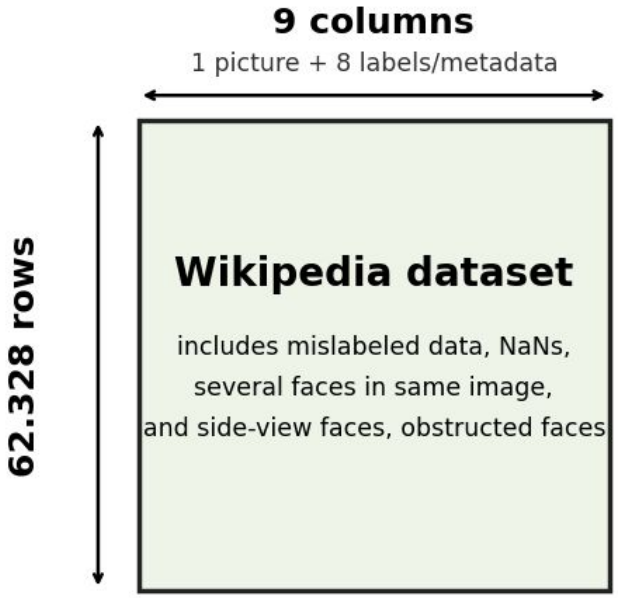
Age & Gender Prediction Using CNN's on Face Images

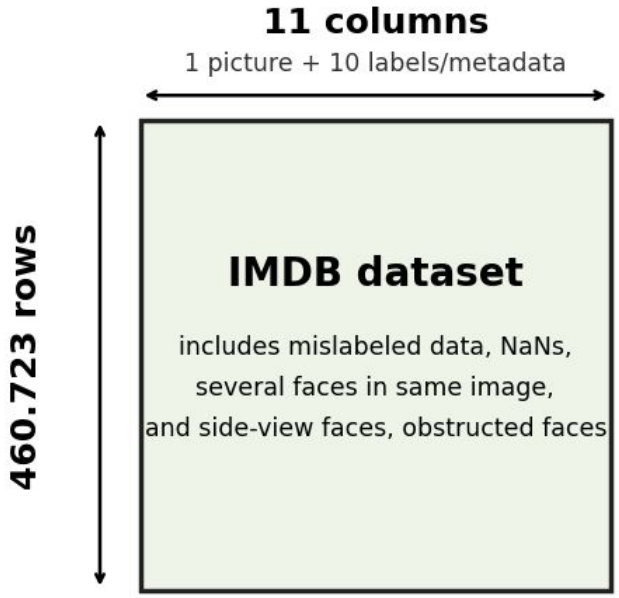
by Victoria, Jona, David, Natalie

https://data.vision.ee.ethz.ch/cvl/rrothe/imdb-wiki/

Data description



labels / metadata:
age, photo_taken, gender, name, face_location, face_score, second_face_score, birth_date



labels / metadata:
birth_date, photo_taken, gender, name, face_location, face_score, second_face_score, birth_date

Targets:

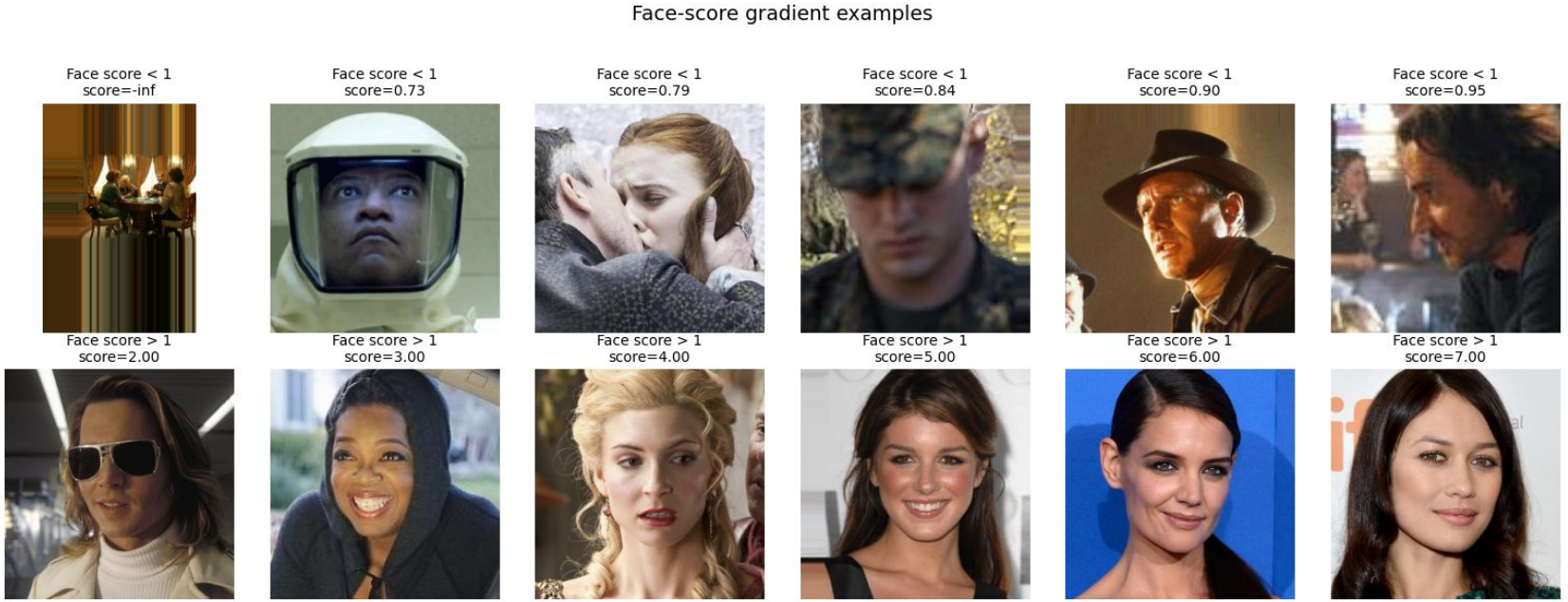
- Gender
- Age

General Cuts

- face_score > 1
- second_face_score is NaN

- Age between 0 and 100
- Gender is **not** NaN

~460k to ~170k images



Clustering

“Research Questions”

2 Models:

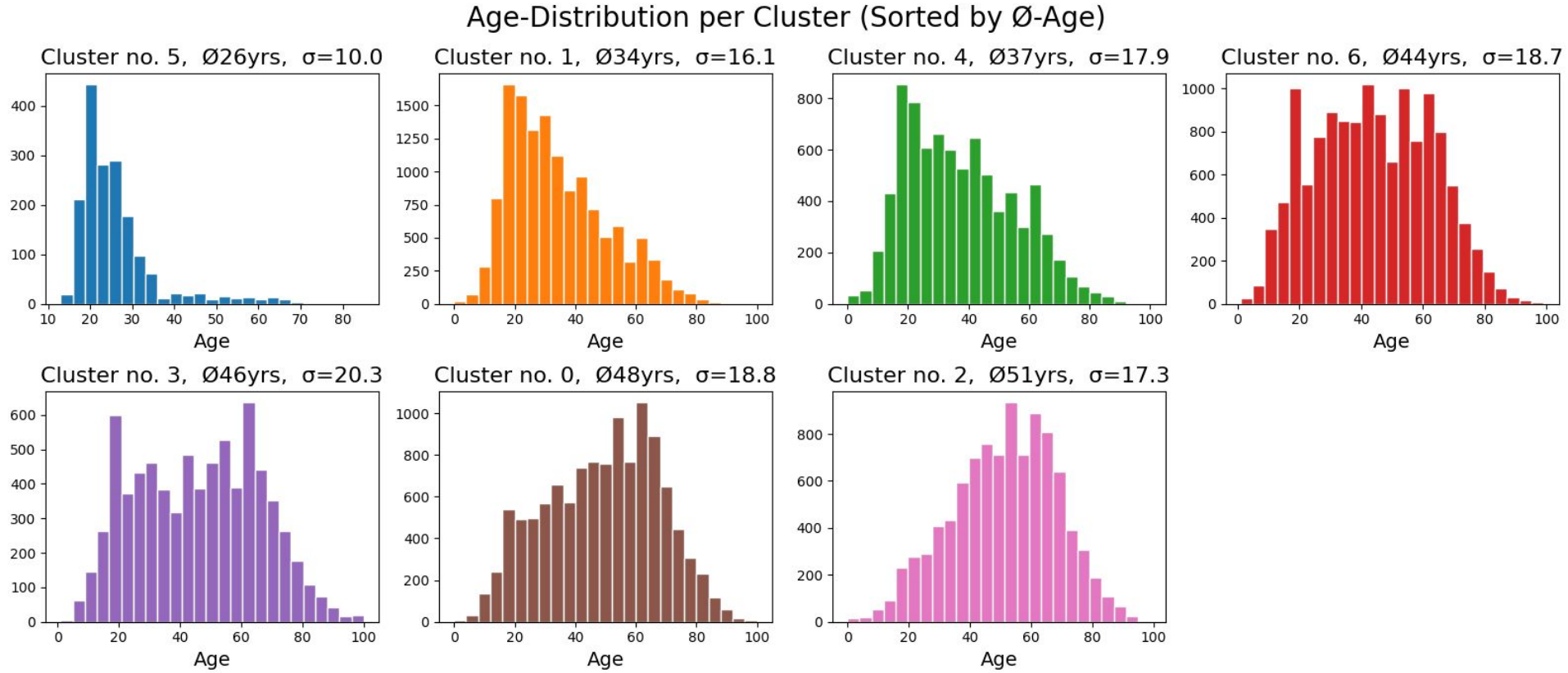
- VGG16
  - Do we find an **underlying structure?**
- Age-Supervised Autoencoder (SAE)
  - Can we **enforce clustering by age?**

General questions:

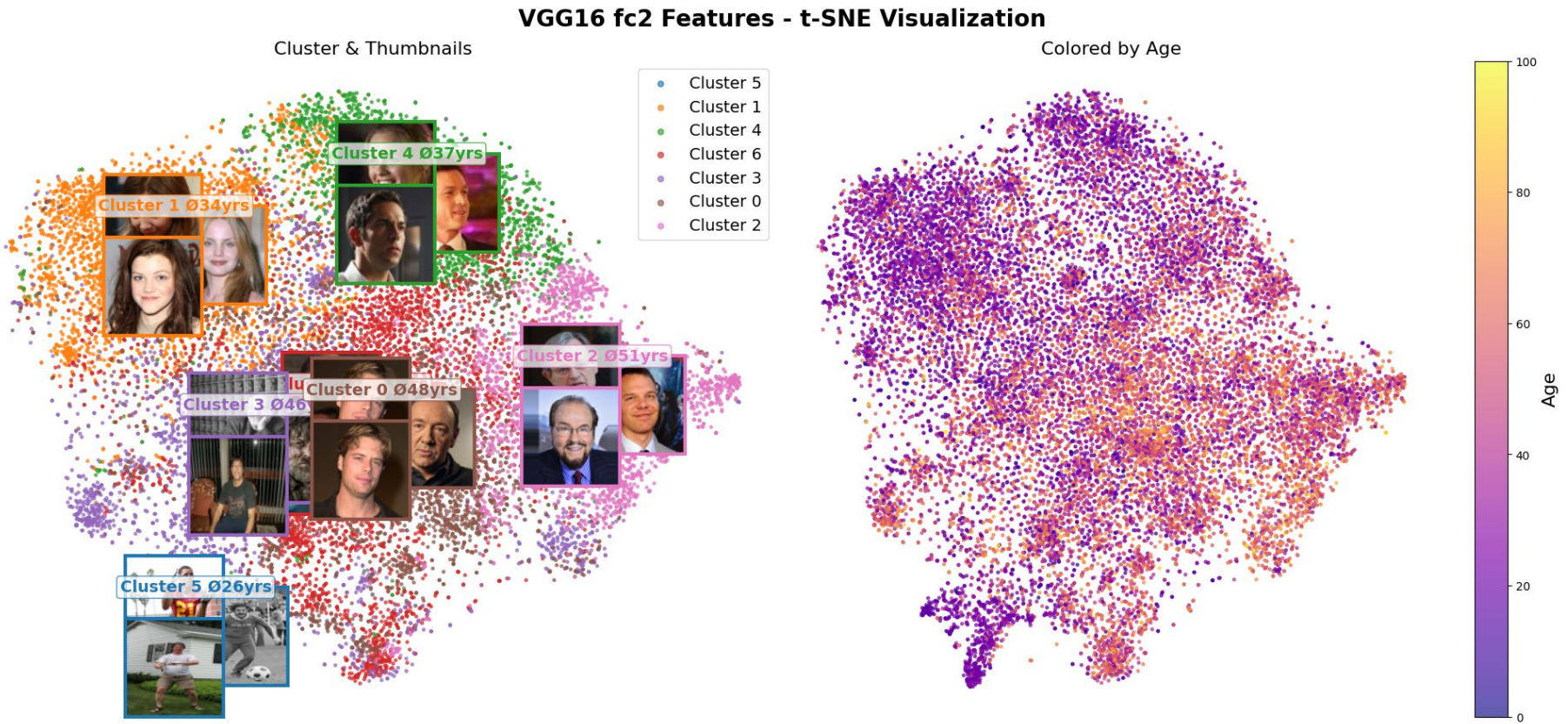
- Influence of the gender?
- Source bias (IMDB vs. Wiki)?

VGG16

VGG16 – K-Means Clustering

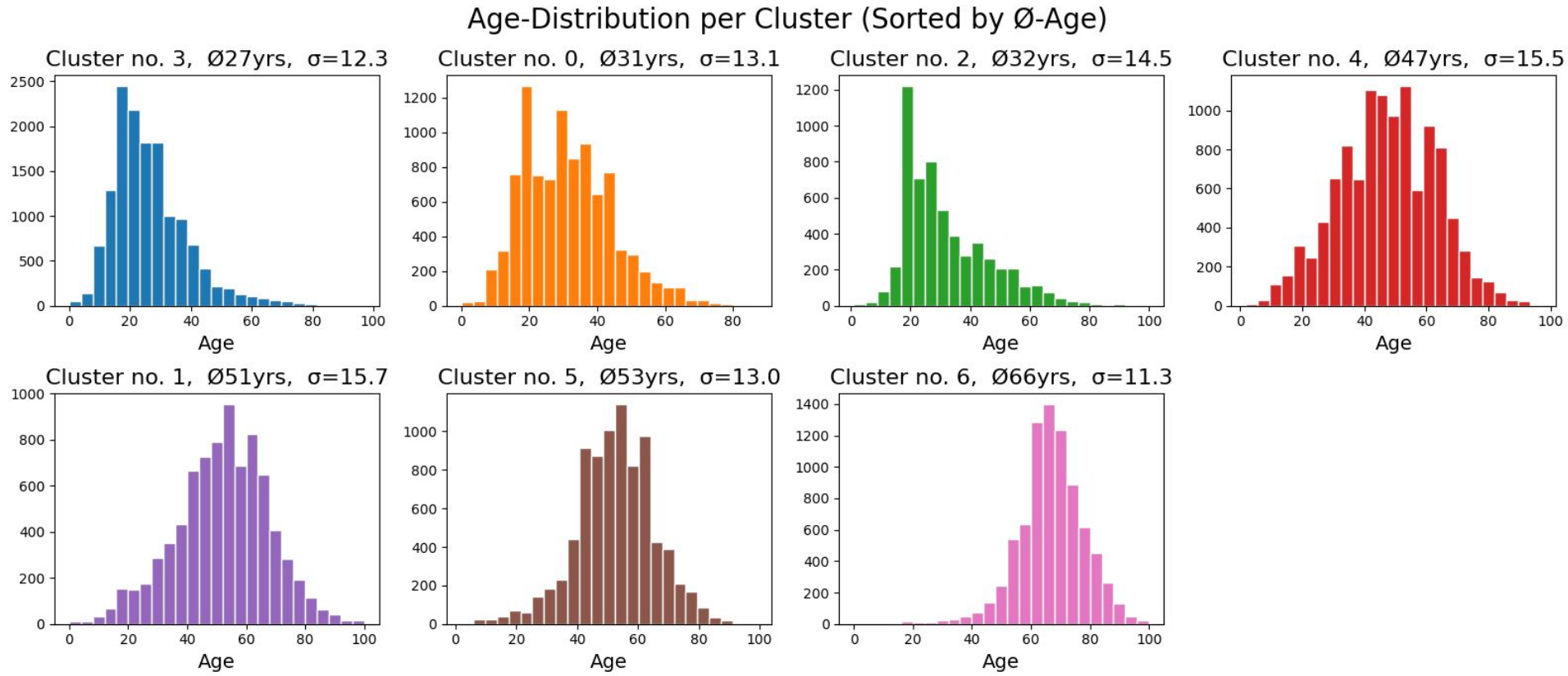


VGG16 (K-Means) - Visualization by t-SNE

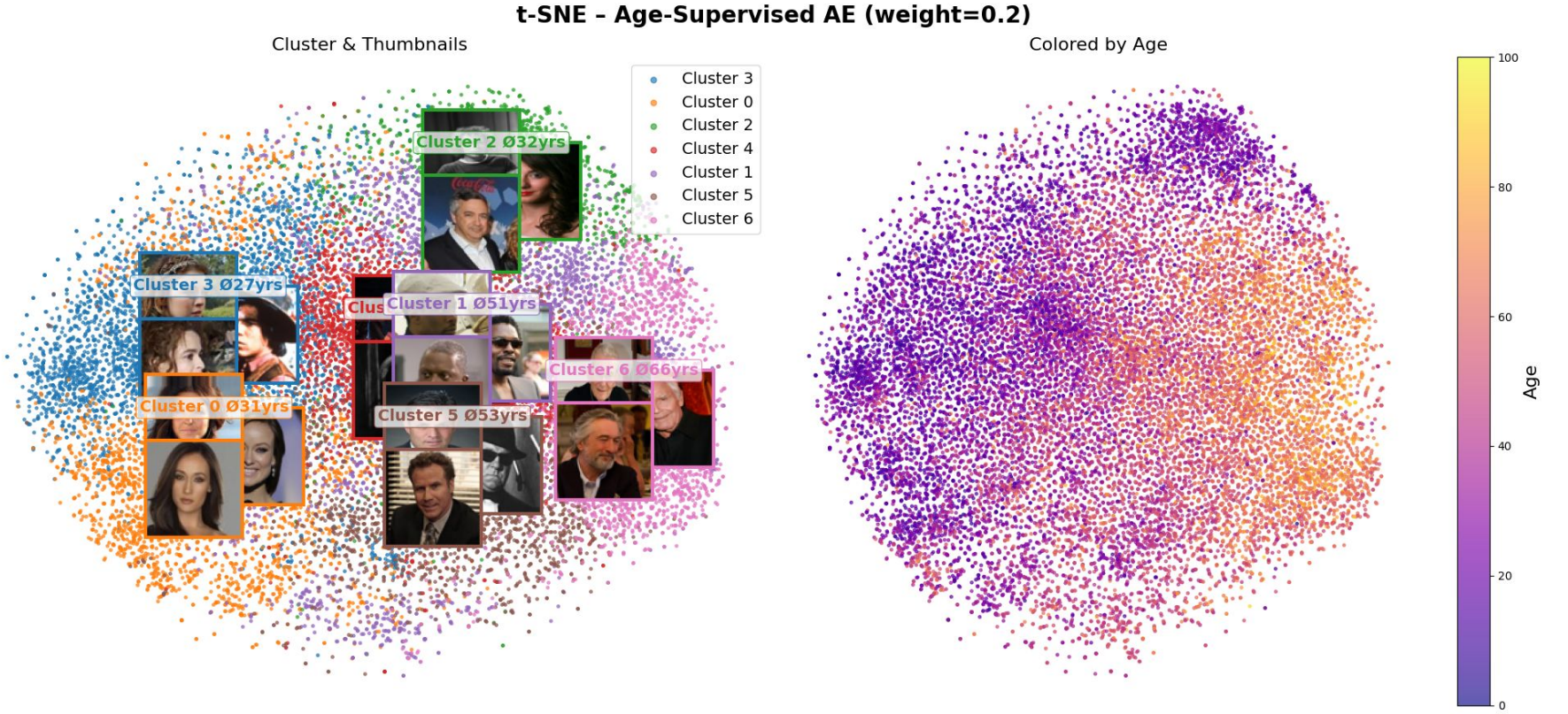


Age-Supervised Autoencoder (SAE)

SAE - Age-Distribution of Clusters

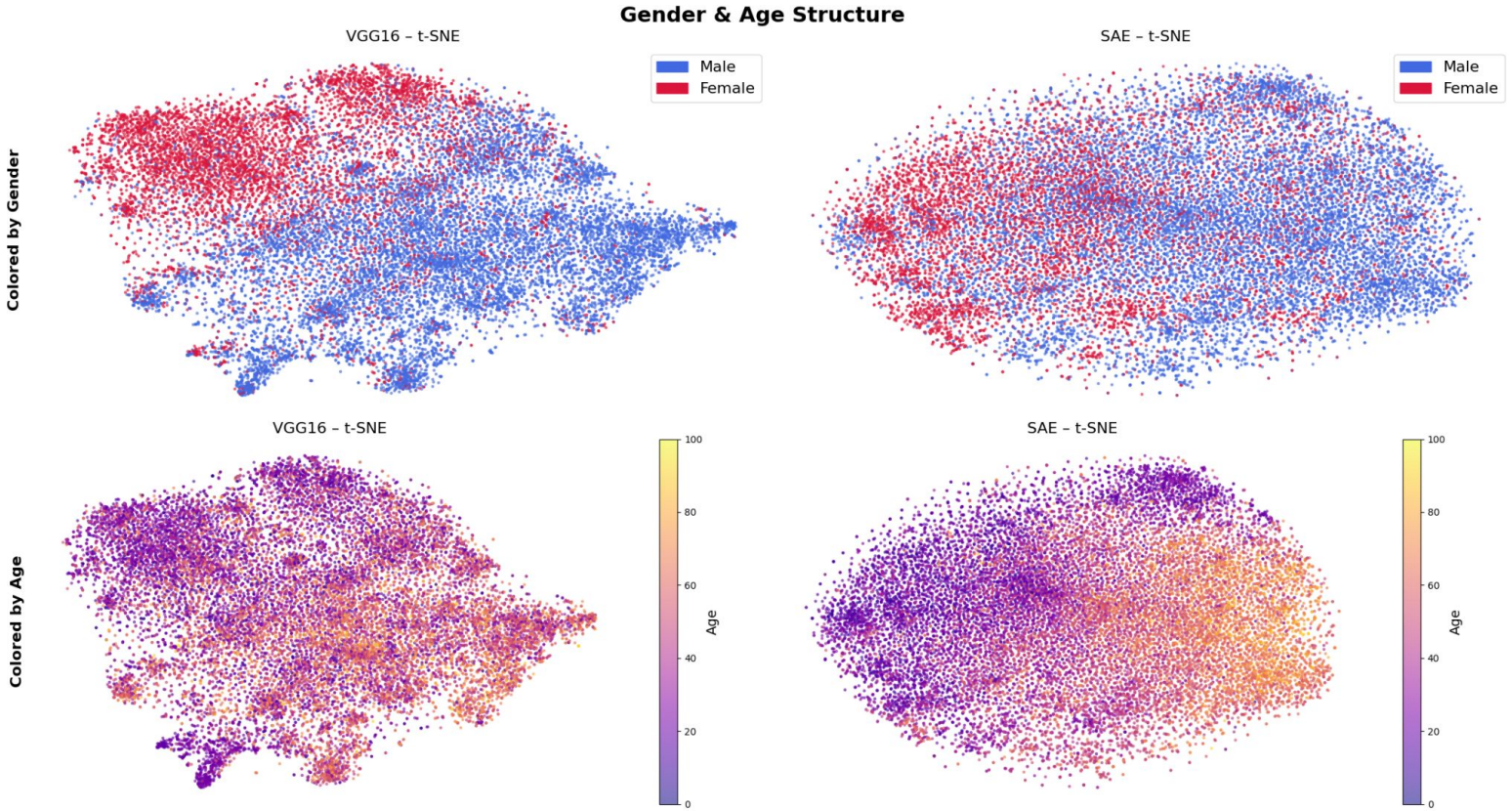


SAE - Visualization by t-SNE

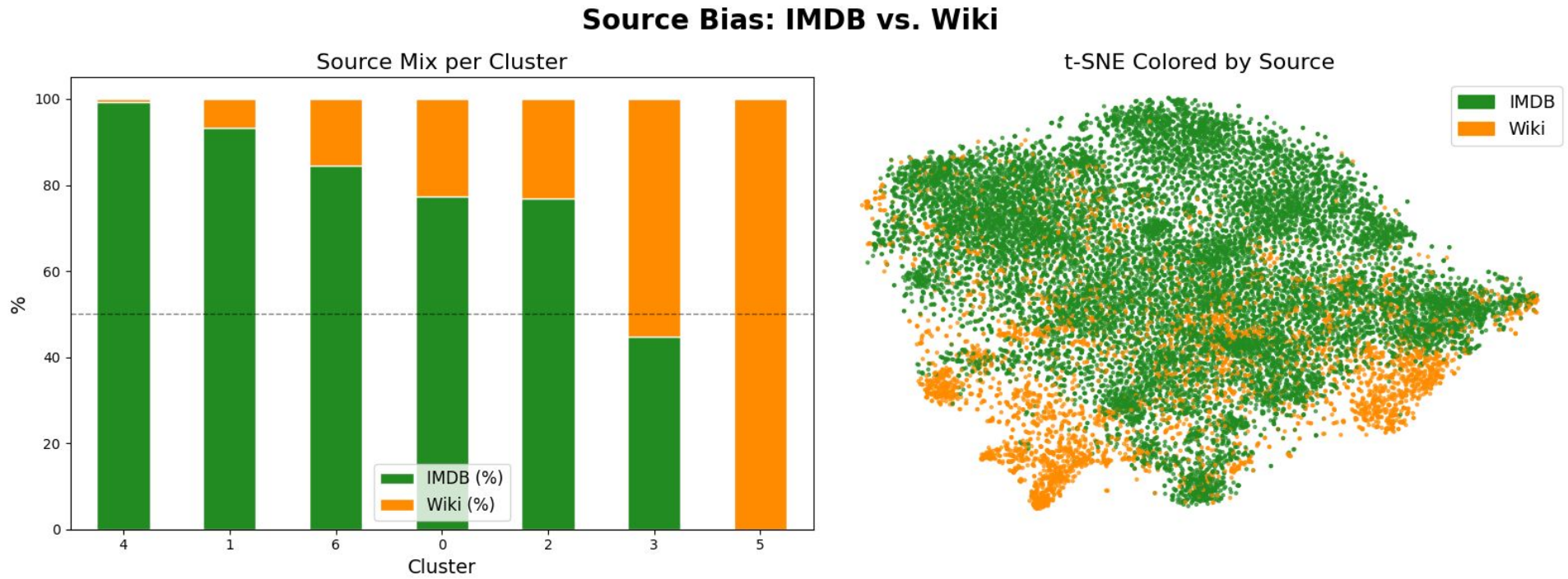


General Questions

Gender bias?


