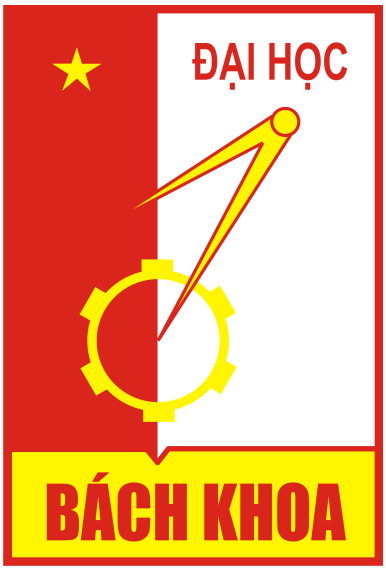
HA NOI UNIVERSITY OF SCIENCE AND TECHNOLOGY
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY
FACULTY OF COMPUTER SCIENCE

o0o



PROJECT REPORT:

Face Recognition

Course: Introduction to Deep Learning - IT3320E
Class Code: 144109
Supervisor: Associate Prof Nguyen Hung Son
Member of our group:

Trinh Hoang Giang	20214893	giang.th214893@sis.hust.edu.vn
Ho Ngoc Anh	20214877	anh.hn214877@sis.hust.edu.vn
Hoang Thanh Dat	20214899	dat.ht214889@sis.hust.edu.vn
Lang Van Quy	20214928	quy.lv214928@sis.hust.edu.vn
Vu Lam Anh	20214876	anh.vl214876@sis.hust.edu.vn

Ha Noi 2023

Contents

1 Introduction

In this project, we want to dive into facial analysis, combining cutting-edge technologies for face detection, age estimation, gender classification, and facial emotion recognition. Our system harnesses advanced machine learning models to not only precisely locate faces but also estimate the age and classify the gender of individuals in real-time. Going beyond demographic insights, our deep learning techniques enable the recognition and interpretation of a spectrum of facial emotions. With applications ranging from personalized user experiences and targeted marketing strategies to enhanced security and surveillance systems, our project give a good approach to decoding human expressions for a myriad of practical applications.

2 Dataset

2.1 Human Faces (Object Detection)[6]

The dataset includes 2204 photos with 3350 faces. A diversified collection of human facial photos representing various ethnicities, age groups, and profiles, with the goal of producing an unbiased dataset including coordinates of facial areas suitable for training object detection algorithms.

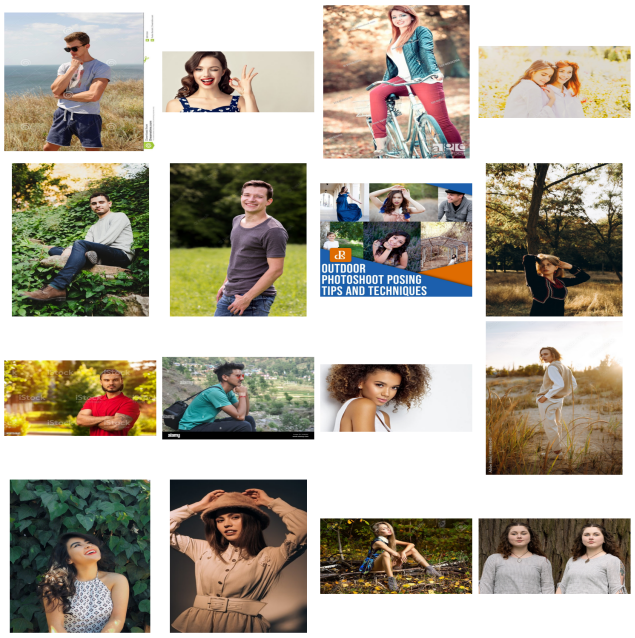


Figure 1: Human Faces datasets

2.2 FER2013[7]

The open source dataset **FER2013.csv**, created for a project by PierreLuc Carrier and Aaron Courville, was publicly shared in the Kaggle competition (2013).

The dataset consists of 35,887 gray images of 48x48 resolution. Kaggle has divided into 28,709 training images, 3589 public test images and 3589 private test images. Each image is assigned one of seven different facial emotion classes, all labeled from 0 – 7 (0 = Anger, 1 = Disgust, 2 = Fear, 3 = Happy, 4 = Sad, 5 = Surprise, 6 = Neutral). Including 8,989 'Happy' images, 6,077 'Sad' images, 6,198 'Neutral' images, 4002 'Surprise' images, 5121 'Scared' images, 547 'Disgust' images and 4593 'Angry' images.

Figure 2: FER datasets

2.3 UTK-Face[8]

UTKFace dataset is a large-scale face dataset with long age span (range from 0 to 116 years old). The dataset consists of over 20,000 face images with annotations of age, gender, and ethnicity. The images cover large variation in pose, facial expression, illumination, occlusion, resolution, etc. Each image is assigned gender label (0 for Male and 1 for Female) and age label as integer number from 0 to 116. This dataset could be used on a variety of tasks, e.g., face detection, age estimation, age progression/regression, landmark localization, etc. Some sample images are shown as following

Figure 3: UTK-Face datasets

3 Data Preprocessing

3.1 Age Gender Recognition

Data Augmentation

The `train_datagen` configuration is a key component of our age and gender recognition project. This generator is designed to augment and diversify the training dataset, promoting better generalization and robustness of the model. The following augmentation techniques were employed:

- Rescaling: Images were rescaled to a range of [0, 1] to ensure consistent pixel values.

