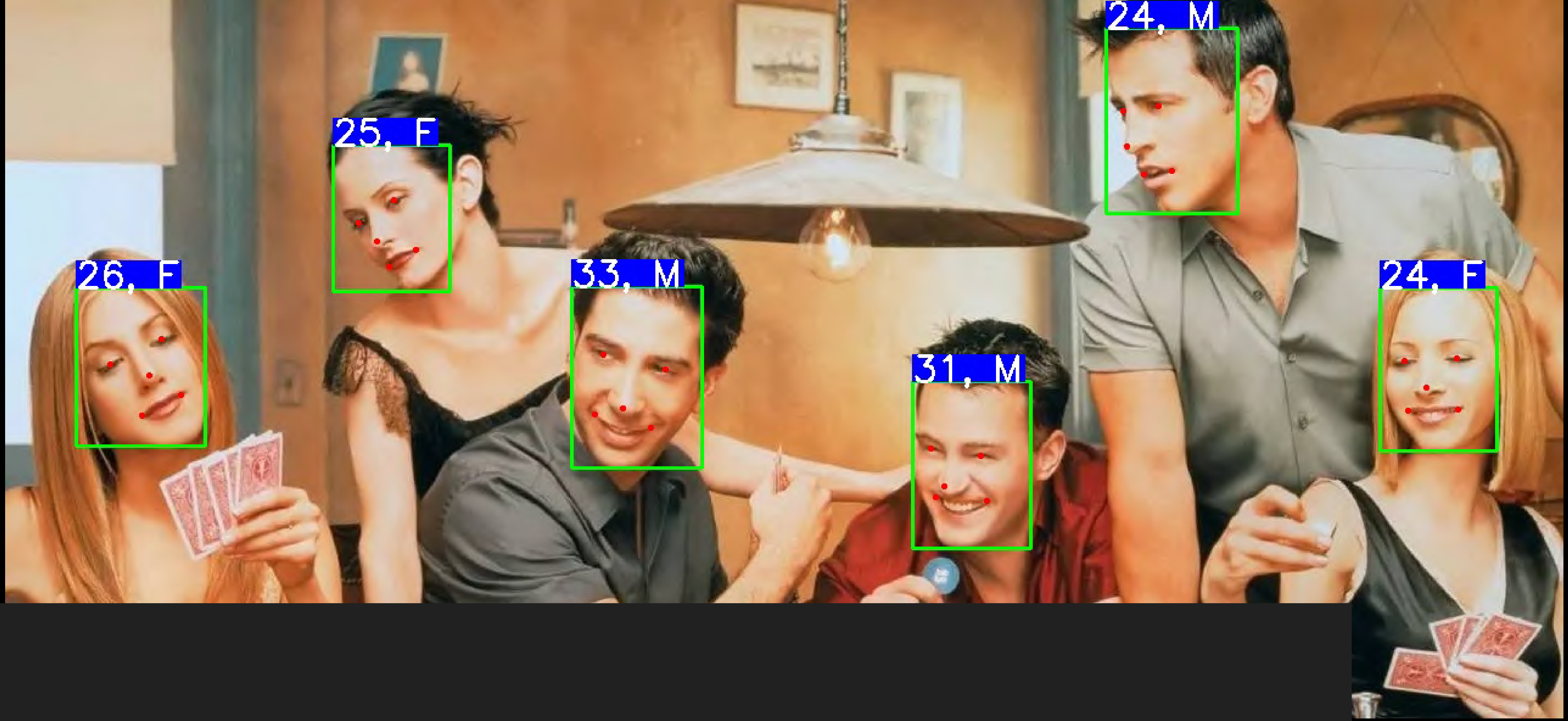


Team 11

Bosch's Age And Gender Detection

Pitch

Objectives

What an ideal solution should comprise

1. **Use maximum information**

   Should use both face and body information as well as information from all frames subject is present in.

2. **Have good fps rate**

   Include optimizations for ensuring near real-time playback of video.

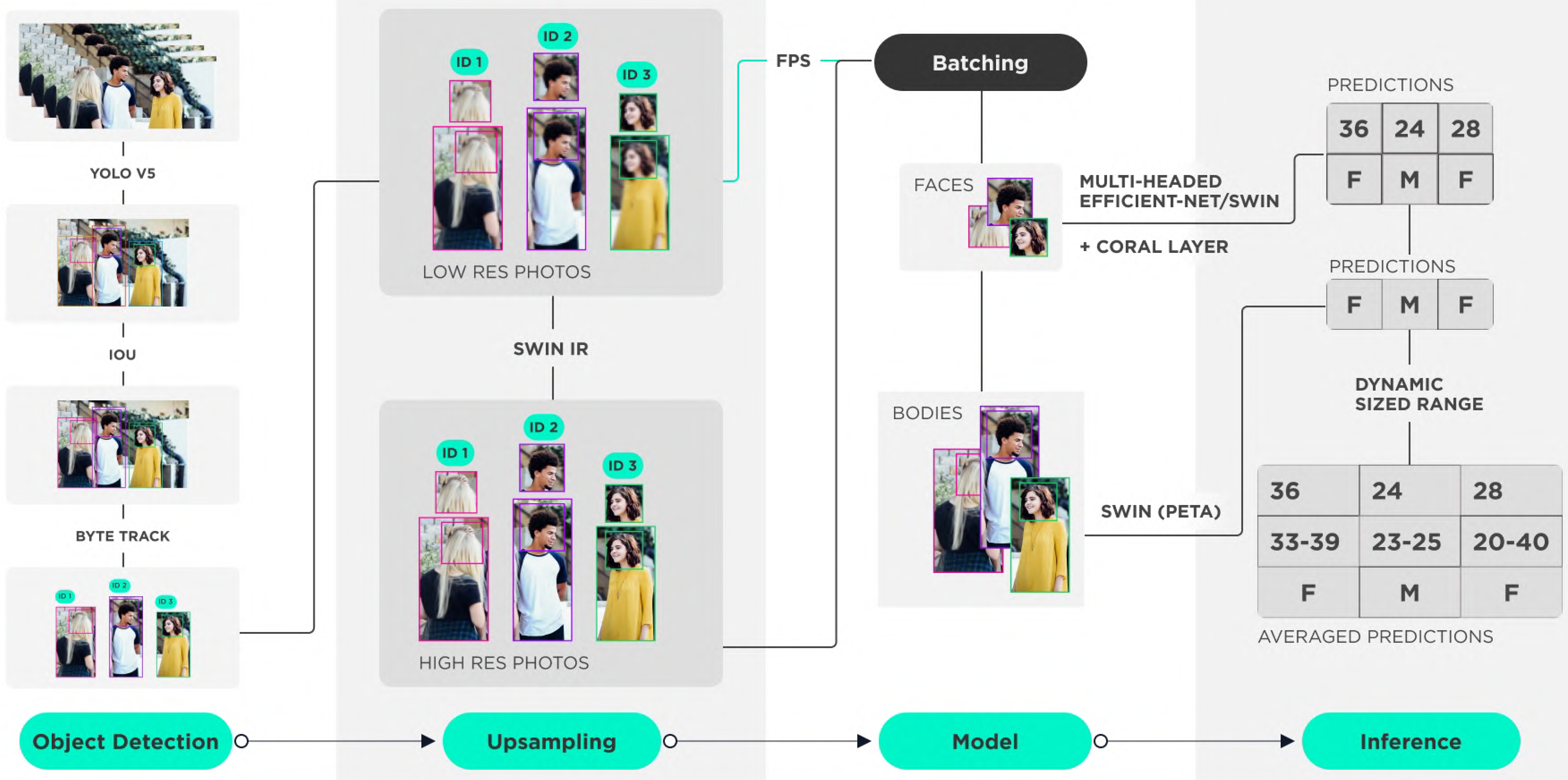
3. **Usable and customizable**

   Results should be usable and interpretable. The pipeline should have modularity and be customizable.

