HEREDITY-AWARE CHILD FACE IMAGE GENERATION WITH LATENT SPACE DISENTANGLEMENT

Xiao Cui, Wengang Zhou & Houqiang Li
University of Science and Technology of China
cuixiao2001@mail.ustc.edu.cn,{zhwg,lihq}@ustc.edu.cn

ABSTRACT

In this paper, we propose ChildGAN to generate a child's face image according to the images of parents with heredity prior. The main idea is to disentangle the latent space of a pre-trained generation model and precisely control the face attributes of child images with clear semantics. We use distances between face landmarks as pseudo labels so as to avoid using external labels. By calculating the gradient of latent vectors to pseudo labels, we figure out the most influential semantic vectors of the corresponding face attributes. Then we disentangle the semantic vectors in three aspects: adding a weight factor in the calculating process, working on the proper resolution layers, and using Schmidt orthogonalization to orthogonalize these vectors. Finally, we fuse the latent vectors of the parents by leveraging the disentangled semantic vectors under the guidance of biological genetic laws.

1 INTRODUCTION

Child image generation aims at synthesizing child face image given the images of parents. This is a challenging task since the generated child face should not only resemble the parents, but inherit the attributes following the known genetic laws. Besides, the children born to the same parents may look quite different, which means there is no unique solution to the problem of child image generation.

There are only a few works studying the child image generation problem. KinshipGAN (Ozkan & Ozkan, 2018) uses a deep face network to generate a child's face based on one-to-one relationship. DNA-Net (Gao et al., 2021) and ChildNet (Pernuš et al., 2023) proposes to use a deep generative Conditional Adversarial Autoencoder for this task. Although some success has been achieved, those methods suffer three non-trivial issues. First, the generated images are usually blur and of low quality. Second, those methods cannot explicitly control the facial attributes in the generated faces, which significantly limits their application scenarios. Third, they ignore the inheritance law from genetic basis. For instance, thin upper lip is controlled by a dominant gene and is very likely to be inherited. Without considering such prior, the generated child images may fail to reflect the inherited attributes.

In this paper, we propose a new framework, *i.e.,* ChildGAN, to generate the face image of the child according to the parents' images under the guidance of genetic laws. To ensure the generated images are of high quality, we leverage a pretrained StyleGAN (Karras et al., 2019) generator and conduct macro fusion. To explicitly control the facial attributes in the generated faces, we identify disentangled semantic directions in the latent space. Our semantic learning method is based on gradient estimation from a large number of samples as well as irrelevant factor reweighting. It finds the important and decoupled semantic vectors in the latent space without the need for manual labels. To the inheritance law from genetic basis, we let the child to inherit attributes of the parents in a micro way under the guidance of genetic laws. This approach enables the zero-shot generation of child images, without training on specific datasets.

2 METHODS

The framework of the proposed ChildGAN is shown in Figure 1 . The face images of the parents are first embeded to the latent space of StyleGAN. Then after some preprocessing, the latent codes of the

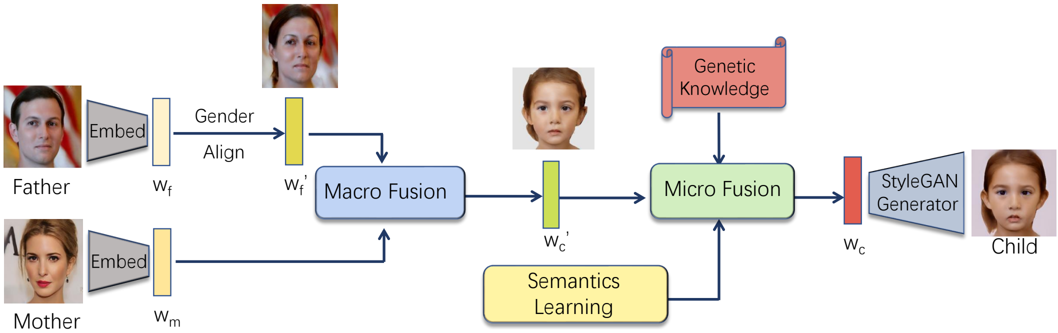

Figure 1: The flowchart of our generation process.

parents are mixed through Macro Fusion. We propose an effective method to identify disentangled semantic directions in the latent space, which allow the child to inherit attributes of the parents in a micro way under the guidance of genetic laws. Finally, the child image is generated from the child's latent code by a pre-tained StyleGAN generator. The details of each component are given in the Appendix A.1 to A.4.