Gender & Age Structure
VGG16 – t-SNE
SAE – t-SNE
Male
Female
Colored by Gender
Colored by Age
Age
0
20
40
60
80
100

Source Bias? (VGG16)



Age Prediction

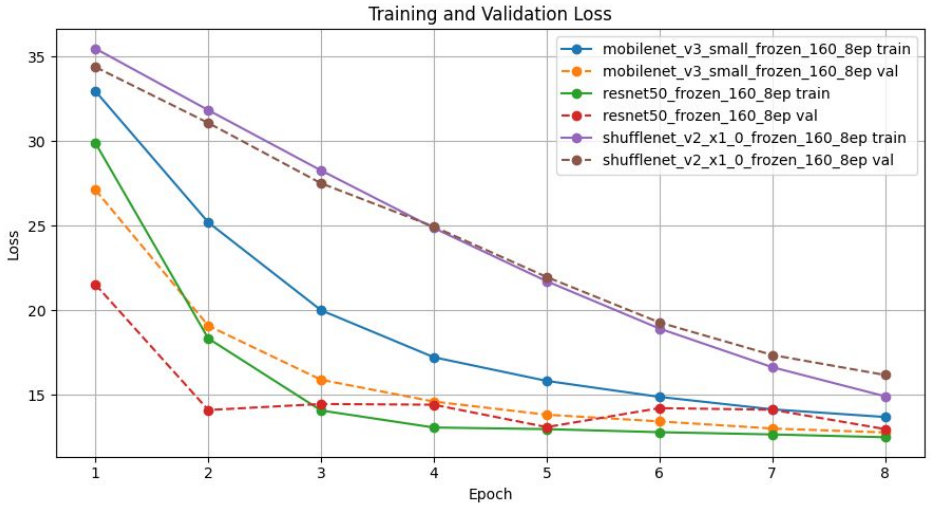
Questions

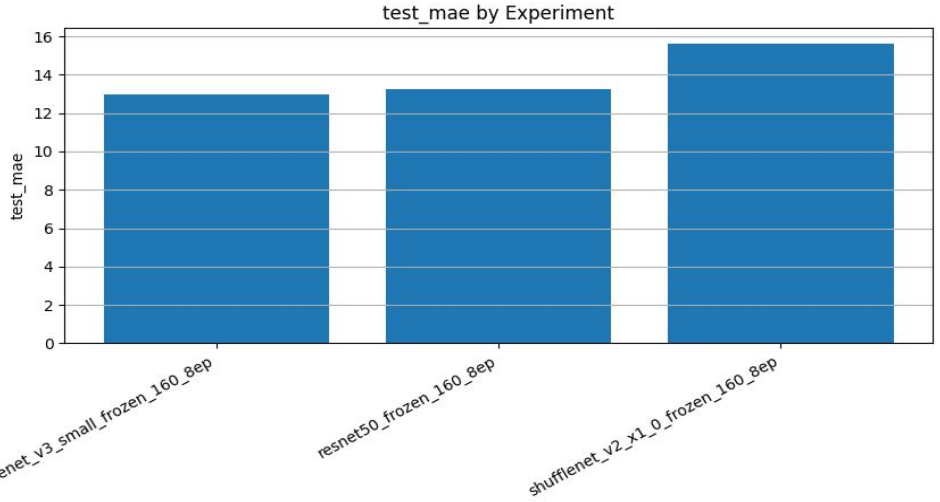
- Which **pretrained CNN** is best suited?
- How does it do with **only one picture** of each person?
- Training on **genders** separately?

Selection of Model

Model Parameters

- **MobileNet V3 Small**
  ~2.5 million
- **ShuffleNet V2 x1.0**
  ~2.3 million
- **ResNet50**
  ~25.6 million
  (4-5 times longer run time)

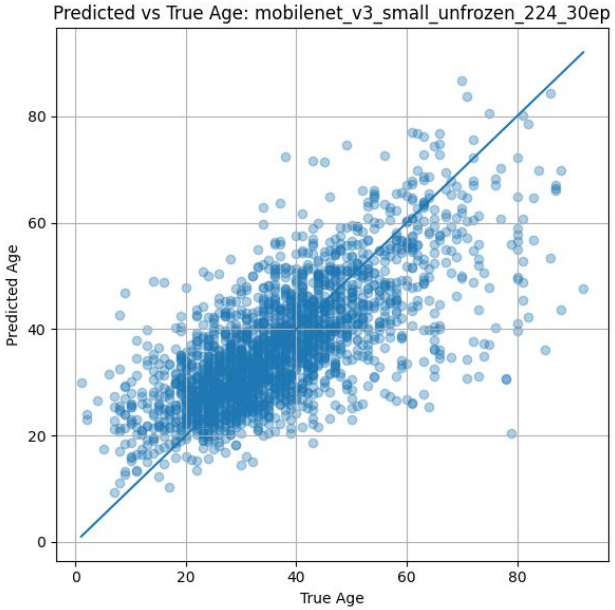




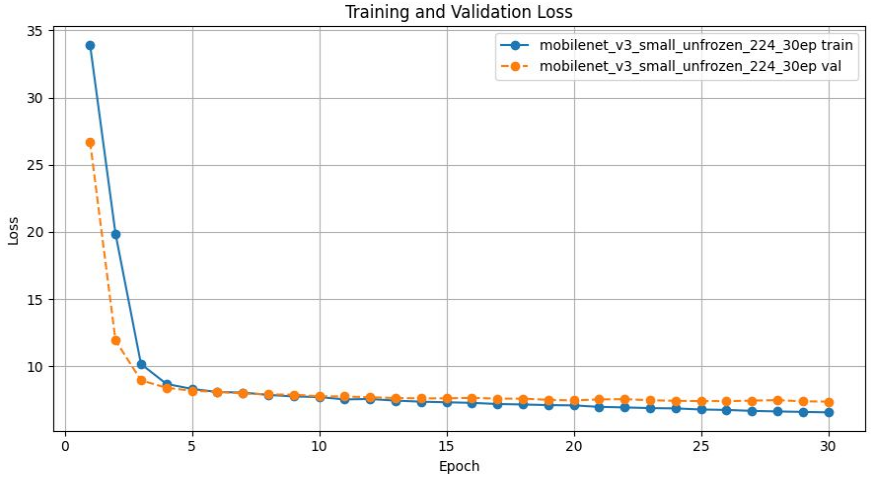
One Picture per Person

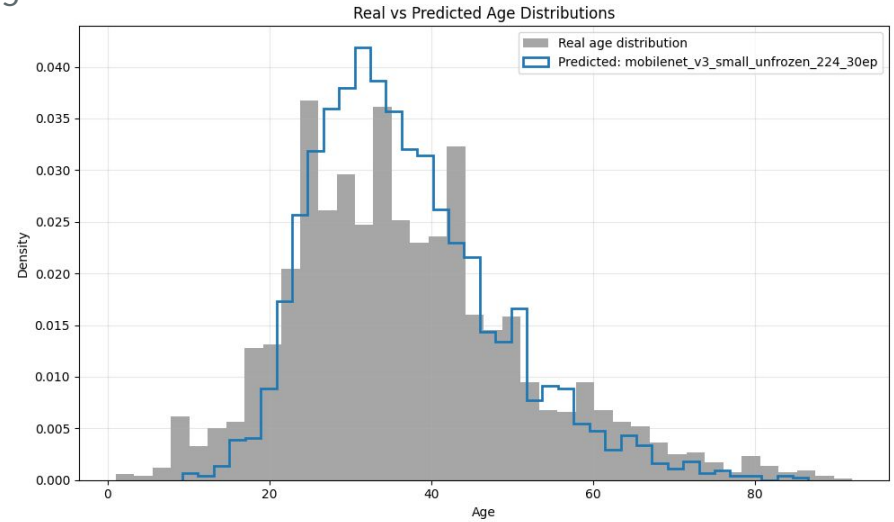
Mobilenet

- Successful learning
- Good train-validation agreement
- Little overfitting

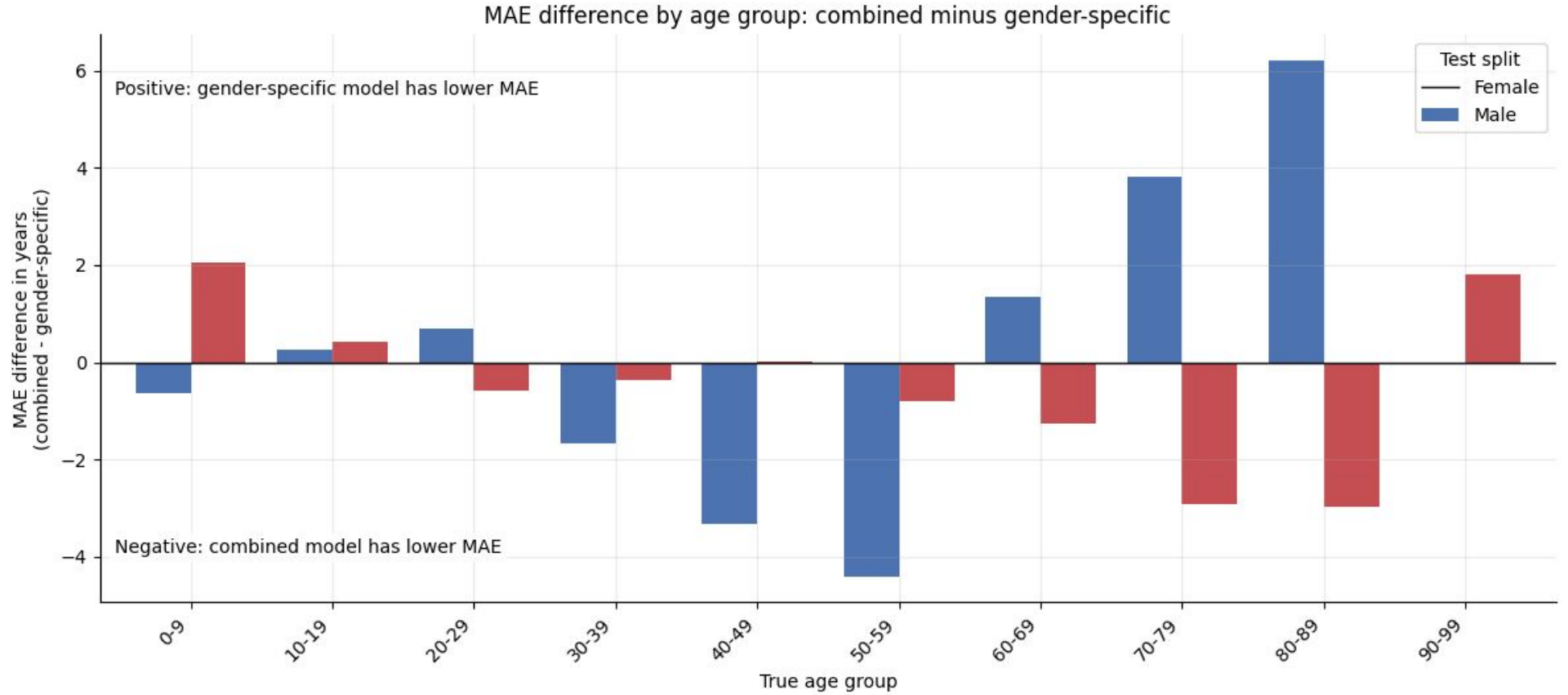


r^2=0.55





Gender-Specific Training



Combined model predictions across facial expressions

1Bandit
Predicted age: 31.0

2Cykelhjelm
Predicted age: 48.7

3Solbriller
Predicted age: 47.7

4Hat
Predicted age: 43.4

5Neutral
Predicted age: 39.8

6Glad
Predicted age: 43.2

7Overrasket
Predicted age: 43.7

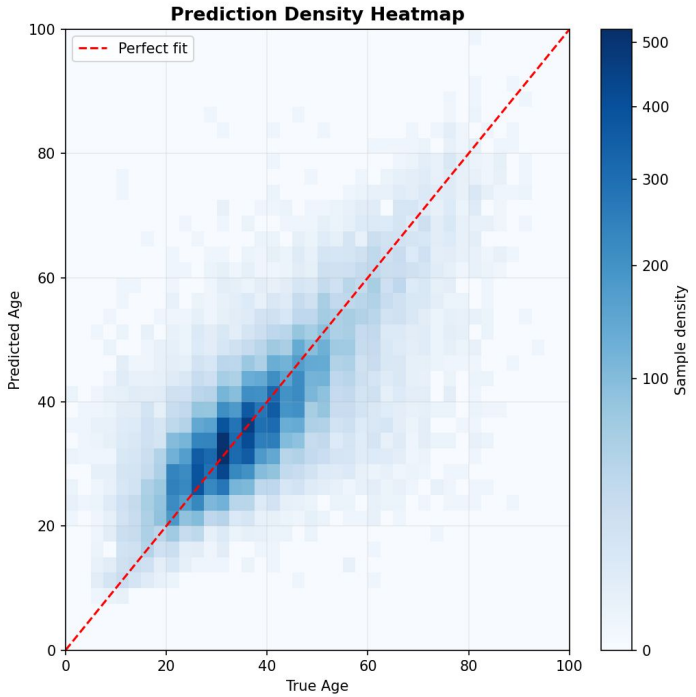
8Sad
Predicted age: 39.7

Summary

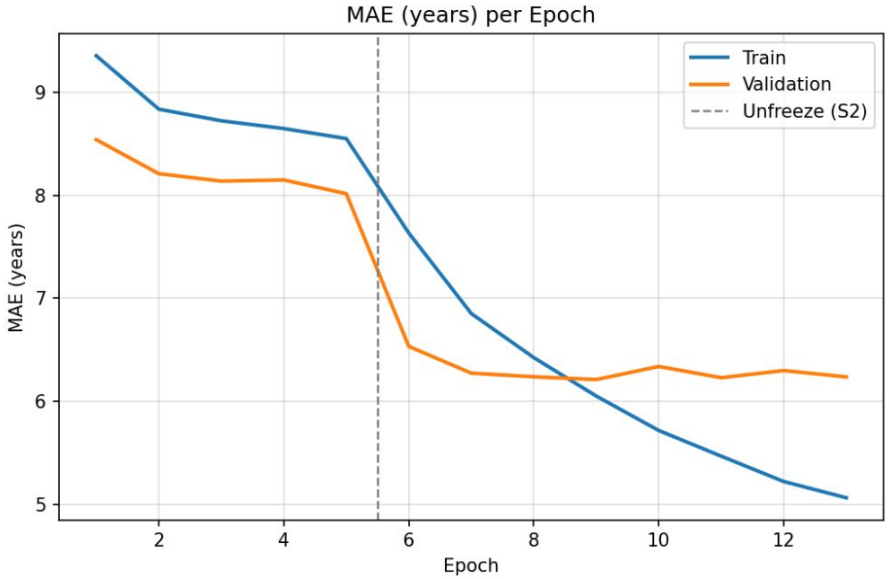
- We decided on Mobilnet and Resnet50
- Age could be predicted with a MAE of 7.3 years
- Training on gender separately didn’t help

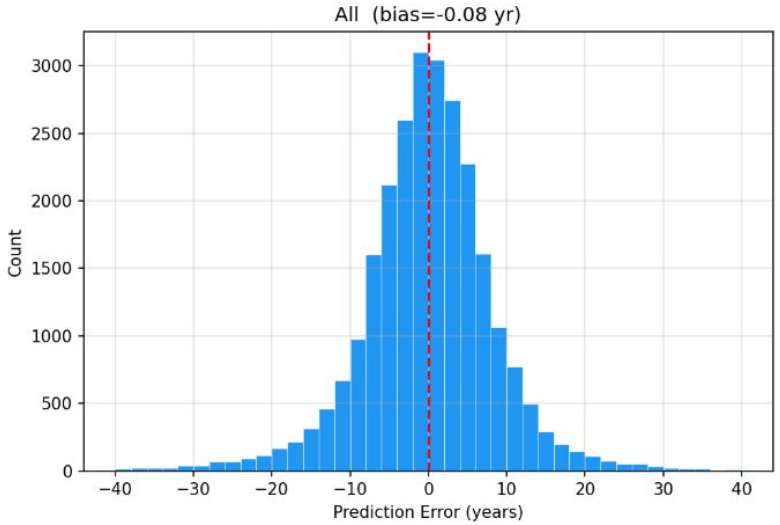
Age Regression Using ResNet50

- Regression head
- **Stage 1:** backbone fully frozen, head-only training (~ 1M Parameters)
- **Stage 2:** unfreeze Layer 3 + Layer 4 at once (~ 23M params)



Age Regression Using ResNet50





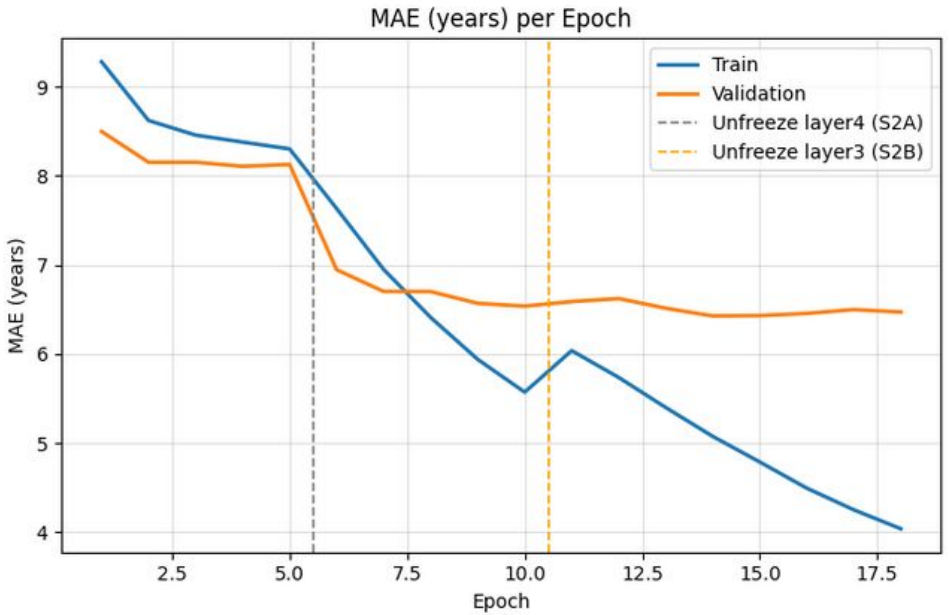
Attempts at Improving Previous Model

- **Stage 1:** same
- **Stage 2A:** unfreeze Layer 4 + head only (~6M params)
- **Stage 2B:** unfreeze Layer 3 + Layer 4 + head (~23M params), same differential LRs

⇒ ResNet50 needs minimal training to adapt

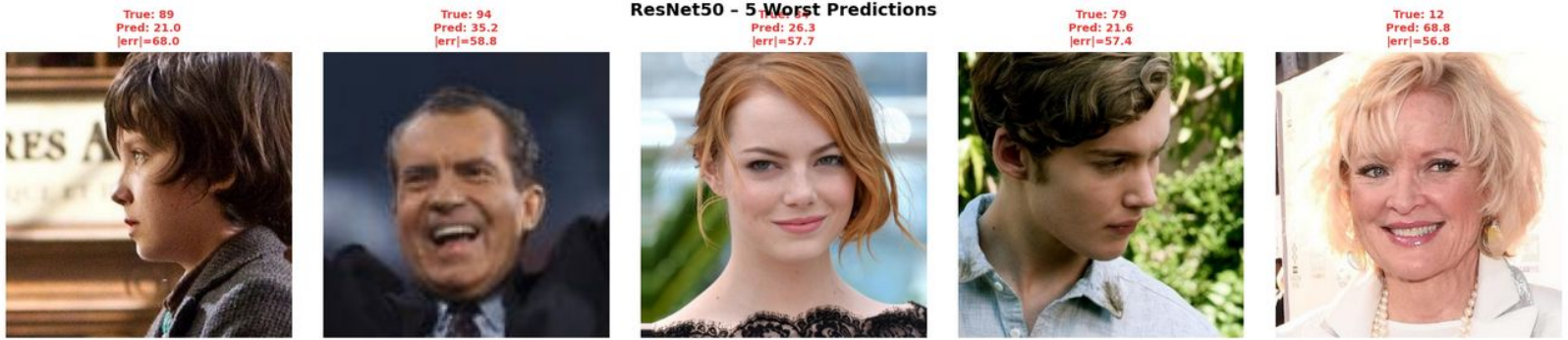
⇒ Overfitting bad labels

⇒ Limit seems to be around MAE ~ 6 yrs



Best and Worst Age Predictions

Saved: ResNet50_best_predictions.png



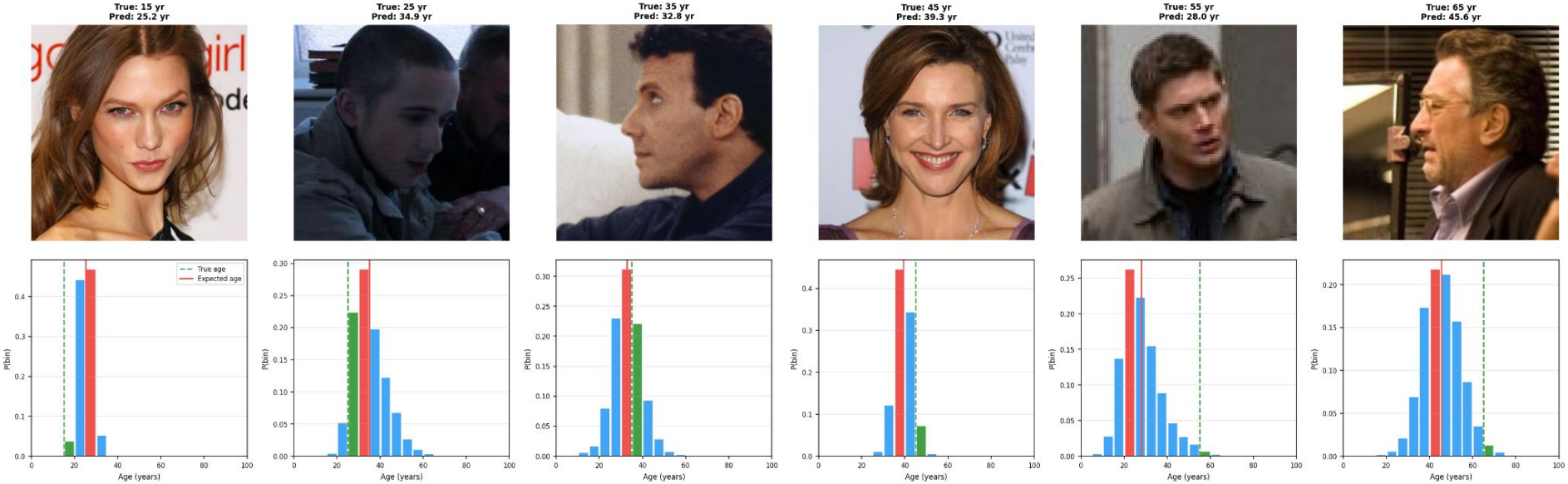
Saved: ResNet50_worst_predictions.png

Classification

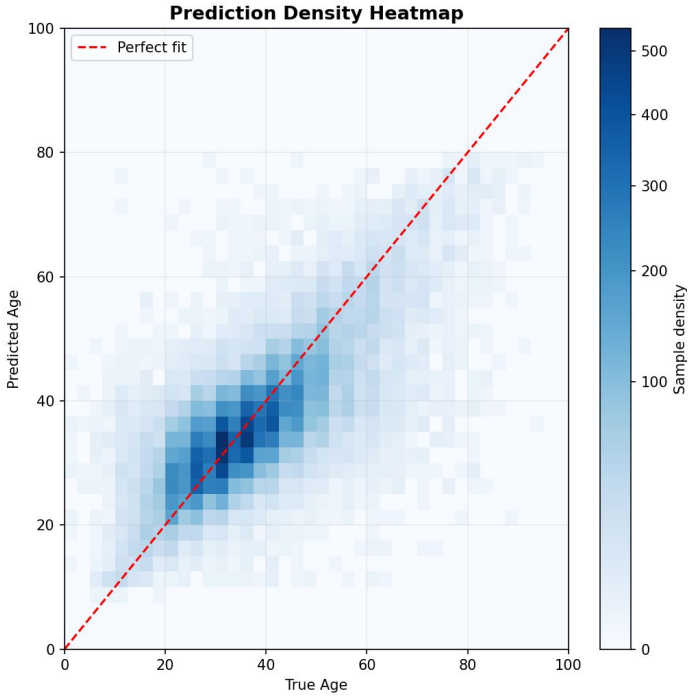
Age Prediction

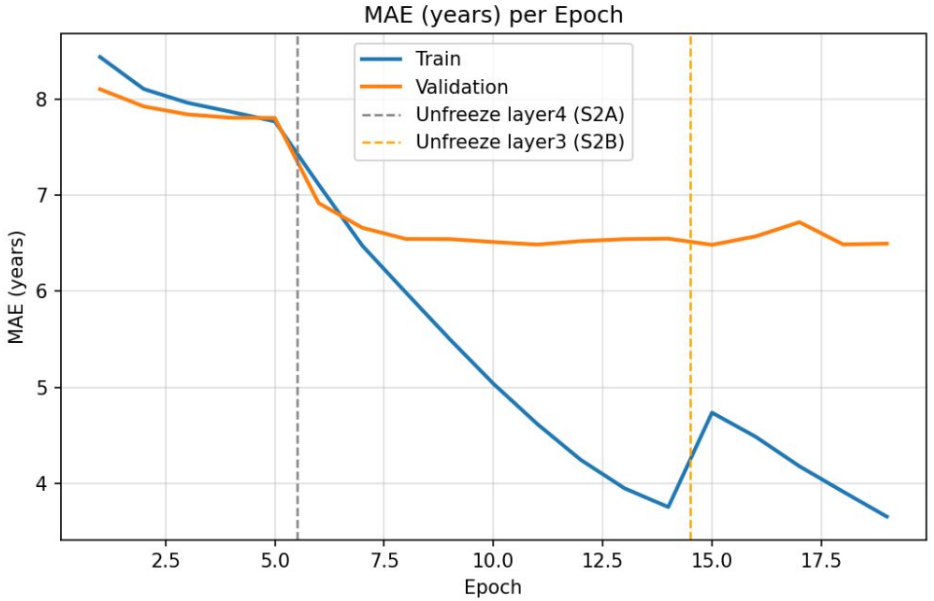
Age Bin Classification Examples

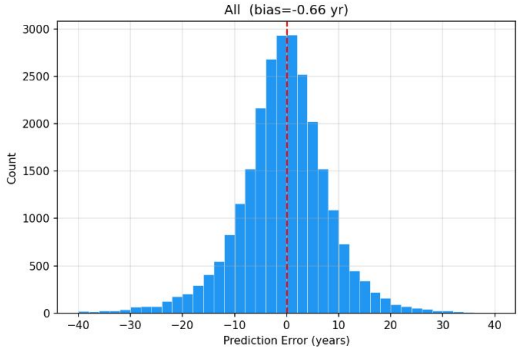
- Same backbone and staged fine-tuning
- Head outputs 20 logits (age-bins)
- Predicted age decoded as expected value