Components

Object detection + Tracking → Upscaling → Models → Inference

Pipeline



Pitch

Detecting body and faces

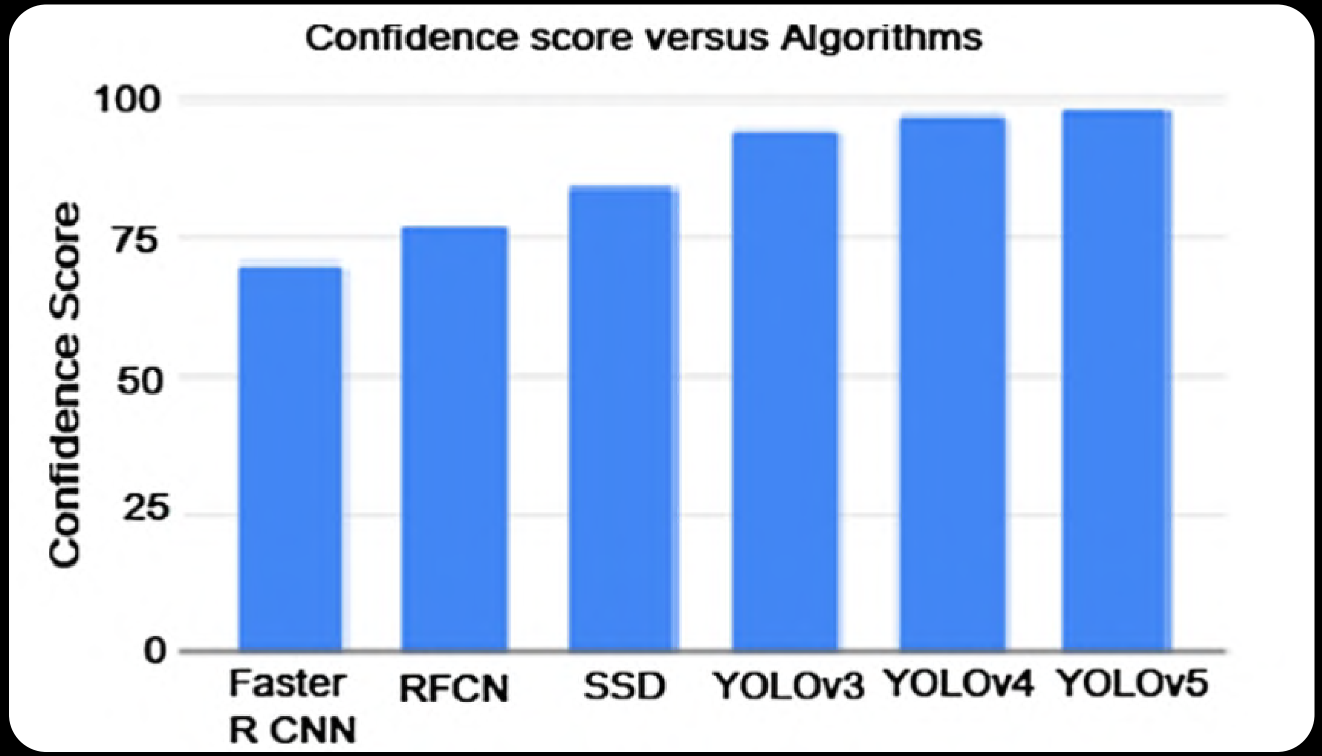
Object Detection + Tracking

YOLOv5

YOLO stands for **"You Only Look Once"**

It is the SOTA algorithm used for multi-object detection tasks which give high accuracy even when applied in real-time on videos.

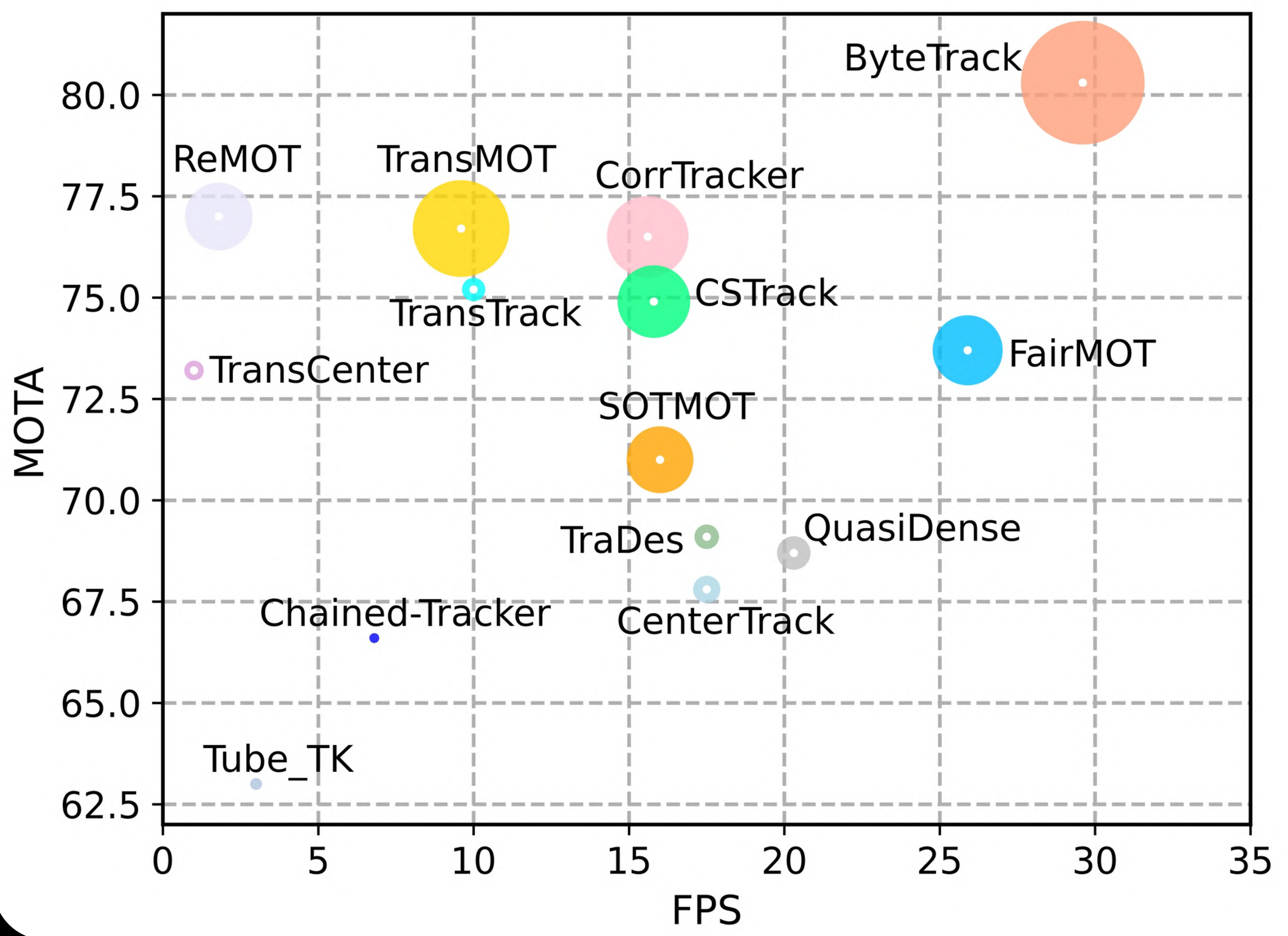
We have fine-tuned YOLOv5-large on CrowdHuman dataset to get better human detections.



ByteTracker

It is the SOTA algorithm for multi-object tracking.

It's a fairly simple and fast algorithm compared to other commonly used trackers like DeepSort.

