Trustworthy Biometrics: Adaptive Uncertainty Quantification in Age Estimation

Sandro Khizanishvili Matricola: 2175979
Emre Yesil Matricola: 2186113
Juan Iñaki Larrea Matricola: 2241722
Mohammadreza Heibati Matricola: 2166497

Advanced Machine Learning
Sapienza University of Rome

December 2025

Abstract

Standard deep learning approaches for age estimation typically operate as point estimators, outputting a single scalar value without quantifying confidence. In applications such as automated age verification, such "black box" predictions create significant liability. This project proposes a robust framework for Trustworthy Biometrics by replacing point estimates with calibrated prediction intervals. We leverage a Swin Transformer V2 backbone modified for Quantile Regression to capture global facial geometry and apply Conformalized Quantile Regression (CQR) to guarantee a strict 95% marginal coverage rate. Our experiments on the UTKFace dataset demonstrate that while baseline models like ResNet-50 achieve safety with an average interval width of 27.2 years, our Swin V2 approach achieves the same 95% safety guarantee with higher efficiency (24.8 year width), demonstrating superior adaptive uncertainty handling.

1 Introduction

Age estimation from facial images is a classic computer vision task with critical applications in access control and automated age verification. However, standard systems face a significant "Issue": they provide point estimates (e.g., Age: 25) that fail to account for the model's inherent uncertainty or the ambiguity of visual age versus biological age. A single-number prediction offers no measure of reliability, which is problematic in legal or security contexts where a wrong estimate constitutes a failure. The goal of this project is to develop a robust framework for Uncertainty Quantification in age estimation. Instead of point estimates, our system outputs calibrated prediction intervals (e.g., [17.5, 20.2]). We specifically target a 95% marginal coverage guarantee using Conformal Prediction, ensuring that the true age falls within the predicted range 95% of the time. By moving from "best-guess" regression to statistically valid intervals, we prioritize reliability and the ability to capture outliers in biometrics.

2 Related Work

Traditional approaches rely on Convolutional Neural Networks (CNNs) [1] like ResNet to regress age as a scalar value [4]. More recently, Transformers such as the Swin Transformer [2] have shown promise by capturing long-range dependencies in images, which is crucial for relating distributed aging features (e.g., forehead wrinkles and jawline sagging). To quantify uncertainty, Romano et al. proposed Conformalized Quantile Regression [3]. Unlike standard conformal prediction which produces fixed-width intervals, CQR combines the adaptivity of quantile regression with the rigorous frequentist guarantees of conformal prediction. This allows intervals to expand or contract based on the difficulty of the specific input image.

3 Methodology

3.1 Model Architecture

We compare two architectures to see if understanding the global context helps the model better quantify its uncertainty, rather than just looking at local details:

- **Baseline (ResNet-50) [1]:** We used this standard CNN as our baseline to compare against. ResNet is excellent at catching small details like skin texture and edges (local features). However, since it looks at the image piece-by-piece, it might sometimes miss the overall shape and structure of the face.

- **Proposed (Swin Transformer V2) [2]:** We chose the *Swin V2 Tiny* model because it is designed to capture long-range dependencies. Its **Shifted Window** mechanism enables the model to attend to different parts of the face simultaneously, capturing global geometry more effectively than CNNs. We specifically picked the V2 version because its **scaled cosine attention** makes the training much more stable for our regression task.

3.2 Quantile Regression

We modified the output layer of both models to predict three values simultaneously instead of a single mean: the lower bound ($q_{0.025}$), the median ($q_{0.5}$), and the upper bound ($q_{0.975}$). We trained the models using the **Pinball Loss**, which is defined as:

$$L_\tau(y, \hat{y}) = \begin{cases} \tau(y - \hat{y}) & \text{if } y \geq \hat{y} \\ (\tau - 1)(y - \hat{y}) & \text{if } y < \hat{y} \end{cases} \quad (1)$$

By minimizing this loss, the model learns to construct a 95% confidence interval along with the age prediction.

3.3 Calibration via CQR

Since the output of Quantile Regression does not guarantee valid coverage, we applied **Conformalized Quantile Regression (CQR)** [3]. We followed these steps:

1. **Splitting:** We split the data into a training set and a separate calibration set.
2. **Scoring:** On the calibration set, we computed conformity scores E_i to measure the error:
$$E_i = \max\left(\hat{q}_{low}(x_i) - y_i, \quad y_i - \hat{q}_{high}(x_i)\right) \quad (2)$$
3. **Correction:** We calculated a correction factor Q, which is the $(1-\alpha)$-th quantile of the scores E_i. Here, we set $\alpha = 0.05$ to ensure a 95% target coverage.
4. **Adjustment:** Finally, we adjust the prediction interval for any new input x using Q:
$$C(x) = [\hat{q}_{low}(x) - Q, \quad \hat{q}_{high}(x) + Q] \quad (3)$$

4 Dataset and Setup

We used the **UTKFace** dataset, which contains over 20,000 aligned face images ranging from 0 to 116 years old. To ensure a fair evaluation, we randomly split the dataset into three parts:

- **Training (60%):** Used to train the model parameters.
- **Calibration (20%):** Used only to compute the CQR scores and the factor Q.
- **Test (20%):** Used strictly for the final evaluation of the model's performance.

