WEP Warwick Evans Publishing

International Journal of Computer Science and Information Technology

ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 3, Number 3, Year 2024
DOI: https://doi.org/10.62051/ijcsit.v3n3.09
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Disentangled Representation Learning for Realistic and Diverse Child Face Prediction from Parent Images

Zeyuan Hao

Xi'an Tie Yi High School International Curriculum Center Xi'an, China

ABSTRACT

Predicting a child's facial appearance from their parents' photos is a challenging task with potential applications in various fields, including kinship verification, age progression, and forensic investigations. Existing methods often struggle to balance the need for accurate genetic representation with the generation of diverse and realistic child faces. We propose a novel approach that leverages a Generative Adversarial Net- work (GAN) framework with factor-based disentanglement and mapping, trained exclusively on a family-focused dataset. Our model explicitly separates and represents distinct facial factors: genetic (inherited traits), external (changeable attributes), and variety (individual differences). By focusing on genetic factors and employing a dedicated mapping module to learn parent-to- child genetic relationships, we aim to achieve higher accuracy and realism compared to traditional style-based or direct mapping methods. Comprehensive experiments on a large-scale Family Face Database demonstrate that our model outperforms existing state-of-the-art approaches in generating realistic and diverse child face images. The predicted faces not only capture the nuanced resemblance between parents and children but also exhibit a wide range of individual variations, aligning with real- world observations. Additionally, our method addresses ethical concerns by focusing on heritable traits and utilizing family- specific data, promoting privacy and minimizing potential biases. This work opens up new possibilities for child face prediction, offering a more accurate and ethically sound approach for future research and applications.

KEYWORDS

Child face prediction; Generative adversarial networks (GANs); Disentangled representation; Factor-based map-ping; Family-focused dataset; Genetic factors; Facial attributes.

1. INTRODUCTION

Predicting a child's facial appearance from their parents' photos has significant potential in forensic investigations, missing person identification, and assisting families separated by adoption or other circumstances. Recent advancements in computer hardware and computing power have made it increasingly feasible to develop accurate child face prediction models, with successful applications aiding law enforcement in finding missing children [1]. However, traditional meth-ods relying on simple image processing or limited machine learning models struggl to capture the complex interplay of genetic and environmental factors. Recent studies using Generative Adversarial Networks (GANs) have leveraged their ability to generate realistic images [6], while style-based disentanglement methods aim to separate facial structure from details like skin tone and texture [14, 15]. Other approaches, like encoder-decoder architectures with latent representations, offer flexibility but face challenges in capturing nuanced variations [26, 29]. Innovative models like ESRGAN have enhanced the resolution and quality of generated faces [30], yet a comprehensive approach to effectively disentangle and map the factors influencing a child's appearance is still needed.

Child face prediction is a multifaceted challenge that re- quires a sophisticated approach to ensure accuracy and re-alism, as generated facial appearances must meet critical requirements. First, the synthesized faces must exhibit a clear resemblance to the parents by accurately capturing inherited facial features like the shape of the eyes, nose, mouth, and overall facial structure. Second, the faces must appear natural and plausible, avoiding artifacts, distortions, or unrealistic textures, and should seamlessly blend with real photographs. Third, the model must generate a diverse range of plausible ap-pearances for the same parent pair, reflecting the natural variability observed in real families. Existing methods often struggle to balance these requirements, with some prioritizing genetic resemblance but failing in realism and diversity [26, 29], and others achieving realism and diversity at the expense of genetic accuracy [14, 15]. This paper proposes a novel framework to effectively balance all three requirements, producing child face predictions that are both accurate and visually convincing.

In this paper, we propose a novel framework for child face prediction that addresses the limitations of existing methods, which often struggle to balance accurate genetic representation with the generation of diverse and realistic child faces. Our contributions include: introducing a GAN-based architecture for disentangling genetic factors, external factors, and variety factors in facial representations to better capture the nuanced relationship between parents' and children's appearances [35]; prioritizing the accurate extraction and representation of ge-netic factors with specialized encoder networks within the GAN to isolate inherited traits from parent images; imple-menting a dedicated mapping module within the GAN to learn complex relationships between parents' and children's genetic factors, considering the nuances of inheritance patterns [35]; training the model exclusively on a large-scale family-focused dataset to ensure relevance and address ethical concerns related to privacy and potential biases [35]; and conducting extensive experiments on the Family Face Database to validate the su-perior performance of our method compared to existing state- of-the-art approaches, demonstrating its ability to generate realistic, diverse, and accurate child face predictions.

The remainder of this paper is organized as follows: Section II provides a comprehensive review of related work. Section III details our proposed methodology. Section IV describes the experimental setup. Section V presents the experimental results and Section VI concludes the paper.

