Dishware Instance Segmentation

Bido Mohamed
1007732483, abdalla.mohamed@mail.utoronto.ca
Google Colab

Introduction

Dish cleaning is a repetitive and time-consuming activity in large scale environments such as cafeterias and commercial kitchens. Automating this workflow would reduce physical strain on staff and support broader efforts in service robotics by enabling systems to identify and sort dishware efficiently. Instance segmentation of dishware represents a key perception requirement for such systems, since it enables the detection and localization of individual items including plates, bowls, cups, and utensils. Accurate segmentation supports downstream tasks such as grasp planning, sorting by category, and placement into dish racks.

The project addresses this problem by training an instance segmentation model to predict bounding boxes, class labels, and pixel level masks for six dishware categories. Learning based approaches have achieved strong performance on related detection tasks due to their ability to extract hierarchical features that remain stable under occlusion, clutter, and variation in lighting. The goal of the project is to evaluate whether a modified Faster R-CNN model can outperform a widely used baseline model on dishware segmentation, and to analyze the strengths and limitations of the system across different object types.

Illustration

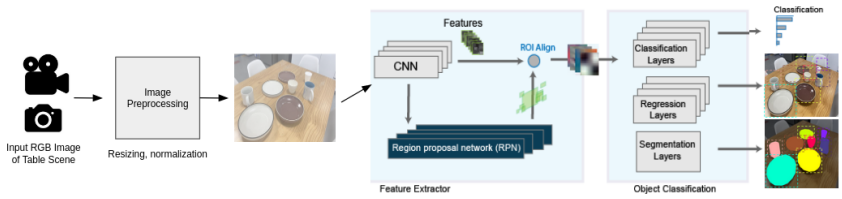

Figure 1: Proposed dishware instance segmentation pipeline.

Background & Related Work

Instance segmentation has been a central problem in computer vision, with recent progress driven by advances in convolutional architectures and large scale annotated datasets. Methods such as Faster R CNN introduced a unified approach to region proposal generation and object classification that significantly improved detection accuracy and runtime performanceRen et al. (2016). Subsequent extensions, including Mask R-CNN, added a parallel branch for predicting pixel level masks while maintaining strong bounding box localization He et al. (2018). These architectures have become standard baselines for segmentation tasks due to their modular design and competitive accuracy. Large scale datasets have played a critical role in enabling these models. The COCO dataset provides diverse annotated scenes and remains a benchmark for both bounding box detection and instance segmentationLin et al. (2015). LVIS expands object diversity and includes rare categories that require strong generalization capabilities Gupta et al. (2019). In parallel, single stage models such as YOLOv3 and later variants introduced real time detection architectures that balance accuracy and efficiency Redmon and Farhadi (2018). Region based approaches tend to perform better on small, thin, or partially occluded items, which is directly relevant for dishware since utensils and stacked

plates present challenging geometric and occlusion patterns. Similar perception challenges arise in robotic dish handling and industrial bin picking systems, where instance level detection under clutter and occlusion is essential for reliable manipulation Wada et al. (2020). These application domains further motivate the need for robust segmentation models and provide context for the evaluation carried out in this project.

Data Processing

The LVIS v1 validation annotations and their linked COCO val2017 images served as the initial data source, and only the six dishware categories used in this project were retained. After filtering and preprocessing, the resulting dataset was split into 245 training images, 45 validation images, and 55 test images. A representative processed training sample is shown in Figure 2.

Processing Step	Description
Mask parsing	Conversion of polygon or RLE annotations into binary masks.
Bounding box validation	Removal of annotations with zero height or width.
Class balancing	Maximum of sixty images retained for each class.
Normalization	Pixel intensities scaled to the zero to one range.
Augmentation	Random flip, random crop and brightness jitter.
Splitting	Train, validation and test split of 70, 15 and 15 percent.

Table 1: Summary of data processing steps.

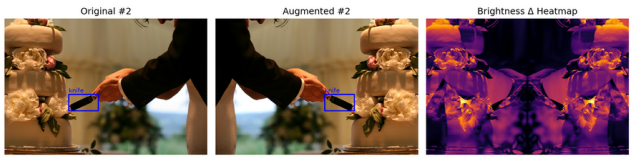

Figure 2: Example of cleaned and augmented training sample.

