CEA Comité Español de Automática

XLVI JORNADAS AUTOMÁTICA CARTAGENA 2025

UNIVERSIDADE DA CORUÑA

Jornadas de Automática

Comparative Analysis of CNNs and Vision Transformers for Age Estimation

Tanveer, Waqar[a,*], Fernández-Robles, Laura[b], Fidalgo, Eduardo[a], González-Castro, Víctor[a], Alegre, Enrique[a], Mirjalili, Milad[a]

[a] *Electrical, Systems and Automation Engineering, Universidad de León, Campus de Vegazana s/n, 24071, León, Spain.*
[b] *Department of Mechanical, Computer Science and Aerospace Engineering, Universidad de León, Campus de Vegazana s/n, 24071, León, Spain.*

To cite this article: Tanveer, Waqar, Fernández-Robles, Laura, Fidalgo, Eduardo, González-Castro, Víctor, Alegre, Enrique, Mirjalili, Milad. 2025. Comparative Analysis of CNNs and Vision Transformers for Age Estimation. Jornadas de Automática, 46. https://doi.org/10.17979/ja-cea.2025.46.12251

Resumen

Los transformadores de visión han adquirido recientemente una importancia significativa en las tareas de visión por ordenador debido a sus mecanismos de autoatención. Anteriormente, las CNN dominaban el campo de la visión por ordenador al lograr resultados notables en diversas aplicaciones como la clasificación de imágenes o el reconocimiento de objetos, entre otras. Sin embargo, con la llegada de los Transformadores de Visión, ha surgido una intensa competencia entre ambos. Este artículo presenta un análisis comparativo del rendimiento de las CNNs y los Transformadores de Visión para la tarea de estimación de la edad en los conjuntos de datos FG-NET y UTKFace. Realizamos la estimación de la edad utilizando seis modelos, incluidos tres modelos de CNN (VGG-16, ResNet-50, EfficientNet-B0) y tres modelos de transformadores de visión (ViT, CaiT, Swin). Nuestros resultados experimentales muestran que el transformador Swin superó tanto a la CNN como a los demás transformadores de visión, alcanzando un error medio absoluto (MAE) de 2,79 años con una puntuación acumulada (CS) del 83,90% en FG-NET y un MAE de 4,37 años con una CS del 68,73% en UTKFace.

Palabras clave: Visión por computadora, CNNs, Transformadores de Visión, VGG-16, ResNet-50, EfficientNet-B0, ViT, CaiT, Swin.

Comparative Analysis of CNNs and Vision Transformers for Age Estimation

Abstract

Vision Transformers have recently gained significant importance in computer vision tasks due to their self-attention mechanisms. Previously, CNNs dominated the computer vision field by achieving remarkable results in various applications such as image classification, object recognition, and more. However, with the arrival of Vision Transformers, an intense competition has emerged between the two. This paper presents a comparative analysis of the performance of CNNs and Vision Transformers for the task of age estimation on the FG-NET and UTKFace datasets. We performed age estimation using six models, including three CNN models (VGG-16, ResNet-50, EfficientNet-B0) and three Vision Transformer models (ViT, CaiT, Swin). Our experimental results show that the Swin Transformer outperformed both CNN and other Vision Transformers, achieving a mean absolute error (MAE) of 2.79 years with a cumulative score (CS) of 83.90% on FG-NET and an MAE of 4.37 years with a CS of 68.73% on UTKFace.

Keywords: Computer vision, CNNs, Vision Transformers, VGG-16, ResNet-50, EfficientNet-B0, ViT, CaiT, Swin.

1. Introduction

Convolutional Neural Networks (CNNs) have dominated the field of computer vision due to their extensive use in various applications, including face recognition (Song and Wang, 2024), age estimation (Hiba and Keller, 2023), image classification (Li et al., 2025), and many others. CNNs became widely adopted because of their ability to autonomously ex-

*Corresponding author: wtan@unileon.es

tract features and patterns from unprocessed images (Agbo-Ajala and Viriri, 2021). Typically, images contain localized patterns, referred to as feature elements, which are distributed systematically. Various filters within the convolutional layers are designed to detect a range of these elements, while pooling layers reduce dimensionality and enhance resilience to variations. However, this localized processing in CNNs can lead to a loss of spatial relationships, potentially affecting their effectiveness when handling larger and more complex patterns (Tomasini et al., 2023).