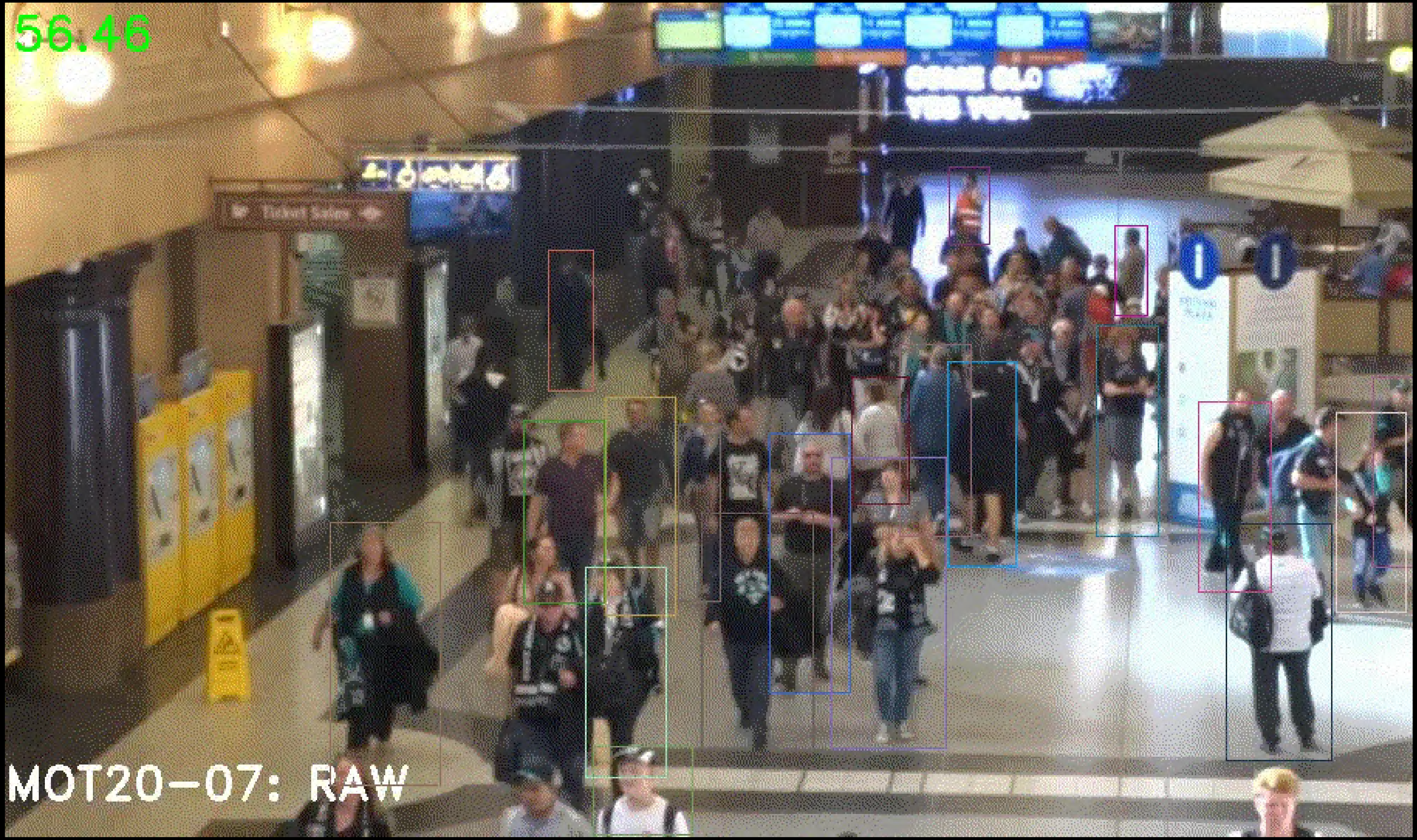


Pitch

ByteTracker

To track the detected objects and assign them a unique ID throughout the frames we need a tracker

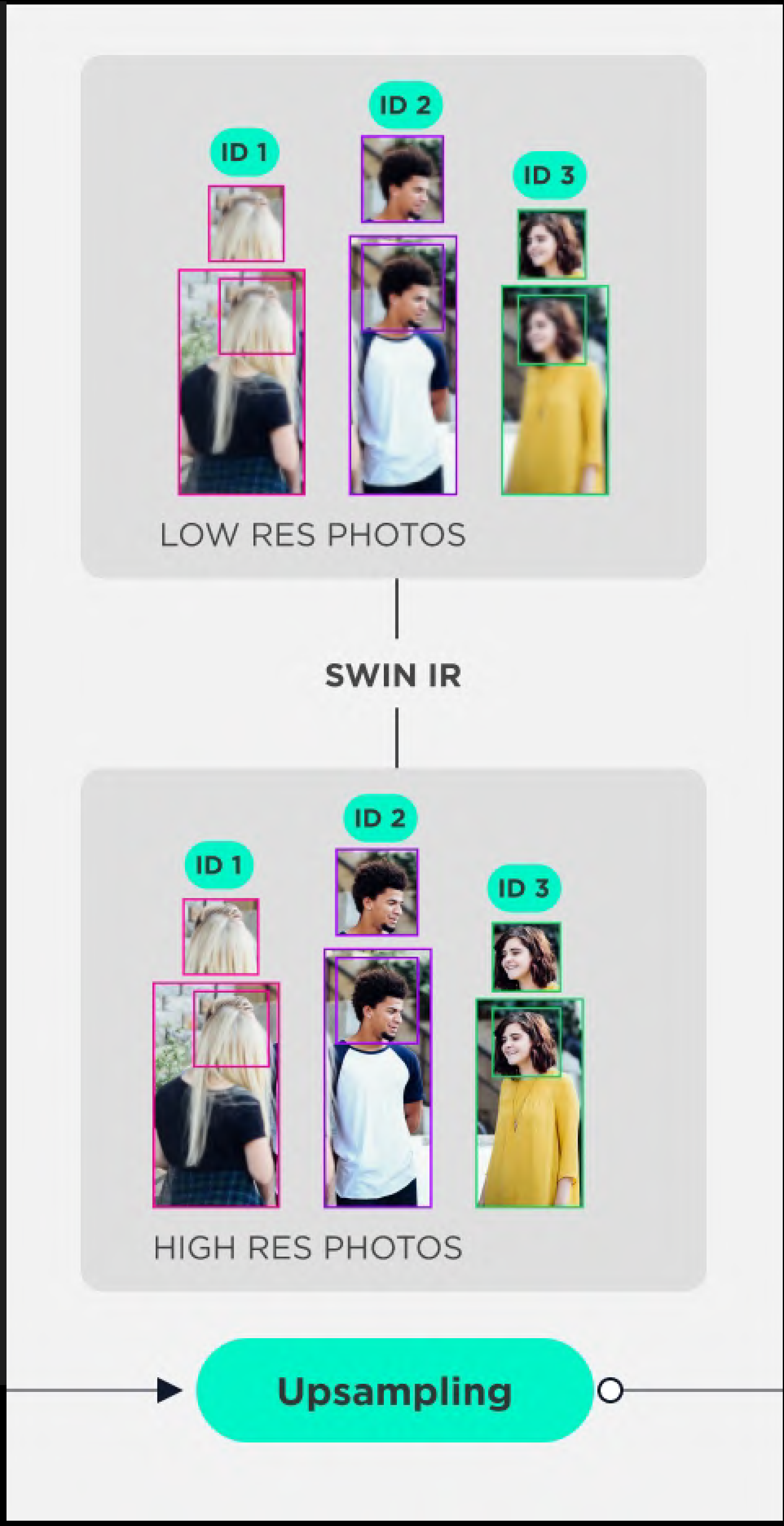
We used modified IOU scores to assign the face to bodies which have the maximum score.