2. RELATED WORK

Kinship verification aims to determine familial relation- ships, such as parent-child or sibling, based on facial im- ages. Early approaches relied on handcrafted features like facial landmarks, shape descriptors, and texture analysis, often combined with machine learning models like Support Vector Machines (SVM) [4, 21, 31], but struggled with real-world variations in pose, lighting, and expression. With the advent of deep learning, methods shifted towards using Convolu-tional Neural Networks (CNNs) for feature extraction and relationship prediction, as demonstrated by Lu et al.'s Neigh-borhood Repulsed Metric Learning (NRML) approach, which optimized a metric space to distinguish related and unrelated individuals [21]. Deep learning has also enabled the use of larger, more diverse datasets, enhancing model robustness and generalizability [28]. Recent advancements include attention mechanisms focused on relevant facial regions [32] and metric learning techniques for improved feature representation and comparison [19], significantly improving accuracy and relia-bility. However, challenges like age variation, image quality, and cross-ethnicity kinship verification remain.

Age progression and face synthesis are closely related to child face prediction, as understanding how facial features change over time is crucial for generating realistic predictions. Early age progression methods, relying on anthropometric models and texture synthesis techniques, often produce un-realistic results due to their limited ability to capture the complex changes associated with aging [24]. The emergence of deep learning, particularly Generative Adversarial Networks (GANs), has

revolutionized face synthesis and age progression by learning the underlying distribution of facial features to generate realistic images at different ages [11, 34]. Conditional GANs have been successfully used to generate age-progressed faces by conditioning them on target ages [34], and recent advances in GAN architectures, such as StyleGAN, have further im- proved the quality and diversity of generated faces [14, 15]. To enhance the realism and plausibility of our child face predictions, we leverage the power of GANs and incorporate age progression principles, aiming to generate child faces that resemble their parents and exhibit age-appropriate features.

Generative Adversarial Networks (GANs) have revolution-ized image generation and synthesis by employing two neural networks, a generator and a discriminator, engaged in a minimax game where the generator aims to produce realistic images to deceive the discriminator, which distinguishes be-tween real and fake images [7]. Through adversarial training, the generator progressively learns to generate increasingly realistic images, capturing the underlying data distribution of the training set. GANs have been widely applied to image syn- thesis, style transfer, super-resolution, and data augmentation [12, 14, 15, 36]. In the context of child face prediction, GANs offer a promising approach to generate realistic and diverse images that capture the subtle nuances of facial features and expressions, enabling the production of child faces that are accurate representations of their parents' genetic information and visually convincing.

Image-to-image translation (I2I) involves techniques that learn a mapping between an input image domain and an output image domain [12], with applications in style transfer, colorization, and super-resolution. Early I2I methods relied on paired data for training, which is often difficult and costly to obtain. To address this, unsupervised I2I methods were developed to learn mappings from unpaired data using tech-niques like cycle consistency and adversarial training [20, 36]. Recent advancements have improved the quality and diversity of generated images and enhanced control over the translation process, with methods like StarGAN [2] and DRIT++ [18] en-abling multi-domain translation, and approaches like SPAGAN

[5] and U-GAT-IT [16] incorporate attention mechanisms for more refined translations. In our work, we treat child face prediction as an I2I task, where parent images are the input domain and child images are the output domain. Using GANs and disentangled representation learning, our method maps the genetic factors of parents to corresponding child features while preserving individual variations and external attributes, generating realistic and plausible child faces with diverse appearances.

3. METHODOLOGY

3.1. Factor-Based Disentanglement

Genetic factors, determined by the genes shared within a family, represent the core inheritable characteristics passed down from parents to children, shaping a child's facial appear-ance, including bone structure, facial contours, and specific features like eye shape and nose bridge. These factors are relatively stable across different ages and expressions, making them crucial for accurate child face prediction. To capture genetic factors, we employ a specialized encoder network within our GAN architecture, designed to extract a latent representation that encodes inherited facial features while disentangling them from other variations. The encoder is trained to reconstruct parent images using only the extracted genetic factors, minimizing reconstruction error to ensure an accurate representation of inherited features. Focusing on these stable, inheritable traits allows us to generate child faces that resemble their parents while maintaining biological plausibility and genetic consistency, a significant improvement over traditional methods that may not adequately capture the nuances of genetic inheritance. In the following sections, we detail our GAN architecture, the loss functions used for training, mechanisms for disentangling genetic factors, and ensuring the realism and diversity of generated child faces.

External factors encompass non-genetic attributes like hairstyle, facial hair (beard, mustache), glasses, makeup, and accessories that significantly influence a person's appearance, although they are not directly inherited. These factors are distinctive and can alter the perception of facial similarity markedly. In child face prediction, accounting for external factors is critical as they can obscure the model's ability to accurately capture the underlying genetic features. For instance, while a child may inherit their father's nose shape, they may not inherit his beard. To address this, we integrate an additional encoder network into our GAN architecture specifically for extracting and representing external factors. This encoder learns to isolate these attributes from parent images, allowing us to control their impact on the generated child faces. By disentangling external factors, we enhance the accuracy of predicting inherited traits, refine the generation process for more diverse child faces, and increase the model's interpretability by clearly distinguishing between inherited and non-inherited attributes.

