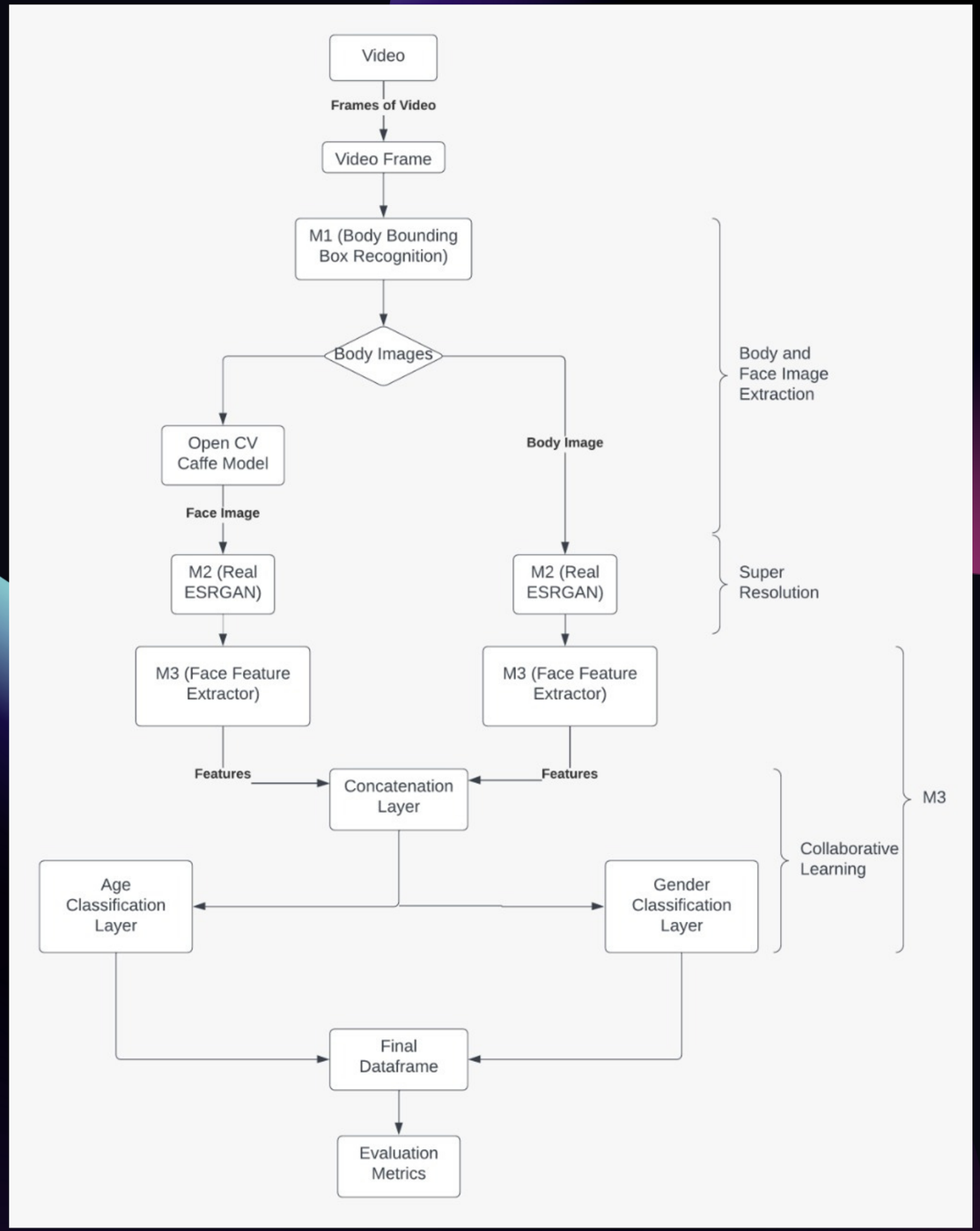
TEAM-13

BOSCH
AGE AND GENDER DETECTION

Problem Statement

Build a solution to estimate the age and gender of people detected from a surveillance footage feed in public areas.

