KinStyle: A Strong Baseline Photorealistic Kinship Face Synthesis with An Optimized StyleGAN Encoder*

Li-Chen Cheng[1], Shu-Chuan Hsu[1], Pin-Hua Lee[1], Hsiu-Chieh Lee[1], Che-Hsien Lin[2], Jun-Cheng Chen[2], and Chih-Yu Wang[2]

[1] National Taiwan University
{b06902128,b06502152,b07303024,b07902030}@ntu.edu.tw
[2] Academia Sinica
{ypps920080,pullpull,cywang}@citi.sinica.edu.tw

Abstract. High-fidelity kinship face synthesis is a challenging task due to the limited amount of kinship data available for training and low-quality images. In addition, it is also hard to trace the genetic traits between parents and children from those low-quality training images. To address these issues, we leverage the pre-trained state-of-the-art face synthesis model, StyleGAN2, for kinship face synthesis. To handle large age, gender and other attribute variations between the parents and their children, we conduct a thorough study of its rich latent spaces and different encoder architectures for an optimized encoder design to repurpose StyleGAN2 for kinship face synthesis. The obtained latent representation from our developed encoder pipeline with stage-wise training strikes a better balance of editability and synthesis fidelity for identity preserving and attribute manipulations than other compared approaches. With extensive subjective, quantitative, and qualitative evaluations, the proposed approach consistently achieves better performance in terms of facial attribute heredity and image generation fidelity than other compared state-of-the-art methods. This demonstrates the effectiveness of the proposed approach which can yield promising and satisfactory kinship face synthesis using only a single and straightforward encoder architecture.

Keywords: Kinship face synthesis · StyleGAN Encoder.

1 Introduction

With the recent popularity of deep image and face synthesis, kinship face synthesis gets increasing attention in the research community of facial analysis. The goal of kinship face synthesis is to render the possible children faces given a pair of parental face images. This facilitates plenty of kinship-related applications,

* This work was supported by the National Science and Technology Council under Grant 108-2628-E-001-003-MY3, 111-2628-E-001 -002 -MY3, 111-3114-E-194-001 -, 110-2221-E-001 -009 -MY2, 110-2634-F-002-051-, 111-2221-E-001-002-, and the Academia Sinica under Thematic Research Grant AS-TP-110-M07-2.

including producing visual effects, delineating the possible facial appearances of a lost child after a long period of time, analyzing the facial traits of a family, etc. However, kinship face synthesis is still a challenging and ongoing research problem as compared with other face synthesis tasks due to a lack of large-scale training data, severe label noise, and poor image quality. In addition, it is also hard to trace the genetic traits between parents and children from those low-quality training images, especially when there is interference caused by the facial variations in illumination, pose, and other factors.

To synthesize a child face, one can use the image from a single reference parental image or images from both parents. For the former, because the information from the other parent is unavailable, there exist ambiguities of mapping a parental face to its child face. Ertuğrul *et al.* [1] propose the first work in this category, but the one-versus-one relation fails to capture enough information to yield promising and satisfactory results. For the latter, although two-versus-one relation between the parents and their children provides a good constraint for better kinship face synthesis, the limited amount of kinship data and data noise still restrict the generative models [2, 3], to synthesize high fidelity child faces. This also usually results in the situations of overfitting or lacking diversity for synthesized faces that are close to the average face when no further regularizations are applied. Lin *et al.* [4] leverage the pre-trained state-of-the-art face synthesis model upon the FFHQ dataset, StyleGAN2 [5], and train an additional encoder to extract latent representations encoding rich parental appearance information from the face images of parents encoding rich for kinship face synthesis to mitigate the training data issues. Their method can effectively utilize the data manifold of the pre-trained StyleGAN2 as a regularization to effectively restrict the possible kinship face distributions and to synthesize meaningful and good child faces. Nevertheless, without considering the issues of attribute data imbalance and feature entanglement, the method still lacks the capability to perform further smooth and effective manipulation over specific facial traits towards the parents, such as face component-wise manipulation over eyes, nose, mouth, etc. Zhang *et al.* [6] proposed to use multiple encoders for each component to realize component-wise manipulation, but this introduces more efforts of training and computational costs than others.

In general, the pipeline for kinship face synthesis can be divided into three stages: parental feature extraction, parental feature fusion, and face rendering. To address above issues of kinship face synthesis, we also leverage the pre-trained StyleGAN2 for rendering due to its encoded rich face prior and superior face synthesis capability. However, due to the complex nature of StyleGAN2 model and large appearance variations (i.e., age, gender, and other facial attributes.) between parents and their children, it requires us to conduct a careful study of its various latent spaces (i.e., Z, W, W+, S spaces.) and encoder architectures for an optimized encoder to repurpose StyleGAN2 for kinship face synthesis. To our knowledge, these have not been well studied for kinship face synthesis in the literature. With thorough evaluations of different design choices as shown in Table 1 (i.e., more details are presented in Section 4.), we propose an encoder

design consisting of an image encoder and a fusion block. The image encoder is further composed of an ID-preserved block with the design of *Encoder for Editing* (e4e) by Tov *et al.* [7] for better disentangled latent representation and editability in addition to an attribute block for normalizing age and gender variations of the parental representations for better synthesis fidelity. The fusion block fuses the latent representations of the parents for the final child representation. The obtained representation from the proposed encoder pipeline with stage-wise training strikes a better balance of editability and synthesis fidelity for identity preserving and attribute manipulations than other compared approaches. With extensive subjective, quantitative, and qualitative evaluations, the proposed approach consistently achieves better synthesis results using the Family-In-the-Wild (FIW) [8] and TSKinFace [9] datasets in terms of facial attribute heredity and image generation fidelity than other compared state-of-the-art methods. Furthermore, with our representation, we can also easily realize a component-wise parental trait manipulation (CW-PTM) through a method proposed by Chong *et al.* [10] to flexibly manipulate any desired face parts or regions of the synthesized face towards the parents through latent interpolation while ensuring the transition is smooth and continuous. Surprisingly, the manipulation results are competitive with other approaches employing multiple facial component encoders for the purpose. This also demonstrates the proposed method can not only yield promising and satisfactory kinship face synthesis but also enable the fine control of facial attributes using only a single and straightforward encoder architecture without the complex multi-encoder structure. This reduces training difficulties such as tuning the hyperparameters of multiple encoders simultaneously.