Age-Bin Classification







Gender Prediction

Gender Classification

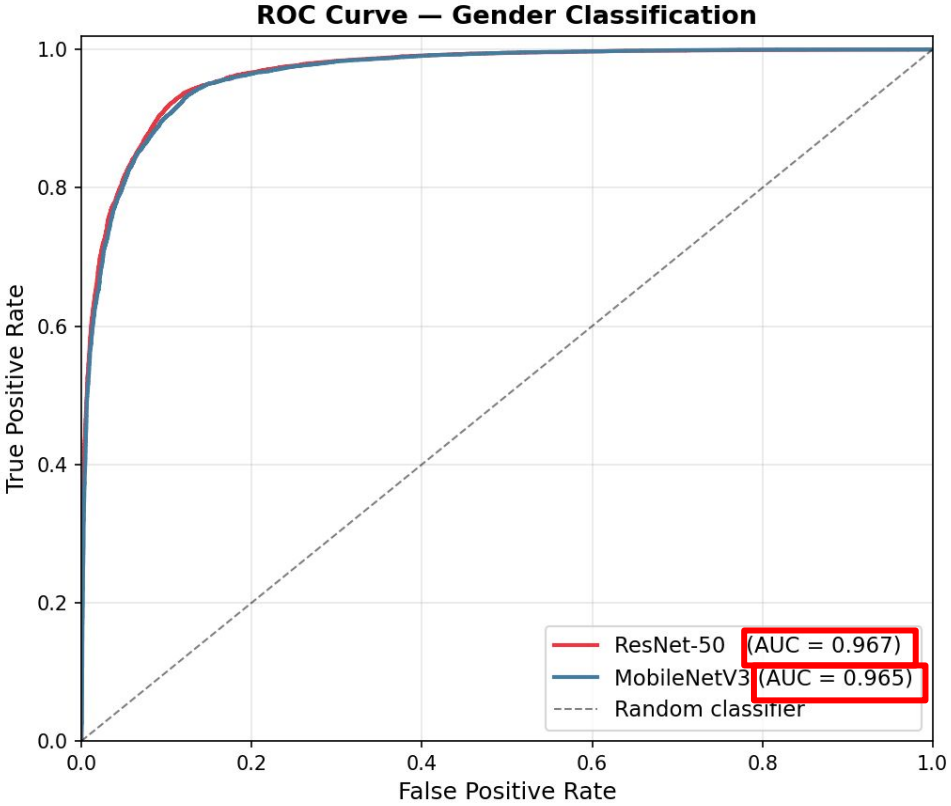
3 Models:

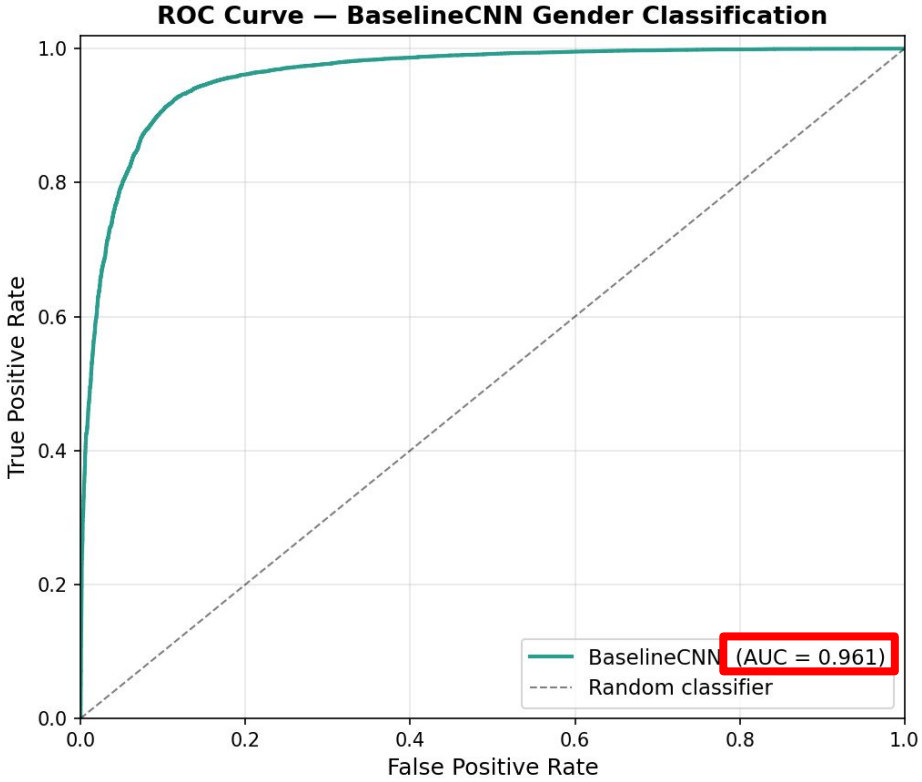
- ResNet50
- MobileNet
- non pretrained CNN

Goal:

- Predict gender based on labeled data
- Study limitations of the models and possible improvements

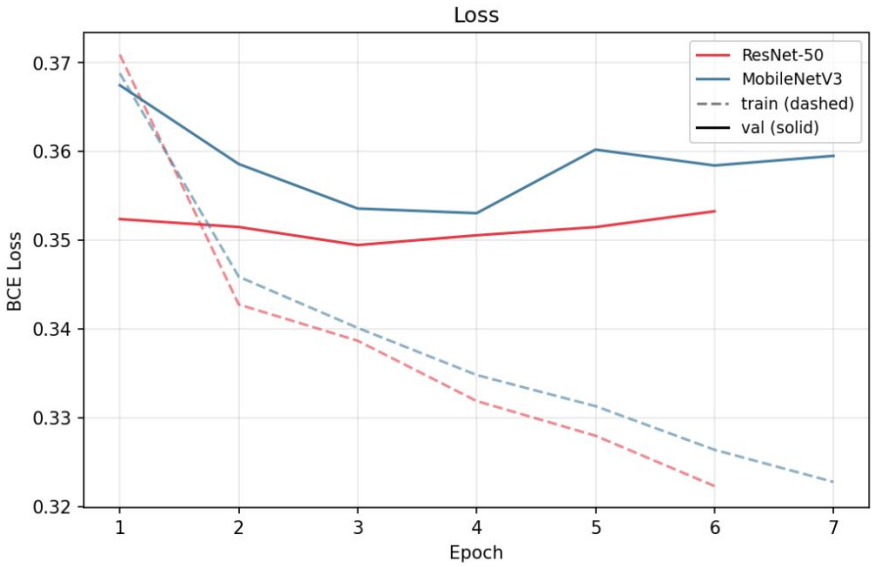
Gender Classification - Model Evaluation



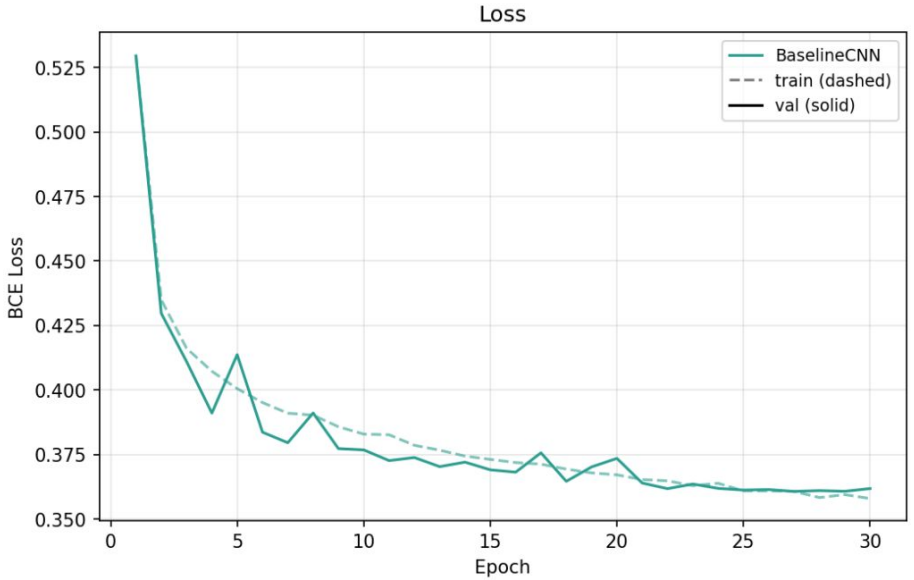


Gender Classification - Model Evaluation

ResNet-50 and MobileNetV3



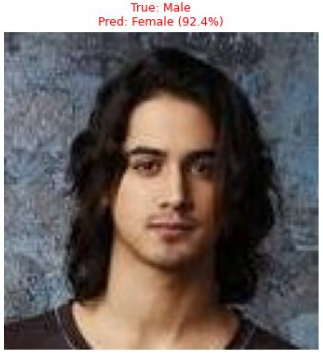
Baseline CNN

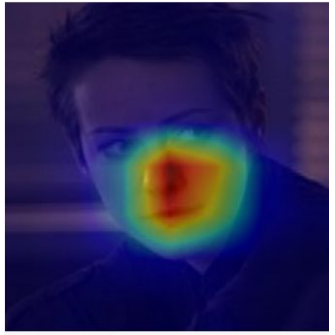


Gender Classification - Model Evaluation

correctly classified:

correctly classified:





Grad-CAM = Gradient-weighted Class Activation Mapping

Discussion

Summary

Clustering:

- Investigated patterns

Age Prediction:

- Best prediction with MAE 6 years

Gender Prediction:

- Best prediction with AUC 0.967

Discussion

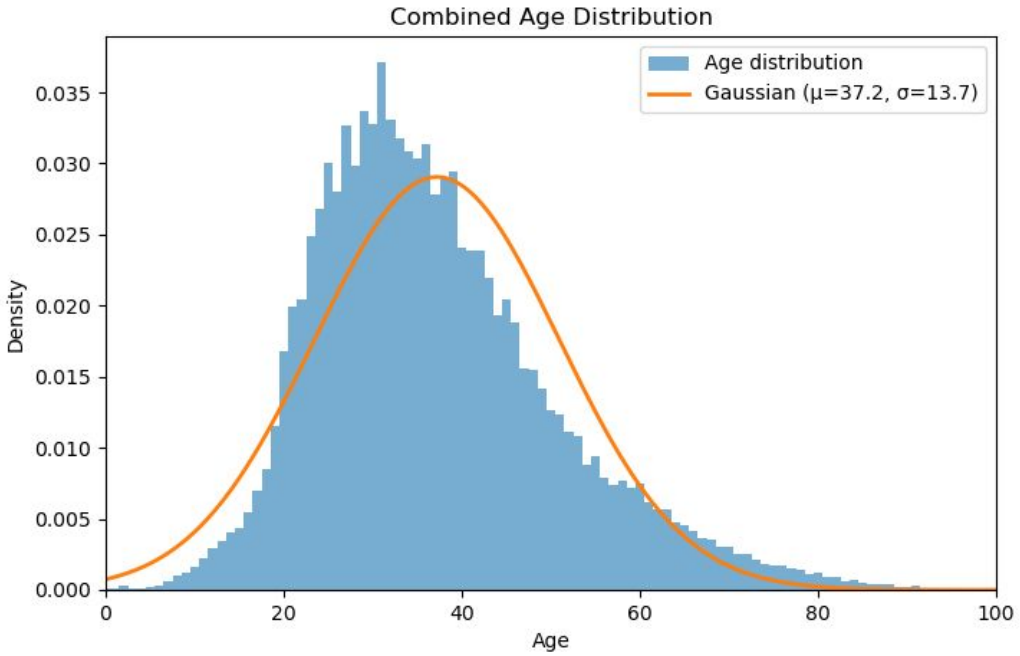
Our Takeaways:

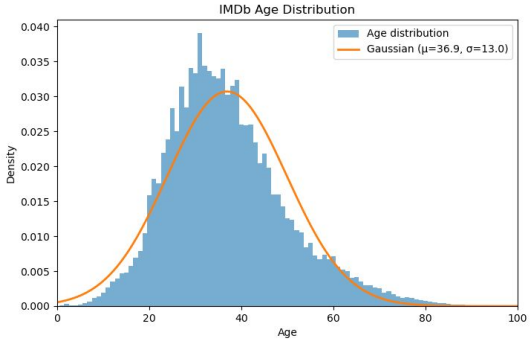
- Load your images effectively!
- Use a GPU! Shoutout to Kaggle Notebooks!
- Watch for noisy labels and data bias!
- Regression and Classification Models:
    - ResNet50 > MobileNet V3 > ShuffleNet V2 > non pretrained CNN

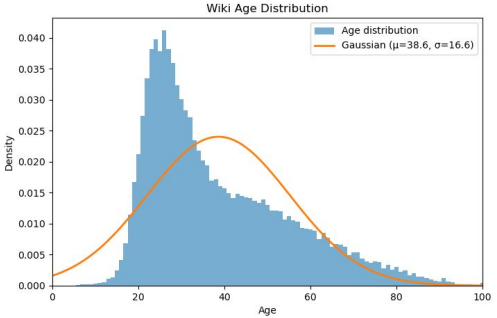
Appendix

Appendix - Data

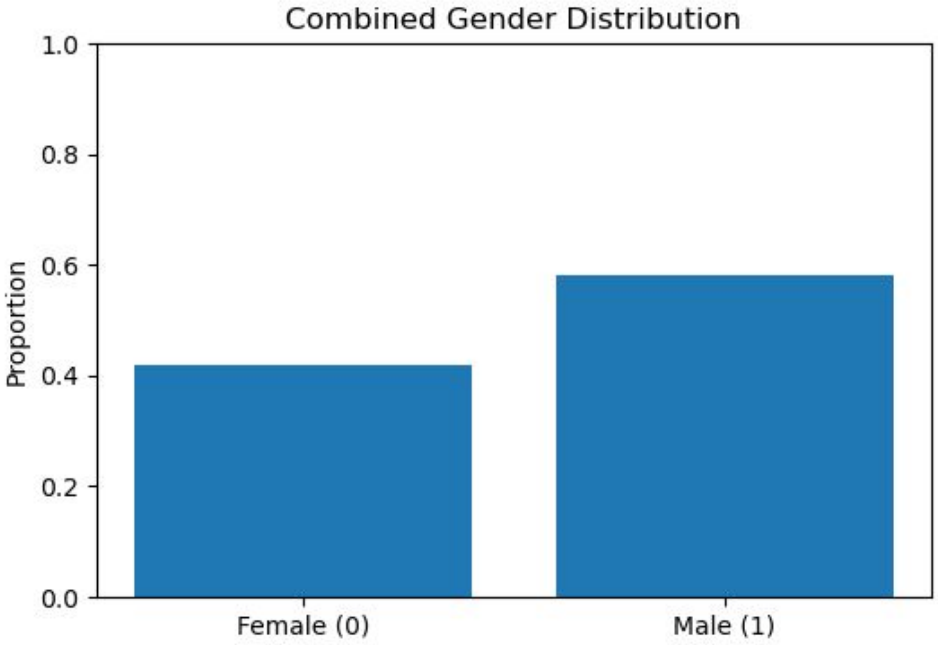
The IMDB-WIKI Dataset - Age Distribution

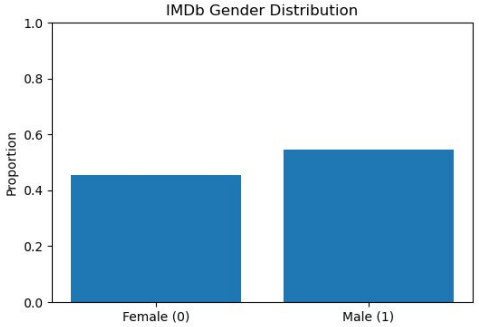


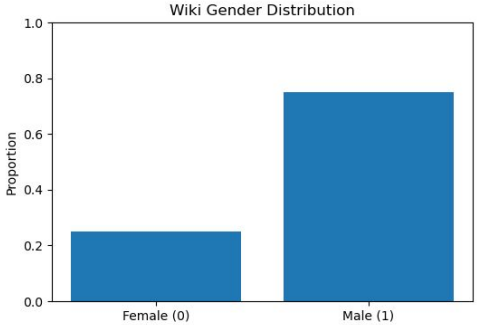




The IMDB-WIKI Dataset - Gender Distribution

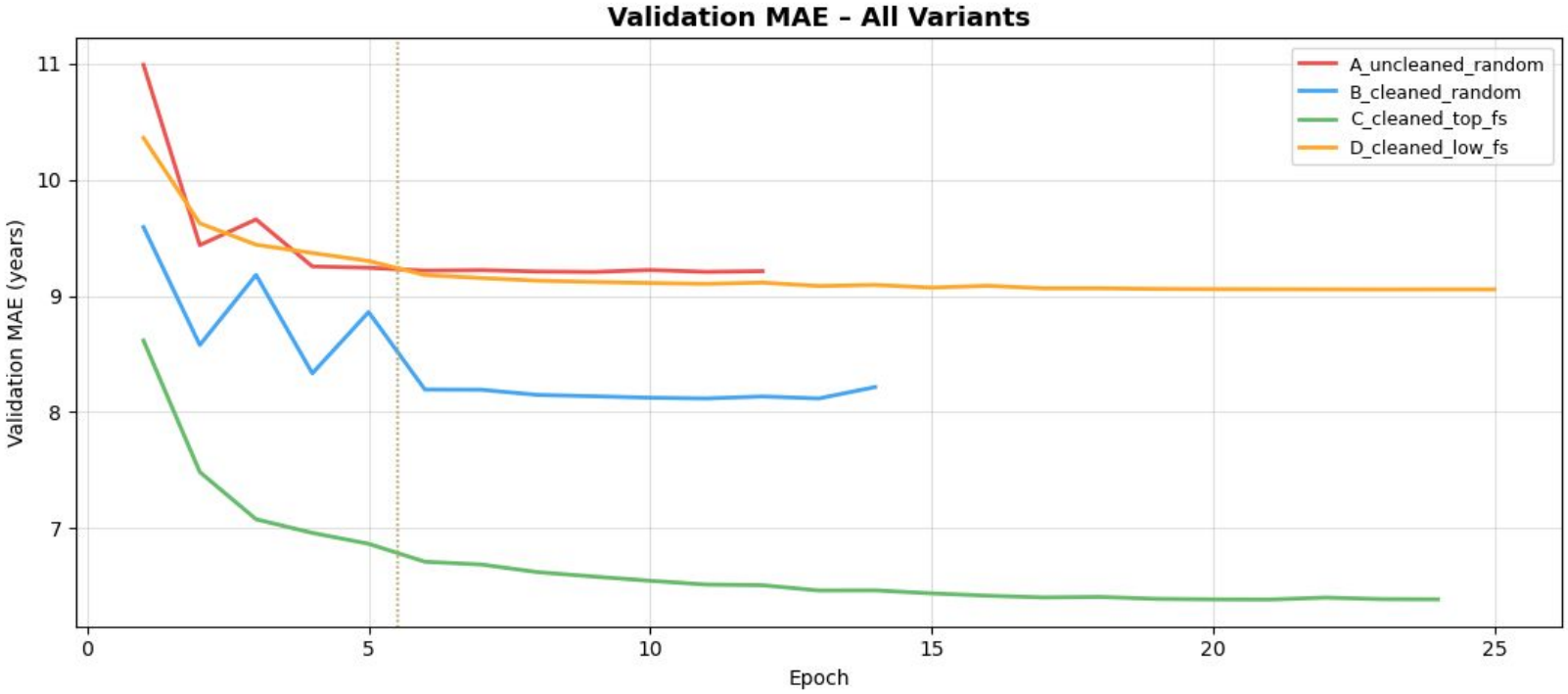






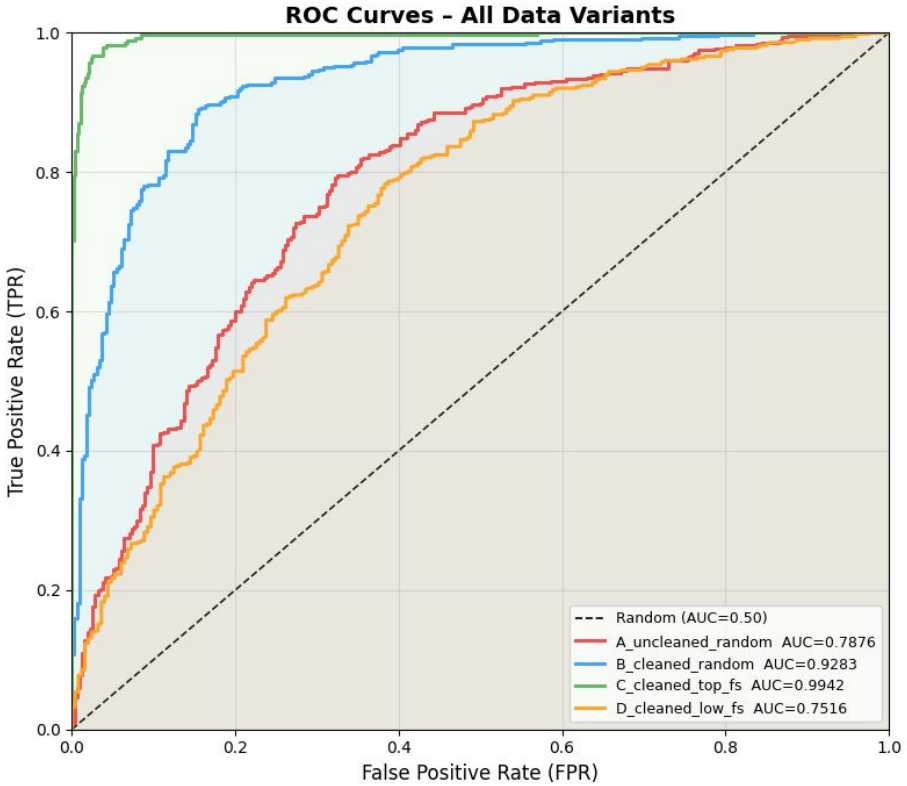
Experiment on Data Quality

light mobilenet model



Experiment on Data Quality

light mobilenet model



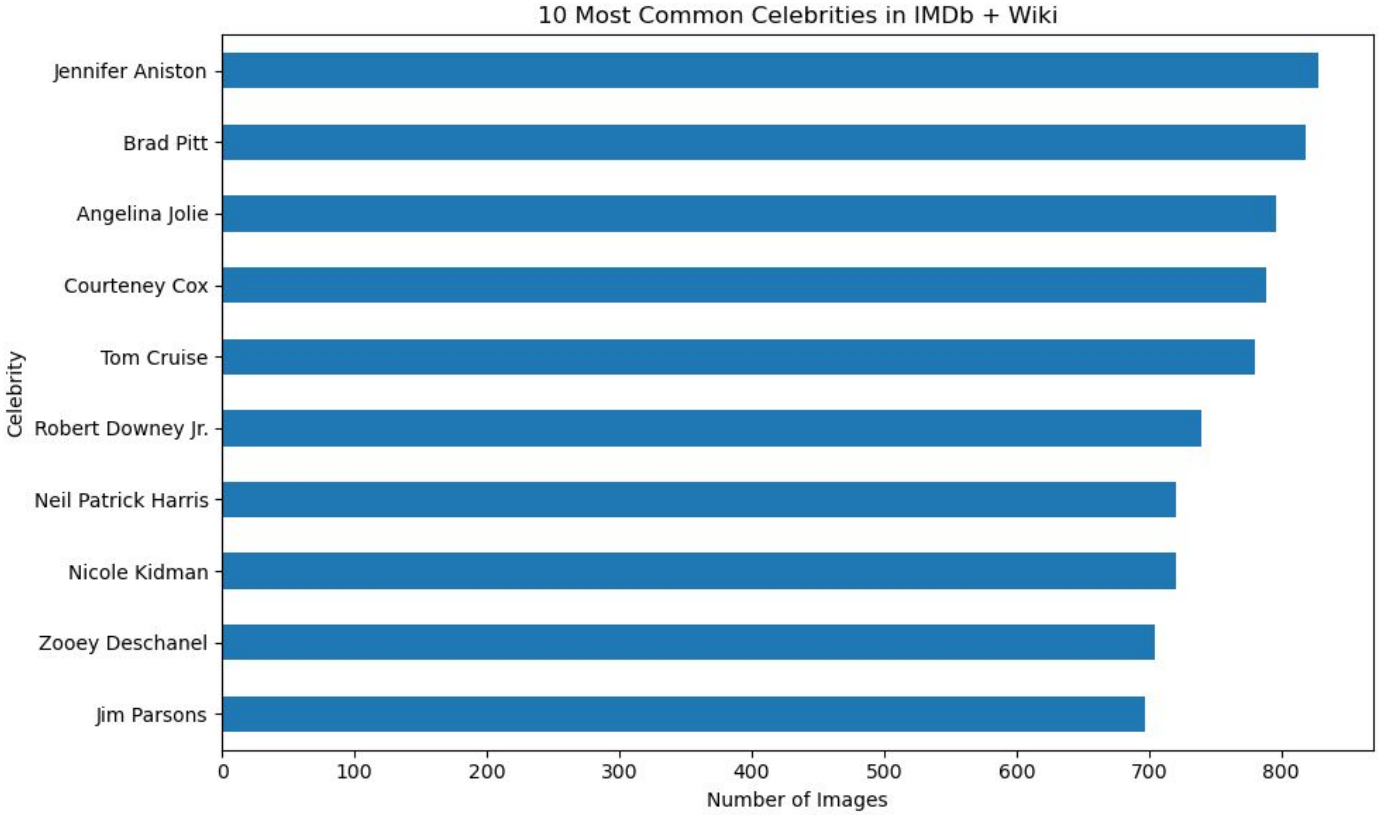
The IMDB-WIKI Dataset - Overview

- Large dataset of celebrities
- Crawled the IMDb website for all images, date of birth & gender
- Crawled all profile images from pages of people from Wikipedia with the same meta information
- Removed all images without a timestamp
- Images can be from movies with extended production time, wrong time stamps
- IMDb 460,723 face images from 20,284 celebrities
- Wikipedia 62,328

The IMDB-WIKI Dataset - Meta Information

- **dob:** date of birth (Matlab serial date number)
- **photo_taken:** year when the photo was taken
- **full_path:** path to file
- **gender:** 0 for female and 1 for male, NaN if unknown
- **name:** name of the celebrity
- **face_location:** location of the face. To crop the face in Matlab run
- **face_score:** detector score (the higher the better). *Inf* implies that no face was found in the image and the *face_location* then just returns the entire image
- **second_face_score:** detector score of the face with the second highest score. This is useful to ignore images with more than one face. *second_face_score* is *NaN* if no second face was detected
- **celeb_names:** list of all celebrity names
- **celeb_id:** index of celebrity name