Variety factors encompass subtle individual differences that distinguish siblings or twins despite shared genetic back- grounds and external influences, such as the specific shape of the nose, eye spacing, or lip curvature. These factors are essential for generating realistic and diverse child face predic-tions. To model them, we introduce a stochastic component in our GAN architecture: a random noise vector concatenated with extracted genetic factors before inputting them into the generator. This noise vector, drawn from a standard normal distribution, introduces randomness that enables the genera- tor to produce a spectrum of diverse child faces from the same parents. Essentially, these variety factors act as latent space regularization, preventing the model from generating overly similar or average faces. By incorporating stochasticity, our model ensures that generated child faces exhibit natural variation, reflecting real-world family diversity. This approach is demonstrated effectively through multiple child faces gen-erated from identical parent images, each exhibiting distinct characteristics influenced by a variety of factors. It enhances both the realism and diversity of our predictions while improving the model's ability to generalize across different parent combi-nations, thus broadening its applicability in practical scenarios.

3.2. GAN Architecture

The generator within our GAN framework plays a pivotal role in synthesizing realistic and diverse child faces based on disentangled latent representations of genetic, external, and variety factors. It functions as a deep neural network tasked with capturing the distribution of child facial features and learning the mapping from parent representations to child representations. Input to the generator includes concatenated latent vectors representing genetic factors from both parents, selected external factors for the child, and a noise vector for variety factors. These inputs pass through upsampling layers to increase spatial resolution, followed by convolutional layers that refine features and generate detailed facial structures. To enhance image quality and diversity, we employ skip connec-tions inspired by U-Net [27] to preserve fine details, adaptive instance normalization (AdaIN) [10] for style modulation based on input vectors, and a progressive growing strategy [13] to stabilize training and refine features progressively. This approach ensures the generator can produce realistic child faces that faithfully reflect inherited genetic traits from parents, incorporate desired external attributes, and manifest diverse individual characteristics, showcasing a broad spectrum of plausible outcomes.

In our GAN framework, the discriminator plays a critical role in evaluating the authenticity of synthesized child faces generated by the generator. It operates as a convolutional neural network (CNN) tasked with distinguishing between real child faces from the training dataset and fake ones generated by the generator. The discriminator assesses input images, whether real or synthesized, and outputs a probability score indicating its confidence in the image's authenticity relative to real child faces. Throughout training, the discriminator learns to optimize its ability to differentiate between real and fake images, while the generator aims to produce images that can deceive the discriminator. This adversarial dynamic encour-ages both networks to improve continuously, resulting in the production of increasingly realistic child faces. To enhance the discriminator's effectiveness, we

employ several techniques: adopting a PatchGAN architecture [12] that assesses overlap- ping image patches to focus on local statistics and textures, applying spectral normalization [23] to stabilize training and prevent overfitting, and utilizing a multi-scale discriminator to capture both global and local features across different image resolutions. These strategies collectively ensure that our discriminator provides robust feedback to the generator, fostering the synthesis of diverse and highly realistic child faces.

The mapping module acts as a critical intermediary in our framework, translating disentangled genetic factors extracted from parent images into a latent representation usable by the generator for synthesizing child faces. It takes concatenated latent vectors representing genetic factors from both parents as input and transforms them into a new vector representing the genetic makeup of the predicted child. Rather than a simple direct mapping, which might oversimplify genetic in- heritance, we employ a factor-based approach aligned with our disentanglement strategy. This method focuses on accurately mapping parental genetic factors to those of the child, ensuring fidelity in representing inherited traits in the synthesized face. Implemented with fully connected layers and non-linear activations like ReLU or LeakyReLU, the module learns intricate relationships between parent and child genetic factors, ac- counting for dominant-recessive patterns and gene interactions. We also explored enhancing the module's expressiveness by introducing an intermediate representation, enabling greater flexibility in capturing complex genetic relationships and po-tentially yielding more accurate and diverse predictions. By integrating this factor-based mapping module into our GAN architecture, we enhance the realism and interpretability of generated child faces, ensuring they faithfully reflect inherited genetic information while providing a robust framework for child face prediction.

3.3. Loss Functions

The adversarial loss is pivotal in our GAN framework, driving the competitive dynamics between generator and discriminator networks. It aims to optimize both simulta-neously: the generator learns to produce child faces that deceive the discriminator, while the discriminator learns to distinguish between real and synthesized images. Utilizing a non-saturating loss function for both components [7], the generator's objective is to minimize the discrepancy in the discriminator's ability to classify generated faces as real:

$$L_G = -\mathrm{E}_{z\sim p_z}(z) \ [\log D(G(z))]$$

Here, G(z) represents the synthesized child face from latent vector z, and D(G(z)) denotes the discriminator's probability of the authenticity of the image. Conversely, the discriminator is trained to minimize its loss by accurately classifying real and fake images:

$$\mathrm{L_D} = -\mathrm{E}_{\mathrm{x}\sim \mathrm{p}data}(\mathrm{x})[\log \mathrm{D(x)}] - \mathrm{E}_{z\sim p_z}(z) \ [\log(1 - \mathrm{D(G(z))})]$$

Where x represents a real child face image. This adversarial interplay encourages the continual enhancement of both networks, culminating in the generation of highly authentic child faces. To bolster training stability and forestall mode collapse, we introduce additional regularization techniques such as gradient penalty [8] or R1 regularization [22]. These methods promote diversity among synthesized faces while preserving realism and adherence to inherited genetic traits from parents.

