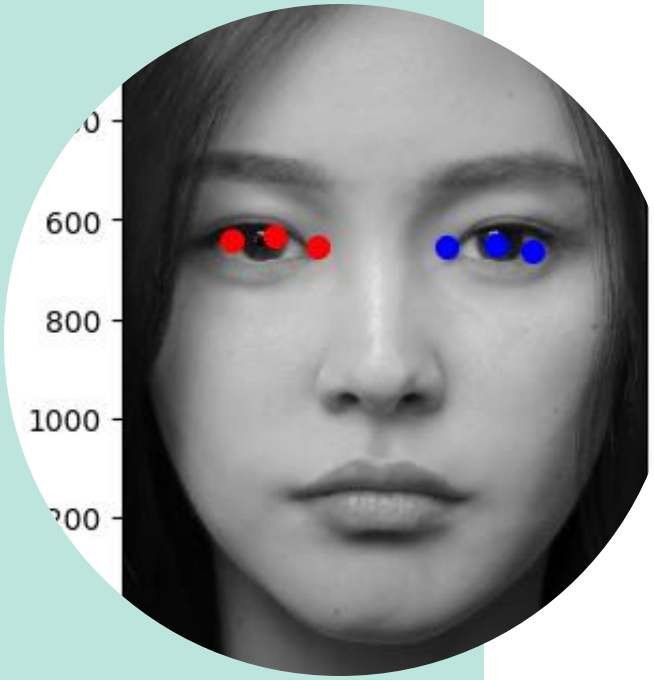
Università Ca'Foscari Venezia

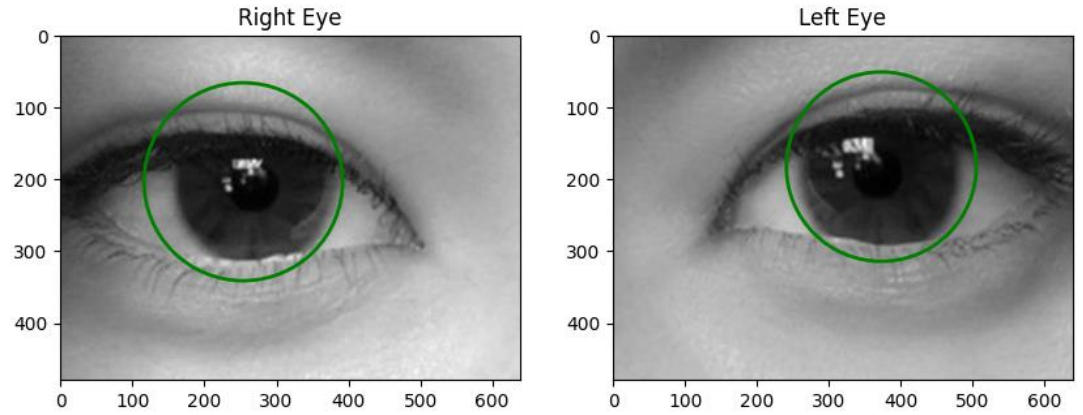
IMAGE AND VIDEO UNDERSTANDING PROJECT

TONETTO DAVIDE - 884585 - 2023/2024

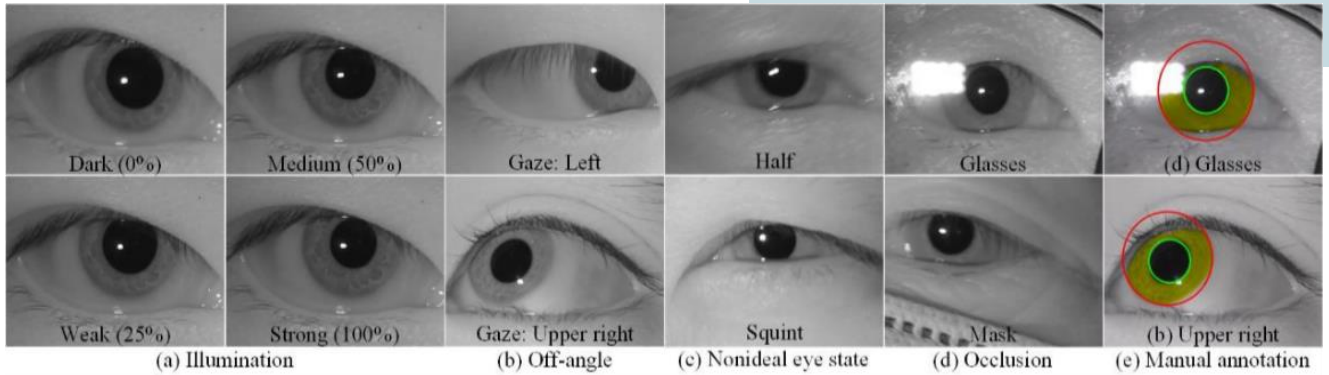


TASK

The task of the project is to develop a system that uses artificial intelligence to track the eyes in a given human face image.



DATASETS



TWO DATASETS ARE USED FOR THE PROJECT:

- CASIA IRIS DEGRADATION
