the MMDetection toolbox [7]. The core design is detailed below:

- **Backbone: ResNet-50**
  A residual convolutional neural network [3] that extracts hierarchical image features. I use the ImageNet-pretrained ResNet-50 to leverage transfer learning for better convergence and generalization.
- **Neck: Feature Pyramid Network (FPN)**
  FPN [4] fuses multi-scale features from different backbone stages, allowing the model to detect cells at various scales—a critical property given the varying sizes of cells.
- **Region Proposal Network (RPN)**
  RPN generates object proposals using anchors of multiple sizes and aspect ratios. It learns to propose bounding boxes that likely contain objects.
- **RoI Head + RoIAlign**
  RoIAlign [1] eliminates the quantization problem in feature extraction by interpolating features at exact float locations. This leads to better alignment and accuracy in the mask head.
- **Mask Head**
  A small Fully Convolutional Network (FCN) is applied to each RoI to predict a binary mask for each instance. This branch operates in parallel with classification and box regression.

2.3 Hyperparameters

- Learning rate: 1e-4
- Optimizer: Adam with CosineAnnealing learning rate scheduler
- Batch size: 2
- Epochs: 20
- Anchor sizes: [4, 8, 16, 32, 64, 128]
- box detections per image: 200
- aspect ratios: (0.5, 1.0, 1.5, 2.0)

3 Results

The model achieves strong segmentation performance on the validation set. As shown in Figure 1, the training loss quickly drops and stabilizes after several epochs. The final validation mean Average Precision (mAP) reaches approximately **0.43**.

Key observations:

- The FPN-augmented ResNet-50 backbone provides robust multi-scale features, leading to accurate cell localization.
- The model successfully segments overlapping and densely packed cells.
- Validation mAP indicates that the model generalizes well without overfitting.

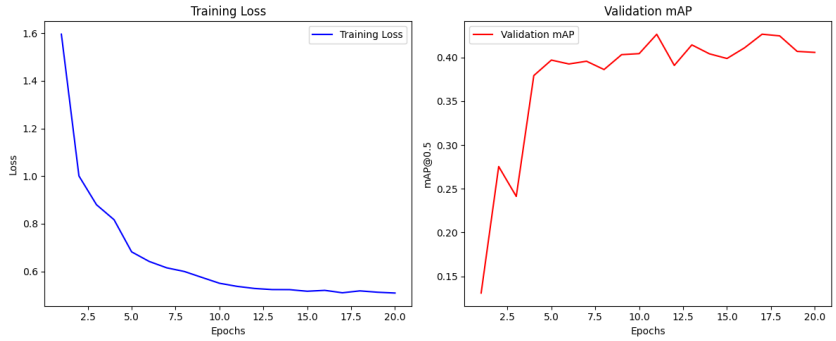

Figure 1: Training loss curve over 20 epochs.

45	110550052	1	2025-05-01 21:51	279738	110550052	0.3209
46	111550020	1	2025-04-30 00:03	278952	111550020	0.3188
47	31354nycu	1	2025-05-03 22:25	280677	313540002	0.3155
48	313551044	1	2025-05-04 17:11	281045	313551044	0.3147
49	110550108	1	2025-05-04 21:56	281186	110550108	0.3116
50	110550128	1	2025-05-02 06:30	279879	110550128	0.3106
51	313540009	1	2025-04-29 18:17	278789	313540009	0.3088
52	S0n9Yu	1	2025-05-05 12:41	281463	111550098	0.3077
53	312540013	1	2025-05-02 15:28	280099	312540013	0.3069

Figure 2: Public leaderboard score: mAP@50 = 0.3116.

4 Additional Experiment: Replacing Mask Head Loss with Dice Loss

4.1 Hypothesis

The default mask loss in Mask R-CNN is pixel-wise Binary Cross Entropy (BCE) loss, which treats each pixel independently. However, in our dataset, the cell regions are sparse compared to the background, leading to a significant foreground-background class imbalance. I hypothesize that using **Dice Loss**, which directly optimizes the overlap between predicted and ground truth masks, may yield better mask predictions by focusing on the global structure of the object and mitigating the effect of class imbalance.

4.2 Mechanism of Action

Dice Loss measures the similarity between the predicted mask and the ground truth by computing the ratio of their intersection over union at the set level. This makes the optimization more robust to class imbalance since it considers the entire object shape rather than individual pixels. While BCE may be dominated by abundant background pixels, Dice Loss assigns more balanced importance to small object areas (e.g. cells).

However, Dice Loss is non-linear and may lead to unstable gradients during early training, particularly when the predicted masks are still inaccurate. Therefore, the improvement is not guaranteed and might depend on good initialization or complementary losses.

4.3 Results and Implications

After replacing the mask head loss with Dice Loss (keeping the rest of the configuration unchanged), the model achieved a slightly improved validation mAP of **0.44**, compared to the baseline of **0.43** with BCE loss. This minor gain supports the hypothesis that Dice Loss better handles pixel imbalance in segmentation tasks involving small or sparse foreground regions.

Implications:

- Incorporating global similarity measures like Dice Loss can enhance segmentation accuracy in dense instance segmentation tasks.
- Loss function choice significantly affects mask quality when objects are small or densely packed.
- Future work may consider combining BCE and Dice (composite loss) or using Tversky Loss to further control false positives/negatives.

5 Conclusion

This project demonstrates the effectiveness of the Mask R-CNN framework in cell instance segmentation. The architecture, enhanced by FPN and pretrained backbone, provides accurate localization and segmentation of individual cells. Replacing the mask head loss with Dice Loss offers marginal improvement in handling class imbalance, reinforcing the importance of loss design in segmentation performance.

References

[1] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 2961–2969.

[2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2015.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016.

[4] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017.

[5] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 740–755.

[6] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation," in *Proc. Int. Conf. 3D Vision (3DV)*, 2016.

[7] X. Chen *et al.*, "MMDetection: Open MMLab Detection Toolbox and Benchmark," `https://github.com/open-mmlab/mmdetection`, accessed Apr. 2025.