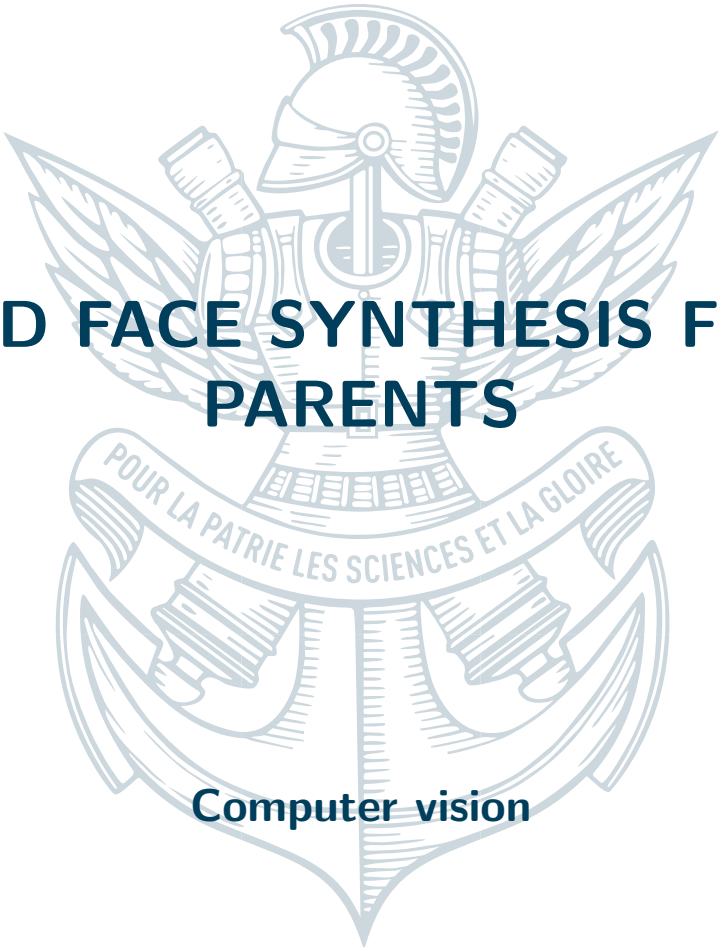

CHILD FACE SYNTHESIS FROM PARENTS

Computer vision

March 25, 2025

Fares Boudelaa, Christelle Clervilsson

CONTENTS

1 INTRODUCTION

Child face synthesis from parents is a fascinating and complex subject in computer vision and deep learning. The goal of this research is to generate the most accurate facial image of a child using images of their parents. With advancements in Generative Adversarial Networks (GANs) and deep learning-based facial analysis, modern approaches leverage large datasets of family images to learn how genetic traits are inherited and predict the most likely child appearance.

This is a challenging task due to the wide variability in traits such as eye shape, hair color, and skin tone, while still maintaining realism and consistency. One of the primary challenges in this domain is the limited availability of high-quality datasets containing parent-child images, as such data is often restricted due to privacy concerns. This task has applications in forensic science, missing person investigations, genealogy studies, and even entertainment.

Driven by curiosity, we decided to explore this subject to see what results we could achieve.

2
BACKGROUND AND RELATED WORK

2.1 Gan & StyleGan

A **Generative Adversarial Network (GAN)** is a class of machine learning frameworks used to generate new data that mimic real-world data. It was introduced in 2014 and is used for applications such as image generation, video generation, and even music creation. It has two main components:

- **Generator**: A neural network that takes random noise as input and generates something that resembles real data.
- **Discriminator**: Another neural network that tries to differentiate between real and fake data.

These two components are trained together, and the ultimate goal is for the generator to become better at producing realistic data, while the discriminator becomes better at detecting fake data. Eventually, the Generator should produce data that is almost indistinguishable from real data by the Discriminator.

StyleGAN, presented in this paper [1], is an advanced generative model developed by NVIDIA researchers, widely used for tasks such as face synthesis. It produces extremely high-fidelity images by improving several key aspects of the original architecture, including stability, image quality, and versatility in generating diverse outputs. StyleGAN enables the generation of highly realistic and diverse images using a style-based generator architecture and advanced training techniques.