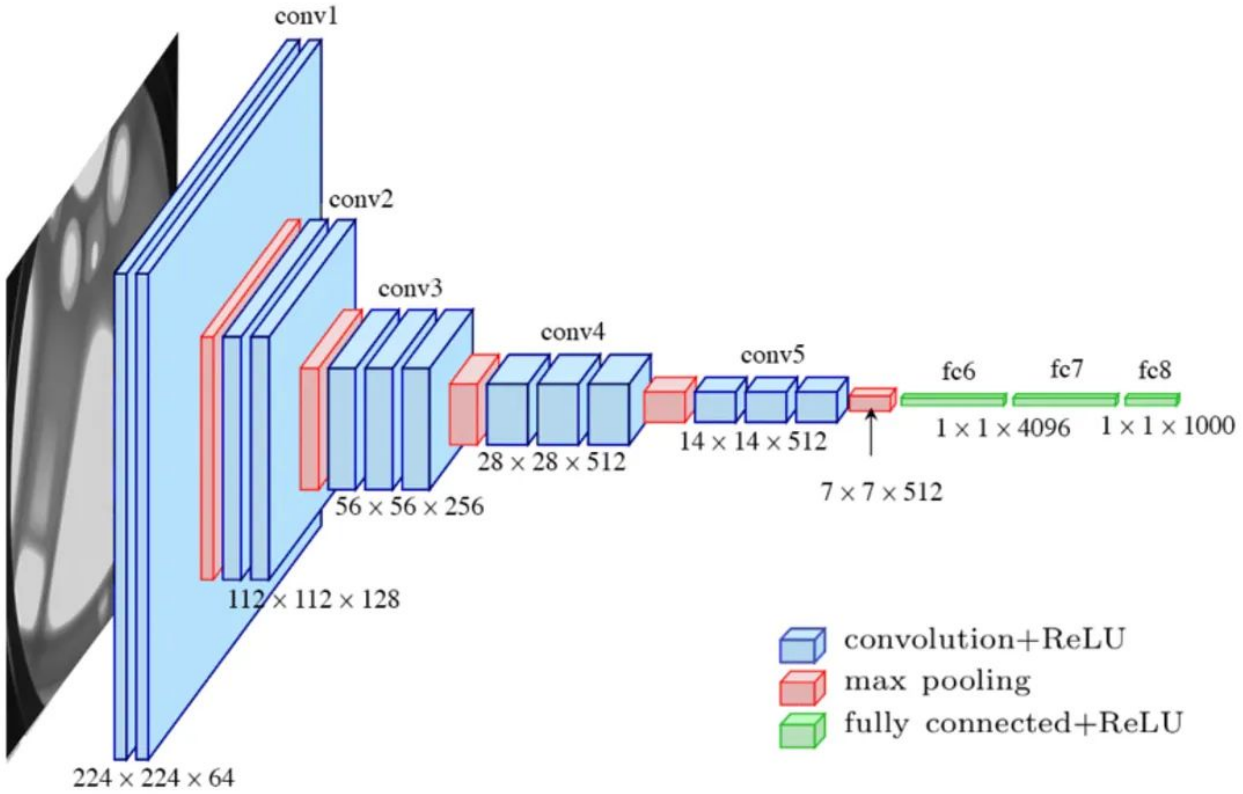
The IMDB-WIKI Dataset - Multiplicity



Appendix - Clustering

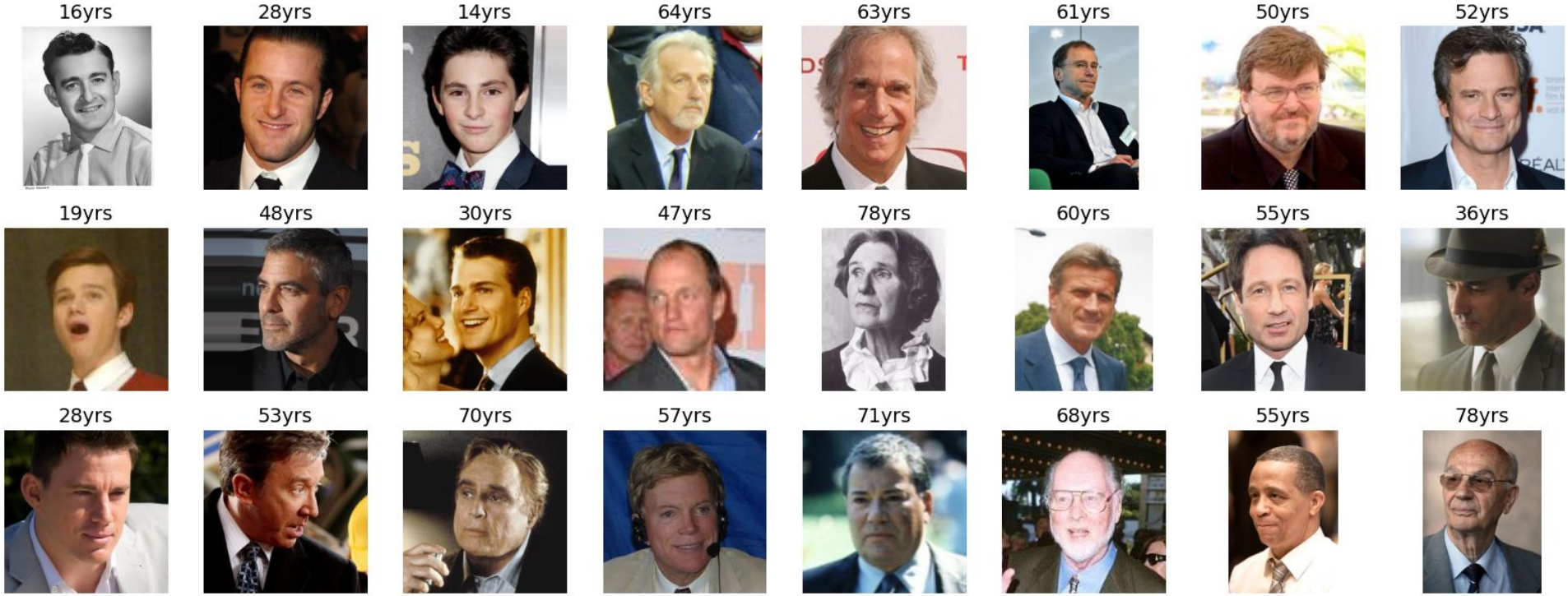
VGG 16

VGG16 – Architecture



Looking at the Clusters (Clear: People in Suits)

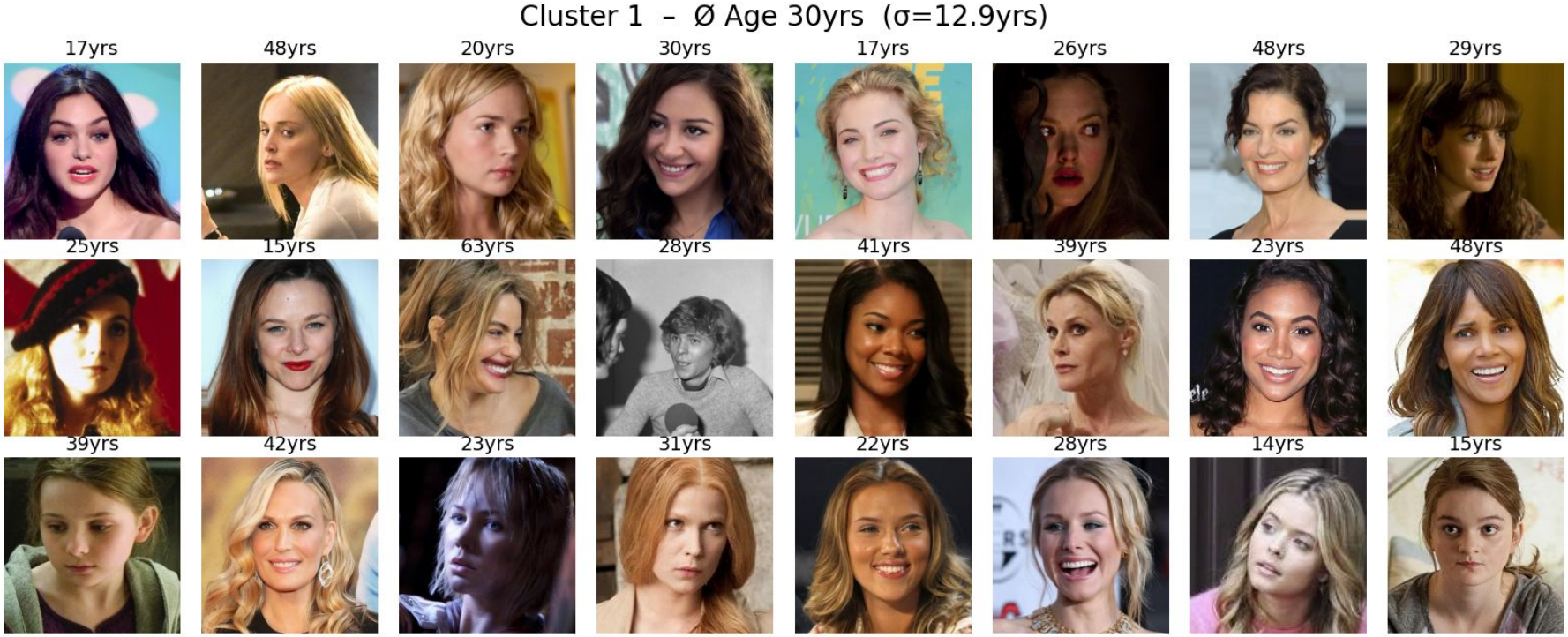
Cluster 2 – Ø Age 50yrs (σ=19.1yrs)



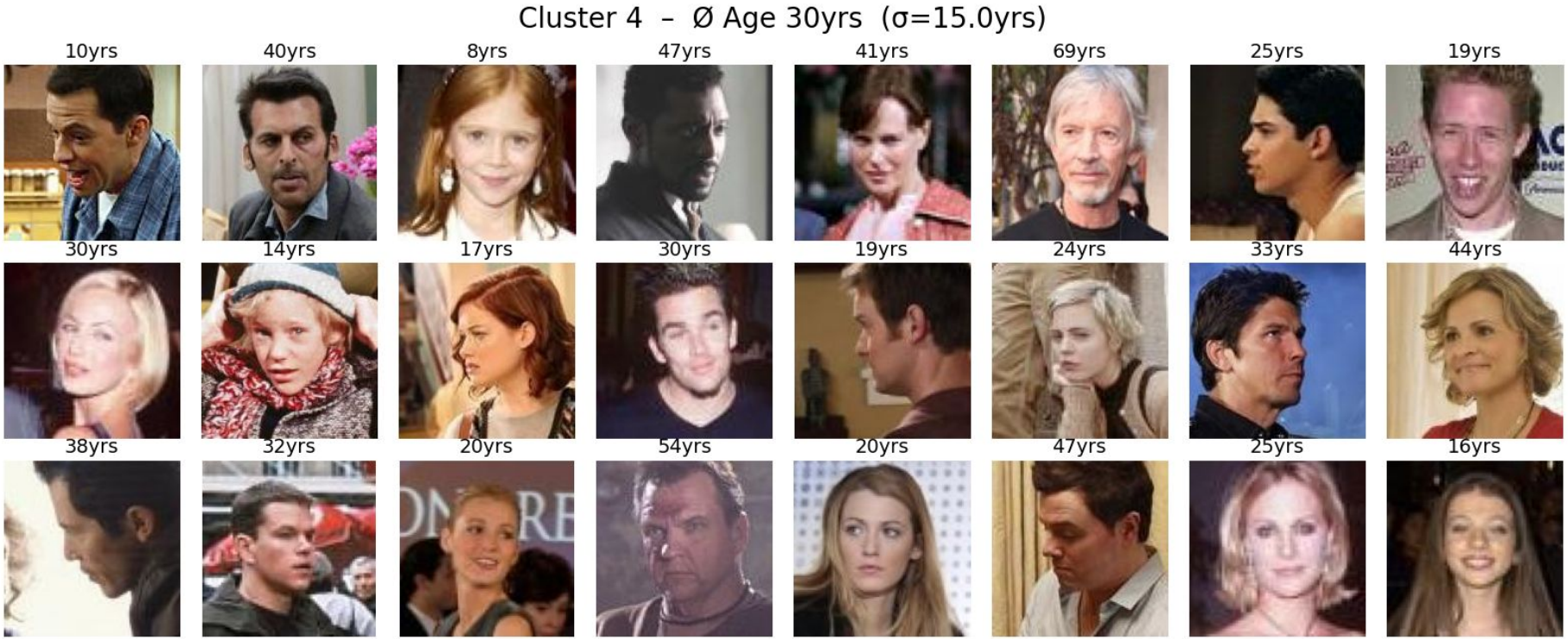
Looking at the Clusters (Clear: Sports People)

Cluster 5 – Ø Age 27yrs (σ=10.1yrs)

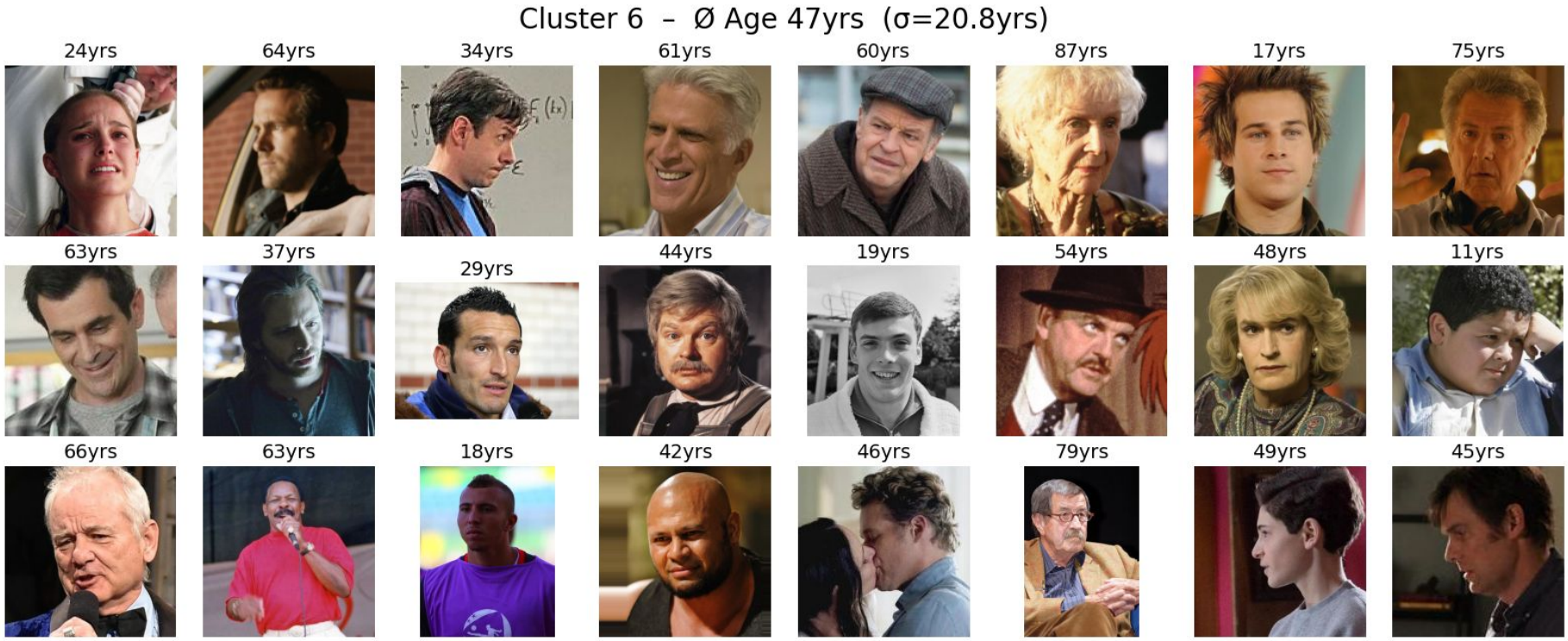
Looking at the (not so clear) Clusters



Looking at the (not so clear) Clusters



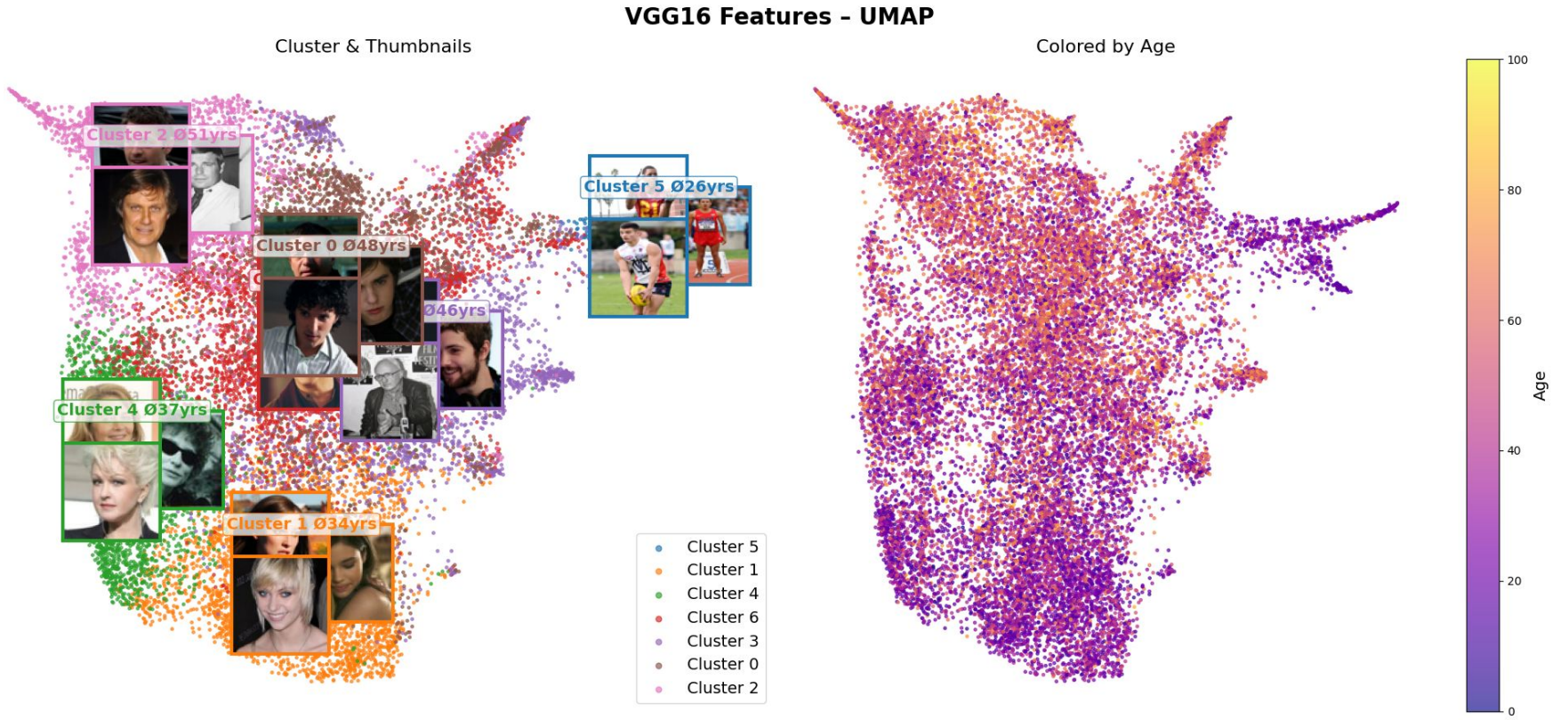
Looking at the (not so clear) Clusters



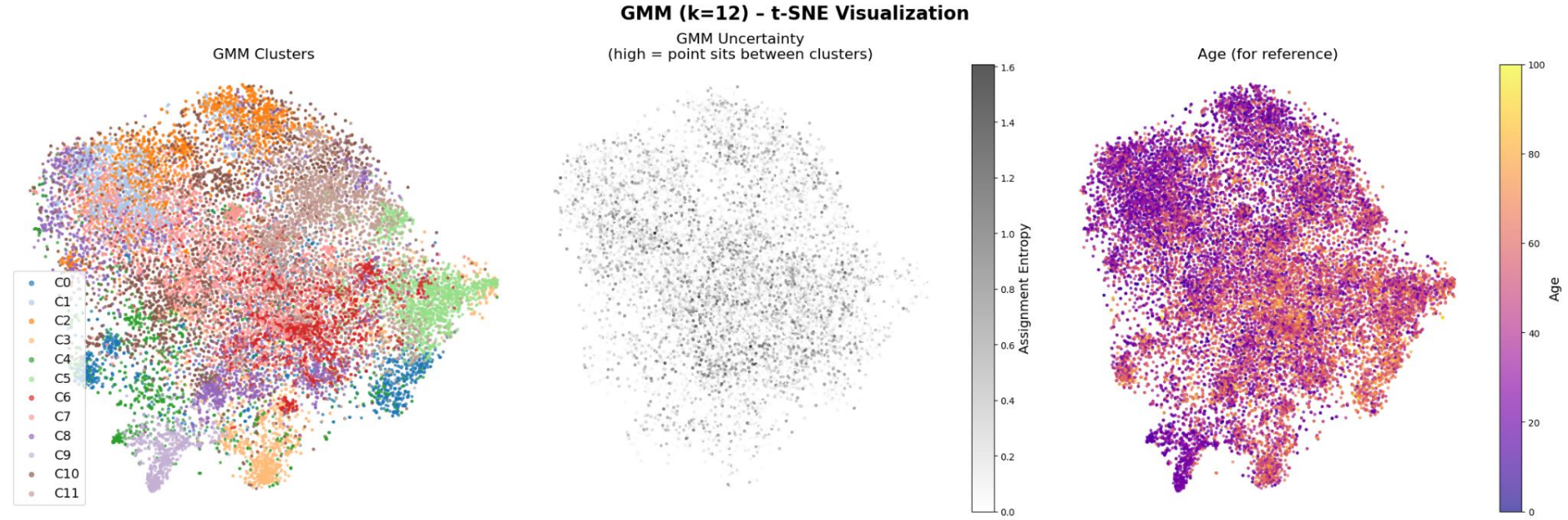
Looking at the (not so clear) Clusters

Cluster 3 – Ø Age 43yrs (σ=20.5yrs)

VGG16 (K-Means) – Visualization by UMAP

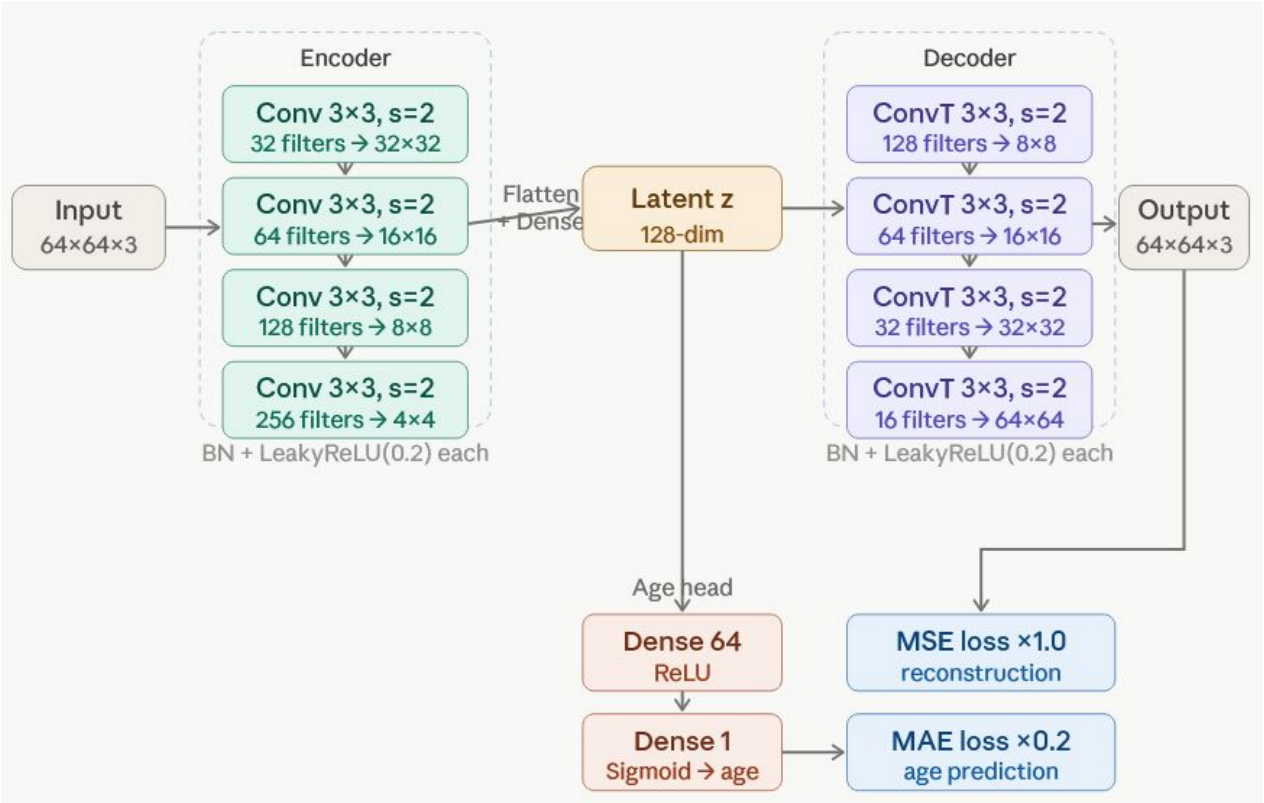


GMM Clustering on VGG16



Age-Supervised Autoencoder

Age-Supervised AE (SAE) - Architecture



Age-Supervised AE (SAE) - Model Description

```
Modell ready  |  AGE_LOSS_WEIGHT=0.2
Model: "age_supervised_ae"
```

Layer (type)	Output Shape	Param #	Connected to
input_layer_2 (InputLayer)	(None, 64, 64, 3)	0	-
encoder (Functional)	(None, 128)	914,752	input_layer_2[0][0]
dense_1 (Dense)	(None, 64)	8,256	encoder[0][0]
age_pred (Dense)	(None, 1)	65	dense_1[0][0]
decoder (Functional)	(None, 64, 64, 3)	921,699	encoder[0][0]

Total params: 1,844,772 (7.04 MB)

Trainable params: 1,843,332 (7.03 MB)

Non-trainable params: 1,440 (5.62 KB)

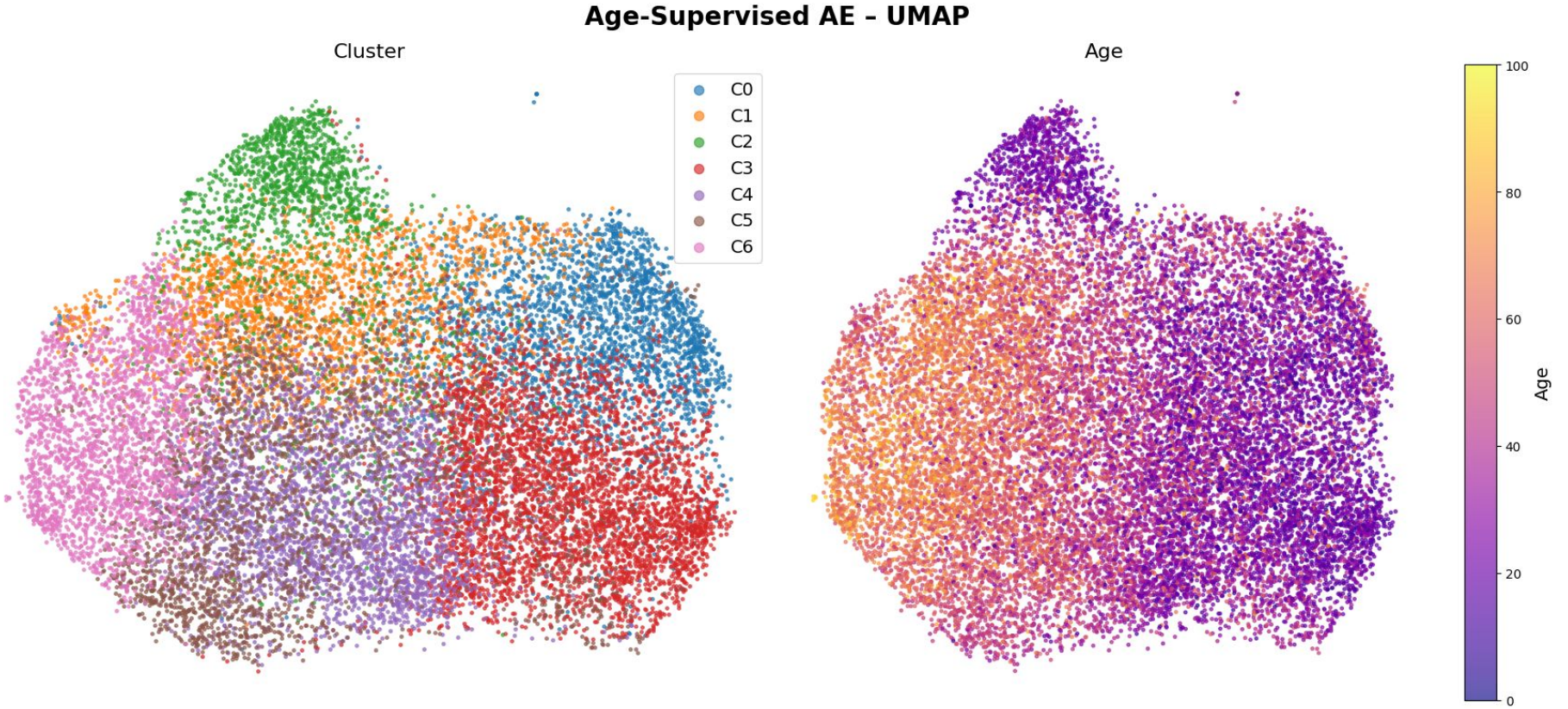
Looking at the Clusters (SAE works well)

