Facial Palsy Early Detection Based on YOLOv5

Yiying Pan[1], Feiya Xiang[2], Edgar J.Lobaton[2]

[1]Department of Psychology and Behavioral Sciences, Zhejiang University, Hangzhou, Zhejiang, 310058, China.

[2]Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC 27695, USA.

Background

Facial palsy, affecting about 1 in 5,000 people annually, causes sudden loss of muscle control on one side of the face, leading to uneven facial movements. While its exact causes are unclear, viral and bacterial infections are known contributors. Early detection is crucial, especially for those with diabetes or HIV, who are at higher risk.Traditionally, facial palsy diagnosis relies on invasive methods like EMG and ENoG. This project aims to use non-invasive images and videos for early detection. By leveraging advancements in AI and computer vision, we are developing and optimizing a face detection algorithm to improve diagnostic accuracy and timely treatment.

Dataset

YouTube Facial Palsy (YFP) Database [1] is a specialized dataset curated for research purposes in the field of Facial Palsy detection and analysis. It comprises a collection of video clips sourced from YouTube, featuring individuals with varying degrees of facial paralysis：

- 32 videos of 21 patients from YouTube, and a few patients have multiple videos.
- As the shortest facial palsy session lasts for a second, each video is converted into an image sequence with 6FPS;
- Manually labeled local palsy regions when the deformation intensity was considered sufficiently high by clinicians.
- The palsy regions were labeled by three independent clinicians, and we used the intersection of the independently cropped regions as the ground truth.
- We labeled the intensity observed in each palsy region as 0.5 for low or 1.0 for high.
- In addition to the syndrome intensity, we also labeled the palsy regions into Classes Eyes or Mouth.

