IFSCC 2025 full paper (2025-866)

Deep Learning Analysis of Perceived Facial Aging and Influential Features Across Evaluator Groups

Fudi Wang [1,2], Siying Fu [2], Baolin Chen [2], Zhiyang LI [2], Eagle Lee [2], Sijia Wang [1,*]

[1] Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences, Shanghai, China; [2] EVELAB INSIGHT (SINGAPORE) PTE. LTD., Singapore, Singapore

1. Introduction

Facial aging has long been recognized as an important biomarker of aging and overall health [1-6]. It is closely linked to numerous age-related diseases, including cardiovascular conditions, metabolic disorders, and neurodegenerative syndromes[7-12]. As a visible and socially relevant indicator, facial aging has received sustained attention from both scientific communities and the public.

Over the years, various approaches have been proposed to quantitatively assess facial aging. Some methods focus on localized features—such as wrinkles, pigmentation, or skin texture—while others rely on perceptual experiments, where evaluators estimate overall facial age based on visual appearance[13-16]. Among these, perceived age assessments have emerged as a widely accepted and effective method for quantifying global facial aging, due to their strong correlation with biological age and health outcomes[1, 17-19].

However, a major limitation of existing perceptual studies is the lack of consideration for individual-level variability among assessors. Demographic attributes such as the evaluator's own age and gender may introduce systematic biases into perceived age ratings, potentially confounding the interpretation of facial aging estimates.

In this study, we address this gap by leveraging a crowdsourced evaluation framework to systematically examine how individual characteristics of assessors influence perceived facial aging outcomes. By modeling these effects, we aim to simulate the perceptual process more objectively and remove potential bias sources.

Subsequently, we develop deep learning and statistical models to replicate human perceptual judgments of age in a more standardized and reproducible manner. Finally, we identify and quantify the facial features most strongly associated with perceived aging, providing insights into the visual cues that drive human judgments of facial age. The overall workflow of this study, from data collection to bias modeling, deep learning simulation, and feature analysis, is summarized in Figure 1.

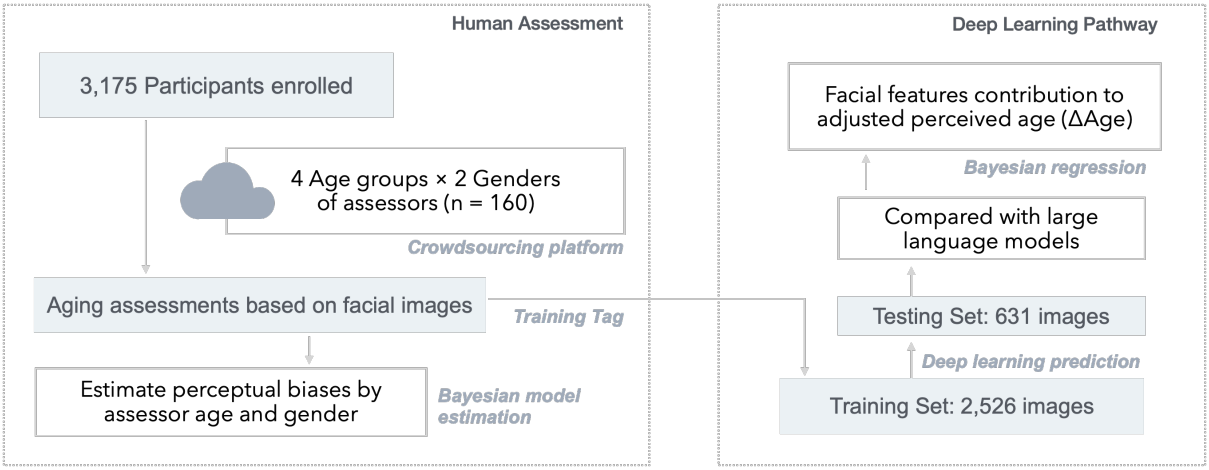

Figure 1. Study Design and Flowchart. The study begins with a large-scale crowdsourced experiment, where assessors rated perceived age for 3,157 facial images. Bayesian modeling was then applied to quantify and remove demographic biases introduced by assessor age and gender. A deep learning model (STDC2-FLD-HR) was trained and compared against large language models (LLMs) to simulate perceptual judgments. The adjusted perceived age (ΔAge) was derived by subtracting chronological age from bias-corrected estimates. Finally, Bayesian regression was used to identify which facial features contributed most significantly to perceived aging.