Recently, a novel class of neural networks, known as Vision Transformers, has been introduced to the field of computer vision (Dosovitskiy et al., 2021). Transformers were initially introduced for Natural Language Processing (NLP) and achieved remarkable performance due to the use of self-attention in their architecture (Vaswani et al., 2017). Unlike traditional CNNs, Vision Transformers divide an image into a sequence of patches and apply a transformation model to them, enabling the model to understand spatial patterns (Han et al., 2023). Vision Transformers are now being considered competitors to CNNs, after demonstrating the ability to capture complex patterns and achieving remarkable results across diverse image datasets in various computer vision tasks (Hatamizadeh et al., 2023; Kuprashevich and Tolstykh, 2023; Liu et al., 2025). Although Vision Transformers have achieved remarkable performance for various computer vision tasks, they also present their challenges, including the requirement for larger datasets and high computing resources for effective training, compared to CNNs (Wang et al., 2022).

Since the emergence of Vision Transformers, only a few studies have been conducted for a comparative analysis between Vision Transformers and CNNs for various computer vision applications, including image classification (Raghu et al., 2021; Maurício et al., 2023), face recognition (Takahashi et al., 2024), and action recognition (Moutik et al., 2023). However, for facial age estimation, the performance evaluation of CNNs and Vision Transformers has yet to be thoroughly explored.

In this paper, we present a comparative analysis of the performance of CNNs and Vision Transformers for the age estimation task on two datasets (FG-NET and UTKFace). Additionally, we compare the performance of both CNNs and Vision Transformers across two age groups (minors and adults) to further illustrate the effectiveness of these models. In the following sections, we present the related work (Section 2), methodology (Section 3), experiments and results (Section 4), and conclusions and future work (Section 5).

2. Related Work

2.1. CNN-based Age Estimation

Age estimation has been mostly addressed by using CNNs. A deep expectation model based on the VGG-16 architecture was proposed to estimate real and apparent age from a single facial image without relying on facial landmarks (Rothe et al., 2018). Zhao et al. (2022) proposed an adaptive mean residue loss for effective facial age estimation, using VGG-16 and ResNet-50 as backbone feature extractors. The proposed mean loss penalizes age probabilities between the mean of the estimated age distribution and the apparent age. Feature reconstruction based on knowledge distillation was proposed to address age estimation in occlusion scenarios, utilizing multiple variants of the ResNet model, including ResNet-34, ResNet-50, and ResNet-101, for feature extraction (Yu and Zhao, 2025). Kuang et al. (2023) proposed the EfficientRF model based on the integration of EfficientNet and Random Forest to estimate age from the facial images.

2.2. Transformer-based Age Estimation

Since the introduction of Vision Transformers, they have been widely adopted for image classification tasks (Dosovitskiy et al., 2021). However, their potential for age estimation remains largely unexplored. Shi et al. (2023) proposed a combination of an attention-based convolution module with a Swin transformer to predict the ages from facial images. Xu et al. (2025) presented the Cross Spatial and Cross-Scale Swin Transformer (CSCS-Swin) to improve facial age estimation by extracting fine-grained age-related features using a Cross Spatial Feature Block (CSFB) for directional wrinkle and craniofacial features and a Cross-Scale Feature Partition (CSFP) for multi-scale feature discrimination. A Feature Enhancement Module (FEM) was also proposed which refines feature representation to resolve ambiguities in distinguishing adjacent ages. A novel Multi-input Vision Transformer model (MiVOLO) was proposed to simultaneously predict age and gender, enhancing performance in scenarios where faces are partially or fully occluded (Kuprashevich and Tolstykh, 2023).

3. Methodology

To analyze the performance of CNNs and ViTs for the age estimation task, we propose an age estimation pipeline, illustrated in Figure 1.

3.1. Face Cropping and Alignment

Face cropping and alignment are performed as preprocessing steps to provide clean, task-relevant input for model training. Facial images are cropped using *dlib*'s facial landmark detection (King, 2009). Face alignment then standardizes facial orientation using *dlib* for landmark detection and OpenCV for image transformation. A 68-point facial landmark predictor extracts coordinates for the left eye (landmark 36), right eye (landmark 45), and nose tip (landmark 30). The rotation angle is calculated via the *arctangent* of eye coordinate differences, aligning the eyes horizontally with their midpoint as the pivot. A 2D rotation matrix is applied using cubic interpolation to produce an aligned facial image.