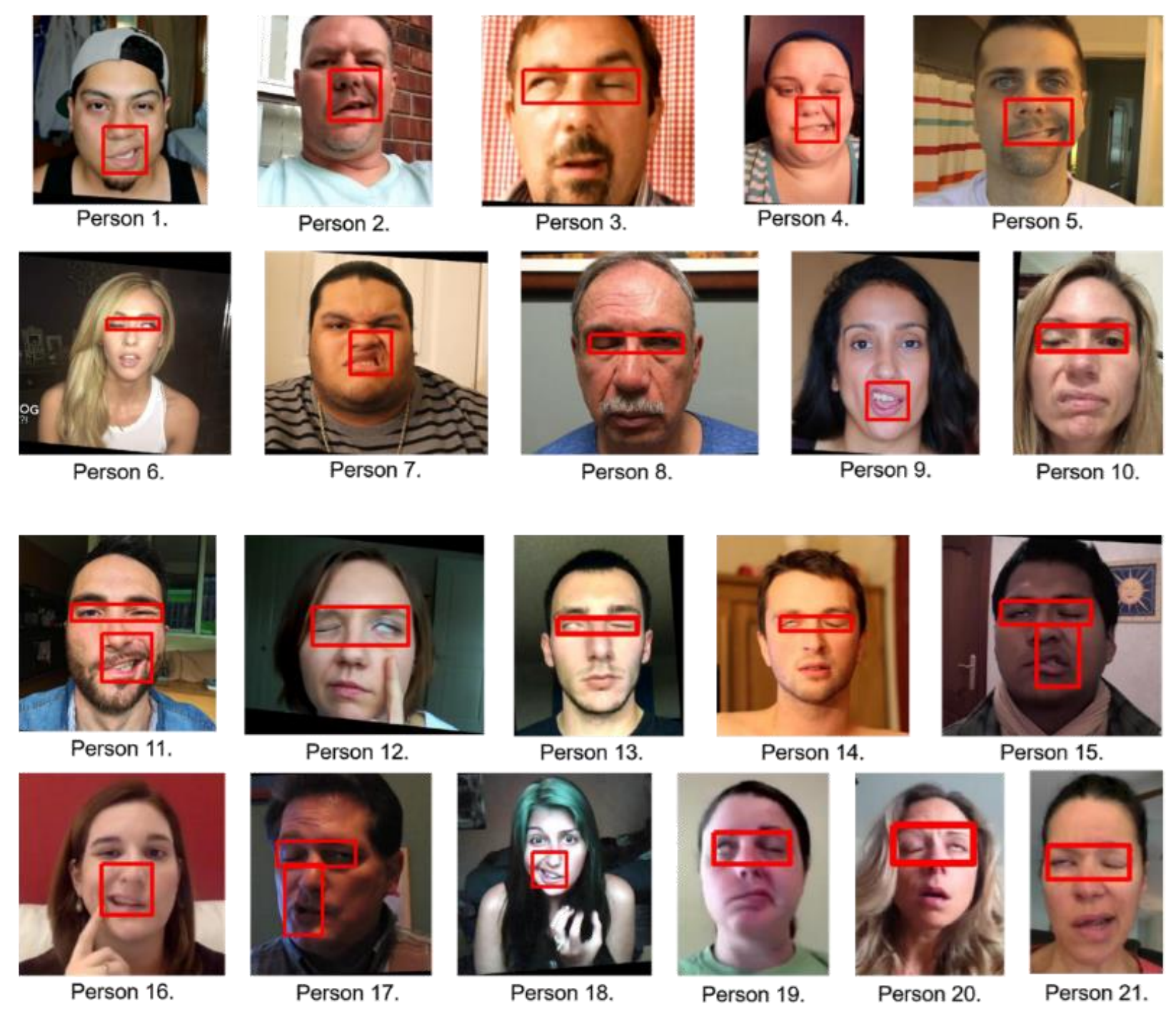


Figure 1 Facial Paralysis Patient Representation

Model: YOLOv5

YOLOv5 (You Only Look Once version 5) is a state-of-the-art, real-time object detection model designed for high-speed and high-accuracy performance. It builds on the YOLO (You Only Look Once) series, featuring several improvements and optimizations.

Architecture:

- **Backbone:** YOLOv5 uses CSPNet (Cross-Stage Partial Network) as its backbone for feature extraction. CSPNet divides the network into two parts, processing them separately and then merging the results. This approach improves training efficiency and feature representation.
- **Neck:** YOLOv5 incorporates PANet (Path Aggregation Network) to enhance feature aggregation across different scales. PANet combines features from multiple layers to better detect objects of various sizes and improve the detection of small objects.
- **Head:** The head of YOLOv5 outputs predictions for object classes, bounding box coordinates, and objectness scores. It predicts bounding box coordinates (x,y)(x, y)(x,y), width www, and height hhh relative to anchor boxes.

Loss Function:

- **CIoU Loss:** YOLOv5 uses Complete Intersection over Union (CIoU) loss for bounding box regression, which considers not only the overlap but also the distance between the centers of predicted and ground truth boxes:

$$\mathrm{CIoU} = \mathrm{IoU} - \frac{\rho^2(b,b^*)}{\mathrm{v}+\varepsilon} + \alpha \mathrm{v}$$

- **Classification Loss:** Utilizes cross-entropy loss to measure the accuracy of class predictions:

$$L_{\text{class}} = -\sum_i y_i \log(\hat{p}_i)$$

- **Confidence Loss:** Measures the accuracy of objectness score predictions with binary cross-entropy:

$$L_{\text{conf}} = -\sum_i \left[y_i \log(\hat{p}_i) + (1-y_i)\log(1-\hat{p}_i) \right]$$

Training and Inference:

- **Training:** Involves data augmentation techniques to improve model robustness and generalization. The model is optimized using stochastic gradient descent (SGD) or its variants.
- **Inference:** During inference, YOLOv5 applies non-maximum suppression (NMS) to eliminate redundant bounding boxes and select the most confident predictions.

Data Preprocessing

- **Convert XML Format Labels:** Transform XML format labels into a format compatible with YOLOv5s model input.
- **Resize Images:** Convert all images to a resolution of 640 × 640 pixels.