Architecture

The model used is a Faster R-CNN detector with a ResNet 50 backbone and Feature Pyramid Network, a well established two stage architecture for instance level recognition Ren et al. (2016). This model was selected because two stage detectors achieve strong performance on small and medium objects and are less prone to missing thin structures compared to one stage detectors Lin et al. (2018). In adapting the architecture to the dishware segmentation task, targeted modifications were introduced. The Region Proposal Network anchor configuration was set to sizes 16, 32, 64, 128, 256, 512 pixels with aspect ratios 0.5, 1.0, 2.0, 4.0 so that both small utensils and elongated shapes were covered by the proposal distribution. ROIAlign output resolution was set to seven by seven to preserve spatial details relevant for separating fine-grained utensil classes. Proposal sampling used 512 proposals per image with a foreground to background ratio of 1 to 3 using IoU at least 0.5 as foreground. COCO pretrained backbone weights were used for initialization and the network was fine-tuned end to end on the dishware dataset.

Baseline Model

The baseline model was YOLOv8n segmentation, a strong reference point for how a lightweight detector performs on dishware scenes, and to isolate the benefit of a two stage region based model under the same dataset and label set Yaseen (2024). The baseline is also commonly used in practical pipelines because it offers stable training behaviour with minimal configuration and provides an accuracy versus speed tradeoff for real time settings.

For reproducibility, the COCO pretrained YOLOv8n segmentation weights were used, and the model was fine-tuned on the dishware dataset for 10 epochs without modifying the architecture or loss functions. Inference was performed at an input resolution of 640 by 640, predictions were filtered

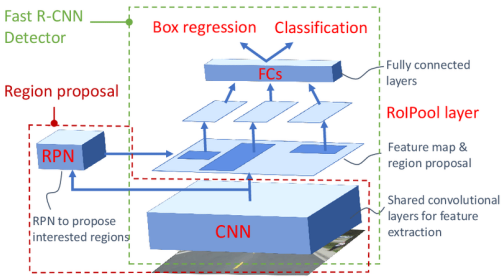

Figure 3: Faster R-CNN architecture overview.

using a confidence threshold of 0.25, and outputs were restricted to the six dishware categories to ensure that evaluation reflected the target task rather than the full COCO label space.

Quantitative Results

Quantitative performance was evaluated using bounding box Average Precision (AP). AP at IoU ≥ 0.50 (AP50), IoU ≥ 0.75 (AP75), and mean AP over IoU thresholds from 0.50 to 0.95 (mAP) were computed on the train and validation splits (Table 2). Across all splits, higher AP50 and mAP values were obtained by Faster R-CNN compared to YOLOv8n, indicating stronger overall detection performance. Both models showed reduced AP75, reflecting the increased difficulty of precise localization, although the relative ranking remained consistent. Per-class AP50 values (Figure 4)indicated the largest gains on plates and the utensil classes, while bowls and cups exhibited similar performance for both models and higher scores than utensils.

Split	AP50		AP75		mAP (0.50:0.95)	
	YOLOv8n	Faster R CNN	YOLOv8n	Faster R CNN	YOLOv8n	Faster R CNN
Train	0.026072	0.429089	0.022902	0.257167	0.020046	0.246515
Val	0.143734	0.269702	0.140345	0.146993	0.120574	0.134144

Table 2: Train and validation average precision for YOLOv8n and Faster R CNN.

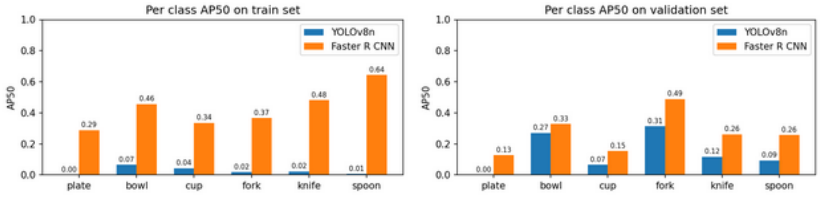

Figure 4: Per-class AP50 comparison on train and validation splits.

