Visual Age Estimation of Infant Photographs Using Deep Neural Networks

Marcelo Fernandez
Stanford University
450 Jane Stanford Way
marcefer@stanford.edu

Edgar Leon
Stanford University
450 Jane Stanford Way
edgar11@stanford.edu

Abstract

This paper explores the challenge of estimating infant age from facial photographs using deep learning. We evaluate three vision architectures: VGG16, a convolutional neural network, ViT-B/16, a Vision Transformer, and two variants of SimCLR-V2 classifier, framing the task as a 3-class classification problem (ages 1, 2, or 3). To address the scarcity of baby-specific data, we construct an expanded dataset by augmenting UTKFace with curated web-sourced baby images and label-preserving transformations. Our models are fine-tuned and benchmarked on a held-out test set, with confusion matrices used to analyze performance by age class. While VGG16 excels at identifying younger infants and ViT achieves stronger generalization and higher overall accuracy, the SimCLR-V2 outperforms both. These findings highlight the promise of transfer learning for fine-grained age estimation in early childhood.

1. Introduction

Estimating the age of individuals from facial images is a widely studied problem in computer vision, with applications in biometrics, content personalization, and healthcare. While some models exist for adult age estimation, the task becomes significantly more challenging when applied to infants and toddlers. Subtle developmental changes, limited facial structure differentiation, and the lack of large annotated datasets make baby age estimation a largely underexplored domain.

This project focuses on predicting the age category of babies from single facial images using deep learning. Specifically, we evaluate the effectiveness of three powerful model families: convolutional neural networks (CNNs), Vision Transformers (ViTs), and two variants of SimCLR with a Resnet50 backbone. The CNN, ViT models were pretrained on ImageNet, while one variant of our model was also pretrained on ImageNet, the other variant was trained with self-supervision on UTKFace dataset before fine tuning. We aim to determine which model generalizes better under real-world conditions and limited data by comparing their performance on the same dataset, both with and without data augmentation.

Given the scarcity of publicly available, baby-specific face datasets, we began with a filtered subset of the UTKFace dataset and augmented it by crawling additional images from the web. To further increase data diversity and quantity, we applied targeted data augmentation techniques including rotation, color jitter, and perspective transformations. This resulted in a balanced and sufficiently large training set, allowing for meaningful evaluation of deep learning models.

Our goal is not only to assess model accuracy but also to analyze where and why different architectures succeed or fail. By reporting validation accuracy, confusion matrices, and qualitative trends in predictions, we provide an empirical foundation for future work in baby age estimation using modern vision architectures.

2. Related Work

Early methods for age estimation relied on handcrafted features and statistical models. Techniques such as Local Binary Patterns (LBP) and Scale-Invariant Feature Transform (SIFT) were employed to capture facial textures and shapes, which were then used in conjunction with classifiers like Support Vector Machines (SVMs) for age prediction. However, these methods often struggled with variations in lighting, pose, and facial expressions, leading to limited accuracy, especially in unconstrained environments.

Deep Learning has revolutionized age estimation tasks. Convolutional Neural Networks (CNNs) have demonstrated superior performance by automatically learning hierarchical features from data [1,7]. Models like VGG-16, ResNet, and Inception have been fine-tuned for age estimation tasks, achieving notable accuracy improvements. For instance, the Deep EXpectation (DEX) model utilized a VGG-16 [1] architecture trained on the IMDB-WIKI dataset, achieving impressive results in age estimation tasks.

Despite these advancements, most deep learning models have been trained and evaluated on datasets predominantly consisting of adult faces, such as MORPH, FG-NET, and IMDB-WIKI [1][3]. This focus limits their applicability to infant age estimation, where facial features differ significantly.

Estimating the age of infants poses distinct challenges. The rapid and subtle changes in facial features during early development require models to be sensitive to minute differences. The variation in growth and development among children of the same age (newborn to three years) tends to be greater than the differences seen across different ages; these are attributed to factors such as genetic, environmental and nutritional differences; furthermore, growth is not linear: babies and toddlers experience spurts in growth rather than steady increases. [15]

Moreover, the scarcity of large-scale, annotated infant face datasets hampers the training of robust deep learning models.