- FACE IMAGES WITH MARKED LANDMARK POINTS

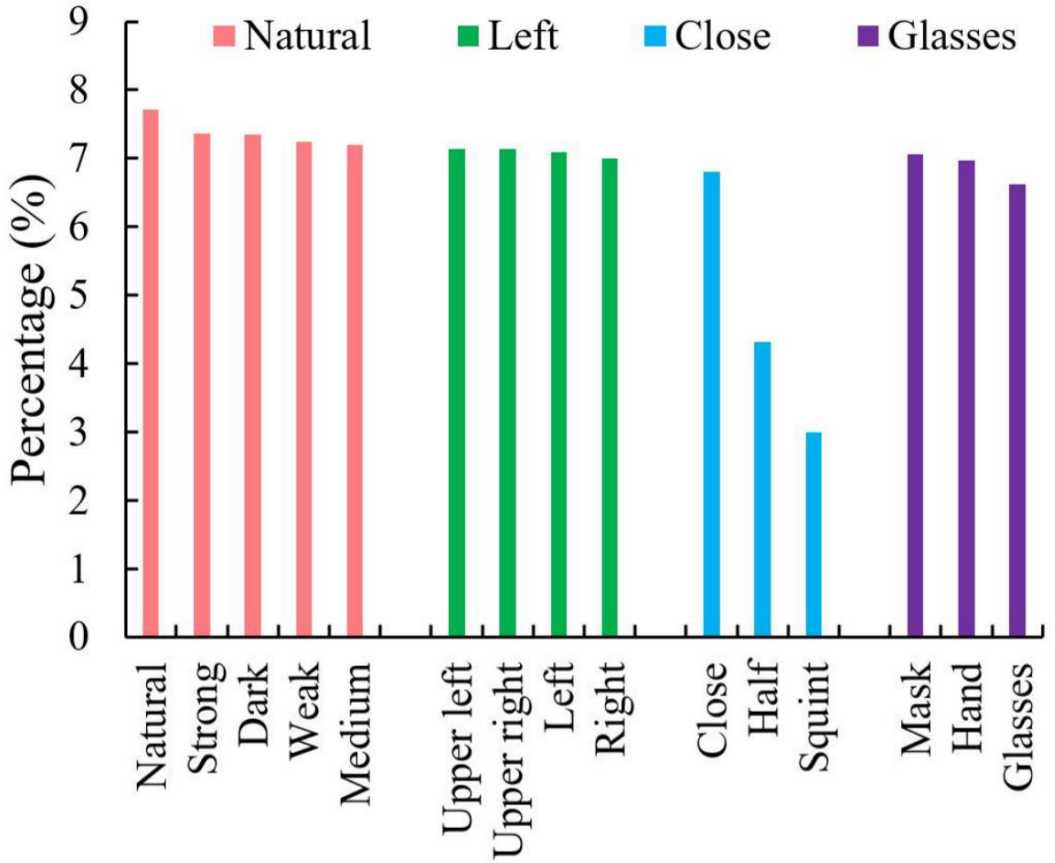
CASIA IRIS DATASET

This dataset is composed by:

- 36539 PNG images containing one eye each
- 32962 INI files containing the following localization parameters for the eye:
  - Pupil x and y coordinates and radius
  - Iris x and y coordinates and radius

The images are in grayscale and have the characteristics shown in the right picture.

