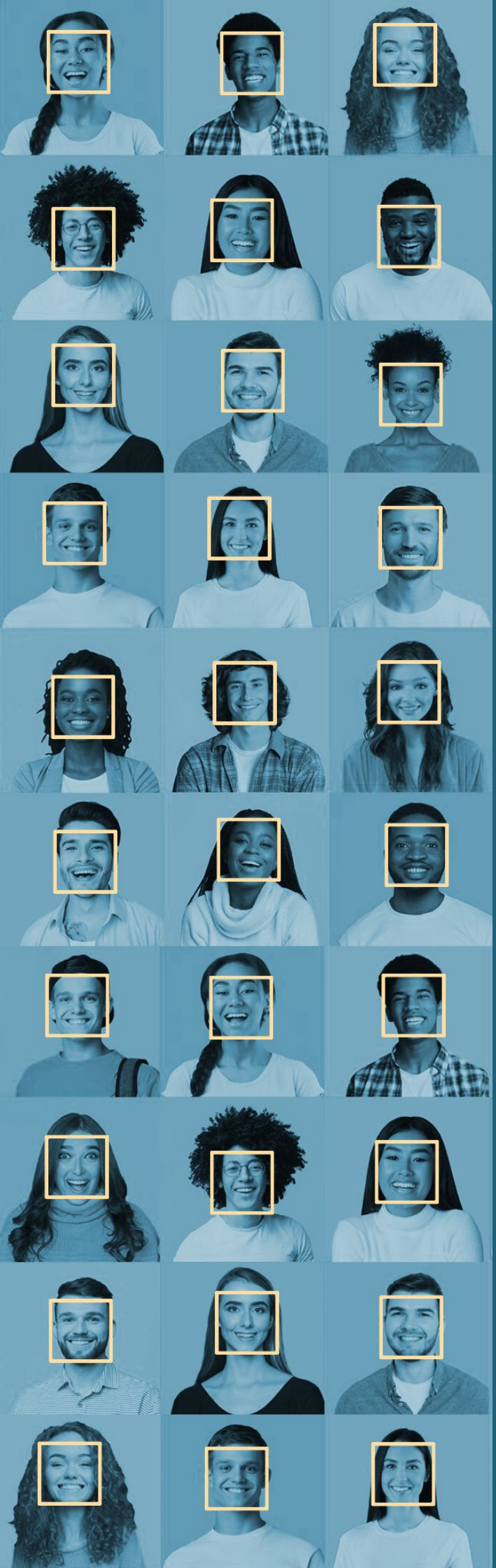


FOURIERISE

TEAM ID 5

BOSCH

Bosch's Age and Gender Detection

Motivation

Data Analytics

We believe that data has the potential to boost industries, in the 21st century. Age and Gender predictions on CCTV footage in public spaces like malls, hospitals, retail stores, and other tourist spots can be used by the stakeholders to analyse their frequent customers. Such analysis can help realign business strategies to increase sales and profit.

Crime Detection

Age and gender prediction can assist in identifying criminals. Many a times only the age, gender and height of the criminals are reported. Age and gender prediction can be used for narrowing down the crowd under observation to locate the criminal more promptly making the process speedy.

Entry Regulation

In an overpopulated country like India, it is difficult for a security personnel to be held solely responsible for differentiating between the two genders to regulate entry points.