2 Related Work

In this section, we briefly review the relevant research works for kinship image synthesis using deep generative models.

Deep Image Synthesis: Many studies of image synthesis using a deep neural network rely heavily on Generative Adversarial Network (GAN) [11–14]. Based on GAN models, StyleGAN [15] was proposed to generate high-level attributes for synthesized images and preserve linearity in the latent space of generative models. StyleGAN encodes the latent z from the latent space Z into the feature space W, then w is chosen from feature space and input to multiple layers of convolution layers in order to control various styles of the output image through the adaptive instance normalization (AdaIN) module. Abdal *et al.* [16] proposed an efficient algorithm to perform inversion of the input image into an extended feature space W^+ instead of W space of a pre-trained StyleGAN for better image reconstruction. StyleGAN2 [5] further improves the details of the synthesized image, such as removing the blob-like artifacts by redesigning the network of StyleGAN. Moreover, StyleGAN2 simplifies the instance normalization process with a weight demodulation operation. The latent space of GANs has been stud-

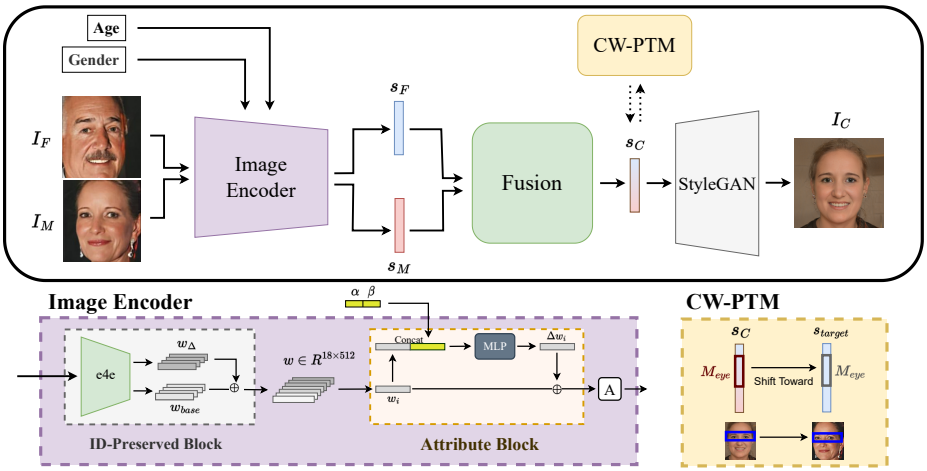


Fig. 1: The overview of the proposed encoder-decoder framework based on StyleGAN2 for kinship face synthesis.

ied carefully in recent years especially in the field of computer vision [17–20]. For facial images, it is ideal to map the source into latent space in an effort to reduce dimension as well as provide image editing in latent space. Tov *et al.* proposed Encoder for Editing [7] that allows manipulation of inverted images.

Kinship Face Synthesis: Kinship face synthesis is a recently commenced problem that aims to generate the child image given the images of parents [21]. Some works studied generating a child image using the image of father or mother as reference [22, 2, 23]. Nevertheless, these approaches suffer from the problem of either low resolution or mode collapse, and thus the generated results are unsatisfactory. Some works use both parents' images as input, such as the methods proposed by Ghatas *et al.* [24] and Zaman *et al.* [25]. Still, the artifacts from the first work are sometimes corrupted, and the second work does not take the child image as guidance. Some recent works have further improved the synthesis results in the kinship face synthesis problem using more advanced architectures and loss designs. Gao *et al.* [3] introduce DNA-Net that leverages conditional adversarial autoencoder to generate the child images. Zhang *et al.* [6] generate child images by assigning inheritance control vector of a facial part so as to let the child inherit the facial region from the mother or father. Lin *et al.* [4] concatenate latent space embedding of a child with age and gender vector to render child images with pre-trained StyleGAN2. ChildGAN [26] extracts the representative semantic vectors and synthesizes the child image by macro fusion and micro fusion. Our proposed method utilizes a novel designed encoder and the method of attribute alignment, our model is capable of synthesizing the child images that inherit designated facial regions, which leads to an outstanding performance as well as diversity in the synthesized child images.

3 METHODOLOGY

Our goal is to build a framework to synthesize a high-fidelity child face from a pair of parental face images, I_F and I_M, while being able to smoothly control the age, gender, and specific facial features of the synthesized face. The overview of the proposed framework is shown in Fig. 1, which comprises three phases. The first phase is to encode face images of the parents into the optimal latent codes, s_F and s_M, while having a good compromise between fidelity and editability for StyleGAN2. The identity information of parental face images is preserved through the ID-Preserved block, and the age and gender attributes are further normalized by the attribute block. In the second phase, we performs weighted average upon the transformed latent representations of the parents into the final representation, s_C, for the child. The weights can be either manually assigned or learned by several multilayer perceptrons (MLPs) layers. In the third phase, with the representation, a flexible and fine manipulation of the selected region towards the parents can be achieved through latent interpolation. In the following, we will describe the details of each component.

