KinStyle: A Strong Baseline for Photorealistic Kinship Face Synthesis with An Optimized StyleGAN Encoder (Supplementary Materials)*

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

A Implementation details

In this section, we show additional details of some components for further elaboration.

A.1 Landmark Loss

The landmark loss is used to further enhance the alignment of the synthesized face while training the ID-preserved block. As shown in Figure 1, we use the landmarks in red to compute the landmark loss.

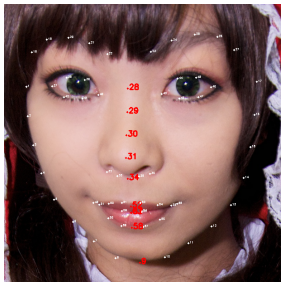


Fig. 1: Red points are the facial landmarks in the center line of a face, which we use for landmark loss.

* 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.

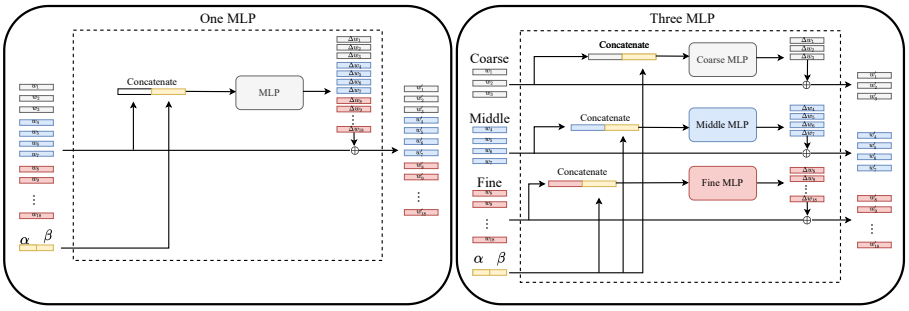


Fig. 2: The detailed architecture of attribute block. The left side is with one MLP. The right side is with three MLPs.

A.2 Attribute Block

For the elaboration of different setups of attribute blocks in the ablation study, we show the illustrations in Figure 2. For the left-hand side, it is the one with a single MLP, and three MLPs for the right-hand side, where a leakyReLU is followed by each MLP. For three MLPs, the first three layers are for coarse-grained image features, 4-7 layers are for middle-grained, and the rest of the layers are for fine-grained.

A.3 Kinship Face Synthesis during Inference

During Inference, for easy comparison purposes, we use our pre-trained age and gender classifiers to predict the child's age and gender of the ground truth. Then, we align the age and gender for the latent codes of the parents based on the estimated age and gender values for later kinship face synthesis (i.e., the proposed method can freely specify the desired age and gender for normalization of parental face images. For the example of teenagers, we suggest specifying the ages ranging from 13 to 19 for the best synthesis results.)

B More Qualitative Results

In this section, we show more qualitative results to demonstrate the strength of the proposed approach.

Table 1: Configurations

		(1)	(2)	(3)	(4)	(5)	(6)	(7)
Encoder	Space	W	W^+	W^+	W^+	W^+	W^+	W^+
	Type	Resnet	e4e	pSp	e4e	pSp	e4e	pSp
Fusion	Space	W	W^+	W^+	S	S	S	S
	Type	Learned	Mean	Mean	Mean	Mean	Learned	Learned

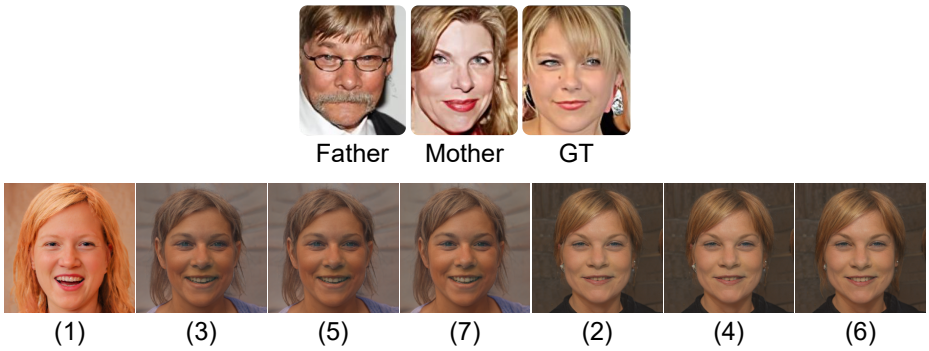


Fig. 3: The qualitative results of the proposed framework with different configurations. The configurations of each synthesized result are shown in Table 1 where we group them according to the encoder architecture. The results of (2)(4)(6) with the e4e encoder architecture are perceptually closer to the ground truth child face than (3)(5)(7) with pSp.