Pitch

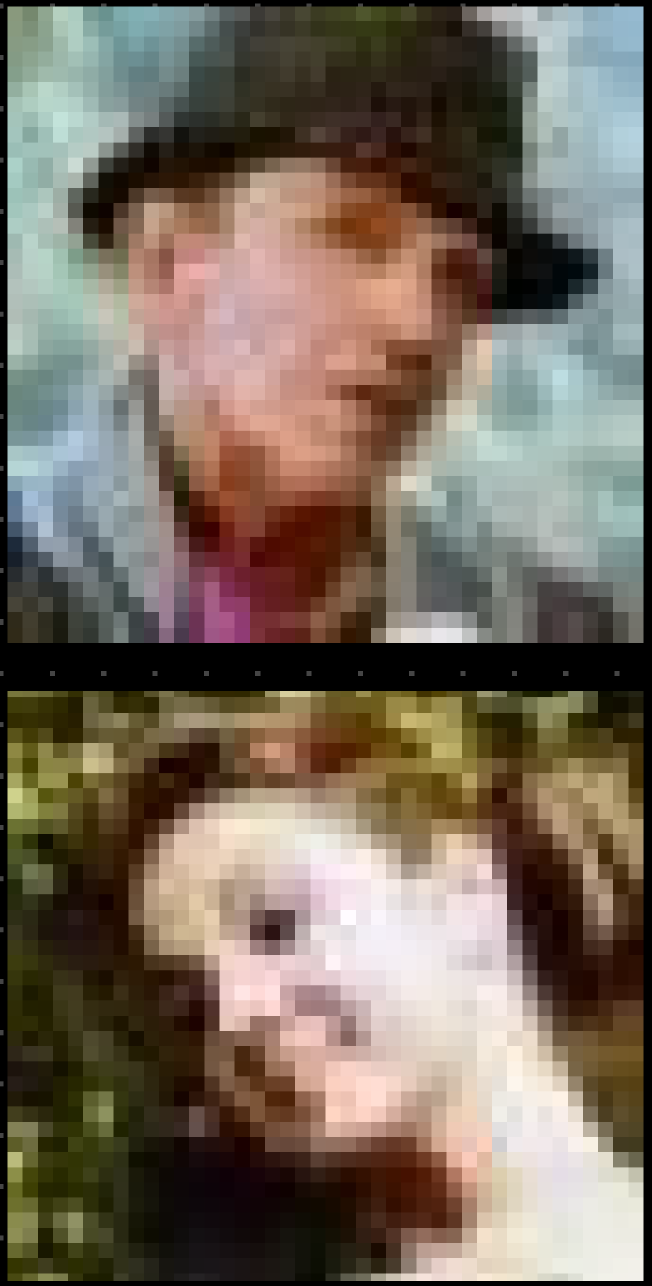
Converting low res faces to high res

Upscaling

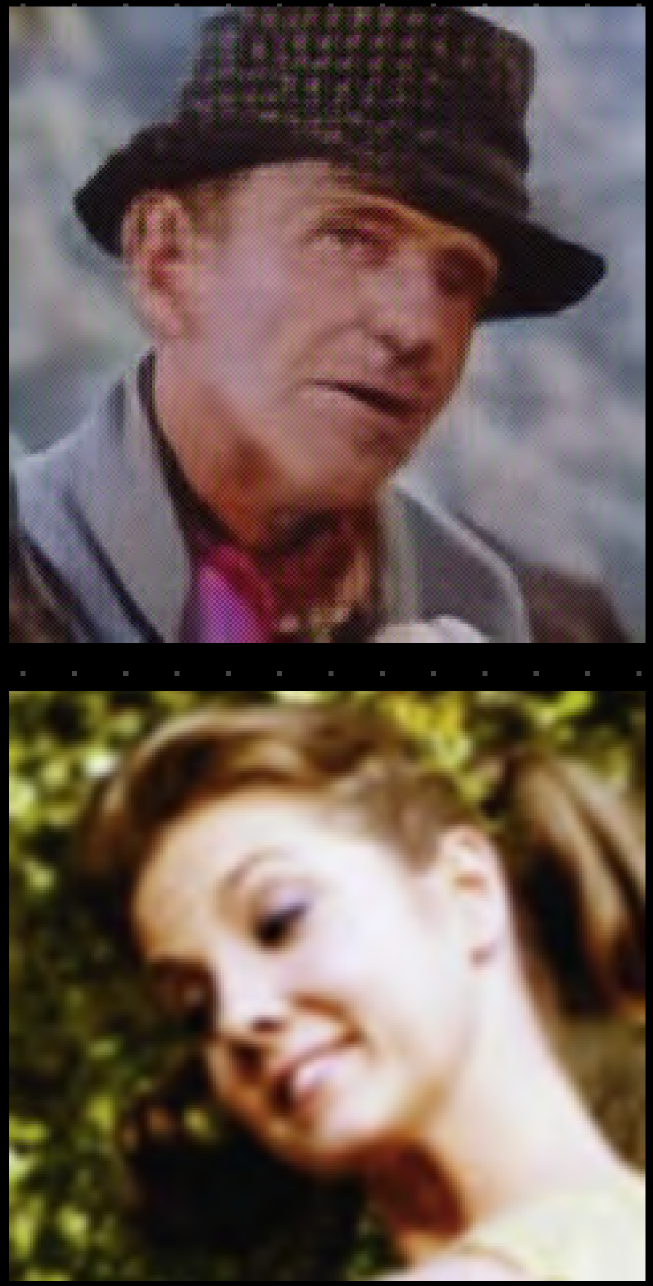


SWIN-IR vs SRGAN

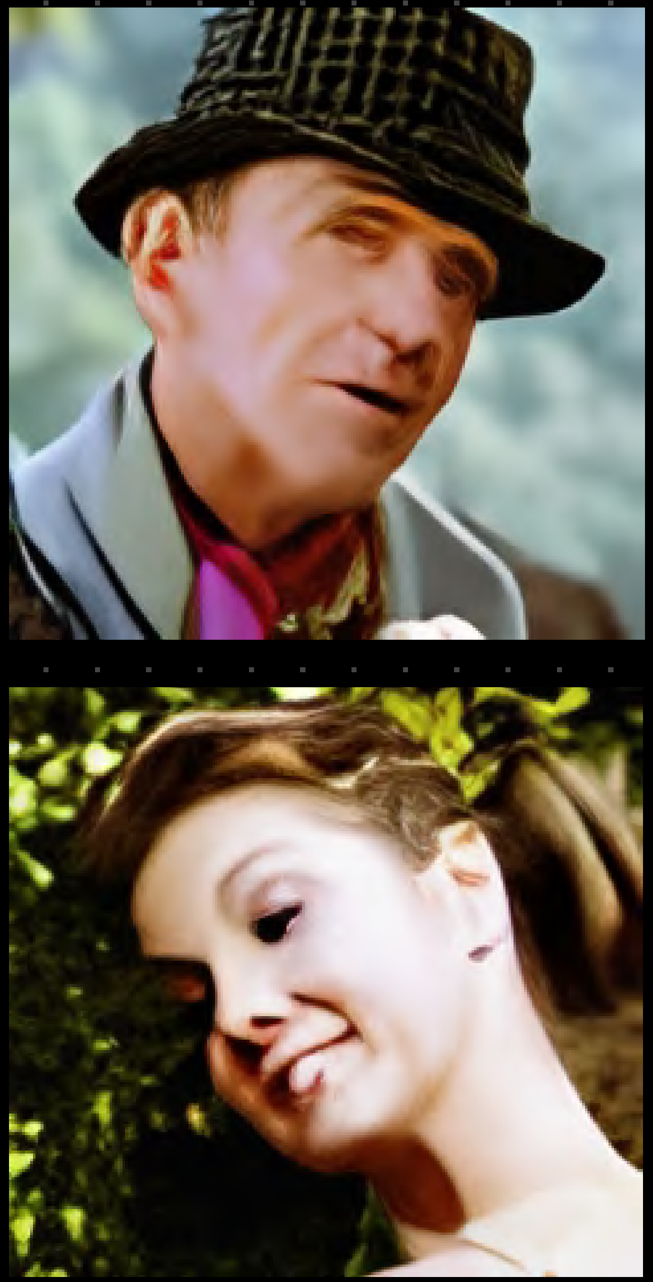
Low Res



SRGAN



SWIN-IR



We fine-tuned the SRGAN on the IMDB Face dataset, the results were good but Swin-IR performed much better.

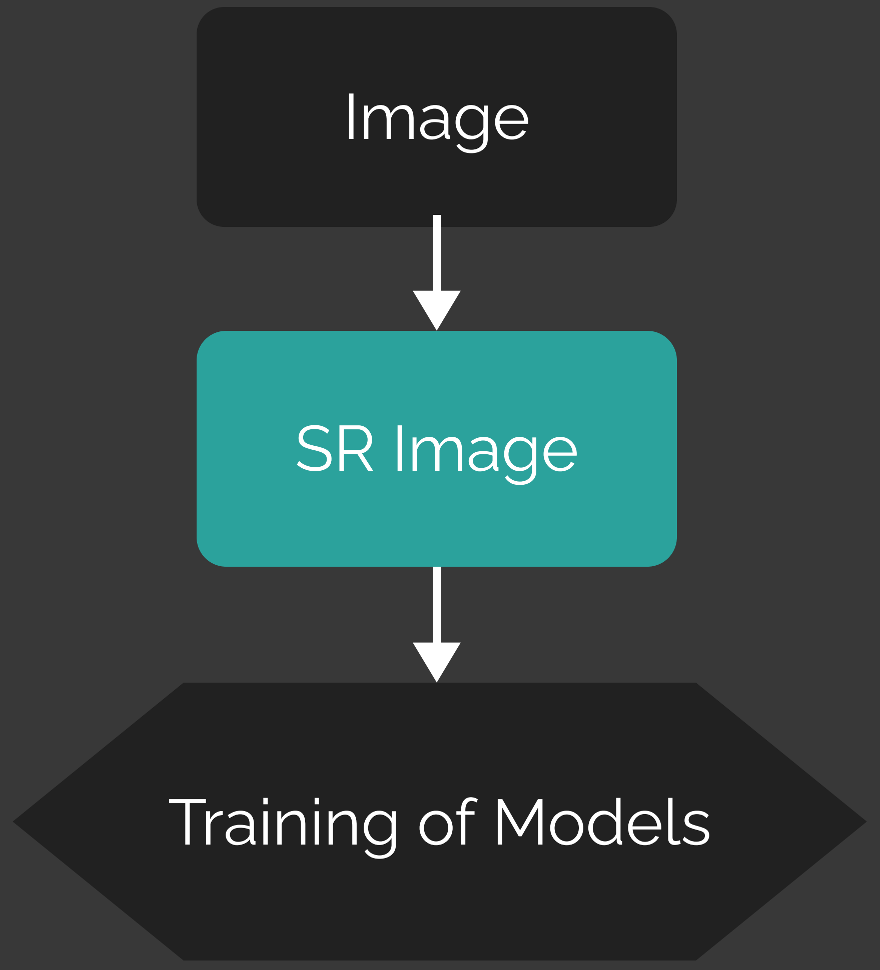
SWIN-IR

SwinIR outperforms SOTA methods on multiple tasks, while the number of parameters can be reduced by up to 67%.