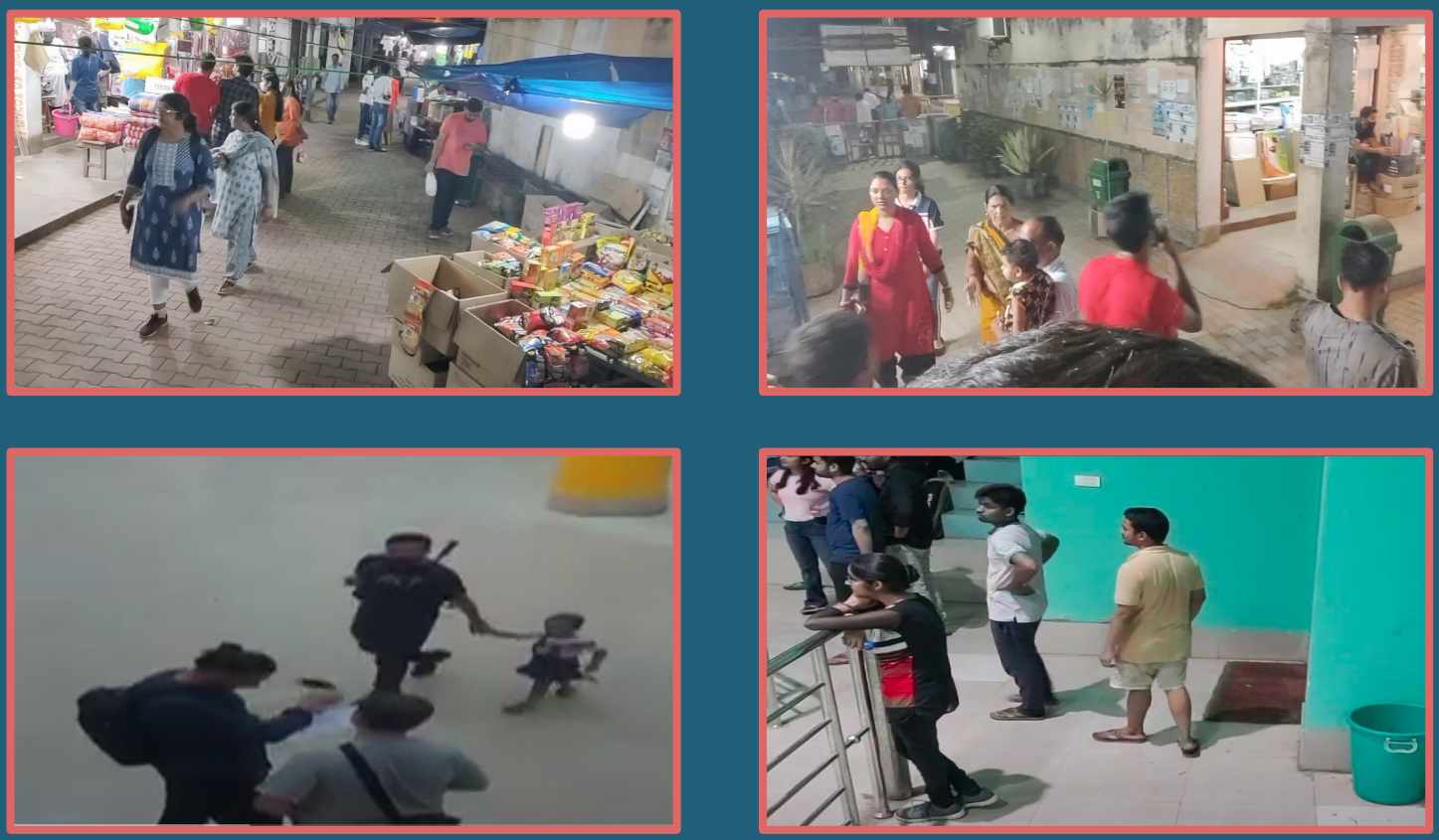
Missing Person Cases

Police departments may use this technology to detect missing people. This technology can even be used at publicly crowded places where missing person reports are frequent .

Dataset

We manually collected surveillance clips from our institute CCTV cameras. The scenes included gym activity, institute market footage and the annual cultural fest footage. We use publicly available UTKFace[1] and Adience[2] dataset.

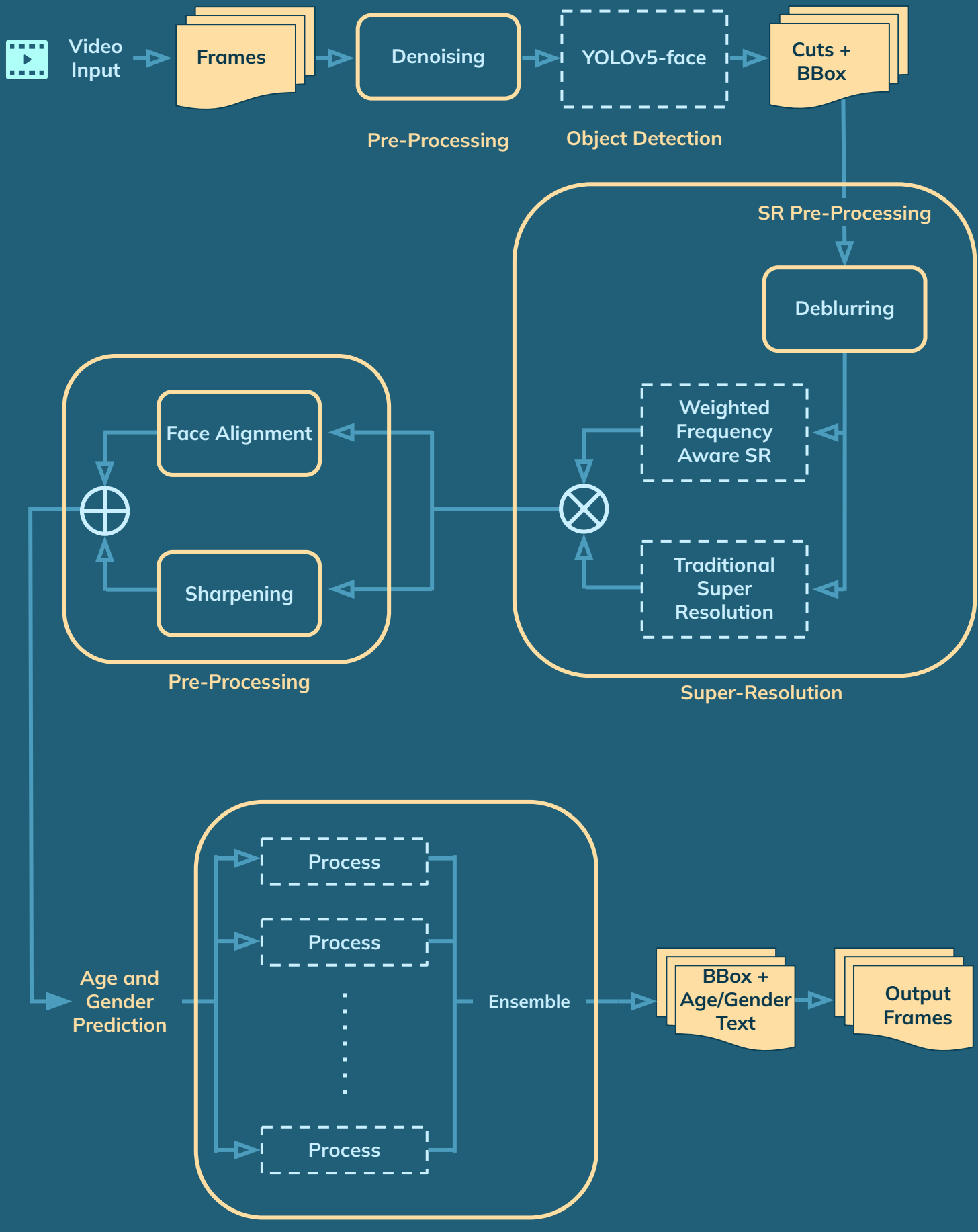
Annotation format - We report the annotations in the output form as desired by the Problem Statement. The format is specified below