2. Materials and Methods

2.1 Study Population

The Jidong cohort (JD) is a community-based, long-term observational cohort study to evaluate health related risk factors[20]. The baseline data were collected from 2013 to 2014 in the Staff Hospital, Jidong Oilfield Branch, China. Approval was obtained from the Ethics Committee of Kailuan General Hospital of Tangshan City and the Medical Ethics Committee, Staff Hospital, Jidong Oilfield Branch, China National Petroleum Corporation in July, 2013 (approval No. 2013 YILUNZI1). In this study, 3,157 individuals (1,712 men and 1,445 women, aged 18-87) have been enrolled after excluding individuals who were unable or unwilling to participate. Written informed consent was obtained from all participants. The facial images were collected in the Staff Hospital for further analysis.

2.2 Crowdsourced Perceived Age Assessment

In this study, we recruited 152 assessors through a crowdsourcing platform, evenly distributed across four age groups (21–30, 31–40, 41–50, and 51–60 years) with equal representation of male and female participants (19 men and 19 women per group). Each assessor was tasked with estimating the perceived age via facial images. In addition, assessors provided localized aging assessments based on cropped facial regions of facial images. The evaluated features included under-eye bags, glabellar lines, nasolabial folds, tear troughs, forehead wrinkles, frown lines, crow's feet and marionette lines[14].

2.3 Deep Learning Model Training

We employed deep learning models to simulate human age estimation processes[21]. The performance of three different network architectures was evaluated: STDC2-FLD-HR and STDC2-FLD (both using a CNN-based backbone), Pvt_v2-FLD (Transformer-based backbone), and Vmamba-FLD (VSSM-based backbone)[22-24]. Model training and evaluation were conducted on an NVIDIA 3090 GPU. To enhance the segmentation accuracy of wrinkle regions, STDC2-FLD-HR incorporated feature maps from Stage 1 into the decoder phase, whereas the other models only integrated features from Stage 2 (Figure 2)[25]. For the decoder, we adopted the flexible and lightweight FLD module from PP-LiteSeg to ensure efficient and accurate feature decoding.

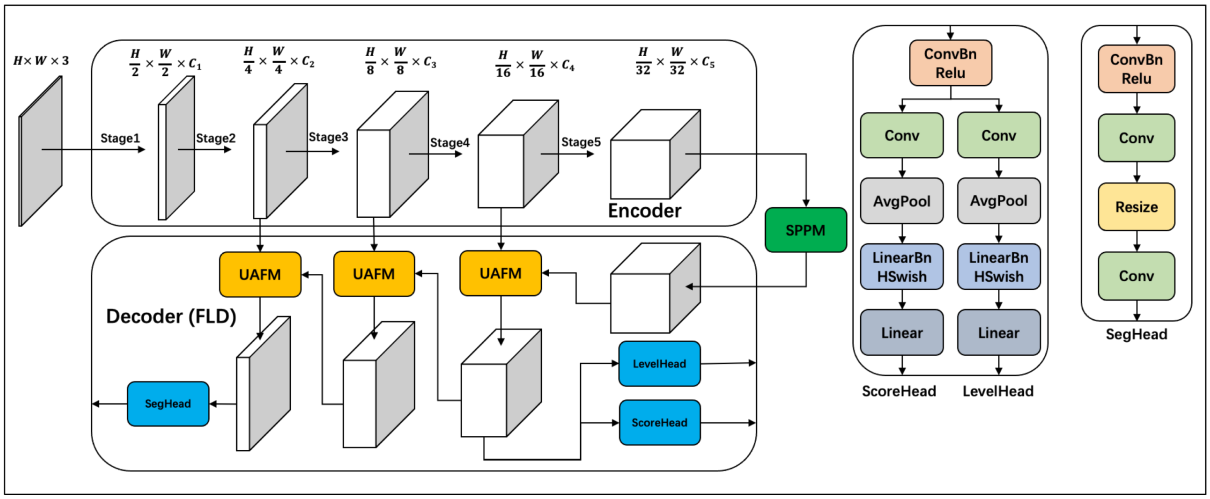

Figure 2. Schematic illustration of the deep learning architecture used for perceived age prediction. The model is based on a Short-Term Dense Concatenate (STDC2) backbone for feature extraction, followed by a lightweight FLD decoder adapted from PP-LiteSeg. High-resolution feature maps from early encoder stages are integrated during decoding (in STDC2-FLD-HR) to enhance fine-grained wrinkle and skin texture representation.