As a **preprocessing** step, we resized all images to 256×256 pixels. We also applied **Random Horizontal Flip** with a probability of 0.5 during training.

Additional Experiment (MORPH-2): We also tried adding the MORPH-2 dataset to our training data. However, MORPH-2 consists mainly of mugshot-style images with controlled lighting, which differs significantly from the "in-the-wild" images in UTKFace. Since this domain gap did not improve performance on the UTKFace test set, we decided not to include MORPH-2 in the final model reported here.

5 Experimental Results

We evaluated the efficacy of our proposed framework by comparing a baseline **ResNet-50** model against our **Swin Transformer V2** approach on the UTKFace dataset. The primary objective was to assess the ability of each architecture to produce reliable uncertainty estimates while maintaining high efficiency, defined as the narrowness of the prediction intervals under a strict 95% coverage constraint.

5.1 Baseline Performance: ResNet-50

The **ResNet-50** baseline demonstrated the limitations of standard CNN architectures in capturing the nuanced uncertainty of age estimation. Prior to calibration, the raw quantile regression model achieved a coverage rate of 93.0% with an average interval width of 25.0 years. Following **Conformalized Quantile Regression (CQR)** calibration, the model successfully met the safety requirement, reaching a valid marginal coverage of 95.5%. However, this safety came at the cost of efficiency; the average interval width expanded to 27.2 years. The distribution of uncertainty widths revealed a heavy tail, indicating that the model frequently resorted to excessively wide intervals to mitigate prediction errors on difficult samples, lacking the necessary adaptivity for precise biometrics.

5.2 Proposed Method: Swin Transformer V2

In contrast, the **Swin Transformer V2** demonstrated superior performance in both precision and adaptivity. Before calibration, the raw model exhibited an average interval width of 17.8 years with a coverage rate of 86.0%. This lower initial coverage suggests the uncalibrated model was highly precise but overconfident. The **CQR** calibration effectively corrected this bias, adjusting the final coverage to 94.7%, which is statistically consistent with the 95% target. Crucially, the final calibrated average width settled at 24.8 years, representing a reduction of 2.4 years compared to the baseline while maintaining the same safety guarantee. Additionally, the Swin model achieved a Mean Absolute Error (MAE) of 4.8 years, surpassing the baseline's 5.2 years.

5.3 Analysis of Heteroscedasticity

A key finding of our study is the model's ability to adapt its uncertainty estimates based on the inherent difficulty of the input, a property known as heteroscedasticity. As illustrated in Figure 2, the **Swin Transformer V2** exhibits distinct performance characteristics across age decades. For the 0–10 age group, the model produced extremely narrow intervals, reflecting high confidence due to the distinct visual features of childhood development. Conversely, for the 80–90 and 90+ age groups, interval widths expanded significantly to over 35 years. This behavior validates that the model effectively captured the aleatoric

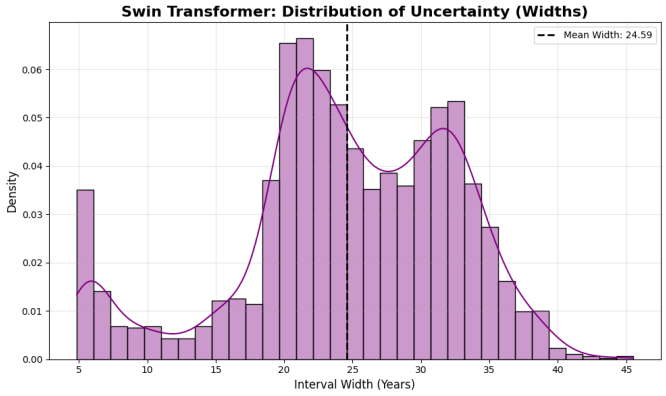

Figure 1: **Distribution of interval widths for the Swin Transformer**. The distribution is shifted towards narrower intervals compared to the baseline, with a mean width of 24.8 years.

uncertainty inherent in the aging process, where visual signs are distinct in youth but highly variable in seniority. While coverage remained robust across most demographics, a degradation was observed in the 90+ age group (dropping below 70%), attributed to the data scarcity of extreme ages in the dataset.