- Validation Split: A validation split of 20% was applied to the dataset, facilitating model evaluation during training.
- Rotation Range: Random rotation up to 20 degrees was applied to introduce variability and improve the model's ability to handle rotated faces.
- Width and Height Shift Range: Horizontal and vertical shifts (up to 10% of the image width and height, respectively) were applied to simulate variations in facial positioning.
- Shear Range: A shear range of 0.5 was used to introduce deformations in the facial structure, enhancing the model's ability to recognize diverse facial expressions.
- Zoom Range: Random zooming (up to 10%) was applied to capture variations in facial scales.
- Brightness Range: Adjustments to brightness were made within the range of 0.5 to 1.5 to account for varying lighting conditions.
- Horizontal Flip: Random horizontal flips were performed to augment the dataset with mirror images.
- Vertical Flip: Vertical flips were disabled to maintain the natural orientation of facial features.
- Fill Mode: The 'nearest' fill mode was used to handle newly created pixels resulting from transformations.

After applying Image Augmentation techniques we obtain the results below:

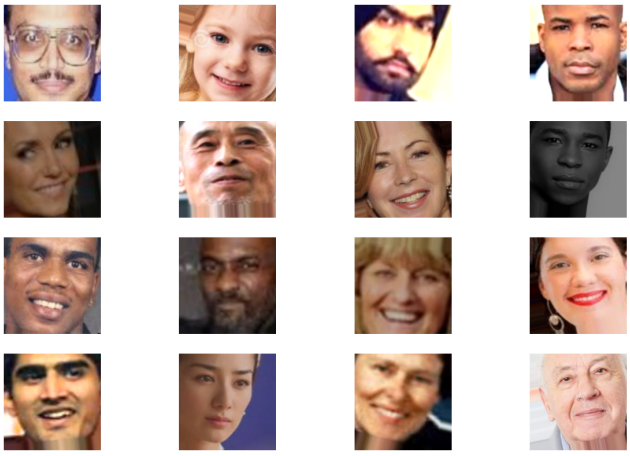


Figure 4: Image Augmentation Samples

3.2 Facial Emote Recognition

In this datasets, although it has quite large number of images: 35,887 images. But each image is only gray images of 48x48 resolution. Therefore we augment datasets with some techniques as follow (Those techniques below we implement was inspired by the paper [5]):

1. RandomResizedCrop:
   We scale the image in height and width by 0.8 and 1.2 then random crop image and resize back to 48x48.

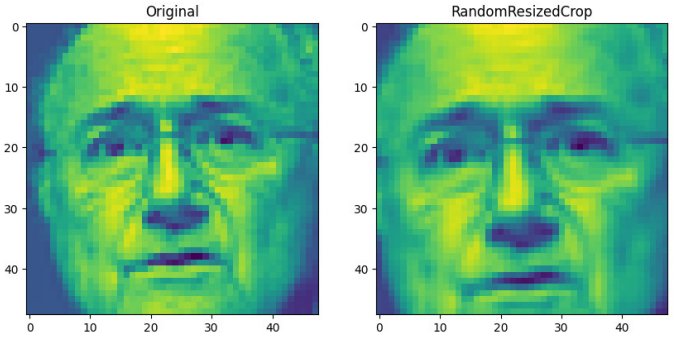


Figure 5: RandomResizedCrop transform

2. ColorJitter:
   We change brightness, contrast, saturation of image with probability p = 0.5

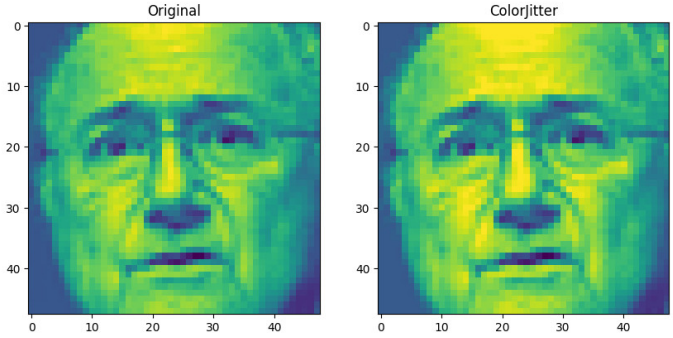


Figure 6: ColorJitter transform

3. RandomAffine:
   We randomly move the image to right or to the left with length = 0.2 * width. Similarly,We randomly move the image up or down with length = 0.2 * height.