Table 1. Details of the DV1 database.



FACE IMAGES WITH MARKED LANDMARK POINTS DATASET

This dataset is composed by:

- 7049 facial images
- Up to 15 keypoints marked on each image

For the given task only the keypoints for the eyes are used.

MODELS

Two main models are used sequentially to solve the task:

- **Eyes detection model**:
  - Given an image containing a human face, it finds the coordinates of the two eyes.
- **Iris tracker model**:
  - Given an image containing an eye it finds the coordinates of the pupil and the iris radius

EYES DETECTION MODEL

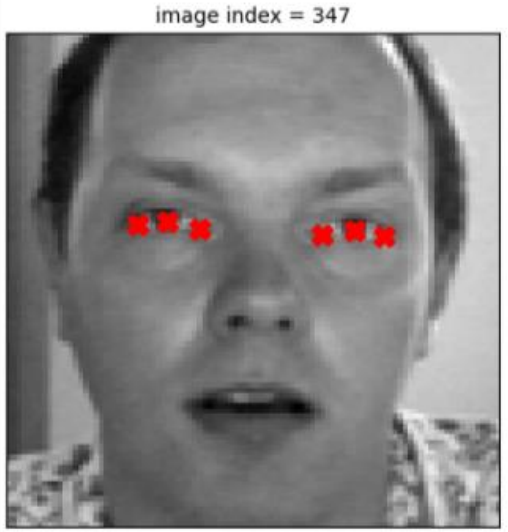
Let's introduce the model used to detect the eyes in a human face.

DATASET PREPARATION

For this step the "Face Images with Marked Landmark Points" dataset is used.

The dataset offers **three keypoints** for the eyes:

- **Inner corner**
- **Outer corner**
- **Center**



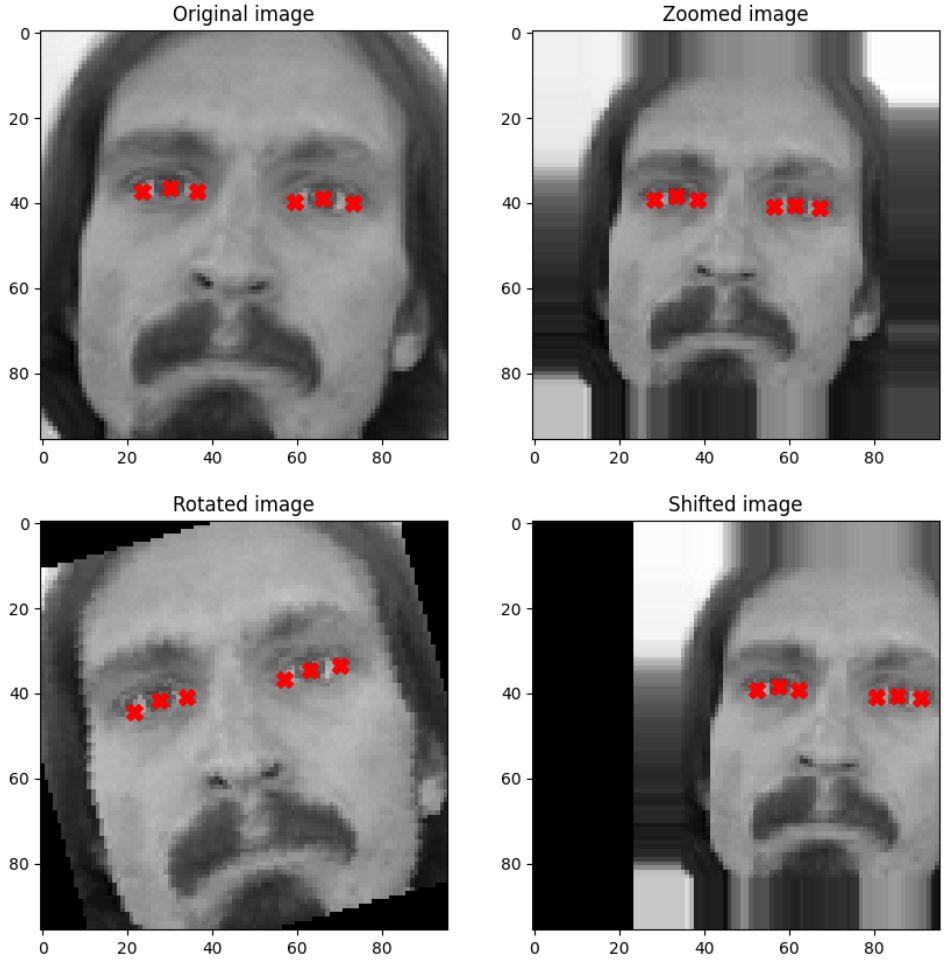
Since **not all the images have all three keypoints assigned**, each sample's missing values are **imputed** using the mean value from **n nearest neighbours** found in the training set.