3.1 Phase 1: Image Encoder

Due to the immense diversity of age, gender, and identity for the face images of the parents and children, it is difficult to train a single encoder at a time to acquire a latent representation that not only preserves the identity information but also is age and gender invariant. Thus, we divide the image encoder into two blocks, one for preserving identity information and the other for normalizing the age and gender of the latent codes of the parents. Meanwhile, we perform stage-wise training, which provides more flexible training strategies and requires less computational resources. The details of each block are described as follows.

Identity-Preserved Block Let $E(\cdot)$ denote identity-preserved encoder block (ID-preserved block), and $G_{W^+}(\cdot)$ refers to StyleGAN2 taking the latent code in W^+ space as input. The objective of the ID-preserved block is to learn the mapping $E : \mathcal{I} \rightarrow W^+$, preserving the identity of the input image. As shown in Table 1, to preserve the identity information in the resultant latent representation while maximizing its editability, we adopt the *Encoder for Editing* (e4e) architecture [7] with the identity loss for the purpose. Given an input image, the encoder returns a base latent and a residual, each having the dimension $\mathbb{R}^{18\times512}$.

The final output latent code is obtained by adding them together. This design can minimize the variance between the latent codes from 18 layers of a W^+ space and make the latent code more editable. In addition, this also facilitates the age and gender normalization and other manipulations of the latent codes of the parents for the second and later stages. Let w denote the latent code encoded from the input image I, w_{base} denote the base latent code, w_{Δ} denote the residual, and I_{syn} denotes synthesized image. That is,

$$w = E(I) = w_{base} + w_{\Delta} \quad , \tag{1}$$

$$I_{syn} = G_{W^+}(w) \quad . \tag{2}$$

For the loss function, to preserve the identity of the input image, we impose a common identity constraint.

$$\mathcal{L}_{ID_1} = 1 - < R(I_{syn}), R(I) > \quad , \tag{3}$$

where $< \cdot , \cdot >$ is cosine similarity and $R(\cdot)$ is a pre-trained ArcFace model that extracts the feature map of a face from the penultimate layer.

Moreover, we add an L2 facial landmark loss to further enhance the alignment of the synthesized face I_{syn} to be centered at the image. The landmarks on the center line of the face between the nose and mouth are used for the landmark loss, which is defined as

$$\mathcal{L}_{land} = || \ E^x_{land}(I_{syn}) - C \ ||_2 \quad , \tag{4}$$

where $E^x_{land}(\cdot)$ is a pre-trained landmark predictor and C is a vector with the values of the x coordinate of the image center which is 512 in the rest of experiments for a 1024×1024 image. Lastly, we also add two additional losses as in [7] for image editability and quality. The first one is the an L2 regularization loss on the residual w_Δ:

$$\mathcal{L}_{reg_1} = || \ w_\Delta \ ||_2. \tag{5}$$

The second loss is the non-saturating GAN loss with R1 regularization, which is used for forcing w not to deviate from the W^+ space. Let $D(\cdot)$ denote a latent discriminator.

$$\begin{aligned} \mathcal{L}^D_{adv} = & - \mathop{\mathbb{E}}_{w_r \sim \mathcal{W}^+} [\log D(w_r)] - \mathop{\mathbb{E}}_{I \sim \mathcal{I}} [\log(1 - D(E(I)))] \\ & + \frac{\gamma}{2} \mathop{\mathbb{E}}_{w_r \sim \mathcal{W}^+} [|| \ \nabla_{w_r} D(w_r) \ ||_2^2] \quad , \\ \mathcal{L}^E_{adv} = & - \mathop{\mathbb{E}}_{I \sim \mathcal{I}} [\log D(E(I))] \quad . \end{aligned} \tag{6}$$

To sum up, the total loss function of ID-preserved block is

$$\mathcal{L}_{enc} = \lambda_{ID_1} \mathcal{L}_{ID_1} + \lambda_{land} \mathcal{L}_{land} + \lambda_{reg_1} \mathcal{L}_{reg_1} + \lambda_{adv} \mathcal{L}_{adv}. \tag{7}$$

Attribute Block After the ID-preserved block, the attribute block is further used to align the age and gender of input parental latent codes, $w_= \{w_i\}_{i=1}^{18}$, where $w_i \in \mathbb{R}^{512}$ in W^+. As mentioned in Section 2, the attribute manipulation can be achieved by shifting the latent code along specific latent directions. Thus, we learn an offset vector for desired modification by employing MLPs with leaky ReLUs followed by each MLP, $M(\cdot)$, which take the concatenation of the latent codes of the parents w, desired age α, and gender β values as input. For α, the input value ranges from 0 to 1, corresponding to 0 years old to 100 years old. For β, 1 represents male, and 0 represents female. Then, we can obtain the modified latent code w' by adding the original w and offsets $\Delta w = M(w, \alpha, \beta)$.