3.2. Convolutional Neural Networks

We utilize three CNNs(VGG-16, ResNet-50, EfficientNet-B0) to estimate age from facial images. VGG-16 is chosen for its deep, uniform architecture, ideal for learning hierarchical facial features across age groups. ResNet-50 is selected for its residual connections, improving gradient flow for robust age estimation. EfficientNet-B0 is picked for its optimized scaling, balancing efficiency and accuracy on diverse datasets. Each model's final fully connected layer is replaced with a linear layer to output a continuous age value $\hat{y}$.

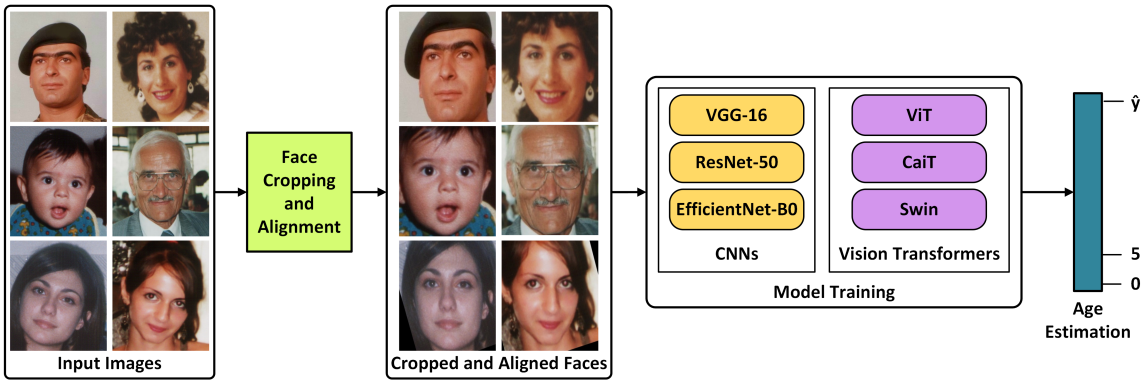

Figure 1: Overview of the proposed pipeline for age estimation. The input images are first cropped and aligned and then processed through CNNs and Vision Trasformers for the age estimation,

3.2.1. VGG-16

Among the CNNs, we first implement the age estimation pipeline using the VGG-16 CNN model. The input image ($X \in \mathbb{R}^{224\times224\times3}$) passes through 13 convolutional layers (3×3 kernels, stride 1, padding 1) organized into five blocks, with max-pooling layers (2×2 kernel, stride 2) reducing spatial dimensions to 112×112, 56×56, 28×28, 14×14, and 7×7. These layers extract hierarchical features, from low-level edges to high-level facial structures. The final convolutional layer's feature maps ($\mathbb{R}^{7\times7\times512}$) are flattened into a 25088-dimensional vector, processed through two fully connected layers (4096 units each, ReLU activations), and output a feature vector ($F \in \mathbb{R}^{1\times4096}$).

3.2.2. ResNet-50

The input image ($X \in \mathbb{R}^{224\times224\times3}$) is processed through five stages. The first stage employs a 7×7 convolutional layer (stride 2) and a 3×3 max-pooling layer (stride 2), thereby reducing the dimensions to 56×56. Four subsequent stages utilize bottleneck blocks with residual connections, resulting in a total of 50 layers that efficiently extract features. Feature maps are reduced to 56×56, 28×28, 14×14, and 7×7, with the number of channels increasing from 256 to 2048. The final stage's feature maps ($\mathbb{R}^{7\times7\times2048}$) are flattened into a 100352 dimensional vector, processed through a fully connected layer to produce a feature vector ($F \in \mathbb{R}^{1\times2048}$).

3.2.3. EffiecientNet-B0

The input image ($X \in \mathbb{R}^{224\times224\times3}$) is processed through nine stages. The first stage uses a 3×3 convolutional layer (stride 2), reducing dimensions to 112×112. Eight subsequent stages use mobile inverted bottleneck blocks (MBConv) with squeeze-and-excitation, totaling 237 layers, with 3×3 and 5×5 kernels for efficient feature extraction. Feature maps reduce to 112×112, 56×56, 28×28, 14×14, 14×14, 7×7, 7×7, and 7×7, with channels increasing from 32 to 1280. The final stage's feature maps ($\mathbb{R}^{7\times7\times1280}$) are flattened into a 62720 dimensional vector, processed through a fully connected layer to produce a feature vector ($F \in \mathbb{R}^{1\times1280}$).

