Supplementary: Photo Dating by Facial Age Aggregation

Jakub Paplhám
Czech Technical University in Prague
paplhjak@fel.cvut.cz

Vojtěch Franc
Czech Technical University in Prague
xfrancv@fel.cvut.cz

Stratified Performance Analysis

Figure 4 presents a detailed performance analysis of our face-based model against the *Scene* baseline, stratified by both the number of faces and the average age of subjects. The analysis reveals that while the performance of the *Scene* baseline is higher for older subjects, our face-based method is significantly more effective for images of younger individuals. This performance advantage for our method further increases with the number of identifiable faces in an image.

It is crucial, however, to interpret the *Scene* baseline performance as an optimistic, best-case scenario, as it was trained and evaluated on in-distribution data from CSFD-1.6M. The impact of this in-domain advantage is starkly evident when evaluating performance on images containing faces for which our pipeline found no matching identities. For images with at least one known identity, the *Scene* baseline achieves a Mean Absolute Error (MAE) of 3.25 years. This error increases sharply to 5.35 years for images containing only unmatched faces (examples in Figure 5).

We hypothesize that the *Scene* model learns not only from general temporal cues in the full image (e.g., fashion, color processing) but also implicitly learns to recognize frequently occurring actors and associate their specific appearance with a given time period. This dual reliance on the training distribution manifests in its failure modes. Performance degrades significantly when the model is confronted with faces that belong to less-known actors or extras. Moreover, as illustrated in Figure 2, the model exhibits a strong temporal bias: its prediction error is lowest for years that are most frequent in the training set and highest for minority years. The ability of our proposed face-based model—which leverages external training data—to match and often outperform this powerful in-domain baseline is a testament to the potential of our approach.

CSFD-1.6M Quality vs. Cleaned IMDB-WIKI

To distinguish the contribution of data quality from that of data quantity, we conducted a size-controlled experiment. We trained a ViT-B/16 age estimation model on a random subset of our dataset, referred to as CSFD-0.3M (≈300,000 faces). To ensure the comparison was against the strongest possible baseline, we trained identical ViT-B/16 models on multiple cleaned versions of the IMDB-WIKI dataset, including the original [3], IMDB-CLEAN [2], and the top-performing EM-CNN version [1]. We selected the EM-CNN version (≈310,000 faces) for the definitive comparison, as it is both the strongest baseline and the most comparable in size to our subset.

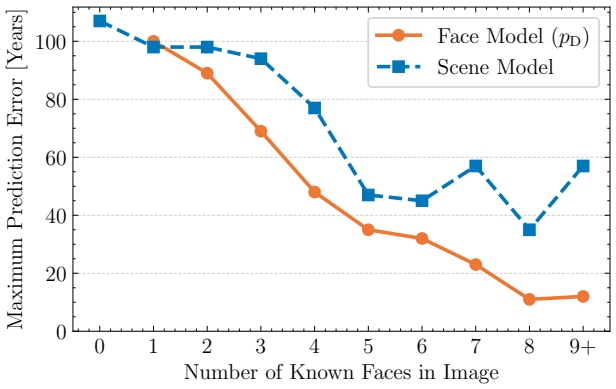


Figure 1. **Worst case error analysis.** Worst MAE ↓ of the *Face* (p_D) and *Scene* methods by the number of faces in the image.

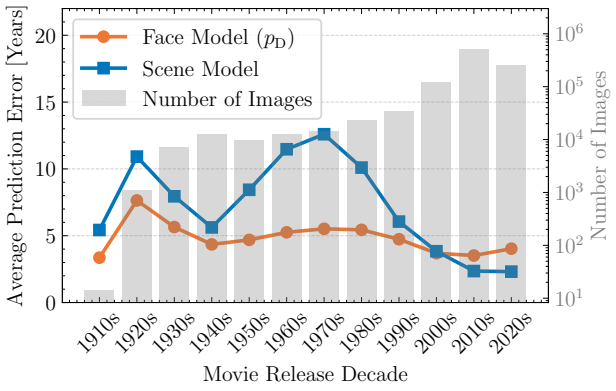


Figure 2. **Error by year analysis.** MAE ↓ of the *Face* (p_D) and *Scene* methods by the image capture year.

Benchmark	Pre-training Dataset			
	ImageNet	**IMDB-WIKI**	**CSFD-0.3M**	**CSFD-1.6M**
AgeDB	7.05 ± 0.29	6.34 ± 0.25	5.46 ± 0.24	$\mathbf{5.25 \pm 0.21}$
AFAD	3.19 ± 0.04	3.10 ± 0.03	3.08 ± 0.01	$\mathbf{3.04 \pm 0.03}$
MORPH	2.98 ± 0.05	2.88 ± 0.07	2.83 ± 0.07	$\mathbf{2.76 \pm 0.05}$
UTKFace	4.84 ± 0.08	4.64 ± 0.06	4.23 ± 0.02	$\mathbf{4.08 \pm 0.03}$
CLAP2016	5.87	4.89	3.53	**3.52**

Table 1. **Age estimation.** MAE $\downarrow$ ($\pm$ std) after pre-training a ResNet-101 model on different datasets. The CSFD-0.3M is size-matched to the cleaned version of IMDB-WIKI, EM-CNN [1].