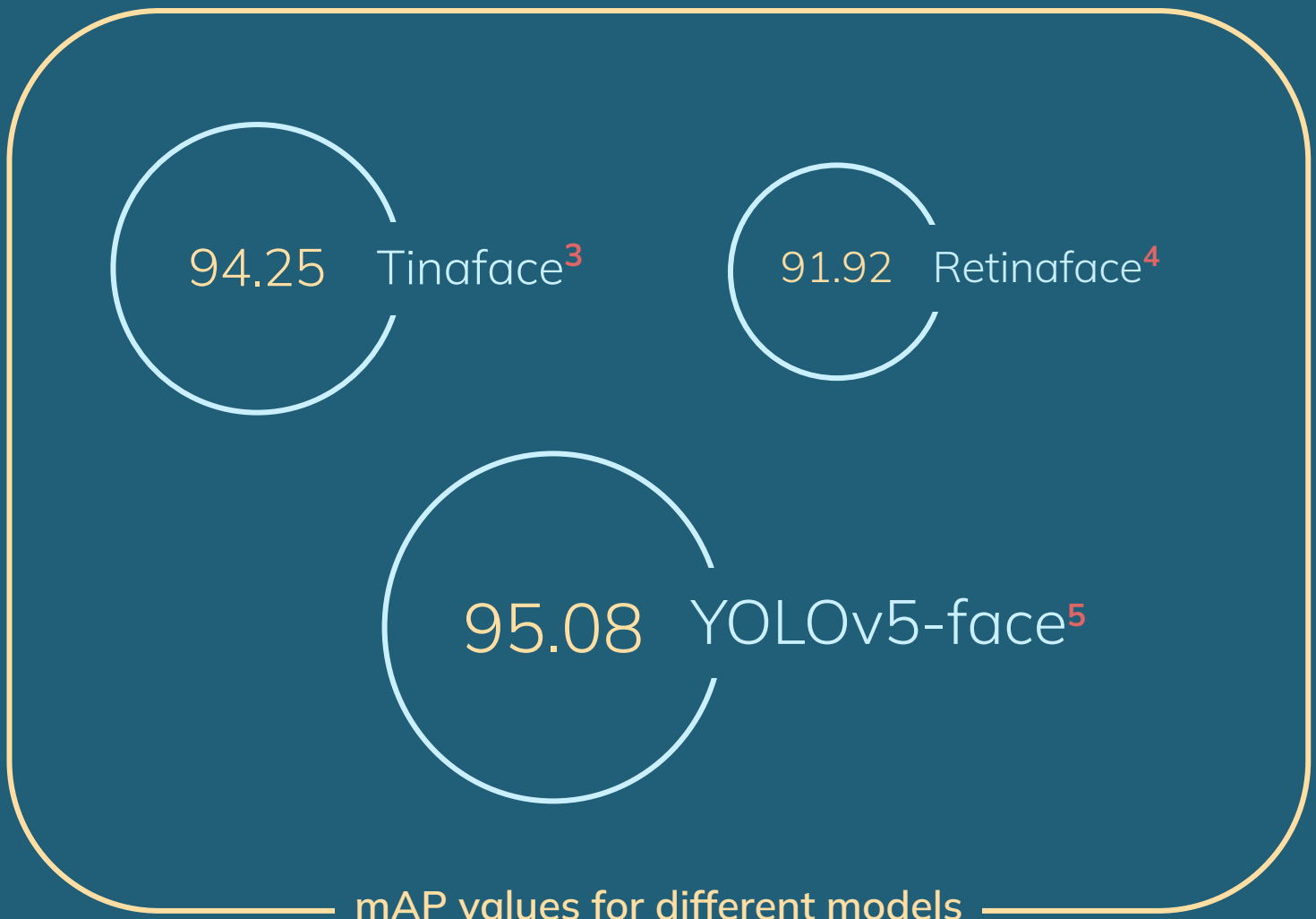
[samples from the dataset]

bb_xmin	bb_ymin	bb_height	bb_width	age_min	age_max	age_actual	gender
197	379	70	53	20	30	24.9	F
250	342	50	44	40	50	46.8	F
430	397	53	44	40	50	49.5	M
453	511	56	39	20	30	24.2	M

Workflow


Video Input
Frames
Denoising
YOLOv5-face
Cuts + BBox
Pre-Processing
Object Detection
SR Pre-Processing
Deblurring
Face Alignment
Sharpening
Weighted Frequency Aware SR
Traditional Super Resolution
Pre-Processing
Super-Resolution
Age and Gender Prediction
Process
Process
Process
Ensemble
BBox + Age/Gender Text
Output Frames