3.3. Vision Transformers

To perform age estimation using Vision Transformers, we utilize the Vision Transformer (ViT), Class-Attention Image Transformer (CaiT), and Swin Transformer (Swin). ViT is chosen for its global attention, modeling long-range facial feature dependencies. CaiT is selected for its class-attention layers, focusing on age-discriminative patterns. Swin is picked for its hierarchical shifted-window attention, capturing multi-scale features for diverse datasets. Each model's final classification head is replaced with a linear layer to output a continuous age value $\hat{y}$.

3.3.1. ViT

For the Vision Transformer (ViT), the input image ($X \in \mathbb{R}^{224\times224\times3}$) is divided into a 14×14 grid of patches (with a size of 16×16 pixels each, 196 patches), each flattened into a 768-element vector ($16\times16\times3$). These vectors are embedded into a 768-dimensional space via linear projection. A class token is appended to the sequence, and positional encodings are added to retain spatial information. The sequence passes through 12 transformer encoder layers, each using multi-head self-attention (12 heads) and feed-forward networks to capture global dependencies. The final layer outputs a sequence of 768-dimensional vectors, from which the class token's representation is extracted as the feature vector ($F \in \mathbb{R}^{1\times768}$).

3.3.2. CaiT

The input image ($X \in \mathbb{R}^{224\times224\times3}$) is divided into a 14×14 grid of 16×16 pixel patches (196 patches), each flattened into a 768-element vector ($16 \times 16 \times 3$), embedded into a 384-dimensional space via linear projection, with a fixed class token appended and positional encodings added for spatial information. The sequence passes through 24 transformer encoder layers, with 22 self-attention layers and 2 class-attention layers, each using multi-head self-attention (6 heads) and feed-forward networks to capture global dependencies. The final class-attention layers produce the class token's feature vector ($F \in \mathbb{R}^{1\times384}$).

3.4. Swin

The input image ($X \in \mathbb{R}^{224\times224\times3}$) is divided into a 56×56 grid of 4×4 pixel patches (3136 patches), each flattened into a 48-element vector ($4 \times 4 \times 3$) and embedded into a 128-dimensional space via linear projection. Relative positional biases are incorporated for window-based attention. The sequence passes through 18 transformer encoder layers in four

Table 1: Comparison of CNNs and Vision Transformers for age estimation.

Models / Datasets		FG-NET		UTKFace	
		MAE	CS (%)	MAE	CS (%)
CNNs	VGG-16 (Simonyan and Zisserman, 2014)	3.52	76.51	4.85	64.70
	ResNet-50 (He et al., 2016)	3.11	80.31	4.81	64.73
	EfficientNet-B0 (Tan and Le, 2019)	3.90	75.38	4.83	64.18
Vision Transformers	ViT (Dosovitskiy et al., 2020)	3.18	81.23	4.40	68.51
	CaiT (Touvron et al., 2021)	3.30	79.79	4.50	67.24
	Swin (Liu et al., 2021)	**2.79**	**83.90**	**4.37**	**68.73**

stages. Stage 1 uses 2 layers with shifted window-based self-attention (window size 7×7) on 3136 patches with 128 dimensions. Stages 2, 3, and 4, each with patch merging, use 2, 6, and 2 layers, processing 784 patches (28×28, 256 dimensions), 196 patches (14×14, 512 dimensions), and 49 patches (7×7, 1024 dimensions), respectively. The final feature map ($\mathbb{R}^{7\times7\times1024}$) is globally average-pooled to produce a feature vector ($F \in \mathbb{R}^{1\times1024}$).

4. Experiments and Results

In this section, we present the details of the experiments conducted and the results obtained.

4.1. Datasets

We utilize two benchmark age estimation datasets, FG-NET (Lanitis et al., 2002) and UTKFace (Zhang et al., 2017), to evaluate the proposed pipeline. The distribution of samples for both datasets is shown in Figure 2.

FG-NET dataset comprises 1,002 images taken from 82 people, covering the ages from 0 to 69 years. It exhibits significant variations in head pose, facial expressions, and lighting conditions, making it a challenging dataset despite its small size.