To address these challenges, researchers have explored various approaches:

- BabyFace Dataset: Dataset comprising over 15,000 images of infants aged 0 to 24 months. The study proposed SSR-Net with an attention mechanism, achieving an age estimation error of less than two months. [17]

- Gestational Age Estimation: Researchers developed a system combining CNNs and Support Vector Regression to estimate gestational age using images of newborns' faces, feet, and ears, achieving an expected error of six days. [18]

- Skull Radiograph Analysis: A study utilized deep learning models on skull X-ray images to predict the postnatal age of infants under 12 months, demonstrating the potential of medical imaging in age estimation.[19]

Recently, Transformer architectures have gained attention in computer vision tasks. Vision Transformers (ViTs) have shown promise in capturing global contextual information, which is beneficial for age estimation. While ViTs have been primarily applied to adult face datasets, their application to infant age estimation remains an emerging area of research. [6]

The field of infant age estimation is still developing, with challenges stemming from limited datasets and the subtlety of facial changes in early development. While traditional methods laid the groundwork, deep learning approaches, particularly CNNs, have significantly advanced the field. The introduction of specialized datasets like BabyFace and the exploration of Transformer-based, CNN and contrastive learning models offer promising directions for future research.

3. Methods

We formulate the task of baby age estimation as a 3-class classification problem, where the input is a single RGB image of a baby's face and the output is a discrete age class: 1 year, 2 years, or 3 years. Let the input image be denoted by $x \in R^{3x224x224}$ and the target label by $y \in \{0,1,2\}$, corresponding to the three age categories. Given an image x, our model outputs a probability distribution $\hat{p} = f\theta(x) \in R^3$ where $\sum i \ \hat{p}i = 1$ via a softmax layer. The predicted age class is then given by $\hat{y} = \arg\max_i \hat{p}i$.

We experiment with four modern deep architectures:

- VGG-16 CNN: A deep convolutional network with 13 convolutional layers and 3 fully connected layers. We use the ImageNet-pretrained weights and finetune only the final classifier layer by replacing the 1000-way output with a 3-class output head. All convolutional layers are frozen during training, and only the classifier is updated.

- ViT-B/16 (Vision Transformer): A transformer-based architecture that divides an image into 16×16 patches and processes them as a sequence of tokens. We use the ViT_B_16_Weights pretrained weights from PyTorch. All transformer blocks are frozen, and we finetune only the classification head. Both models are trained using cross-entropy loss.

- SimCLRv2 r50_1x_sk1: A self-supervised learning model designed to leverage large amounts of unlabeled data while requiring minimal labeled examples for fine-tuning. Built with Resnet50 with 35 million parameters and pretrained on ImageNet.

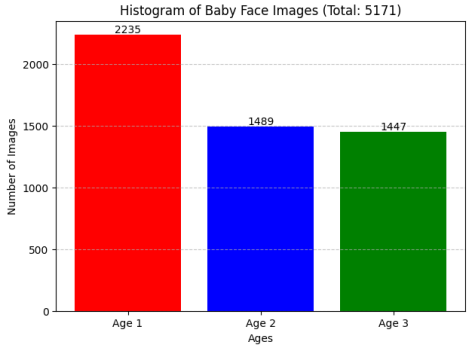
- SimCLRv2 UTKFace: A new model trained on contrastive learning with UTKFace dataset for 270 epochs, and a classifier added on top.

4. Dataset and Features

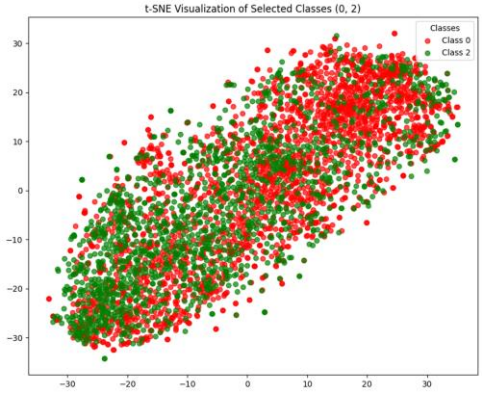
Our experiments are based on a combination of the UTKFace dataset [11] and a supplementary collection of baby images crawled from the Internet. The

