Object Detection and Depth Estimation

Gokul Edakke Puram 22 541 618 MM Parth Vipul Patel 11 141 617 MM

June 4, 2024

Acronyms

FOV Field of View

IoU Intersection over Union

KITTI Karlsruhe Institute of Technology and Toyota Technological Institute

YOLO You Only Look Once

Abstract

The distance between the camera of ego vehicle and the vehicles detected by You Only Look Once (YOLO) object detector (limited to class: cars) in each of the selected scene from the Karlsruhe Institute of Technology and Toyota Technological Institute (KITTI) dataset is calculated using only the camera information and compared to the distances given in the ground truth. The depth estimation have been performed and evaluated on all the 20 different scenes that were provided for evaluation.

1 Introduction

In this task, the camera calibration matrix (intrinsic matrix) and ground-truth from KITTI dataset that contains the bounding box information of the vehicles and their estimated distance to the camera were provided for each selected scene.

The vehicles in the scenes were detected by using YOLOv8x (see 2.1) and the detections were filtered out selecting only the vehicle detections whose information were present in the labels (see 2.2).

Distance to each vehicles were calculated after bringing the point selected on the bounding box to 3D world coordinate and evaluated by plotted against the ground truth (see 2.4). The images of the 20 evaluated scenes are included in Section 3 of this report.

2 Goals of Task

2.1 Detection of Vehicles using YOLO

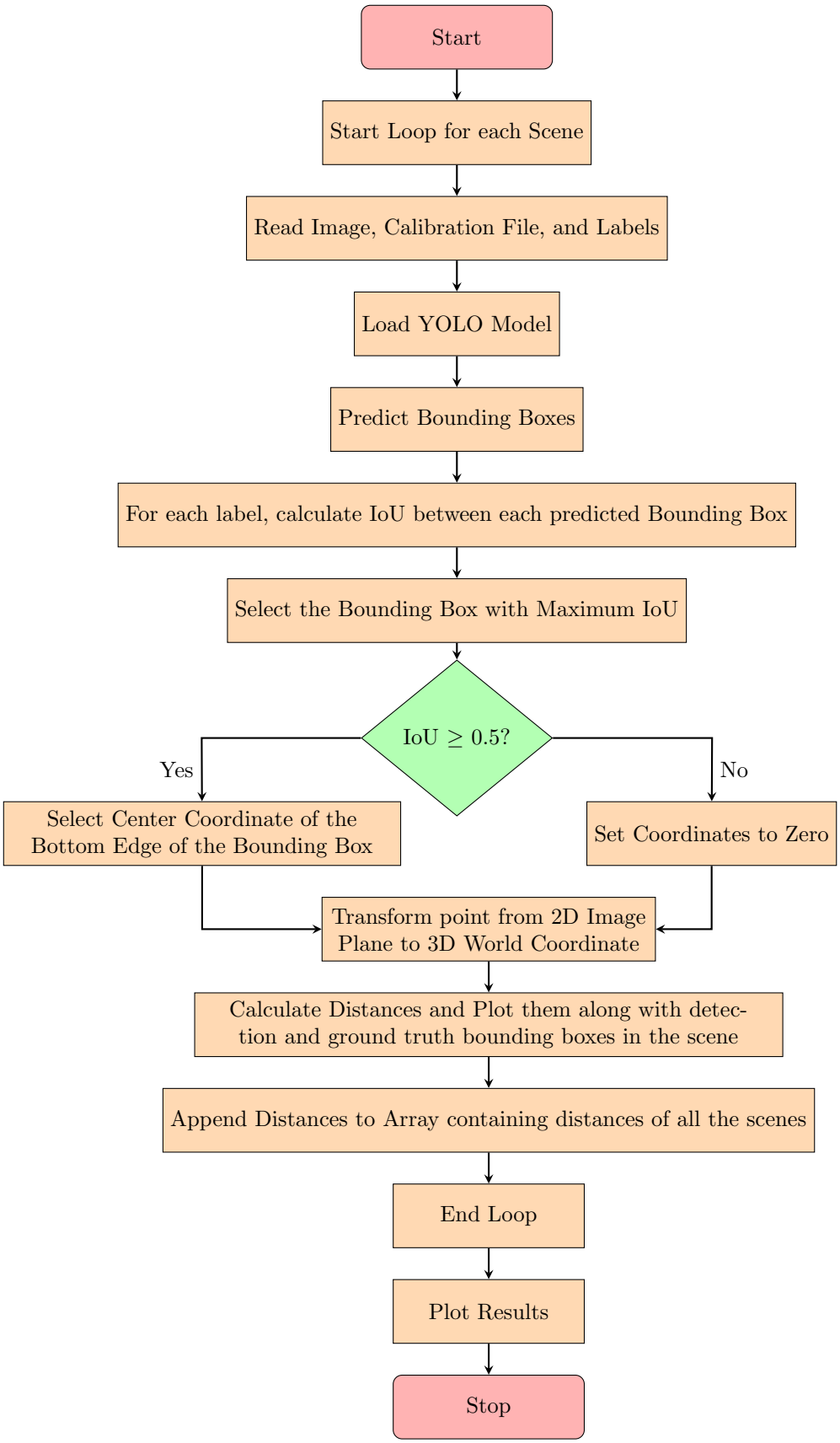
The version 8x of YOLO object detector (Jocher et al.) was used to detect the vehicles in each of the 20 scenes provided. The classes array [2,5,7] corresponding to vehicles (cars, bus, truck) were provided as argument to the predict function of YOLO.