UTKFace dataset has more than 23000 images covering an age range from 0 to 116 years. It also contains significant variations in head pose, facial expressions, and lighting conditions.

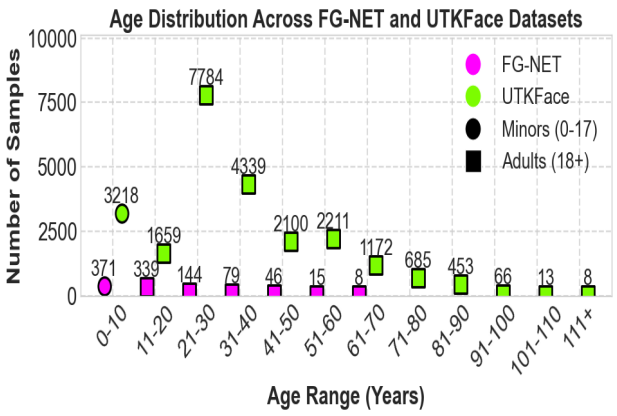

Figure 2: Age distribution of FG-NET and UTKFace datasets.

4.2. Evaluation Metrics

To assess the performance of the proposed pipeline, we utilize the Mean Absolute Error (MAE) and Cumulative Score (CS). MAE measures the average absolute difference between the true age y and the predicted age $\hat{y}$. CS represents the percentage of images where the absolute difference between y and $\hat{y}$ is below a threshold I, set to 5 in this study. MAE is computed as $\frac{1}{N}\sum_{i=1}^{N}|y-\hat{y}|$, where N is the total number of images. CS is calculated as $\frac{N_I}{N}\times100\%$, where N_I is the count of images with $|y-\hat{y}| < I$.

4.3. Implementation Details

We begin by cropping and aligning the input facial images using the dlib facial landmark detector King (2009). The images are then resized to 224×224 pixels. Datasets are split into 60% training, 20% validation, and 20% test sets. We evaluate three CNNs (VGG-16, ResNet-50, EfficientNet-B0) and three Vision Transformers (ViT, CaiT, Swin), pre-trained on ImageNet, with the final classification layer replaced by a linear layer for age regression. Models are fine-tuned using the AdamW optimizer (learning rate 5e-5, weight decay 5e-4), with batch sizes of 8 (FG-NET) and 16 (UTKFace). A Cosine Annealing Learning Rate Scheduler is utilized to dynamically adjust the learning rate for stable convergence. Training epochs are set to 200, with an early stopping criterion of 10 epochs, to mitigate overfitting. Huber Loss is used as the loss function with $\delta = 1$ due to its robustness to outliers. Five-fold cross-validation is applied, and the final prediction averages the outputs of five models.

4.4. Results

We evaluated our proposed pipeline using three CNN models (VGG-16, ResNet-50, EfficientNet-B0) and three Vision Transformers (ViT, CaiT, Swin Transformer). Table 1 illustrates the performance evaluation of both CNNs and Vision Transformers for the task of age estimation.

Among the CNN models, ResNet-50 outperformed VGG-16 and EfficientNet-B0 on the FGNET dataset, achieving an MAE of 3.11 years and a CS of 80.31%, compared to an MAE of 3.52 years and CS of 76.51% for VGG-16, and an MAE of 3.90 years and CS of 75.38% for EfficientNet-B0. On the UTKFace dataset, ResNet-50 slightly outperformed EfficientNet-B0 and VGG-16, with an MAE of 4.81 years and CS of 64.73%, compared to 4.83 years and 64.18% for EfficientNet-B0, and 4.85 years and 64.70% for VGG-16. The higher CS of ResNet-50 on FGNET indicates a greater proportion of predictions within the acceptable age tolerance, reflecting its robustness on the smaller, less diverse FGNET dataset. On UTKFace, the CS values are closer across CNNs, suggesting similar performance in capturing predictions within the tolerance range, despite the dataset's greater diversity in age, ethnicity, and lighting conditions. The residual connections in