Proposed Pipeline



Pipeline

Model 1

Body and Face Detection

- CCTV video Input processing
- Body Detection
- Face Detection

Pipeline

Model 2

Real-ESRGAN Architecture

- Super Resolute the cropped images

Pipeline

Model 3

- Age and Gender Prediction
- Final Output
- Evaluation Metrics

Model 1

Body and Face Detection Architecture

Faster RCNN Model

- Pretrained on CoCo Dataset and fine-tuned on:
  a. Pytorch Official Datasets
  b. Widerperson Dataset

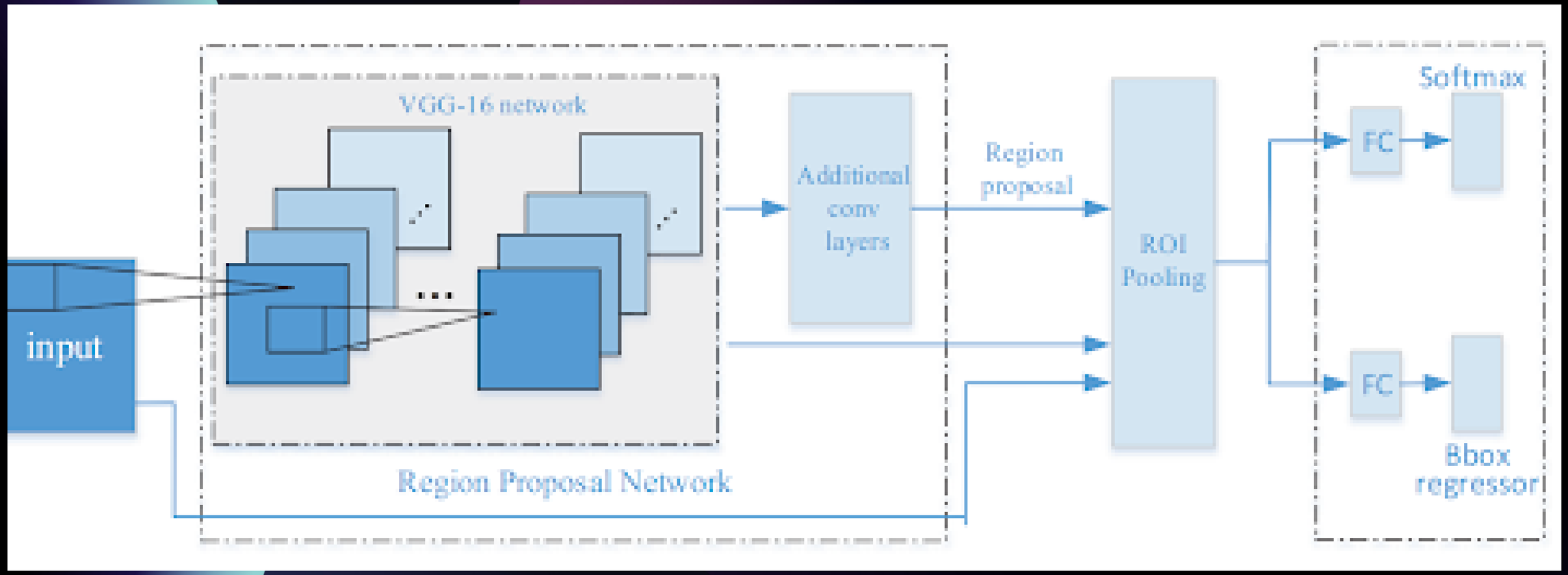
Fig: Widerperson Dataset

Fig: CoCo Dataset

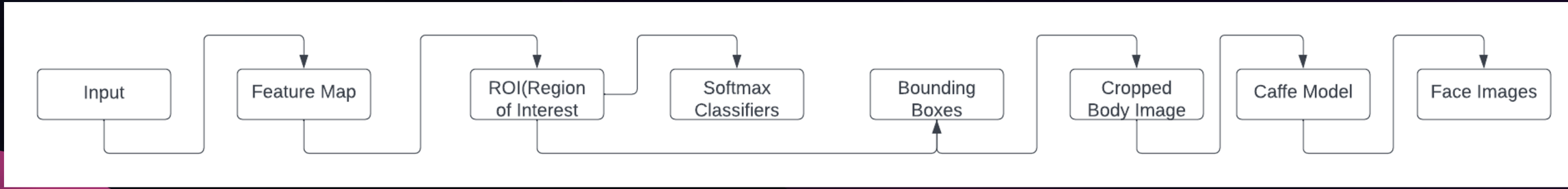
Why Faster RCNN?

- Directly finds region proposals and outputs regression bounding boxes.
- The size of the input images can be arbitrary, whereas most CNN-based models adjust the image to a specific size.