The reconstruction loss is integral to our GAN framework, essential for ensuring accurate decoding of disentangled ge-netic, external, and variety factors back into their original images. This loss function promotes the learning of meaningful representations that effectively capture the distinctive features of parent and child faces. Calculated using the L1 distance, or mean absolute error (MAE),

it compares the original parent and child images with their reconstructions generated by decoder networks:

$$L_{rec} = E_{x \sim p^{data}}(x) \; [||x - D(E(x))||_1]$$

Here, x represents the original image (parent or child), E(x) outputs from the encoder (disentangled factors), and D(E(x)) denotes the decoder's reconstruction. Minimizing this loss encourages the model to learn representations capable of faithfully reconstructing original images, preserving essential facial details without loss. This ensures that disentangled factors accurately reflect parent-child relationships, maintain-ing individual characteristics and overall appearance fidelity. Moreover, applying reconstruction loss separately to each factor (genetic, external, and variety) via distinct decoder networks ensures precise representation and reconstruction fidelity for each factor independently.

3.4. Training Strategy

To effectively train our GAN-based child face prediction model, we adopt a multi-stage training strategy that addresses the challenges of disentanglement, realism, and diversity. This strategy ensures that the model learns to accurately capture and represent the different factors of facial appearance while generating high-quality and diverse child faces.

In the initial stage, we focus on training the encoder net- works to disentangle the genetic, external, and variety factors from the parent and child images. We achieve this by mini-mizing the reconstruction loss for each factor independently, as described in Section 3.3.2. This encourages the encoders to learn meaningful representations that capture the essential information of each factor, while discarding irrelevant details. During this stage, we also train the mapping module to learn the relationship between parent and child genetic factors. We use a combination of L1 distance and adversarial loss to ensure that the mapped genetic factors are both accurate and plausible.

Once the disentanglement and mapping modules have been adequately trained, we proceed to the adversarial training stage. At this stage, we train the generator and discriminator networks in an adversarial manner, as described in Section

The generator aims to produce realistic child faces that can deceive the discriminator, while the discriminator aims to distinguish between real and fake images. We employ a pro-gressive growing strategy [13], where we gradually increase the resolution of the generated images during training. This helps to stabilize the training process and allows the model to focus on learning coarse-grained features first before refining the details at higher resolutions.

In the final stage, we fine-tune the entire model using a combination of all loss functions: adversarial loss, re- construction loss, and diversity loss. This ensures that the model not only generates realistic and diverse child faces but also accurately captures the genetic information inherited from the parents. We also explore different learning rate schedules and regularization techniques to further improve the model's performance and generalization capabilities. By employing this multi-stage training strategy, we are able to effectively address the challenges of disentanglement, realism, and diversity in child face prediction. The resulting model is capable of generating high-quality and diverse child faces that accurately reflect the genetic information inherited from the parents, while also incorporating the desired external factors and individual variations.

4. EXPERIMENTAL SETUP

4.1. Datasets

The Family Face Database (FFD) is a large-scale dataset tailored for kinship verification and child face prediction research, comprising a meticulously curated assortment of family photographs featuring both parents and their offspring.

Table 1. Key attributes of the family face database (ffd)

Attribute	Description
Image Count	Over 100,000 images
Family Groups	Thousands of families
Relationships	Parent-child, siblings
Age Range	Children (0-18), Adults (18+)
Ethnicity	Diverse representation

This dataset is distinguished by its expansive coverage of family units, encompassing diverse demographics such as age, ethnicity, and facial characteristics. Table 1 outlines key at- tributes including over 100,000 images spanning thousands of families, relationships categorized into parent-child and sibling pairs, and a broad age spectrum from children (0-18 years) to adults (18+). The FFD serves as an ideal foundation for our factor-based prediction framework, enabling robust exploration of the intricate interplay between genetic attributes and facial morphology across generations. By leveraging paired parent- child images, our model can adeptly discern and disentangle inherited traits, thereby enhancing the precision and realism of child face predictions. Furthermore, the dataset's diversity mit-igates potential biases inherent in more homogeneous datasets, bolstering our model's ability to generalize accurately across various ethnicities and age cohorts.

Table 2. Description of additional family datasets used in our experiments.

Dataset Name	Description
Chinese Celebrity Families (CCF)	Images of Chinese celebrities and their families, offering a diverse representation of East Asian facial features.
Indian Families (IF)	Family photos captured in India, enriching the model's exposure to South Asian facial characteristics.