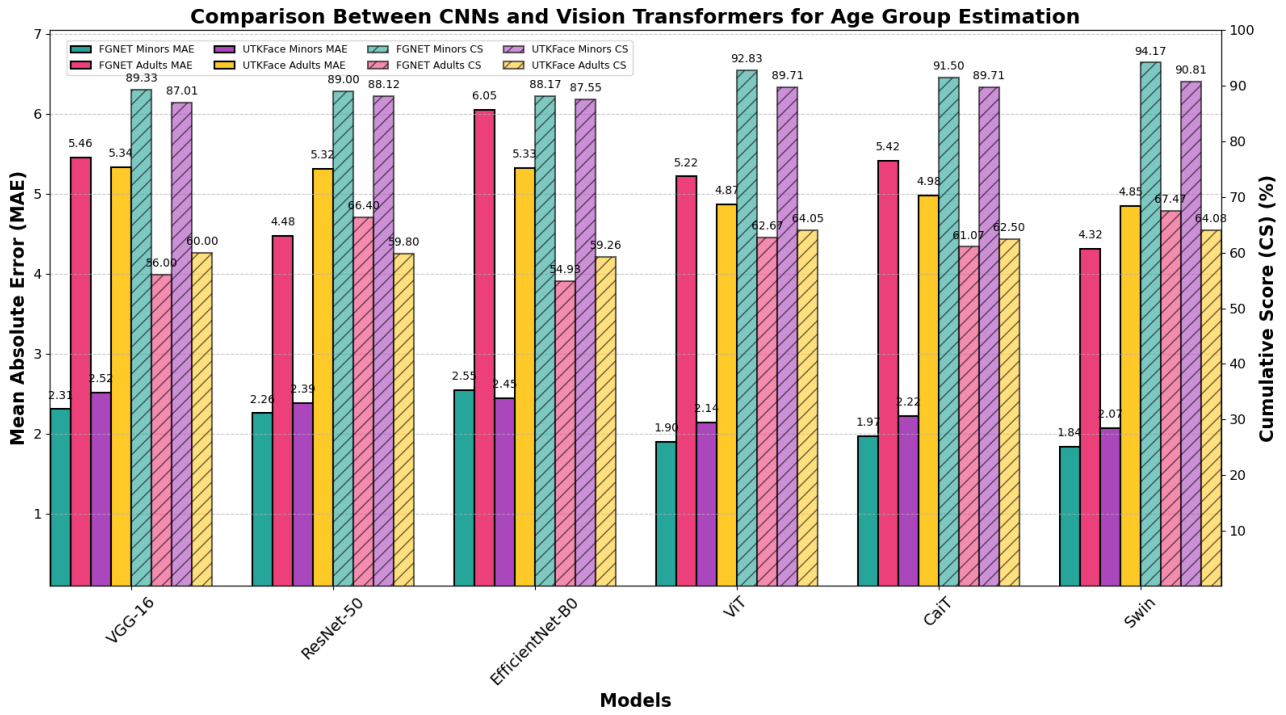

Figure 3: Performance comparison of CNNs and Vision Transformers for age estimation in minor and adult age groups.

ResNet-50 enable better gradient flow and training stability, contributing to its superior MAE and CS on both datasets.

In the case of Vision Transformers, The Swin Transformer outperformed ViT and CaiT on both FGNET and UTKFace datasets. On FGNET, Swin achieved a MAE of 2.79 years and a cumulative score (CS) of 83.90%, compared to ViT (3.18 years, 81.23%) and CaiT (3.30 years, 79.79%). On UTKFace, Swin recorded an MAE of 4.37 years and a CS of 68.73%, slightly ahead of ViT (4.40, 68.51%) and CaiT (4.50, 67.24%). Its hierarchical architecture with shifted window attention effectively captures both local and global features, especially benefiting smaller datasets like FGNET. On UTKFace, while all models face challenges due to data complexity, Swin's multi-scale feature processing provides a modest advantage. Overall, Swin not only surpassed ViT and CaiT but also outperformed the best CNN model, ResNet-50, reducing MAE by 1.11 years and improving CS by 3.59% on FGNET, and lowering MAE by 0.46 years with a 4.00% higher CS on UTKFace.