2.2 Matching of Detections from YOLO to the Ground Truth Objects

The Intersection over Union (IoU) between all the detected bounding boxes and the ground truth bounding boxes were calculated and the detection with the maximum IoU with a threshold value of 0.5 was selected as the detected True Positive bounding box for the object in the ground truth. This approach was effective in filtering out all the False Positives and the vehicles that were absent in the ground truth.

There were some False Negatives (9 in total to be specific), the distance of these were mapped to 0 when plotting the estimated distance against the ground truth and can be seen in Figure 2.

2.3 Flowchart of the Program Algorithm



2.4 Calculation of Distance Using Camera Information

The point situated at center of the bottom edge of the bounding box detected by YOLO object detector was selected and transformed into the 3D coordinate system from the image coordinate system using the camera's intrinsic matrix, thereby establishing a direction vector for a line extending from the camera in 3D space. The transformation is done by performing a dot product between the selected point and the inverse of the camera intrinsic matrix, mapping the 2D point from the image plane into the 3D coordinate. The z coordinate is assumed as 1, augmenting it to get a 3x1 vector in order to perform the dot product operation with the intrinsic matrix (3x3 matrix) and solving for the constant t (refer to line equation).

Line equation:

$$\mathbf{r}(t) = \mathbf{r}_0 + t\mathbf{d}$$

where $\mathbf{r}(t) = (x', y', z')$ is the position vector of any point on the line. $\mathbf{r}_0 = (0, 0, 0)$ is the position vector of a specific point on the line, taken here as the position of the camera (origin). $\mathbf{d} = (x, y, z)$ is the direction vector which is the point selected on the bounding box (center point of the bottom edge), and t is a scalar parameter.

The driving plane is assumed to be perpendicular to the image plane and parallel to the horizon, at a distance of 1.65m from the camera.

Plane equation:

$$y = 1.65$$

The line equation and the plane equation is solved to get the constant t, which when multiplied with x and z gives us the point of intersection between the line and the plane (see Figure 1).

The distance to the vehicle from the camera is estimated by calculating the distance to this point.