Our methodology employs the extended $\mathbb{W}$ space in StyleGAN, namely $\mathbb{W}+$ space, which is a concatenation of 18 different 512-dimensional $\mathbf{w}$ vectors, to carry out our experiment. We use 10,000 images generated randomly by the pre-trained StyleGAN to learn the semantic vectors. For each face image, we get the pseudo labels for the attributes, such as the size of the eyes, the size of the nose, and the thickness of the lip, by first detecting the face landmarks in the image and then computing the distances between corresponding landmarks pair.

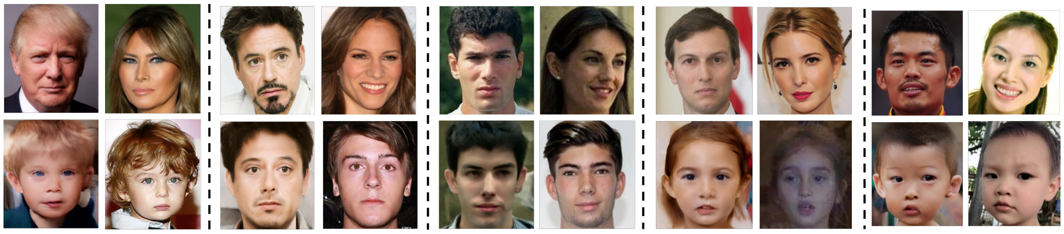

Figure 2: Examples of children images generated by the proposed method. In each group, the images in the first row represent the parents, and the second row displays the child generated (left) alongside the real image of their child (right).

3 Experiments

Table 1: Kinship verification scores.

Method	Acc (E=10)	Acc (E=20)
DNA-Net	0.541	0.583
ChildNet	0.602	0.613
Ours	**0.739**	**0.870**

Table 2: The average rank of different approaches in user study.

	DNA-Net	ChildNet	Ours
Avg. rank	2.40	2.24	**1.36**

Figure 2 presents a subset of our generated results. It is evident that the facial attributes of the generated children are blends of those of the corresponding parents, with a possible bias toward one parent due to genetic factors. For an objective evaluation, we trained a kinship verification network using the Families In the Wild (FIW) dataset (Robinson et al., 2018). This network was employed to verify the parent-child resemblance, using models trained for 10 and 20 epochs. The results of this automated evaluation are presented in Table 1. Additionally, we undertook a human evaluation involving 30 individuals from varied backgrounds, who assessed each image based on its realism and resemblance to the purported parents. Each participant responded to 30 questions, and the average rankings are compiled in Table 2. We conduct all the objective evaluations on the FIW dataset. The outcomes of these evaluations suggest that the ChildGAN effectively simulates kinship relationships, closely mirroring those in real-life parent-child pairs.

URM Statement

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2024 Tiny Papers Track.

Ethics Statement

In developing ChildGAN, a technology designed to generate child face images based on parental images with heredity considerations, we have carefully addressed ethical implications and potential biases, particularly concerning privacy, consent, and fairness. Acknowledging the sensitivity of using generative models for human images, we have implemented robust measures to mitigate biases inherent in our training datasets and ensured that our work does not support or endorse discriminatory practices or applications, such as partner selection based on generated images. Our commitment to ethical research practices includes thorough consideration of human subjects, data privacy, bias mitigation, and the transparent disclosure of any potential conflicts of interest, aiming to uphold the highest standards of research integrity and societal impact.

References

Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (ToG)*, 2021.

Mohamed Elgharib Ayush Tewari, Florian Bernard Gaurav Bharaj, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, and Christian Theobalt. StyleRig: Rigging StyleGAN for 3D control over portrait images. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 6141–6150, 2020.