DATA AUGMENTATION

To improve the accuracy and the generality of the model, some data augmentation is applied to a percentage of random images in the following ways:

- **Zoom out** - zoom out the image by a random factor
- **Rotation** - rotate the image by a random factor
- **Shift** - shift the image by a random factor in a random direction

At the end, the train set contains 13503 images, test set 4220 and validation set 3376.



ARCHITECTURE

Two different architectures of the network are proposed and compared:

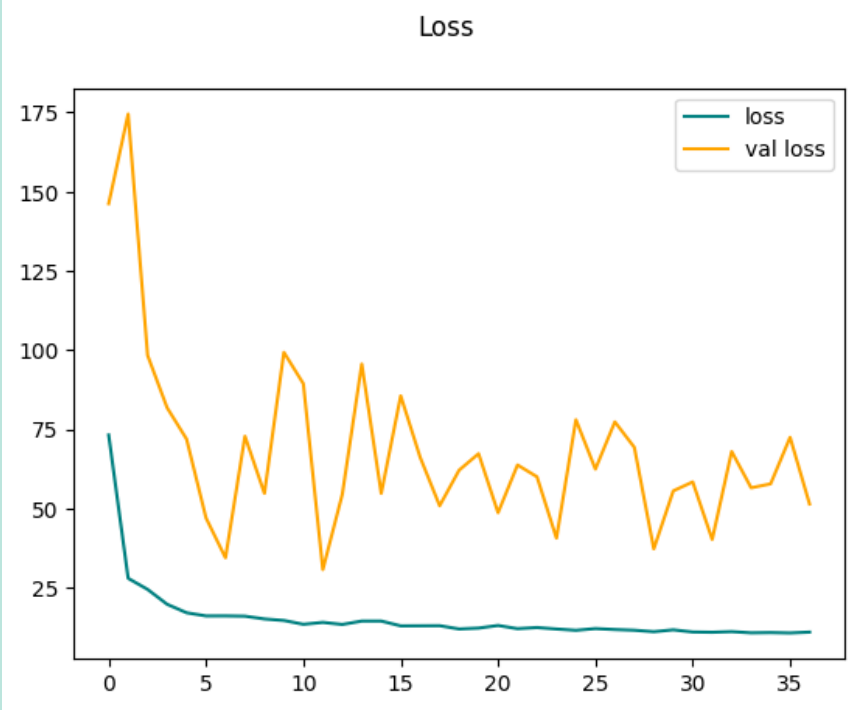
- **Convolutional Neural Network (CNN)**
- **Residual Network (ResNet)**

Both are made by a **first part** composed of Max Pooling 2D, Batch Normalization and 2D Convolution layers **used to extract features**, followed by some **fully connected layers** at the end.