The results are presented in Table 1. The model trained on CSFD-0.3M consistently and significantly outperforms this best-performing IMDB-WIKI-trained model across all evaluation benchmarks. Since the training set sizes are directly comparable, this performance gap provides strong empirical evidence that CSFD-1.6M contains a substantially higher-quality training signal for age estimation, validating its contribution beyond mere scale.

Scalability and Computational Cost

The proposed method is scalable. The initial *feature extraction*, involving the computation of face embeddings and age posteriors, represents a one-time cost and is highly parallelizable on GPU hardware. The dating logic, which includes the similarity search and marginalization over assignments, is performed on a CPU. In our experiments, the dating process on a single CPU core achieved an average inference speed of 0.5 seconds per image, a figure that includes the computation for both the *Full* and *Top-1* models. Since the dating of each image is an independent task, the overall throughput can be significantly increased for batch processing via standard multiprocessing. The primary computational bottleneck is the similarity search against the $\approx$46,000 identities in our database. For intended applications such as dating genealogical photographs, the set of known identities would be substantially smaller, leading to a considerable reduction in inference time.

Qualitative Analysis of Failure Cases

A key advantage of the proposed facial-age-based model is the interpretability of its failure modes. It is possible to differentiate whether a prediction error originates from an incorrect identity assignment or from an inaccurate age estimation. An analysis of the highest MAE failure cases, presented in Figure 6, reveals that errors predominantly stem from the face recognition component. Recognition failures are typically caused by challenging conditions such as masks, glasses, heavy makeup, extreme facial expressions, or difficult poses including profiles and long-distance shots.

In contrast, the failure cases of the *Scene* model are not readily interpretable due to its black-box nature. However the largest prediction errors correlate strongly with the temporal distribution of the training data. As visualized in Figure 2, the model performs best on years that are most represented in the dataset. Further visualizations of the error characteristics for both models, stratified by the number of known faces, are provided in Figure 1.

Ethical Considerations

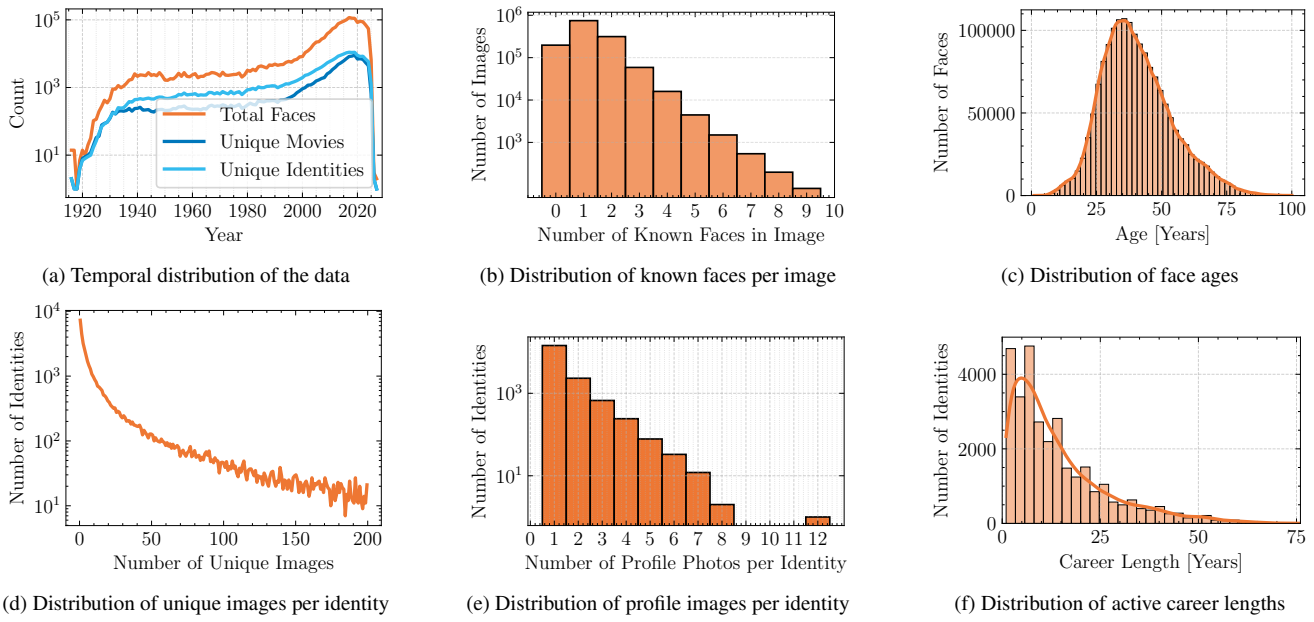
We release our contributions to foster research in temporal media analysis and encourage their responsible application, however, we recognize the associated ethical duties. The use of public data scraped without explicit consent, the demographic bias from the Czecho-Slovak Movie Database, and the potential for dual-use are significant concerns. To mitigate these while supporting reproducible research, we don't distribute the original images, but only the annotations, pre-computed features and URLs.