The **Encoder4Editing** [2] is a pre-trained neural network designed to encode images in the latent space of StyleGAN, which captures key features of those images. The mathematical structure of e4e's latent space balances the image reconstruction fidelity with semantic editability. The latent space is a 512-dimensional representation ($\mathbb{R}^{512}$) that maps an initial random vector through a learned network, producing a foundational latent vector that can be systematically replicated across 18 layers. Rather than treating each of the 18 layers as completely independent, e4e introduces a unified base vector with small, constrained modifications. This approach is mathematically formulated to encourage consistency across layers while maintaining proximity to the well-behaved W space. A delta-regularization loss further ensures that these offsets remain minimal, effectively guiding the latent representation to retain its desirable editability properties. This layer-wise replication allows for nuanced control over facial features, with different layers capturing distinct levels of detail, from global features in coarse layers to details in fine layers.

2.2 Transformers & Attention

The **Transformer** model was introduced in 2017 in the paper *"Attention is All You Need"* [3], which revolutionized the way we approach sequence data. They are the foundation of models like BERT and GPT. Unlike RNNs and LSTMs, transformers do not process data sequentially, allowing for faster training. The structure consists of Multi-Head Self-Attention, which allows the model to focus on different parts of the input sequence, feedforward networks that are simple fully connected layers that process the output of the attention layers, and positional encoding to determine the order of the tokens because Transformers do not process the data in order.

The **attention** mechanism is central to the Transformer model. It allows the model to focus on the relevant parts of the input sequence while processing each element.

Self-Attention In self-attention, which is the type of attention used in Transformers, each token attends to all other tokens in the sequence to capture the relevant context. It calculates the attention scores between each pair of tokens, which helps the tokens determine how much importance they should give to each other. To compute the attention score, we used the following :

- **Query (Q)**: Represents the word to focus on.
- **Key (K)**: Represents the words to be compared with.
- **Value (V)**: The actual data used to compute the output once attention is applied.

The output of the attention mechanism is a weighted sum of the value vectors, where the weights are the attention scores. To calculate the attention score, we compute the softmax of the dot product of the query and the key, followed by scaling.

Multi-Head Self-Attention Instead of using just one attention mechanism, Multi-Head Self-Attention uses multiple attention mechanisms (heads) in parallel. Each head learns different attention patterns, allowing the model to focus on different parts of the input sequence simultaneously. The outputs of all heads are combined and passed through a linear transformation.

2.3 Related Work

Generative Adversarial Networks (GANs) [4] consist of a generator and a discriminator trained in competition to generate realistic data. Despite their success, early GANs faced challenges such as mode collapse and training instability.

StyleGAN [5] introduced a style-based generator architecture, improving control over image synthesis. StyleGAN2 [1] further improved image quality by reducing artifacts and improving feature modulation, making it a state-of-the-art model for high-resolution image generation.

Transformers [3] revolutionized deep learning by introducing the self-attention mechanism, allowing models to process entire sequences in parallel rather than sequentially like RNNs. This significantly improved efficiency in natural language processing (NLP) and beyond.

3 METHODOLOGY

This section introduces our approach for the synthesis of a face based on the faces of two parents. We use the CelebA-HQ dataset for high-quality face images, Encoder4Editing to project the images onto an information-dense latent space, a learned linear projection for latent blending, a Transformer encoder for feature refinement, and a StyleGAN decoder pre-trained for high-quality face synthesis. In addition, we introduce a specialized loss function to optimize the generation process. For evaluation, we employ the Fréchet Inception Distance (FID) and the Inception Score (IS) to measure the performance of the model.

3.1 Dataset

Since we needed high-resolution images for our project, we sought a dataset that provided high-quality facial images while also ensuring that sensitive or private pictures not were included. This is why we chose the **CelebA-HQ dataset**.

CelebA-HQ is a high resolution version of the original CelebA (CelebFaces Attributes) dataset, specifically designed to improve *face synthesis and image generation tasks*. It was introduced by **Karras et al. (2017)** in their research paper [6]. This dataset is widely used for training **Generative Adversarial Networks** and other deep learning models that require detailed ,high-qualityfacial images.