B.1 Qualitative Evaluations of Different Configurations

In Figure 3, we show the offspring images synthesized by encoders with different configurations. The results are consistent with the quantitative results in Section 4.1 of the main paper. Encoding images to W space results in offspring faces with low identity similarity to the ground truths. Furthermore, the methods with an e4e backbone allow us to generate faces with higher quality and closer to the ground truths than others. Lastly, the learning-based fusion achieves a slight improvement over the manual one.

B.2 Ablation Study

In Figure 4, we show more visualization results for the ablation study of different encoder components. We can see that the result has the highest fidelity when we apply all components of our framework.

B.3 Component-wise Parental Trait Manipulation (CW-PTM)

In Figure 5, we demonstrate the capability to manipulate more than one face part in our proposed framework. For instance, the synthesized offspring image can have the mother's eyes and father's nose at the same time, which is (a) in Figure 5. Thus, with the flexibility, we can synthesize many possible offspring faces, including the one that resembles the real offspring, as (c) and (d) shown in Figure 5.

B.4 Attribute Manipulation

We show another example of attribute manipulation in Figure 6. Our proposed framework has smoother and better control on age and gender attributes along with parental trait manipulation of the region of interest than StyleDNA.

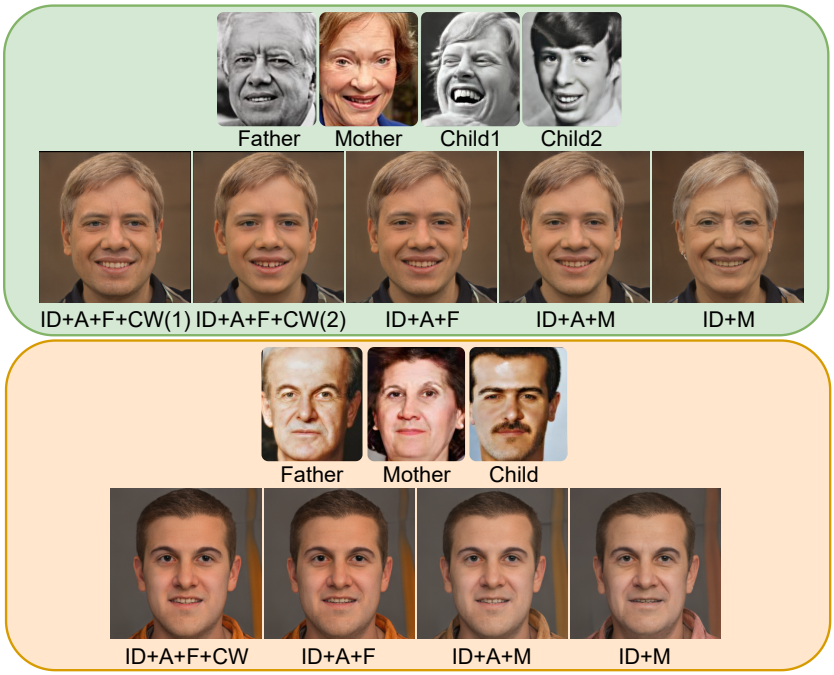


Fig. 4: More qualitative results for ablation study. ID stands for ID-preserved block, A for attribute block, F for learned fusion, and CW for component-wise parental trait manipulation. M represents directly averaging parent latent codes without fusing them with a learned network.