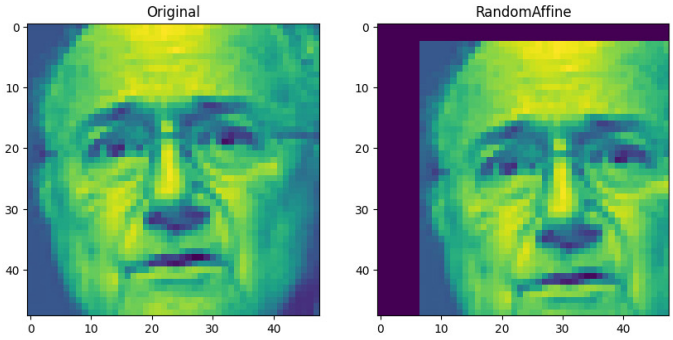


Figure 7: RandomAffine transform

4. RandomHorizontalFlip:
   We randomly flip the image horizontally with probability p = 0.5.

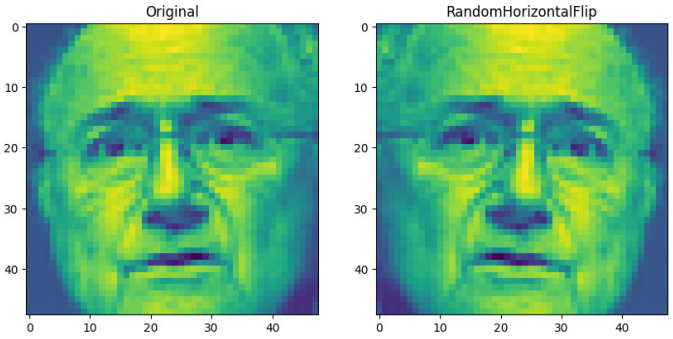


Figure 8: RandomHorizontalFlip transform

5. RandomRotation:
   We randomly rotate image 10 degrees to right or left with probabilty p = 0.5.

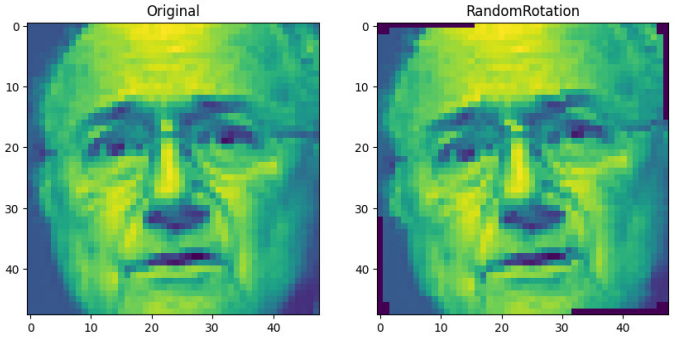


Figure 9: RandomRotation transform

6. FiveCrop:
   We randomly crop original image into five images with 40x40 pixels so now we have augmented data five times. We stack them together then use it to train the model later.

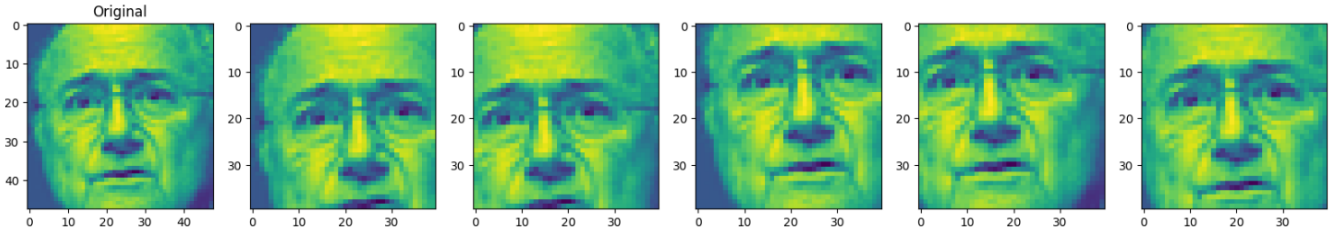


Figure 10: FiveCrop transform

7. Normalize: We normalize the image by dividing the image by 255.

4 Modelling

4.1 VGG

The VGG (Visual Geometry Group) model represents a pivotal advancement in the field of deep learning, particularly in the domain of computer vision. Developed by the Visual Geometry Group at the University of Oxford, VGG gained prominence for its simplicity and remarkable performance in image recognition tasks. Introduced in 2014, the VGG architecture is characterized by its straightforward design, featuring a series of convolutional layers with small receptive fields and max-pooling layers. The model's depth, with configurations such as VGG16 and VGG19, contributed to its effectiveness in capturing intricate features within images, making it a key player in various computer vision applications, including object detection and image classification.

4.1.1 Model Architecture

VGG16 consists of 16 weight layers, including 13 convolutional layers and 3 fully connected layers.

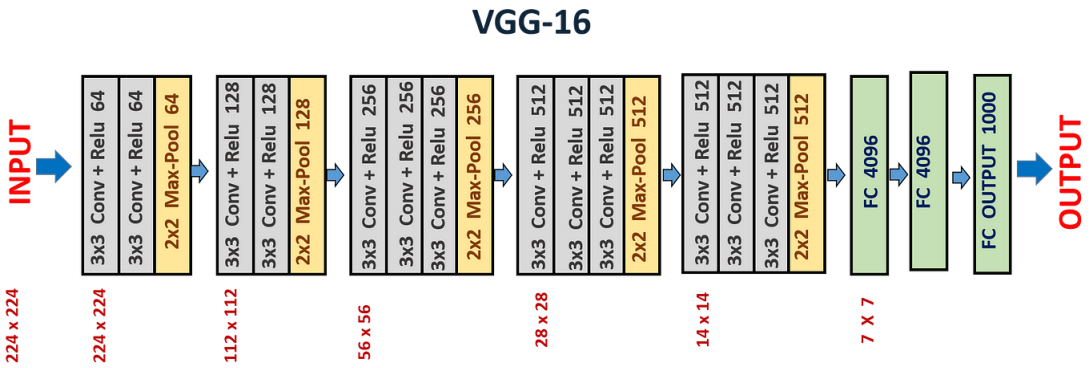


Figure 11: VGG16 Architecture

4.1.2 Activation Function

ReLU (Rectified Linear Unit) activation functions are used after each convolutional layer to introduce non-linearity.

