Mid-Term Project Report

Team Members:

- Aniket Patel
- Matthew Kasper
- Matej Popovski

What has changed:

- Narrowing down the scope of our project. The main idea is to control certain computer actions with our hand, replacing the function of the mouse + any custom additional actions.
- While previously not having any dataset, we created our own dataset, including 711 hand photos of different gestures (taken by the participants in the project).
- From the 6 different methods we had listed in the proposal, we specified one method that we use - YOLOv12 nano. For this project we require inference to be run in real time, which eliminated consideration for some methods such as a cascade classifier. Using YOLO also allowed us to leverage existing general models trained on the COCO image dataset, making training time feasible (30 minutes on a NVIDIA T4 GPU) as we only need to finetune the model on our small dataset.

Current Progress:

- Created a custom dataset of 711 images with annotations for 6 gestures: closed hand, closed hand thumbs out, open hand, palm out, thumbs up, thumbs down.
- Finetuned an existing general model of Yolo v12n on our custom dataset.
- Created a program that captures a live video feed and annotates the hand with the gesture label in real time.

Difficulties:

- Our current training dataset does not have an equal distribution of images for each gesture. This has led to model predicting the over-represented gesture when it should not.
    - In order to tackle this, we plan to add more images to our dataset for the gestures that are underrepresented. At the end, we should have an equal number of images for each gesture.
- Our current testing set also did not have an equal distribution of images from each gesture, which has led to misleading results.
    - To fix this issue, we plan to include an equal distribution of images for each gesture in our testing dataset to get an accurate representation on how the model performs on a variety of our data.
- We currently do not have a thorough testing process. We simply run our model of our testing data once and use that to calculate the accuracy.
    - We plan to implement a more robust testing process: 5-fold cross-validation.
- Our current training dataset does not have enough variety, making our model lack robustness for handling different people, lighting conditions, backgrounds
    - We plan to fix this by adding more images to our dataset, varying these parameters to build robustness to these different scenarios.

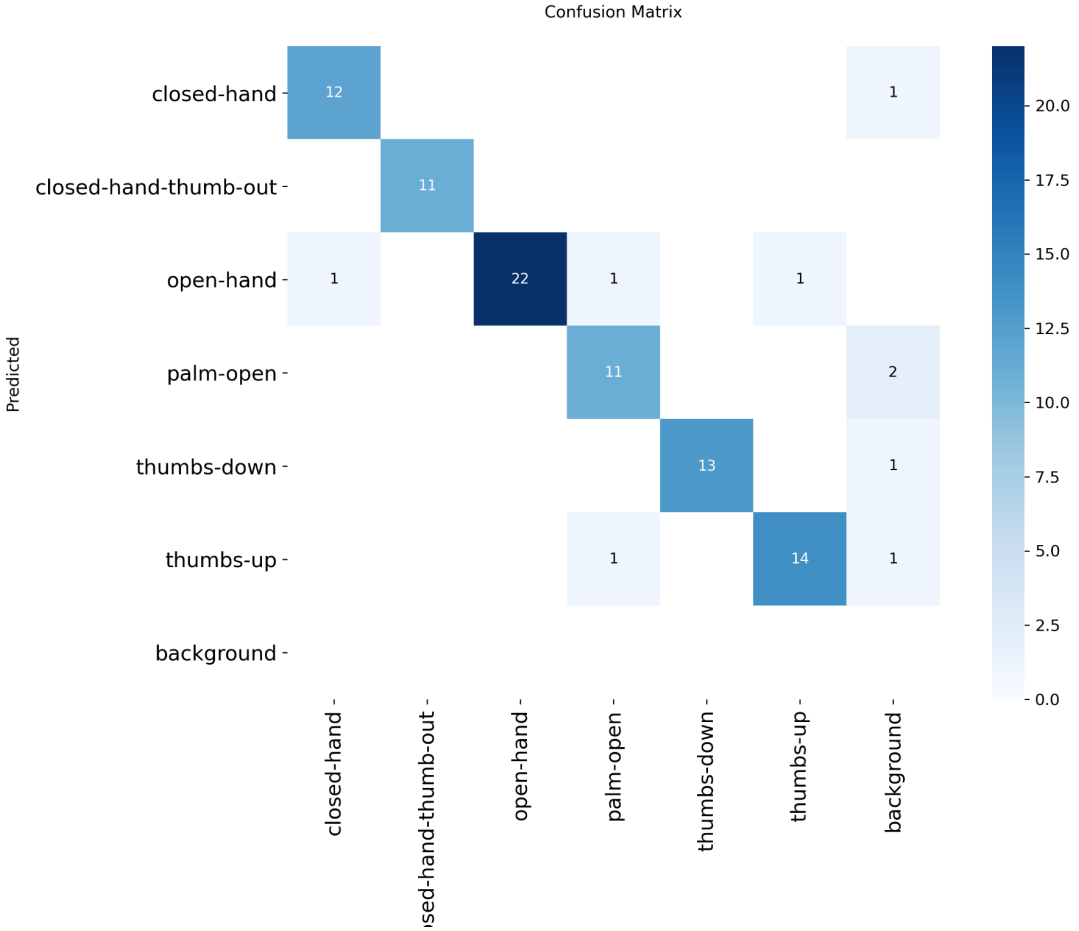
    - We can also create more robustness via image augmentation, such as rotating and flipping the image and adjusting its hue, saturation, and exposure, which will create more robustness in detection.
- We realized that even among the 6 gestures, different people may do the same gesture in slightly different ways. So, we need a way to account for that.
    - We plan to add more training images to account for this.