Pengyu Gao, Joseph Robinson, Jiaxuan Zhu, Chao Xia, MIng Shao, and Siyu Xia. Dna-net: Age and gender aware kin face synthesizer. In *IEEE International Conference on Multimedia and Expo*, pp. 1–6, 2021.

G Ainsworth Harrison. The measurement and inheritance of skin colour in man. *The Eugenics review*, 49(2):73, 1957.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

Heyi Li, Jinlong Liu, Xinyu Zhang, Yunzhi Bai, Huayan Wang, and Klaus Mueller. Transforming the latent space of stylegan for real face editing. *The Visual Computer*, pp. 1–16, 2023.

Hongyu Liu, Yibing Song, and Qifeng Chen. Delving stylegan inversion for image editing: A foundation latent space viewpoint. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 10072–10082, 2023a.

Zhian Liu, Maomao Li, Yong Zhang, Cairong Wang, Qi Zhang, Jue Wang, and Yongwei Nie. Fine-grained face swapping via regional gan inversion. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 8578–8587, 2023b.

Victor A McKusick. *Mendelian inheritance in man: catalogs of autosomal dominant, autosomal recessive, and X-linked phenotypes*. Elsevier, 2014.

Dorothy Osborn. Inheritance of baldness: Various patterns due to heredity and sometimes present at birth—a sex-limited character—dominant in man—women not bald unless they inherit tendency from both parents. *Journal of Heredity*, 7(8):347–355, 1916.

Savas Ozkan and Akin Ozkan. KinshipGAN: Synthesizing of kinship faces from family photos by regularizing a deep face network. In *International Conference on Image Processing*, pp. 2142–2146, 2018.

Hamza Pehlivan, Yusuf Dalva, and Aysegul Dundar. Styleres: Transforming the residuals for real image editing with stylegan. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 1828–1837, 2023.

Martin Pernuš, Mansi Bhatnagar, Badr Samad, Divyanshu Singh, Peter Peer, Vitomir Štruc, and Simon Dobrišek. Childnet: Structural kinship face synthesis model with appearance control mechanisms. *IEEE Access*, 2023.

Joseph P Robinson, Ming Shao, Yue Wu, Hongfu Liu, Timothy Gillis, and Yun Fu. Visual kinship recognition of families in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(11):2624–2637, 2018.

Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of GANs for semantic face editing. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 9243–9252, 2020.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11), 2008.

Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. Gan inversion: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 3121–3138, 2022.

A Appendix

In the appendix, we first introduce the method of macro fusion to achieve the inheritance of macro and relatively rough characteristics. Then we discuss the scientific knowledge for child generation. Next, we introduce the extraction and decoupling of important semantics. After that, we present how to use the genetic knowledge and orthogonal semantic vectors for micro fusion. Finally, we present more evaluation results of our methods.

A.1 Details of Macro Fusion

In order to generate a child's image, we start by making a rough mix of the parents' faces to get a preliminary image of the child. We call this process *macro fusion*. Before our fusion process, we map the parents' images to vectors in the latent space, which called latent codes. Given an image of the father, we first crop and align the face in it. Then we do GAN inversion (Xia et al., 2022; Liu et al., 2023b;a), finding the optimal latent code $\mathbf{w}_f$ by minimizing the reconstruction loss between the image generated from $\mathbf{w}_f$ and the real image. The latent code $\mathbf{w}_m$ for the mother is obtained in the same way. To produce a child with a specific gender and reduce the background artifacts, we change the gender character of one parent by moving the corresponding latent code along a pre-learned orientation in the latent space which mainly controls the gender attribute. That is to say, if we want the genarated child to be a girl, we need to move the father's latent code $\mathbf{w}_f$ forward in this orientation (about 2 units in length as that's the average difference between the latent codes of men and women in this orientation) , and if we want the child to be a boy, we need to move the mother's latent code $\mathbf{w}_m$ backward in this orientation. After this adjustment, we still refer to the latent codes of the father and mother as $\mathbf{w}_f$ and $\mathbf{w}_m$.