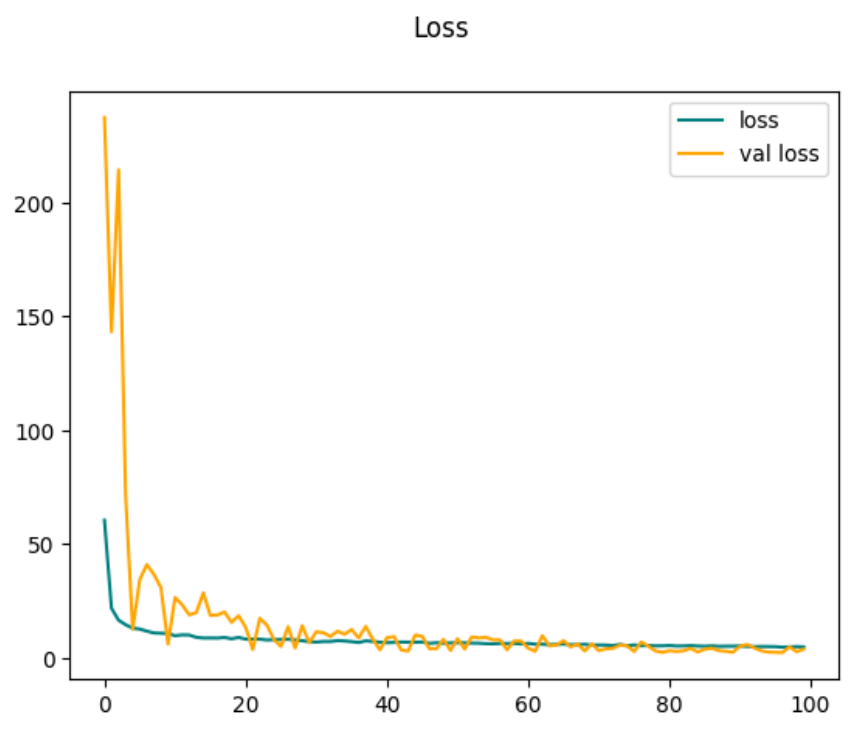
The **loss metric** used is **mean squared errors**. The **early stopping** technique is applied.

COMPARISON

CNN



RESNET

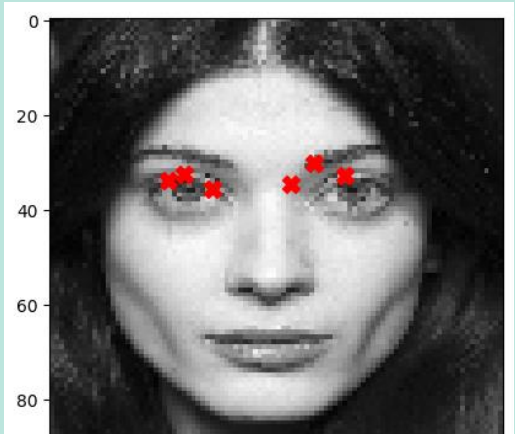


RESULTS

CNN

Test set score:

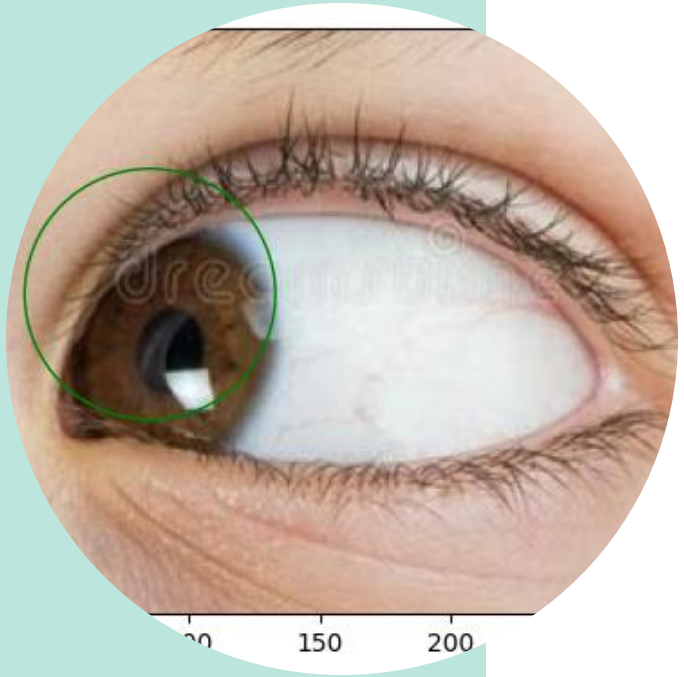
- **Loss:** 30.9289
- **Mean absolute error:** 4.6922



RESNET

Test set score:

