Pose Estimation for Robotic Assembly from Instruction Manuals
Baseline results
Team 24

Aldrin Inbaraj Augustin Ponraj
CSE 598: Perception in Robotics
ASU ID: 1226200393
aaugus11@asu.edu

April 4, 2025

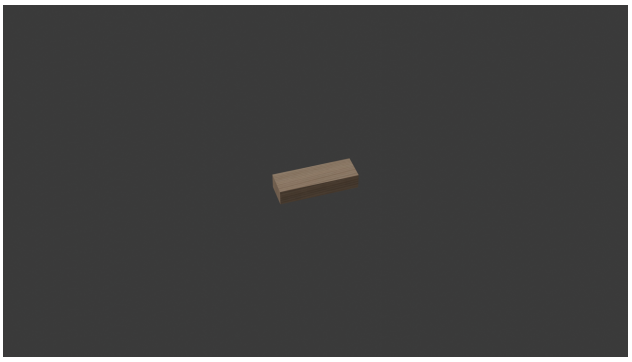
1 Experimental Setup

A 3D model of a Jenga block with its texture was imported into Blender. A Python script was developed to rotate the camera around the block in 360° increments along both the azimuth (θ) and elevation (ϕ). At each position, the script captures an RGB image and saves it as `rgb.png`, along with a label file (`.txt`) containing eight 2D coordinates that describe the object's bounding polygon in the image plane.

An example of the 8 coordinates output by the Blender script is:

```
Jenga_Block: [(1063, 507), (1064, 470), (1092, 521), (1090, 559),
              (829, 557), (827, 520),  (850, 574),  (852, 611)]
```

Below is a sample RGB image produced by this process:



These coordinates are then converted into the YOLOv11n-compatible format, which requires normalized corner coordinates of the Oriented Bounding Box (OBB). Each line of this format starts with a class label (e.g., 0), followed by four (x, y) coordinate pairs in normalized image coordinates. The format ensures that each oriented box is represented consistently for training.

2 YOLOv11n Training

After preparing the dataset, which consists of RGB images and corresponding OBB labels, a YOLOv11n-OBB model is trained using the Ultralytics YOLO API. The training configuration includes both geometric and photometric augmentations.

Geometric augmentations applied include small random rotations, translations, scaling, shearing, perspective changes, horizontal and vertical flips, as well as mosaic, mixup, and copy-paste techniques. These augmentations help in simulating a variety of image conditions and orientations.

Photometric augmentations modify hue, saturation, and brightness values to enhance robustness under different lighting conditions.

The training is carried out for 100 epochs on 640×640 images. The final model weights are saved as `best.pt`, which are then used for inference on the test dataset.

3 Baseline Calculation

To evaluate the model's performance, 10 images from the test folder are selected. The trained model is loaded and used to run inference on these test images. Across the 10 detections, the model achieves an average confidence score of approximately 0.95, suggesting high certainty in correctly locating the Jenga block.

Below is an example detection with a confidence score of 0.94 from the image `155_pred.png`:

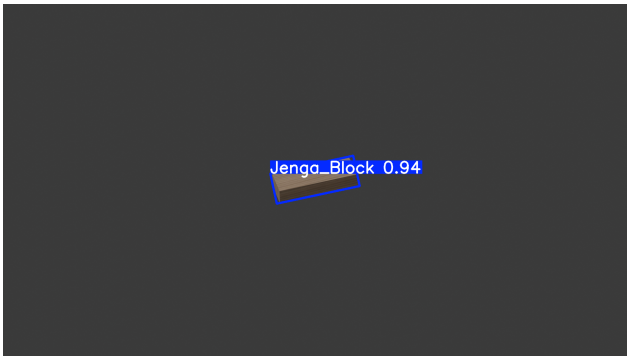


Figure 1: *
Example prediction showing Jenga block detection with confidence 0.94

Each prediction is stored in YOLOv11n's 8-point OBB format. These predicted bounding boxes are compared against ground-truth bounding boxes generated using `cv2.minAreaRect`. The comparison involves calculating the following metrics:

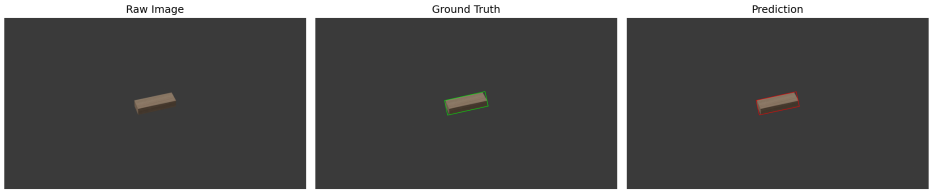
- Intersection over Union (IoU)
- Centroid distance
- Orientation angle difference

The IoU is computed using polygon-based intersection and union areas, while the centroid distance is calculated from the mean of the corner points of each box. The angle difference metric evaluates the deviation in rotation between predicted and ground-truth boxes.