There are two alternatives in macro fusion. First, as a simple method, we can use linear combination: $\mathbf{w}'_c = (1-\lambda)\,\mathbf{w}_f + \lambda\mathbf{w}_m$, where λ is a parameter between 0 and 1. In this way, every resolution layer of the child will be a mixture of the parents. Alternatively, for each resolution layer of $\mathbf{w}'_c$, we can take the corresponding dimensions of $\mathbf{w}_f$ or $\mathbf{w}_m$ respectively as the value of this resolution layer of $\mathbf{w}'_c$. For example, if we want rough features such as posture, hairstyle, and facial contour to be inherited from the father, while more subtle features such as facial components are inherited from the mother on a macro level, we can let the first two resolution layers of $\mathbf{w}'_c$ copy from $\mathbf{w}_f$, while the other layers copy from $\mathbf{w}_m$. At the end of the macro fusion, we just need to adjust $\mathbf{w}'_c$ along the pre-learned age vector to generate the child of an expected age.

A.2 Details of Inheritance Prior for Child Image Generation

We adopt some genetic evidence in biology (McKusick, 2014) to make our results more scientific (here we only consider Mendelian inheritance).

1. Skin color (Harrison, 1957): The skin color of a child always follows the natural law of "neutralizing" the skin colors of the parents.
2. Eyes: Big eyes are inherited in a dominant way, so as long as one parent has big eyes, the child is more likely to have big eyes.
3. Nose: Generally speaking, large, high noses and wide nostrils are in dominant inheritance. If one of the parents has a big nose, it is likely to be inherited by the child.
4. Jaw: As dominant inheritance, if one parent has a prominent big chin, the child will be more likely to grow into a similar chin.
5. Lip thickness: A thin upper lip is a dominant inheritance, while a thicker lower lip is a dominant inheritance.
6. Baldness (Osborn, 1916): Alopecia is caused by an autosomal dominant gene. Bald men may be heterozygous (Bb) or homozygous (BB), while bald women are homozygous (BB)

If we use B to present dominant gene and b for recessive gene, then the genotype with dominant trait is Bb or BB (in our work, we assume that they're equally likely), and the genotype with recessive trait is bb. If one parent presents a recessive trait and the other presents a dominant trait, the probability of the child presenting a recessive trait is: $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$, while the probability of presenting a dominant trait is $\frac{3}{4}$. If both parents display dominant traits, the probability of the child presenting a recessive trait is: $\frac{1}{4} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{16}$, while the probability of presenting a dominant trait is $\frac{15}{16}$. If both parents present recessive traits, their child will surely present a recessive trait.

In order to classify facial attribute values corresponding to biological traits, we compare the value of the attribute with a threshold value. We get the value for each attribute by by first detecting the face landmarks in the image and then computing the distances between corresponding landmarks. Each threshold is set as the average value of the attribute in a large number of samples. For attributes that cannot be classified by size, we use the projection value of the latent vector on the attribute vector's orientation instead of distance difference as the value of the attribute.

A.3 Details of Semantics Learning

As discussed in Section III.B, a genetic rule usually describes how a certain facial attribute is passed down from the parents to the child. To generate the child face according to the genetic laws, we need to identify these attributes in the latent space $\mathbb{W}+$ of StyleGAN, in which we fuse the faces of the parents. However, this is not readily available, since each $\mathbf{w}$ vector usually relates to multiple attributes. Some previous works (Shen et al., 2020; Ayush Tewari et al., 2020; Abdal et al., 2021; Pehlivan et al., 2023; Li et al., 2023) have shown that there are directions in the latent space of StyleGAN that correspond to different attributes of a face. In this section, we propose an effective method to identify semantic directions (or semantics in short) in the $\mathbb{W}+$ space that separately correspond to the attributes covered by the heredity laws. To ensure that moving a latent vector along one semantic direction affects other attributes as little as possible, we further make these semantic directions orthogonal to each other.