4.1.3 Training

- Loss Function: CrossEntropy for gender prediction and MAE for age prediction
- Optimizer: Adam

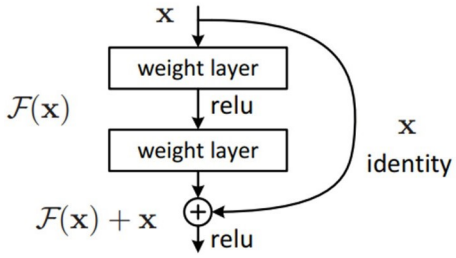
4.2 ResNet

The common problem with model in deep learning and also other CNN model like VGG is "vanishing gradient" problem. A vanishing gradient occurs during backpropagation. When the neural network training algorithm tries to find weights that bring the loss function to a minimal value, if there are too many layers, the gradient becomes very small until it disappears, and optimization cannot continue.

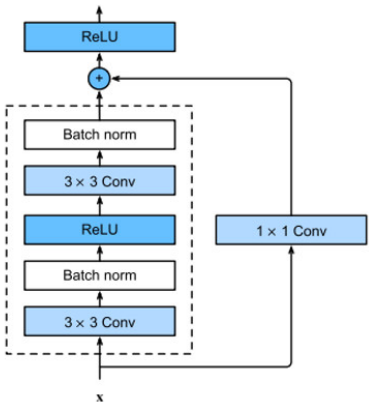
In this section, we introduce the Residual Network (ResNet)[1] was published in 2016, is a Convolutional Neural Network (CNN) architecture that overcame the "vanishing gradient" problem, making it possible to construct networks with up to thousands of convolutional layers, which outperform shallower networks.

4.2.1 Residual Block

The idea of ResNet is inspired by the technique is called skip connections. The skip connection connects activations of a layer to activations of further layers by skipping some layers in between (as image (a) below).



(a) Skip connection



(b) Residual Block.Source [2]

Those implementation of idea practically in ResNet is called Residual Block. we can see here the input x go through the Residual Block with two 3x3 convolutional layers and at the end, it add the outputs with the original input which has been go through 1x1 convolution. These help we overcome the gradient vanishing and possible to explore additional parts of the feature space which would have been missed in a shallow convolutional network architecture.

4.2.2 ResNet Architecture

ResNet are made by stacking these residual blocks together. We can construct a ResNet-18 and ResNet-34 with by stacking the residual blocks in section above together. Here is the image of ResNet-18 architecture:

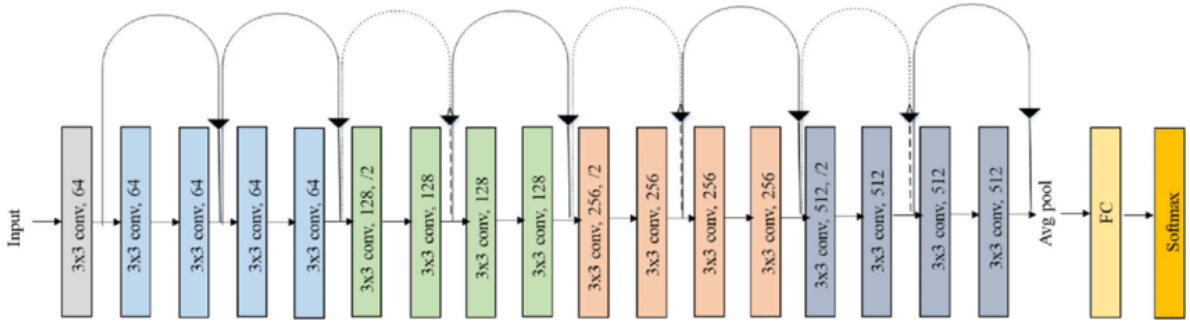


Figure 13: ResNet 18 Architecture

Moreover, we can go deeper maybe ResNet-50, ResNet-101 or ResNet-152 with Bottleneck block which consists more convolutional layers.

4.3 YOLOv5

4.3.1 Model Architecture

YOLOv5's architecture is divided into three sections: Backbone, Neck, Head

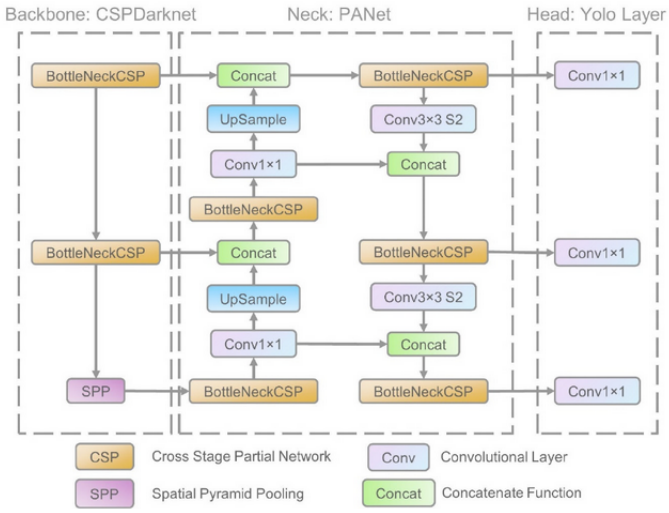


Figure 14: YOLOv5 Architecture.Source [3]

A. Backbone This is the main body of the network. For YOLOv5, the backbone is designed using the CSP-Darknet53 structure, a modification of the Darknet architecture using CSPDenseBlock instead of DenseBlock.