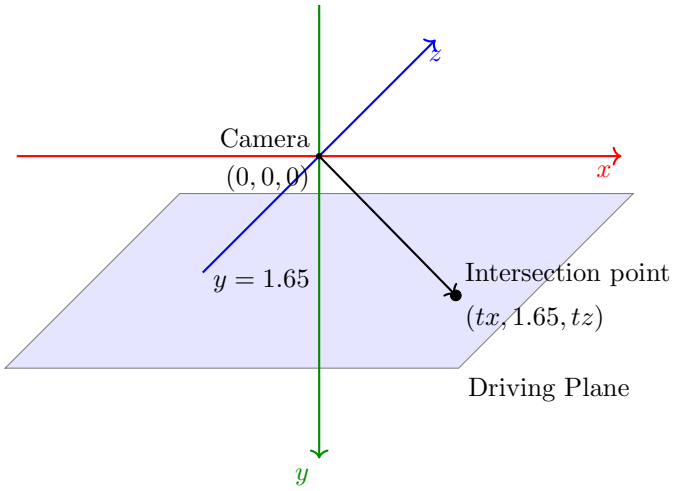


Figure 1: Graphical representation of method used for distance estimation from camera information

2.5 Evaluation of Estimated Distances with Ground Truth Distances

In the following graph, Figure 2, we can notice that most of the estimated distances of the vehicles that are under 40 meters lie close to the ideal line (ideal is when ground truth distance = estimated distance), meaning that the error is low, whereas for vehicles that are present over 40 meters have a larger error and vary further away from the ideal line. The vehicles that are close to the camera also varies a lot from the ground truth, because the vehicles are outside the Field of View (FOV) of the camera and YOLO can only generate a bounding box around object that is actually visible in the image. This causes a lower limit to the distances that can be estimated by this method (see 3.1).

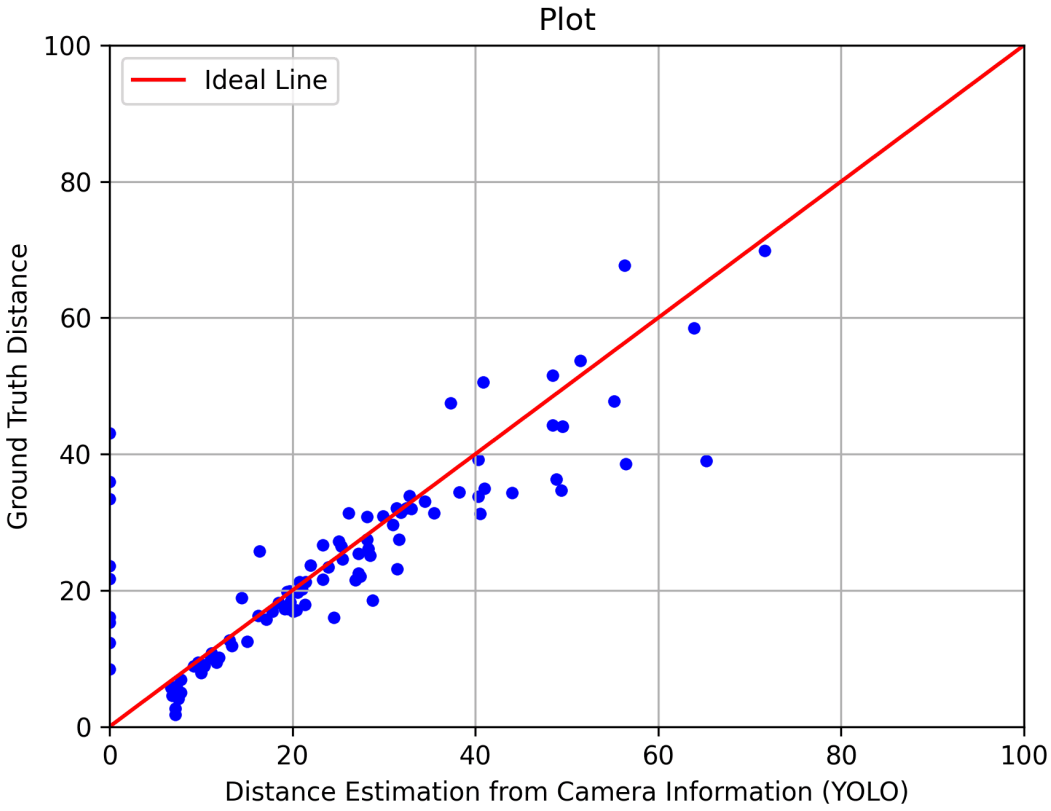


Figure 2: Plot Between Estimated Distance against Ground Truth Distance

The error in estimating distances by this approach for vehicles at further distance (over 40m in this case) is because the intersection of the assumed driving plane and the line constructed by the point that are at a larger distance from the camera are at a distance larger than the points close by, Figure 3 depicts this scenario visually.