Cluster 6 – Ø Age 66yrs (σ=10.9yrs)

Looking at the Clusters (SAE works not so well)

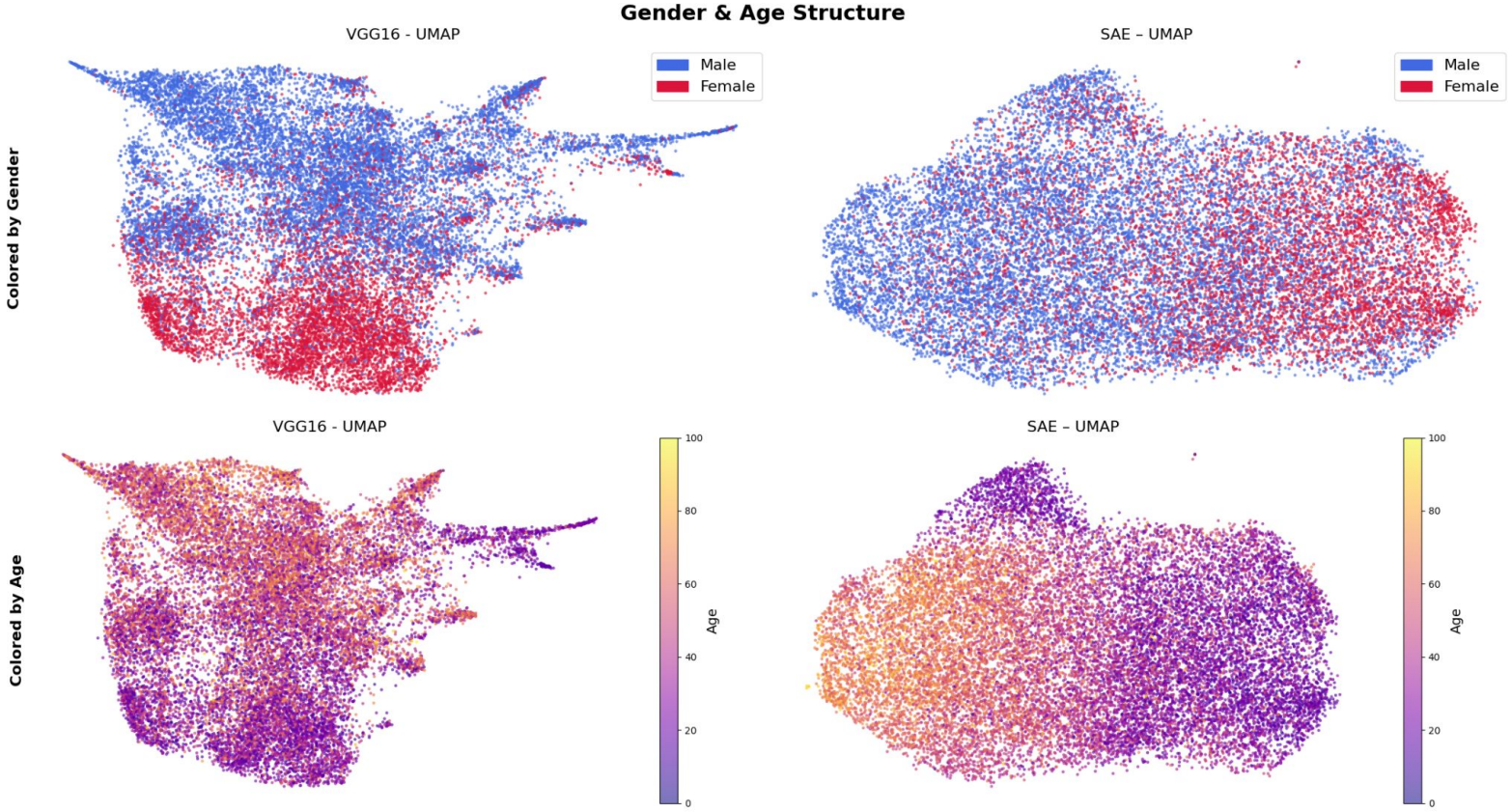
Cluster 2 – Ø Age 32yrs (σ=16.8yrs)

SAE – Visualization by UMAP

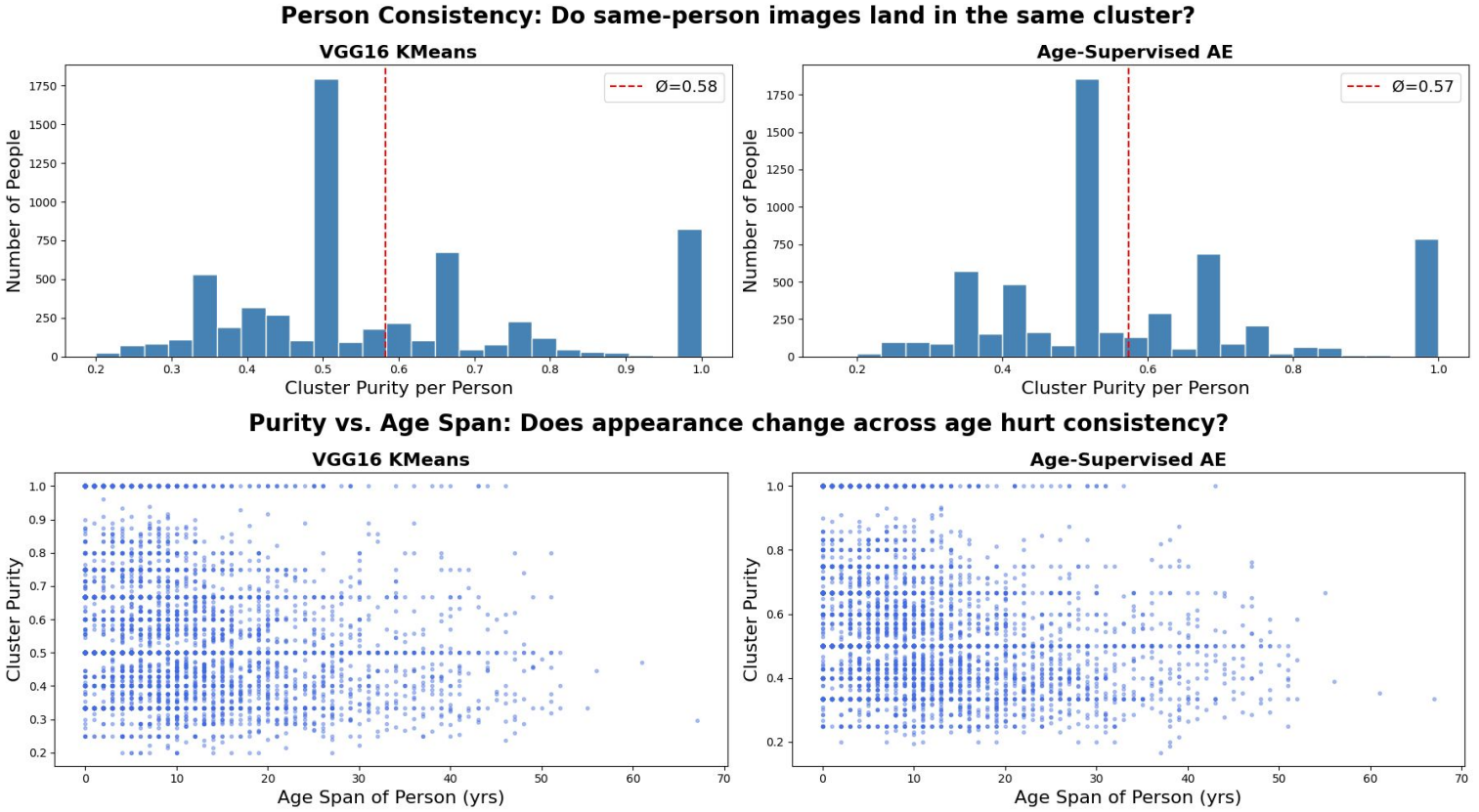


General Questions

Gender bias? – UMAP Visualization



Images of the Same Person?

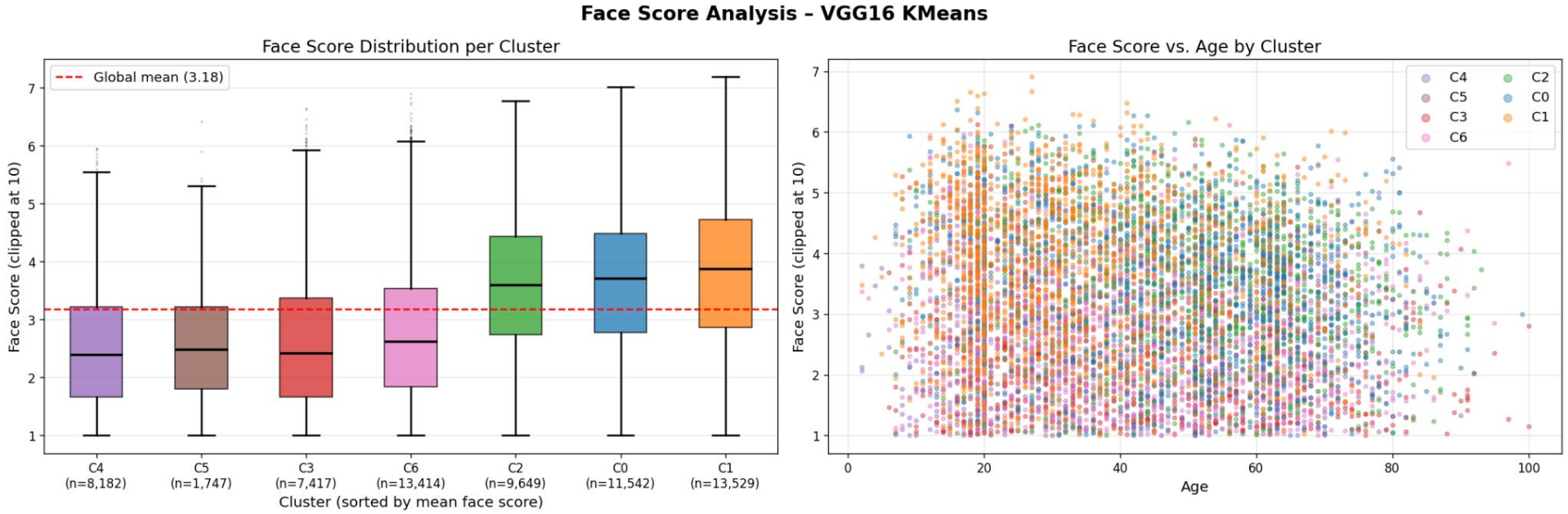


Explanation to “Images of the Same Person?”

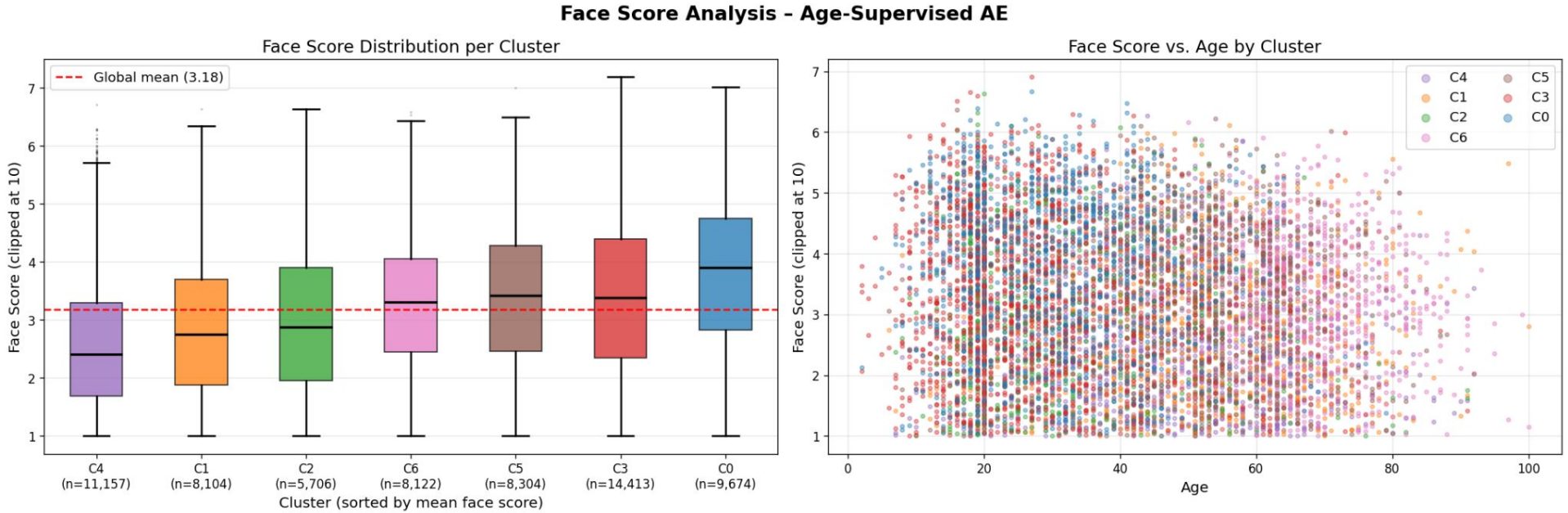
- Cluster Purity: Fraction of the person’s images in the same cluster
  - 1: all images land in the same cluster ⇒ rather identity-based clustering
  - → 0: all images in different clusters ⇒ rather age-based clustering

- Clustering vs. Age-Span:
  - Flat line: clustering is not affected by how the person ages → rather gender, image style, identity, …
  - Negative slope: the clustering is picking up on the age-related appearance change

VGG16 – Facescore Bias?



SAE - Facescore Bias?



Explanation to “Facescore Bias?”

Left - FaceScore vs. Cluster:

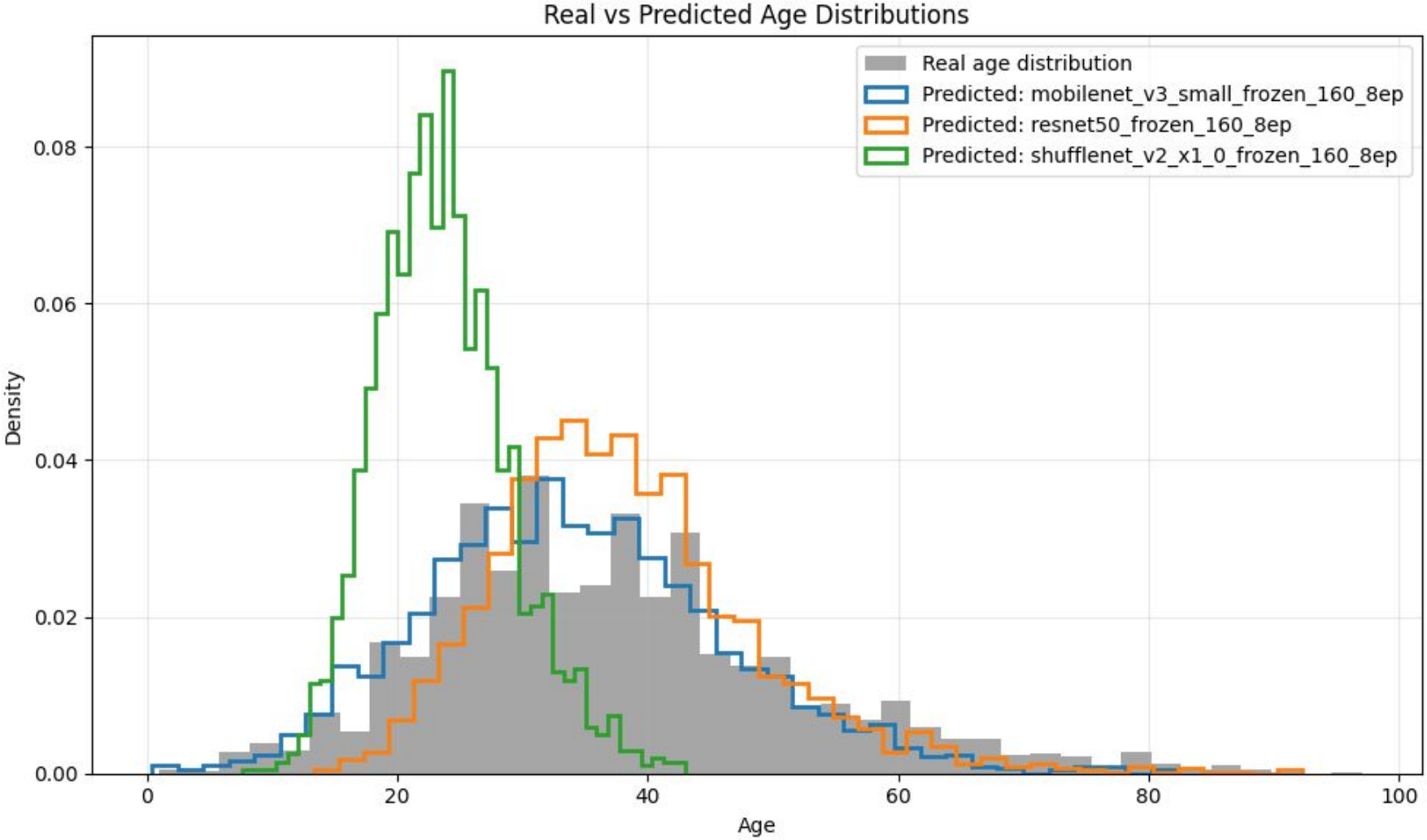
- FaceScore of a cluster completely below mean: maybe grouping images that were hard to detect (blurry, side-profile, occluded) rather than by age, identity, …
- Low number and low FaceScore might indicate a cluster-category of “bad images”

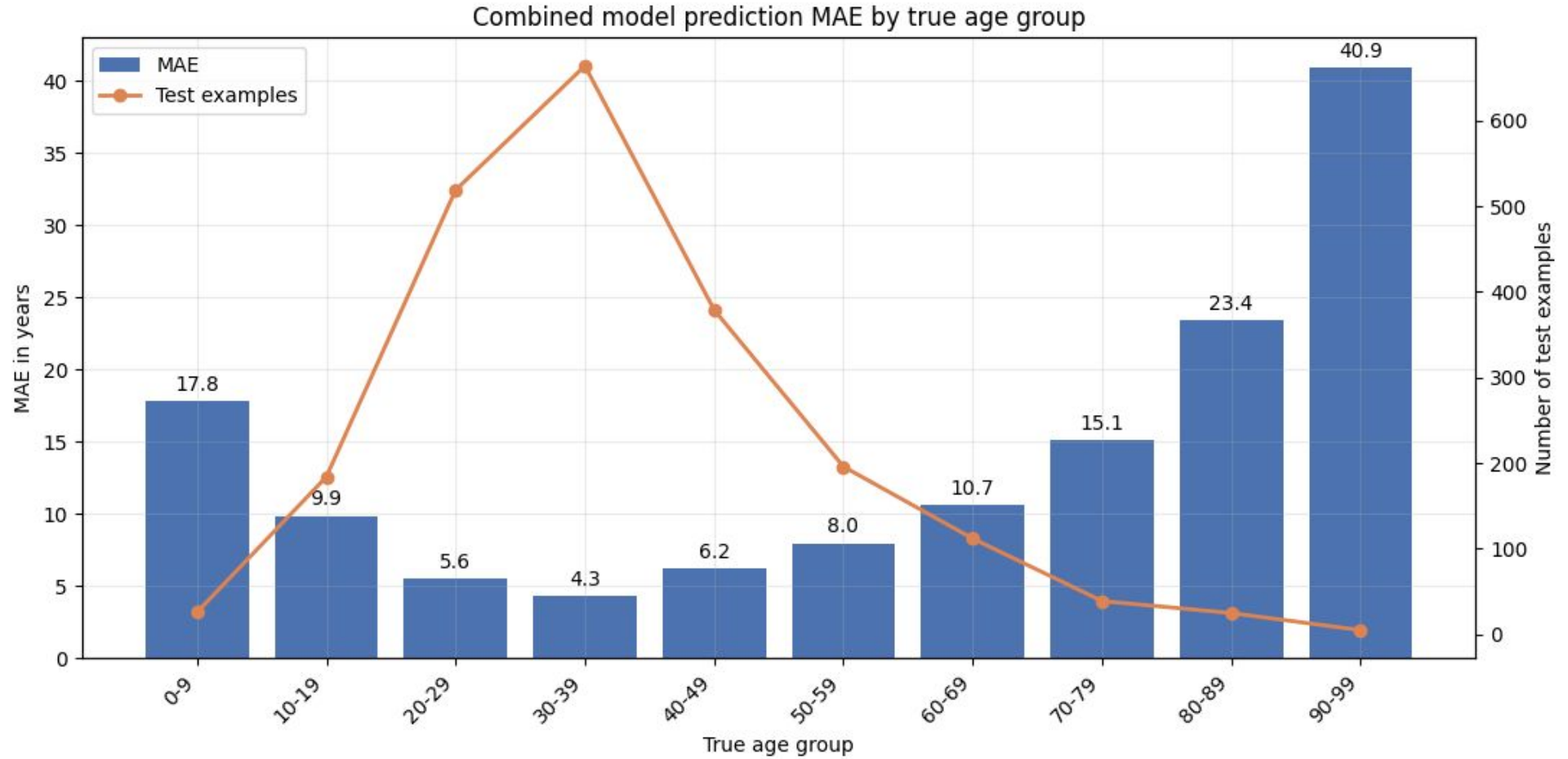
Right: Face-Score vs. Age

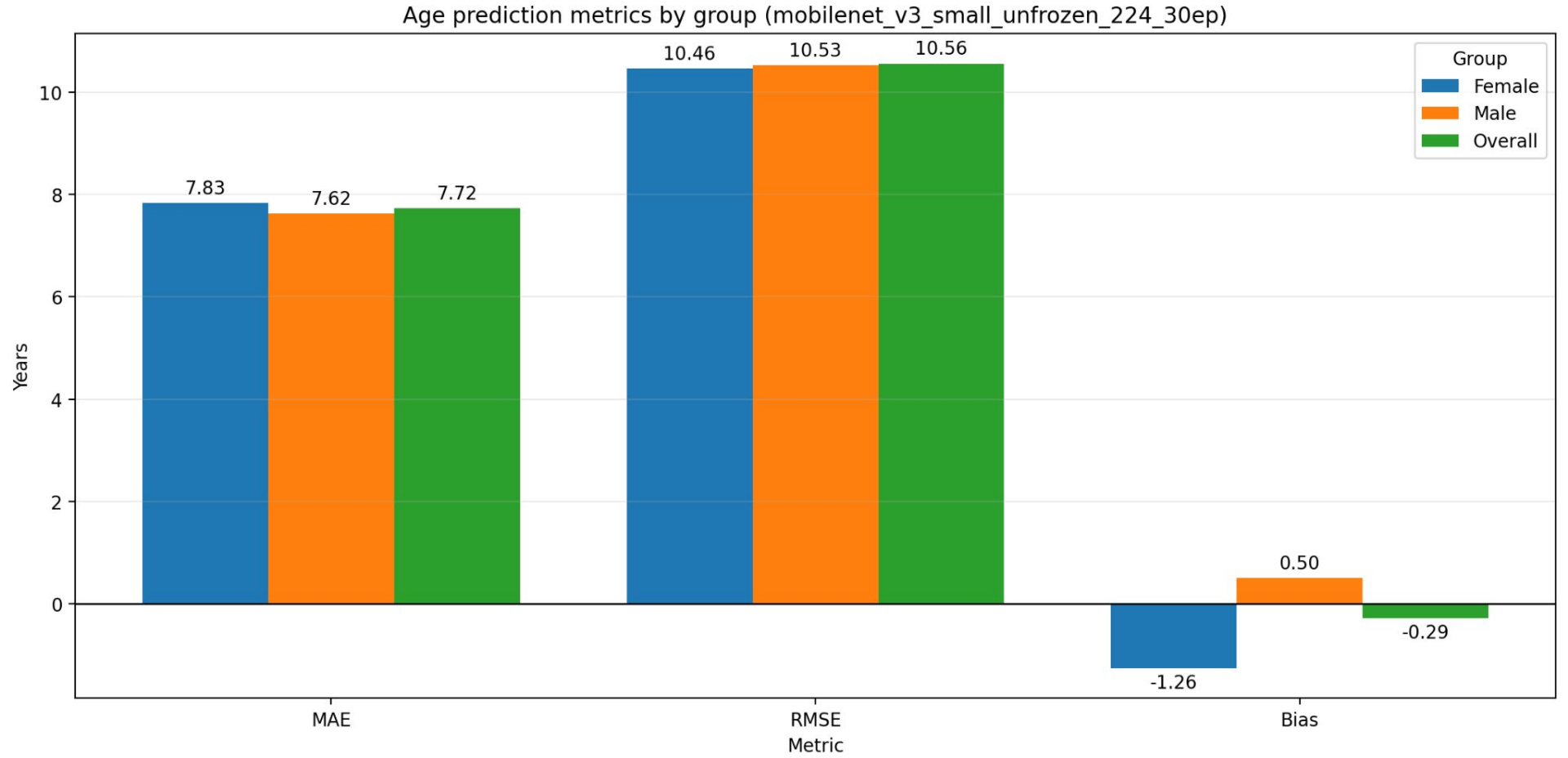
- Diagonal pattern (lower left triangle): might indicate that age and image quality are confounded in the cluster
- Uniform distribution: clustering rather not driven by image artifacts