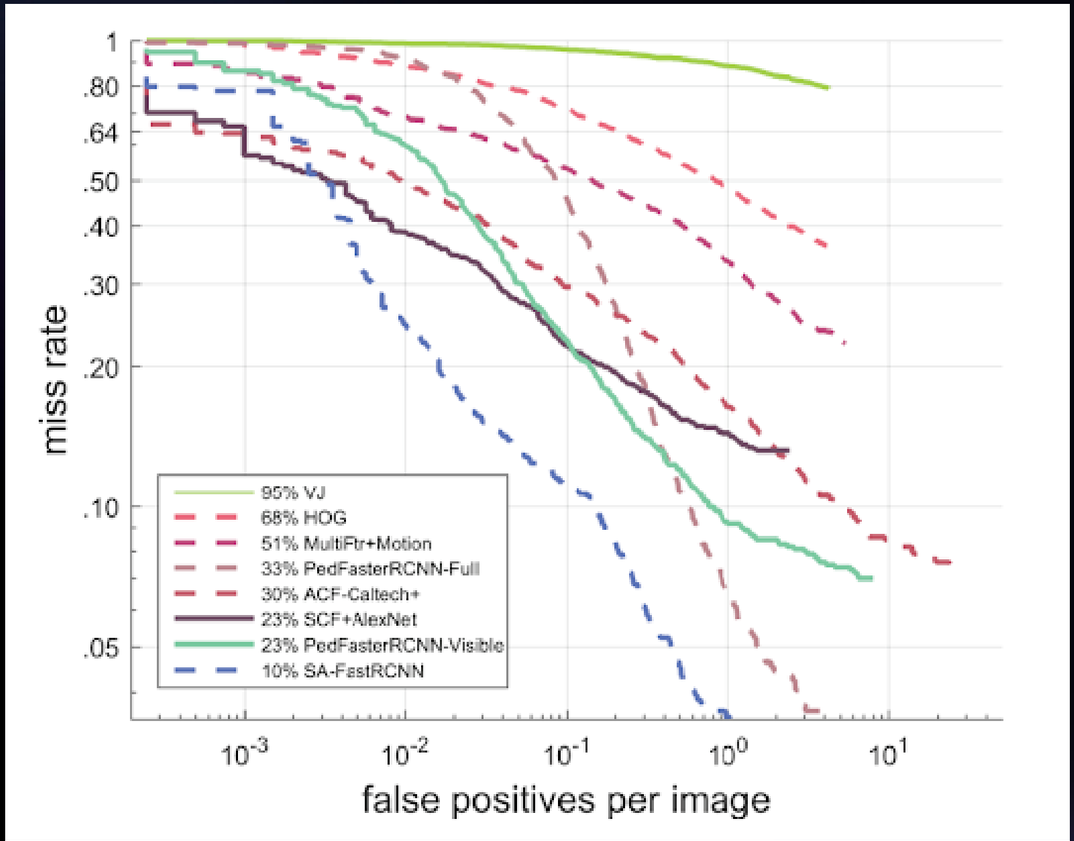
Faster RCNN Architecture



Model 1 Workflow



Result of the used Model



Method	Test Time(s)	Miss Rate
Proposed	0.26	23
SOTA RCNN	0.37	10
RCNN	5.31	13

Faster R-CNN

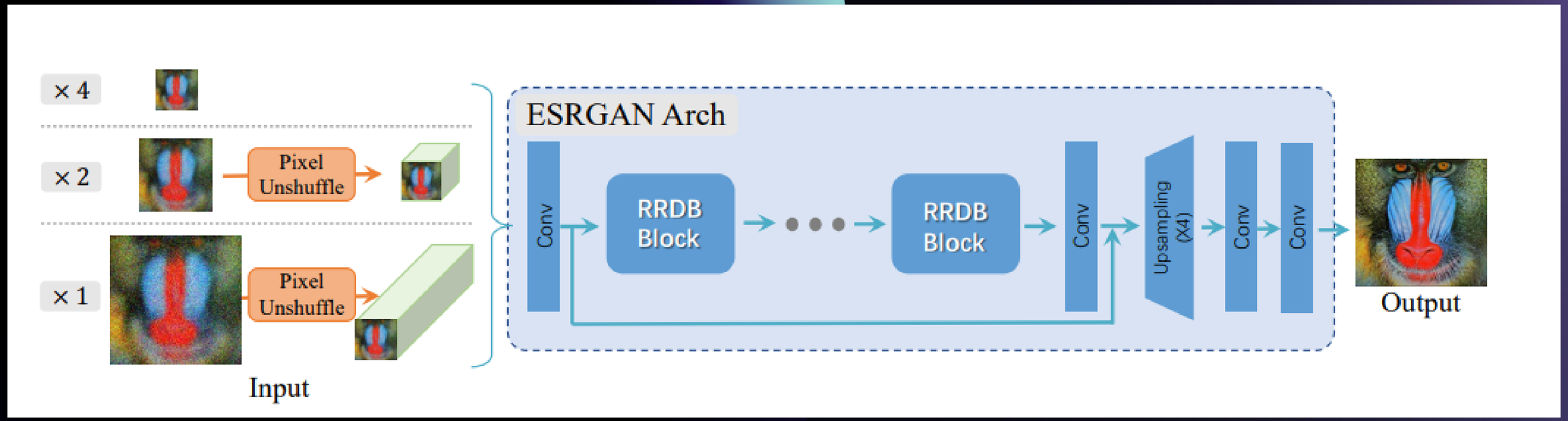
OpenCV Caffe Model

- It makes the face bounding boxes from the body cutouts.