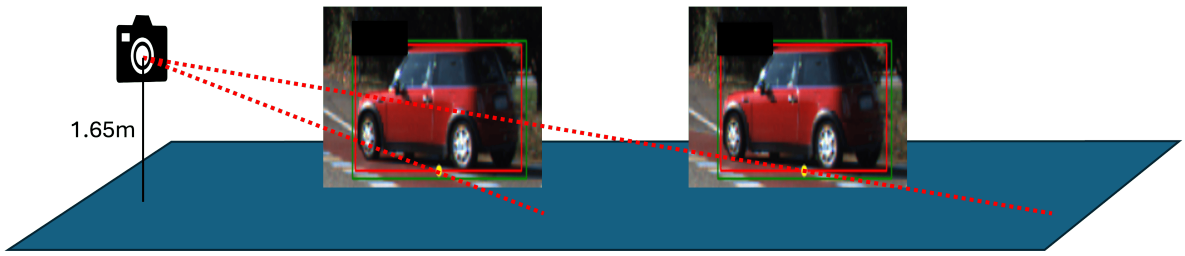


Figure 3: Difference between the distance estimated when the point is closer and further away from camera

3 Images of the Evaluated Scenes

3.1 Scenes with bad Estimations

In Scene 006310, there is an estimated distance of 328.61m for a car whose ground truth distance is 67.33m refer Figure 4. In this scene, we can see that the vehicle is far away from the camera and is on a road which is elevated and the gradient of the road is increasing. The point selected on the detected bounding box is closer to the horizon, this causes the intersection point between the line constructed by this point as the direction vector and the assumed driving plane to be at a point very much further away than the actual position of the vehicle.

This is one of the flaws of our approach where we assume the actual driving plane to be straight, parallel to the horizon and offset by a distance of 1.65m below the camera whereas in this particular scene it clearly is not. Similar scenario with scene 006098 (see Figure 7)

Figure 4: Image of Scene 006310. Zoomed in for better clarity between crowded cars(bottom). Red - YOLO detection, Green - Ground truth, Yellow dot - Point selected for distance estimation

In scene 006097, there are two cars that are very close to the ego vehicle and not completely visible in the FOV of the camera. Due to this, we cannot accurately determine the distance only from camera information. YOLO can only generate the bounding box around the object that are visible in the image. In this scenario, we can clearly see that the cars with bad estimation are too close and partially outside the FOV of the camera (see Figure 5).

This can be seen in scene 006329 also, where vehicle on the left is partially out of the FOV of the camera and is having a bad estimation of distance. The van on the right also have a bad estimation even though it is close to the ego vehicle for a good estimate, but it is parked outside the driving plane further away from the curb and is at a height, due to which the estimation is not accurate. (see 6).

Figure 5: Image of Scene 006097. Red - YOLO detection, Green - Ground truth, Yellow dot - Point selected for distance estimation. (All units in meters)