In the training process, α is sampled from $Uniform[0,1]$, and β is sampled from $Bernoulli(0.5)$. Besides arbitrary attributes, we can also use ground truth attributes to construct a reconstruction loss and a cycle consistency loss on w. We can further obtain three modified latent codes for each w.

$$\begin{aligned} w'_{syn} &= w + M(w, \alpha, \beta) \\ w'_{rec} &= w + M(w, \alpha_{gt}, \beta_{gt}) \\ w'_{cyc} &= w'_{syn} + M(w'_{syn}, \alpha_{gt}, \beta_{gt}) \end{aligned} \tag{8}$$

where α_{gt}, β_{gt} denote the ground truth age and gender labels. For each w' vector, we can obtain the corresponding synthesized images $I_{syn}, I_{rec}, I_{cyc}$ by passing it into $G_{W+}(\cdot)$. For the loss function of the attribute block, we first employ the following age and gender losses.

$$\begin{aligned} \mathcal{L}_{age} &= ||\alpha - C_a(I_{syn})||_2 + ||\alpha_{gt} - C_a(I_{rec})||_2 + ||\alpha_{gt} - C_a(I_{cyc})||_2 \\ \mathcal{L}_{gen} &= H(\beta, C_b(I_{syn})) + H(\beta_{gt}, C_b(I_{rec})) + H(\beta_{gt}, C_b(I_{cyc})) \ , \end{aligned} \tag{9}$$

where $C_a(\cdot)$ denotes a pre-trained age classifier, $C_b(\cdot)$ denotes a pre-trained gender classifier, and $H(\cdot)$ denotes the cross-entropy loss. Similarly, we also use identity loss for attribute block training. However, since the identity of a person may become obscure as the person ages, Alaluf *et al.* [27] proposed an identity loss decayed with the age difference between the prediction and the ground truth. We further extend the idea to both age and gender. The identity loss can be formulated as

$$\begin{aligned} \mathcal{L}_{ID_2} = \xi \cdot (1 - < R(I_{syn}), R(I) >) + \\ (1 - < R(I_{rec}), R(I) >) + (1 - < R(I_{cyc}), R(I) >), \end{aligned} \tag{10}$$

where ξ is the decay coefficient, and we set $\xi = 0.45 + 0.35 \cdot cos(|\alpha - \alpha_{gt}| \cdot \pi) + 0.2 \cdot cos(|\beta - \beta_{gt}| \cdot \pi)$. Moreover, to make training faster and prevent latent codes from deviating from the original W^+ space, we utilize perceptual similarity losses on the images and use L2 regularization on the offsets.

$$\begin{aligned} \mathcal{L}_{reg_2} =& ||M(w, \alpha, \beta)||_2 + ||M(w, \alpha_{gt}, \beta_{gt})||_2 + \\ & ||M(w'_{syn}, \alpha_{gt}, \beta_{gt})||_2, \\ \mathcal{L}_{per} =& ||P(I_{enc}) - P(I_{syn})||_2 + ||P(I_{enc}) - P(I_{rec})||_2 + \\ & ||P(I_{enc}) - P(I_{cyc})||_2 \ , \end{aligned} \tag{11}$$

where $I_{enc} = G_{W^+}(E(I))$ is the reconstructed image by passing the latent after ID-preserved block to StyleGAN2, and $P(\cdot)$ is a pre-trained AlexNet feature extractor upon the ImageNet dataset. Lastly, the total loss function is as follows.

$$\mathcal{L}_{attr} = \lambda_{ID_2}\mathcal{L}_{ID_2} + \lambda_{age}\mathcal{L}_{age} + \lambda_{gen}\mathcal{L}_{gen} + \lambda_{reg_2}\mathcal{L}_{reg_2} + \lambda_{per}\mathcal{L}_{per}. \tag{12}$$

As suggested in [28], the latent codes in S space of StyleGAN2 result in a better style mixing performance. We follow the idea and apply the affine transform layers, $A(\cdot)$, available in the StyleGAN2 model to convert the latent representation of each parent, w' vector, into s for the next stage as $s = A(w')$.

3.2 Phase 2: Fusion

Once the transformed latent representations of the parents in S space are obtained, we perform weighted average to blend them into one final child latent code as follows:

$$s_C = \gamma \circ s_M + (1 - \gamma) \circ s_F, \tag{13}$$

where s_F and $s_M \in \mathbb{R}^{9088}$ denote two latent codes in S space for the parents, s_C denotes the resultant child code, $\circ$ is the element-wise product, and $\gamma \in [0, 1]^{9088}$ is the blending coefficient which can be either manually specified or learned by a fusion network where we employ an MLP layer that takes the concatenated vector of s_F and s_M as the input trained with the $L2$ reconstruction and ID losses similar to the ID-preserved block. Besides S space, we also perform the blending in W^+ space for comparison, and more details can be found in Section 4.1.

3.3 Phase 3: Component-wise Parental Trait Manipulation (CW-PTM)

With our optimized encoder pipeline of kinship face synthesis in Phase 1 and 2, the resultant latent code for the child is suitable for editing through latent interpolation as compared with other methods, like StyleDNA. We can easily apply a similar approach as in [10] to realize component-wise parental trait manipulation (CW-PTM) in a single encoder-decoder framework. The relation between each dimension of the latent code and specific facial features, such as eyes, nose, and mouth, are obtained with K-means clustering. Then, a mask corresponding to the facial features over the latent code can be derived accordingly. A specific facial attribute can be transferred from one image to another by shifting the original latent code towards the target one based on the mask. For example, suppose the mask for the eyes is denoted by M_{eye}, where $M_{eye} \in \{0, 1\}^{9088}$ and one stands for the position of the latent vector controlling the synthesized face's eyes, as shown in Fig. 1. Therefore, we can shift the part of s_C within M_{eye} toward s_{target} by a coefficient ϵ. That is,

$$s'_C = s_C + \epsilon \cdot M_{eye} \circ (s_C - s_{target}) \ , \tag{14}$$

where s'_C is the modified latent code, and $\circ$ denotes the element-wise product. The step size ϵ will determine how similar the eye is to the target. Leveraging this method, we can manipulate the latent code to make the synthetic child inherit specific parental traits. The proposed approach is not only more memory and computation efficient but also allows more flexible and smooth manipulation towards any selected regions of the parents by latent interpolation as shown in Fig. 6b than other similar works, such as [6] which employs fixed multiple component-wise encoders. In addition, the proposed approach avoids using multiple encoders, which increases the training difficulties due to plenty of model parameters and hyperparameter tuning. The multiple region manipulations can be achieved through recursively applying the same procedure.