We also evaluated the performance of CNN and Vision Transformer models for two age groups: minors (0-17) and adults (18+), as shown in Figure 3. For minors, ResNet-50 achieved the lowest MAEs of 2.26 and 2.39 years on FG-NET and UTKFace, respectively, with CS scores of 89.00% and 88.12%, slightly outperforming VGG-16 and EfficientNet-B0. Although VGG-16 achieved a marginally higher CS of 89.33% on FG-NET, ResNet-50's lower MAE reflects superior precision, benefiting from residual connections that enhance feature extraction, particularly in the 0–20 age range prevalent in both datasets (see Figure 2). Among adults, ResNet-50 again led with an MAE of 4.48 years and a CS of 66.40% on FG-NET, outperforming VGG-16 and EfficientNet-B0, which struggled due to fewer high-age samples. On UTKFace, ResNet-50 slightly outperformed the others in MAE, with CS values remaining close across models due to that dataset's more balanced distribution across higher age ranges, except for few age groups.

Among Vision Transformers, the Swin Transformer outperformed ViT and CaiT for both minor and adult age groups on the FG-NET and UTKFace datasets. For minors, it achieved the lowest MAEs (1.84 years on FG-NET and 2.07 years on UTKFace) and highest CS scores (94.17% on FG-NET and 90.81% on UTKFace), surpassing ViT and CaiT on both datasets. For adults, Swin also led, outperforming ViT and CaiT by up to 1.10 years in MAE and 1.33% in CS on FG-NET, with slight advantages on UTKFace.

Overall, both CNNs and Vision Transformers performed better for minors than for adults, achieving lower MAEs and higher CS scores. Predictable developmental features in minors, such as facial growth patterns, enhance model accuracy, whereas adult predictions are less accurate due to varied aging signs influenced by lifestyle and environmental factors. Additionally, imbalanced age distributions in both datasets contribute to higher errors and lower consistency in adult age estimation.

5. Conclusions and Future Work

In this paper, we present a comparative analysis of the performance of CNNs and Vision Transformers for age estimation on two benchmark datasets (FG-NET, UTKFace). We evaluate CNN models (VGG-16, ResNet-50, EfficientNet-B0) and Vision Transformers (ViT, CaiT, Swin) under identical settings for age estimation and age group estimation for minors (0-17 years) and adults (18 years and older). Experimental results demonstrate that the Swin Transformer outper-

formed other Vision Transformers and CNNs in both tasks, achieving MAEs of 2.79 and 4.37 years with CS scores of 83.90% and 68.73% for overall age estimation on FG-NET and UTKFace, respectively. For age group estimation, the Swin Transformer achieved MAEs of 1.84 and 2.07 years with CS scores of 94.17% and 90.81% for minors, and MAEs of 4.32 and 4.85 years with CS scores of 67.47% and 64.08% for adults on FG-NET and UTKFace, respectively. In future work, we aim to explore techniques to address dataset imbalances in age estimation and enhance robustness in occluded scenarios for real-world applications. These advancements will support multiple practical applications, including border security systems, personalized healthcare diagnostics, and age-based content filtering.

Acknowledgments

This work has been funded by the Recovery, Transformation, and Resilience Plan, financed by the European Union (Next Generation), thanks to the LUCIA project (Fight against Cybercrime by applying Artificial Intelligence) granted by INCIBE to the University of León.

References

Agbo-Ajala, O., Viriri, S., 2021. Deep learning approach for facial age classification: a survey of the state-of-the-art. Artificial Intelligence Review 54 (1), 179–213.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16x16 words: Transformers for image recognition at scale.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., Yang, Z., Zhang, Y., Tao, D., 2023. A survey on vision transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (1), 87–110.
DOI: 10.1109/TPAMI.2022.3152247

Hatamizadeh, A., Yin, H., Heinrich, G., Kautz, J., Molchanov, P., 2023. Global context vision transformers. In: Proceedings of the 40th International Conference on Machine Learning. ICML'23.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778.
DOI: 10.1109/CVPR.2016.90

Hiba, S., Keller, Y., 2023. Hierarchical attention-based age estimation and bias analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (12), 14682–14692.
DOI: 10.1109/TPAMI.2023.3319472

King, D. E., 2009. Dlib-ml: A machine learning toolkit. The Journal of Machine Learning Research 10, 1755–1758.

Kuang, H., Huang, X., Ma, X., Liu, X., 2023. Efficientrf: Facial age estimation based on efficientnet and random forest. In: 2023 IEEE 3rd International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA). Vol. 3. pp. 196–200.
DOI: 10.1109/ICIBA56860.2023.10165244

Kuprashevich, M., Tolstykh, I., 2023. Mivolo: Multi-input transformer for age and gender estimation. In: International Conference on Analysis of Images, Social Networks and Texts. pp. 212–226.