References

[1] Vojtech Franc and Jan Cech. Learning cnns from weakly annotated facial images. *Image and Vision Computing*, 2018. 1, 2

[2] Yiming Lin, Jie Shen, Yujiang Wang, and Maja Pantic. Fp-age: Leveraging face parsing attention for facial age estimation in the wild. *IEEE Transactions on Image Processing*, 34:4767–4777, 2025. 1

[3] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, 126(2-4):144–157, 2018. 1



(a) Temporal distribution of the data

(b) Distribution of known faces per image

(c) Distribution of face ages

(d) Distribution of unique images per identity

(e) Distribution of profile images per identity

(f) Distribution of active career lengths

Figure 3. **Overview of the CSFD-1.6M dataset statistics.**

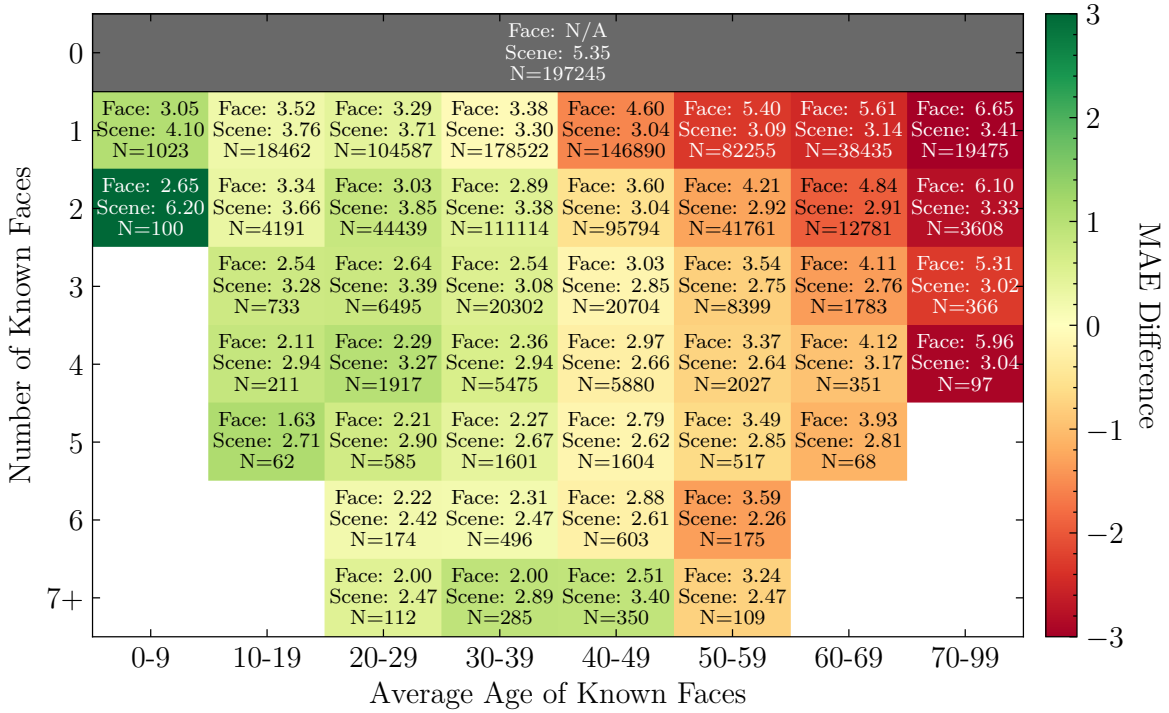


Figure 4. **Stratified performance.** MAE ↓ on CSFD-1.6M of the *Face* (p_D) and *Scene* methods, where N denotes the sample-size. The results reflect that age-estimation is more precise for younger subjects.

Figure 5. **Examples of Images With 0 Matches (Known Faces).**

(a) Failure cases caused by incorrect identification due to masks, heavy makeup or extreme expressions.

(b) Failure cases caused by incorrect identification due to actors being viewed from the side.

Figure 6. **Examples of Dating Failure Cases.**

Figure 7. **Examples of Face Matching Results.**

Figure 8. **Examples of Face Matching Results.**

Figure 9. **Examples of Face Matching Results.**