UNIVERSITY OF SCIENCE AND TECHNOLOGY OF HANOI

Computer Vision

REPORT

Group 19

Face Detection with YOLO

Member

BA11-007:	Nguyễn Tuấn Anh
BA11-079:	Lương Khôi Nguyên
BA11-093:	Trần Minh Trung
BA11-011:	Nguyễn Gia Bách

Department ICT

Hanoi, March 2024

Table Contents

I. Introduction

Before the function of detecting and recognizing faces, there were many problems that had to be solved in a way that was not expensive but also took time and effort. Such as security for a device such as a computer, phone, managing people entering and leaving the company or even tracking criminals,... First for security, we must use text-based passwords, number. However, alphanumeric passwords often require us to remember. As there are more and more devices and more accounts, we will be required to remember more, which leads to a lot of trouble. worth having. Next, when we want to manage employees, check everyone's attendance, previously we had to call them by name, even now at USTH it still takes a lot of time to do that. Therefore, in the 21st century, with the continuous development of AI, people have created tools to identify and recognize human faces, helping to do tasks that previously took a lot of time. This becomes more convenient, faster, even more accurate. However, during the learning and testing process, we realized that solving the face recognition problem is not simple. Therefore we will split the problem into many small parts. And in this report we will address the issue of face detection, an important step in the face recognition process.

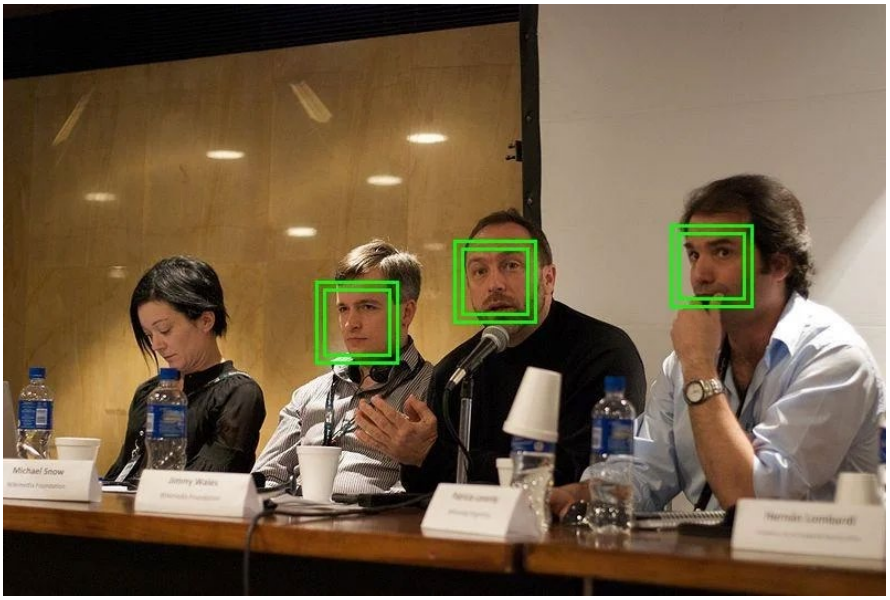


Figure 1: Face Detection for detecting a human face.

II. System Components

Face Detection

Related Work

Face detection[1] has become a cornerstone in the field of computer vision with applications spanning from security to social media which is an Artificial Intelligence (AI) technology that identifies human face within digital images and videos. This technology leverages deep learning (DL) and artificial neural networks (ANNs) to distinguish faces from other object and features within an images.

In order to estimate the human face, active detection systems frequently make use of cameras such as in the mobile phone, security camera and embedded peripherals. However, the cost computing and human resources required to construct accurate and handle each situation in the problem are typically very high.

In the proposal-based framework, the Face detection with R-CNN [2] are composed of several correlated stages with combination of features extracted from CNN, classification, and bounding box regression, which are usually trained separately. Because of handling various stability accuracy which becomes an issue depending on human face. Single-stage systems like YOLO and SSD, which directly translate detected regions of an image into bounding box coordinates and class probabilities through global regression/classification, can significantly cut down on processing time.