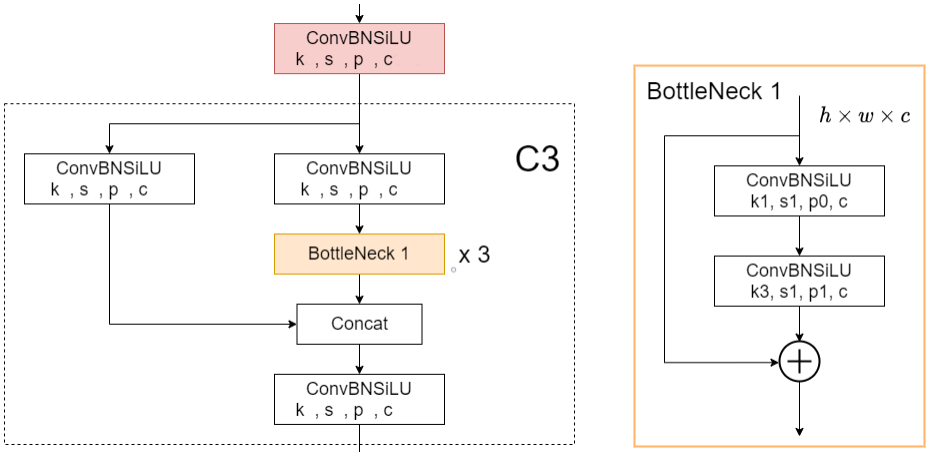


Figure 15: CSPDenseBlock [3]

B Neck This part connects the backbone and the head. In YOLOv5, SPPF and CSP-PAN structures are utilized.

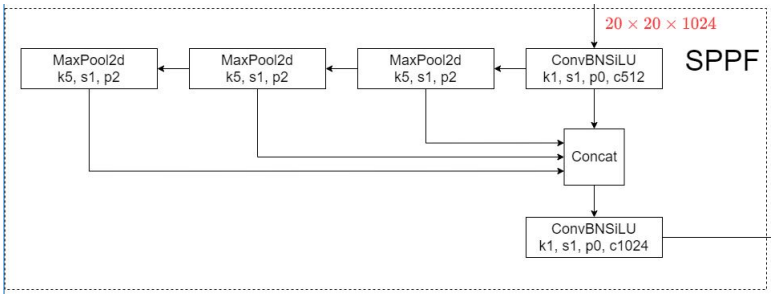


Figure 16: SPPF Architecture [3]

C. Head YOLOv5 have 3 detection layers in head: 80 x 80x 256 for detect small size object, 40 x 40 x 512 for detect medium size object, and 20 x 20 x 512 for detect large size object.

4.3.2 Data Augmentation

YOLOv5 applies several augmentation techniques including:

- Mosaic Augmentation: An image processing technique that combines four training images into one in ways that encourage object detection models to better handle various object scales and translations.

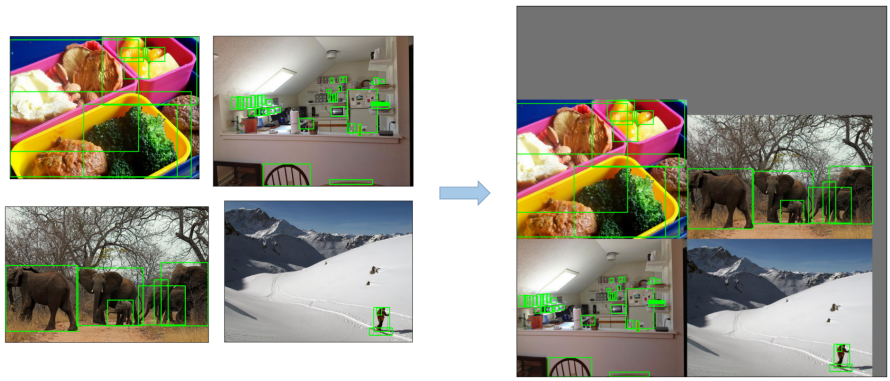


Figure 17: Mosaic Augmentation [4]

- Copy-Paste Augmentation: An innovative data augmentation method that copies random patches from an image and pastes them onto another randomly chosen image, effectively generating a new training sample.

Figure 18: Copy-Paste Augmentation [4]

- Random Affine Transformations: This includes random rotation, scaling, translation, and shearing of the images.

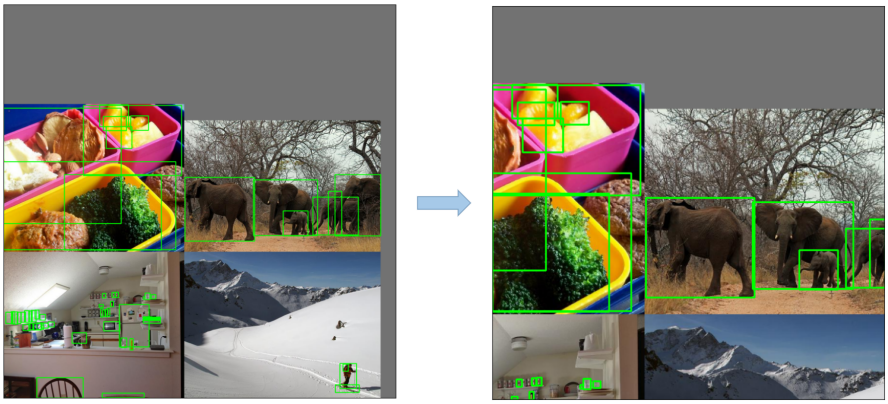


Figure 19: Random Affine Transformations [4]

- MixUp Augmentation: A method that creates composite images by taking a linear combination of two images and their associated labels.