Object Detection

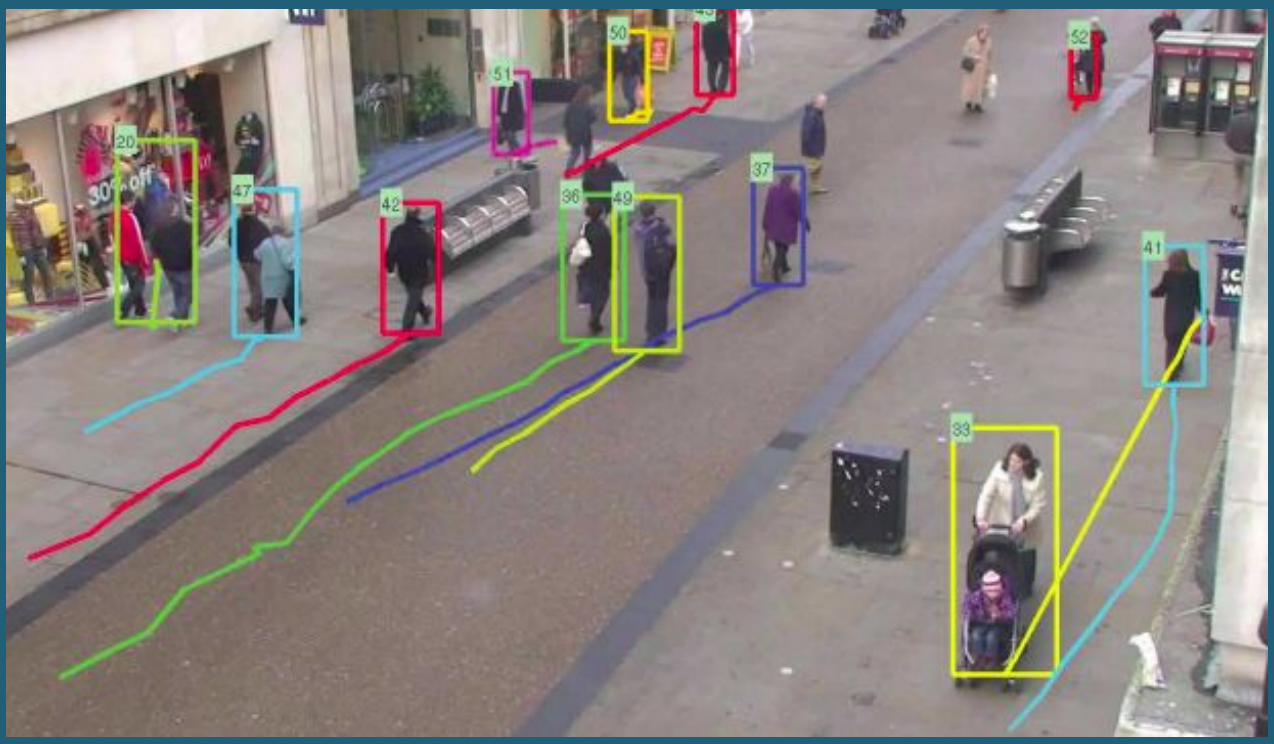


mAP values for different models

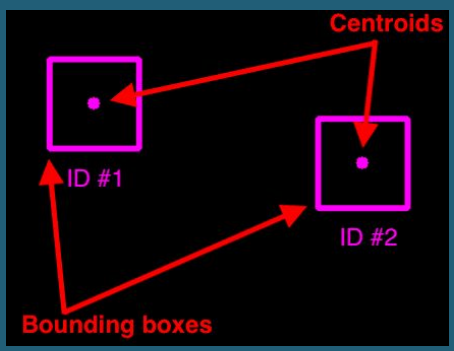
Decision

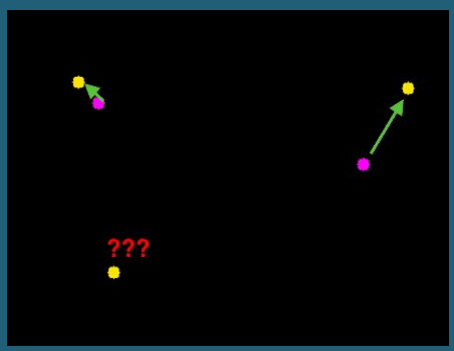
We observed that YOLOv5-face performs well across a plethora of datasets as compared to the other state of the art object detection models, as represented by the mAP values.

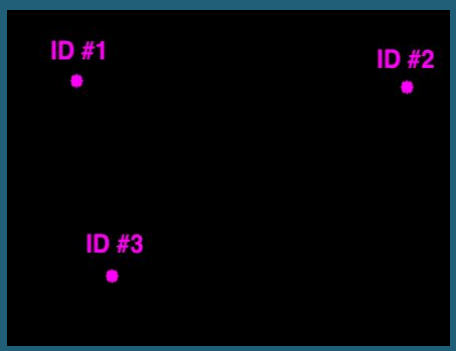
Object Tracking



After detecting faces in each frame, object tracking algorithms can be used to assign a unique identity to each detected face for each frame.







Step 1

Get centroids for each detected face's bounding box

Step 2

Match centroid for each bounding box of new frame to nearest centroid from last frame