YOLOv5

Overview

You Only Look Once (YOLO)[3] is a powerful object detection approach that is widely used in computer vision because of its high processing power. The newest model in the YOLO series is known as YOLOv8, created by Ultralytics, the same firm that built the well-known and industry-defining YOLOv5 model. The reason we chose YOLOv5 is because the stability of the statement and the output of the face detection problem need to have a certain stability to recognize human faces. This model is capable of image processing even on devices with low resources. In terms of stability, YOLOv5n outperforms YOLO5s, YOLOv7, YOLO8n, and R-CNN, emphasizing its suitability for face detection applications.

YOLOv5[4] was trained on the COCO dataset, a large-scale image recognition dataset used for object detection, segmentation, tracking, and labeling tasks that comprises more than 330,000 photos, each labeled with 80 classes and 5 descriptions that describe the scenario.

Architecture

In the YOLOv5 is following compose of three parts: Backbone, Neck, and Head[5]

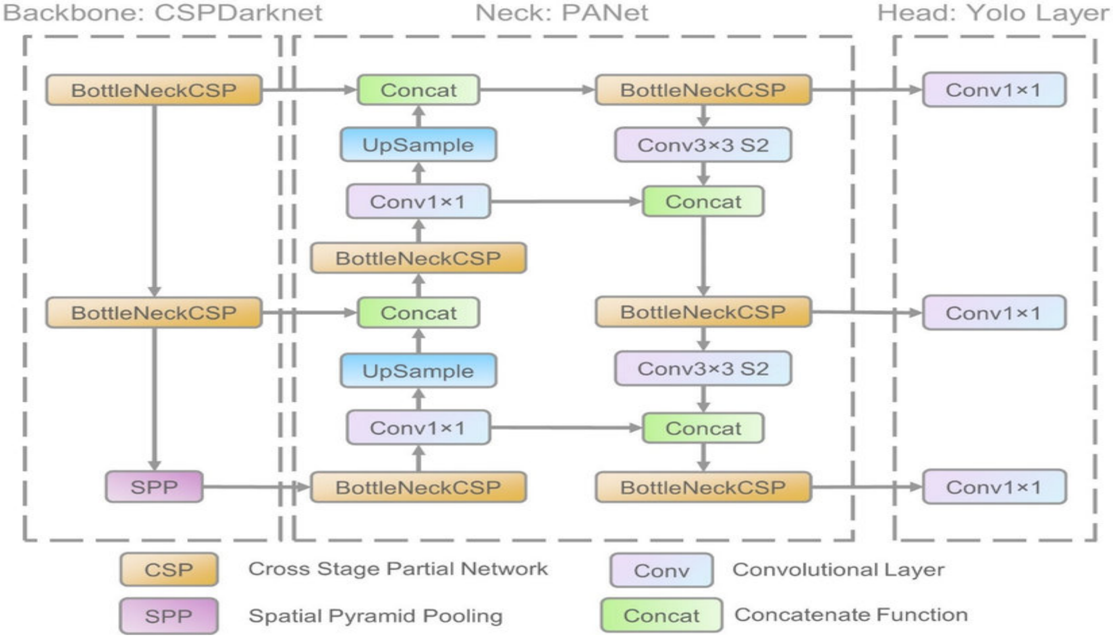


Figure 2: YOLOv5's Architecture

- **Backbone**

The backbone is typically employs a convolution neural network (CNN) as its backbone, often based on CSPDarknet architectures which stands for Cross-Stage Partial Darknet. The Darknet is convolution neural network architecture known for its simplicity and efficiency. It consists of multiple convolutional layers followed by down sampling procedures such as max-pooling or stride convolution to extract hierarchical features from the input image. The Cross-Stage Partial Connections (CSP) provide the idea for enhancing feature reuse and facilitating information flow across different stages of the network. This works by connecting the output of one stage directly to the input of another, enabling information to bypass certain stages while reducing total computing costs.

- **Neck**

Continuing from the backbone is the neck which is primary function of enhancing the features extracted from the backbone. In YOLOv5 uses neck architecture is Path Aggregation Network (PAN) which solving a leaking information existing. It allows PAN to capture richer contextual information and facilitate long-range dependencies across different spatial resolutions. These modules combine features from multiple pyramid levels using operations such as summation or concatenation, followed by additional convolutional layers to refine the fused features. This process enhances the discriminative power of the feature representations and improves the model's ability to localize and classify objects accurately.

- **Head**

To the final part is the component of architecture is the head. It consists of multiple convolutional layer responsible for predicting bounding boxes and class probabilities for objects in the image. By utilizing a single-stage object identification method, YOLOv5 does not require the use of intermediary stages or separate area proposal networks to forecast bounding box coordinates and class scores.