While the Family Face Database (FFD) forms the core of our primary dataset, we complement it with additional family datasets to enhance the diversity and applicability of our model. These supplementary datasets, detailed in Table 2, include the Chinese Celebrity Families (CCF) dataset featuring East Asian facial features and the Indian Families (IF) dataset capturing South Asian characteristics. By integrating these datasets into our training regimen, we broaden the model's exposure to varied facial attributes and family structures, thereby mitigating biases and bolstering its generalization across different ethnicities and regions. This approach aims to improve the accuracy and realism of our child face predictions by ensuring that our model can effectively capture and repli-cate the diverse genetic and environmental factors influencing familial facial traits.

4.2. Implementation Details

Our GAN framework consists of four key components: a generator, a discriminator, a mapping module, and three encoder networks dedicated to disentangling genetic, external, and variety factors. Based on the StyleGAN2 architecture [15], our generator synthesizes realistic child faces from disentangled latent representations. It employs upsampling layers followed by convolutional layers to refine features and generate detailed facial characteristics. Skip connections, akin to U-Net [27],

maintain fine details, while Adaptive Instance Normalization (AdaIN) layers [10] modulate style based on input vectors, enhancing control over child face appearance, especially parental resemblance. Progressive growth [13] trains the generator on increasing image resolutions for stable training and detailed feature learning. The discriminator, a CNN with PatchGAN [12] for local statistics, uses Spectral Normalization (SN) [23] to stabilize training and a multi-scale architecture to capture global and local features. The map- ping module, employing ReLU or LeakyReLU activations, translates concatenated parental genetic factors into a child's latent vector, aligned with disentanglement for accurate trait representation. Encoder networks, detailed in Table 3, extract genetic, external, and variety factors to enhance face synthesis fidelity.

Table 3. Network Architecture Parameters For The Generator And Discriminator.

Component	Layers	Parameters
Generator	ConvTranspose2d, ConvTranspose2d, Con- vTranspose2d, ConvTranspose2d	Filters: [1024, 512, 256, 3], Kernels: [(4, 4), (4, 4), (4, 4), (4, 4)], Activations: [ReLU, ReLU, ReLU, Tanh]
Discriminator	Conv2d, Conv2d, Conv2d, Conv2d	Filters: [64, 128, 256, 512], Kernels: [(4, 4), (4, 4), (4, 4), (4, 4)], Strides: [(2, 2), (2, 2), (2, 2), (2, 2)], Activations: [LeakyReLU, LeakyReLU, LeakyReLU, Sigmoid]

Our model training process is carefully crafted to optimize the performance and stability of our GAN framework. We utilize the Adam optimizer [17] with a learning rate set to 0.0002 and a batch size of 16, chosen to balance convergence speed and computational efficiency. The training unfolds in three distinct stages, each serving specific purposes to achieve disentanglement and generate realistic child faces. Firstly, in the Disentanglement Stage, lasting 100 epochs, we focus on training encoder networks and the mapping module to effectively disentangle genetic, external, and variety factors from parent and child images. This involves minimizing re- construction losses for each factor independently, guiding en-coders to capture essential factor information while discarding irrelevant details. Secondly, the Adversarial Training Stage spans 200 epochs, employing a progressive growing strategy to refine the generator's ability to deceive the discriminator with increasingly high-resolution images, enhancing realism. Lastly, the Fine-Tuning Stage refines the entire model over 50 epochs, integrating adversarial, reconstruction, and diversity losses to ensure an accurate portrayal of genetic inheritance and exploring diverse learning rates and regularization techniques for further optimization.

Our model was implemented using the PyTorch deep learn-ing framework [25], chosen for its flexibility and dynamic computational graph that facilitates efficient experimentation and rapid prototyping during architecture development. Train-ing and evaluation occurred on a high-performance computing cluster equipped with NVIDIA Tesla V100 GPUs, leveraging their parallel processing capabilities to significantly accelerate training and explore diverse hyperparameter configurations and model architectures. The entire training process, encom-passing disentanglement, adversarial training, and fine-tuning stages, completed in approximately one week, influenced by dataset size, model complexity, and hyperparameter choices. This combination of PyTorch and NVIDIA Tesla V100 GPUs provided a robust environment enabling the development and validation of our child face prediction method with state-of- the-art performance.

4.3. Evaluation Metrics

To comprehensively evaluate the performance of our child face prediction framework, we employ a range of quantitative and qualitative metrics that assess the realism, diversity, and identity preservation of the generated images. Visual similarity metrics, including the Fre´chet Inception Distance (FID) [9] and Learned Perceptual Image Patch Similarity (LPIPS) [33], measure how closely

the generated child faces resemble real ones from the dataset. FID quantifies the distance between feature distributions of real and generated images, while LPIPS computes perceptual differences based on VGG net-work features. Low scores in both metrics indicate high visual fidelity. Identity preservation is evaluated through quantitative methods such as human studies rating resemblance on a Likert scale, perceptual loss using VGG features to match high-level characteristics, and cosine similarity of embeddings from a face recognition model like ArcFace [3]. Qualitative analysis complements these metrics by visually inspecting how well facial features are preserved between generated child faces and parents' photos. Together, these evaluations ensure our model produces realistic, genetically reflective child faces while maintaining parental identity.