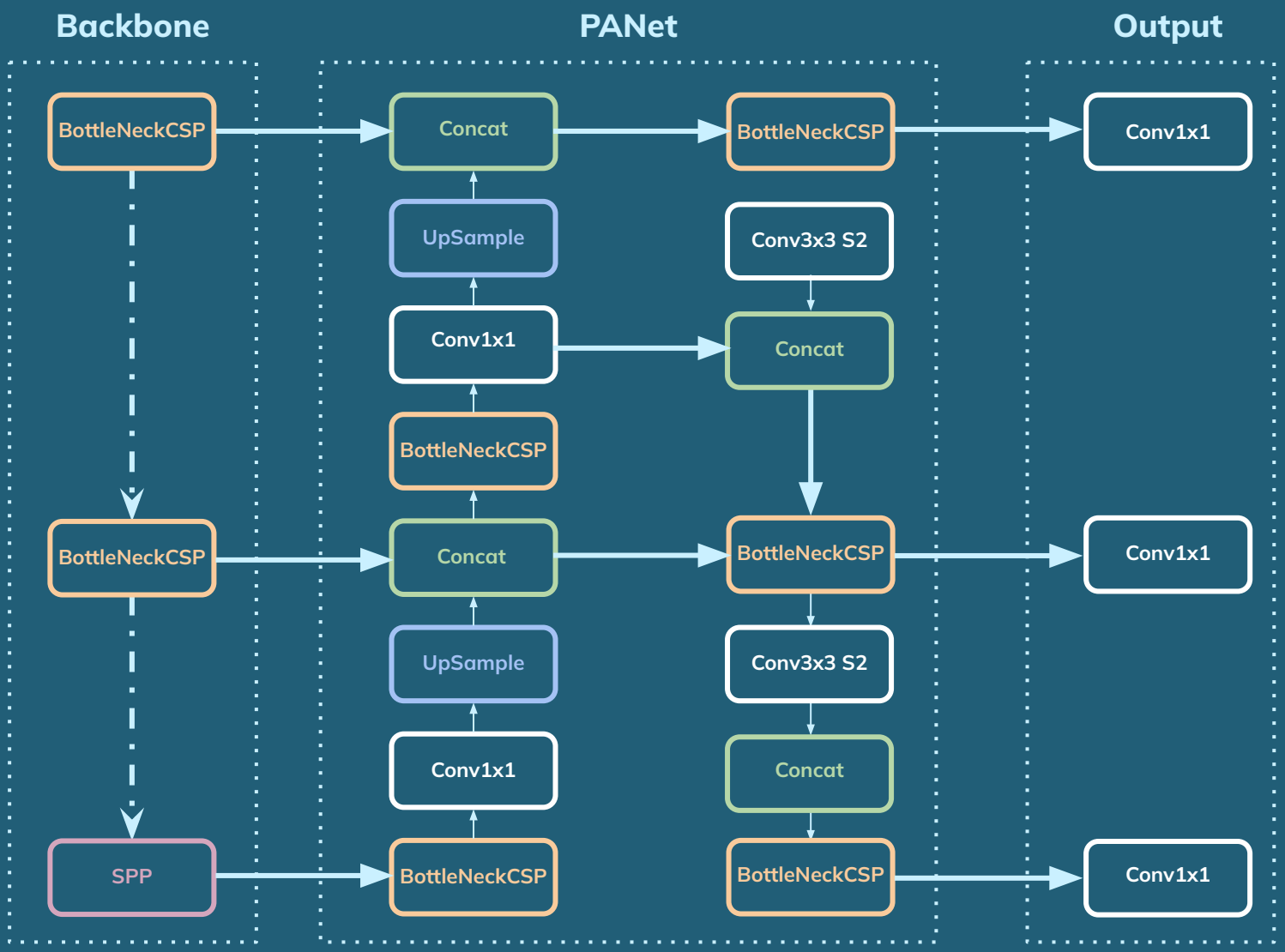
Step 3

Assign same id as the matched centroid from previous frame. Assign new id to unmatched centroid

YOLOv5-face[5]

We tested multiple face detection algorithms through a wide variety of manual as well as open source datasets. This experiment gave us an insight into the fact that face detection is a well solved problem even under the most adverse perturbations in the image. Even after adding salt and pepper noise the state of the art models were able to detect most of the faces in the frame.

Taking into account the mAP(mean Average Precision) scores and wall time we chose yolov5-face face detection algorithm. This gave us cropped facial images and also did not require the input image to be resized which could have caused significant data loss as in the case of some other models.



Pre-Processing

Denoise[6]

Sharpen[7]

Deblur[8]