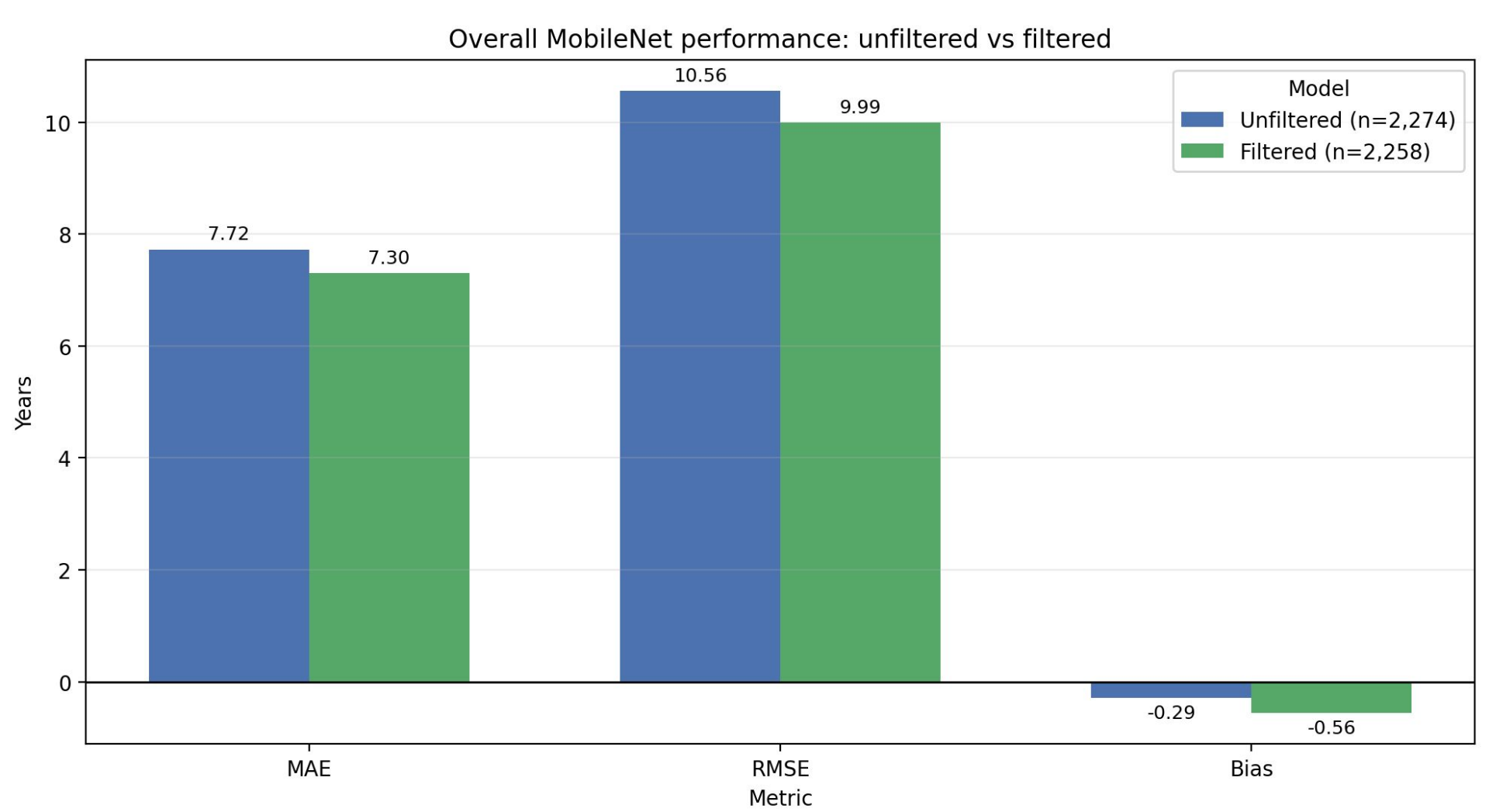
Comparison VGG16 and SAE: SAE seems to be a bit more age-driven, while the clusters C4, C5, C3 in VGG16 seem to be biased by the facescore and/or image-quality

Appendix - Regression


Real vs Predicted Age Distributions
Real age distribution
Predicted: mobilenet_v3_small_frozen_160_8ep
Predicted: resnet50_frozen_160_8ep
Predicted: shufflenet_v2_x1_0_frozen_160_8ep
Density
0.00
0.02
0.04
0.06
0.08
0
20
40
60
80
100
Age


Combined model prediction MAE by true age group
MAE
Test examples
MAE in years
Number of test examples
True age group
0-9
10-19
20-29
30-39
40-49
50-59
60-69
70-79
80-89
90-99
17.8
9.9
5.6
4.3
6.2
8.0
10.7
15.1
23.4
40.9


Age prediction metrics by group (mobilenet_v3_small_unfrozen_224_30ep)
Group
Female
Male
Overall
Years
10
8
6
4
2
0
7.83
7.62
7.72
10.46
10.53
10.56
-1.26
0.50
-0.29
MAE
RMSE
Bias
Metric


Overall MobileNet performance: unfiltered vs filtered
Model
Unfiltered (n=2,274)
Filtered (n=2,258)
Years
Metric
MAE
RMSE
Bias
7.72
7.30
10.56
9.99
-0.29
-0.56
0
2
4
6
8
10

Images with Bad Predictions

true=80.0, pred=39.7, err=-40.3

true=88.0, pred=43.7, err=-44.3

true=92.0, pred=47.7, err=-44.3

true=79.0, pred=20.4, err=-58.6

true=73.0, pred=31.1, err=-41.9

true=78.0, pred=30.8, err=-47.2

true=85.0, pred=36.0, err=-49.0

true=78.0, pred=30.6, err=-47.4

Pretrained Models and ResNet50

- Training a CNN from scratch requires millions of labeled images and significant computational resources.
- Pretrained CNNs have already learned rich visual representations from large-scale datasets
- Transfer learning allows these learned features to be adapted to new tasks with much less data
- Widely used pretrained models include VGG, Inception, EfficientNet, DenseNet, and ResNet
- REsNEt50 was introduced 2015 by Microsoft

ResNet50 Architecture

- Initial 7 x 7 convolutional layer followed by max pooling
- 4 sequential residual stages with increasing feature depth
- Built using bottleneck residual blocks
  - 1 x 1 convolution (dimensionality reduction)
  - 3 x 3 convolution (feature extraction)
  - 1 x 1 convolution (feature expansion)
- Skip connections add the input directly to the block output
- Ends with Global Average Pooling followed by a Fully Connected classification layer
- Total depth: 50 layers and 25 million trainable parameters

ResNet50 Training

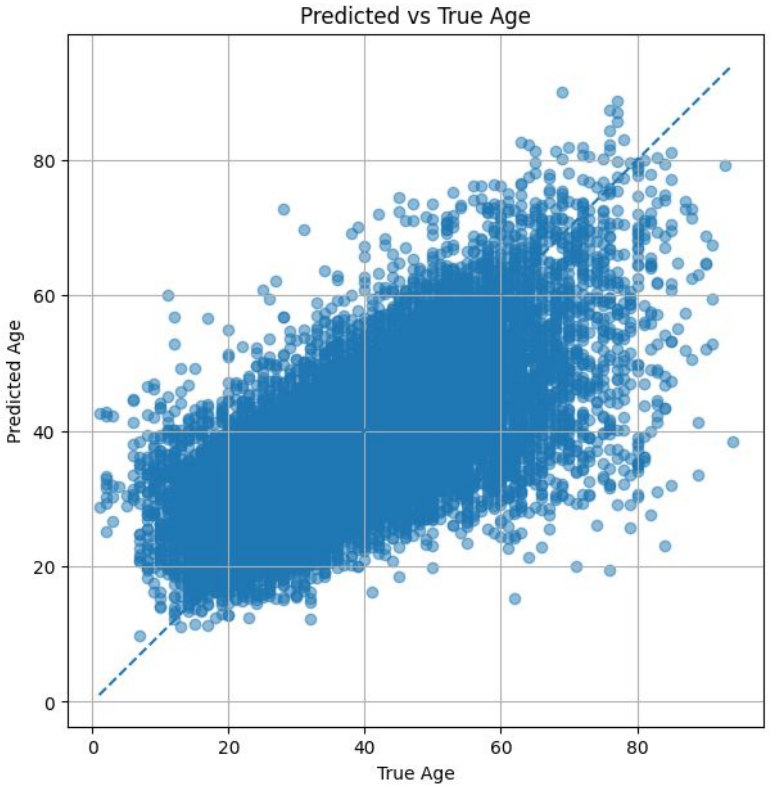
- Originally trained on the ImageNet dataset
- Imagenet contains over 1 million labeled images
- Covers 1.000 objects and categories
- Training uses supervised learning with stochastic gradient descent (SGD)
- Data augmentation techniques improve robustness and generalization

How Did We Apply ResNet50

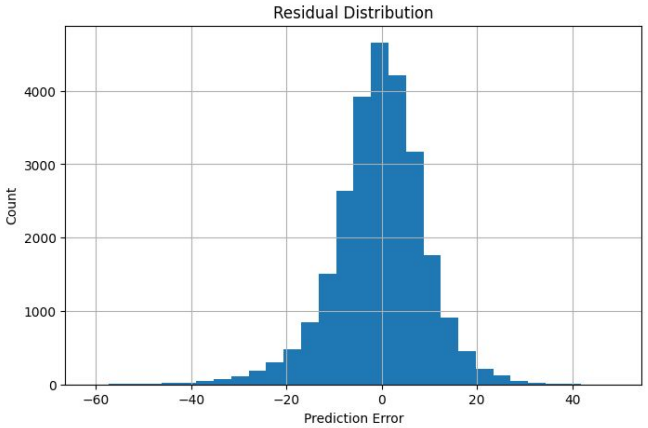
- Initialize the network using pretrained weights instead of random initialization
- Replace the final classification layer with task-specific output layers (Age Regression and Gender Classification)
- Freeze early convolutional layers to preserve general visual features
- Fine-tune deeper layers to learn facial characteristics relevant to age and gender

CNN from Scratch Architecture

- 4 convolutional blocks
- 8 convolutional layers
- Batch Normalization after every convolution
- ReLU activation through out
- Feature expansion: 32 → 64 → 128 → 256
- Parameters

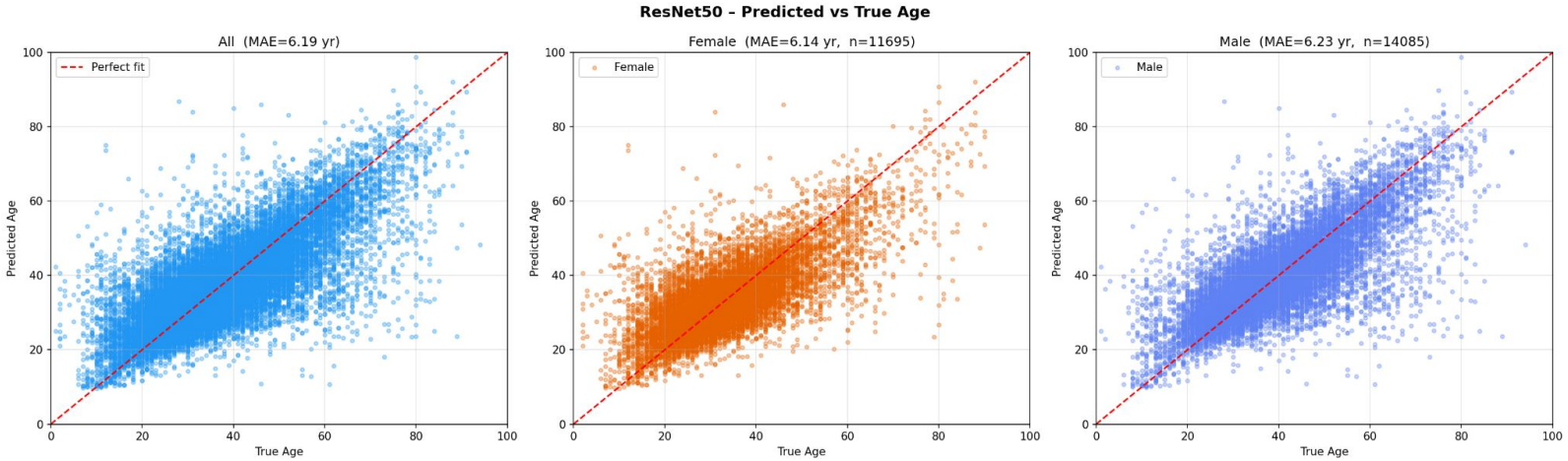


CNN for Age Regression from Scratch



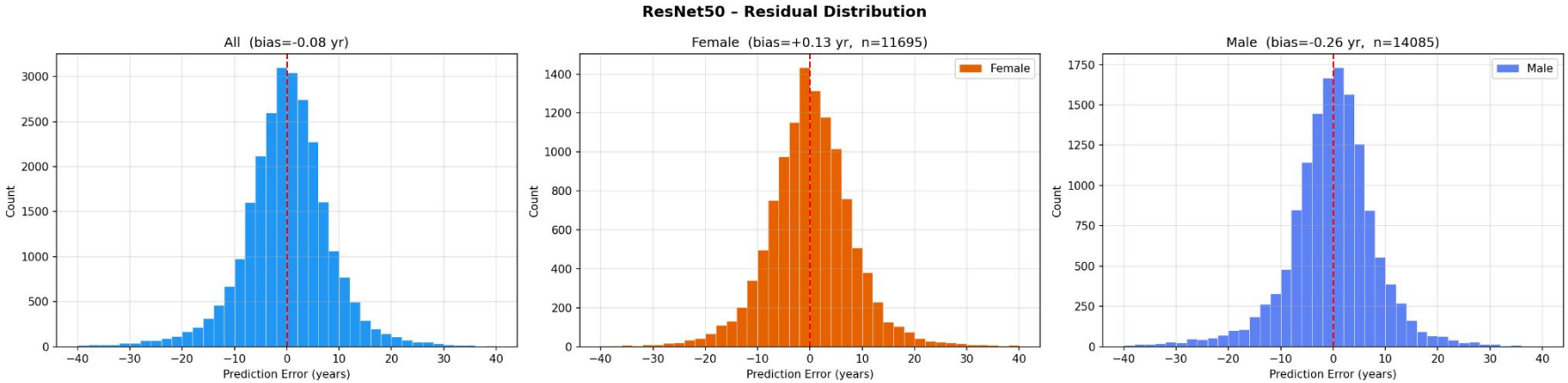
Age Regression ResNet50

True age vs predicted age by gender



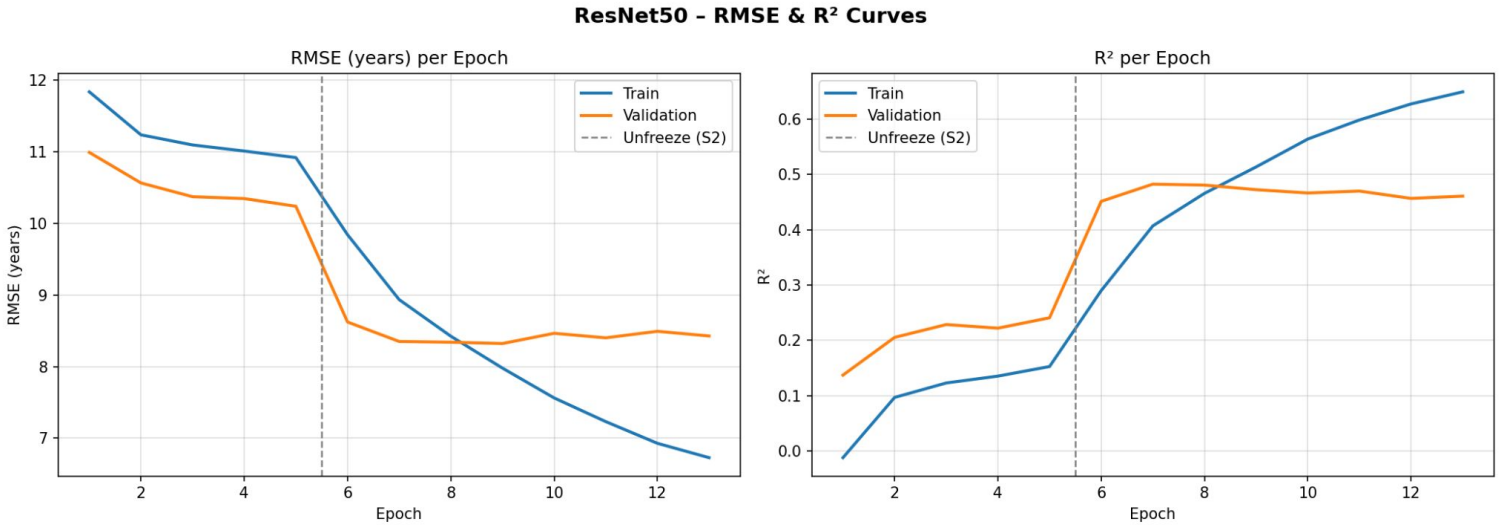
Basic ResNet50 Regression Split by Gender

Residuals by Gender



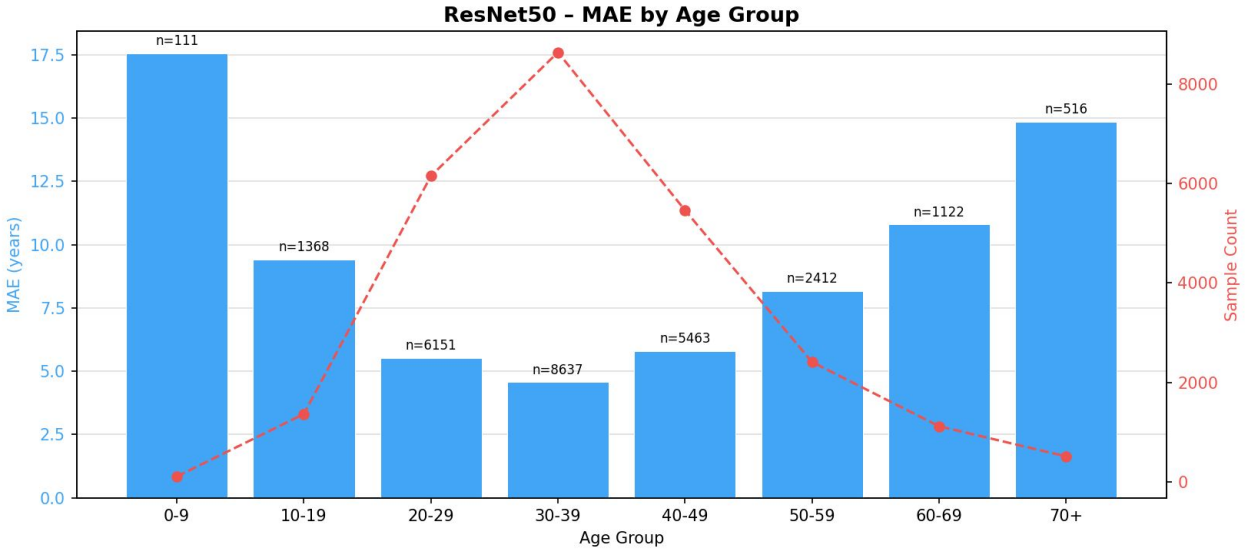
Age Regression ResNet50

Some more metrics, show the same behaviour as MAE



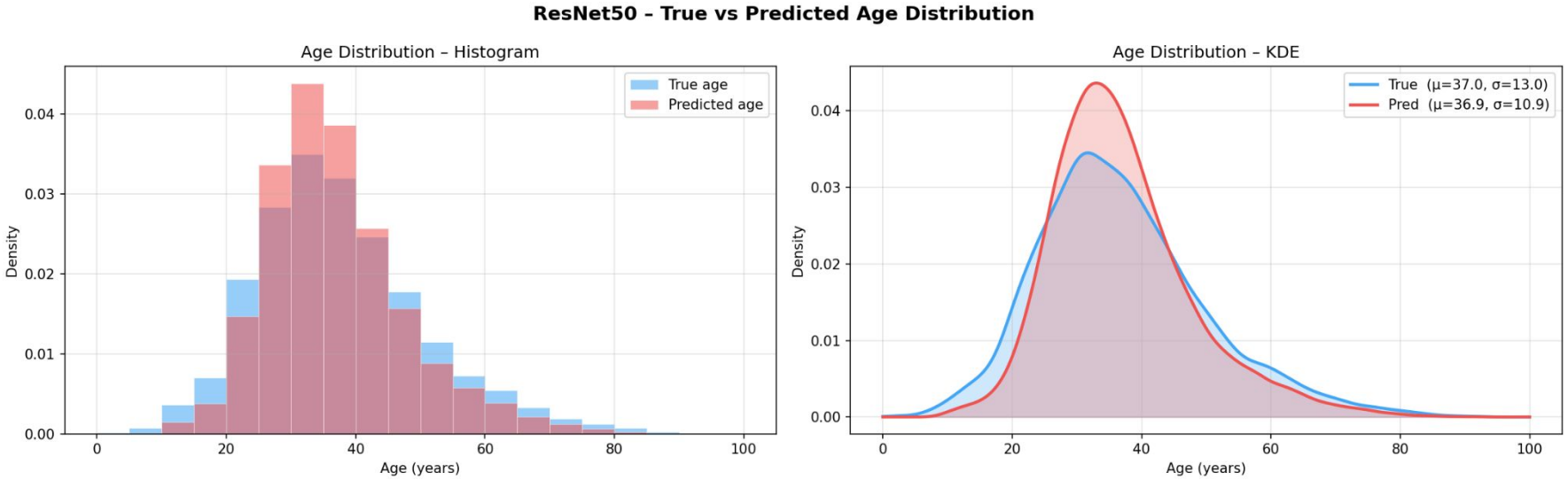
Age Regression Using ResNet50

MAE compared to the number of samples in this age group (inverse correlation)



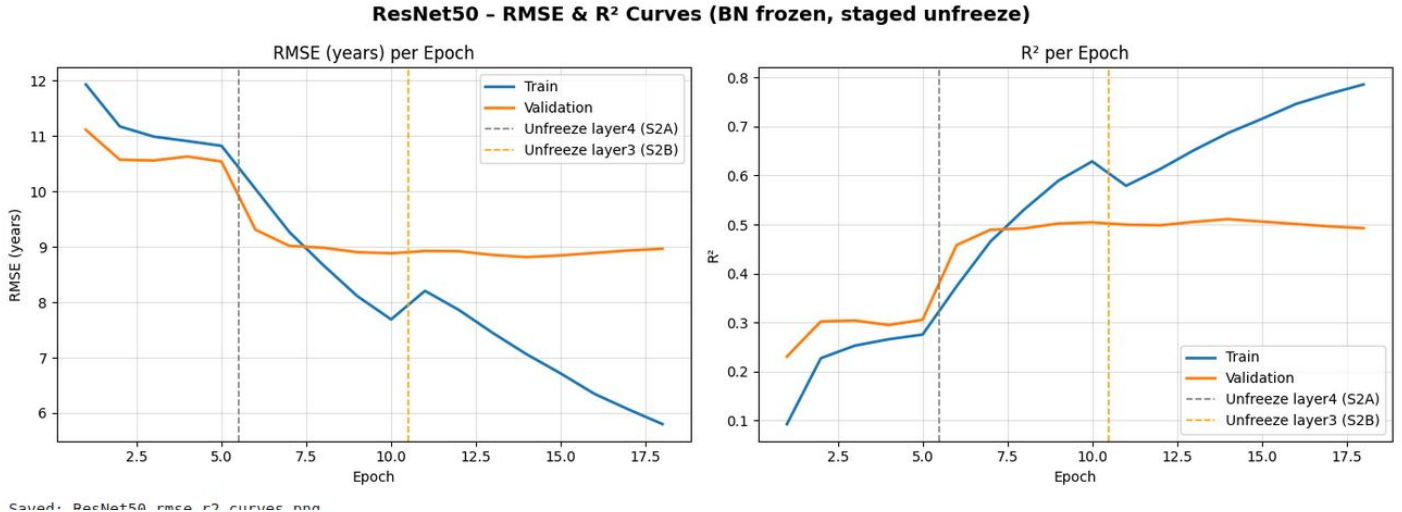
Age Regression Using ResNet50

Predicted vs actual distribution of ages. Left out for time reasons



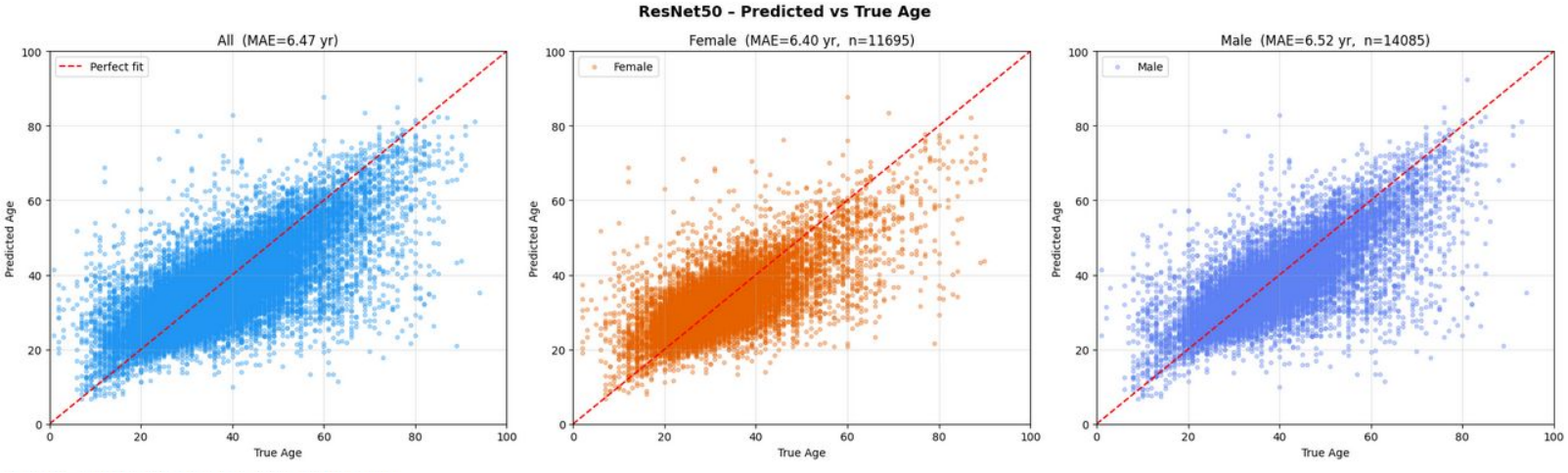
Age Regression Using ResNet50 Improved

Some more metrics, show the same behaviour as MAE (multiple stage case)