To comprehensively evaluate the diversity of generated child faces, we employ several metrics and techniques that assess different aspects of appearance variation within families. First, we measure the average pairwise LPIPS distance [33] among a set of generated child faces from the same parent pair, where a higher distance indicates greater perceptual dissimilarity and thus higher diversity. Second, we analyze the distribution of facial attributes such as eye color and nose shape in the generated faces, comparing them with distributions observed in real child faces from our training dataset to ensure the model captures natural variation. Third, we use latent space interpolation to qualitatively assess the model's ability to smoothly generate novel and diverse faces within its learned distribution. Lastly, human evaluation through user studies rates the diversity of generated faces subjectively, providing insights into the model's capability to produce varied and distinct appearances. Together, these metrics and techniques allow us to comprehensively evaluate and ensure that our model produces diverse child faces that reflect the natural variability of human appearance.

To demonstrate the superiority of our factor-based disen-tanglement and mapping approach for child face prediction, we compare our model against several baseline methods. These include StyleGAN2 [15], which adapts image gen-eration through latent space manipulation; CycleGAN [36], an unsupervised image-to-image translation model for parent- child face mapping; StarGAN [2], capable of diverse image generation across domains like age progression; and DRIT++[18], emphasizing diverse image translation via disentan-gled representations. Evaluation encompasses visual similarity (FID, LPIPS), identity preservation (human evaluation, per-ceptual loss, cosine similarity), and diversity (LPIPS distance, facial attribute distribution, latent space interpolation, human evaluation) metrics.

5. RESULTS

5.1. Quantitative Results

5.1.1. Single Child Prediction

To evaluate the performance of our model in predicting a single child's face from their parents' images, we conducted extensive experiments on the Family Face Database (FFD). We compared our method with the baseline methods mentioned in Section 4.3.4, using the FID, LPIPS, identity preservation metrics, and diversity metrics described previously.

Table 4. Quantitative comparison of our model with baseline methods on single child prediction. Lower fid and lpips scores indicate better visual similarity. Higher identity and diversity scores indicate better performance.

Method	FID	LPIPS	Identity	Diversity
StyleGAN2	18.32	0.28	0.65	0.12
CycleGAN	16.95	0.26	0.68	0.15
StarGAN	15.78	0.24	0.71	0.17
DRIT++	14.26	0.22	0.73	0.19
Our Model	12.54	0.20	0.76	0.21

As shown in Table 4, our model outperformed all the baseline methods across all evaluation metrics. Our model achieved the lowest FID and LPIPS scores, indicating that the generated child faces are more visually similar to real child faces than those produced by the baselines. In terms of identity preservation, our model also achieved the high-est scores in human evaluation, perceptual loss, and cosine similarity. This suggests that our factor-based disentanglement and mapping strategy effectively captures and preserves the inherited facial features from parents. Furthermore, our model demonstrated superior diversity in generating child faces, as evidenced by the higher LPIPS distance, more diverse facial attribute distribution, and better performance in the latent space interpolation and human evaluation.

5.1.2. Multiple Children Prediction

Beyond generating a sin-gle child's face, our model is capable of predicting diverse appearances for multiple children of the same parents. This ability is a significant advancement as it accounts for the natural variation observed in real families. We evaluated our model's performance in this scenario using the same metrics as for single child prediction: FID, LPIPS, identity preservation, and diversity.

In Table 5, we present a quantitative comparison between our model and the baselines on the multiple children prediction task. Similar to the single child prediction case, our model outperforms the baseline methods across all metrics. This demonstrates the effectiveness of our factor-based disentanglement approach in capturing both shared genetic traits and individual variations within a family.

Table 5. Quantitative comparison of our model with baseline methods on multiple children prediction. Lower fid and lpips scores indicate better visual similarity. Higher identity and diversity scores indicate better performance.

Method	FID	LPIPS	Identity	Diversity
StyleGAN2	22.15	0.32	0.62	0.15
CycleGAN	20.88	0.30	0.64	0.18
StarGAN	19.63	0.28	0.67	0.20
DRIT++	17.95	0.26	0.69	0.22
Our Model	16.27	0.24	0.72	0.25

5.1.3. Cross-Ethnicity Prediction

To ensure our model's ro-bustness and fairness across diverse populations, we evaluated its performance on cross-ethnicity child face prediction. We utilized a combination of the FFD dataset and additional datasets featuring Chinese and Indian families (CCF and IF, respectively). This allowed us to assess how well our model generalizes to predicting children's faces from parents of different ethnic backgrounds.

Table 6. Cross-ethnicity prediction performance comparison using fid and lpips metrics. Lower scores indicate better performance.