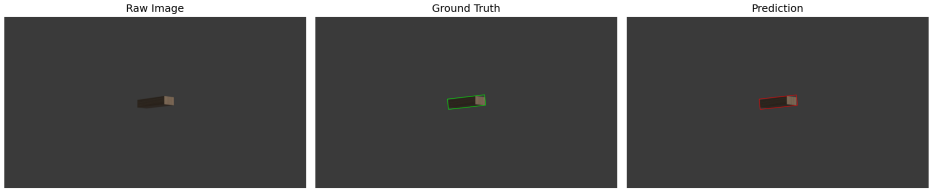
For each of the 10 test images, a four-column comparison figure is created to illustrate:

1. The raw input image,
2. The ground-truth bounding box,
3. The predicted bounding box,
4. A summary text block showing IoU, centroid distance, and angle difference.

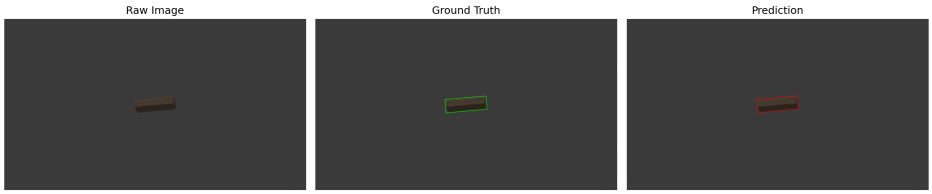
All ten results are stacked vertically in one figure, as shown in Figure 2. No individual captions are included to maintain visual continuity.



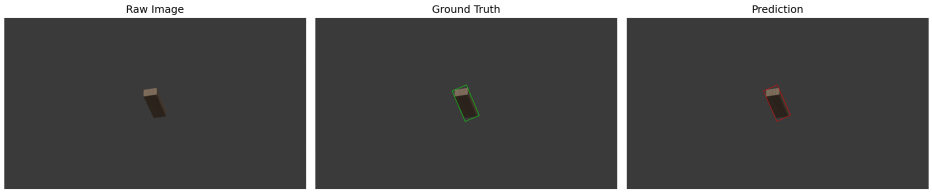
Box #1:
IoU: 0.9786
Dist(px): 0.83
AngleDiff(deg): 0.23



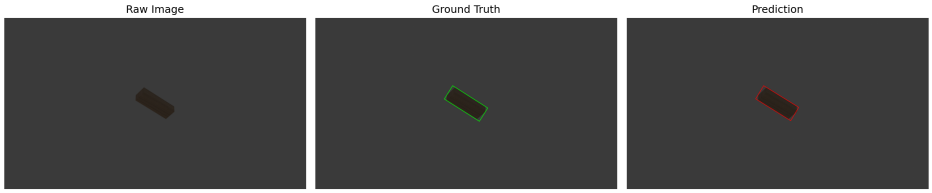
Box #1:
IoU: 0.9631
Dist(px): 1.73
AngleDiff(deg): 0.38



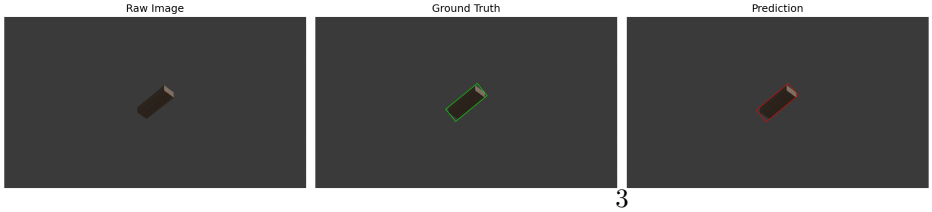
Box #1:
IoU: 0.9636
Dist(px): 1.61
AngleDiff(deg): 0.14



Box #1:
IoU: 0.9579
Dist(px): 1.20
AngleDiff(deg): 0.89



Box #1:
IoU: 0.9599
Dist(px): 2.87
AngleDiff(deg): 0.63



Box #1:
IoU: 0.9443
Dist(px): 4.15
AngleDiff(deg): 0.60

4 Conclusion

The baseline experiments demonstrate that the YOLOv11n-OBB model can accurately detect the Jenga block, consistently achieving high confidence scores around 0.95 and producing bounding boxes that closely align with the ground-truth.

The evaluated metrics—IoU, centroid distance, and angle difference—confirm that the model is effective in both spatial localization and orientation estimation.

For future work, I plan to extend the evaluation to a larger dataset with multiple object classes, compute mean Average Precision (mAP) at various IoU thresholds, and explore more sophisticated augmentation strategies to further enhance robustness in complex and diverse scenes.