High Fidelity Human Behavior Modeling and Prediction

Lixuan Luo; Jianbo Shi, PhD
University of Pennsylvania (SEAS 2025); Department of Computer & Information Science

PURM 2022

ABSTRACT

- We are attempting to apply AI to the mental model level: recognizing intention, personality, and moments of confusion, etc.
- We are studying this subject by constructing an escape-room experiment. Subjects follow clues placed around the room in attempts to either mentally or physically assemble objects.
- We record from two egocentric cameras: a GoPro and gaze-tracking Tobii glasses.
- Created a detailed 3D map of the room using a Matterport Scanner.
- The recorded information is then undistorted and reconstructed into the 3D scene.
- An one-stage, real-time object detection and recognition model is implemented.

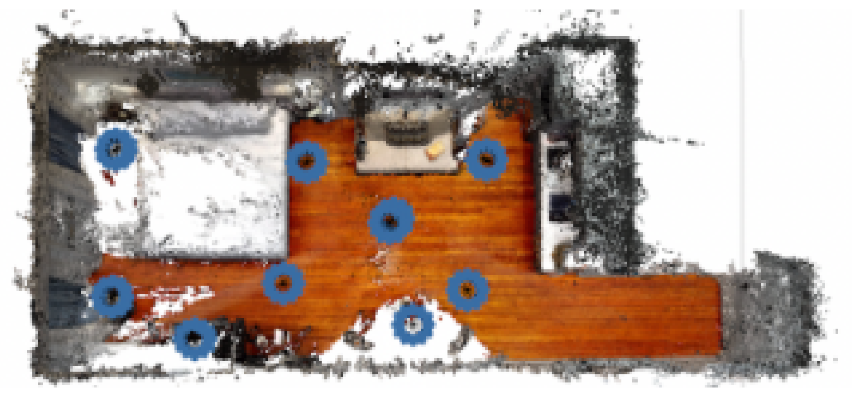

Figure. Dense-cloud 3D model recreated from Matterport Scanner data

UNDISTORTION

- GoPro cameras and Tobii glasses were used to capture a wider view, which resulted in distortions that could negatively affect the 3D reconstruction.
- We used MATLAB and Python scripts to perform radial and fisheye undistortions by reversing coordinates of the input image to the undistorted pixel locations through a mathematical model.