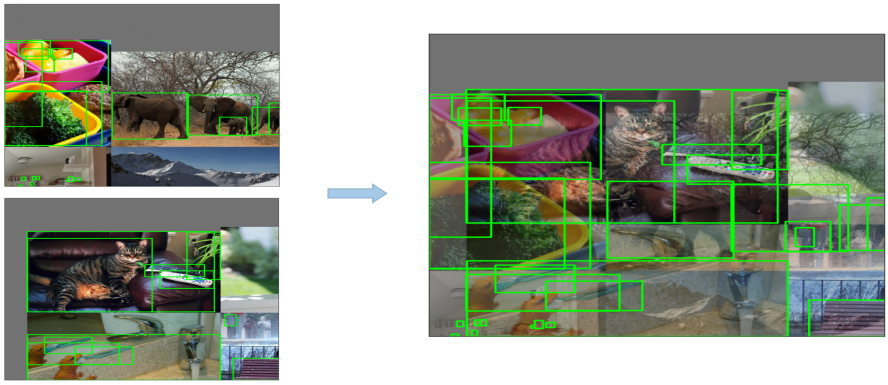


Figure 20: MixUp Augmentation [4]

- Albumentations: A powerful library for image augmenting that supports a wide variety of augmentation techniques.

- HSV Augmentation: Random changes to the Hue, Saturation, and Value of the images.

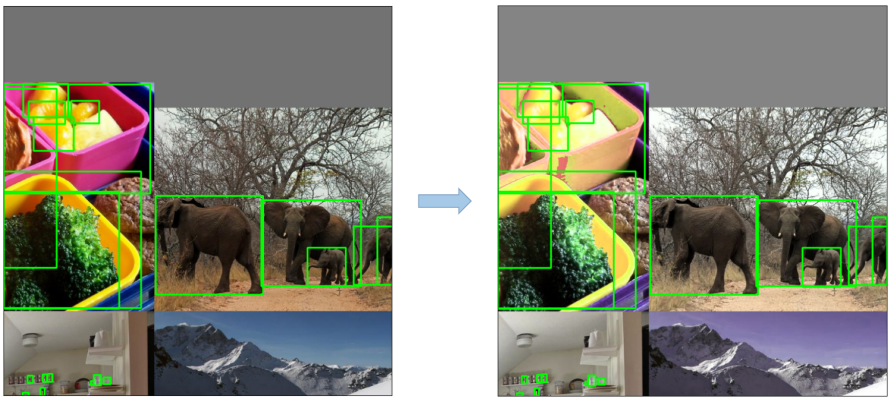


Figure 21: HSV Augmentation [4]

- Random Horizontal Flip: An augmentation method that randomly flips images horizontally.

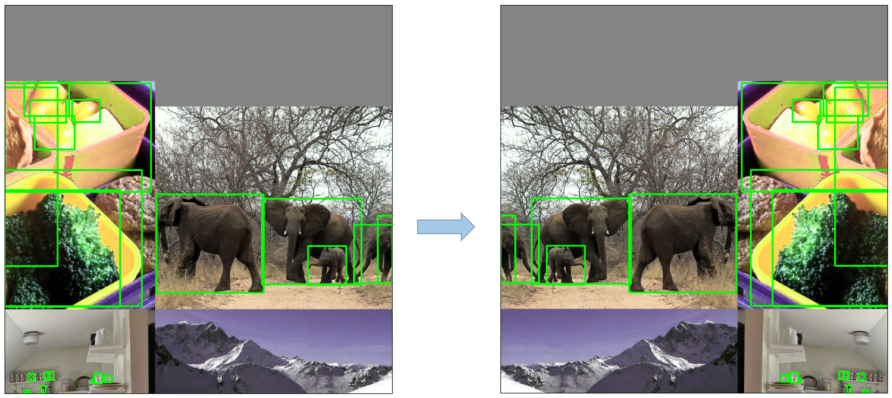


Figure 22: Random Horizontal Flip [4]

4.3.3 Training

YOLOv5 applies several sophisticated training strategies to enhance the model's performance. They include:

- Multiscale Training: The input images are randomly rescaled within a range of 0.5 to 1.5 times their original size during the training process.
- AutoAnchor: This strategy optimizes the prior anchor boxes to match the statistical characteristics of the ground truth boxes in your custom data.
- Warmup and Cosine LR Scheduler: A method to adjust the learning rate to enhance model performance.
- Exponential Moving Average (EMA): A strategy that uses the average of parameters over past steps to stabilize the training process and reduce generalization error.
- Mixed Precision Training: A method to perform operations in half-precision format, reducing memory usage and enhancing computational speed.

- Hyperparameter Evolution: A strategy to automatically tune hyperparameters to achieve optimal performance.