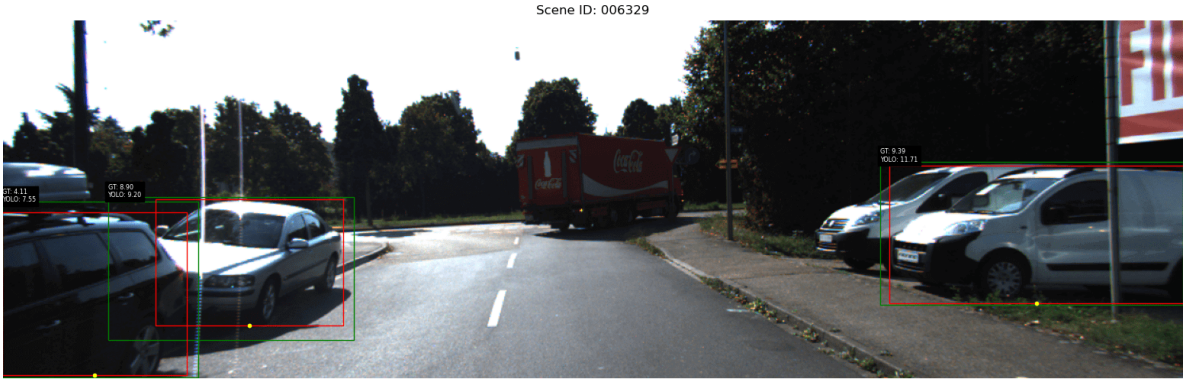


Figure 6: Image of Scene 006329. Red - YOLO detection, Green - Ground truth, Yellow dot - Point selected for distance estimation. (All units in meters)

Figure 7: Image of Scene 006098. Red - YOLO detection, Green - Ground truth, Yellow dot - Point selected for distance estimation. (All units in meters)

3.2 Crowded Scenes with False Negatives

The scenes where there are false negatives are included in this section. YOLO failed to detect them accurately either because they were under the threshold value of 0.5 or occluded in the scene by another object. There were two instances where the vehicles were very close to the vehicle, and the vehicle lying partially outside the FOV of the camera, resulting in bad estimation (see Figure 8, 11, 12). There are also instances where the distance estimations had huge errors when the cars were far away from the camera in scene 006312 and scene 006048 (see Figure 9, 12).

Figure 8: Image of Scene 006291(above). Zoomed in for better clarity between crowded cars(bottom-left and bottom-right). Red - YOLO detection, Green - Ground truth, Yellow dot - Point selected for distance estimation. (All units in meters)

Figure 9: Image of Scene 006312(above). Zoomed in for better clarity between crowded cars(bottom). Red - YOLO detection, Green - Ground truth, Yellow dot - Point selected for distance estimation. (All units in meters)

Figure 10: Image of Scene 006037(above). Zoomed in for better clarity between crowded cars(bottom). Red - YOLO detection, Green - Ground truth, Yellow dot - Point selected for distance estimation. (All units in meters)

Figure 11: Image of Scene 006211(above). Zoomed in for better clarity between crowded cars(bottom). Red - YOLO detection, Green - Ground truth, Yellow dot - Point selected for distance estimation. (All units in meters)

Figure 12: Image of Scene 006048. Red - YOLO detection, Green - Ground truth, Yellow dot - Point selected for distance estimation. (All units in meters)

3.3 Scenes with Good Estimations

This section contains all the scenes where all cars have been detected and have low error between the estimates and the ground truth. Almost all the cars in these scenes lie in a driving plane almost parallel to the horizon with no inclination. The scenes where there are no cars present are also included in this section.

In Figure 13, we have two cars, the red one on the right have a very good distance estimation because it is near to the camera and the driving plane is straight and parallel to the assumed driving plane. The car on the left have a bad estimation with almost 10m off the actual distance, this is because the car is partially occluded by the divider and YOLO could not detect the whole vehicle making the bottom edge of the bounding box at a much higher location on the vehicle.