Model 2

RealESRGAN

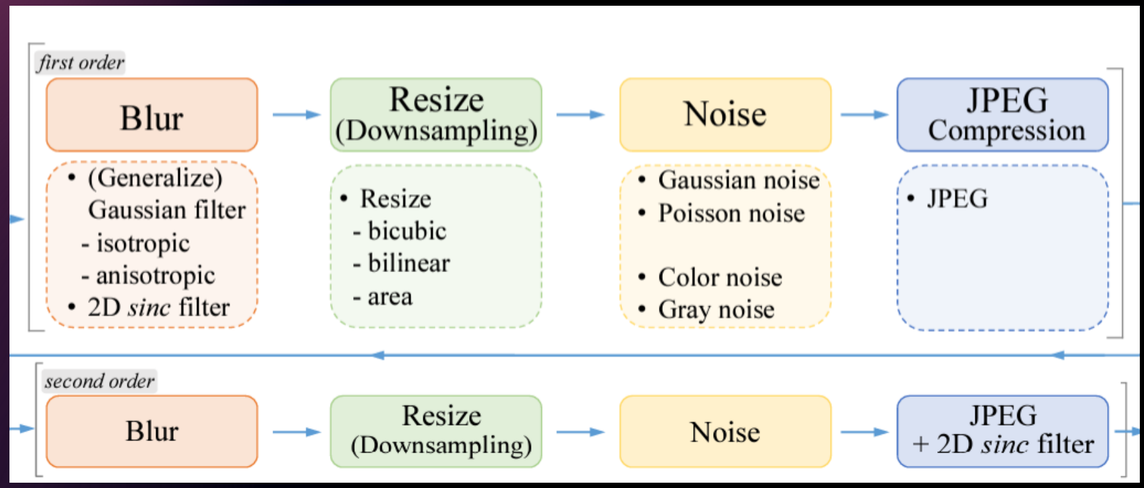
Real-ESRGAN



Real-ESRGAN adopt the same generator (SR network) as ESRGAN, i.e., a deep network with several residual-in-residual dense blocks (RRDB)