- Characteristics of the CelebA-HQ Dataset
  - **30,000 images** selected from the original CelebA dataset.
  - Each image is **1024×1024 pixels**, making it suitable for high-resolution tasks.
  - Denoised, enhanced, and high-quality images, ensuring improved clarity and consistency.
  - Facial landmarks & attributes are available, inherited from the original CelebA dataset.
  - Designed for tasks such as face generation, image super-resolution, and attribute-based facial editing.

Unlike many other datasets, CelebA-HQ provides publicly available celebrity images, reducing concerns about privacy while still offering diverse facial characteristics. This makes it a valuable resource for AI-driven facial analysis and synthesis.

3.2 Encoder: Obtaining Latent Space

For child face synthesis, the e4e latent space proves especially appropriate due to its structured, semantically meaningful embedding. The layer-wise control enables sophisticated feature inheritance operations, such as weighted interpolation or advanced latent mixing techniques. While there is a slight trade-off between reconstruction precision and editability, the perceptual quality remains high, ensuring synthesized faces look natural and representative of parental genetic traits.

Moreover, e4e offers practical advantages beyond its high-quality latent representation. By using the pretrained encoder on FFHQ, we can control nuanced details of the generated face, such as age and smile, through provided directional manipulations in the latent space.ay.

Our inputs are two parent images (a father and a mother), and we feed each of them into an encoder. We used **Encoder4Editing** from StyleGAN.

3.3 Latent Blending

We propose three distinct methods for blending parents' faces: a simple average, multiplication, and a learned linear projection. Each approach offers unique advantages and limitations.

The simple average method is the most intuitive, leveraging the linear properties of the latent space. Since latent representations typically encode features like age, eye shape, and facial hair along specific directions, averaging provides a straightforward way to blend parental characteristics. This approach is computationally efficient and captures basic linear relationships between facial features.

Matrix multiplication is a blending technique in which each feature of one latent vector interacts with all features of the other latent vector. This method creates a more general blending of the latents, capturing complex interactions between them. While it can effectively model high-level relationships between inherited features, it may not capture detailed, specific interactions, potentially losing some finer information.

The learned linear projection emerges as the most sophisticated approach. Instead of superficially combining parent embeddings, this method uses a neural network to learn an optimal mapping between parental genetic information. By concatenating the two parent embeddings and processing them through a projection layer, the approach can model complex trait inheritance patterns. The projection can adaptively learn fusion strategies that account for genetic dominance and subtle feature interactions.

While the learned projection offers the most sophisticated blending, it comes with trade-offs. It is computationally more expensive and requires more complex training compared to simpler averaging methods. We initially explored using a multilayer perceptron for blending but found no significant performance improvement over the simple projection technique.

3.4 Refinement: Transformer Encoder with Self-Attention

We implement a Transformer-encoder model designed to capture the complex interactions between parental genetic features while maintaining the rich informational structure of the latent space. By conceptualizing the 18 vectors of the latent space as a sequence, with the first vector representing the overall face and the subsequent vectors capturing progressively finer details, we use the self-attention mechanisms to model genetic inheritance.

Sinusoidal Positional encoding is used in our approach, addressing the importance of vector order by adding critical contextual information to the input tensors. We generate the positional embeddings that enable the model to understand inherent hierarchical structures within the latent space.

The encoder architecture was specifically chosen for the bidirectional self-attention. Unlike transformer decoders that rely on masked self-attention, the encoder considers the entire context simultaneously. This approach generates learned nonlinear relationships that complement the linear projections, creating better representation of parental genetic mixture.

Configured with eight attention heads, two transformer layers, and a 1024-dimensional feedforward network, the encoder is precisely calibrated to capture intricate interactions between facial features. A 0.1 dropout rate helps prevent overfitting, while layer normalization ensures training stability.

3.5 Loss Function

Due to the challenge of the lack of a real dataset containing pairs of parents with their offspring, we must choose a loss function that doesn't rely on labeled data but on our assumptions and the latent space. We have a total loss function that balances critical objectives during latent space blending, ensuring generated faces maintain identity consistency, follow probabilistic priors, and exhibit smooth interpolation between parental characteristics.

The identity consistency loss preserves facial identity by measuring cosine similarity between the blended face and original parental faces using the ArcFace [7] feature extractor. This approach captures nuanced facial attributes beyond pixel-level comparisons while ignoring futile background pixels.