4.3.4 Loss Function

YOLOv5 computes the classes and objectness losses using Binary Cross Entropy, and the location loss using Complete Intersection over Union (CIoU).

$$Loss = \lambda_1 L_{cls} + \lambda_2 L_{obj} + \lambda_3 L_{loc} \tag{1}$$

$$L_{obj} = \sum_{i=1}^{S^2} \sum_{j=1}^{B} 1_{ij}^{obj} \sum_{k=1}^{C} BCE(y_k, \hat{y}_k) \tag{2}$$

The objectness loss is weighted differently for small, medium, and large objects.

$$L_{obj} = 4L_{obj}^{small} + 1L_{obj}^{medium} + 4L_{obj}^{large} \tag{3}$$

CIoU loss can be computed as follow:

$$L_{CIoU} = 1 - IoU + \frac{Euclide_Distance(b, b^{gt})}{c^2} + \alpha v \tag{4}$$

where

1. b, b^{gt} are center point of the predicted bounding box and ground truth bounding box.
2. c represents the diagonal distance of the minimum closure area that contains both the prediction box and the ground truth.
3. α is the positive trade-off parameter.
$$\alpha = \frac{v}{1 - IoU + v} \tag{5}$$
4. v measures the consistency of aspect ratio.
$$v = \frac{4}{\pi^2} \left(\arctan\left(\frac{\omega^{gt}}{h^{gt}}\right) - \arctan\left(\frac{\omega}{h}\right) \right)^2 \tag{6}$$

5 Experiment and Results

5.1 Age Gender Recognition

Training configuration:

- Size of input image : 200x200x3
- Number channel of input image: 3
- Number outputs: 2 labels for Gender and an integer for Age
- Learning rate: 0.0003
- Batch size: 32

- Number of Epochs: 150
- Optimizer: Adam
- Loss Function: MAE for Age and Binary CrossEntropy for Gender

The table below presents a comparison of validation losses among different models:

Model	Age Prediction	Gender Prediction
ResNet-18	7.375	0.2114
ResNet-34	6.221	0.2161
ResNet-50	6.681	0.2291
ResNet-101	7.318	0.1939
ResNet-152	7.357	0.1957
VGG-16	8.9967	0.4808
VGG-19	8.1687	0.4208

Based on the findings presented in the aforementioned table, we have concluded that ResNet-34 is the optimal model, striking a favorable balance between validation losses in both Age Prediction and Gender Prediction tasks.

Here are a few instances showcasing our prediction results:

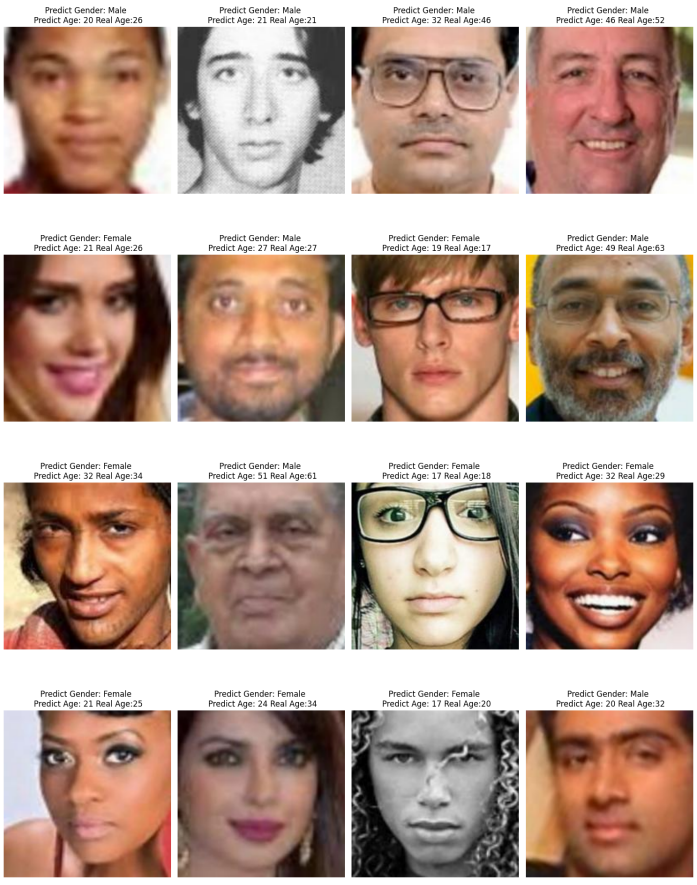


Figure 23: Age Gender Predict Results

5.2 Facial Emote Recognition

With this problem, we setup the training configuration as follow:

- Size of input image : 40x40x1
- Number channel of input image: 1
- Number outputs: 7
- Learning rate: 0.001
- Batch size: 64 (We crop the 48x48 pixels image into 5 crops 40x40 image. Therefore, we have batch size 64x5 = 320 of 40x40 image)
- Number of Epochs: 150
- Optimizer: Adam
- Weight decay : 1e-4
- Scheduler : ReduceLROnPlateau. We wil reduce the learning rate by 0.5 if there are no improvement of the accuracy in validation datasets after 5 epochs.
- loss function: CrossEntropyLoss

Below is the table show the test accuracy of each model compare to each others:

Model	Test Accuracy(%)
ResNet-18	71.25
ResNet-34	70.92
ResNet-50	70.78
ResNet-101	70.85
ResNet-152	71.12
VGG-16	65.17
VGG-19	63.25

From the table above, the model ResNet-18 has the best performance. Then we can use it to predict the Facial Emote Recognition.