- **Algorithms for detection**

Intersection over Union (IoU)

The IoU[6] is an essential indicator for assessing how well bounding box predictions perform during object identification tasks. For a given item in an image, IoU calculates the overlap between the predicted bounding box and the ground truth bounding box. This is done by simply adding a scaling factor as the following formula.

$$\text{IoU} = \frac{\textit{Intersction Area}}{\text{Union Area}}$$

Intersection Area: calculate the intersection area between the ground truth bounding box and the predicted bounding box. The region where the two bounding boxes overlap is represented by this area.
Union Area: The two bounding boxes' union area is then computed. The combined size of the ground truth bounding box and the anticipated bounding box, including the area where they overlap, is known as the union area.
Here is the figure 3 show sample that:

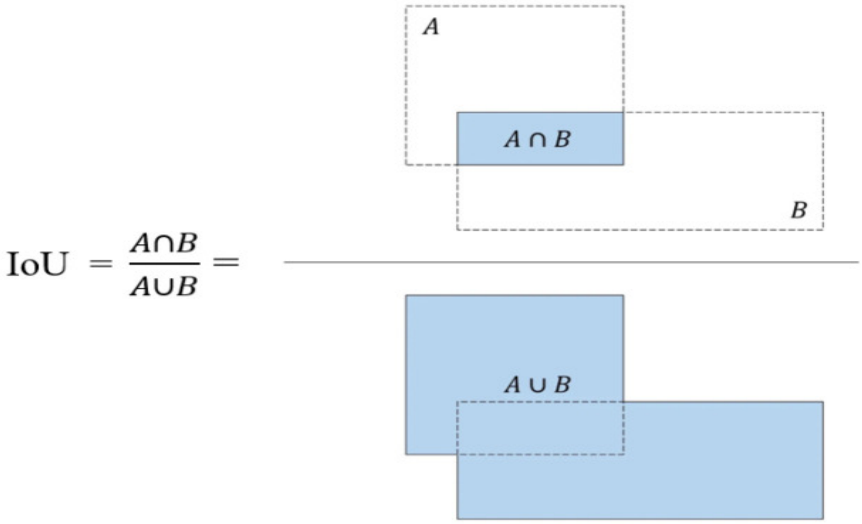


Figure 3: The IoU formula

Bounding box loss

The GIOU_Loss[7] function is used to calculate the bounding box loss which expanding the area of the union and then optimizing the IoU. For any two bounding boxes *A* and *B*, first, we find the minimum bounding box *C* that can cover them; then, the *GIOU*_Loss function is defined by Equations and (5):

$$L_{GIOU} = 1 - \text{IoU} - \frac{C-(A \cup B)}{C}$$

However, as shown in Figure , the GIoU_Losss is degraded to the IoU when the prediction box contains the actual box. Moreover, the convergence speed is slow in the horizontal and vertical directions when the two boxes intersect. This will lead to inaccurate detection results[6].

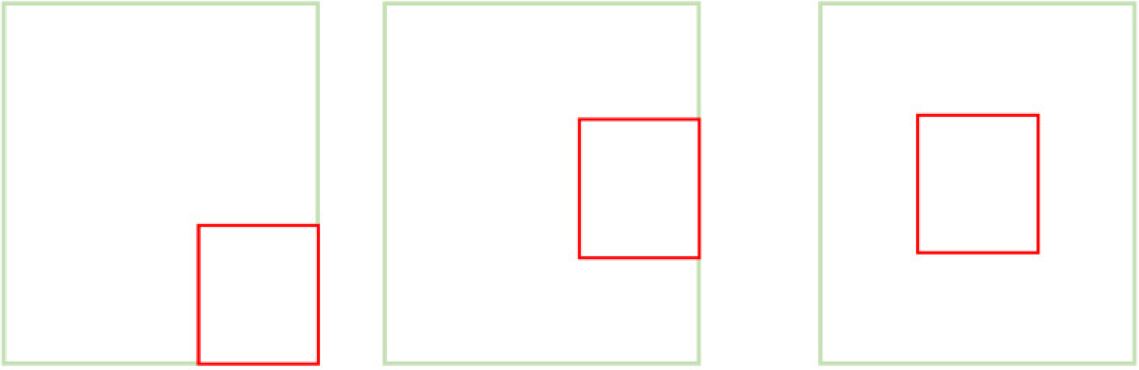


Figure : GIoU degraded to IoU

For small viewed faces that have a small bounding box which is below the threshold and don't match any of our sub-frames, we will then use the center of the face to indicate the face's position.