2.4 Deep Learning Visualization

To visualize the information captured by the deep learning models, we first extracted downsampled feature maps and performed Principal Component Analysis (PCA) to reduce their dimensionality. The top three principal components were then mapped to the RGB channels to generate interpretable feature visualizations. Additionally, the low-resolution feature maps were upsampled to higher resolutions to provide clearer and more detailed visual representations.

2.5 Bayesian Rating Model Construction

To account for subjective biases in perceived age ratings, we constructed a hierarchical Bayesian model. The model assumes that evaluators from different age groups and genders may systematically over- or under-estimate the perceived age of facial images. These latent biases are captured as random effects and inferred through Bayesian estimation. Specifically, the observed score y_{ij} given by evaluator j to subject i is decomposed into a global mean, a subject-specific fixed effect, and a rater-specific bias term that varies by age and gender group. Random noise is added to model unexplained variance.

The rating bias β_j is modeled as the sum of random effects from the evaluator's age group and gender:

$$\beta_j = \gamma_{age}[k_j] + \gamma_{gender}[g_j]$$

where $\gamma_{age}[k_j] \sim \mathcal{N}(0,0.2)$ and $\gamma_{gender}[g_j] \sim \mathcal{N}(0,0.2)$ represent the random effects associated with the evaluator's age group and gender, respectively.

3. Results

3.1 Sample Characteristics

A total of 3,157 facial images were included in this study, representing 1,712 males and 1,445 females aged between 18 and 87 years (mean age = 48.12). These images were used as stimuli in the perceived age evaluation experiments. Initially, 152 assessors were recruited to perform the age assessments. The distribution of perceived age ratings was examined across all age and gender groups and found to approximately follow a normal distribution (P>0.05). Table 1 shows the standard deviations of perceived age ratings within each subgroup, compared to the combined results from the full evaluator samples. The results indicate a high degree of consistency across subgroups, with coefficient of variation (CV) values ranging between 0.11 and 0.13. mean absolute error (MAE) values remained relatively stable, generally between 7.16 and 7.50 years, suggesting that assessors provided similarly accurate age estimations across each subgroup.

Table 1. Summary of perceived age rating variability stratified by assessor groups

Gender [1]	Age Group [2]	Mean (years)	SD (years)	CV (%)	MAE
Male	**20-30**	48.67	9.84	0.11	7.27
	30-40	48.66	9.84	0.12	7.40
	40-50	48.80	9.85	0.12	7.44
	50-60	49.13	10.27	0.11	7.50
Female	**20-30**	47.92	10.11	0.13	7.16
	30-40	48.66	9.91	0.12	7.26
	40-50	48.30	10.26	0.12	7.25
	50-60	49.00	10.17	0.12	7.50
Total	**20-30**	48.27	9.92	0.12	7.21
	30-40	48.64	9.83	0.12	7.33
	40-50	48.53	9.99	0.12	7.35
	50-60	49.05	10.17	0.12	7.50

[1] The gender of assessors; [2] The age group of assessors. The table reports the mean perceived age, standard deviation (SD), coefficient of variation (CV, %) and mean absolute error (MAE) for each sub-group. Lower CV values indicate greater consistency among assessors within the corresponding group.

Table 2 summarizes the inter-assessor consistency in perceived age ratings using intraclass correlation coefficients (ICC) across different demographic groups. The total ICC across all assessors was 0.72, indicating substantial agreement. Subgroup analyses revealed similar reliability across age groups and genders, with slightly higher ICC observed in the 50–60 age

group (ICC = 0.74, 95% CI: [0.73, 0.75]). These results suggest that perceived age assessments were stable across demographic subgroups.

Table 2. Intraclass Correlation Coefficients (ICC) for perceived age ratings

Group	ICC[1]	F-value	p-value	95% CI
Total	0.72	399.80	<0.001	[0.71, 0.73]
20–30	0.72	101.01	<0.001	[0.71, 0.73]
30–40	0.72	99.51	<0.001	[0.71, 0.73]
40–50	0.72	103.35	<0.001	[0.71, 0.73]
50–60	0.74	113.29	<0.001	[0.73, 0.75]
Male	0.72	202.55	<0.001	[0.71, 0.73]
Female	0.72	202.58	<0.001	[0.71, 0.73]

[1] Intraclass correlation coefficients (ICC) reflecting inter-assessor reliability across demographic subgroups. ICC values were computed using a two-way random effects model for absolute agreement. Higher ICC indicates stronger consistency among assessors.