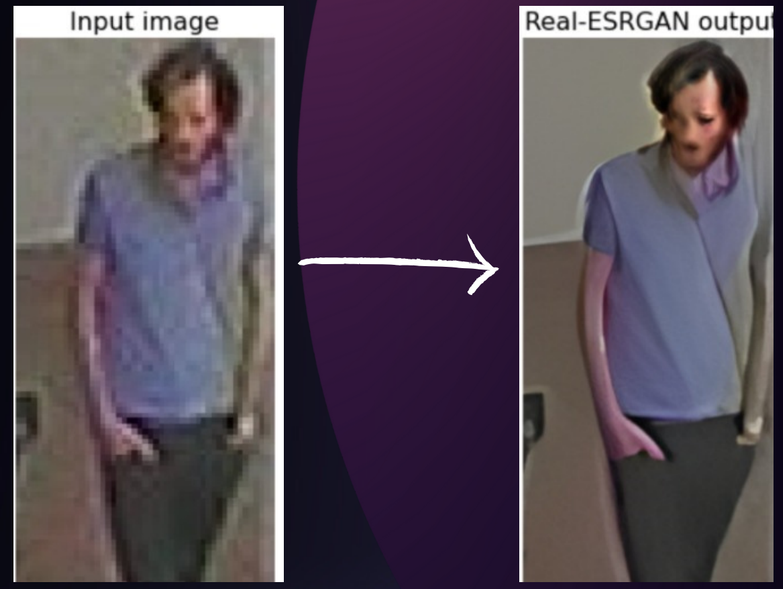
Benefits of using Real-ESRGAN

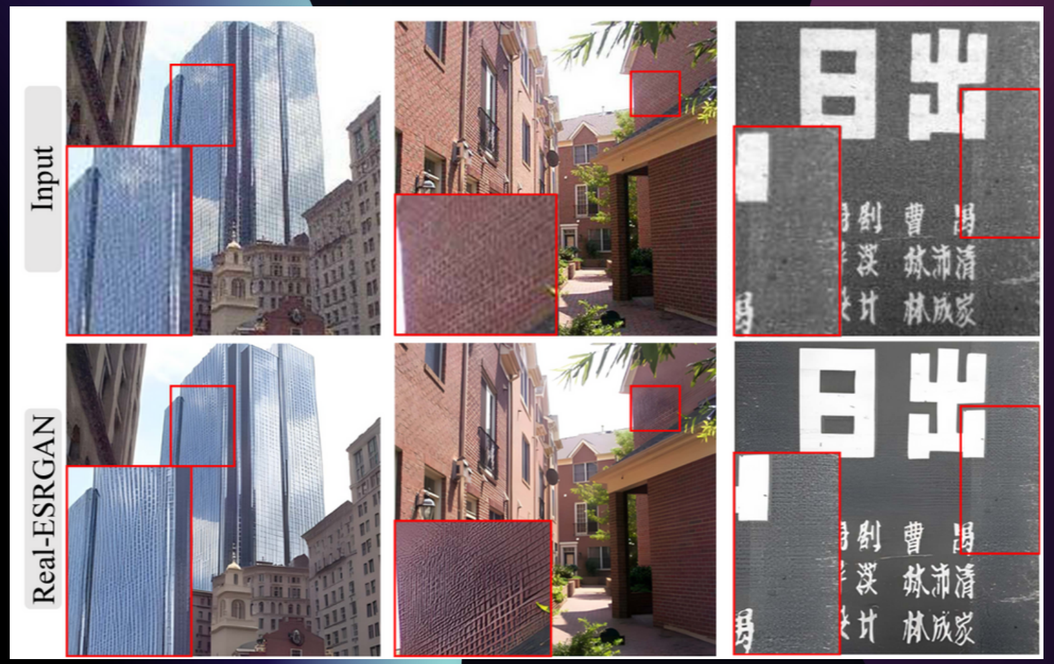
- Real-ESRGAN uses a high-order degradation process to model practical degradations.
- Real-ESRGAN employs essential modifications to increase discriminator capability and stabilize the training dynamics.
- Real-ESRGAN also restores realistic textures for real-world samples, while other methods either fail to remove degradations or add unnatural surfaces.



Limitations of Real-ESRGAN

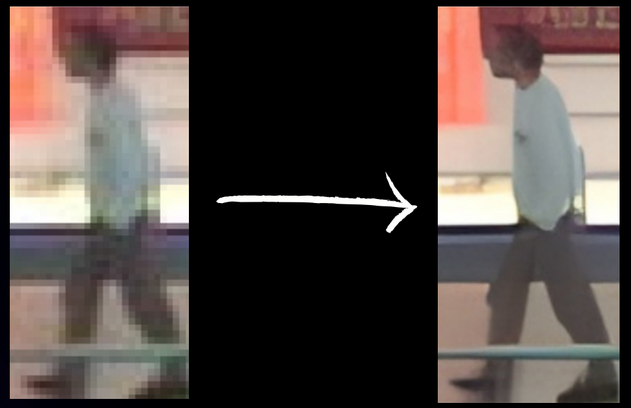
- Some restored images have twisted lines due to aliasing issues.
- GAN training introduces unpleasant artifacts on some samples.





- It could not remove out-of-distribution complicated degradations in the real world.
- Even worse, it may amplify these artifacts.
- Smoothening the image instead of super resolving