It is observed in StyleGAN that we can change the semantics contained in a synthesis continuously by linearly interpolating two latent codes. Let $\{\mathbf{v}_1, \cdots, \mathbf{v}_k\}$ be a set of vectors, each representing a semantic direction in the latent space. Since we chose to operate in $\mathbb{W}^+$ space, $k = 512 \times 18 = 9216$. These semantic directions will contain all the semantics we want to control. The difference between the latent codes of two images can be represented as a linear combination of the semantic vectors $\{\mathbf{v}_k\}$. The weight for each $\mathbf{v}_k$ is proportional to the change of the value for the corresponding attribute. Without loss of generality, we have:

$$\mathbf{w}_i - \mathbf{w}_j = \sum_k (u_{i,k} - u_{j,k}) \times \mathbf{v}_k, \tag{1}$$

where $\mathbf{w}_i$ and $\mathbf{w}_j$ are the latent codes of two images. $u_{i,k}$ and $u_{j,k}$ denote the values of the face attribute corresponding to $\mathbf{v}_k$. If $u_{i,k} - u_{j,k} \neq 0$, we have:

$$\frac{\mathbf{w}_i - \mathbf{w}_j}{u_{i,l} - u_{j,l}} = \mathbf{v}_l + \sum_{k \neq l} \frac{u_{i,k} - u_{j,k}}{u_{i,l} - u_{j,l}} \times \mathbf{v}_k. \tag{2}$$

With a large number of image pairs available, we can compute the expectation as:

$$E\left(\frac{\mathbf{w}_i - \mathbf{w}_j}{u_{i,l} - u_{j,l}}\right) = \mathbf{v}_l + \sum_{k \neq l} E\left(\frac{u_{i,k} - u_{j,k}}{u_{i,l} - u_{j,l}}\right) \times \mathbf{v}_k, \tag{3}$$

where $E(p)$ represents the statistical expectation of p.

Consider the most ideal situation, that is, if different semantics are independent of each other, when we have enough samples, we can estimate $\mathbf{v}_l$ through:

$$\mathbf{v}_l^e = \frac{2}{N \times (N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \frac{\mathbf{w}_i - \mathbf{w}_j}{u_{i,l} - u_{j,l}}, \tag{4}$$

where N is the total number of images available for learning, and $\mathbf{v}_l^e$ is an estimation of $\mathbf{v}_l$.

But in fact, the distribution of values of different attributes in sample pictures is not independent of each other. For example, people with larger eyes have larger mouths on average. That makes $E\left(\frac{u_{i,k} - u_{j,k}}{u_{i,l} - u_{j,l}}\right) > 0$. According to Equation 3, the vector $\mathbf{v}_l^e$ we find based on Equation 4 will have components not only in the $\mathbf{v}_l$ direction, but also in other directions, which can be written as $\mathbf{v}_l + \sum_{k \neq l} \frac{u_{i,k} - u_{j,k}}{u_{i,l} - u_{j,l}} \times \mathbf{v}_k$ as the sample size approaches infinity. This means that when we want to change the l-th attribute, the rest of the attributes will change as well.

In order to reduce the proportion of irrelevant components $\mathbf{v}_k$ $(k \neq l)$ in the extracted vector and reduce the influence of other attributes when changing the l-th feature, we add an additional weight to the terms in Equation 4:

$$\mathbf{v}_l^e = \frac{\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \frac{\mathbf{w}_i - \mathbf{w}_j}{u_{i,l} - u_{j,l}} \times e^{-\left|\frac{u_{i,m} - u_{j,m}}{u_{i,l} - u_{j,l}}\right|}}{\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} e^{-\left|\frac{u_{i,m} - u_{j,m}}{u_{i,l} - u_{j,l}}\right|}}, \tag{5}$$

where m is the index of the attribute which we want to reduce its entanglement with target semantic $\mathbf{v}_l^e$ due to the large value of $E\left(\frac{u_{i,m} - u_{j,m}}{u_{i,l} - u_{j,l}}\right)$. By introducing this weight factor, $\mathbf{w}_i - \mathbf{w}_j$ would be weighted heavier if the difference of m-th attribute between the two images is small, and $\mathbf{w}_i - \mathbf{w}_j$ would be given a lower weight if the m-th attribute of the two images are quite different. This also makes the variance of the restricted directional component smaller and accelerated the convergence speed. Equation 5 can be further extended if we want to do this disentanglement on more than one attribute. We only need to multiply the weight factors corresponding to these attributes together.