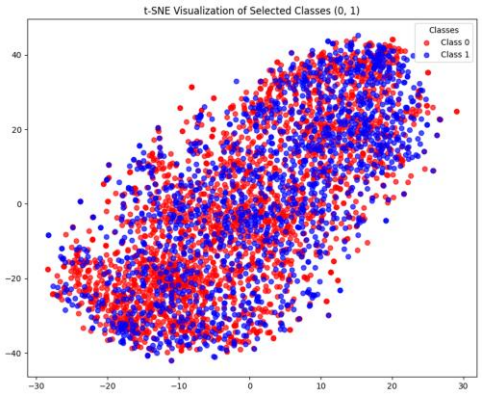
UTKFace dataset is a large-scale facial dataset labeled with age, gender, and ethnicity, from 0 - 116 years old. Unfortunately, infant images represent only a small fraction (8.5%) of the dataset and are highly imbalanced across age labels. Furthermore, this dataset also presents significant inconsistencies in images between the three age groups: Age 1: 1282, Age 2: 531, Age 3: 318, with Age 1 representing 60% of the data. To mitigate this limitation, we constructed a curated superset and later applied extensive data augmentation to artificially expand our training set.

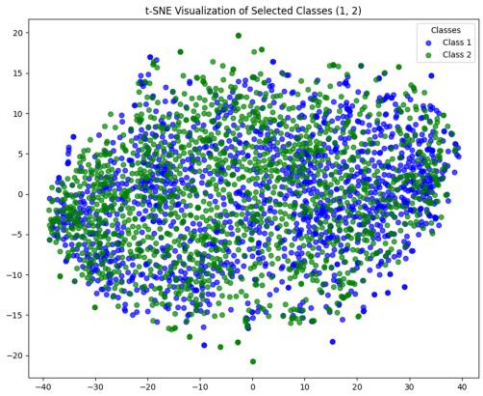
We selected only those images labeled with age 1, 2, or 3, based on the filename convention in UTKFace (e.g., 1_...jpg indicates ages between 0 - 1), and after adding the crawled images, this resulted initially in a dataset of: 2,235 images for age 1, 1489 for age 2, and 1,447 for age 3, for a total of 5,171 original baby face images.



Inspecting the dataset with t-SNE analysis we found Age 1 (Class 0) to have the most separation from Age 3 (Class 2) and Age 2 (Class 1) to have the least separation from the other two categories. This is not surprising since Age 2 is in a developmental stage between Age 1 and Age 3:







We therefore expect our models to have special difficulties identifying Age 2 category.

VGG-16 and ViT-B/16

For VGG and ViT models it is essential to have a large dataset, to increase the dataset size we applied three forms of label-preserving augmentation to every image:

- Random Rotation (±15°) — to simulate head pose variation
- Color Jitter (brightness, contrast) — to simulate lighting variability
- Perspective Warp — to simulate slight viewpoint shifts

After augmentation, we obtained an expanded dataset of:

- 5,688 images per class
- Total: 17,064 images

Each image was resized and cropped to a resolution of 224 × 224 pixels using the MTCNN face detector [8],

with a 20% margin added to preserve surrounding context (e.g. head and ears)
We did not extract hand-crafted features like HOG, SIFT, or PCA-reduced embeddings. Instead, we relied entirely on end-to-end feature learning from raw pixels using VGG-16 and ViT. No additional engineered features (e.g., landmarks or age-specific ratios) were used. The effectiveness of these deep features was validated by training only the final classification layer on our custom dataset.

5. Baseline Experiments / Results / Discussion

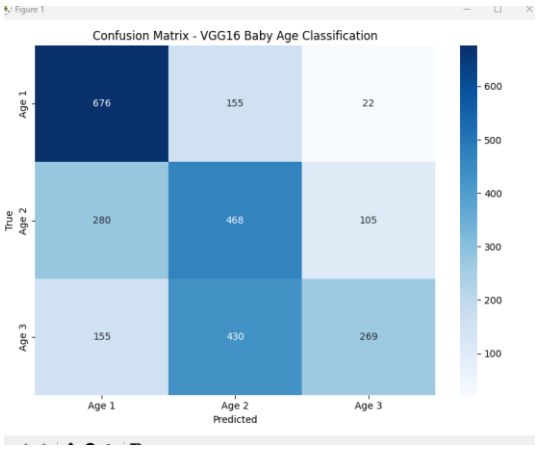
In the initial experiments (baseline), we trained both models on a smaller subset of the UTKFace dataset (with fewer than 1300 images total). The performance was promising but limited by data scarcity, several images with inaccurate ages (2 and 3) and class imbalance, with VGG-16 and Vit-B/16 resulting in validation accuracy of 66.7% and 65.4% respectively. Despite decent results, both models exhibited overfitting and reduced generalization due to the unbalanced nature of the initial data (60% Age 1), and thus, we discarded those results.