Model 2 Results



Model 3

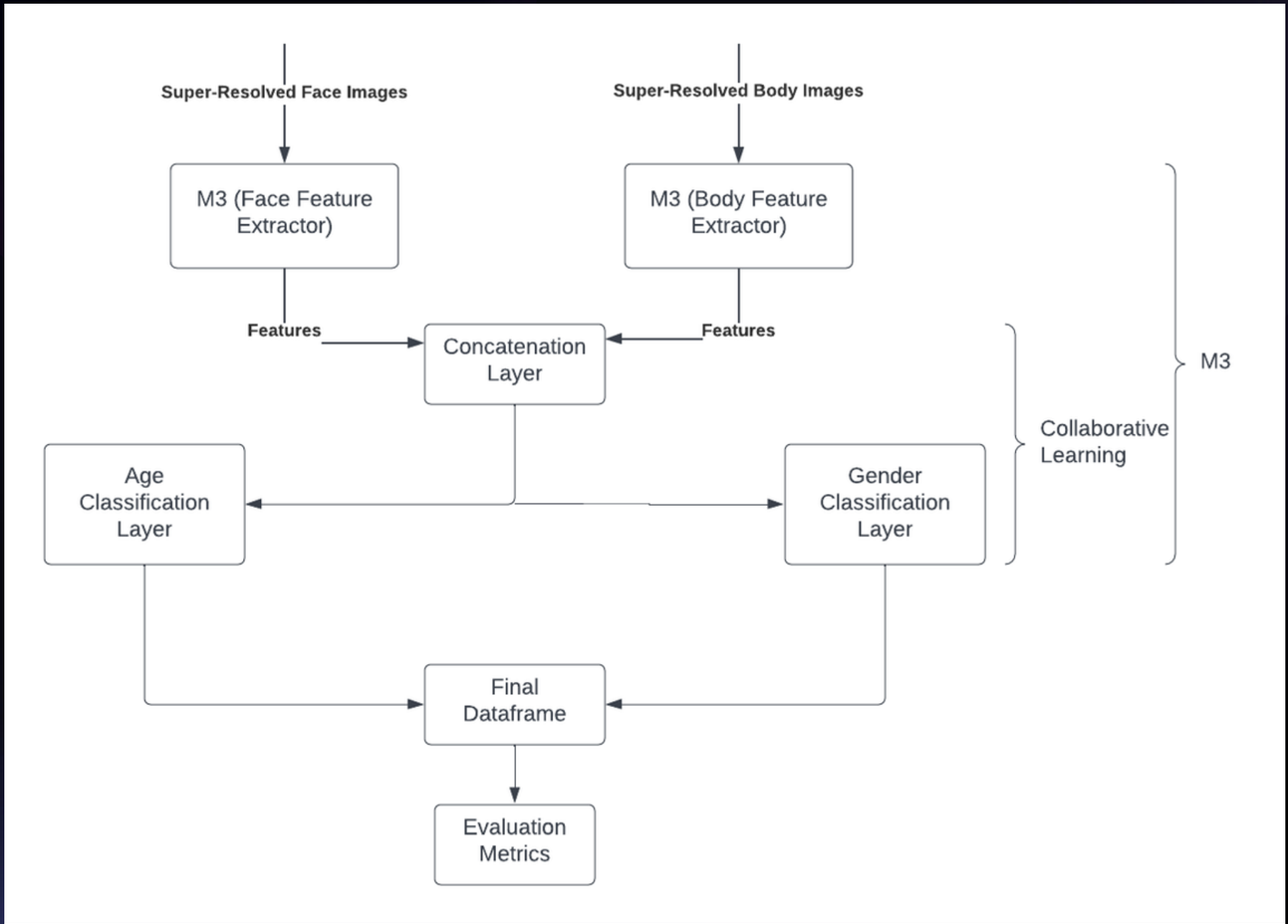
Age and Gender Detection

Model 3

- Model Three was built to predict the age and gender of the person.

- The main purpose was to extract features from CNN layers and we used famous architectures as well as scratch implementations to achieve the same task.