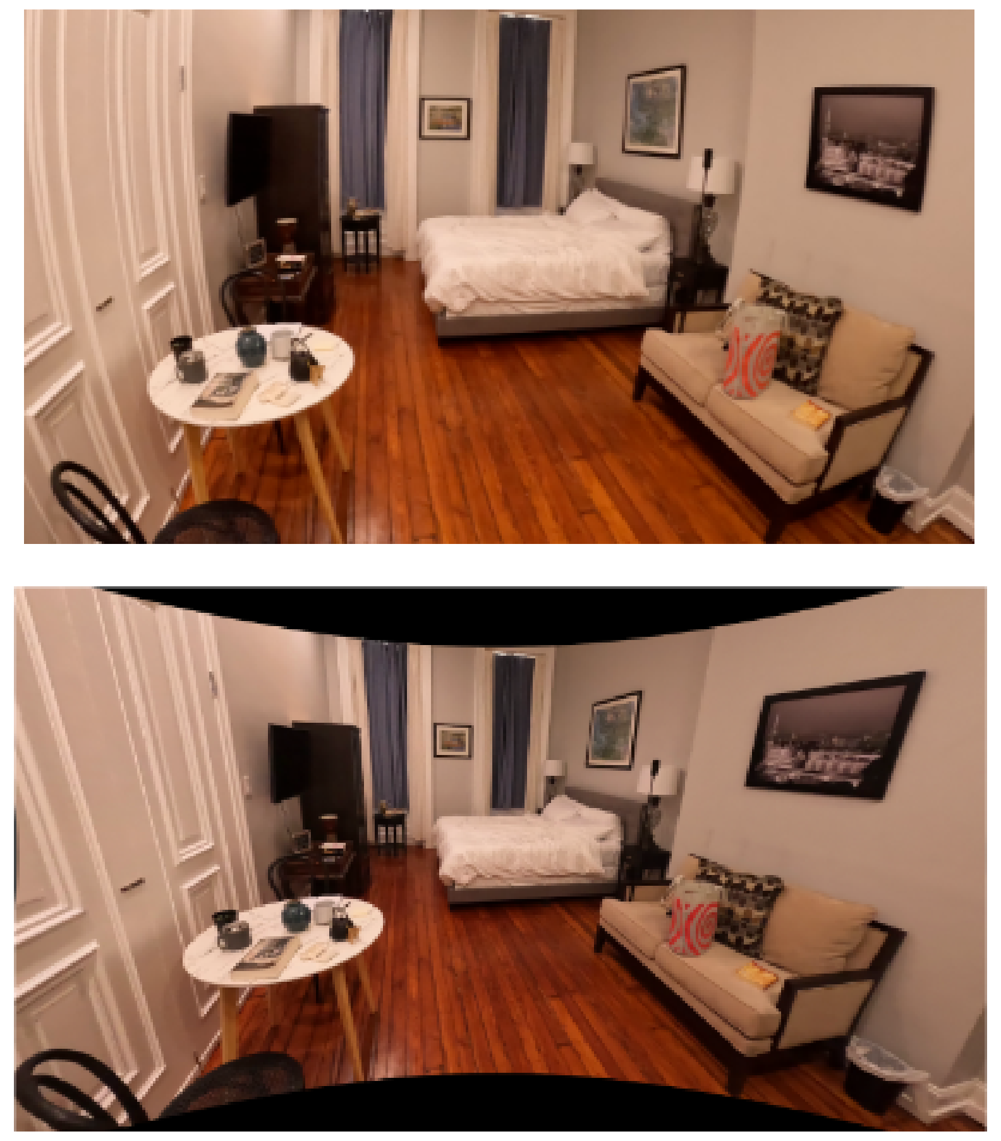

Figure. Frame of video captured by a GoPro camera worn on a participant's head (above) and same frame of video after passing through the undistortion script (below)

3D RECONSTRUCTION

- Once videos of subjects performing tasks were taken, we wanted to reconstruct the environment they were working in to better understand and analyze their interactions with the surroundings.
- Given consecutive 2D images with only 2 dimensions, we were required to produce a model that also included a third dimension - depth.

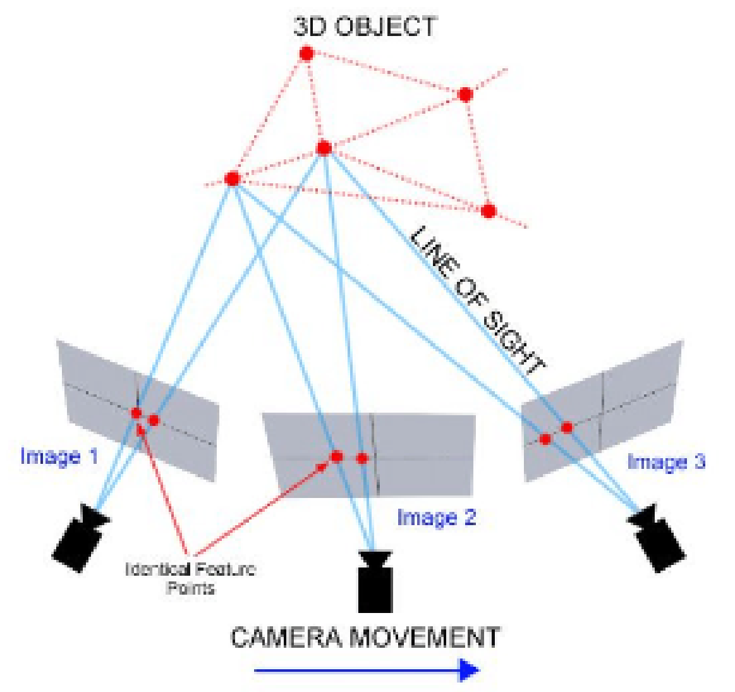

Figure. 3D reconstruction from multiple images, wherein point coordinates are calculated by the intersection of the two or more projection rays (Mason 2017).

- We extracted frames of videos at a rate of 5 frames per second (FPS).
- Images were then fed into Metashape for photo alignment, dense cloud building, mesh building, and texture building.