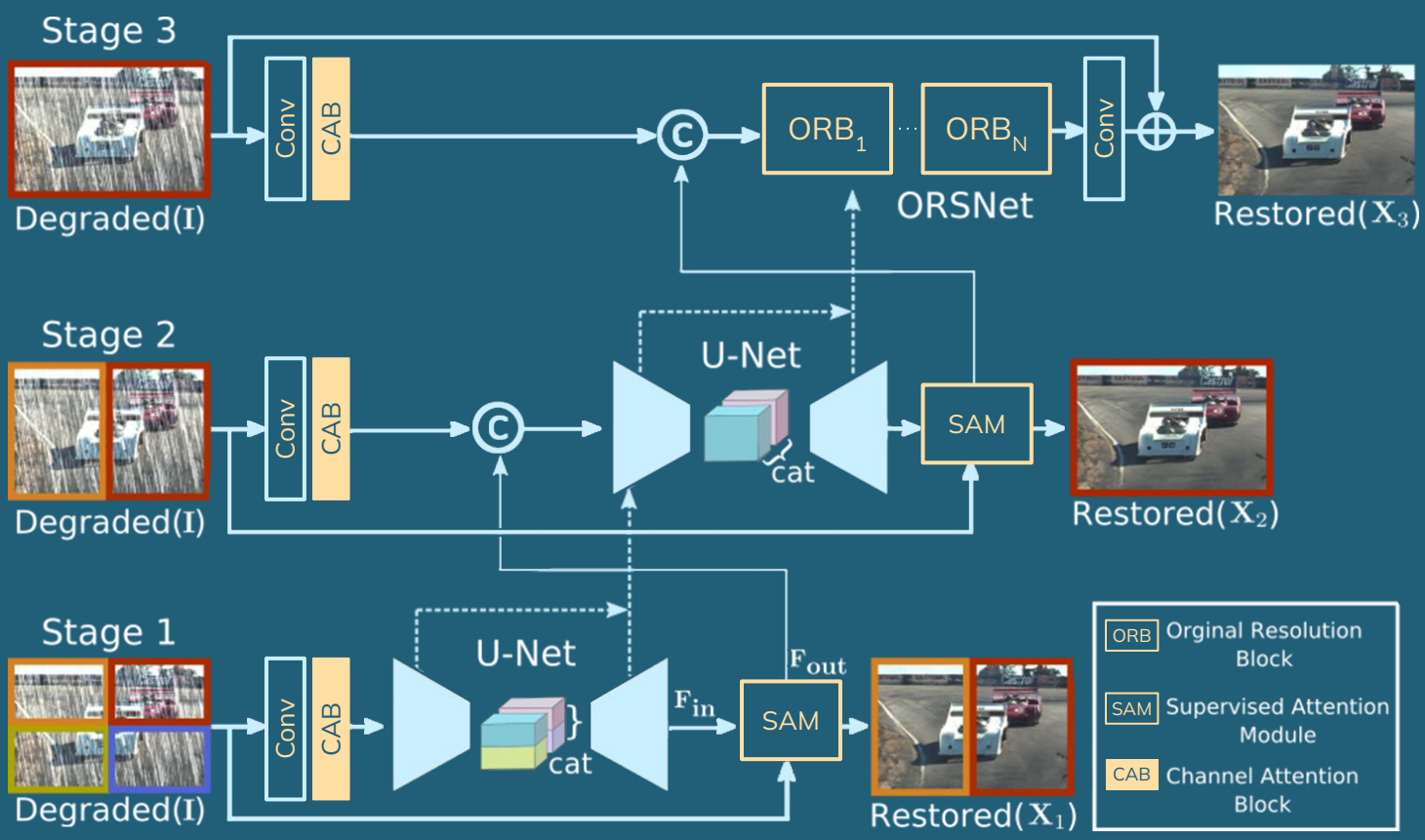
Pre-Processing

Deburring process

Deblurring is the process of removing blurring artifacts from images [input image say B which is blurred image which generally happens due to camera shake or some other phenomenon].

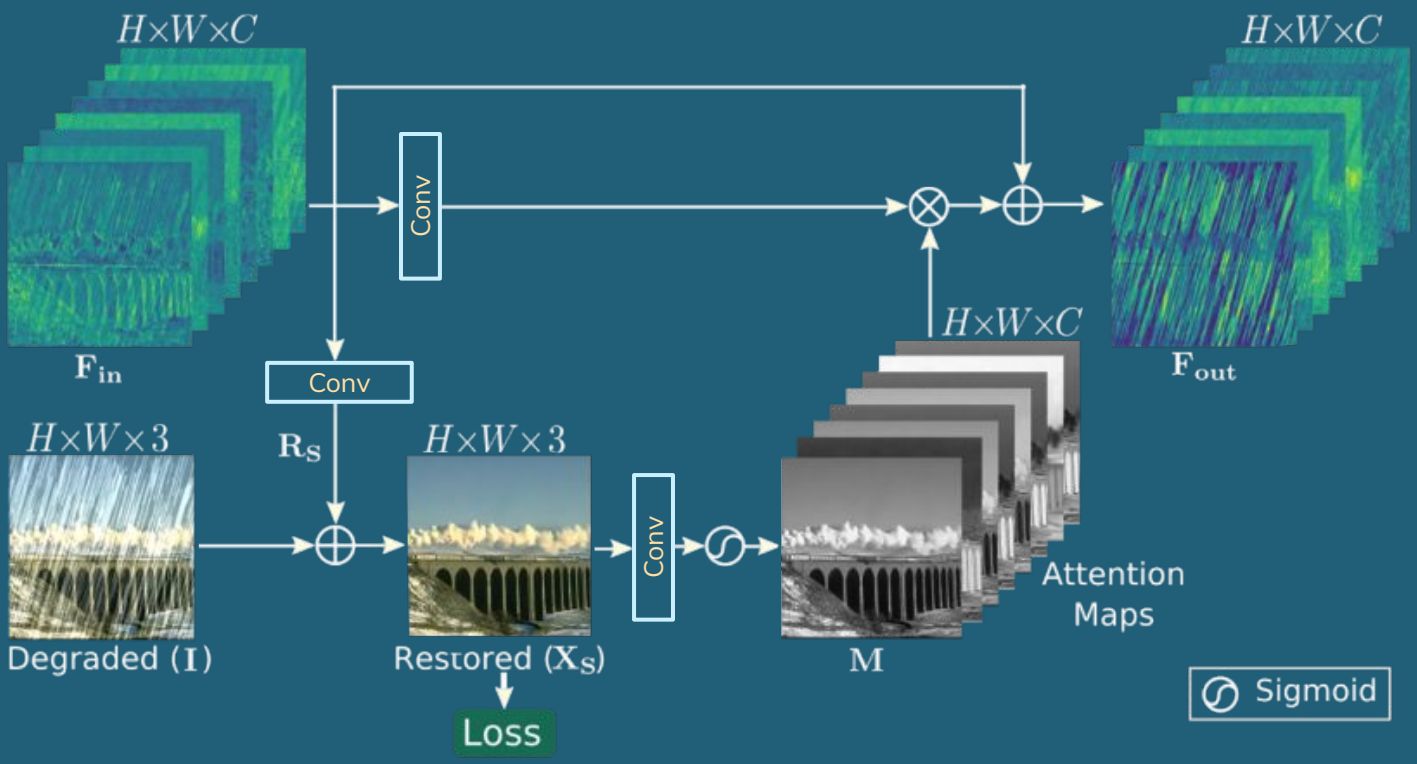
For this preprocessing part we have used MPRnet[8] or Multi-Stage Progressive Image Restoration

Model Architecture



Pre-Processing

This is the overall framework of MPRNet and the supervised attention model can be visualized below :



The model learns the contextualized features using encoder-decoder architectures and later combines them with a high-resolution branch that retains local information. The resulting tightly interlinked multi-stage architecture, named as MPRNet, delivers strong performance gains on ten datasets across a range of tasks including image deraining, deblurring, and denoising.