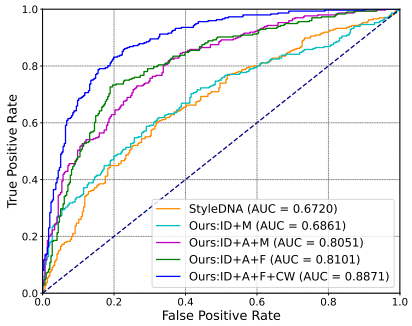


Fig. 2: It illustrates the ROC curves for ablation studies and the comparisons between the proposed approach and the baseline, StyleDNA. ID stands for ID-preserved block, A for attribute block, F for learned fusion, and CW for component-wise parental trait manipulation. M represents directly averaging parent latent codes without fusing them with a learned network.

4 EXPERIMENT

In this section, we show the results of the proposed approach with several recent most representative state-of-the-art methods for quantitative, qualitative, and subjective comparisons.

Implementation Details: We use a pre-trained and fixed StyleGAN2 model upon the FFHQ dataset as our decoder. For the encoder, we train both ID-preserved and attribute blocks using the FFHQ-Aging dataset [29], which contains images with age and gender label information. We also perform weighted sampling based on the number of training instances per age and gender. Then, the learned fusion block is trained with the FIW dataset [8], which contains approximately 2,000 tri-pairs of kinship images. Since images in the FIW dataset are low-resolution, we preprocess them with GFP-GAN for super-resolution [30] before training. The facial landmarks are extracted using the MobileFaceNets [31]. We pre-train the age and gender classifiers following the same setting as [4]. Instead of one-hot vectors, age and gender conditionals are transformed to $[0, 1]$ and $\{0, 1\}$ respectively and then duplicate 50 times each in order to facilitate the stable training. For hyperparameters of training, we set batch size as 6, use a standard Ranger optimizer with a learning rate 0.0001, and set the loss weights as follows. We set $\lambda_{ID_1} = 1,\ \lambda_{land} = 0.0008,\ \lambda_{reg} = 0.0002,\ \lambda_{adv} = 0.5$ for ID-preserved block; $\lambda_{ID_2} = 0.5,\ \lambda_{gen} = 1,\ \lambda_{age} = 5,\ \lambda_{per} = 0.5,\ \lambda_{reg} = 0.05$ for attribute block; $\lambda_{ID_3} = 1,\ \lambda_2 = 1$ for the learned fusion block.

Training Pipeline: To facilitate the training process and handle the data imbalance issues, we perform stage-wise training to train a component at a time

while freezing the model weights in prior stages. This enables us not only to apply relevant losses for the optimal training for each encoder block according to their task characteristics but also to train the model using much less GPU resources than other approaches.

4.1 Quantitative Evaluations of Different Encoder Configurations

Due to the complex nature of the StyleGAN2 model, for an optimized encoder design, we first perform an investigation of the encoder design choices of different latent spaces and network architectures for kinship face synthesis in terms of AUC of the ROC curve for face verification and FID scores for image synthesis where we measure the facial similarity between the synthesized child face and its corresponding ground truth child face. The pre-trained ArcFace model for face verification is used to extract the latent representations of both the prediction and ground truth followed by cosine similarity computation. We randomly sample 100 positive and negative pairs from the test set of the FIW dataset for the similarity computation. From Table 1, we find that encoding images to W^+ space or S space can generate offspring faces that are more realistic and more similar to ground truths. In addition, the editability of encoded latent codes needs to be considered in our styleGAN pipeline. The results show that the methods with another popular pSp backbone [32] which does not account for editability attained lower AUC and higher FID scores compared to the ones with e4e. Lastly, although the learning-based fusion can achieve slightly better performance improvement than the manual one, there is not much difference between them if the encoder is well designed. Then, performing fusion in W^+ space or S space attained similar AUC and FID scores. In the rest of the experiments, we select the configuration of (6) to further conduct qualitative and subjective evaluation, since it has the highest AUC for the best identity preservation and also yields good perceptual quality image. For the configuration of (6), we further compare the ROC curves of the proposed method with StyleDNA due to the public availability of its source code. As shown in Fig. 2, we can see the proposed approach achieves the best performance with AUC 0.8101. We also show the blue curve for the most promising result with AUC 0.8871 after applying CW-PTM using ID loss from the ground truth as the guidance. The improved number also shows the strength of CW-PTM to explore a more resembled face to the ground truth child face. Also, the learning-based fusion can achieve slightly better improvement than the manual one. In the rest of the experiments, we select the configuration of (6) to further conduct qualitative and subjective evaluation, since it has the highest AUC and yields good perceptual quality image.

4.2 Subjective Evaluation

To further compare the generation quality of different methods, we conduct the subjective evaluation in two independent online sessions, 186 participants for the first and 131 for the second. Each session contains 9 and 13 questions in total respectively. For each session, we ask the participants to rank the synthesized

Table 1: The quantitative results of pipelines with different combinations of the encoder and the fusion method.