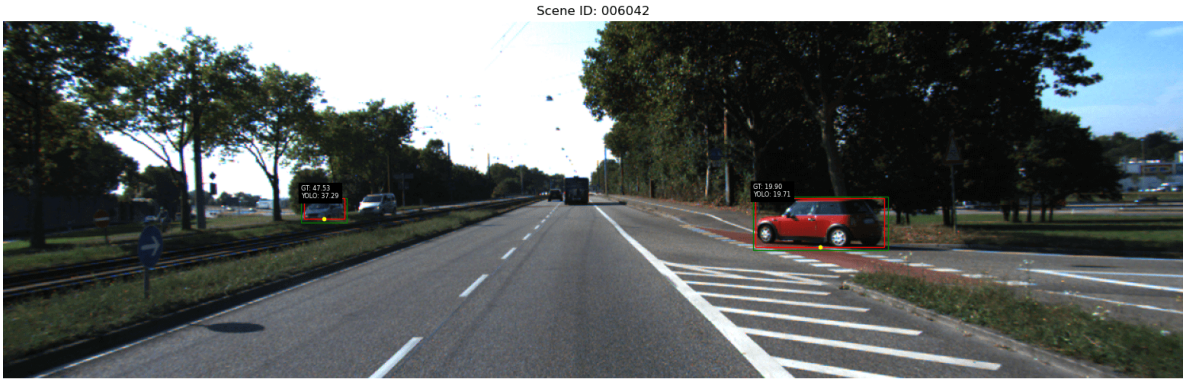


Figure 13: Image of Scene 006042. Red - YOLO detection, Green - Ground truth, Yellow dot - Point selected for distance estimation. (All units in meters)

Figure 14: Image of Scene 006067(above). Red - YOLO detection, Green - Ground truth, Yellow dot - Point selected for distance estimation. (All units in meters)

In scene 006054, there is one vehicle close to the ego vehicle and outside the FOV of the camera, giving a bad estimate, but all the other vehicles, even though were crowded together, gave a good estimate due to which this scene was classified under scenes with good estimation (see Figure 22).

In scene 006121 and scene 006130, there were no vehicles present in the ground truth to be detected (see Figure 16, 17).

Figure 15: Image of Scene 006374. Red - YOLO detection, Green - Ground truth, Yellow dot - Point selected for distance estimation. (All units in meters)

Figure 16: Image of Scene 006121.

Figure 17: Image of Scene 006130.

Figure 18: Image of Scene 006206. Red - YOLO detection, Green - Ground truth, Yellow dot - Point selected for distance estimation. (All units in meters)

Figure 19: Image of Scene 006227. Red - YOLO detection, Green - Ground truth, Yellow dot - Point selected for distance estimation. (All units in meters)

Figure 20: Image of Scene 006315. Red - YOLO detection, Green - Ground truth, Yellow dot - Point selected for distance estimation. (All units in meters)

Figure 21: Image of Scene 006253(above). Zoomed in for better clarity between crowded cars(bottom). Red - YOLO detection, Green - Ground truth, Yellow dot - Point selected for distance estimation. (All units in meters)

Figure 22: Image of Scene 006054. Red - YOLO detection, Green - Ground truth, Yellow dot - Point selected for distance estimation. (All units in meters)

In scene 006059, we have good estimations of distance except for cars that are lying on a different road next to the rail tracks see Figure 23. The cars on the same road as the ego vehicle have good estimations as the plane equation is a good assumption to the driving plane, whereas the road next to the tracks seems to be at a different height compared to the road on which the ego vehicle is present, resulting is wrong estimations.

Figure 23: Image of Scene 006059. Red - YOLO detection, Green - Ground truth, Yellow dot - Point selected for distance estimation. (All units in meters)

4 Conclusion

The distances estimated when the vehicles are relatively close to the camera and when lying on a flat driving plane with minimal or no inclination were very close to the ground truth distances. Whereas in scenes where the driving plane was inclined and vehicles far away from the camera had very large errors between the estimated distance and the ground truth.

There were scenes where YOLO could not detect the vehicles (False Negatives) and did not provide us with a distance estimation. This was in scenes where the vehicles were too crowded together.

There is a lower limit to the distance calculated by the approach used in this task, as the vehicles too close to the camera had big errors too. This was mainly due to the fact that the complete car is not present in the image and YOLO only generates the bounding box around objects that are actually present in the image.

From these 20 scenes, we can conclude that this method to estimate the distance performed well with low error as long as the vehicle lies on a flat driving plane which is parallel to the horizon and is not too far away (within 40m) and is not too close (not under 7m).

Works Cited

Jocher, Glenn, et al. *Ultralytics YOLO*. Version 8.0.0, Jan. 2023, github.com/ultralytics/ultralytics.