- **Loss:** 2.2519
- **Mean absolute error:** 1.0830



EYE TRACKING MODEL

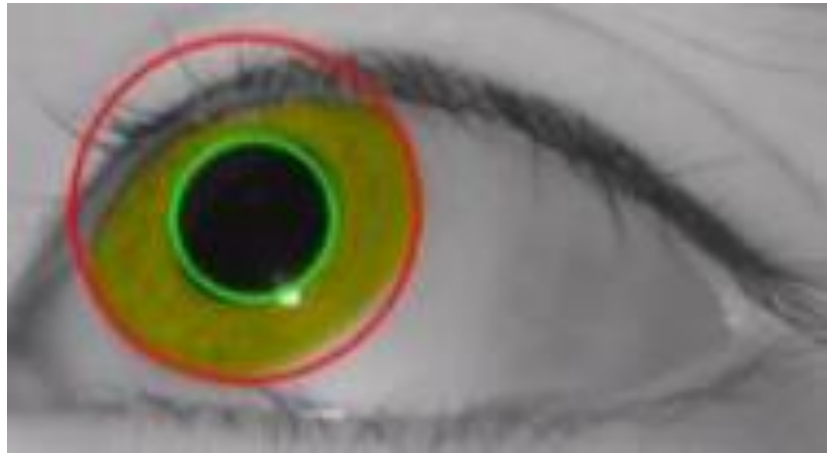
Let's introduce the model used to track the iris in a human eye.

DATASET PREPARATION

For this step the "CASIA Iris Degradation" dataset is used.

Also in this case, some **data augmentation** is introduced. At the end we have:

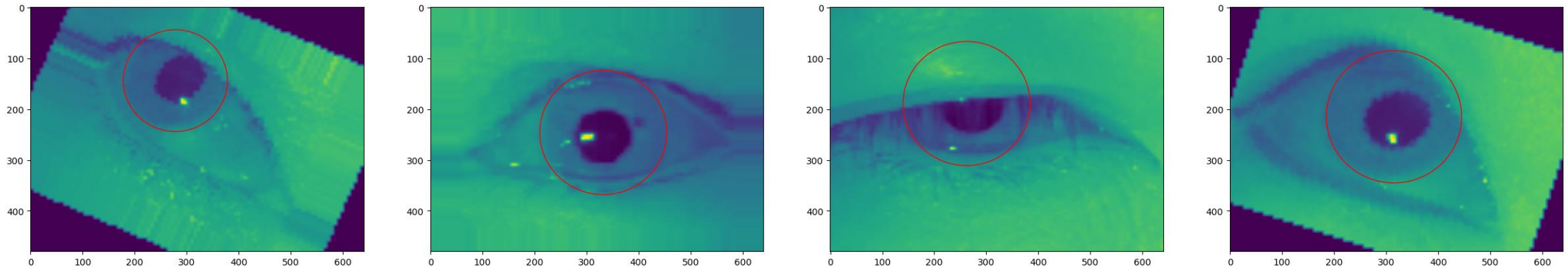
- **Train set** with 46577 images
- **Test set** with 9980 images
- **Validation set** with 9983 images



DATA AUGMENTATION

To improve the accuracy and the generality of the model, some data augmentation is applied to a percentage of random images in the following ways:

- **Zoom out** - zoom out the image by a random factor
- **Rotation** - rotate the image by a random factor
- **Shift** - shift the image by a random factor in a random direction



ARCHITECTURE

Also in this case, two different architecture of the network are proposed and compared:

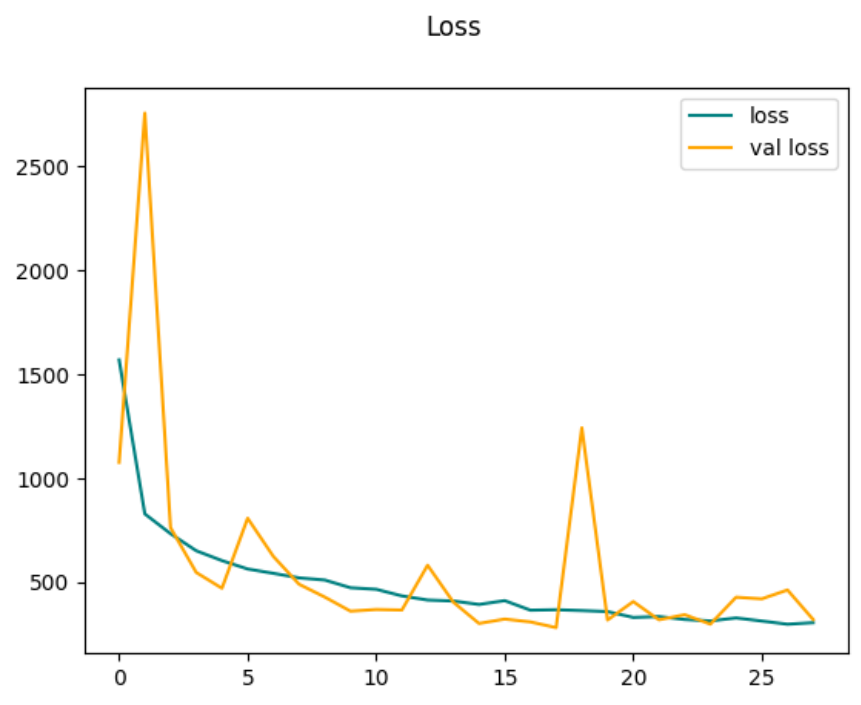
- **Convolutional Neural Network (CNN)**
- **Residual Network (ResNet)**

Both comprise Max Pooling 2D, Batch Normalization and 2D Convolution layers **used to extract features** with a **Reshape layer at the end** in order to obtain the three real values needed.

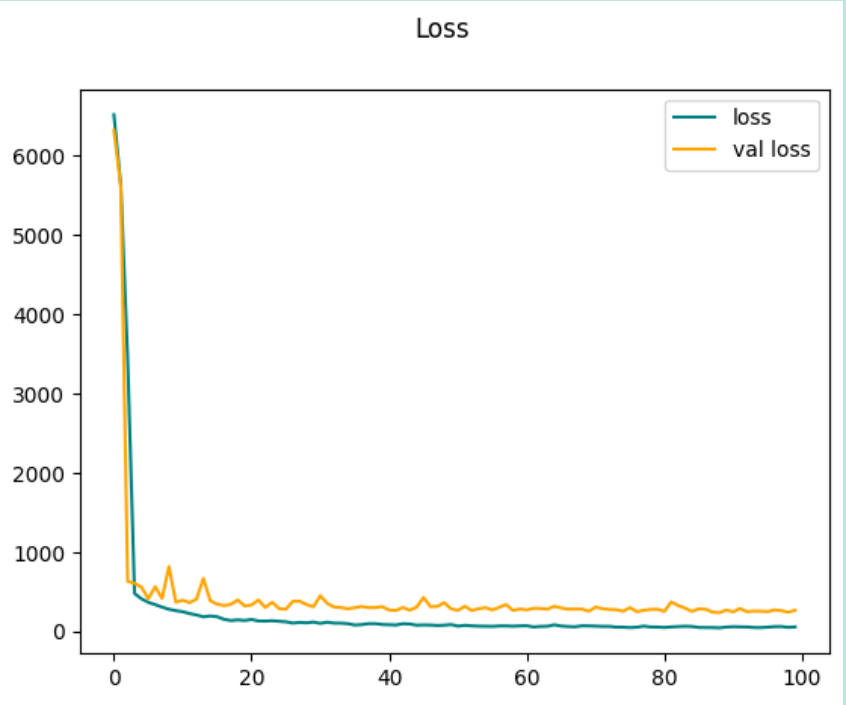
The **loss metric** used is **mean squared errors**. The **early stopping technique** is applied.