III. System Evaluation

1. Dataset and pre-processing

In this project, we trained and evaluated our models using WIDER FACE[8]. This dataset was chosen because it includes a large number of conditions for face detection training which focus on face detection and multi-face in the image. It is a benchmarked suite designed to all major face scene interpretation tasks which is focused on 2D scene and a large of face human interpretation.

In WIDER FACE consists of 3 components:

- The multi-face: a lot of people or human faces in an image.
- Annotations: the dataset is annotated with 2D and bounding box for face.
- Images quality: RGB-Images, the resolution from medium to high like the noise and without noise, and a lot of images with group human face

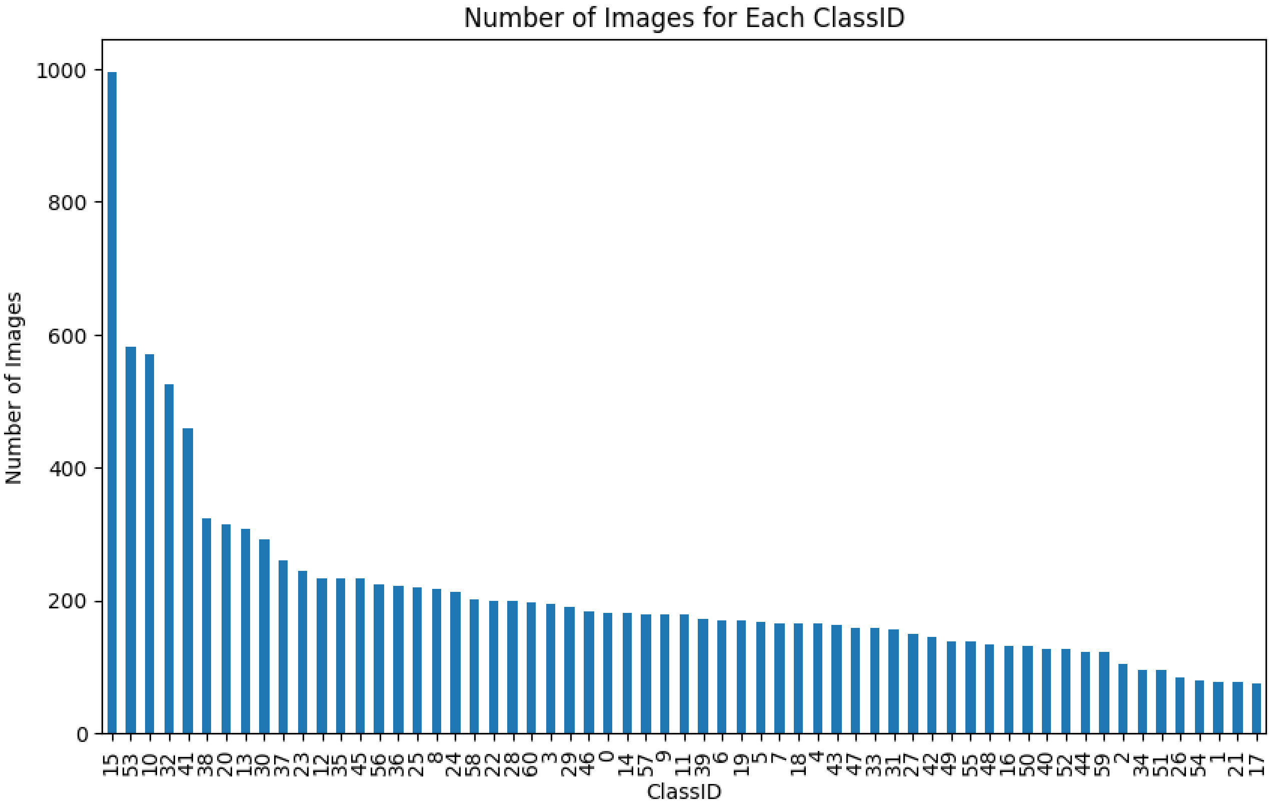


Figure 4: The number of Image for each classes

The dataset originally included 60 categories relevant to human faces environment. This is the list of our classes: Parade, Handshaking, Demonstration, Riot, Dancing, Car_Accident, Funeral, Cheering, Election_Campain, Press_Conference, People_Marching, Meeting, Group, Interview, Traffic, Stock_Market, Award_Ceremony, Ceremony, Concerts, Couple, Family_Group, Festival, Picnic, Shoppers, Soldier_Firing, Soldier_Patrol, Soldier_Drilling, Spa, Sports_Fan, Students_Schoolkids, Surgeons, Waiter_Waitress, Worker_Laborer, Running, Baseball, Basketball, Football, Soccer, Tennis, Ice_Skating, Gymmastics, Swimming, Car_Racing, Row_Boat, Aerobics, Ballconist, Jockey, Matador_Bullfighter, Parachutist_Paratrooper, Greeting, Celebration_Or_Party, Dresses, Photographers, Raid, Rescue, Sport_Coach_Trainer, Voter, Angler, Hockey, people_driving_car, Street_Battle. However, a class's object number is not distributed equally; many classes have less than 200 images such as the Demonstration class (Figure show the ClassID is 15) highest than other classes. It is consequence of unbalanced datasets.

2. Performance Evaluation of The System Components