	Encoder		Fusion		AUC (↑)	FID (↓)
	Space	Type	Space	Type		
(1)	W	Resnet	W	Learned	0.6720	197.9197
(2)	W^+	e4e	W^+	Mean	0.8050	133.3718
(3)	W^+	pSp	W^+	Mean	0.7738	176.4417
(4)	W^+	e4e	S	Mean	0.8051	133.3681
(5)	W^+	pSp	S	Mean	0.7738	176.4349
(6)	W^+	e4e	S	Learned	0.8101	138.2808
(7)	W^+	pSp	S	Learned	0.7783	173.5681

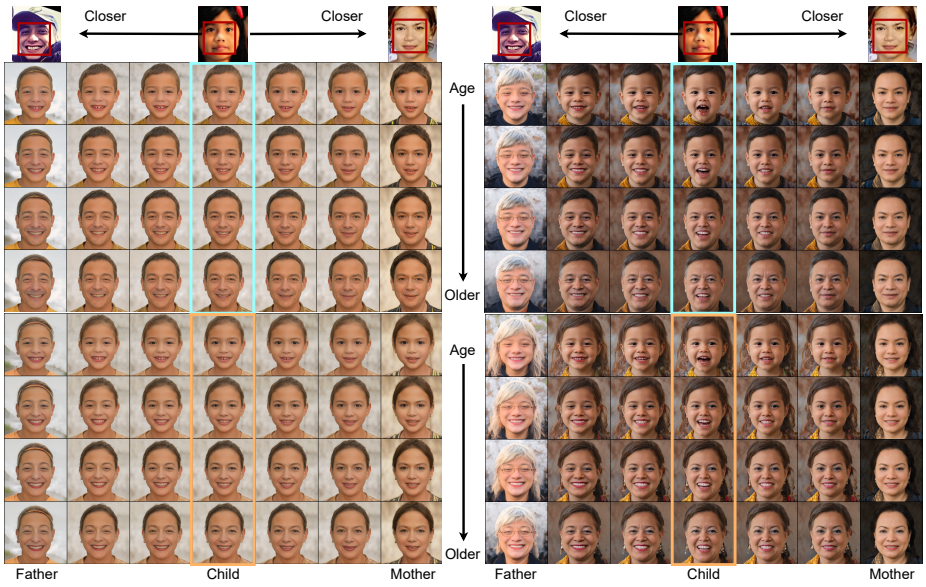
Table 2: (a) The weighted average rank for the resemblance between the synthesized child faces using different approaches and a pair of parental face images. (b) The weighted average rank of naturalness and photo-realism for the synthesized child faces of different approaches.

	GT	styleDNA [4]	ChildGAN [26]	Ours
Session I	2.83	2.92	2.33	**1.92**
Session II	2.27	2.82	2.18	**1.55** (CW-PTM)

(a)

	styleDNA [4]	ChildGAN [26]	Ours
Avg rank	**1.57**	2.59	1.84

child faces by different methods, ground truth and other state-of-the-art methods, StyleDNA [4] and ChildGAN [26] along with the ground truth according to their perceived resemblance to the given reference face images of the parents, where a lower rank represents a higher score (*i.e.*, one is the most likely and four is the least likely.). Since the code of [6] which employs multiple facial component encoders is not publicly available and the image quality of the pdf file is low, we do not use it for subjective evaluation. In addition, we also asked participants to answer the extent of naturalness and photo-realism among different synthesized child faces. As shown in Table 2, we compute the weighted average ranks of rank for the resemblance between the synthesized child and the reference parental faces. The proposed method consistently achieve the best subjective performances as compared with other methods. For the photo-realism test in Table 2 for the first session, the average weighted rank of the proposed approach is close to StyleDNA. Instead of ranking, we further conducted the mean opinion score to compare the photo-realism of the proposed approach with StyleDNA in the second session, where the score ranges from 1 to 5, and a higher value is better. The proposed approach achieves better average score of **3.53** than



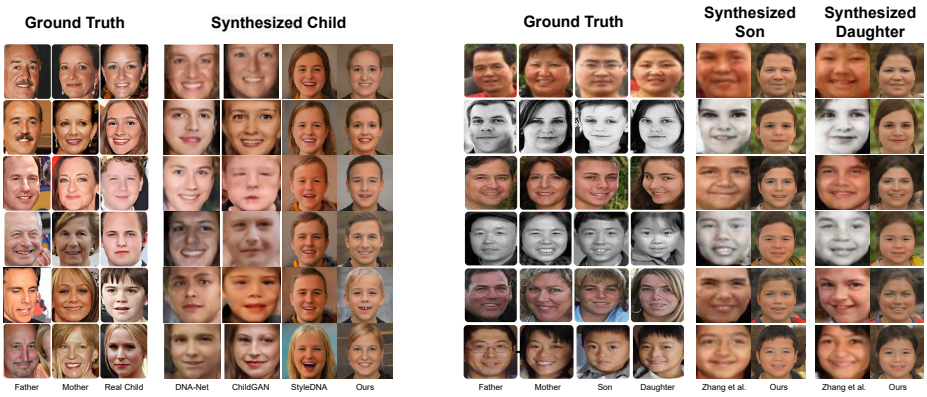
(a) Ours (b) StyleDNA

Fig. 3: The qualitative comparisons between the proposed approach and StyleDNA after performing fine attribute manipulation towards the parents using CW-PTM upon the respective extracted latent representations according to the selected facial component or region of the parent. The top row shows the ground truth faces of the parents and their child. The proposed approach achieves more continuous and better manipulation results than the compared baseline, keeping the age and gender intact while performing the manipulation.