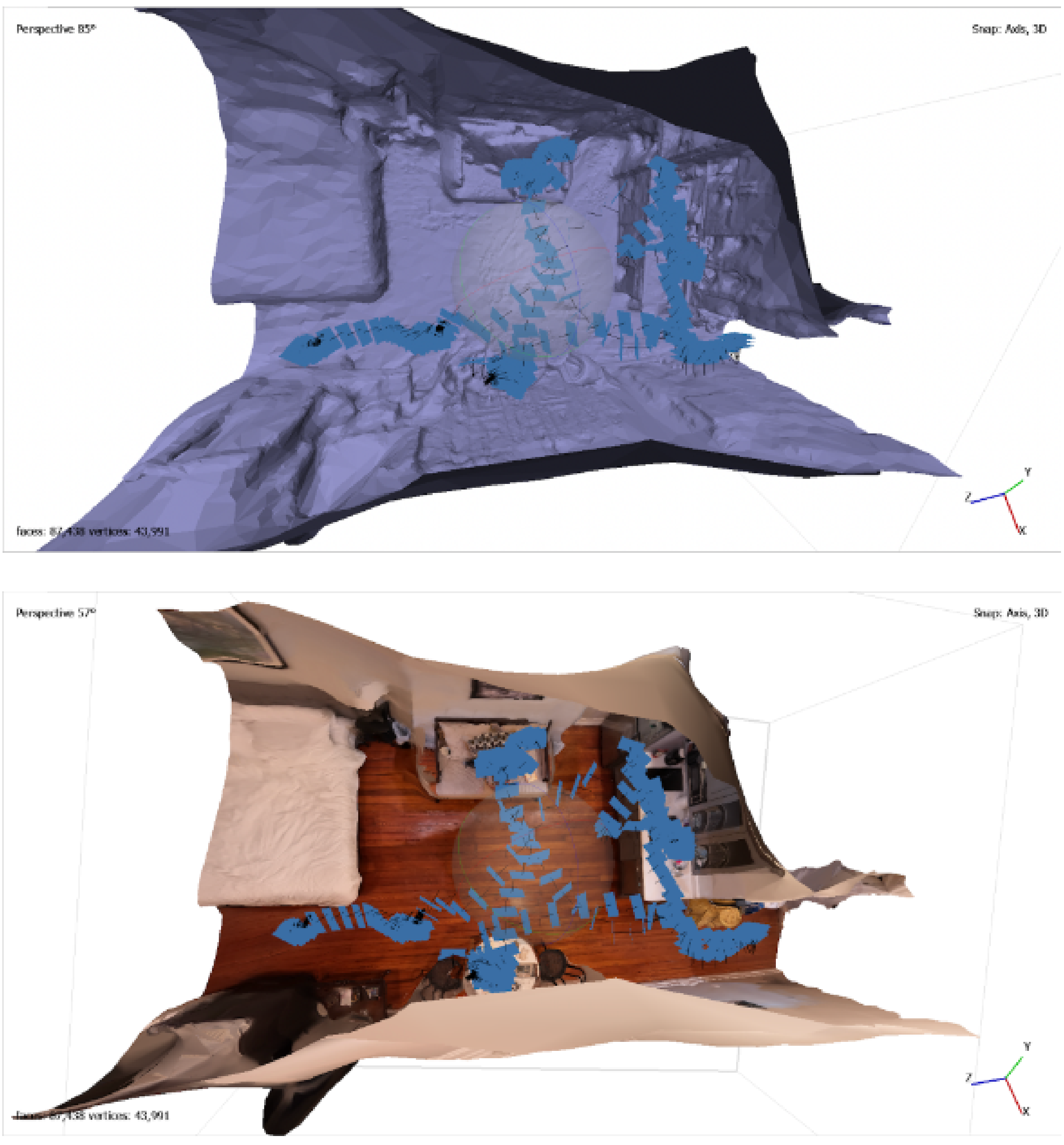

Figure. The mesh model (above) and textured model (below) of the experiment room reconstructed using Metashape.

DETECTION & RECOGNITION

- A real-time object detection and recognition model was preferred. However, many two-stage object detection networks, such as Faster R-CNN, can only process images up to 18 FPS (Redmon et al., 2016).
- We implemented YOLO V1, a one-stage object detection and recognition model that is able to process images at 45 FPS and reaches an accuracy (mAP) of 63.4% (Redmon et al., 2016).