After an exhaustive search for an adequate dataset, we opted for augmenting to expand the dataset to 17,064 total images (5,688 per class), we retrained both models and observed marked improvements.

Final Validation Accuracy (Based on Confusion Matrix Evaluation):

Model	Correct Predictions	# Val Images	Final Accuracy
VGG-16	1411	2560	55.2%
ViT-B/16	1467	2560	57.3%

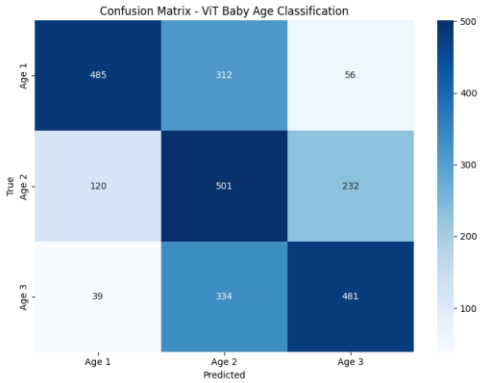
VGG-16 Confusion Matrix



Strong on Age 1: 79.2% accuracy (676/853)

Weak on Age 3: only 31.5% accuracy (269/854)
High Age 3→2 confusion observed

ViT Confusion Matrix



More balanced across all classes
Notably stronger on Age 3: 56.3% accuracy (481/854)
Moderate confusion between adjacent ages (1→2, 2→3)

Detailed Comparison

Metric	VGG-16	ViT
Total val .images	2560	2560
Correct predictions	1413	1467
Val accuracy	55.2%	57.3%
Age 1 accuracy	79.2%	56.9%
Age 2 accuracy	54.9%	58.7%
Age 3 accuracy	31.5%	56.3%
Bias observed	Overconfident on Age 1	Balanced, slightly Age 2 heavy
Generalization	Weaker on Age 3	Stronger overall
Confusion Trend	High Age 3 -> Age 2 confusion (UTKDataset shows this issue too)	Smoother Age 1→2→3 transitions

Qualitative Error Analysis

Both models struggle most with distinguishing between Ages 2 and 3, confirming our prediction from t-SNE analysis and the confusion matrix findings, and the mislabeled data we identified with the original UTKFace Dataset. These qualitative samples highlight the visual ambiguity of certain faces and the

importance of fine-grained cues for infant age estimation.

Overfitting and generalization

VGG16 showed signs of overfitting to younger age classes, particularly Age 1, where it achieved high accuracy but deteriorated sharply on Age 3.

ViT, while less accurate on Age 1, generalized better across the entire age spectrum. This aligns with our hypothesis that ViT's attention mechanism better captures fine-grained facial features when given sufficient data.

Insights

VGG16 tended to rely on low-level features, leading to underperformance on less distinctive faces of older babies.

ViT's patch-based encoding led to smoother predictions, but still showed mild bias toward Age 2, potentially due to data augmentation artifacts or inherent ambiguity in that developmental stage.

Data augmentation significantly increased both models' robustness.

SimCLRv2 r50_1x_sk1

Next, we evaluated the dataset using two variations of SimCLRv2. The first variation was provided by google-research/simclr at github.com.

The model has been pretrained on the ImageNet dataset, providing high-quality feature embeddings from its backbone. Its layers are frozen, therefore fine-tuning the backbone layers is not possible, but given the limited size of our dataset, fine-tuning would not be promising. However, the model's strong performance in Linear Evaluation (74.6%) suggests that adding a classifier on top can effectively leverage its pretrained features for promising results.

We tried both, expanding our dataset, and using only the 5,171 images. We also tried extracting the face from each image and omitting extraction from our data pipeline. For the final classifier, we tried both a shallow classifier and alternatively a deep layer classifier. We also added minor image augmentation to the data pipelines which improved the speed of each training epoch.