3.2 Perceptual Biases by Assessor Demographics

To examine potential subjective biases in perceived age assessments, we applied a hierarchical Bayesian model to estimate group-level deviations based on assessor age and gender. As shown in Table 3, the estimated effects (γ) for different assessor age groups demonstrated a clear trend: older assessors tended to give higher perceived age ratings. In particular, the 50–60 age group showed a positive bias (posterior mean = 0.41), while younger groups had slightly negative or near-zero estimates from -0.01 to -0.37.

Regarding gender, female assessors on average provided slightly lower perceived age ratings, with a posterior bias estimate of -0.34 compared to male assessors (0.003). The directional trends suggest potential demographic influences on perceptual judgments.

Table 3. Posterior summaries of age and gender group-level bias effects

	Mean	SD	$\text{Mean}_{\text{MCSE}}$[1]	SD_{MCSE}
$\boldsymbol{\gamma_{\text{age}}[20-30]}$	-0.37	1.03	0.09	0.07
$\boldsymbol{\gamma_{\text{age}}[30-40]}$	-0.01	1.03	0.09	0.07
$\boldsymbol{\gamma_{\text{age}}[40-50]}$	-0.12	1.03	0.09	0.07
$\boldsymbol{\gamma_{\text{age}}[50-60]}$	0.41	1.03	0.09	0.07
$\boldsymbol{\gamma_{\text{gender}}[\text{Male}]}$	0.003	1.45	0.10	0.07
$\boldsymbol{\gamma_{\text{gender}}[\text{Female}]}$	-0.34	1.45	0.10	0.07

[1] MCSE: Monte Carlo standard errors. Results are derived from a hierarchical Bayesian model.

3.3 Deep Learning Model Performance in Simulating Human Perception

To assess the ability of deep learning models to simulate human perceptual judgment of facial age, we evaluated the performance of four model architectures using the Pearson correlation coefficient (PCC) between predicted and human-assessed perceived age, stratified by assessor age and gender groups. As shown in Table 4, all models achieved statistically significant correlations with human ratings ($p < 0.001$ across all subgroups), confirming their predictive validity.

Among the models tested, STDC2-FLD-HR consistently achieved the highest performance, with PCC values ranging from 0.94 to 0.95 across all gender and age combinations, demonstrating superior ability to approximate human perception. This was followed by STDC2-FLD (PCC = 0.91–0.94), Pvt_v2-FLD (PCC = 0.86–0.88), and Vmamba-FLD (PCC = 0.83–0.87). STDC2-FLD-HR was used in further analysis.

Table 4. Pearson correlation coefficients (PCC) between model-predicted and human-perceived ages across assessor gender and age groups.

Gender	**Female**				**Male**			
Age Group	**20-30**	**30-40**	**40-50**	**50-60**	**20-30**	**30-40**	**40-50**	**50-60**
STDC2-FLD-HR	0.95[1]	0.95	0.94	0.94	0.94	0.94	0.94	0.94
STDC2-FLD	0.94	0.94	0.94	0.94	0.94	0.93	0.93	0.91
Pvt_v2-FLD	0.88	0.87	0.87	0.86	0.86	0.86	0.86	0.86
Vmamba-FLD	0.87	0.87	0.87	0.87	0.86	0.86	0.86	0.83

[1] Values in the table represent the Pearson correlation coefficients (PCC) between predicted ages from deep learning models and perceived ages assessed by human raters. Higher PCC values indicate stronger alignment between model predictions and human perception. All reported correlations are statistically significant ($p < 0.001$).

To further evaluate the capability of deep learning models in simulating human-like age perception, we compared the performance of a visual model (STDC2-FLD-HR) with two large language models (Qwen and LLaMA) prompted to perform perceived age estimation from facial images.