Model 3

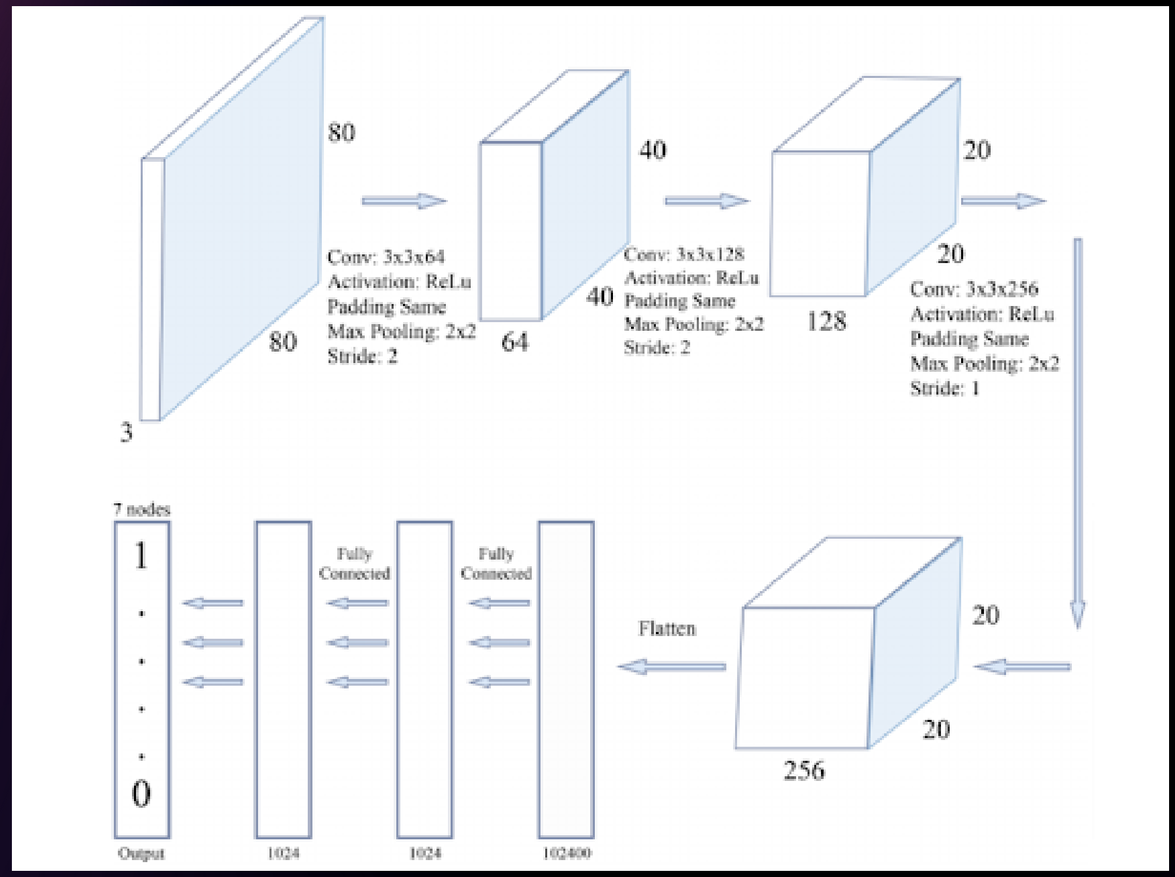

Super-Resolved Face Images
Super-Resolved Body Images
M3 (Face Feature Extractor)
M3 (Body Feature Extractor)
Features
Features
Concatenation Layer
Age Classification Layer
Gender Classification Layer
Collaborative Learning
M3
Final Dataframe
Evaluation Metrics

Body Only Model Scratch Model

Dataset and Hyperparameters:

- Dataset : PETA(Body Images Dataset)
- Batch Size = 512
- Image Size = 160,60
- Optimizer = ADAM
- Learning rate = 0.001
- Weight decay = default
- Criterion = cross entropy loss
- Epochs = 20

Architecture:



Body Only Model Accuracy

Model Name	Scratch
Accuracy	66.8 %

Face Only Model training Hyper-Parameters

- Dataset : UTK Face Dataset
- Batch Size = 512
- Image Size = (160, 60)
- Optimizer = ADAM
- Learning rate = 0.001
- Weight decay = default(0)
- Criterion = cross entropy loss
- Epochs = Varied