The model's performance can be described by average with mAP, Precision, and Recall. The value of average Accuracy method which describing a moment when detected faces human. The SGD is used as the optimizer, with a weight decay of 0.0005 and momentum of 0.9. The warm-up method is used to initialize the learning rate, and the cosine annealing algorithm is adopted to update the learning rate. The size of the input image is 640 × 640 pixels, the batch size is 64, and the initial learning rate is 0.001. The total number of training epochs is 100.

mAP-50

Mean Average Accuracy (mAP)[9] is a performance metric used for evaluating class prediction of object detection models which ranges from 0 to 100. We chosen a metric is 50 for better performance of based on the average precision-recall curve.

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i$$

AP_i: the AP of class i

N: the number of classes

And, mAP@50 is the mean average precision calculated at an intersection over union (IoU) threshold of 0.50.

Confusion Matrix

A confusion matrix [10] is a table used to visualize and quantify the performance of a classification algorithm by indicating where the algorithm classified a value as compared to the ground truth. While confusion matrices are very intuitively calculated from classification models because their only output is a single class, confusion matrices can become more complicated in terms of their underlying computation for more complex computer vision tasks, especially in object detection.

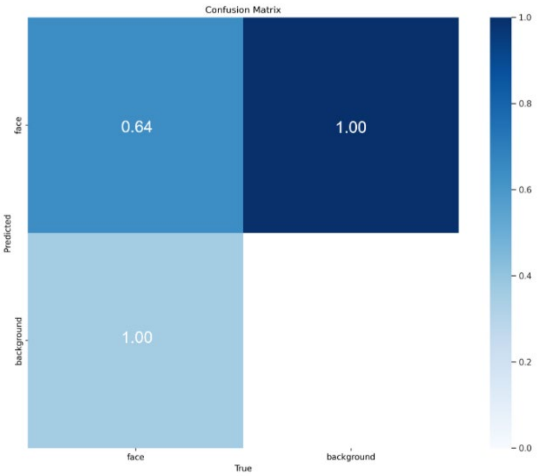


Figure 5: Confusion Matrix of YOLOv8s

During the evaluation, the confusion matrix provided important information about the model's performance. This matrix is used to specifically analyze the cases where the model predicted correctly and the cases where it made mistakes.

$$\mathbf{P} = \frac{\mathbf{TP}}{\mathbf{TP+FP}}\ ,\ \mathbf{R} = \frac{\mathbf{TP}}{\mathbf{TP+FN}}$$

when *P* denotes precision and *R* denotes recall
True positive (*TP*) denotes the number of positive samples that are correctly predicted as positive, false positive (*FP*) represents the number of negative samples that are incorrectly predicted as positive, and false negative (*FN*) denotes the number of positive samples that are mistakenly predicted as negative.