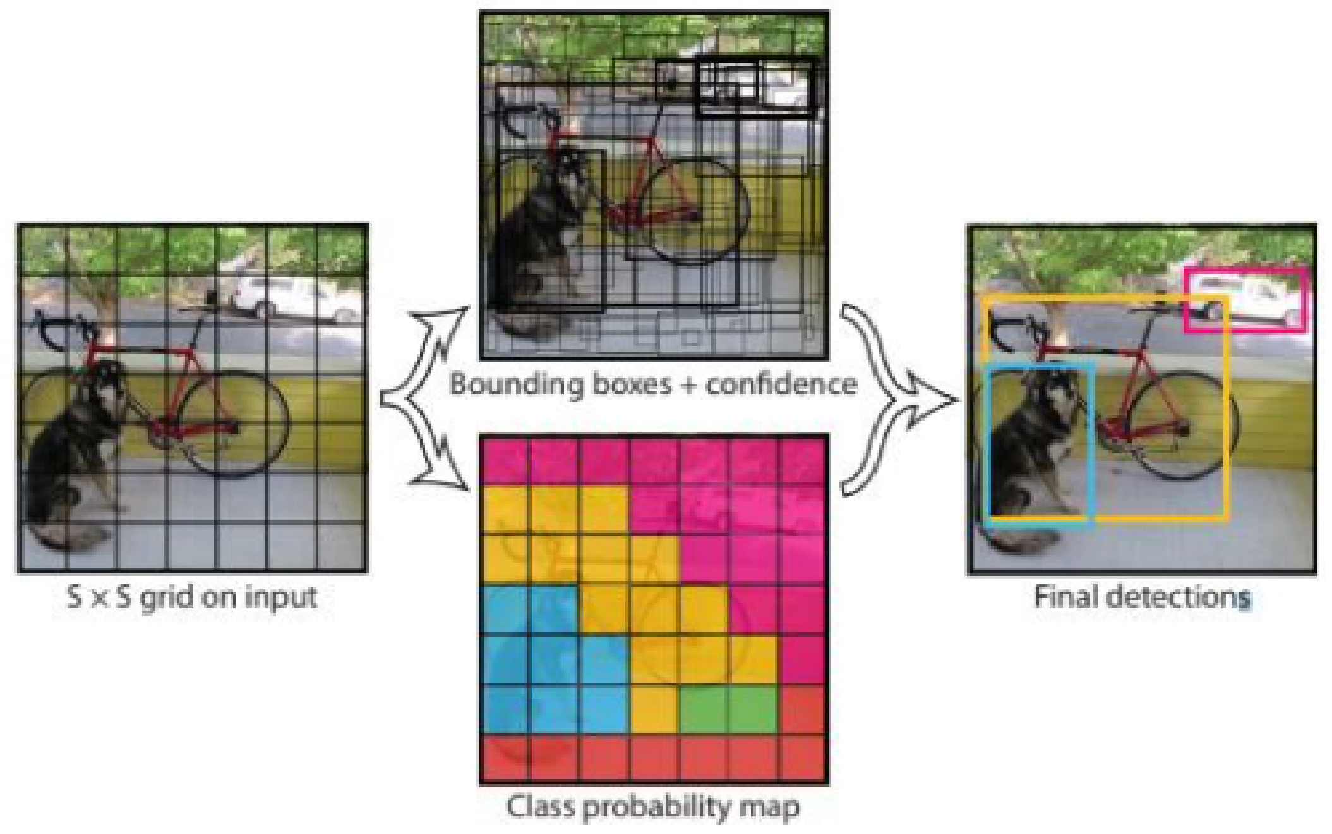

Figure. A demonstration of YOLO V1 's process for object detection and recognition. YOLO V1 divides each input image into an *S x S* grid and then predicts *B* bounding boxes, the confidence for those boxes, and the classification *C* for each box. Sourced from You Only Look Once: Unified, Real-Time Object Detection (Redmon et al., 2016)

- The model was then trained with 10 000 images for 4+ hours
- An Hue-Saturation-Value (HSV) heatmap was then used for debugging purposes to understand the model's performance against both accuracy of the object class and confidence value. The hue of each grid represented the most likely class while the saturation represented the confidence value.