All faces of a frame are passed as batches instead of individual face to save time.

This step can be skipped to increase FPS.

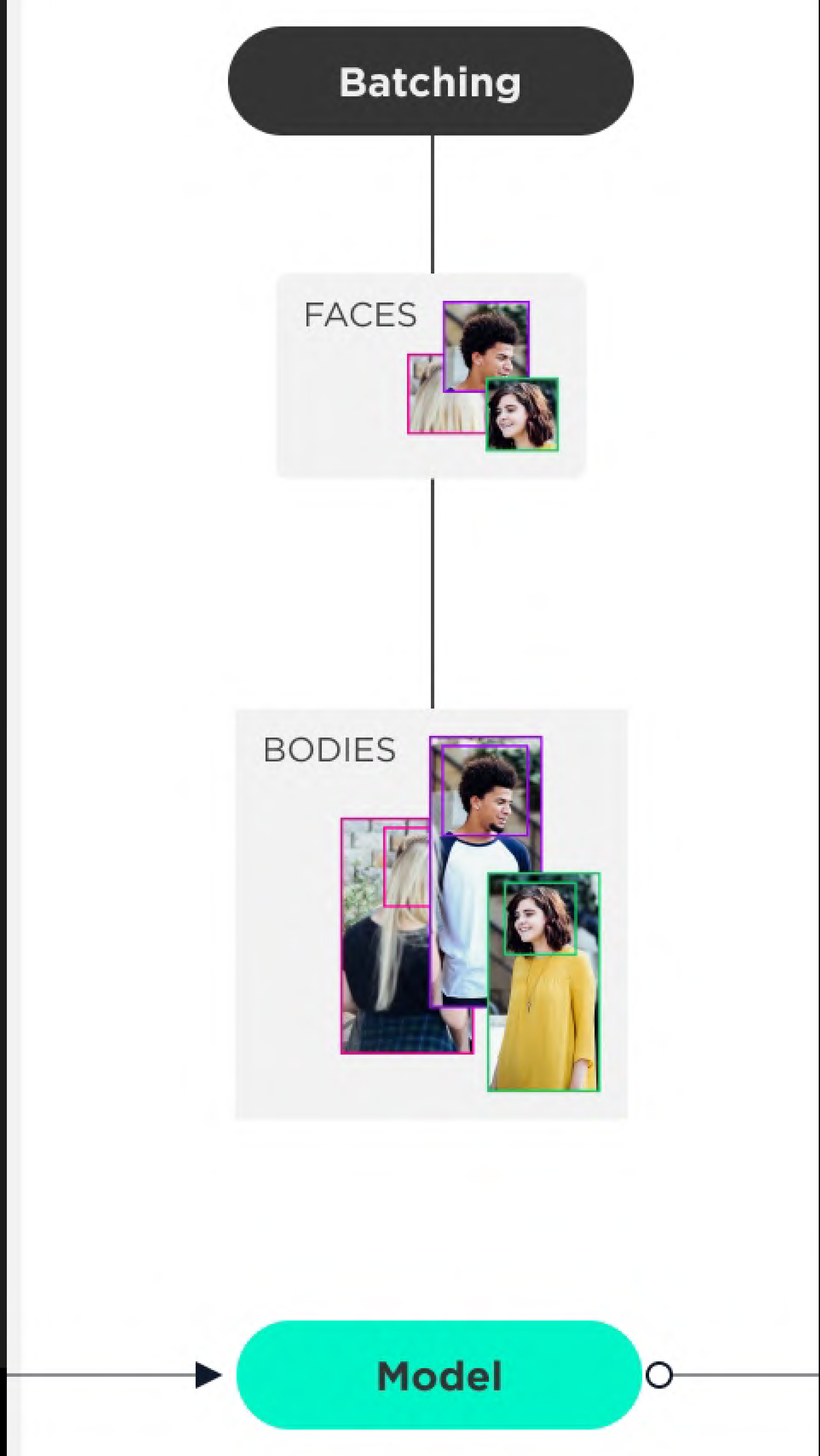
Models were trained on the outputs of the SWIN-IR model.



Pitch

Models used for inference

Models



SWIN Transformers

The SOTA general-purpose image-based transformer model.

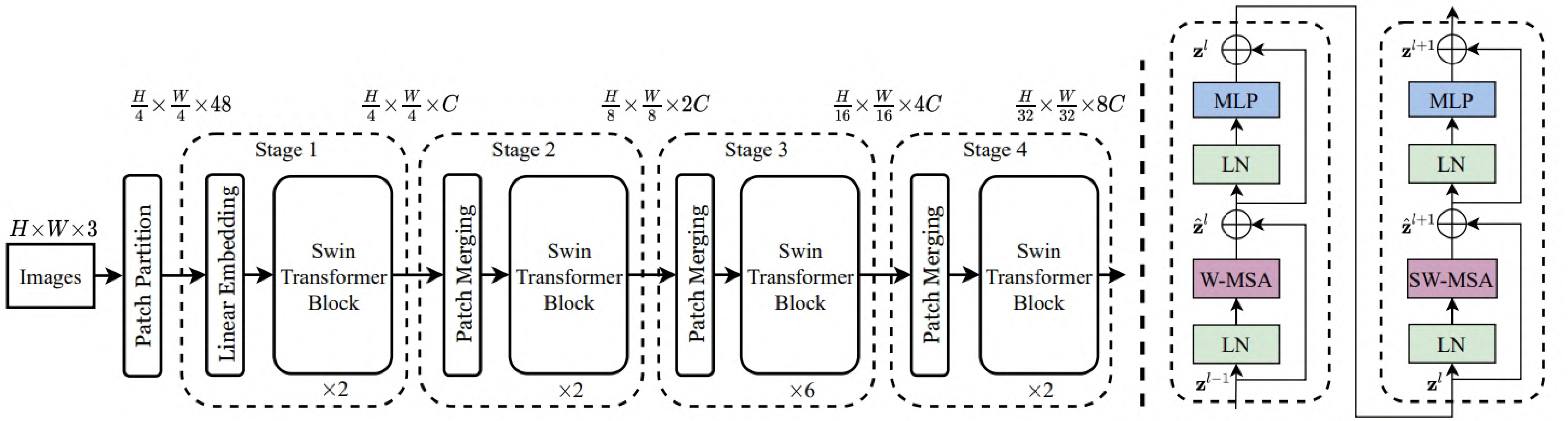
Outperforms all other CNN-based models for multiple tasks with minimum parameters.

Swin 1: Trained on the outputs of SwinIR on UTK Face Dataset, for age and gender prediction.

Swin 2: Trained on PETA Dataset for gender prediction.

(a) Regular ImageNet-1K trained models

method	image size	#param.	FLOPs	throughput (image / s)	ImageNet top-1 acc.
RegNetY-4G [48]	224^2	21M	4.0G	1156.7	80.0
RegNetY-8G [48]	224^2	39M	8.0G	591.6	81.7
RegNetY-16G [48]	224^2	84M	16.0G	334.7	82.9
EffNet-B3 [58]	300^2	12M	1.8G	732.1	81.6
EffNet-B4 [58]	380^2	19M	4.2G	349.4	82.9
EffNet-B5 [58]	456^2	30M	9.9G	169.1	83.6
EffNet-B6 [58]	528^2	43M	19.0G	96.9	84.0
EffNet-B7 [58]	600^2	66M	37.0G	55.1	84.3
ViT-B/16 [20]	384^2	86M	55.4G	85.9	77.9
ViT-L/16 [20]	384^2	307M	190.7G	27.3	76.5
DeiT-S [63]	224^2	22M	4.6G	940.4	79.8
DeiT-B [63]	224^2	86M	17.5G	292.3	81.8
DeiT-B [63]	384^2	86M	55.4G	85.9	83.1
Swin-T	224^2	29M	4.5G	755.2	81.3
Swin-S	224^2	50M	8.7G	436.9	83.0
Swin-B	224^2	88M	15.4G	278.1	83.5
Swin-B	384^2	88M	47.0G	84.7	84.5