The Performance of Metrics

Model	Size	mAP-50	Precision	Recall	Accuracy
YOLOv5n	640	0.654	0.844	0.581	0.846
YOLOv5s	640	0.716	0.867	0.642	0.84
YOLOv7	640	0.783	0.8	0.759	0.847
YOLOv8n	640	0.799	0.86	0.763	0.891
R-CNN	640	0.568	0.686	0.436	0.809

Table 2: The Average of Performance Evaluation of YOLO version and other model

From Table 2 compares our model's performance to other fine-tuned models on the custom dataset, it appears that the YOLOv8n model performs the best overall, with the highest Average Accuracy of 0.891. However, the best model for a given task depends on the specific requirements of the task, such as whether precision or recall is more important. For example, if false positives are particularly costly, a model with higher precision might be preferred. Conversely, if missing a true positive is more costly, a model with higher recall might be preferred. It’s also important to consider the computational resources required by each model, as some models might be too computationally intensive to run in certain environments or in real-time applications.

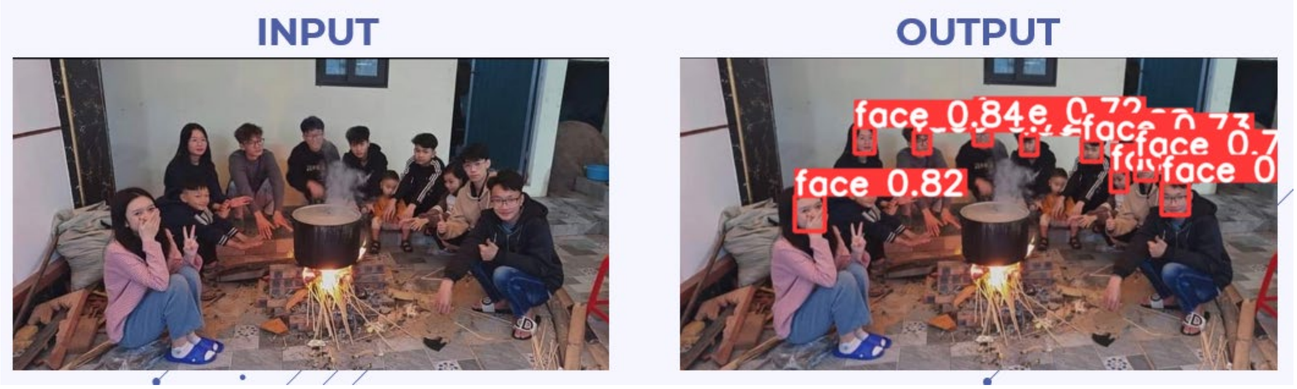
3. Inference of The System

a. Experiment Setup

We are using a computer with an CPU Intel i5 Processor, 16GB RAM, camera Android phone with 64MP, and we inference using Python 3.10 and Pytorch for running model.