COMPARISON

CNN



RESNET

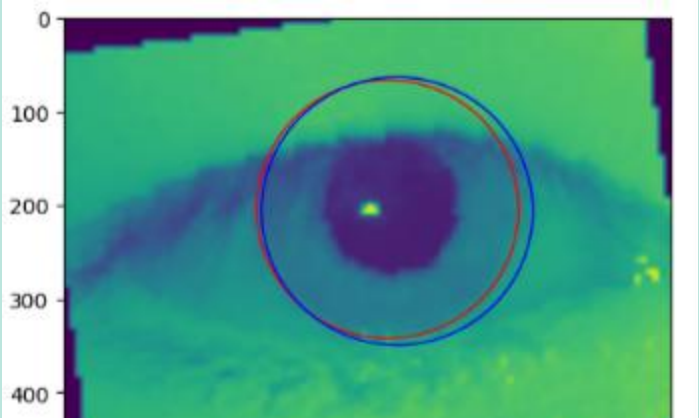


RESULTS

CNN

Test set score:

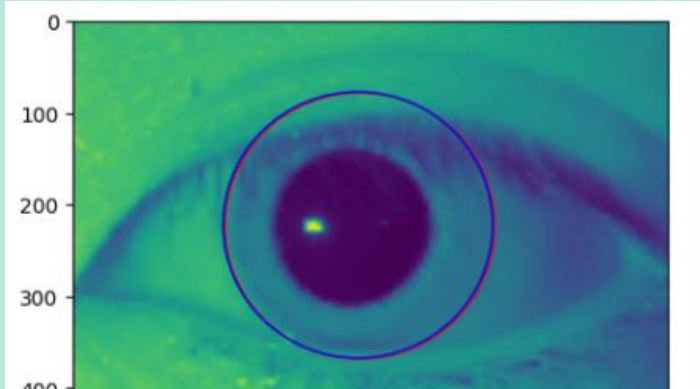
- **Loss:** 298.1567
- **Mean absolute error:** 6.9925



RESNET

Test set score:

- **Loss:** 283.4295
- **Mean absolute error:** 4.9103



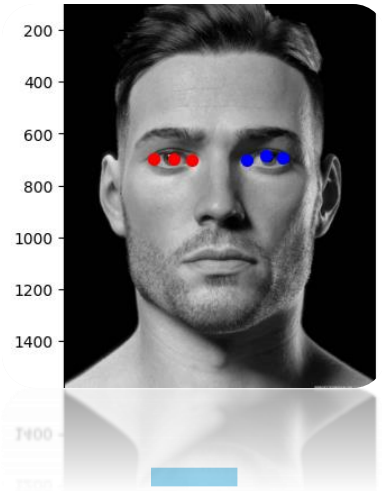
USAGE OF THE TWO MODELS

Given the **image of a single human face,** those operations are applied:

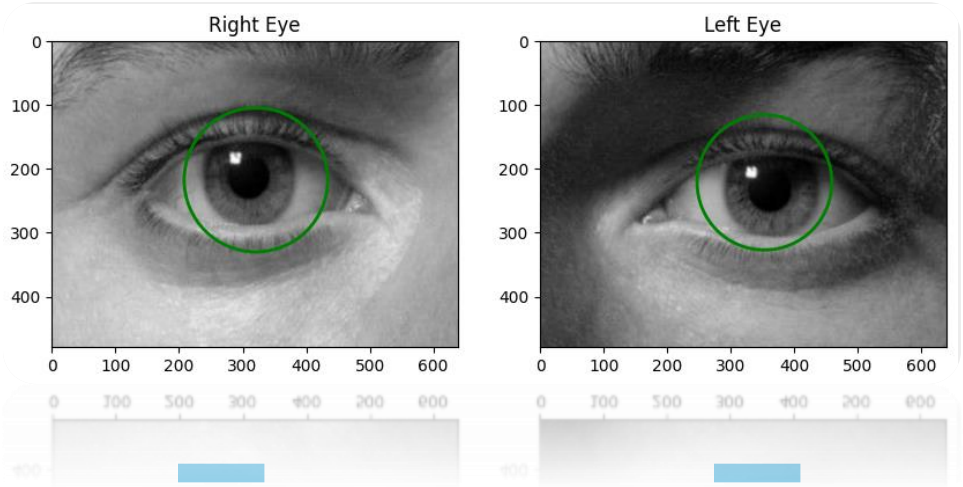
- **Apply the eyes detector model** to find the position of the eyes in the image
- **Extract a bounding box for each of the two eyes** using the three coordinates obtained in the previous step
- **Apply the iris tracker model** to each eye in order to get the position of the irises

EXAMPLE

INPUT IMAGE



EYES POSITION



EYES BOUNDING
BOXES AND
IRIS POSITION