Super Resolution

Experiments

MODEL	Custom Metric (∑(HR-LR))	Wall Time	PSNR
WDSR[16]	345.50	**15.85**	**30.36**
EDSR[17]	347.66	2.11	30.04
SRGAN[18]	354.71	9.19	29.42
FSRCNN[19]	430.68	0.37	23.69
ESPCN[20]	362.02	0.41	25.14
FSRCNN_trained	575.18	0.80	21.94
WFASR[21]	349.65	**2.34**	**30.22**

Decision

	Wall Time	PSNR
WDSR		
WFASR		

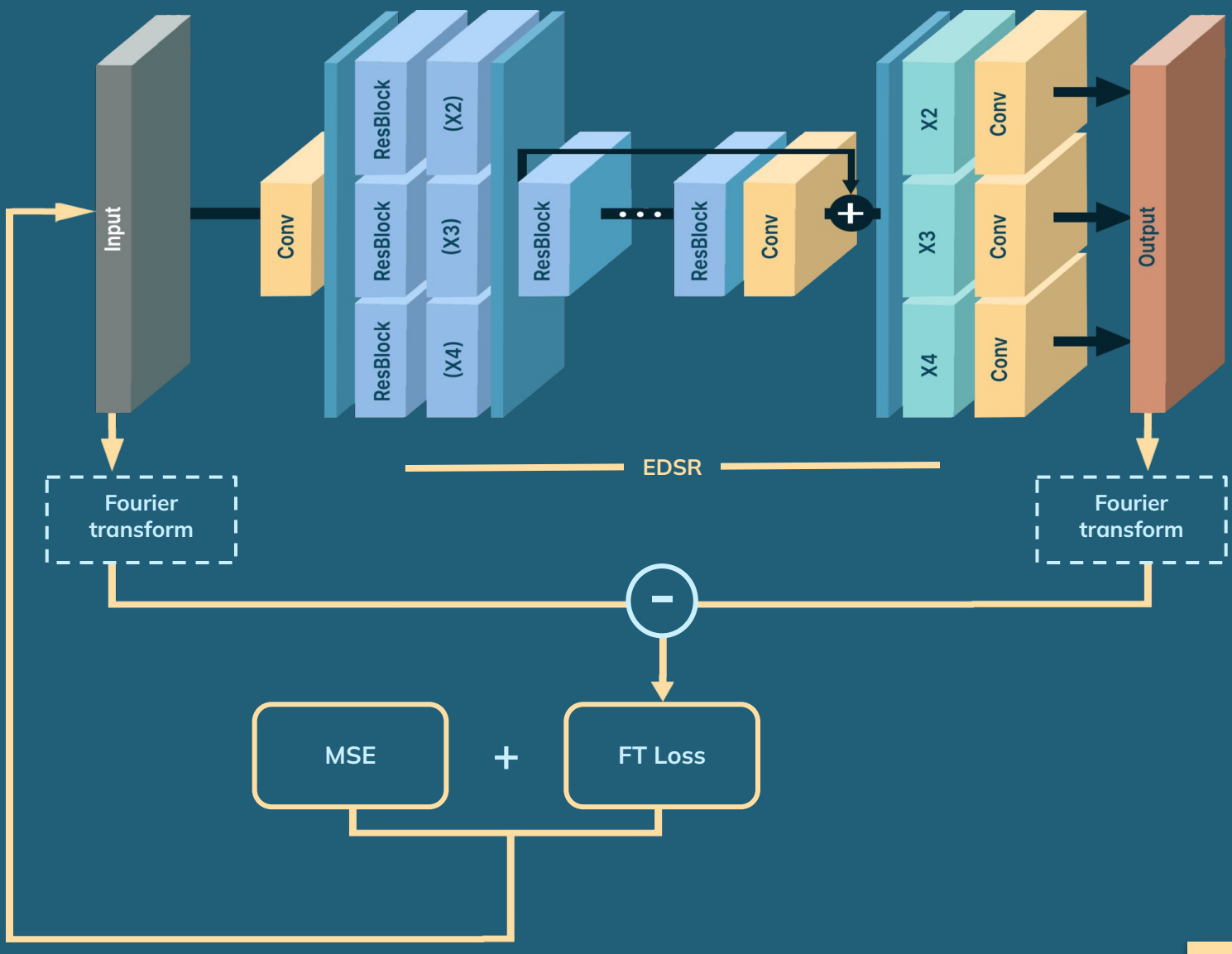
Despite having less PSNR than WDSR, WFASR augments our pipeline by saving us crucial inference time which can then be diverted to other computationally expensive processes such as denoising or age and gender estimation.

Weighted Frequency Aware SR

The pain point of the problem statement mentioned that the existing super resolution methods despite increasing the PSNR value, do not take into account the loss of high frequency details from the image. This causes the image to be perceptually uncomfortable and makes restored video un-human like.

Thus to address this issue we came up with a novel concept of introducing a *FT(Fourier transform[9])-Loss* to minimize the difference of the FT of the input and output image weighted by the distance of the pixel from the origin. **The weights ensure that the high frequency information is preserved in the output image.**

This however could cause the network to become parameter-heavy. However in order to alleviate this problem we decided to train the original architecture and just add this FT-Loss on the top of the MSE Loss. The MSE Loss takes care of the original colour and HSV reconstruction whilst the FT-Loss preserves the edge information, aka the high frequency details.



Gender Prediction

Experiments

MODEL	DATASET	ACCURACY	PRECISION	RECALL	F1	WALL TIME
Facelib[10]	Adience[2]	0.73	0.73	0.73	0.73	0.100
	UTK[1]	0.78	0.79	0.79	0.78	
VGG Face[11]	Adience[2]	0.74	0.76	0.75	0.74	0.003
	UTK[1]	**0.91**	**0.92**	**0.92**	**0.90**	
MLP[12]	Adience[2]	0.48	0.49	0.49	0.47	5.620
	UTK[1]	0.93	0.93	0.94	0.92	
MLP[12] + MTCNN[13] embeddings	Adience[2]	0.85	0.86	0.87	0.84	0.640
	UTK[1]	0.92	0.91	0.93	0.91	

Decision

Gender Prediction is a relatively easy task, being a binary classification problem. As depicted in the table above, **VGG face** is observed to perform the well as compared to the other models evaluated across different datasets. Despite MLP having higher accuracy, it was a relatively larger model with higher prediction time making the pipeline more delayed than desired.