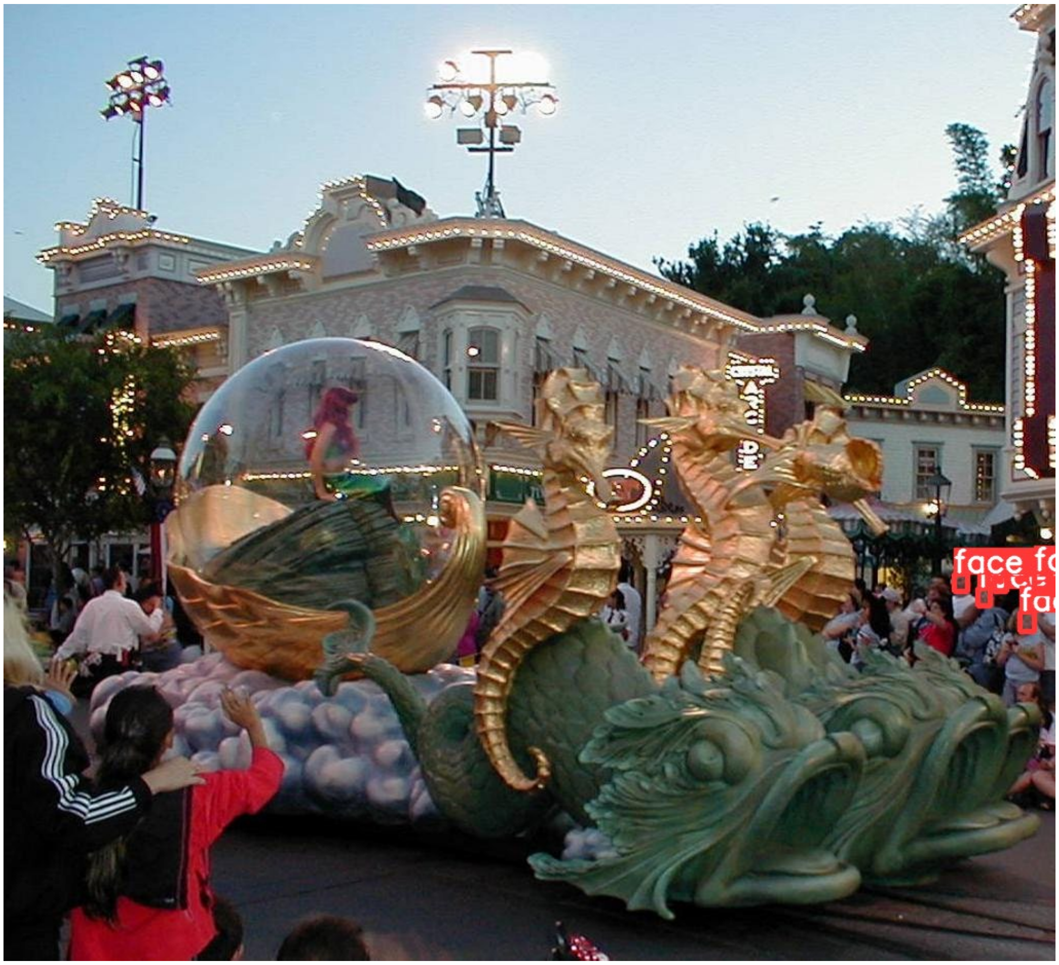
b. Result and Discussion

The Indoor environment



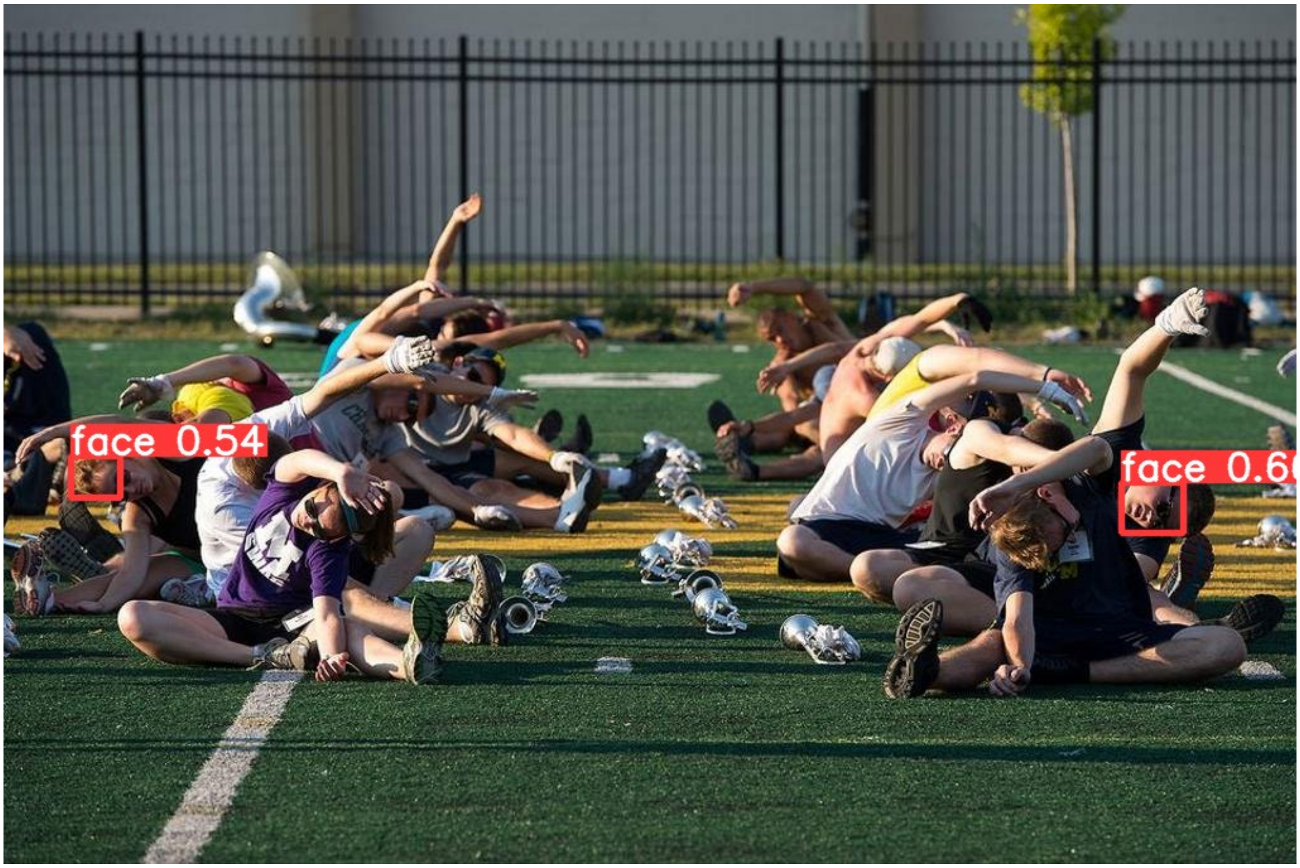
In the “INPUT” image, a group of people is seen sitting around a fire indoors and ambient conditions are quite dark. In the “OUTPUT” image, detecting faces with bounding boxes and confidence scores. After filtering, most of the adult face can be determined, but the face of children cannot be determined. The reason may be due to the light of adults more, or due to the difficulty of the young child's face, or the dataset is missing a faces children.