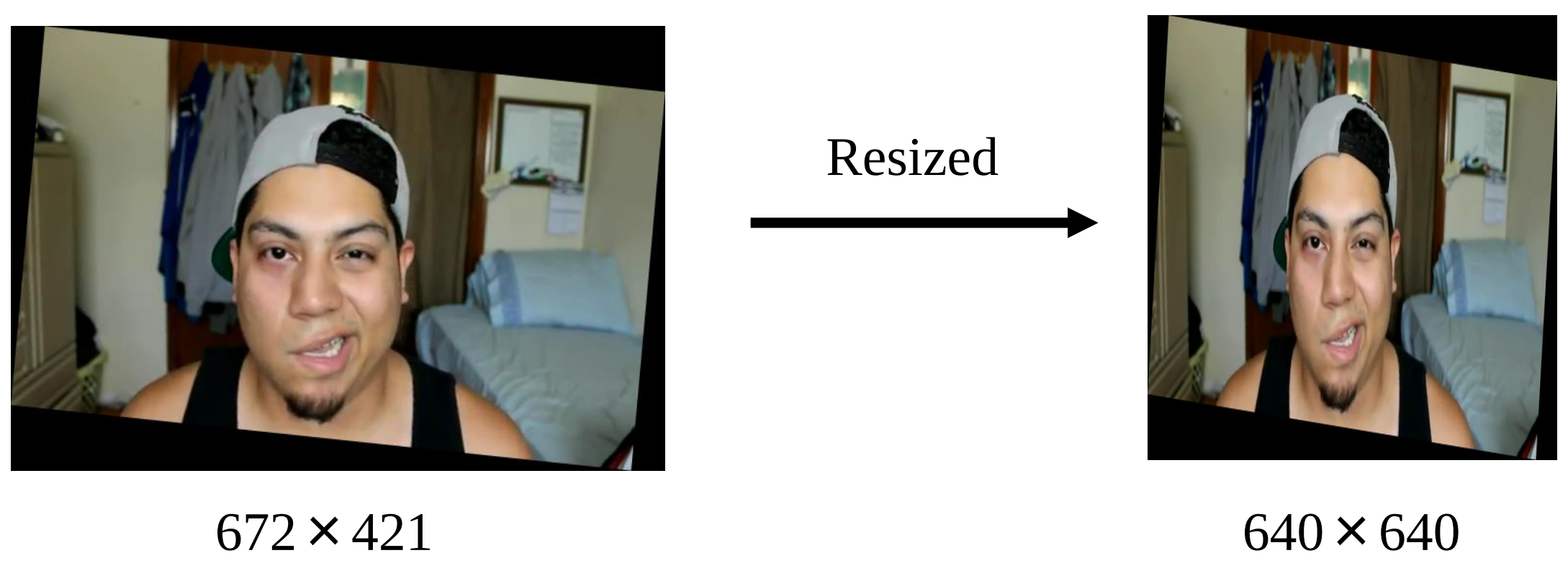


Figure 2 Example of Dataset Image Conversion

- **Split Video Data:** Use videos 1-28 as the training set, videos 29-30 as the validation set, and videos 31-32 as the test set

Result

Train:

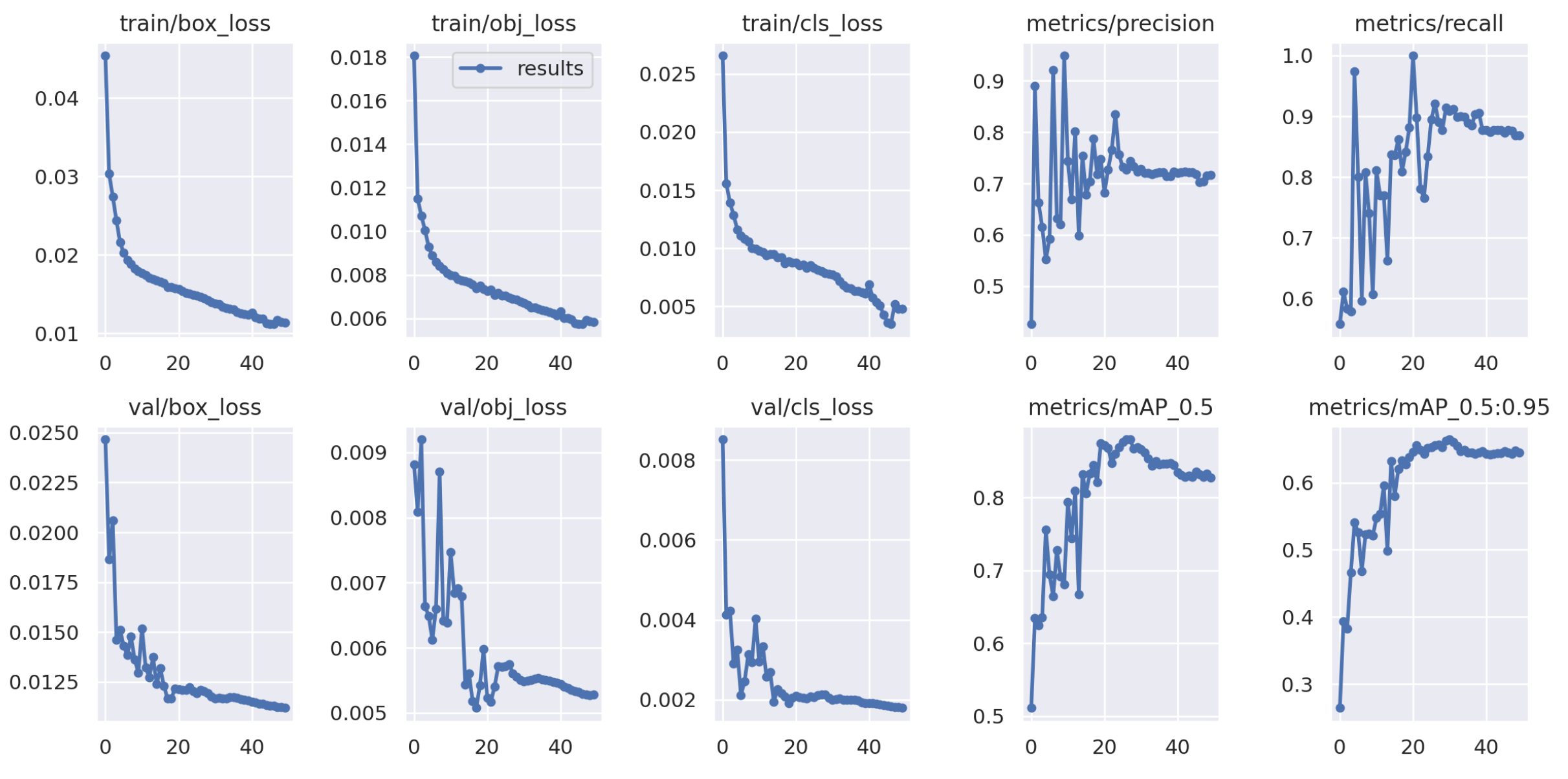
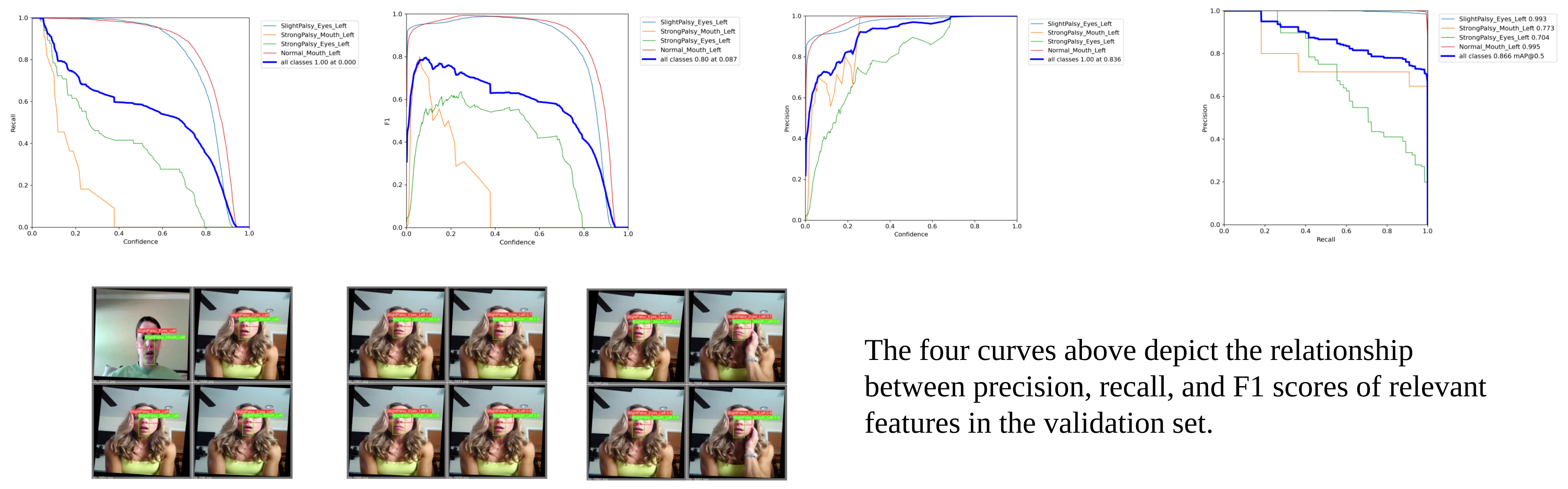