StyleDNA, **3.23**. It is worth noting that the proposed approach even outperforms the ground truth. These results further demonstrate the strength of the proposed approach. For more details of the questions and rank scores, we refer interested readers to the supplementary materials.

4.3 Qualitative Evaluation

Finally, we also show various visual samples to compare the proposed approach with other methods in terms of the capability of various attribute manipulations, including age, gender, and parental traits. From Fig. 3, we find the synthesized faces by the proposed approach change more smoothly when adjusting the values of the corresponding latent codes as compared with StyleDNA (*i.e.*, the inference code of StyleDNA is publicly available, and StyleDNA is thus chosen as the main comparison target for the qualitative analysis.). This further demonstrates the advantages of the proposed approach to synthesize high fidelity kin faces while allowing smooth component-wise manipulation towards any selected facial



(a) FIW (b) TSKinFace

Fig. 4: The results of ground truth face images along with the synthesized child images from the proposed and other compared methods. Left most three or four columns of (a) for FIW and (b) for TSKinFace respectively depict the ground truth images and right four columns depict the synthesized child images, where the compared methods include DNA-Net [3], ChildGAN [26], StyleDNA[4], and Zhang *et al.* [6].

components or regions of the parents as shown in Fig. 6b. Employing multiple encoders usually introduces additional computational costs and training difficulties. We also show the synthesis results in Fig. 4 using the face images of the FIW and TSKinFace datasets. The proposed method can synthesize the faces with better fidelities than other approaches.

4.4 Ablation Studies

In this section, we perform ablation studies to understand the effectiveness of each component where we indicate ID for ID-preserved block, A for attribute block, F for learned fusion, M represents directly averaging parent latent codes without fusing them with a learned network, and CW means the proposed component-wise parental trait manipulation. As shown in Fig. 6a, the fidelity of the synthesized child faces gets improved with applying more components of the proposed framework. With the assistance of CW-PTM, we can further flexibly make the synthesized child faces closer towards either the father or the mother or both with different facial components or regions at the same time. Surprisingly, from Fig. 2, the AUC scores of all the ablated results are better than StyleDNA. This shows the strength of the proposed optimized encoder architecture and the stage-wise training.

On the other hand, for attribute block, besides a single MLP, we further divide the encoded latent code from the eighteen layers of the StyleGAN2 model into three groups with a corresponding MLP for transformation where the first

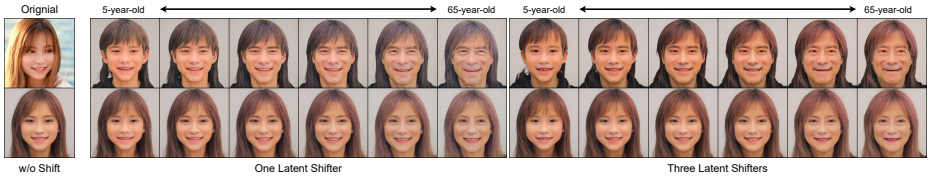
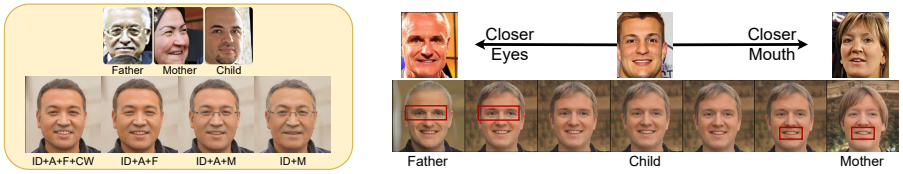


Fig. 5: This illustrates the effects using different numbers of MLP layers in the attribute block. For the left, it uses a single MLP and three MLPs for the right. The results using three MLPs show more prominent facial manipulation than the one using a single MLP.



(a) (b)

Fig. 6: (a) It shows qualitative results using different combinations of our network components for the ablation study. (b) It illustrates the proposed approach allows flexible parental trait manipulation with any selected facial components or regions towards parents.

three for coarse-grained, fourth to seventh layers for middle-grained, and the rest for fine-grained detail control. Although the differences in the quantitative results on FID of two settings are small, which are 37.056 and 36.663. However, we can see the three MLPs results in face images with better attribute manipulation than a single MLP one, as shown in Fig. 5.

5 Conclusion

The main contribution of our work is to conduct a thorough study of different encoder choices of different latent spaces and encoder architectures to repurpose StyleGAN2 for kinship face synthesis. The proposed optimized encoder striking a better balance of editability and synthesis fidelity for kinship face synthesis than other compared methods while allowing smooth and continuous face trait manipulation. With extensive subjective, quantitative, and qualitative evaluations, the proposed approach consistently achieves better performance in terms of facial attribute heredity and image generation fidelity than other state-of-the-art methods. This demonstrates the effectiveness of the proposed approach, which can yield promising and satisfactory kinship face synthesis using only a single straightforward encoder architecture.