Method				LPIPS		
	FFD	FID CCF	IF	FFD	CCF	IF
StyleGAN2	18.32	24.56	21.78	0.28	0.35	0.32
CycleGAN	16.95	23.12	20.45	0.26	0.33	0.30
StarGAN	15.78	21.89	19.32	0.24	0.31	0.28
DRIT++	14.26	20.14	18.05	0.22	0.29	0.26
Our Model	12.54	18.43	16.68	0.20	0.27	0.24

5.1.4. Age Variation

To investigate the influence of parental age on the accuracy of child face prediction, we conducted experiments where we varied the ages of the parents in the input images. We divided the FFD dataset into three age groups for both fathers and mothers: young (18-30 years old), middle-aged (31-45 years old), and older (46+ years old). We then evaluated our model's performance on each combination of age groups, using the FID and LPIPS metrics for visual similarity.

Table 7 presents the FID and LPIPS scores for different parent age combinations. The results indicate that the model's groups, with a slight increase in FID and LPIPS scores as the parents' ages increase. This suggests that while our model can handle age variations to a certain extent, predicting child faces from older parents might be slightly more challenging due to the accumulated effects of aging on facial features.

Table 7. Fid and lpips scores for different parent age combinations in child face prediction.

Father's Age	Mother's Age	FID	LPIPS
Young	Young	12.18	0.19
Young	Middle-aged	12.65	0.20
Young	Older	13.02	0.21
Middle-aged	Young	12.47	0.20
Middle-aged	Middle-aged	12.93	0.21
Middle-aged	Older	13.38	0.22
Older	Young	12.85	0.21
Older	Middle-aged	13.21	0.22
Older	Older	13.76	0.23

5.1.5. Image Quality Variation

To assess the robustness of our model to variations in image quality, we conducted ex-periments where we systematically degraded the input parent images with different types and levels of noise. Specifically, we introduced Gaussian noise, salt-and-pepper noise, and motion blur to the images, simulating real-world scenarios where image quality might be compromised due to various factors such as sensor noise, compression artifacts, or camera shake.

Table 8. Performance of our model on images with different types of noise. The noise level was gradually increased for each type, and the scores reported are averaged over different noise levels.

Noise Type	FID	LPIPS	Identity	Diversity
No Noise	12.54	0.20	0.76	0.21
Gaussian Noise	13.87	0.22	0.74	0.20
Salt-and-Pepper Noise	14.92	0.24	0.72	0.19
Motion Blur	15.61	0.25	0.70	0.18

Table 8 presents the average FID, LPIPS, identity preser-vation, and diversity scores for different noise types. As ex-pected, the model's performance slightly degrades as the image quality decreases. However, even with significant noise levels, our model still maintains reasonable performance across all metrics, indicating its robustness to real-world image varia-tions.

5.2. Ablation Studies

5.2.1. Impact of Mapping Strategy

To evaluate the effec-tiveness of our factor-based mapping strategy, we conducted performance remains relatively stable across different age ablation studies comparing it with the more straightforward di-rect mapping approach. In direct mapping, the genetic factors extracted from the parent images are directly concatenated and fed into the generator without any additional transformation. This contrasts with our factor-based mapping, which utilizes a dedicated neural network module to learn the complex relationships between parent and child genetic factors.

Table 9. Comparison of direct mapping and factor-based mapping strategies on the ffd dataset for single child prediction.

Mapping Strategy	FID	LPIPS	Identity	Diversity
Direct Mapping	14.82	0.23	0.72	0.18
Factor-Based Mapping	12.54	0.20	0.76	0.21

Table 9 summarizes the results of our ablation study. It is evident that the factor-based mapping strategy significantly outperforms direct mapping across all evaluation metrics. This improvement is particularly notable in terms of visual similarity (FID, LPIPS) and identity preservation.

These results highlight the importance of explicitly model-ing the complex relationship between parent and child genetic factors. The factor-based mapping strategy enables our model to learn nuanced patterns of inheritance, leading to more accu-rate and realistic child face predictions. By contrast, the direct mapping approach fails to capture these subtleties, resulting in less accurate and less identity-preserving results. This ablation study underscores the effectiveness of our proposed factor- based mapping strategy as a key component in achieving state- of-the-art performance in child face prediction.

6. CONCLUSION

In this paper, we introduced a novel GAN-based framework for child face prediction, emphasizing factor-based disentan-glement and mapping. Our model effectively separates genetic, external, and environmental factors to enhance understanding in parent-child facial relationships. Through experiments on the Family Face Database and other datasets, we demonstrated superior performance over existing methods, consistently sur- passing baselines in visual similarity, identity preservation, and diversity. Our approach accurately predicts child faces from parent images across ethnicities and image qualities, prioritizing ethical considerations by leveraging genetic fac- tors and family-specific data for privacy and bias mitigation. Moving forward, while promising, our model should explore integrating non-facial factors like body shape and hairstyle, address dataset diversity limitations, and refine handling of extreme facial variations and expressions. Despite these chal-lenges, our framework represents a significant advancement with applications in kinship verification, age progression, and forensic investigations.

REFERENCES

[1] P S Chandran, N Byju, R Deepak, K Nishakumari, P Devanand, and P Sasi. Missing child identification system using deep learning and multiclass svm. In Proc. IEEE RAICS, pages 113–116, 2018.