Next Steps:

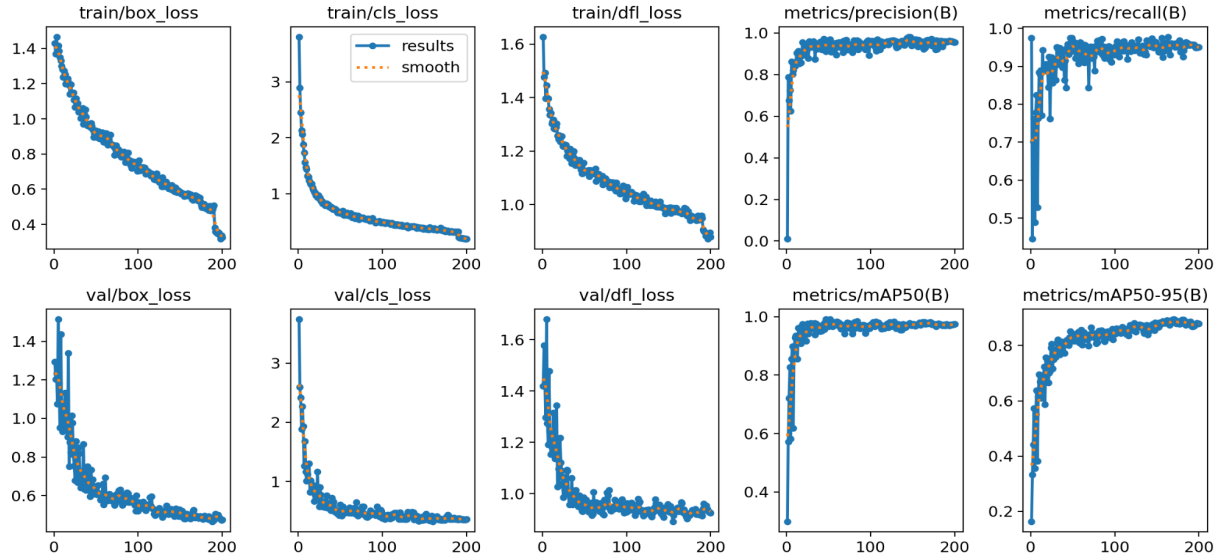
- Make the changes specified in the Difficulties section.
- Map the 6 detected gestures to specific computer actions such as Pause/Play, Volume Up, Volume Down, Move Mouse, Left-Click, Right-Click

Training Results

Confusion Matrix



Training Graphs



Sample of Testing Images Annotated By Trained Model