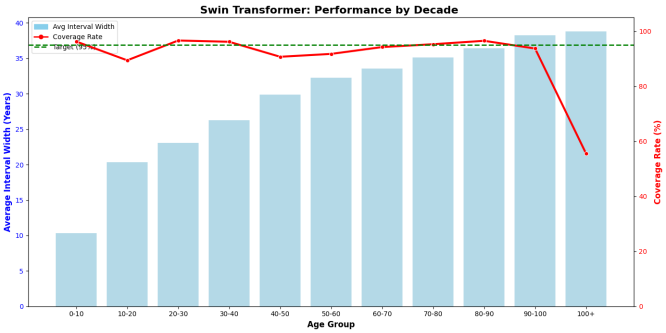

Figure 2: **Performance by age decade**. Blue bars represent average interval width; the red line represents coverage. The model demonstrates high precision for children (0-10) and appropriately increased uncertainty for seniors (80+).

6 Conclusion and Future Work

In this work, we have successfully developed and validated a robust framework for trustworthy age estimation. By shifting the paradigm from point estimation to uncertainty quantification, we addressed the critical "Black Box Liability" inherent in standard regression models. Our proposed method, which integrates a Swin Transformer V2 backbone with Conformalized Quantile Regression (CQR), achieved the target 95% marginal coverage rate with a significantly higher efficiency than the baseline ResNet-50 model. The reduction in average interval width by 2.4 years demonstrates that capturing global facial geometry is essential for tighter, more informative uncertainty bounds. Furthermore, the analysis of heteroscedasticity confirmed that our model intelligently adapts to the biological ambiguity of aging, providing high-precision estimates for children while appropriately expanding safety margins for seniors.

Despite these successes, our analysis revealed a disparity in performance for the 90+ age demographic, where coverage dropped due to extreme data scarcity. Future work will focus on rectifying this through **Fairness Correction**. Specifically, we plan to implement **Group-Conditional Conformal Prediction** to enforce validity guarantees within specific demographic subgroups (e.g., ethnicity and age brackets) rather than just globally.

References

[1] K. He et al., "Deep Residual Learning for Image Recognition," CVPR 2016.

[2] Z. Liu et al., "Swin Transformer V2: Scaling Up Capacity and Resolution," CVPR 2022.

[3] Y. Romano, E. Patterson, and E. Candès, "Conformalized Quantile Regression," NeurIPS 2019.

[4] G. Antipov et al., "Apparent Age Estimation from Face Images Combining General and Children-Specialized Deep Learning Models," CVPR Workshops 2016.

A Appendix: Baseline Results

To complement the experimental analysis, we provide the visual performance metrics for the ResNet-50 baseline model. The heavier tail in the uncertainty distribution and wider intervals across older age groups highlight the limitations of the CNN architecture compared to the Swin Transformer.

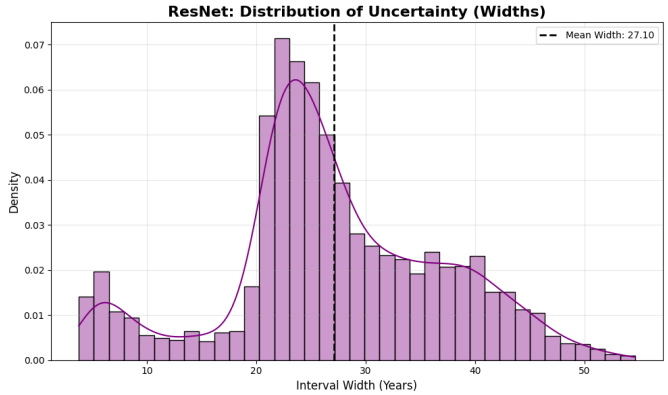

Figure 3: Distribution of prediction interval widths for the ResNet-50 baseline. Note the heavier tail and larger mean width compared to the Swin Transformer.

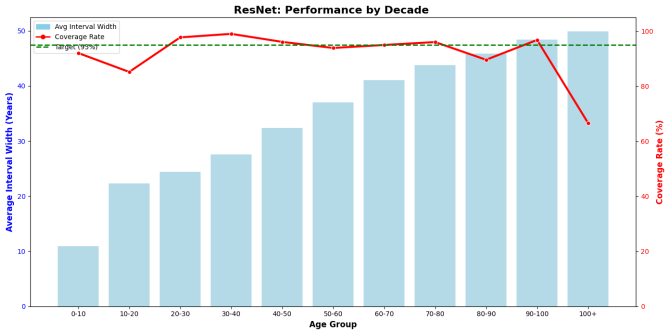

Figure 4: Performance by age decade for ResNet-50. While coverage is achieved, the interval widths for older groups are significantly wider than the proposed method, indicating lower efficiency.