(a) Architecture

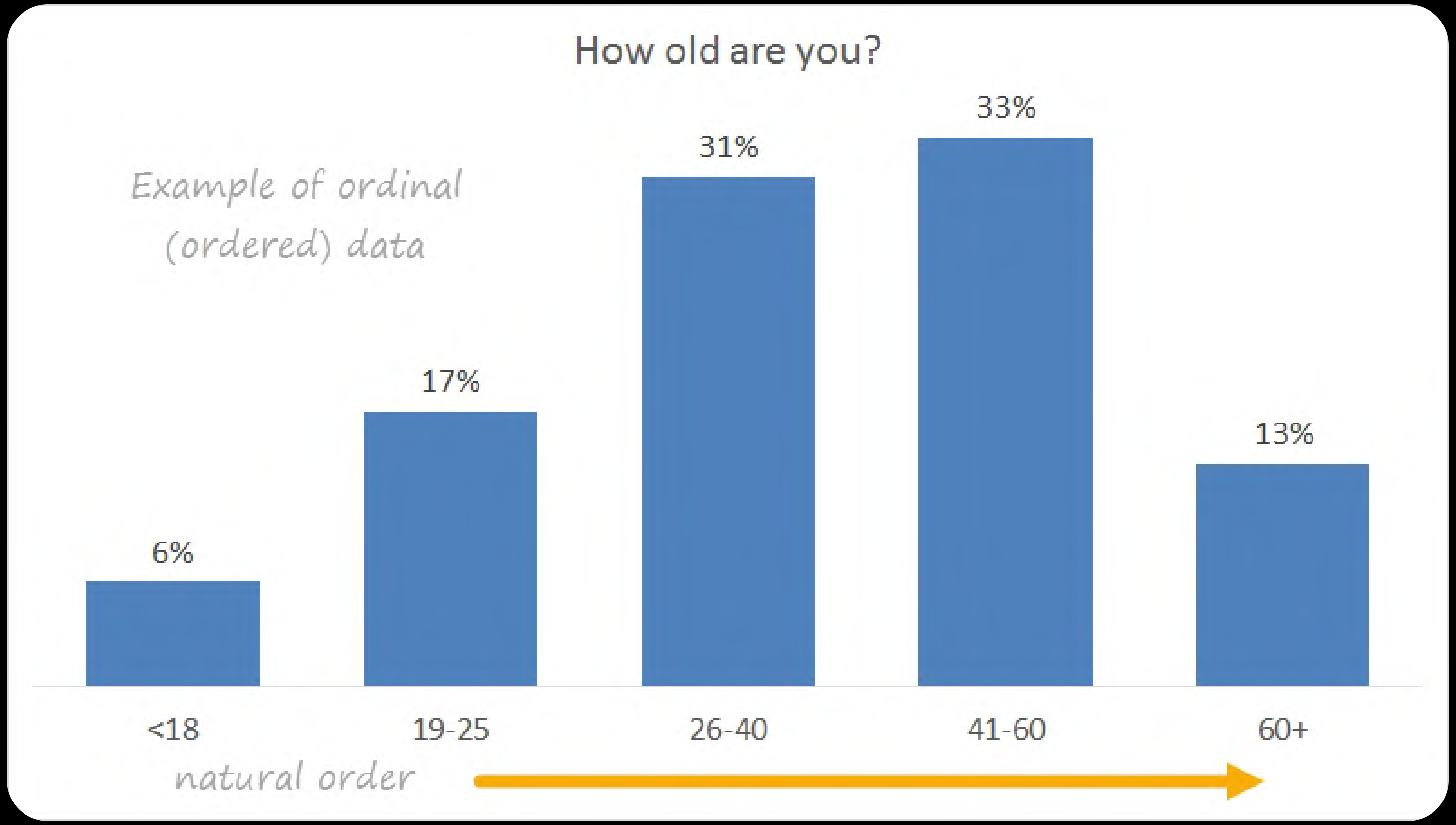
(b) Two Successive Swin Transformer Blocks

Ordinal Regression not classification

Used for predicting variables that exist on an arbitrary scale where only the relative ordering between different values is significant.

Due to the fact that the facial aging process is a non-stationary process, one reliable information we can use would be the relative order among the age labels in addition to their exact values. Hence, the age estimation is cast as an **ordinal regression** problem

We have implemented ordinal regression using Coral Layer.



CORAL Layer

Added as the last layer of all models

Consistent Rank Logits Framework (CORAL)

Used for Ordinal Regression of Age Prediction

Much better than the classification of images for age prediction.