Performance was evaluated based on mean absolute error (MAE), Pearson correlation coefficient (PCC), and linear regression R^2 between predicted and true chronological age. The STDC2-FLD-HR model demonstrated superior performance across all metrics, achieving the lowest MAE (5.56), highest PCC (0.87), and R^2 (0.76). In contrast, Qwen achieved moderate performance, with an MAE of 9.895, PCC of 0.779, and R^2 of 0.608, suggesting that while it captured some meaningful patterns, its estimations were less precise and less aligned with

true age distributions. LLaMA showed significantly poorer performance, with an MAE of 11.045, PCC of 0.392, and R^2 of 0.154, reflecting limited capacity to extract and interpret visual age-related cues.

3.4 Drivers of Perceived Aging After Bias Adjustment

To isolate intrinsic facial aging effects from potential demographic biases, we first adjusted the perceived age scores predicted by the deep learning model using a Bayesian correction model accounting for assessor age and gender. The resulting metric, referred to as ΔAge (delta age), represents the deviation between perceived age and actual chronological age.

We then examined the relationship between ΔAge and a set of facial aging features using a Bayesian regression model. As shown in Table 5, the most influential features contributing to higher ΔAge values were nasolabial folds (β = 0.09), eye bags (β = 0.08), and pigmentation spots (β = 0.08), suggesting these features were most strongly associated with an older perceived appearance. These findings highlight specific regional markers that drive perceived facial aging, independent of demographic rater effects.

Table 5. Facial feature effects on perceived age deviation (ΔAge)

Aging Features	β[1]	Aging Features	β
Nasolabial_fold	0.086	Wrinkle around eyes	0.054
Eye bags	0.083	Tear_troughs	0.035
Pigmentation spots	0.079	Wrinkle forehead	0.031
Crows' feet	0.064	Eye sagging	0.007
Marionette lines	0.060	Frown lines	0.004

[1] Bayesian regression coefficients (β) for facial aging features contributing to ΔAge (Perceived age – chronological age). ΔAge represents perceived age deviation from chronological age after demographic bias correction. Higher β values indicate stronger influence on perceived aging.

4. Discussion

This study sought to better understand the process of human facial age perception and improve its objectivity by modeling demographic influences and leveraging deep learning technologies. Building on the notion that facial aging is a socially and biologically relevant biomarker, we first confirmed that perceived age assessments can be systematically influenced by assessor characteristics—particularly age and gender. Through a large-scale crowdsourced experiment and hierarchical Bayesian modeling, we identified measurable perceptual biases associated with these demographic variables.

We developed a visual deep learning model (STDC2-FLD-HR) that demonstrated strong alignment with human ratings while maintaining high consistency across gender and age

groups. This model significantly outperformed large language models (LLMs) when tasked with estimating perceived age from facial images. Moreover, by removing assessor-level bias and analyzing the remaining discrepancy between predicted and actual chronological age (ΔAge), we identified several facial features—such as nasolabial folds, eye bags, and pigmentation—that strongly contribute to the faical aging. These features may serve as reliable visual markers for future studies of skin aging.

Despite these promising findings, our study has several limitations. First, the dataset and assessors were limited to Han Chinese populations, without accounting for ethnic diversity. Given that facial aging patterns and perceptual tendencies vary significantly across populations, future work will aim to expand the dataset to include participants of different ethnic backgrounds. Second, our analysis of language model performance was restricted to two models (Qwen and LLaMA) available before January 2025. As the field of multimodal AI rapidly evolves, subsequent versions of LLMs may offer improved performance and should be evaluated in future research.

In addition, while our current sample includes over 3,000 individuals, further expansion in both sample size and demographic diversity will enhance the robustness and generalizability of our findings. Incorporating cross-cultural evaluations and larger datasets will be critical for developing universally reliable tools for facial age estimation.

5. Conclusion

This study demonstrated that facial age perception is influenced by both skin-related features and assessor demographics such as age and gender. By correcting for these biases through Bayesian modeling and applying a deep learning framework, we achieved accurate and objective facial age estimation. Our model outperformed language-based models and revealed that features like nasolabial folds, pigmentation spots, and crow’s feet play key roles in perceived aging. Future work will expand to more diverse populations and AI models to enhance generalizability.