In addition to focusing on the overall characteristics of latent space, differences in different resolution layers in latent space facilitate further decoupling. We found that the first resolution layer of StyleGAN mainly controls camera elevation and horizontal angles, while the last four resolution layers mainly control color and background. In order to decouple the extracted facial semantics from attributes that we are not interested in, we can choose to only work on the middle three resolution layers.

The selection of the resolution layers and the estimation of the semantic vectors $\{\mathbf{v}_l\}$ have to a large extent disentangled the semantics. However, there are still some coupling of attributes in $\{\mathbf{v}_l\}$. To achieve more precise control, we further orthogonalize these vectors. In our work, we use the Gram-Schmidt process to made the semantic vectors orthogonal to each other. Starting with $\mathbf{n}_1 = \mathbf{v}_1$, we have:

$$\mathbf{n}_l = \mathbf{v}_l - \sum_{i=1}^{l-1} \frac{\langle \mathbf{v}_l, \mathbf{n}_i \rangle}{\langle \mathbf{n}_i, \mathbf{n}_i \rangle} \mathbf{n}_i, \tag{6}$$

where $\mathbf{v}_l$ is the semantic vector found by Equation 5 and $\mathbf{n}_i$ represents the orthogonal vector.

With the disentangled semantics identified by the method shown in this section, we can decompose the latent vectors of the parents by projecting them onto these semantic directions. Then an attribute of the child will be determined by picking a point in each semantic direction based on the genetic laws. We call this process the micro fusion of the parents.

A.4 Details of Micro Fusion

Now that we have obtained the decoupled semantic vectors that correspond to key attributes of the face, we can inherit face components of the parents according to the genetic laws. Based on the preliminary child latent code obtained after macro fusion, we further adjust it in the semantic directions. For each semantic vector, we first project the parents' and preliminary child's latent codes onto it. For the case that one parent presents a dominant trait while the other parent presents a recessive trait, we get the child's phenotype according to probability, and move the child's latent code to the father or mother's projection. If both parents show dominant traits but the child should show the recessive character according to probability, we move the child's latent code across the less obvious dominant side and move on until it becomes recessive. In the case of parents and child all showing dominant traits, we make the child's latent code move randomly under the restriction of parents' projection in this direction (the same for both parents with recessive traits or when there is no clear genetic rule to guide the semantic).

After dealing with each semantic direction in accordance with the above methods, we resynthesize $\mathbf{w}_c$ by:

$$\mathbf{w}_c = \hat{\mathbf{w}}'_c + \sum_l p_l \mathbf{v}_l, \tag{7}$$

where $\{\mathbf{v}_l\}$ are the semantic vectors, p_l is the projection component of the child image's latent code on each semantic direction, and $\hat{\mathbf{w}}'_c$ is what's left of $\mathbf{w}'_c$ (the child image's latent code after macro fusion) after been decomposed. After that, we send $\mathbf{w}_c$ into the StyleGAN Generator to obtain the final child image.

A.5 More Results

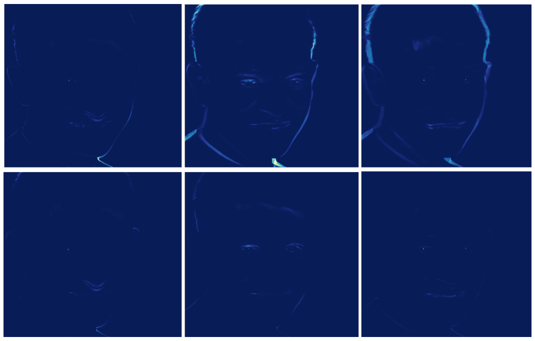

Figure 3: Heat maps of the mean squared error between the edited outputs and original image. The first and second rows are the results of using the basic method and the improved method, respectively. The edited attributes are the nose size (left), eye size (middle) and upper lip thickness (right).