CORAL layer also has as associated loss known as CORAL loss

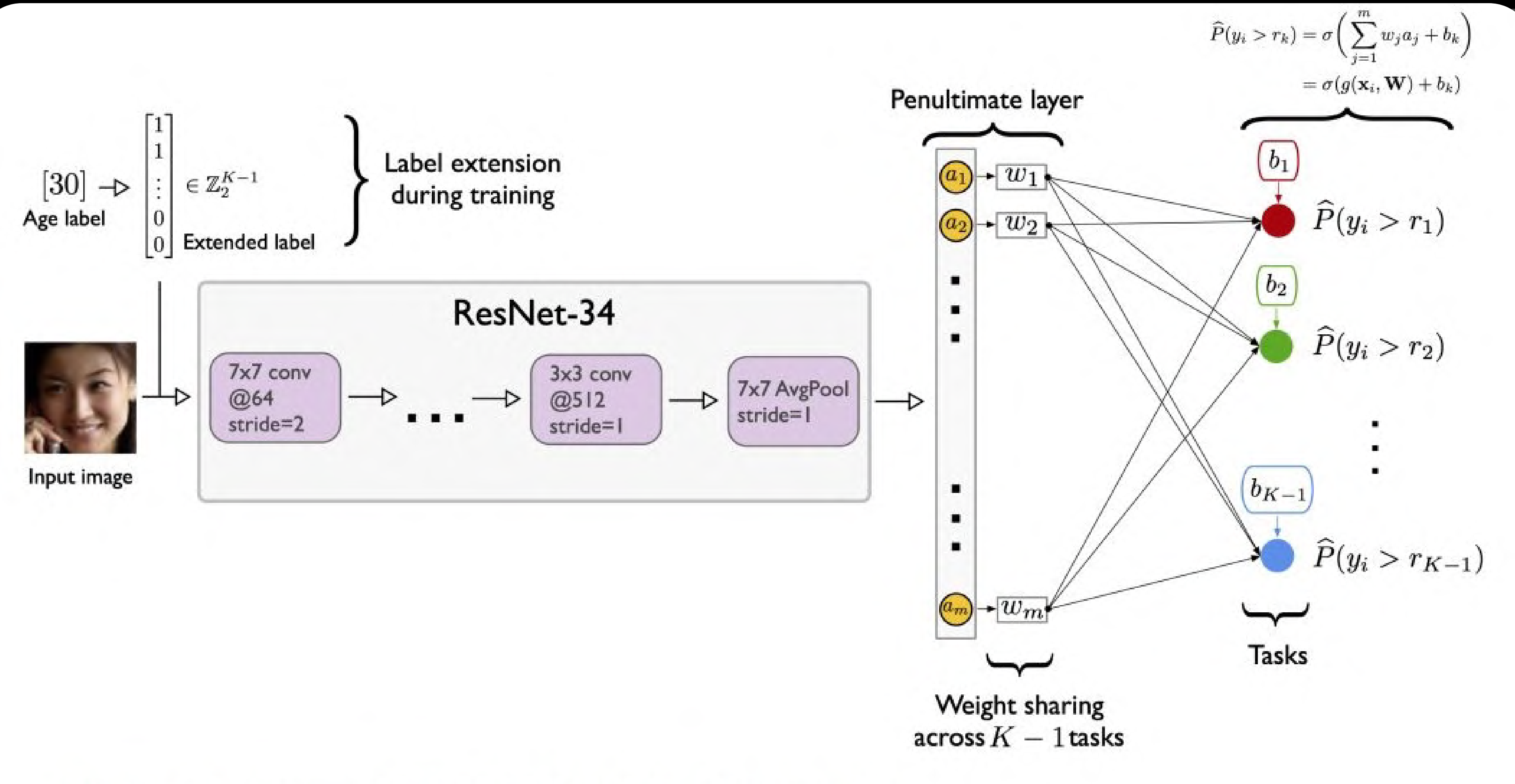


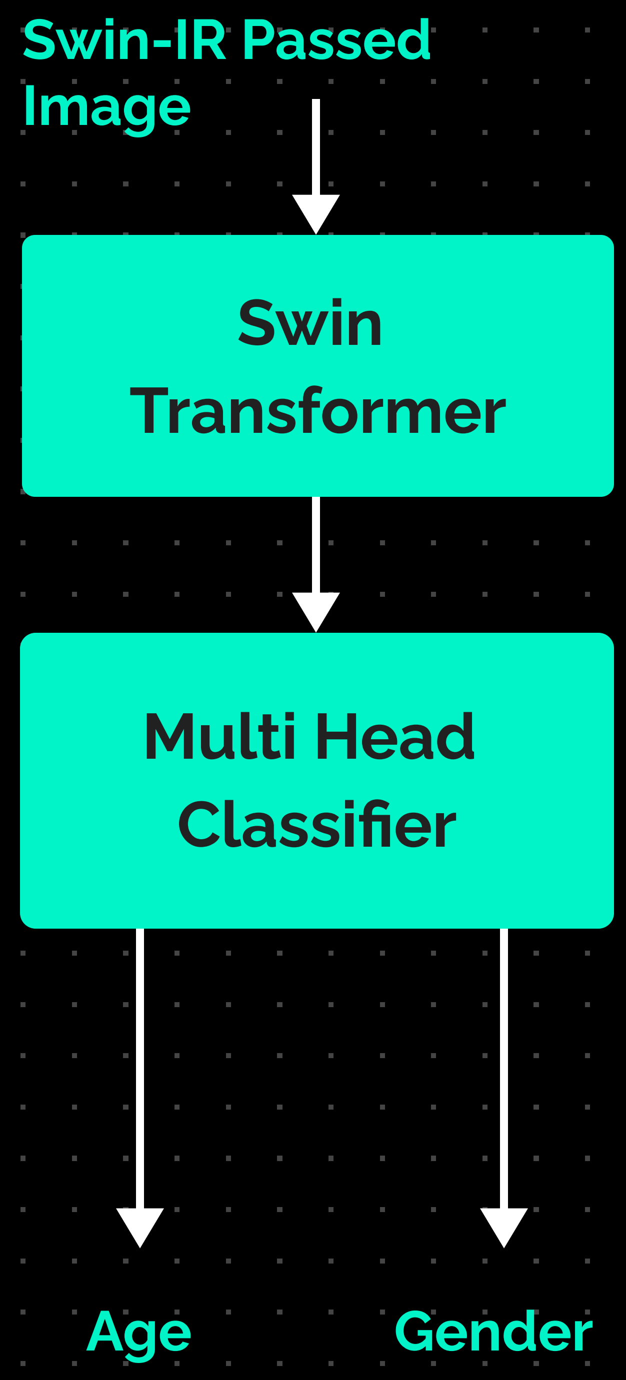
Illustration of the consistent rank logits CNN (CORAL-CNN) used for age prediction.

Multi-Head Classifier

Used for combined prediction of Age and Gender from a single backbone of Swin

Reduces the training as well as prediction time

Increased Scores for both tasks.



Comparisons

Model	MAE (Age)	Accuracy (Gender)
Multi-Head ResNet50 + Coral	7.069	89.7%
Multi-Head EfficientNet + Coral	6.271	91.9%
Mutli-Head Swin + Coral	5.756	93.8%
Multi-Head Swin + Coral (Trained on outputs of SwinIR)	**5.294**	**94.2%**

Multi-Head Models

Model	Accuracy
Baseline Model	85.5%
EfficientNet	90.9%
Resnet50	91.2%
Swin	**94.4%**

Single-Head Models

Datasets

Dataset	Size	Description	Model
CrowdHuman	470K	Contains head, human visible-region, and human full-body bounding box.	YOLOv5 for Human Body and Face Detection
IMDB	1.7M	Fine-tuning of SRGAN model.	SRGAN for SR Task
UTKFace	23K	Images with age, gender, and ethnicity.	Multi head Swin with CORAL for Age and Gender
PETA	19K	Pedestrian images from CCTV Surveillance	Swin Classifier for Gender