Reference

1. Liu, F., et al., *The MC1R Gene and Youthful Looks.* Curr Biol, 2016. **26**(9): p. 1213-20.
2. Bonfante, B., et al., *A GWAS in Latin Americans identifies novel face shape loci, implicating VPS13B and a Denisovan introgressed region in facial variation.* Sci Adv, 2021. **7**(6).
3. Xiong, Y., et al., *Prevalence and associated factors of metabolic syndrome in Chinese middle-aged and elderly population: a national cross-sectional study.* Aging Male, 2021. **24**(1): p. 148-159.
4. Gurovich, Y., et al., *Identifying facial phenotypes of genetic disorders using deep learning.* Nat Med, 2019. **25**(1): p. 60-64.
5. Xiong, Z., et al., *Combining genome-wide association studies highlight novel loci involved in human facial variation.* Nat Commun, 2022. **13**(1): p. 7832.
6. Yu, Z., et al., *Thermal facial image analyses reveal quantitative hallmarks of aging and metabolic diseases.* Cell Metab, 2024. **36**(7): p. 1482-1493 e7.
7. Peng, Q., et al., *3D facial imaging: a novel approach for metabolic abnormalities risk profiling.* Sci China Life Sci, 2025.
8. Cox-Brinkman, J., et al., *Three-dimensional face shape in Fabry disease.* Eur J Hum Genet, 2007. **15**(5): p. 535-42.
9. Kim, M.K., et al., *Associations of Variability in Blood Pressure, Glucose and Cholesterol Concentrations, and Body Mass Index With Mortality and Cardiovascular Outcomes in the General Population.* Circulation, 2018. **138**(23): p. 2627-2637.
10. Meng, T., et al., *Identifying Facial Features and Predicting Patients of Acromegaly Using Three-Dimensional Imaging Techniques and Machine Learning.* Front Endocrinol (Lausanne), 2020. **11**: p. 492.
11. Lin, S., et al., *Feasibility of using deep learning to detect coronary artery disease based on facial photo.* Eur Heart J, 2020. **41**(46): p. 4400-4411.
12. Kong, X., et al., *Facial recognition for disease diagnosis using a deep learning convolutional neural network: a systematic review and meta-analysis.* Postgrad Med J, 2024. **100**(1189): p. 796-810.
13. Liu, Y., et al., *Genome-wide scan identified genetic variants associated with skin aging in a Chinese female population.* J Dermatol Sci, 2019. **96**(1): p. 42-49.
14. Vierkotter, A., et al., *The SCINEXA: a novel, validated score to simultaneously assess and differentiate between intrinsic and extrinsic skin ageing.* J Dermatol Sci, 2009. **53**(3): p. 207-11.
15. Saffari, P.S., et al., *Facial Aging in Thyroid Eye Disease: Quantification by Artificial Intelligence.* J Craniofac Surg, 2025.
16. Estler, A., et al., *Quantification of Facial Fat Compartment Variations: A Three-Dimensional Morphometric Analysis of the Cheek.* Plast Reconstr Surg, 2023. **152**(4): p. 617e-627e.
17. Sun, N., et al., *Self-perception of aging and perceived medical discrimination.* J Am Geriatr Soc, 2023. **71**(10): p. 3049-3058.
18. Nkengne, A., et al., *Influence of facial skin attributes on the perceived age of Caucasian women.* J Eur Acad Dermatol Venereol, 2008. **22**(8): p. 982-91.
19. Flament, F., et al., *Changes in facial signs due to age and their respective weights on the perception of age and skin plumpness among differently aged Korean women.* Skin Res Technol, 2021. **27**(4): p. 526-536.

20. Wang, F., et al., *A Genome-Wide Scan on Individual Typology Angle Found Variants at SLC24A2 Associated with Skin Color Variation in Chinese Populations.* J Invest Dermatol, 2022. **142**(4): p. 1223-1227 e14.
21. Fan, M.a.L., et al., *Rethinking BiSeNet For Real-time Semantic Segmentation*, in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021. p. 9711-9720.
22. Wang, W., et al., *Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions*, in *IEEE/CVF international conference on computer vision*. 2021: Montreal, QC, Canada. p. 568-578.
23. Wang, W., E. Xie, and X. Li, *PVT v2: Improved baselines with pyramid vision transformer.* Computational Visual Media, 2022. **8:3**: p. 415-424.
24. Bian, J., M. Feng, and W. Dong, *Locally Aware Visual State Space for Small Defect Segmentation in Complex Component Images.* IEEE Transactions on Industrial Informatics, 2025: p. 1-12.
25. Shreve, M., R. Bala, and W. Wu, *Region-wise Modeling of Facial Skin Age using Deep CNNs*. 2019 16th International Conference on Machine Vision Applications (MVA). 2019.