Qualitative Results

Figure 5 presents 3 representative validation images selected to highlight model behaviour specifically on dishware-oriented outputs rather than human-centric content. Images were therefore chosen only when dishware objects were clearly visible, ensuring that qualitative patterns reflected the detector's handling of plates, bowls, cups, and utensils rather than incidental features. The good example (recall = 1.00) shows complete detection of all dishware with tight, well-aligned boxes. The interesting example (recall = 0.67) displays partial success, where large objects are accurately detected while small or boundary items are missed or merged. The bad example (recall = 0.50) illustrates common failure modes, including duplicated boxes, confusion between similar categories, and missed thin utensils.

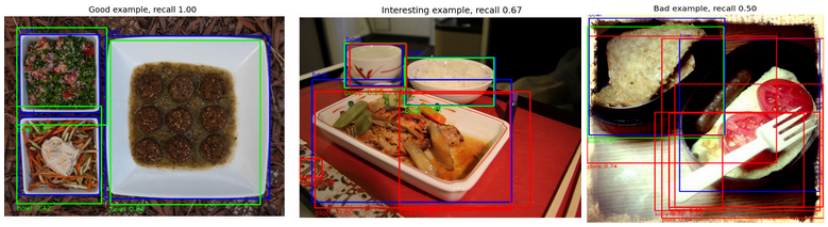

Figure 5: Qualitative detection results across success and failure cases.

Model Performance on New Data

A set of six dishware-focused images was collected from outside the COCO/LVIS dataset to assess model generalization under distribution shift. The images were chosen to cover diverse real-world settings, including formal dining arrangements, patterned table surfaces, cluttered kitchen sinks, countertop assortments, and cafeteria environments. To ensure that none of the images had been used during training or tuning, filename matching against the COCO index and SHA-256 hash comparisons against all cached COCO images were performed, confirming that the new samples were entirely unseen.

When evaluated on this dataset, the model produced an average of 13.17 detections per image. Detection scores remained concentrated above 0.60, and many exceeded 0.80, indicating stable confidence across the new scenes. Representative outputs illustrate consistent detection of major dishware items such as plates, bowls, and cups, while more visually complex scenes produced a wider spread of prediction scores.

Discussion

The results indicate that the model performs strongly for the primary dishware categories. Plates, bowls, and cups achieved the highest AP values and were detected reliably across both the validation images and the unseen scenes. This behaviour is consistent with their larger size, clearer geometric structure, and stronger representation in the training data. It was also observed that the model maintained stable confidence even when evaluated on visually distinct environments such as patterned tablecloths, cafeteria trays, and overhead dining scenes.

The weakest performance occurred for utensils, which showed lower AP values and more variable detections on the new data. This can be attributed to the small scale of these objects, limited anchor coverage, and the reduced number of utensil examples available during training. Overlapping configurations and cluttered backgrounds also made utensils more susceptible to suppression during non-maximum filtering. These observations suggest several clear improvement paths. Increasing the quantity and diversity of utensil examples, incorporating training strategies that emphasize small objects, or adopting models with higher-resolution feature maps would likely enhance detection of thin and partially occluded items.

Ethical Considerations

Ethical considerations were addressed by ensuring that the project focused exclusively on non-sensitive household objects and did not attempt to infer any personal or demographic information about individuals visible in the COCO or LVIS images. It was acknowledged that the training data do not equally represent cultural dishware styles, which may limit performance on items common in underrepresented regions and could introduce unintended bias in real-world applications.The model's limitations in cluttered or utensil-dense scenes further emphasize the need for careful integration to avoid operational errors in practical systems.

References

Gupta, A., Dollár, P., and Girshick, R. (2019). Lvis: A dataset for large vocabulary instance segmentation.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2018). Mask r-cnn.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2018). Focal loss for dense object detection.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollár, P. (2015). Microsoft coco: Common objects in context.

Redmon, J. and Farhadi, A. (2018). Yolov3: An incremental improvement.

Ren, S., He, K., Girshick, R., and Sun, J. (2016). Faster r-cnn: Towards real-time object detection with region proposal networks.

Wada, K., Okada, K., and Inaba, M. (2020). Joint learning of instance and semantic segmentation for robotic pick-and-place with heavy occlusions in clutter.

Yaseen, M. (2024). What is yolov8: An in-depth exploration of the internal features of the next-generation object detector.