Kullback-Leibler (KL) divergence regularization ensures latent representations follow the standard normal distribution, preventing unrealistic facial variations and avoiding the uncanny valley effect by penalizing vectors that deviate from the learned generative distribution.

The smoothness regularization term addresses interpolation between parental latent vectors, ensuring a natural transition of facial features and stabilizing training.

Each loss component is weighted by a hyperparameter (λ) to balance contributions. With a batch size of 24, the KL divergence hyperparameter is set low ($\lambda = 0.001$), with larger batch sizes requiring higher values.

The total loss function is given by:

$$\mathcal{L} = \lambda_{\text{id}}\mathcal{L}_{\text{id}} + \lambda_{\text{kl}}\mathcal{L}_{\text{kl}} + \lambda_{\text{smooth}}\mathcal{L}_{\text{smooth}}$$

Where:

- Identity Consistency Loss :

$$\mathcal{L}_{\text{id}} = 1 - \frac{1}{2}\left(\frac{F(D(f(z_1, z_2))) \cdot F(D(z_1))}{\|F(D(f(z_1, z_2)))\|\|F(D(z_1))\|} + \frac{F(D(f(z_1, z_2))) \cdot F(D(z_2))}{\|F(D(f(z_1, z_2)))\|\|F(D(z_2))\|}\right)$$

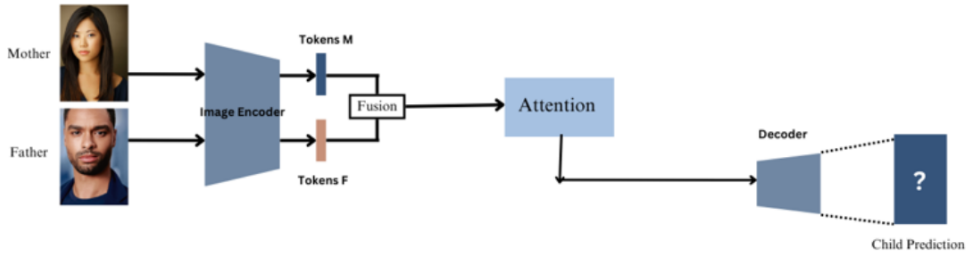

Figure 1: Model structure

- KL Divergence Regularization :

$$\mathcal{L}_{\text{kl}} = D_{KL}(f(z_1, z_2)||\mathcal{N}(0, I))$$

- Smoothness Regularization :

$$\mathcal{L}_{\text{smooth}} = ||f(z_1, z_2) - \alpha z_1 - (1 - \alpha)z_2||_2$$

  Where:

  - $f(z_1, z_2)$ is the output of the model blending of the latent vectors z_1 and z_2.
  - $F(\cdot)$ is the face feature extractor (ArcFace).
  - α is a blending weight, either chosen randomly or set at $\alpha = 0.5$.

4 RESULTS AND ANALYSIS

4.1 Qualitative Results

We present a detailed analysis of offspring generation using two randomly generated parent faces, as illustrated in Figures 2 and 3. The performance from the initial epoch demonstrates promising results, with a notable evolution of facial details observed after five epochs of training.

In the earlier epochs, the synthesized child's face retained direct characteristics from the parents—including parental wrinkles and specific features like the father's teeth. However, subsequent training epochs revealed a significant advancement in the model's capability. The self-attention mechanism demonstrated its sophisticated ability to combine intricate details, generating unique facial characteristics such as personalized wrinkle patterns.

Notably, the model exhibited exceptional performance in capturing eyebrow features from the initial stages. As training progressed, the overall level of detail and nuance in the generated faces continuously improved, showcasing the model's capacity to learn and synthesize complex genetic trait interactions.

This progression highlights the effectiveness of our approach in creating genetically plausible facial representations that go beyond simple linear inheritance, instead capturing the complex, non-deterministic nature of genetic feature transmission.

Figure 2: Parent 1

Figure 3: Parent 2

Figure 4: Evolution of Generated (linear projection model) Offspring Over Training Epochs

4.2 Quantitative results

The implementation of the inception score and Fréchet Inception Distance (FID) follows an incremental computation approach to prevent memory overflow. Rather than storing all images in random access memory, the metrics are updated batch-by-batch using the PyTorch library.