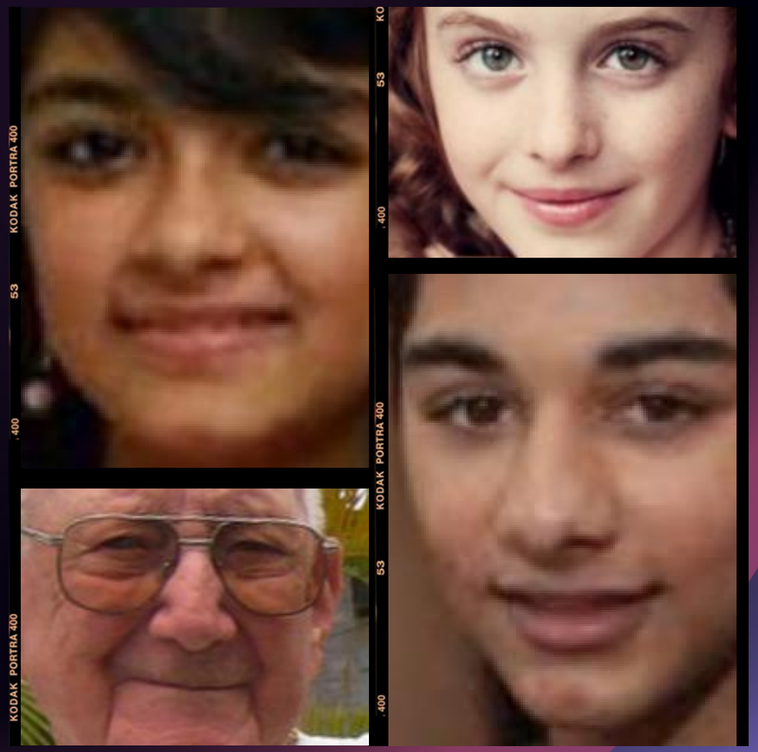
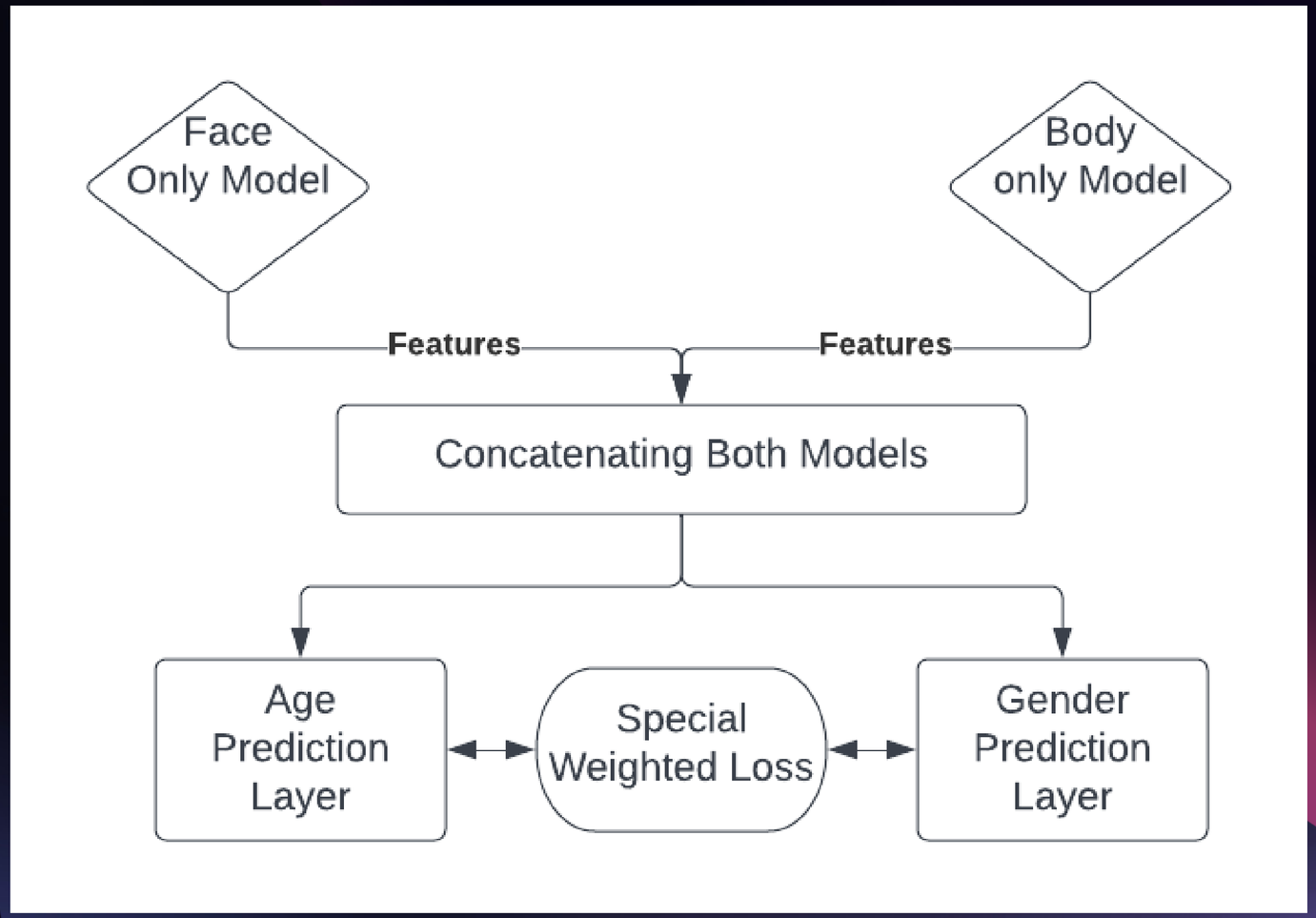


Fig: UTK Dataset

Face Only Models Accuracy

Model Name	VGG-16	Facenet	Resnet 18	Resnet 50
Accuracy	23 %	35.4 %	31.9 %	30.87 %

Combined Model



Collaborative Learning

- Through collaborative learning, we want the face-only and body-only models to learn simultaneously.

$$L_{total} = a.L_{gender} + (1 - a).L_{age}$$

$$a = (0, 1)$$

- For this, we employed a weighted loss function and depending upon the metrics of the two predictions, we adapt the weights of the individual losses.

Future Improvements

- Create a relatable dataset for gender and age classification from both body and face images.
- Model 1 : Improve the detection of people far away and prevent the false boxes due to the background.
- Model 2 : Customise and train it to enable better feature enhancement of the input images from Model 1.

Model 3 :

- Try more SOTA architectures. Train the models on augmented data. Tackle the problem caused by the loss function and arrive at optimal solution which would capture the prime essence of collaborative learning.

Thank you !!