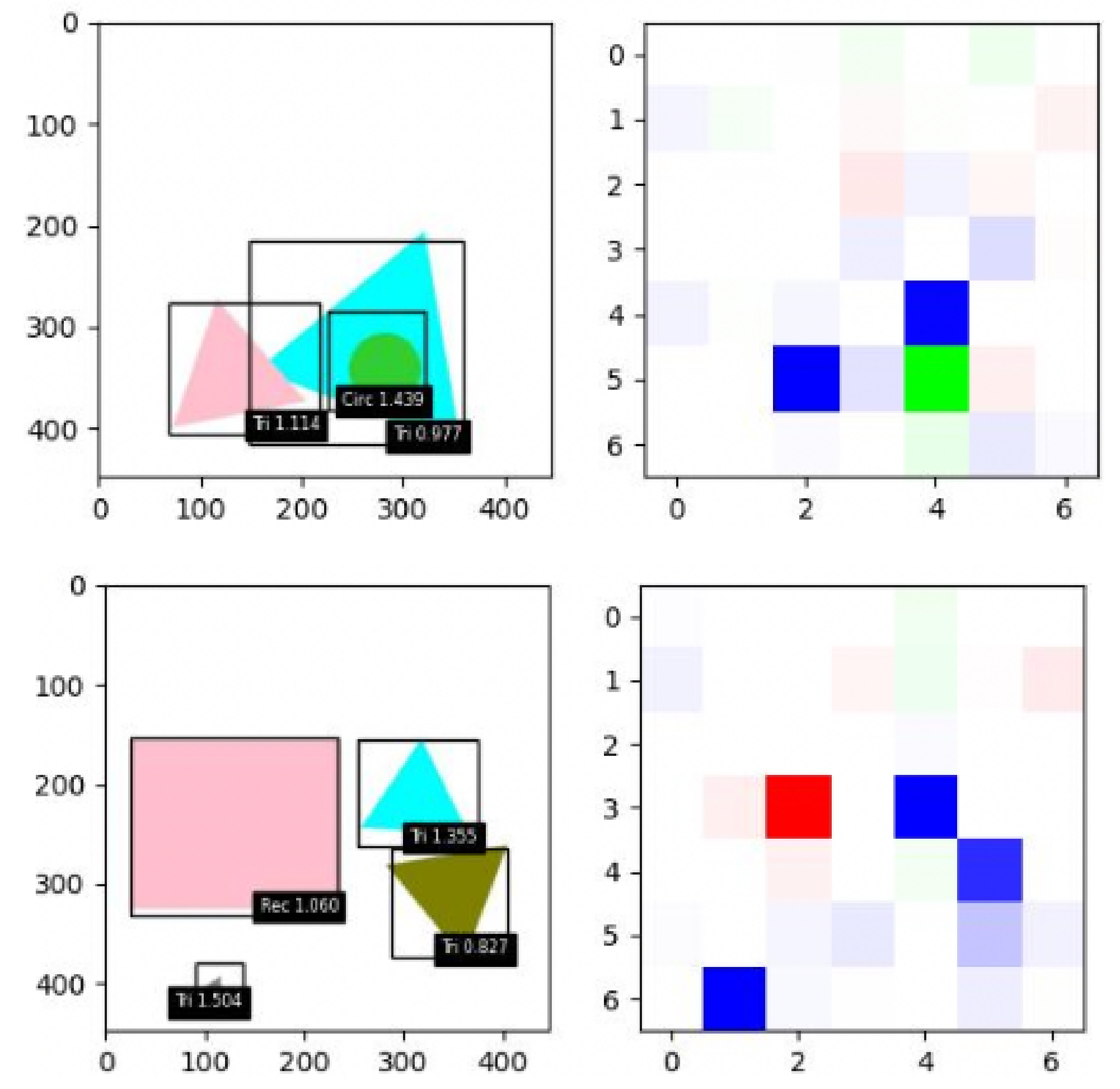

Figure. Results from testing images as predicted by the created YOLO V1 model. The left hand side include the objects (shapes), labels for such objects, and their confidence values, while the right hand side shows the confidence and class for each grid in the YOLO V1 model.

FINE TUNING MODEL

- Once the object detection and recognition model was created, it was also important to fine tune the model for more accurate results.
- We were able to raise the shape-recognition model's accuracy from 0.69 to 0.72 by changing the learning rate, the amount of data inputted as the training set, and the amount of time it was trained for to pick the best version.