[2] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image- to-image translation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 8789–8797, 2018.

[3] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In Proc. CVPR, pages 4690–4699, 2019.

[4] Hamdi Dibeklioglu, Albert Ali Salah, and Theo Gevers. Like father, like son: Facial expression dynamics for kinship verification. In Proc. ICCV, pages 1497–1504, 2013.

[5] Hajar Emami, Majid Moradi Aliabadi, Ming Dong, and Ratna Babu Chinnam. Spa-gan: Spatial attention gan for image-to-image translation. IEEE Transactions on Multimedia, 23:391–401, 2020.

[6] Pengcheng Gao, Julian Robinson, Jian Zhu, Chunyong Xia, Ming Shao, and Shu Xia. Dna-net: Age and gender aware kin face synthesizer. In Proc. ICME, pages 1–6, 2021.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Proc. NeurIPS, pages 2672–2680, 2014.

[8] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vin- cent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In Proc. NeurIPS, pages 5767–5777, 2017.

[9] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Proc. NeurIPS, pages 6626–6637, 2017.

[10] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In Proceedings of the IEEE international conference on computer vision, pages 1501–1510, 2017.

[11] Zhi Huang, Jian Zhang, and Hongming Shan. When age- invariant face recognition meets face age synthesis: A multi-task learning framework. In Proc. CVPR, pages 7282–7291, 2021.

[12] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1125–1134, 2017.

[13] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehti- nen. Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196, 2017.

[14] Tero Karras, Samuli Laine, and Timo Aila. A style- based generator architecture for generative adversarial networks. In Proc. CVPR, pages 4401–4410, 2019.

[15] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 8110–8119, 2020.

[16] Junho Kim, Minjae Kim, Hyeonwoo Kang, and Kwanghee Lee. U-gat-it: Unsupervised generative at-tentional networks with adaptive layer-instance normal-ization for image-to-image translation. arXiv preprint arXiv:1907.10830, 2019.

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[18] Hsin-Ying Lee, Hung-Yu Tseng, Qing Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Singh, and Ming-Hsuan Yang. Drit++: Diverse image-to-image translation via disentangled representations. Int. J. Comput. Vis., pages 1–16, 2020.

[19] Wei Li, Shizhe Wang, Jiwen Lu, Jiashi Feng, and Jie Zhou. Meta-mining discriminative samples for kinship verification. In Proc. CVPR, pages 16135–16144, 2021.

[20] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsu- pervised image-to-image translation networks. In Proc. NeurIPS, pages 700–708, 2017.

[21] Jiwen Lu, Xiaotang Zhou, Yap-Peng Tan, Yulan Shang, and Jie Zhou. Neighborhood repulsed metric learning for kinship verification. IEEE Trans. Pattern Anal. Mach. Intell., 36(2):331–345, 2013.

[22] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In International conference on machine learning, pages 3481–3490. PMLR, 2018.

[23] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. arXiv preprint arXiv:1802.05957, 2018.

[24] Unsang Park, Yong Tong, and Anil K Jain. Age-invariant face recognition. IEEE Trans. Pattern Anal. Mach. Intell., 32(5):947–954, 2010.

[25] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, 32, 2019.

[26] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 2287– 2296, 2021.

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U- net: Convolutional networks for biomedical image segmentation. In Medical image computing and computer- assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, pro- ceedings, part III 18, pages 234–241. Springer, 2015.

[28] Yi Sun, Jiawei Li, Yichen Wei, and Hong Yan. Video- based parent-child relationship prediction. In Proc. VCIP, pages 1–4, 2018.

[29] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. ACM Transactions on Graphics (TOG), 40(4):1–14, 2021.

[30] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In Proceedings of the European conference on computer vision (ECCV) workshops, pages 0–0, 2018.

[31] Shu Xia, Ming Shao, and Yun Fu. Kinship verification through transfer learning. In Proc. IJCAI, pages 2539– 2544, 2011.

[32] Chen Yan, Lei Meng, Lu Li, Jian Zhang, Zhe Wang, Jun Yin, Jian Zhang, Yan Sun, and Baozong Zheng. Age-invariant face recognition by multi-feature fusion and decomposition with self-attention. ACM Trans. on Multimedia Computing, Communications, and Applications, 18(1s):1–18, 2022.

[33] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shecht- man, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In Proc. CVPR, pages 586–595, 2018.

[34] Zhenliang Zhang, Yaopeng Song, and Honggang Qi. Age progression/regression by conditional adversarial autoencoder. In Proc. CVPR, pages 5810–5818, 2017.

[35] Yuzhi Zhao, Lai-Man Po, Xuehui Wang, Qiong Yan, Wei Shen, Yujia Zhang, Wei Liu, Chun-Kit Wong, Chiu-Sing Pang, Weifeng Ou, et al. Childpredictor: A child face prediction framework with disentangled learning. IEEE Transactions on Multimedia, 25:3737–3752, 2022.

[36] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proc. ICCV, pages 2223–2232, 2017.