The classifiers were built as follows:

Shallow Classifier:

- Images (224×224×3) undergo random flipping, rotation, and zoom to improve generalization.
- SimCLRv2 feature extractor model to extract the feature embeddings (proj_head_input).
- A 128-unit dense layer refines extracted features.
- Batch Normalization stabilizes training, while Dropout (0.3) prevents overfitting.
- A Softmax classifier predicts one of three classes

Deep Classifier:

- Input images also undergo random flipping, rotation, and zooming to improve generalization.
- SimCLRv2 feature extractor model to extract the feature embeddings (proj_head_input).
- First dense layer (512-unit, ReLU, dropout) refines extracted features.
- Second dense layer (256-unit, BatchNorm, dropout) introduces a skip connection:
    - If feature dimensions match, they are added.
    - Otherwise, they are concatenated to preserve feature integrity.
- Third dense layer (128-unit, BatchNorm, dropout) refines learned representations.
- A Softmax classifier predicts one of three classes

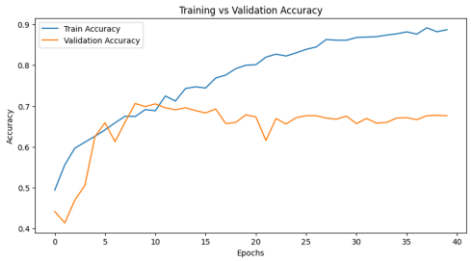
The training was performed with 20 epochs, we explored learning rates of 0.0005 to .0001, with a learning rate drop of 0.75*LR validation loss callback with patience = 3. All runs with a batch size of 64.

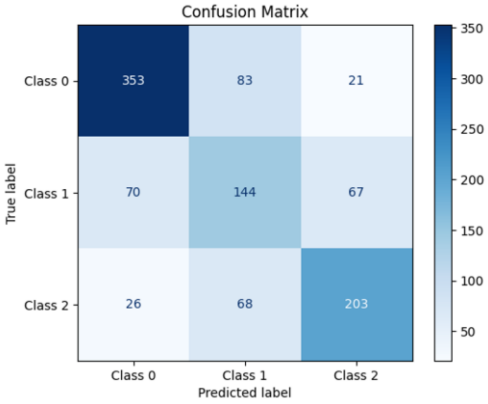
Table with results of our various approaches:

Classifier	Face Crop	Aug / exp	No Crop
Shallow	65.31	64.27	68.6
Deep	64.45	68.3	70.6

The above results of 68.6 show that despite the simplicity of the shallow classifier it performs close to the deep classifier, and no additional pre-processing on the images is needed. The shallow classifier performs better than VGG and ViT on less data.

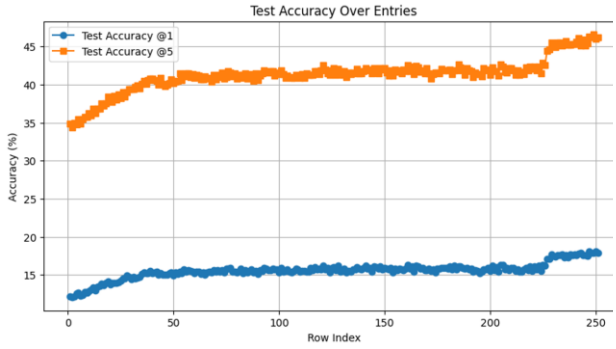
Below are the results of the SimCLRv2 with Deep Classifier:



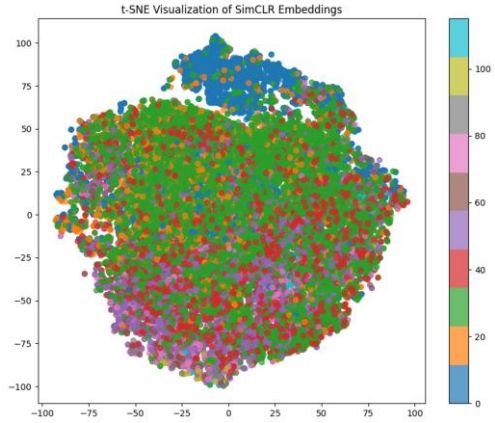


SimCLRv2 UTKFace

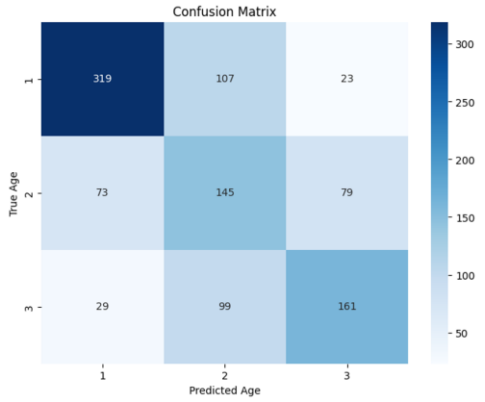
Next, we built a SimCLRv2 model from the ground up by doing self-supervision training with the UTKFace dataset. The main objective was to determine if we could surpass the results from r50_1x_sk1 by leveraging fine tuning. The UTKFace dataset consist of approximately 25,000 images, with age categories from 0 to 116, with various data sizes per category. The Restnet50 architecture was used as the backbone with image size set to 224x224, and normalization transformation applied to each image based on the calculated mean and standard deviation specific to this dataset.
Approximately 250 epochs were completed in 12 hours and achieved an accuracy of ~16%.