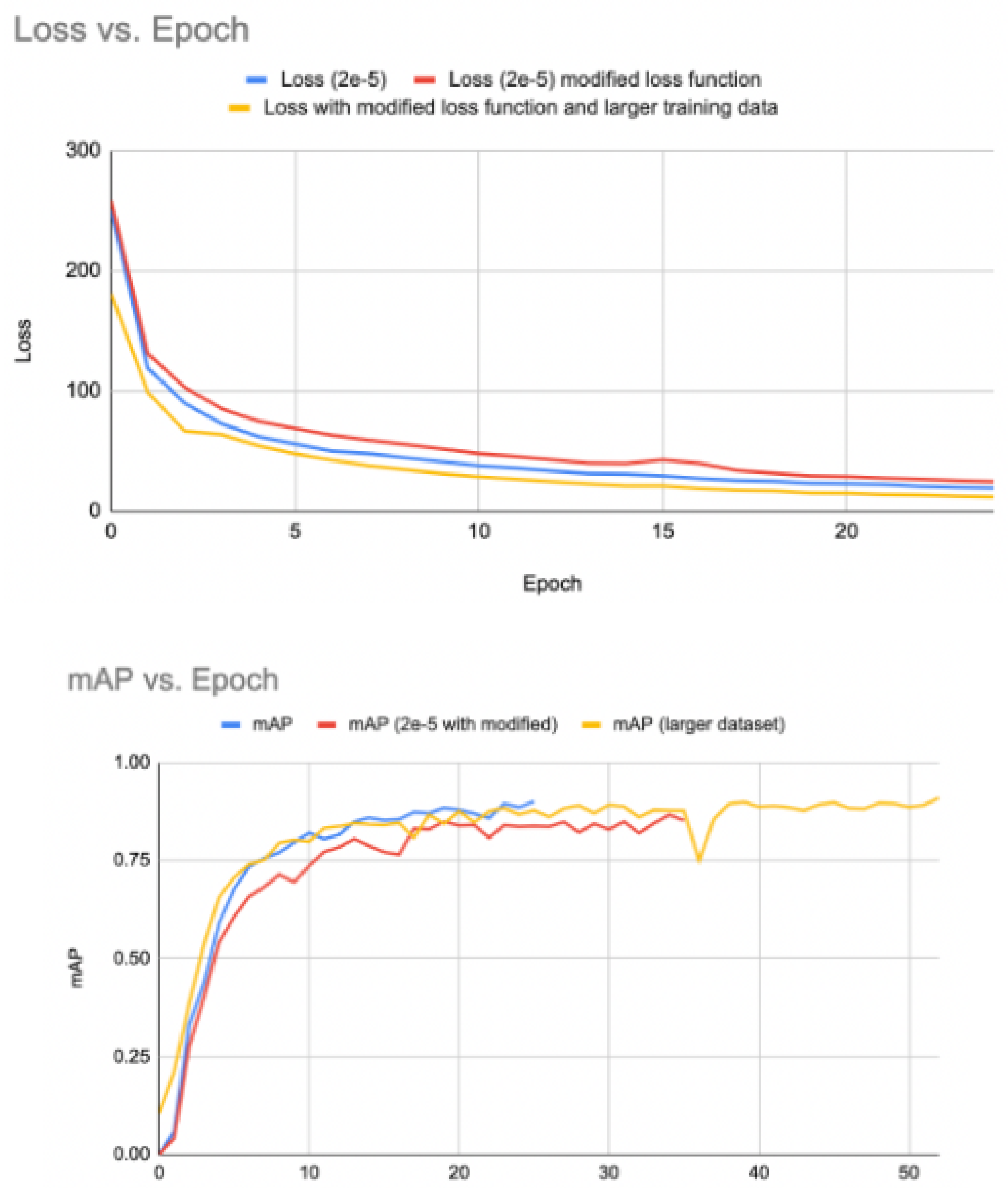

Figure. The mean average precision (mAP) vs. the number of epoches trained (above) and the total loss per epoch vs. epoches trained for different learning rates and loss functions

REFERENCES

- Redmon, Joseph, et al. "You only look once: Unified, real-time object detection." (2016).

- Three-Dimensional UAV-Based Photogrammetric Structural Models for Rock Slope Engineering: Slope Stability: Case Histories, Landslide Mapping, Emerging Technologies - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/3D-reconstruction-from-multiple-images-wherein-point-coordinates-are-calculated-by-the_fig1_327234684

- Lee, Charles B.. "RADIAL UNDISTORTION AND CALIBRATION ON AN IMAGE ARRAY." (2000).