Here are some example, we try to predict the facial emotion of image in the test datasets:

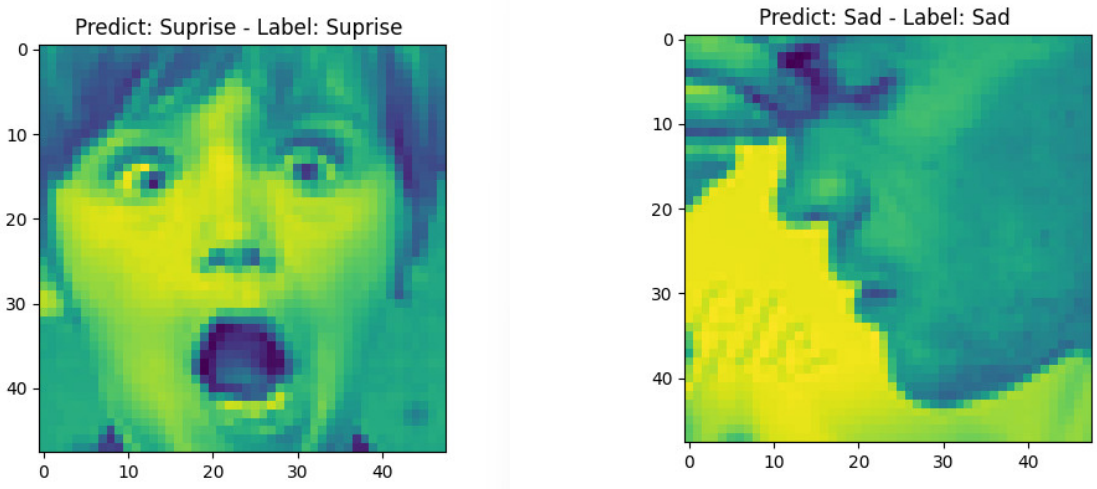


Figure 24: Predict the facial emotion for some example images in the test datasets

5.3 Face Detection

We use YOLOv5 model from ultralytics[4]. Choosing IoU=0.5, Confidence Threshold=0.001, after training 100 epochs, we have the performance of YOLOv5 in Face Detection task showed in the below table

	Precision	Recall	mAP50	mAP50-95
Train	0.872	0.969	0.964	0.81
Valid	0.872	0.98	0.965	0.816

With excellent precision, recall, and mAP scores, the model appears to perform well on both the training and validation sets. The fact that the training and validation sets performed similarly shows that the model did not overfit to the training data. The mAP50 and mAP50-95 scores show that the model is stable across different confidence threshold and IoU ranges.

(a) Image with actual bounding box

(b) Image with predicted bounding box

Figure 25: Comparision between ground truth and predicted bounding box

6 Conclusion and Future work

In the task facial emotion recognition, we build and train the family of the ResNet model and after comparing those models to each other, we find that the ResNet-18 has the best performances with 71,25% in the test datasets. The model recognized emotions equally or slightly better than humans did (65% $\pm$ 5%). For the future work, we want to try some stronger model like: Residual Masking Network, PAtt-Lite, EmoNeXt,... for better performance.

In Age Gender Recognition task, ResNet-34 stands out as the preferred model for Age and Gender Prediction, showcasing a balanced performance with a Mean Absolute Error (MAE) of 6.221 for age prediction and a Binary Crossentropy of 0.2161 for gender prediction. Its dual-task efficiency makes it a promising choice for applications requiring simultaneous age and gender recognition.

In Face Detection task, the YOLOv5 model demonstrates outstanding performance on both the training and validation sets, as evidenced by excellent precision, recall, and mAP scores. The similarity in performance between the two sets suggests that the model successfully avoided overfitting to the training data, showcasing its generalization capabilities.

7 Work Distribution

1. Proposal: Dat
2. Implement VGG for Age Gender Recognition: Quý
3. Implement VGG for Facial Emote Recognition: Ánh
4. Implement Resnet for Age Gender Recognition: Dat
5. Implement Resnet for Facial Emote Recognition: Lâm Anh
6. Implement YOLOv5 for Face Detection: Giang
7. Evaluate VGG for Age Gender Recognition: Quý
8. Evaluate VGG for Facial Emote Recognition: Ánh
9. Evaluate Resnet for Age Gender Recognition: Dat
10. Evaluate Resnet for Facial Emote Recognition: Lâm Anh
11. Evaluate YOLOv5 for Face Detection: Giang
12. Report:
    (a) Introduction: Lâm Anh
    (b) Human Faces (Object Detection) Dataset: Giang
    (c) FER2013 Dataset: Ánh
    (d) UTK-Face Dataset: Dat
    (e) Data Preprocessing:
        i. Age Gender Recognition: Dat

ii. Facial Emote Recognition: Lâm Anh

(f) VGG Model: Quý

(g) ResNet: Lâm Anh

(h) YOLOv5: Giang

(i) Experiment and Result:

i. Age Gender Recognition: Dat, Quý

ii. Facial Emote Recognition: Lâm Anh, Ánh

iii. Face Detection: Giang

13. Slide: Giang, Ánh, Lâm Anh, Quý, Dat

References

[1] He, Kaiming and Zhang, Xiangyu and Ren, Shaoqing and Sun, Jian (2016): Deep Residual Learning for Image Recognition

[2] Zhang, Aston and Lipton, Zachary C and Li, Mu and Smola, Alexander J (2023): Dive into Deep Learning

[3] OpenGenus IQ: YOLO v5 model architecture [Explained]

[4] Ultralytics: YOLOv5 Architecture Summary

[5] Khaireddin, Yousif and Chen, Zhuofa(2021): Facial emotion recognition: State of the art performance on FER2013

[6] Human Faces (Object Detection) Dataset Kaggle: Human Faces Dataset

[7] FER2013 Datasets Kaggle: Facial Expression Recognition(FER)Challenge

[8] UTK-Face Dataset UTK-Face Dataset

[9] Wendong Gai, Yakun Liu, Jing Zhang & Gang Jing (2021) An improved TinyYOLOv3 for real-time object detection, Systems Science & Control Engineering, 9:1, 314-321, DOI: Complete Intersection over Union Loss