After extracting the features and analyzing with t-SNE, we can see that the categories from 0 – 10 (aqua blue) are being separated from the bulk of the data. These capture the age categories of interest in this investigation.



After adding the Deep Classifier on top, and unfreezing the backbone layers for fine tuning, it achieved a validation accuracy of 63.9%. The below results are with the original 5K dataset, without any image preprocessing:



We also tried the Shallow Classifier, and other schemes to improve the validation accuracy but are here omitted for brevity.

SimCLRv2 r50_1x_sk1 Error Analysis

The picture below shows examples of false predictions from the SimCLRv2 r50_1x_sk1model for babies of Age 1 predicted falsely as Age 2 (first row), and the True predictions (second row). When compared, the babies who are wrongly categorized appear in general older.

Furthermore, the picture below shows similar examples for false predictions for babies of Age 3 predicted as Age 2 (first row). In contrast to the previous example, these babies appear in general younger.

In both cases, the model is mis-categorizing images because it's picking up based on the fact that in the real world, in the span of a year, babies of Age 2 sometimes appear similar to Age 1 and sometimes appear similar to Age 3. If the model is recognizing facial details, higher resolution images in the train data might provide the granularity to train the model to recognize more subtle details about Age 2 and categorize it appropriately.

6. Conclusion / Future Work

Overall results:

	Val acc (%)			
Model	Overall	Age 1	Age 2	Age 3
VGG-16	55.2	79.2	54.9	58.7
ViT-B/16	57.3	56.9	58.7	56.3
SimCLRv2 UTKFace	63.9	71	49	55
SimCLRv2 r50_1x_sk1	70.63	77	51	70

Above we can see that all models struggle to predict the Age 2 category. It is surprising that the homebuilt SimCLRv2 model also outperformed VGG and ViT (63,9%). Overall r50_1x_sk1 model is the top performer according to validation accuracy, although the ViT seems to generalize better and provide more balance results.

In this project, we explored the challenge of estimating the relative age of babies from facial photographs using deep learning. This task is especially difficult due to the subtle visual changes between early childhood stages, compounded by a lack of large, annotated datasets specific to infant faces. To overcome this, we constructed an expanded dataset using UTKFace combined with crawled and augmented baby images, resulting in over 17,000 labeled samples across three target age classes: 1, 2, and 3 years old.

We evaluated three prominent deep learning architectures, VGG16, a convolutional neural network, ViT-B/16, a Vision Transformer model, and SimCLRV2, a contrastive learning self-supervision model. Although VGG16 and ViT-B/16 shine really well when there are large datasets available, in the real world we hardly find those datasets waiting for us. Unfortunately, augmenting and expanding artificially did not bring about the best results. However, SimCLRv2 conveyed to us its power by demonstrating its practicality in real world scenarios when freely abundant datasets are not available. By transferring learning, a powerful classifier can be built with a very small amount of data.

Significant challenges had to be overcome to bring this paper to fruition. However, by learning the ability to scrape images of the internet, and leverage self-supervised models and transfer learning, the skills acquired have opened the door to a world of unlimited data and deep learning.

If more time and compute were available, several promising directions could be explored:

- Fine-Grained Age Labels: Instead of discrete class labels (1, 2, 3), train on continuous age values with more precise annotations to improve granularity and evaluate regression-based formulations. The Mixup technique could be explored in this direction.

- Facial Landmark-Aware Models: Introduce explicit facial keypoint guidance or hybrid CNN-graph architectures to capture spatial growth patterns in baby faces.
- Bias and Robustness Analysis: Systematically evaluate the models across ethnicity, gender, lighting, and pose variations to ensure fair and stable performance in real-world deployment.
- Additional training to SimCLR UTKFace model: Explore tweaks to increase the effectiveness and efficiency of feature extraction, and increase its epochs.