The Outside environment with afternoon



The second scenario, the image depicts a lively scene during an evening or afternoon parade with a large float adorned with golden seahorses is the central focus. Within the float, a person is enclosed in a clear bubble, surrounded by decorations resembling waves. Spectators, including both children and adults, are watching the parade. The background features illuminated buildings, creating an enchanting atmosphere. The result of after runned model is not considered a face, so it remains unmarked by the face detection model which people in the front could not be determined by turning their backs so they could not determine the face. The girl who looks like a mermaid cannot identify her face due to her hair and poor light. The faces of the person in the back cannot be determined due to poor light, due to the angle, the model can not determine the face and due to far away, so the ratio of the face is small, so it is difficult to determine.

The outside environment with morning and full-light



Overall, The two people on the left and the right have decided the model to identify the face, but the left person only identified half of the face because he was wearing sunglasses and poor corner. The remaining people make the model unable to determine the face due to most sunglasses or partially covered or the face. People who are far away are not covered by their faces, but the percentage of small face and poor light should not be determined

IV. Future works and Conclusion

This work presents the results of initial efforts to develop a low-cost computer vision-based system as a first step in human face recognition. Overall, our system can detect multiple faces in a photo or frame with relatively high accuracy. However, our system also faces some challenges. Here are some possible future directions of work that could help address these issues: First, with low-resolution photos as well as when the face is covered a lot, the model cannot meet expectations. Therefore, to improve this, we plan to use more data sets containing more diverse information, such as unclear photos, and also go back to the pre-processing step. Eliminate poor quality photos. Second, we want to deploy a system that can identify faces in real time through security cameras, phones, computers,... to get closer to our original desire. is to assist in security, confidentiality as well as in taking attendance at the school. Finally, to improve the accuracy of our object detection models, we try to balance the distribution of classes in our dataset by including more object instances. We can also add more layers to help our system recognize more internal objects. Furthermore, with advanced machine learning algorithms, we can design and test our system in crowded environments where faces are more difficult to identify.

V. References

[1] T. W. Nick Barney, "Face Detection," https://www.techtarget.com/searchenterpriseai/definition/face-detection.

[2] H. Jiang and E. Learned-Miller, "Face Detection with the Faster R-CNN," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, 2017, pp. 650–657. doi: 10.1109/FG.2017.82.

[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788. doi: 10.1109/CVPR.2016.91.

[4] S. Tang, S. Zhang, and Y. Fang, "HIC-YOLOv5: Improved YOLOv5 For Small Object Detection." 2023.

[5] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4 Optimal Speed and Accuracy of Object Detection." 2020.

[6] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang, "UnitBox: An Advanced Object Detection Network," in *Proceedings of the 24th ACM international conference on Multimedia*, in MM '16. ACM, Oct. 2016. doi: 10.1145/2964284.2967274.

[7] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019.

[8] S. Yang, P. Luo, C. C. Loy, and X. Tang, "WIDER FACE: A Face Detection Benchmark," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[9] Kukil, "Mean Average Precision (mAP) in Object Detection," https://learnopencv.com/mean-average-precision-map-object-detection-model-evaluation-metric/.

[10] John Hoang, "How to Evaluate Computer Vision Models with Confusion Matrix," Datature.