Lanitis, A., Taylor, C., Cootes, T., 2002. Toward automatic simulation of aging effects on face images. IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (4), 442–455.
DOI: 10.1109/34.993553

Li, X., Wang, L., Zhu, R., Ma, Z., Cao, J., Xue, J.-H., 2025. Srml: Structure-relation mutual learning network for few-shot image classification. Pattern Recognition 168, 111822.
DOI: https://doi.org/10.1016/j.patcog.2025.111822

Liu, P., Qian, W., Huang, J., Tu, Y., Cheung, Y.-M., 2025. Transformer-driven feature fusion network and visual feature coding for multi-label image classification. Pattern Recognition 164, 111584.
DOI: https://doi.org/10.1016/j.patcog.2025.111584

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9992–10002.
DOI: 10.1109/ICCV48922.2021.00986

Maurício, J., Domingues, I., Bernardino, J., 2023. Comparing vision transformers and convolutional neural networks for image classification: A literature review. Applied Sciences 13 (9).
DOI: 10.3390/app13095521

Moutik, O., Sekkat, H., Tigani, S., Chehri, A., Saadane, R., Tchakoucht, T. A., Paul, A., 2023. Convolutional neural networks or vision transformers: Who will win the race for action recognitions in visual data? Sensors 23 (2).
DOI: 10.3390/s23020734

Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., Dosovitskiy, A., 2021. Do vision transformers see like convolutional neural networks? In: Advances in Neural Information Processing Systems.

Rothe, R., Timofte, R., Van Gool, L., 2018. Deep expectation of real and apparent age from a single image without facial landmarks. International Journal of Computer Vision 126 (2), 144–157.

Shi, C., Zhao, S., Zhang, K., Wang, Y., Liang, L., 2023. Face-based age estimation using improved swin transformer with attention-based convolution. Frontiers in Neuroscience Volume 17 - 2023.
DOI: 10.3389/fnins.2023.1136934

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Song, Y., Wang, F., 2024. Coreface: Sample-guided contrastive regularization for deep face recognition. Pattern Recognition 152, 110483.
DOI: https://doi.org/10.1016/j.patcog.2024.110483

Takahashi, S., Sakaguchi, Y., Kouno, N., Takasawa, K., Ishizu, K., Akagi, Y., Aoyama, R., Teraya, N., Bolatkan, A., Shinkai, N., et al., 2024. Comparison of vision transformers and convolutional neural networks in medical image analysis: a systematic review. Journal of Medical Systems 48 (1), 84.

Tan, M., Le, Q., 09–15 Jun 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. In: Chaudhuri, K., Salakhutdinov, R. (Eds.), Proceedings of the 36th International Conference on Machine Learning. Vol. 97. pp. 6105–6114.

Tomasini, U. M., Petrini, L., Cagnetta, F., Wyart, M., 2023. How deep convolutional neural networks lose spatial information with training. Machine Learning: Science and Technology 4 (4), 045026.

Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., Jégou, H., 2021. Going deeper with image transformers. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 32–42.
DOI: 10.1109/ICCV48922.2021.00010

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., Polosukhin, I., 2017. Attention is all you need. In: Advances in Neural Information Processing Systems. Vol. 30.

Wang, X., Zhang, L. L., Wang, Y., Yang, M., 2022. Towards efficient vision transformer inference: a first study of transformers on mobile devices. In: Proceedings of the 23rd Annual International Workshop on Mobile Computing Systems and Applications. p. 1–7.
DOI: 10.1145/3508396.3512869

Xu, L., Hu, C., Shu, X., Yu, H., 2025. Cross spatial and cross-scale swin transformer for fine-grained age estimation. Computers and Electrical Engineering 123, 110264.

Yu, S., Zhao, Q., 2025. Improving age estimation in occluded facial images with knowledge distillation and layer-wise feature reconstruction. Applied Sciences 15 (11).
DOI: 10.3390/app15115806

Zhang, Z., Song, Y., Qi, H., 2017. Age progression/regression by conditional adversarial autoencoder. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4352–4360.
DOI: 10.1109/CVPR.2017.463

Zhao, Z., Qian, P., Hou, Y., Zeng, Z., 2022. Adaptive mean-residue loss for robust facial age estimation. In: 2022 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6.
DOI: 10.1109/ICME52920.2022.9859703