Age Prediction

Experiments

MODEL	VARIATION	MAE	ACCURACY
NDF[14]	Fine tuned (regression)	4.85	0.626
	Fine tuned (classification)	2.64	0.945
MLP[12] + MTCNN[13] embeddings	Trained (original)	5.37	0.602
	Trained (gender prior)	3.89	0.849
VGG Face[11]	Trained (aug data)	7.12	0.582
	Trained (modified architecture)	6.21	0.613

Decision

Age prediction from face images is a nuanced regression problem which was performing suboptimally in all of our experiments. Thus in order to make it more robust we performed ensembling of various models trained on multiple datasets. We incorporated several innovations in our training paradigm such as augmentations, architecture modification, introducing priors, etc.

Age Prediction

VGG Face[11]

VGG Face Model which is very apt at face recognition tasks was used for age estimation. The face-feature-vectors produced by the VGG Face model in the fourth last layer was routed to a set of classification layers which finally produces the probability distribution corresponding to ages 0-100. The whole model other than the classification layer was frozen during its training.

MLP[12] with gender prior

MLP initially used Arcface to extract the face embeddings which was later used to predict age and gender in MLP. Since extraction of face embeddings came out as a time intensive task, it was replaced by Inception Resnet 1 to extract face embeddings which was utilized to give gender and age predictions.

NDF[14] Regression & Classification

Neural Decision Forests (NDF) is a novel approach unifying the representation power of Neural Networks and divide-and-conquer technique of traditional decision trees. The NDF model differs from convention deep networks because a decision forest provides the final predictions using proposed, joint and global optimization of split and leaf node parameters.

Challenges

1

Problem

Frame extraction from videos returns perceptually uncomfortable frames.

Solution

To be able to reconstruct the information and reproduce static facial features we used deblurring technique. It attempts to maintain a complex relationship between spatial details and high-level contextualized information while recovering images.

2

Problem

Due to the position and unfavourable surroundings, the cctv footage might end up with undesirable noise.

Solution

We used denoising techniques to adapt to these unfavorable perturbations

- Salt and pepper
- Gaussian

3

Problem

Faces captured in videos might not be aligned vertically, as a result the model faces a challenge detecting the facial features.

Solution

We use deep learning based face alignment techniques to facilitate age and gender prediction based on the above experimental learning.

4

Problem Even the SOTA models for age prediction were not performing well on datasets having a distribution different from the train set.

Solution To ensure robustness of our model we implemented ensemble technique on multiple datasets and models

5

Problem Age and Gender are heavily researched problems. Gender being a binary classification task is well solved however, age is a complex problem to solve.

Solution We modified the architecture of MLP to use the high accuracy gender predictions as prior to the age prediction embeddings.

6

Problem MLP model despite being SOTA with high accuracy had high inference time due to the utilization of Arcface embeddings

Solution To reduce the inference time from 8 seconds to 0.2 seconds we utilised Insightface Resnet 1 embeddings of the image instead of Arcface. The massive improvement in inference time is at the cost of a minor loss in accuracy.

7

Problem In the quest of having a higher signal to noise ratio the SR techniques loose the high frequency information leading to loss of information.

Solution

- We designed a novel architecture utilising a customised fourier transform based loss function.
- We also incorporated image sharpening to balance the smoothness imposed by denoising and SR in our pipeline.

Metrics

MAE

$= \sum(|x_i - x|)/n$

Mean absolute error is the the most prevalent metric used to evaluate the performance of age prediction. It is the mean deviation of the prediction with respect to the ground truth.

PSNR[15]

$= 10 \log_{10}(L^2/ MSE)$

Peak Signal to Noise Ratio is the most recommended technique used to determine the quality of results. It can be calculated directly from the MSE using the formula below, where L is the maximum pixel value possible (255 for an 8-bit image).

Custom Metric

$= (\sum(HR-LR))$

Calculated by dividing the sum of absolute value of difference between corresponding pixels of high resolution image and image obtained by super resolution of the downsampled high resolution image by the number of pixels in the image.

References

1. https://susanqq.github.io/UTKFace/
2. https://arxiv.org/pdf/1806.02023.pdf
3. https://github.com/Media-Smart/vedadet/tree/main/configs/trainval/tinaface
4. https://github.com/serengil/retinaface
5. https://github.com/elyha7/yoloface
6. https://docs.opencv.org/3.4/d5/d69/tutorial_py_non_local_means.html
7. https://www.analyticsvidhya.com/blog/2021/08/sharpening-an-image-using-opencv-library-in-python/
8. https://arxiv.org/abs/2102.02808
9. https://docs.opencv.org/3.4/de/dbc/tutorial_py_fourier_transform.html
10. https://github.com/sajjjadayobi/FaceLib
11. https://github.com/rcmalli/keras-vggface
12. https://github.com/tae898/age-gender
13. https://arsfutura.com/magazine/face-recognition-with-facenet-and-mtcnn/
14. https://github.com/Nicholasli1995/VisualizingNDF
15. https://en.wikipedia.org/wiki/Peak_signal-to-noise_ratio
16. https://arxiv.org/abs/1808.08718
17. https://arxiv.org/abs/1707.02921
18. https://arxiv.org/abs/1609.04802
19. https://arxiv.org/abs/1608.00367
20. https://arxiv.org/abs/1609.05158