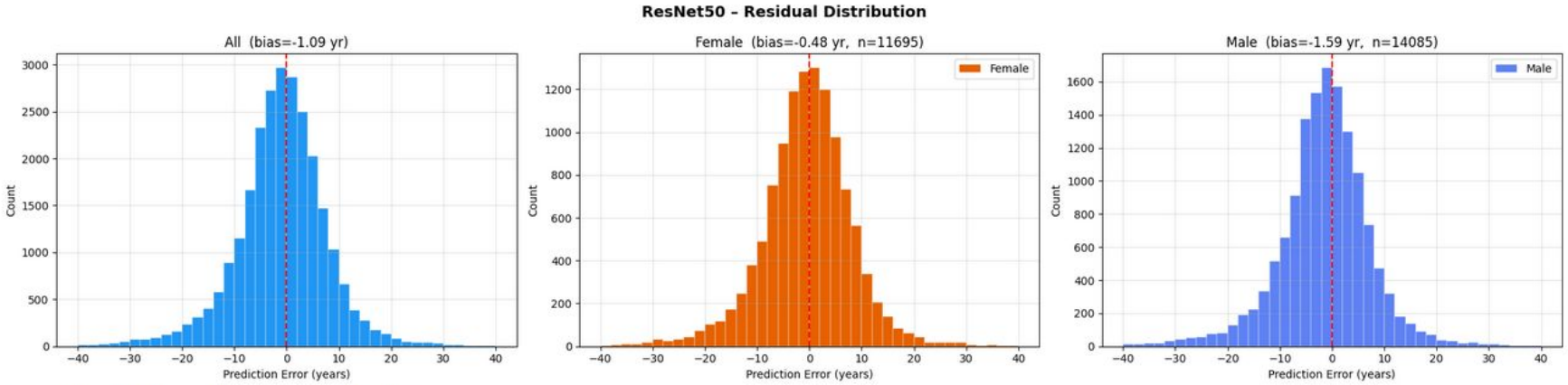
Age Regression ResNet 50 Improved

True age vs predicted age for “improved” unfreezing by gender



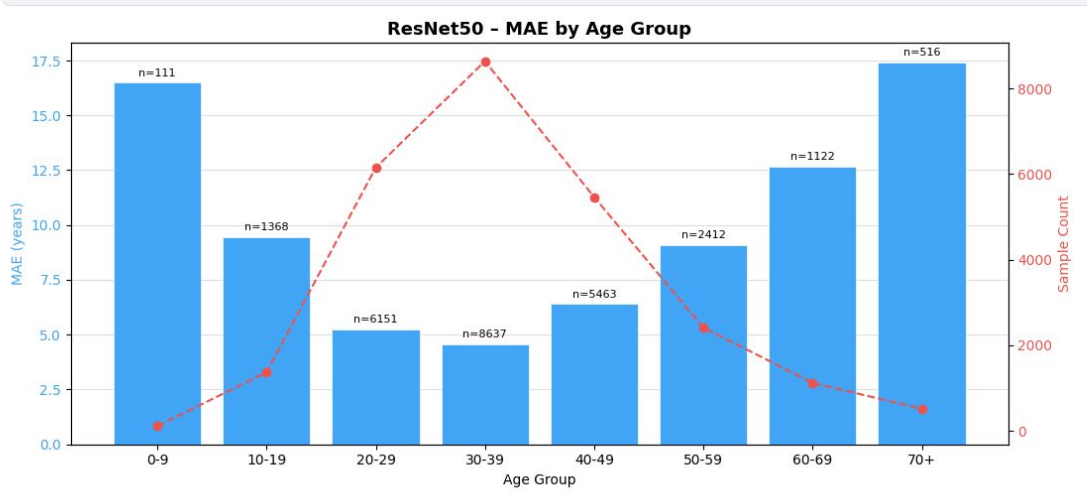
Age Regression ResNet 50 Improved

Residuals for “improved” unfreezing by gender



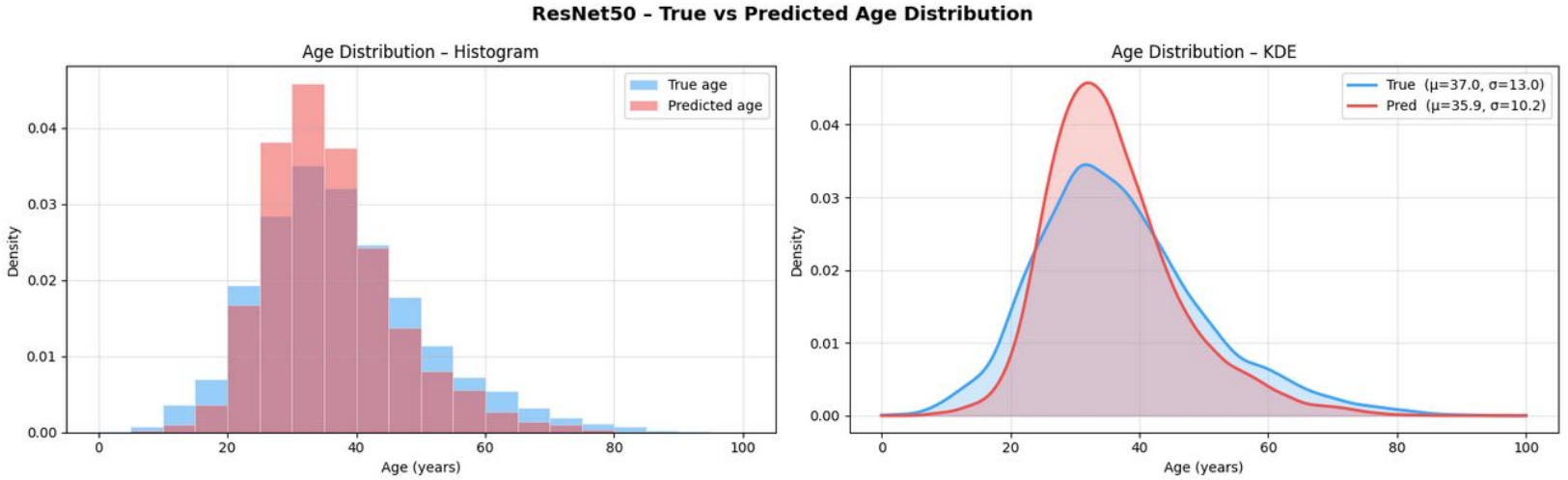
Age Regression ResNet 50 Improved

MAE compared to the number of samples in this age group (inverse correlation) for “improved” unfreezing



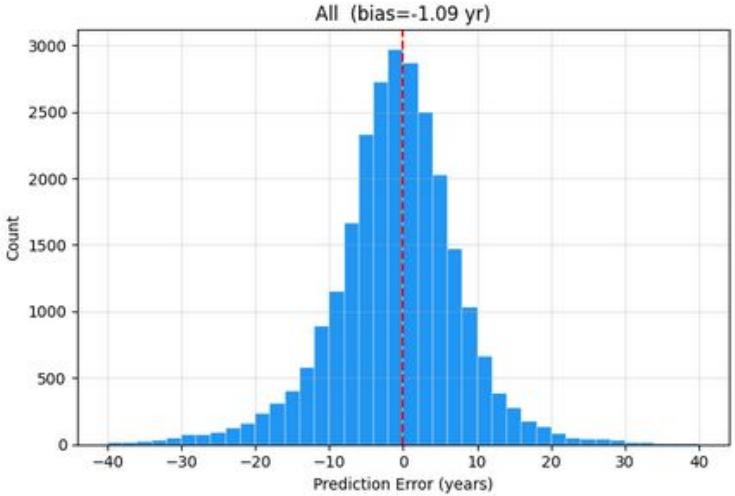
Age Regression ResNet 50 Improved

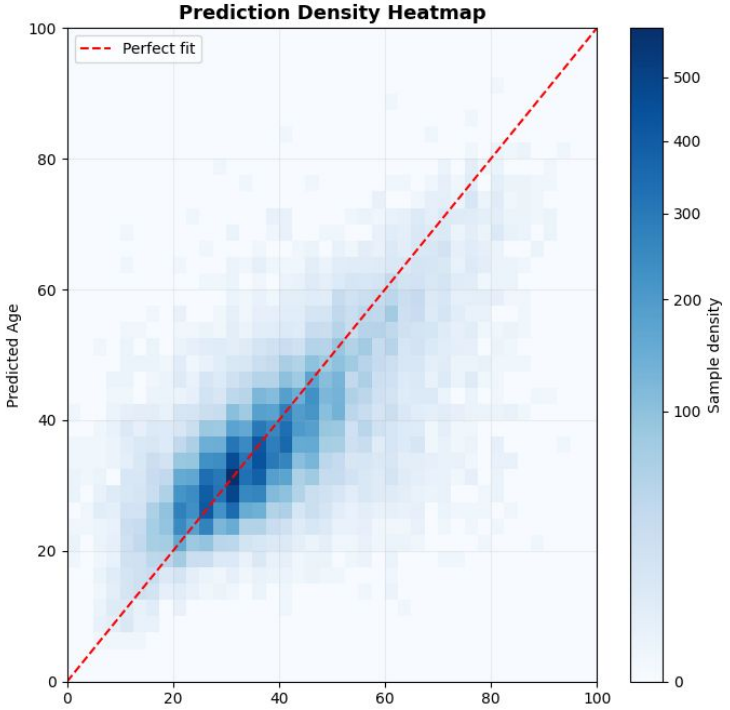
Predicted vs actual distribution of ages for “improved” unfreezing. Left out for time reasons



Appendix - Age Classification

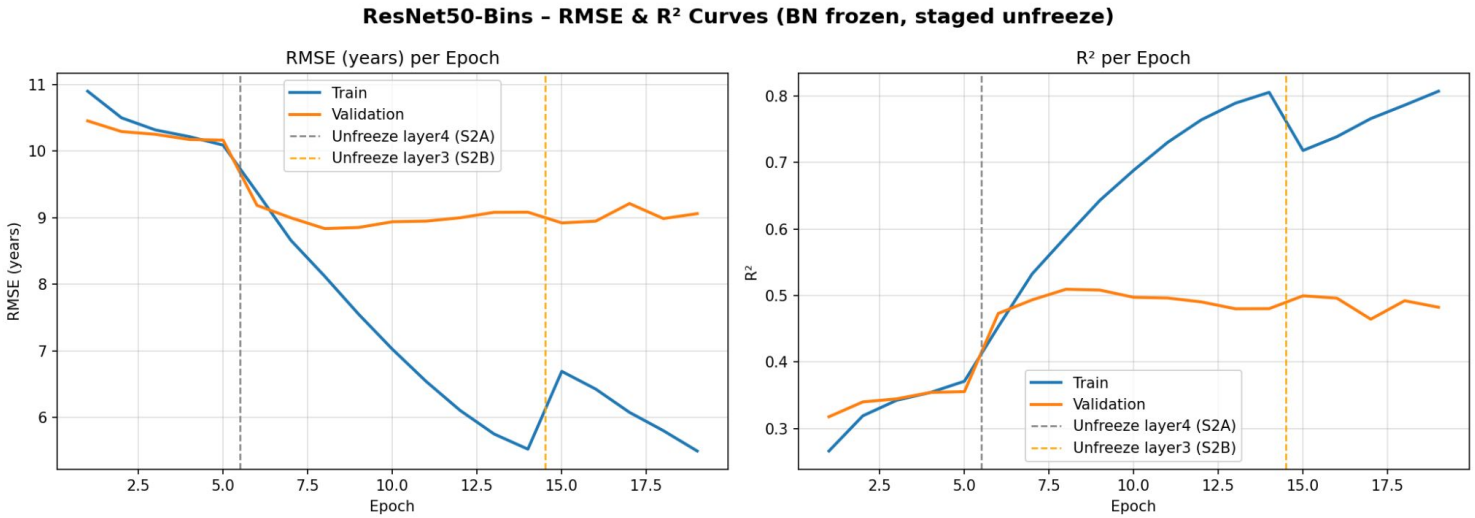
Age Binning


All (bias=-1.09 yr)
Count
Prediction Error (years)
3000
2500
2000
1500
1000
500
0
−40
−30
−20
−10
0
10
20
30
40


Prediction Density Heatmap
Perfect fit
Predicted Age
Sample density
100
80
60
40
20
0
0
20
40
60
80
100
500
400
300
200
100
0

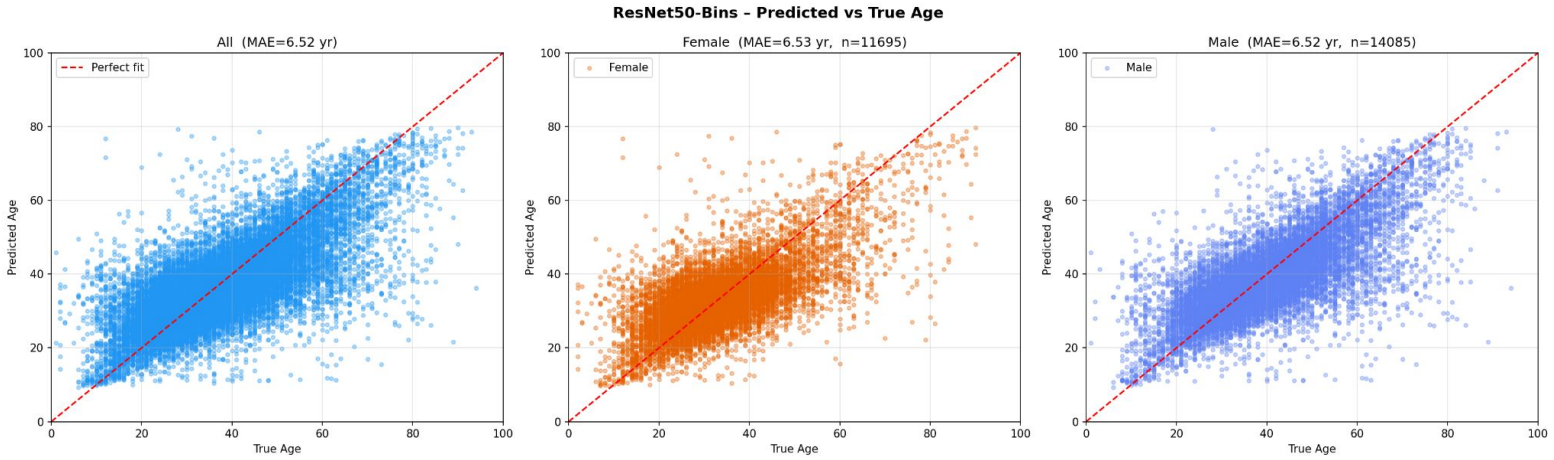
Age Binning

Some additional metrics



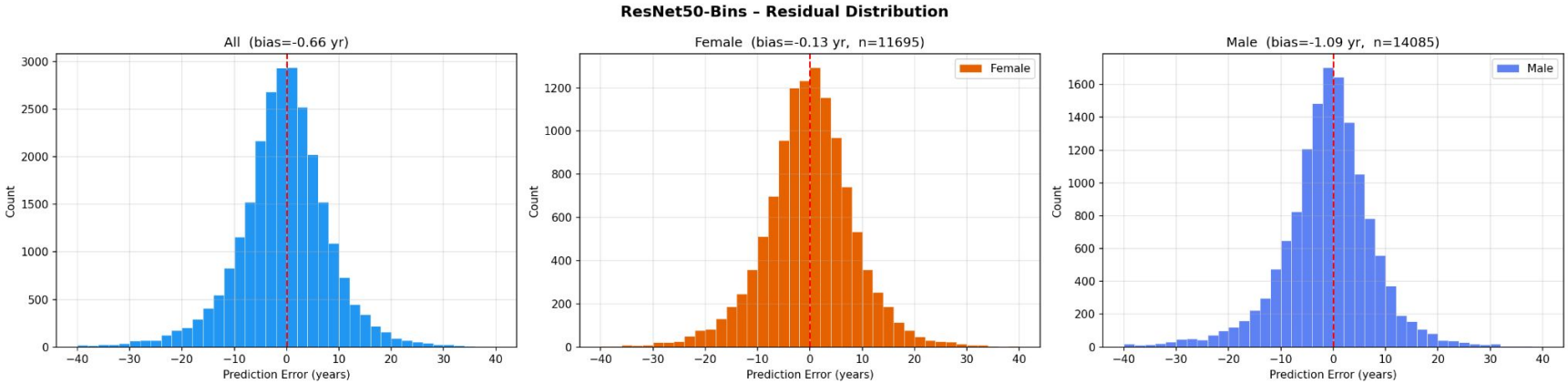
Age Binning

True vs predicted age



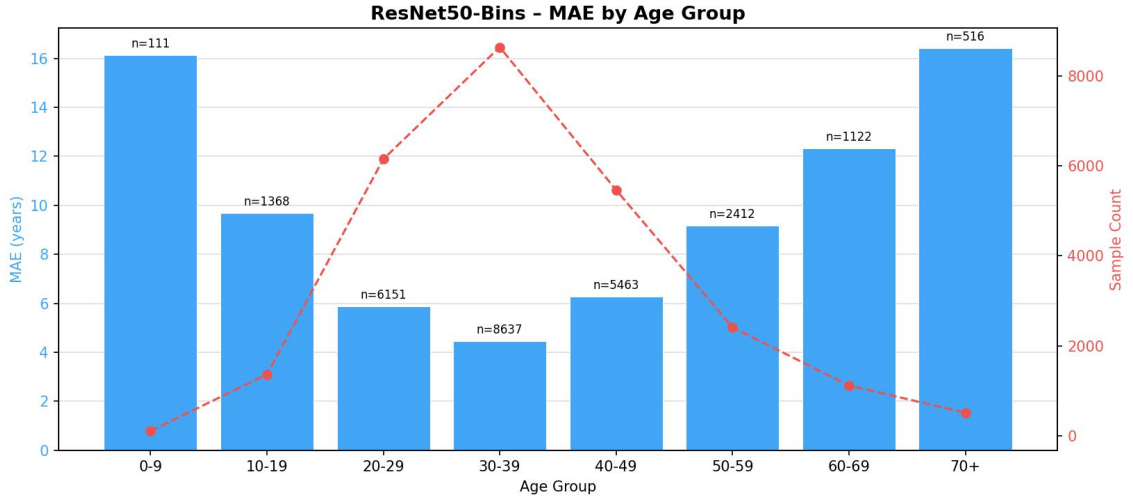
Age Binning

Residuals by gender for the age binning setting



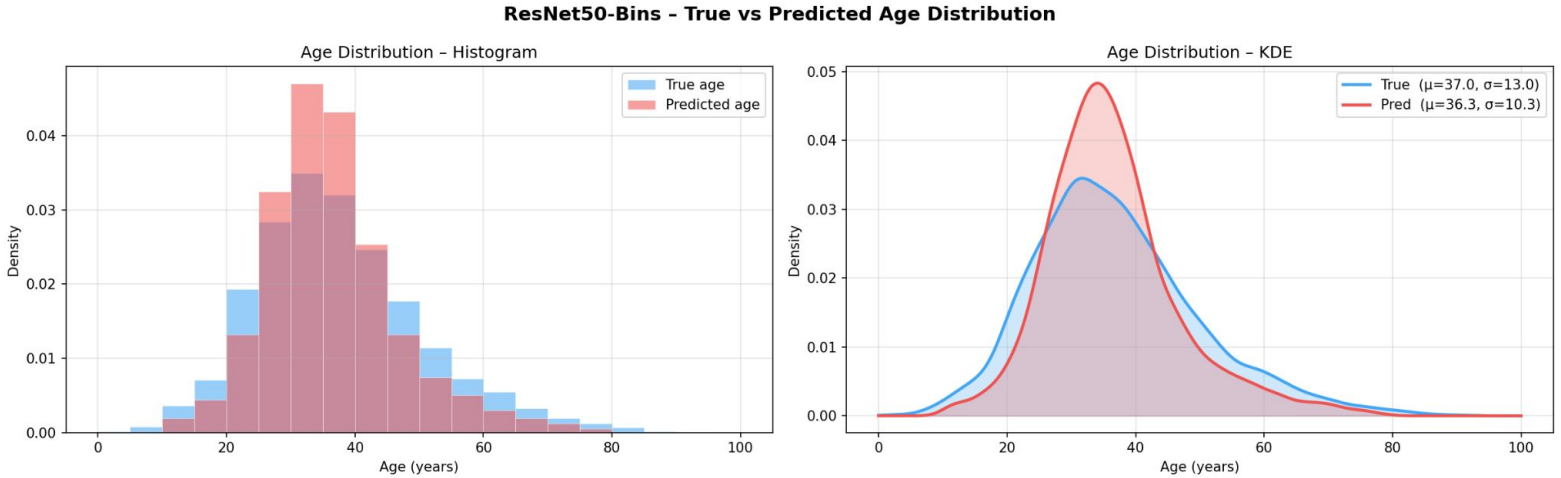
Age Binning

MAE compared to the number of sample for the classification approach

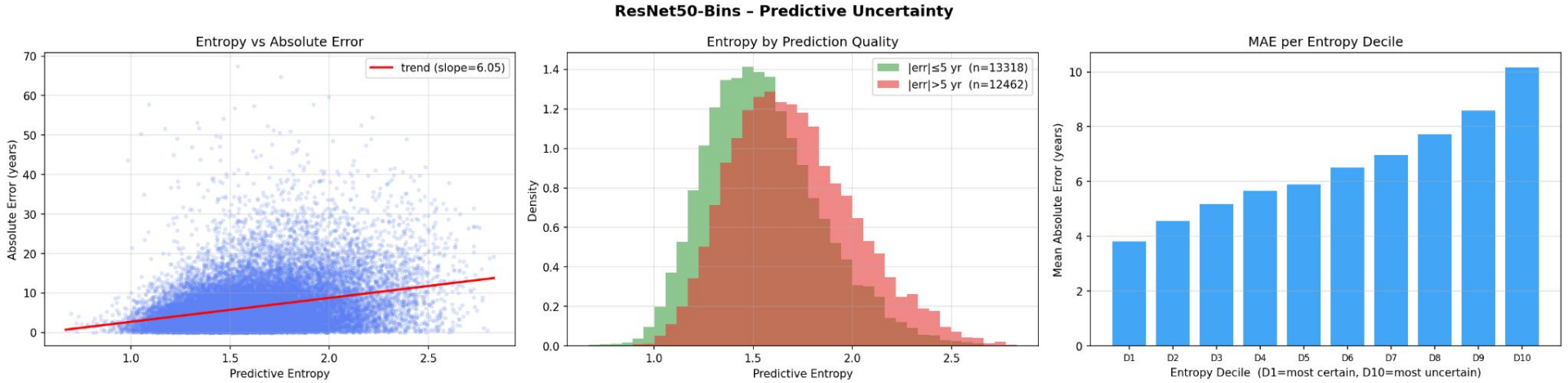


Age Binning

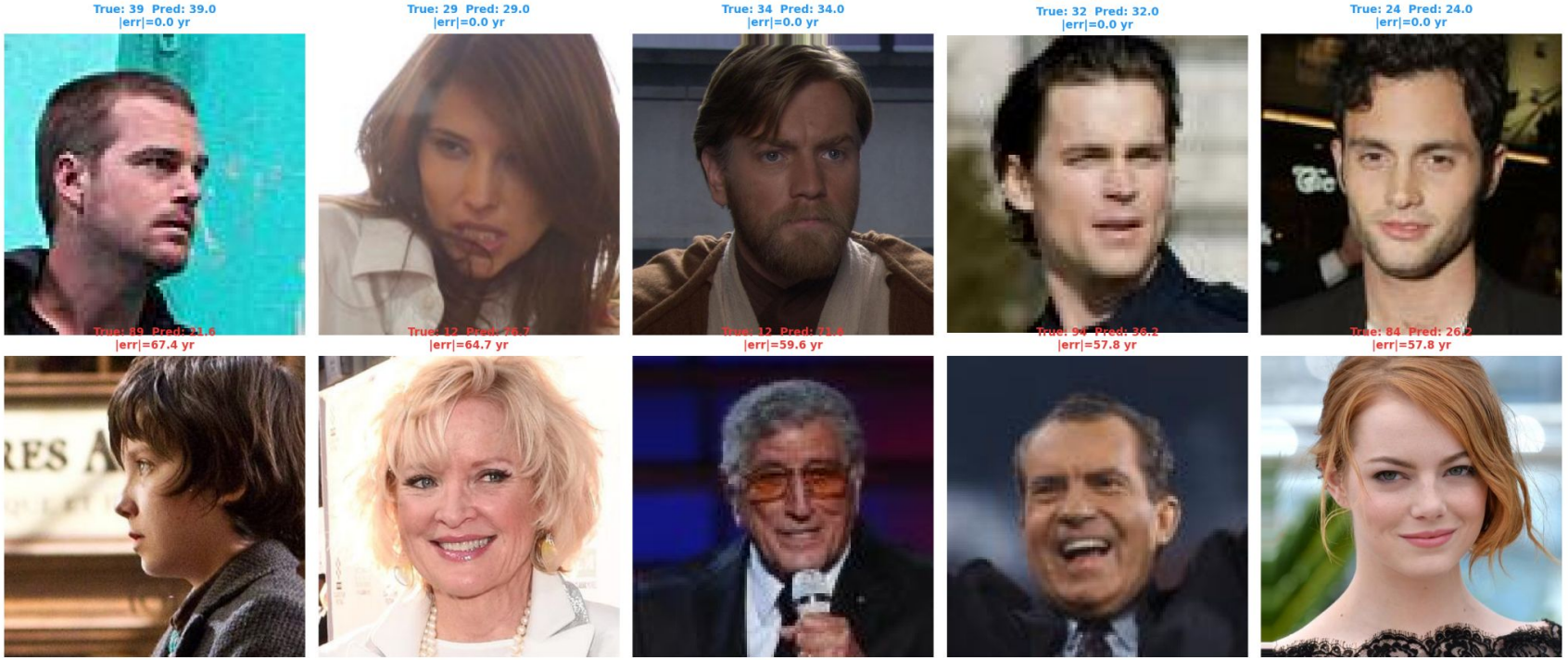
True vs predicted age distribution by the classification approach



Age Binning


ResNet50-Bins – Predictive Uncertainty
Entropy vs Absolute Error
trend (slope=6.05)
Absolute Error (years)
Predictive Entropy
Entropy by Prediction Quality
|err|≤5 yr (n=13318)
|err|>5 yr (n=12462)
Density
Predictive Entropy
MAE per Entropy Decile
Mean Absolute Error (years)
D1
D2
D3
D4
D5
D6
D7
D8
D9
D10
Entropy Decile (D1=most certain, D10=most uncertain)

Best and Worst Age Predictions



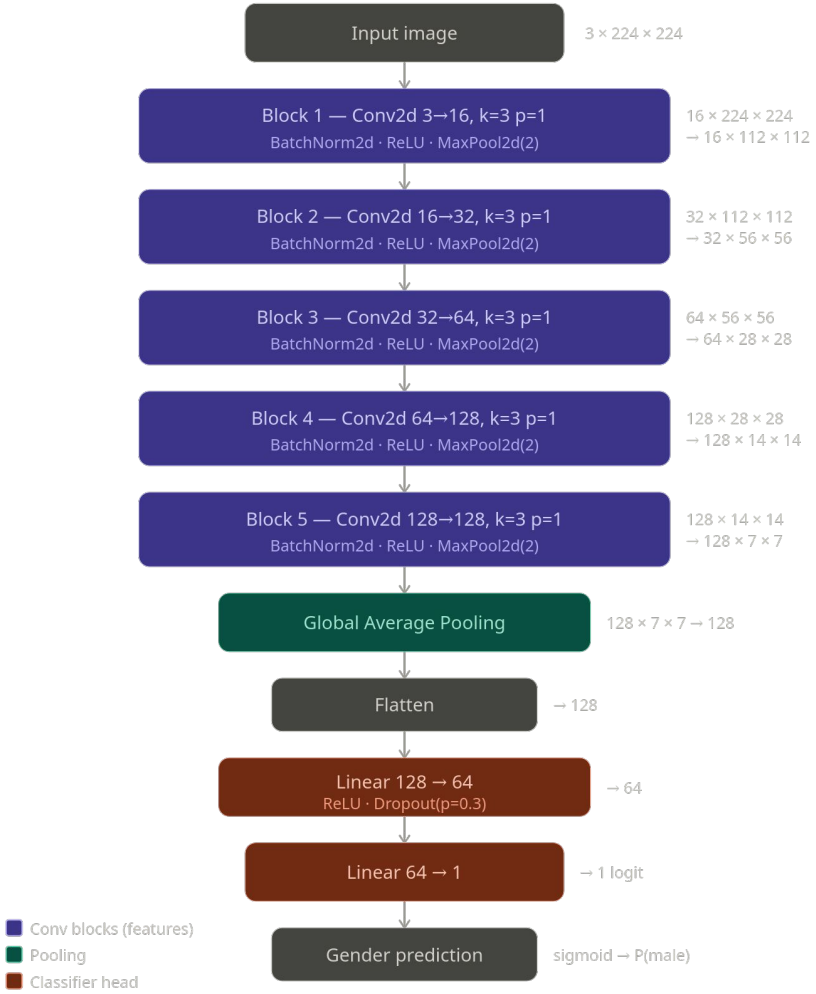
Appendix - Gender Classification

Baseline CNN

Runtime

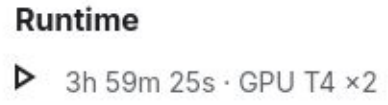
▷ 4h 57m 31s · GPU T4 ×2

total params: 253,713



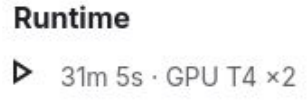
ResNet50 and MobileNetV3

ResNet50 and MobileNetV3 for full dataset:



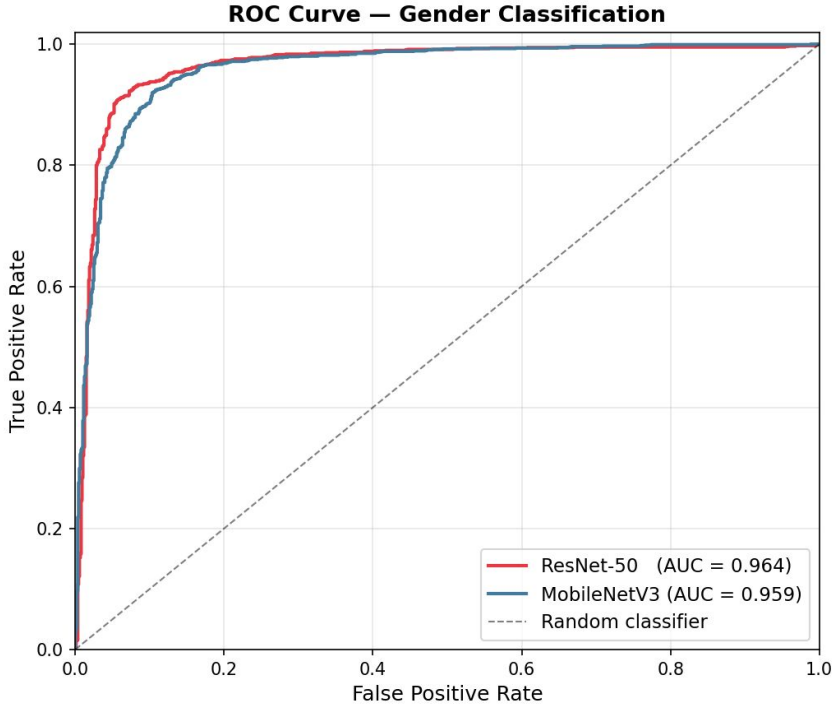

```
ResNet-50         total: 23,510,081, trainable: 23,510,081
MobileNetV3       total: 4,203,313  trainable: 4,203,313
```

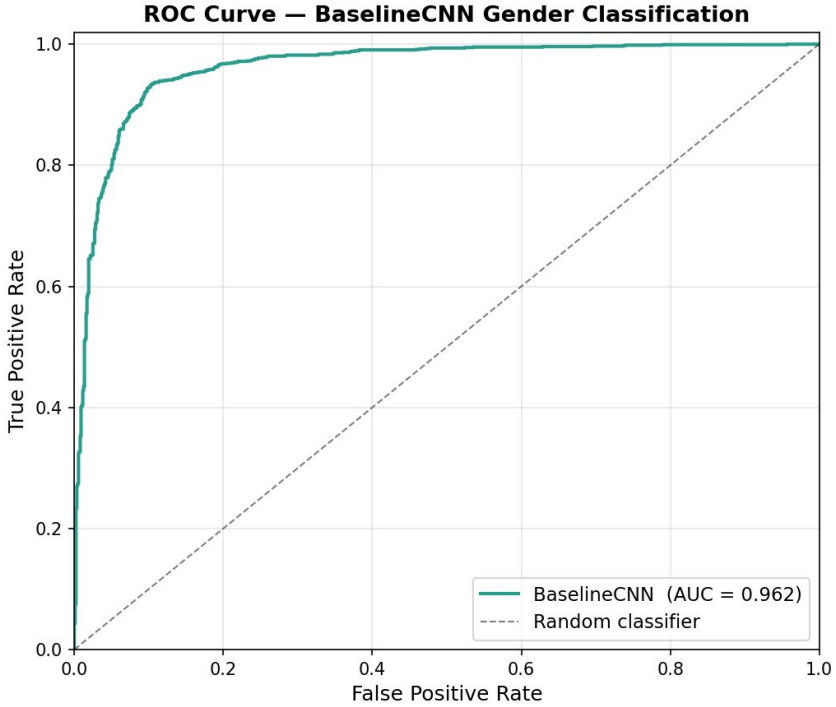
ResNet50 and MobileNetV3 for 1 pic per celeb dataset:



ResNet-50 total: 23,510,081 trainable: 23,510,081
MobileNetV3 total: 4,203,313 trainable: 4,203,313

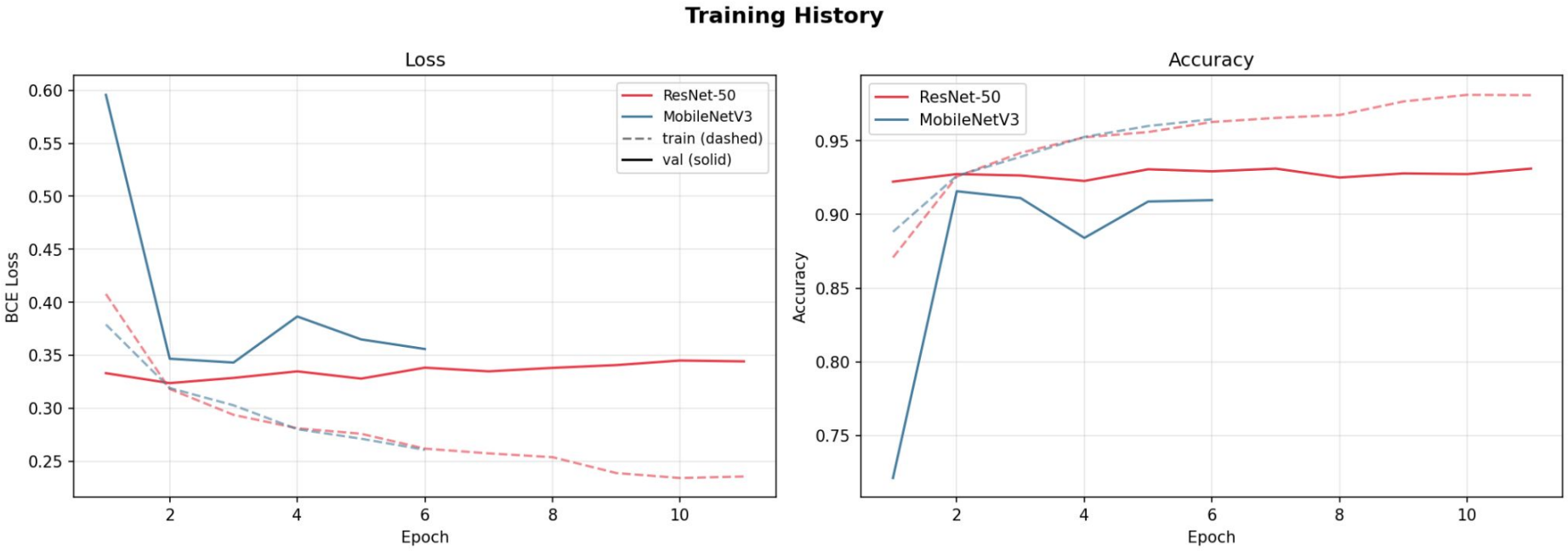
Gender Classification - Model Evaluation





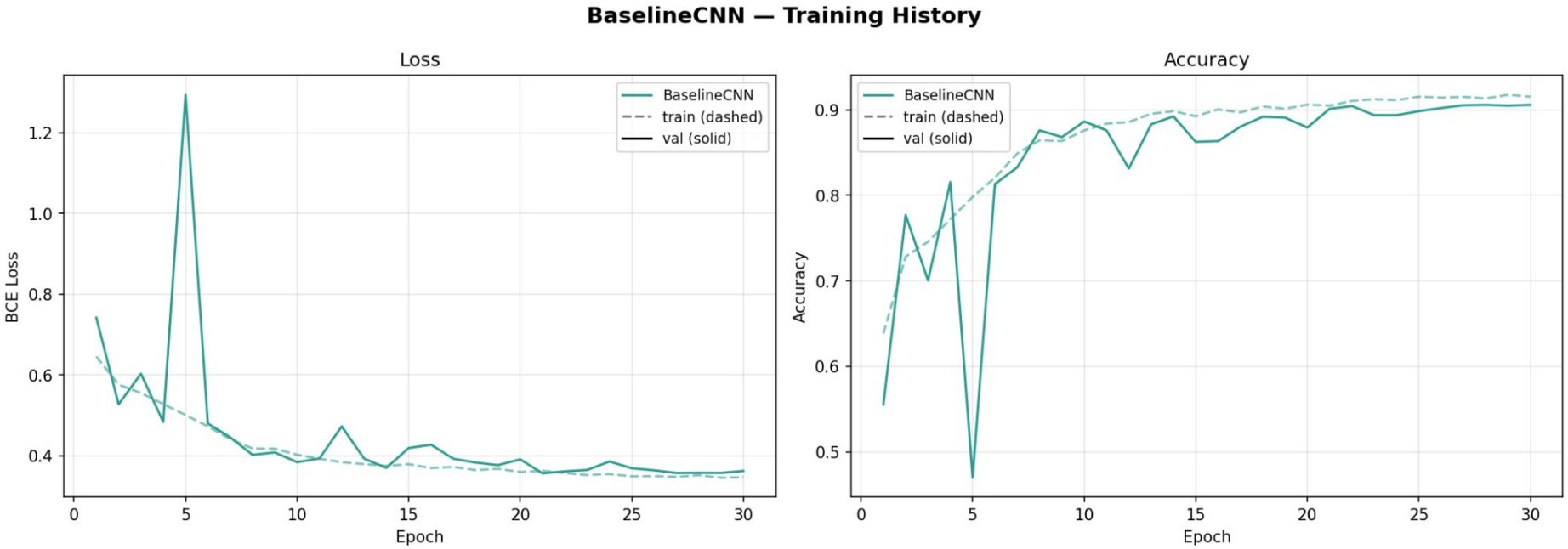
dataset: best face score picture by celeb ~15000 pictures

Gender Classification - Model Evaluation



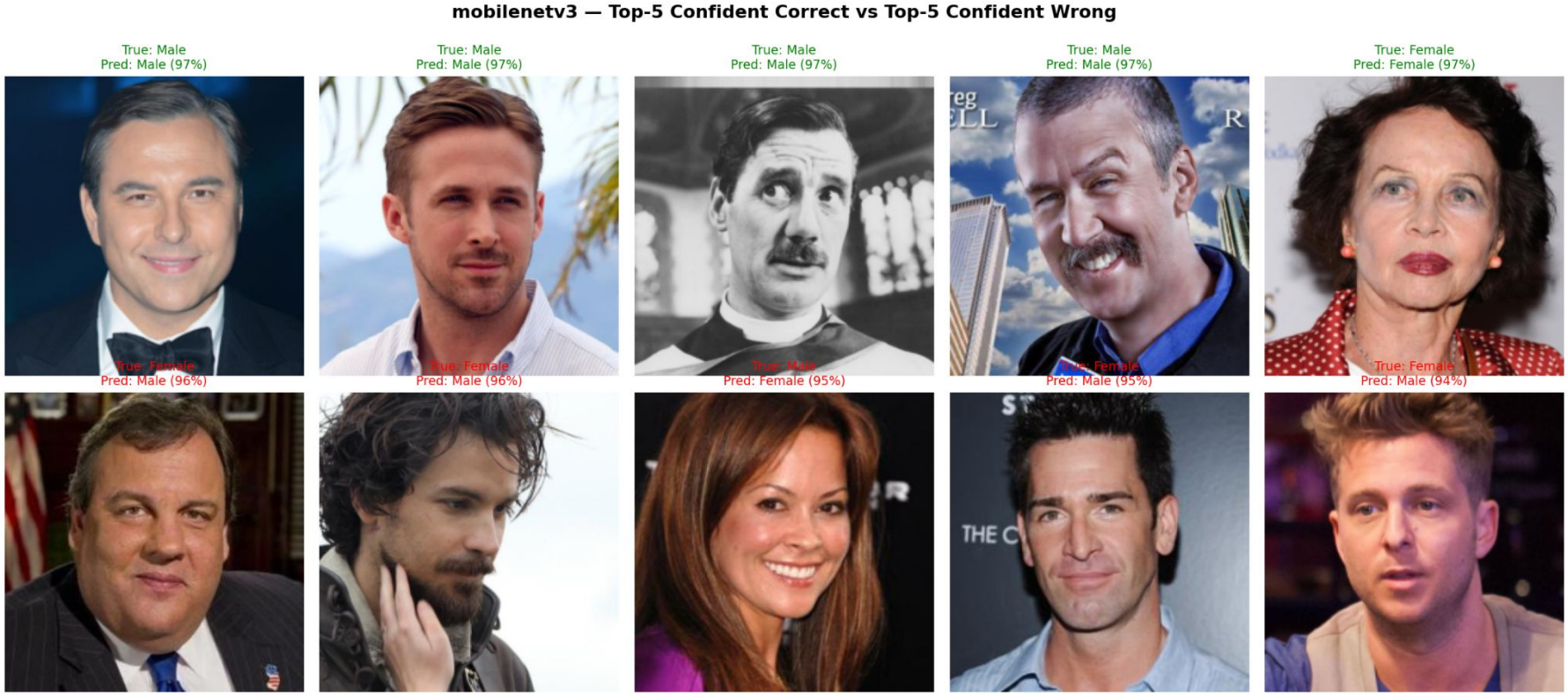
dataset: best face score picture by celeb ~15000 pictures

Gender Classification - Model Evaluation



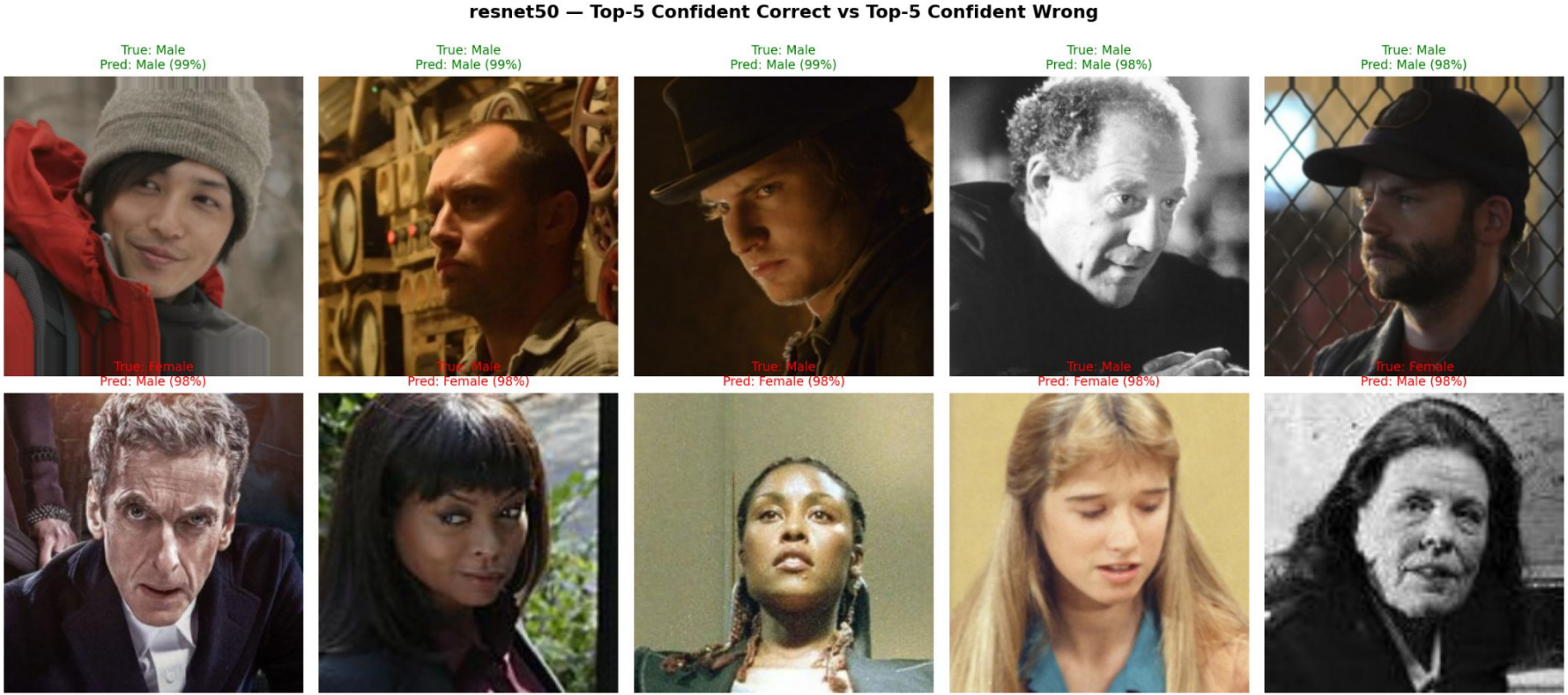
dataset: best face score picture by celeb ~15000 pictures

Gender Classification - mobilenet evaluation



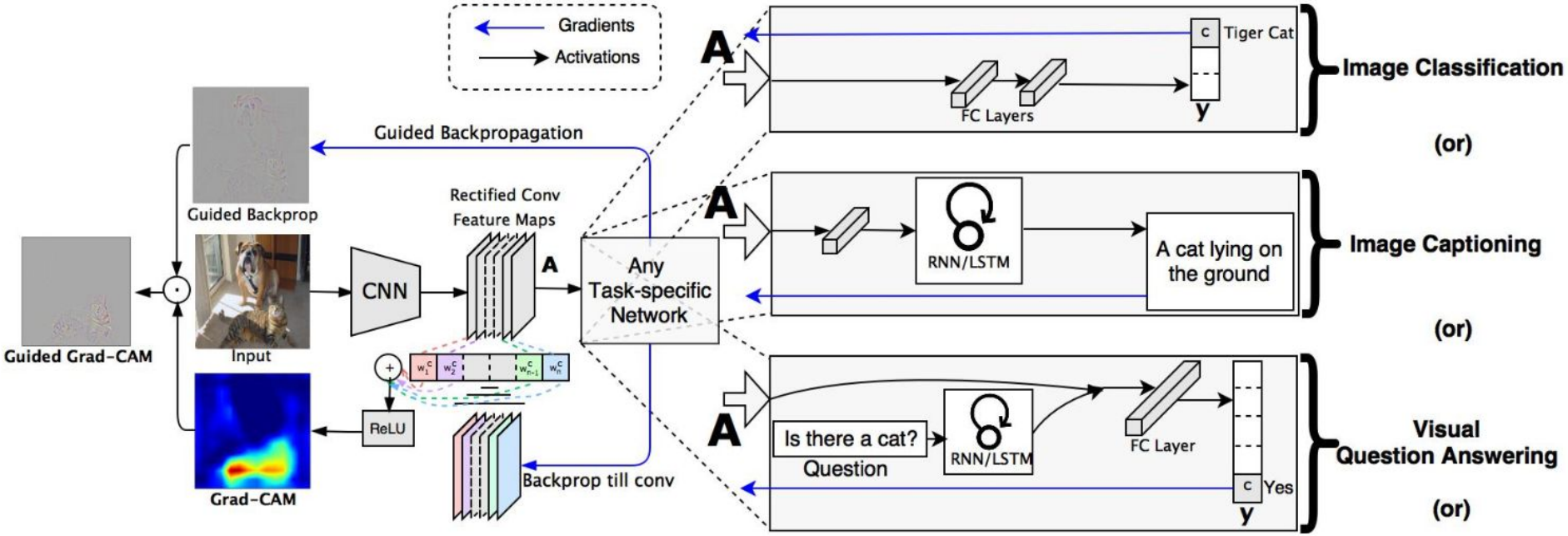
Gender Classification – Baseline CNN Evaluation

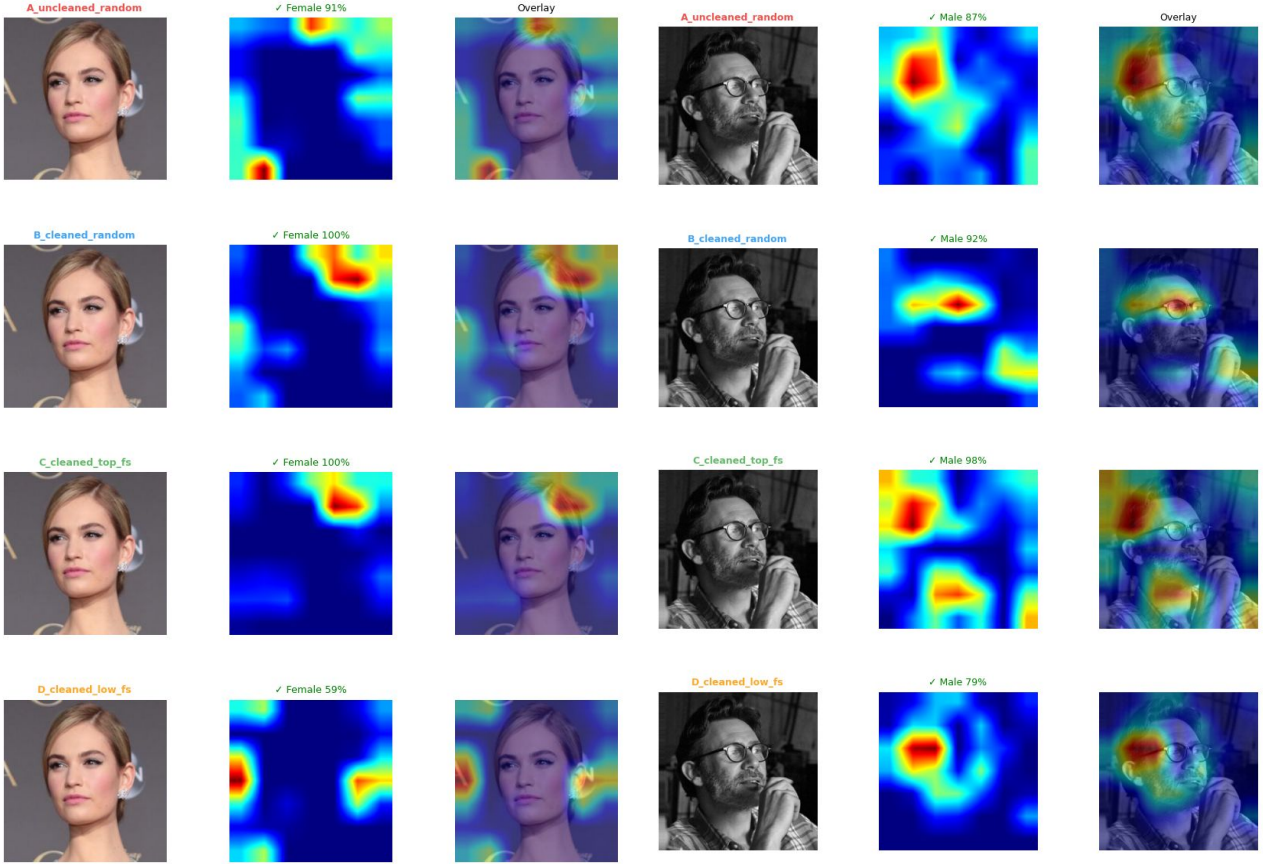
Gender Classification - ResNet Evaluation



https://arxiv.org/abs/1610.02391

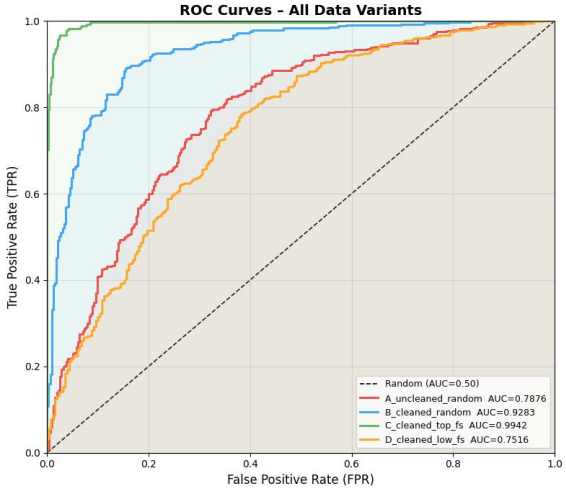
Grad-CAM Architecture





referring to slide 12 models

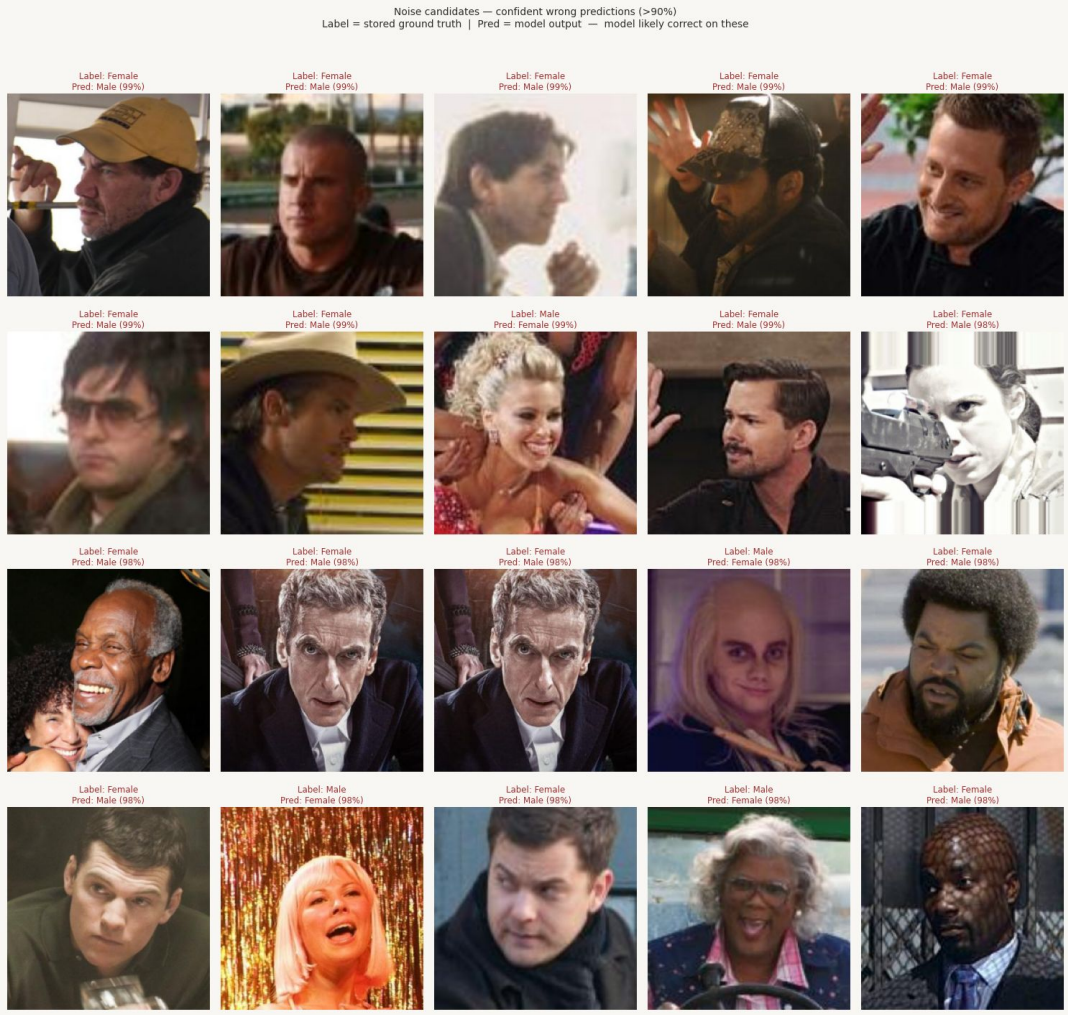
light mobilenet



Gender Classification - Improving Noisy Labels

Procedure:

- Sort out confidently correct classified pictures by model with label mismatch
- Evaluate model again



ResNet50 very confidently off predictions

Gender Classification - Improving Noisy Labels