Datasets

CrowdHuman

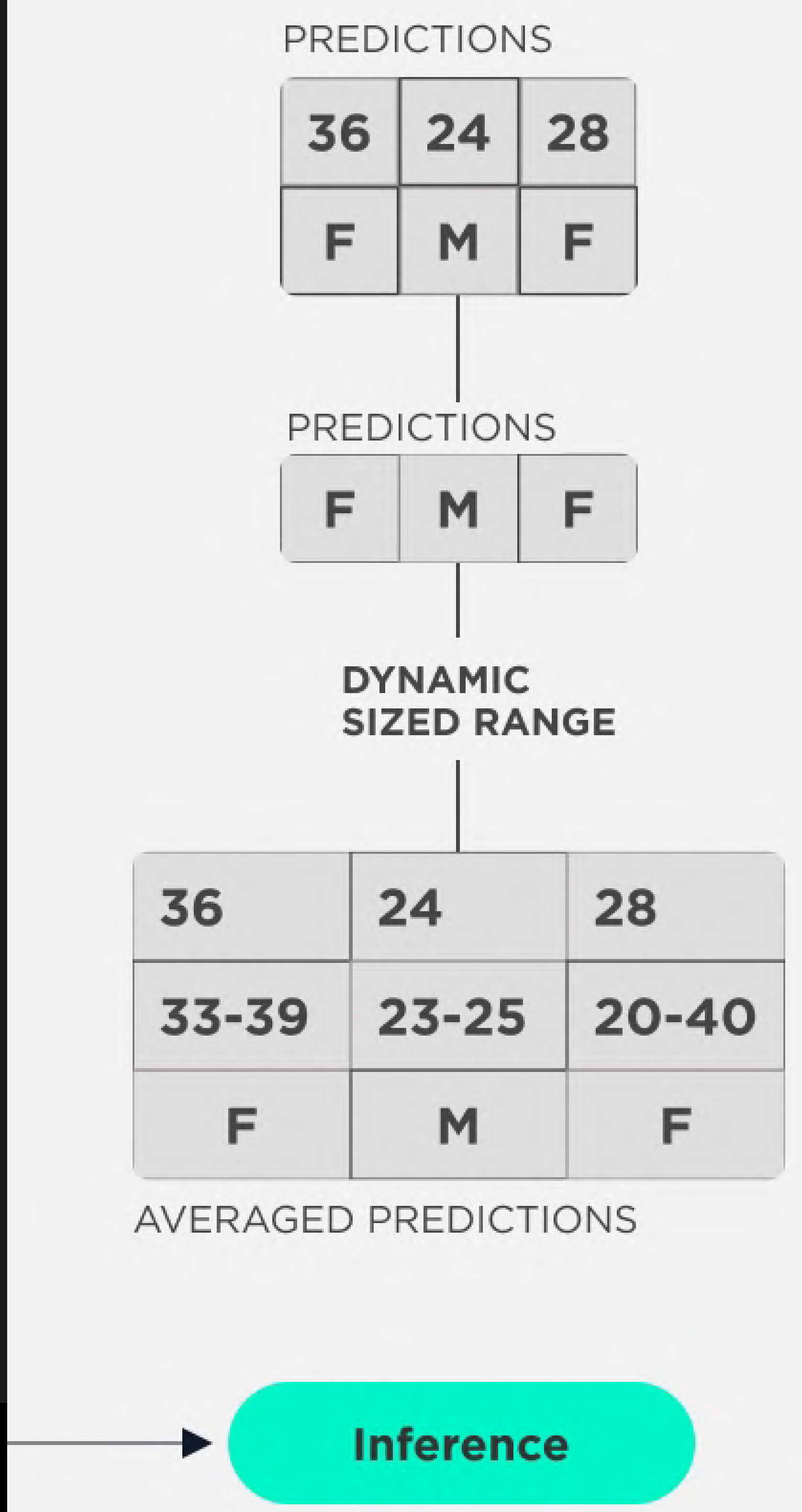
Imdb Image

UTKFace

PETA

Combining all information for final prediction

Inference



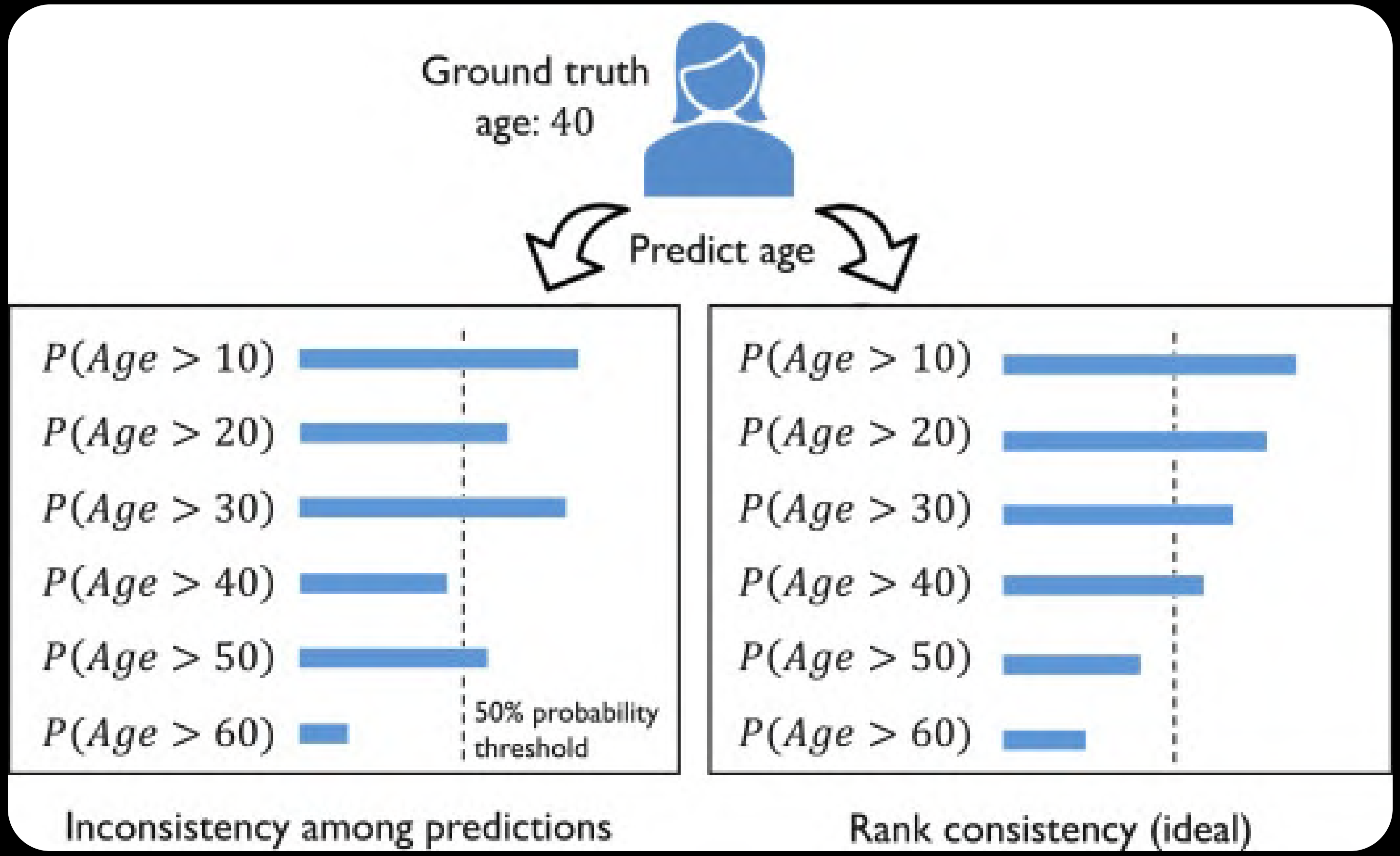
Pitch

Dynamically sized Age ranges

Ordinal regression gives consistently ranked probability scores for each class in the form of a cumulative distribution function

From these scores, we can infer dynamically sized boxes based on how confident we are in our prediction

This allows us to get different range lengths based on our confidence



Averaging over all frames

Frame 1		Frame 2		Frame 3		Frame 3
ID 1 : 31, M	+	ID 1 : 32, M	+	ID 1 : 30, F	=	ID 1 : 32, M
ID 2 : 20, F		ID 2 : 22, M		ID 2 : 21, F		ID 2 : 21, F
ID 3 : 56, M		ID 3 : 55, F		ID 3 : 54, M		ID 3 : 55, M

The age detected in each frame have been averaged over all the previous frame detections.

The most frequent gender detections (mode) over all the frames have been used

Novelties

1. Using State-of-the-art ByteTracker for tracking humans across consecutive frames of the video.

2. Stabilizing the age prediction by taking a moving average across multiple frames.

3. Using body features as well was facial features for enhanced predictions

4. Used the architecture of **SWIN-IR** which outperforms SRGAN and gives more viable results for improving video resolution of CCTV footage.

Novelties

5. Using SWIN-IR model architecture as a teacher for Gender and age models.

6. **Multi-head classifier** for parallely predicting age and gender together.

7. Using **CoRaL layer** for predicting the age using **Ordinal Regression**

8. **Dynamically Sized Age Buckets** based on confidence of logits thus having more accurate predictions.

Future Works

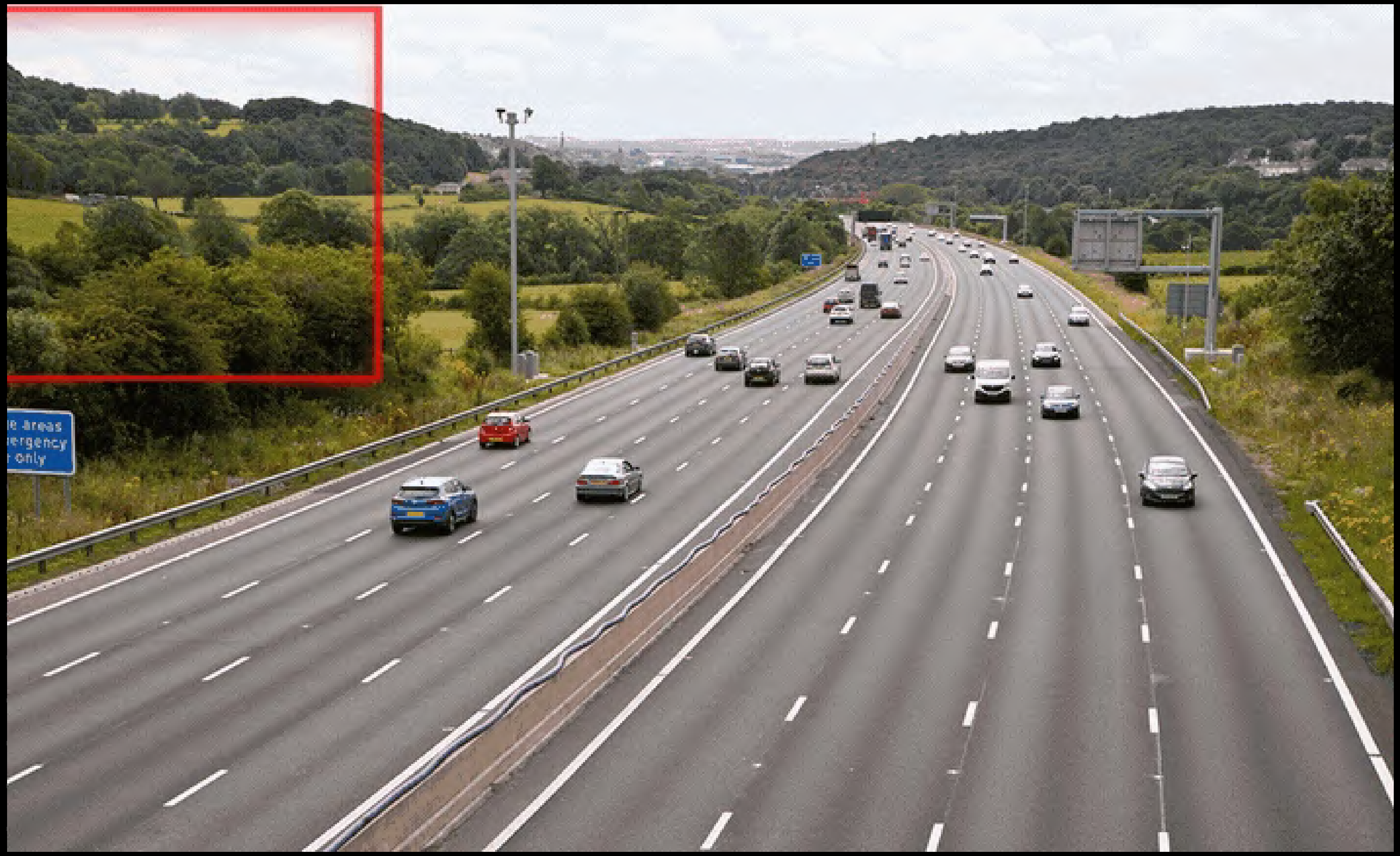
SAHI

YOLO has a drawback when it comes to detecting smaller objects in crowded places..

This will help YOLO detect smaller objects.

Generic slicing aided inference and fine-tuning pipeline for small object detection.

Video super resolution methods to pin point accurate ages for a person spanning across multiple frames



References

He, Xin, et al. "Swin Transformer Embedding UNet for Remote Sensing Image Semantic Segmentation." IEEE Transactions on Geoscience and Remote Sensing (2022).

Cao, Wenzhi, Vahid Mirjalili, and Sebastian Raschka. "Rank consistent ordinal regression for neural networks with application to age estimation." Pattern Recognition

Yifu Zhang, Peize Sun "ByteTrack: Multi-Object Tracking by Associating Every Detection Box"(2021).

Zheng, Zhaohui, et al. "Distance-IoU loss: Faster and better learning for bounding box regression." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. No. 07. 2020.

Takano, Nao, and Gita Alaghband. "Srgan: Training dataset matters." arXiv preprint arXiv:1903.09922 (2019).