Ultimately, this work represents a foundational step toward building AI systems that understand early facial development—a complex, underexplored domain with significant emotional and technical relevance.

7. Contributions & Acknowledgements

Marcelo:
- Data Preprocessing
- Image crawler
- Coded face detection script (via MTCNN), cropping, and resizing.
- Researched best transformers alternatives options and coded script to augment data.
- Implemented custom scripts to perform realistic and label-preserving augmentations.
- Finetuned both VGG16 and ViT.
- Trained and validated models using PyTorch, selecting hyperparameters and loss functions. (VGG and ViT)
- Implemented confusion matrix visualizations, validation accuracy tracking, and classification reporting. (VGG and ViT)
- Compared model performance across age groups and summarized findings through quantitative and qualitative analysis (VGG and ViT)
- Wrote and contributed to the final report, including the abstract, introduction, methods, dataset, experiments, discussion, conclusion, and citations.

Edgar:
-Data preprocessing
-Exploratory Data Analysis
-Image crawler
-Coded SimCLR model
-Researched pretrained SimCLR models
-Constructed, trained and validated SimCLR models.
-Implemented confusion matrix visualizations, validation accuracy tracking, and classification reporting. (SimCLR)
-Wrote and contributed to final report, including the abstract, introduction, methods, dataset, experiments, discussion, conclusion, and citations.
-Hand labeled new data
-Error analysis SimCLR

No other external collaborators or third-party contributions were involved in the execution of this work.

8. References / Bibliography

[1] P. Antipov, M. Baccouche, and J.-L. Dugelay. Facial age estimation using multi-stage deep neural network. *Electronics*, 13(16):3259, 2024. https://doi.org/10.3390/electronics13163259

[2] S. Chen, C. Zhang, and M. Dong. Fine-grained age estimation in the wild with attention LSTM networks. *arXiv preprint arXiv:1805.10445*, 2018. https://arxiv.org/abs/1805.10445

[3] M. K. Dobie. Age classification from facial images. *Computer Vision and Image Understanding*, 74(1):1–21, 1999. https://www.researchgate.net/publication/386508844_Age_Prediction_from_Facial_Images_Using_Deep_Learning_Architecture

[4] X. Shen, H. Shuai, and X. Bai. Deep regression forest for age estimation. *arXiv preprint arXiv:1712.07195*, 2017. https://arxiv.org/abs/1712.07195

[5] S. E. Choi et al. Age estimation using a hierarchical classifier based on global and local facial features. *Pattern Recognition*, 44(6):1262–1281, 2011. https://www.sciencedirect.com/science/article/abs/pii/S0031320310005704

[6] A. Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*, 2021. https://arxiv.org/abs/2010.11929

[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. https://arxiv.org/abs/1512.03385

[8] Z. Zhang et al. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016. https://arxiv.org/abs/1604.02878

[9] F. Schroff et al. FaceNet: A unified embedding for face recognition and clustering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[10] Torchvision. VGG and ViT pretrained models documentation. https://pytorch.org/vision/stable/models.html

[11] Z. Zhang, *UTKFace Dataset*, University of Tennessee. https://susanqq.github.io/UTKFace/
[12] A. Paszke et al. PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019. https://arxiv.org/abs/1912.01703
[13] M. Waskom. Seaborn: Statistical data visualization. https://seaborn.pydata.org
[14] P. Virtanen et al. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17:261–272, 2020. https://www.nature.com/articles/s41592-019-0686-2
[15] Julieana Nichols, MD, MPH; Teresa K. Duryea, MD; Alison G. Hoppin, MD. "Normal Growth Patterns in Infants and Prepubertal Children." UpToDate, 2025, [https://www.uptodate.com/contents/normal-growth-patterns-in-infants-and-prepubertal-children].

[16] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, Geoffrey Hinton. *"Big Self-Supervised Models Are Strong Semi-Supervised Learners."* 2006, [https://doi.org/10.48550/arXiv.2006.10029].
[17] Is it easy to recognize baby's age and gender? http://iccvm.org/2021/papers/S8-2-JCST.pdf
[18] Postnatal gestational age estimation of newborns using Small Sample Deep Learning https://www.sciencedirect.com/science/article/pii/S0262885618301483
[19] Estimating infant age from skull X-ray images using deep learning https://www.researchgate.net/publication/382363550_Estimating_infant_age_from_skull_X-ray_images_using_deep_learning