References

1. Ertugrul, I.Ö., Dibeklioglu, H.: What will your future child look like? modeling and synthesis of hereditary patterns of facial dynamics. In: IEEE International Conference on Automatic Face and Gesture Recognition (FG). (2017)
2. Ozkan, S., Ozkan, A.: Kinshipgan: Synthesizing of kinship faces from family photos by regularizing a deep face network. In: IEEE International Conference on Image Processing (ICIP). (2018)
3. Gao, P., Robinson, J., Zhu, J., Xia, C., Shao, M., Xia, S.: Dna-net: Age and gender aware kin face synthesizer. In: IEEE International Conference on Multimedia and Expo (ICME). (2021)
4. Lin, C.H., Chen, H.C., Cheng, L.C., Hsu, S.C., Chen, J.C., Wang, C.Y.: Styledna: A high-fidelity age and gender aware kinship face synthesizer. In: IEEE International Conference on Automatic Face and Gesture Recognition (FG). (2021)
5. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). (2020)
6. Zhang, Y., Li, L., Liu, Z., Wu, B., Fan, Y., Li, Z.: Controllable descendant face synthesis. arXiv preprint arXiv:2002.11376 (2020)
7. Tov, O., Alaluf, Y., Nitzan, Y., Patashnik, O., Cohen-Or, D.: Designing an encoder for stylegan image manipulation. ACM Transactions on Graphics (TOG) **40** (2021) 1–14
8. Robinson, J.P., Shao, M., Wu, Y., Liu, H., Gillis, T., Fu, Y.: Visual kinship recognition of families in the wild. IEEE Transactions on pattern analysis and machine intelligence **40** (2018) 2624–2637
9. Qin, X., Tan, X., Chen, S.: Tri-subjects kinship verification: Understanding the core of a family. In: 2015 14th IAPR International Conference on Machine Vision Applications (MVA). (2015) 580–583
10. Chong, M.J., Chu, W.S., Kumar, A., Forsyth, D.: Retrieve in style: Unsupervised facial feature transfer and retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). (2021)
11. Liu, S., Sun, Y., Zhu, D., Bao, R., Wang, W., Shu, X., Yan, S.: Face aging with contextual generative adversarial nets. In: Proceedings of the 25th ACM International Conference on Multimedia. (2017)
12. Tang, H., Bai, S., Sebe, N.: Dual attention gans for semantic image synthesis. In: Proceedings of the 28th ACM International Conference on Multimedia. (2020)
13. Brock, A., Donahue, J., Simonyan, K.: Large scale GAN training for high fidelity natural image synthesis. In: International Conference on Learning Representations (ICLR). (2019)
14. Zhao, J., Xiong, L., Karlekar Jayashree, P., Li, J., Zhao, F., Wang, Z., Sugiri Pranata, P., Shengmei Shen, P., Yan, S., Feng, J.: Dual-agent gans for photorealistic and identity preserving profile face synthesis. Advances in neural information processing systems **30** (2017)
15. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). (2019)
16. Abdal, R., Qin, Y., Wonka, P.: Image2stylegan: How to embed images into the stylegan latent space? In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2019)

17. Zhang, L., Bai, X., Gao, Y.: Sals-gan: Spatially-adaptive latent space in stylegan for real image embedding. In: Proceedings of the 29th ACM International Conference on Multimedia. (2021) 5176–5184
18. Sainburg, T., Thielk, M., Theilman, B., Migliori, B., Gentner, T.: Generative adversarial interpolative autoencoding: adversarial training on latent space interpolations encourage convex latent distributions. arXiv preprint arXiv:1807.06650 (2018)
19. Mukherjee, S., Asnani, H., Lin, E., Kannan, S.: Clustergan: Latent space clustering in generative adversarial networks. In: Proceedings of the AAAI conference on artificial intelligence. Volume 33. (2019) 4610–4617
20. Bojanowski, P., Joulin, A., Lopez-Paz, D., Szlam, A.: Optimizing the latent space of generative networks. arXiv preprint arXiv:1707.05776 (2017)
21. Robinson, J.P., Shao, M., Fu, Y.: Survey on the analysis and modeling of visual kinship: A decade in the making. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
22. Ertuğrul, I.Ö., Jeni, L.A., Dibeklioğlu, H.: Modeling and synthesis of kinship patterns of facial expressions. Image and Vision Computing **79** (2018) 133–143
23. Sinha, R., Vatsa, M., Singh, R.: Familygan: Generating kin face images using generative adversarial networks. In: European Conference on Computer Vision Workshops (ECCVW). (2020)
24. Ghatas, F.S., Hemayed, E.E.: Gankin: generating kin faces using disentangled gan. SN Applied Sciences **2** (2020) 166
25. Zaman, I., Crandall, D.: Genetic-gan: Synthesizing images between two domains by genetic crossover. In: European Conference on Computer Vision Workshops (ECCVW). (2020)
26. Cui, X., Zhou, W., Hu, Y., Wang, W., Li, H.: Heredity-aware child face image generation with latent space disentanglement. arXiv preprint arXiv:2108.11080 (2021)
27. Alaluf, Y., Patashnik, O., Cohen-Or, D.: Only a matter of style: Age transformation using a style-based regression model. ACM Transactions on Graphics (TOG) **40** (2021) 1–12
28. Kafri, O., Patashnik, O., Alaluf, Y., Cohen-Or, D.: Stylefusion: A generative model for disentangling spatial segments. arXiv preprint arXiv:2107.07437 (2021)
29. Or-El, R., Sengupta, S., Fried, O., Shechtman, E., Kemelmacher-Shlizerman, I.: Lifespan age transformation synthesis. In: Proceedings of the European Conference on Computer Vision (ECCV). (2020)
30. Wang, X., Li, Y., Zhang, H., Shan, Y.: Towards real-world blind face restoration with generative facial prior. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2021)
31. Chen, S., Liu, Y., Gao, X., Han, Z.: Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. CoRR **abs/1804.07573** (2018)
32. Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., Cohen-Or, D.: Encoding in style: A stylegan encoder for image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). (2021)