Figure 3 Evolution of Model Training Performance Across Epochs

Horizontal Axis: Epoch (Indicates the current training iteration)
Vertical Axis:

- **train/box_loss:** Loss associated with bounding box predictions during training
- **train/obj_loss:** Loss related to objectness predictions during training
- **train/cls_loss:** Loss associated with class classification during training
- **metrics/precision:** representing the ratio of correctly predicted positive instances among all positive predictions**metrics/recall:** indicating the ratio of correctly detected instances among all true positive instances in the dataset
- **metrics/mAP_0.5:** Average Precision (mAP) at an IoU threshold of 0.5
- **metrics/mAP_0.5:0.95:** Average Precision across multiple IoU thresholds (from 0.5 to 0.95 with a step size of 0.05)
- **val/box_loss:** Loss associated with bounding box predictions on the validation dataset.
- **val/obj_loss:** Loss related to objectness predictions on the validation dataset.
- **val/cls_loss:** Loss associated with class classification on the validation dataset.

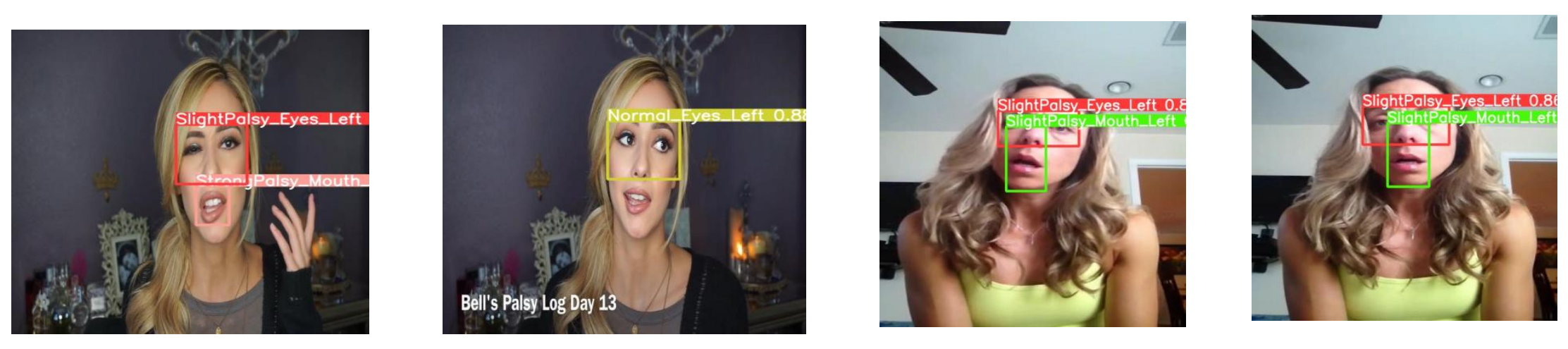
Validation:



The four curves above depict the relationship between precision, recall, and F1 scores of relevant features in the validation set.

Figure 4 Validation Set Results

Test:



The figure on the right showcases the sequential detection outcomes from the final two videos, delineating the precise positions and associated confidence levels of facial paralysis in the ocular and oral regions.

Figure 5 Testing Set Results

Conclusions

In this project, the YOLOv5 model was employed for facial paralysis detection on the YFP dataset. The model trained for 50 epochs achieved an accuracy of 71.62% and a recall rate of 86.834%. The experimental outcomes indicate that the framework applied is an effective solution for automated quantitative analysis of facial paralysis.

References

[1] “YouTube Facial Palsy (YFP) Database.” Accessed: Aug. 27, 2024. [Online]. Available: https://sites.google.com/view/yfp-database

[2] Hsu, G.S.J., Kang, J.H. and Huang, W.F., 2018. Deep hierarchical network with line segment learning for quantitative analysis of facial palsy. IEEE access, 7, pp.4833-4842.