Prior to metric calculation, real and generated images are denormalized, scaling pixel values from the range [-1, 1] to [0, 255] as unsigned 8-bit integers (uint8). The computation proceeds by processing images in batches, which allows for analysis of statistically significant portions of the dataset.

While this batch method enables computational efficiency, we suspect it may negatively influence the metric scores. In our experiment, 10,000 random offspring were generated and compared against the dataset. The results were notably unsatisfactory: Fréchet Inception Distance (FID): Abnormally high, exceeding 15. Inception Score: Remarkably low, falling below 5

Despite these challenging initial results, we observed a consistent decrease in both FID and inception scores throughout the computation process which indicates overall improvement. We only tested on the linear projection model due the high computational cost.

Figure 5: Parent 1

Figure 6: Parent 2

Figure 7: Multiplication blending Result after 1 epoch

5 DISCUSSION AND CONCLUSION

Our research explores child face synthesis using advanced deep learning techniques. We focused on Generative Adversarial Networks (GANs), StyleGAN, and Transformer architectures. By developing a novel approach with sophisticated latent space encoding and transformer-driven refinement, we demonstrated the potential for generating genetically plausible facial representations.

Our methodology integrated Encoder4Editing for precise latent space representation. We used a transformer encoder with self-attention to refine features and developed a carefully designed loss function. This approach balanced identity consistency, distribution regularization, and smooth interpolation between parental traits.

Quantitative analysis initially revealed significant challenges with Fréchet Inception Distance (FID) and Inception Score metrics. However, we observed consistent improvement throughout our computational process. This suggests promising potential for generating nuanced, genetically informed facial syntheses. Our qualitative analysis further supported these findings with satisfying visual results.

The research faced limitations, including the computational complexity of our projection method and the inherent challenges of capturing genetic inheritance. Future work should focus on additional training, refining blending techniques, and developing more sophisticated genetic modeling approaches.

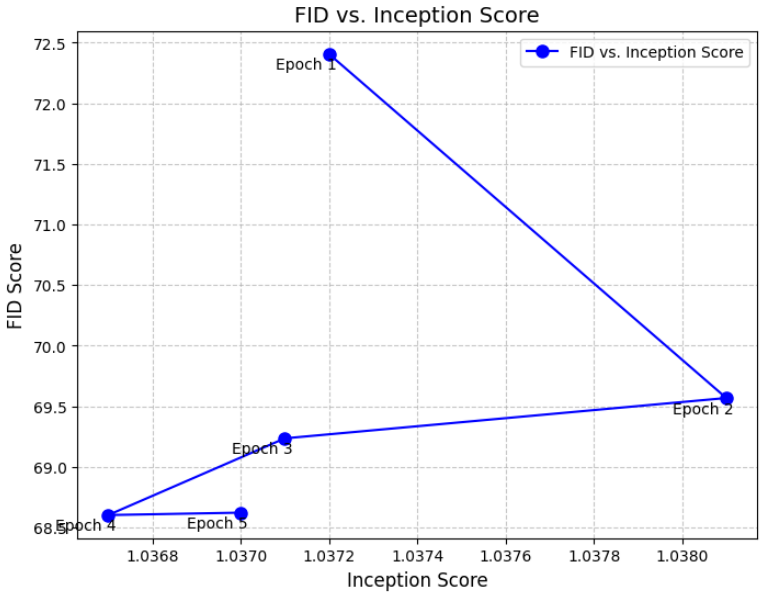

Figure 8: FID vs. Inception Score Over Training Epochs for the linear projection model

REFERENCES

[1] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," 2020. [Online]. Available: https://arxiv.org/abs/1912.04958

[2] O. Tov, Y. Alaluf, Y. Nitzan, O. Patashnik, and D. Cohen-Or, "Designing an encoder for stylegan image manipulation," 2021. [Online]. Available: https://arxiv.org/abs/2102.02766

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023. [Online]. Available: https://arxiv.org/abs/1706.03762

[4] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014. [Online]. Available: https://arxiv.org/abs/1406.2661

[5] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," 2019. [Online]. Available: https://arxiv.org/abs/1812.04948

[6] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," 2018. [Online]. Available: https://arxiv.org/abs/1710.10196

[7] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," 2019. [Online]. Available: https://arxiv.org/abs/1801.07698

6 APPENDIX


Parent 1
Parent 2
Blended