Perception Engineer Take-Home: Bowl Detector Report

1. Data Analysis

- **Format & Quality:** Original images and YOLO-format `.txt` labels (class_id, x_center, y_center, width, height).
- **Issues Identified:**
  - Inconsistent class IDs and missing labels on occluded/truncated bowls.
  - Near-duplicate frames leading to potential data leakage across splits.
  - Bounding boxes misaligned on rotated bowls.
- **Remediation:**
  - Relabeled all 25 train + 5 test images with `0=empty` and `1=full`.
  - Used Makesense.ai for consistent annotation and corrected box coordinates on rotated trays.
  - Ensured no filename overlap in cross-validation folds (K-Fold, K=4).
- **Scaling Strategy:**
  - Collect diverse camera angles and lighting via automated rig in a commercial kitchen.
  - Crowdsource annotation with clear guidelines, review consensus, and use semi-automated tools (e.g. Label Studio).

2. Model Training

- **Architecture:** Faster R-CNN with ResNet-50 backbone + 5-level FPN.
  - Pretrained on COCO for feature generalization.
  - FPN supports multi-scale detection (small vs. large bowls).
  - Two-stage ROI head yields precise localization vs. single-stage detectors.
- **Training Details:**
  - Input size: 480×640, batch size 4 (GPU RTX 4070).
  - Augmentations: random horizontal flip (p=0.5), color jitter (brightness/contrast/saturation), to improve robustness.
  - Loss: classification + Smooth L1 + auxiliary GIoU penalty (weight 2.0) for tighter boxes.
  - Optimizer: SGD (LR 1e-3 → 1e-4 at epoch 10), 30 epochs, StepLR decay.
  - Cross-validation: 4-fold stratified on empty/full ratio.

3. Evaluation

- **Metrics:**
  - mAP@0.5 and mAP@[0.5:0.95] via torchmetrics on val folds and held-out test set.
- **Results (Best Model):**
  - **Cross-val (4-fold):** mAP@0.5=0.464±0.024, mAP@[.5:.95]=0.334±0.019.
  - **Held-out Test:** mAP@0.5=0.3828, mAP@[.5:.95]=0.2833.
- **Analysis:**

    - Augmentation + GIoU marginally improved stability (lower σ) and localization precision.
    - Fold-4 anomaly (mAP@0.5≈1.0) traced to near-duplicate backgrounds—mitigated by dataset split correction.
- **Limitations & Improvements:**
    - Small sample size → high variance.
    - Uniform backgrounds limit generalization to new kitchens.
    - Future work: anchor generator tuning, Cascade-RCNN, one-stage detectors (YOLOv5) with Mosaic, CIoU, and focal loss.

4. Extension

- **Orientation Prediction:**
    - Add a small regression head on ROI features to predict bowl rotation (angle), trained with smooth L1 on annotated angles.
    - Alternatively, discretize orientation into bins and use classification loss.
- **Ingredient Recognition:**
    - Crop detected bowl regions and pass through a lightweight CNN classifier (e.g. MobileNet) to label contents.
    - Explore multi-task learning: joint detection + semantic segmentation in a unified Panoptic-Mask-RCNN.

5. Deliverables

1. **Source Code:** Colab notebook (`.ipynb`) with preprocessing, training, evaluation, and export cells.
2. **Trained Artifacts:**
    - PyTorch weights (`fasterrcnn_bowl_best.pth`).
    - ONNX model (`bowl_detector.onnx`) for on-device inference.
3. **Report:** This concise document summarizing methodology, results, and future directions.