As shown from the heatmaps in Figure 3, our improved method focuses on the component of interest better than our basic method, and it no longer modifies the face contour. This demonstrates that we can manipulate the facial attributes better with less attribute coupling.

In Figure 4, we show how the known genetic laws act on specific characteristics. The results demonstrate that the dominant traits of parents are more likely to be passed on to their children, which makes our generation process more scientifically sound and reliable.

In Figure 5 we show the diversity of our generated results. We can generated children with different ages and genders. Also, since the heredity of various characteristics follows the Mendelian inher-

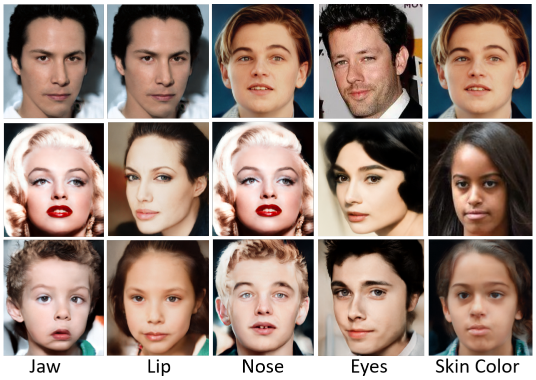

Figure 4: Examples of the role of genetic laws. Each column is the result of considering one genetic factor, and all children inherit the corresponding dominant traits.

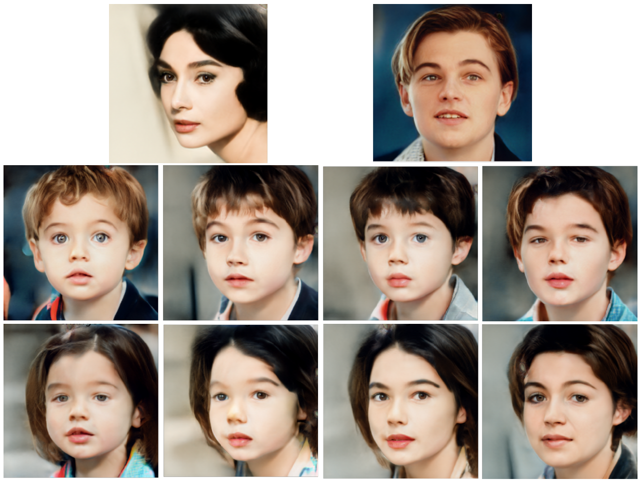

Figure 5: An example about the diversity of generated results. The first row shows the images of the parents, the second and the third rows are the results of children with different genders, ages, and genetic patterns.

itance law, it is not deterministic but with a certain probability that different children of the same parents do not look exactly the same in appearance. Our results show this diversity.

T-Distributed Stochastic Neighbor Embedding (Van der Maaten & Hinton, 2008) is a machine learning algorithm for dimensional reduction. We use it to reduce the dimensionality of high-dimensional facial features to 2-dimensional for visualization. Using this method, we draw a series of face images of parents, real children and the children generated by our method as Figure 6. As shown in Figure 6, unlike in DNA-Net (Gao et al., 2021) where the features of the generated children are concentrated on one side and far away from real ones, the features of children generated by us are evenly distributed. The feature distributions of child image generated by us is similar to that of real children's images. Also, we can see that the feature distributions of the generated child face is close to the faces of the parents (some are closer to father's features and some are closer to mother's features). This result is consistent with our vertification results.

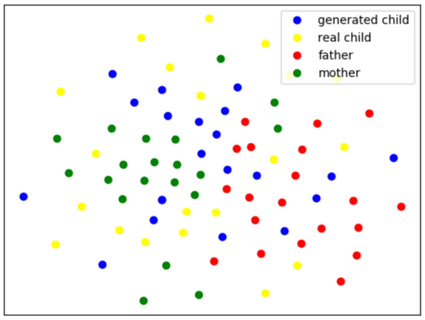

Figure 6: Visualization of facial feature distribution of fathers, mothers, children, and generated ones. Red points represent the feature of fathers, green for mothers, yellow for real children, blue for generated children, respectively. Best viewed in color.