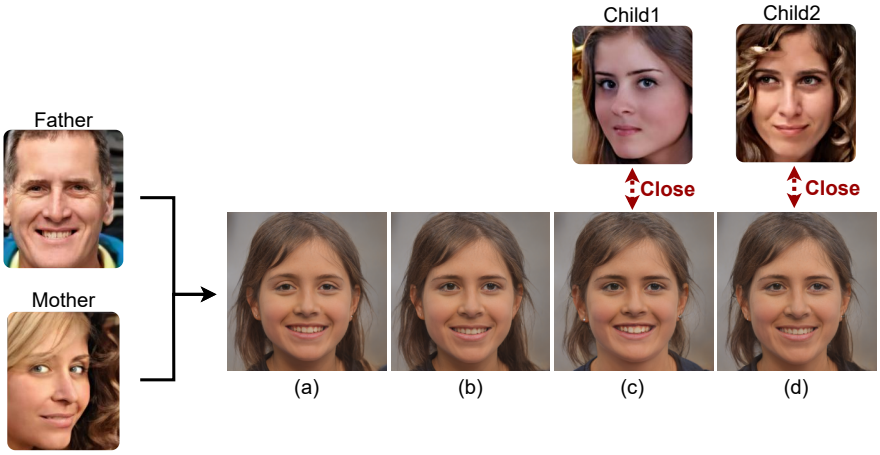
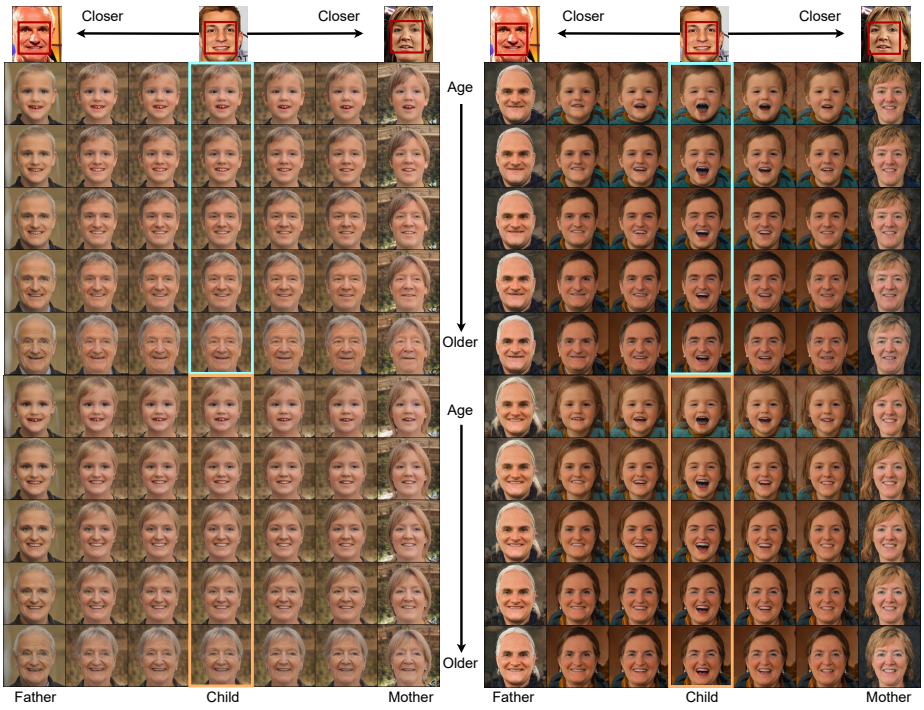


Fig. 5: Demonstration of flexibility in component-wise manipulation. (a) synthesized child with mother's eyes and father's nose. (b) synthesized child with father's eyes and mother's nose and mouth. (c)(d) The other synthesized results close to the ground-truth child 1 and 2.



(a) Ours (b) StyleDNA

Fig. 6: Another qualitative comparison between the proposed approach and StyleDNA using CW-PTM 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.

C Subjective Evaluation

In this section, we show the details of the subjective evaluation.

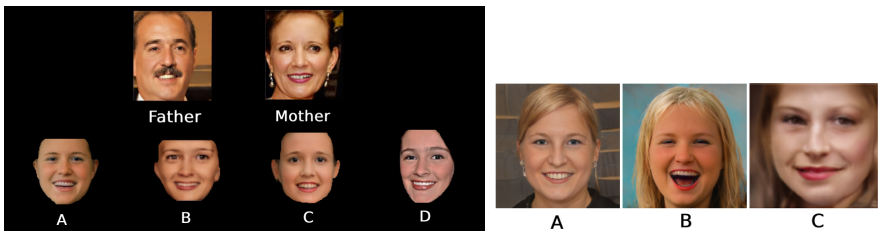
C.1 Weighted Rank Calculation

To calculate the weighted average rank, we set the weight of each rank as the number of people who choose the rank and then average them by the weight. For example, suppose there are three images in a question, and the rank is from 1 to 3. An image is ranked as 1 by ten people, ranked as 2 by thirty people, and ranked as 3 by twenty people. The weighted average rank of the image will be

$$\frac{1 \times 10 + 2 \times 30 + 3 \times 20}{60} = 2.17 \ . \tag{1}$$

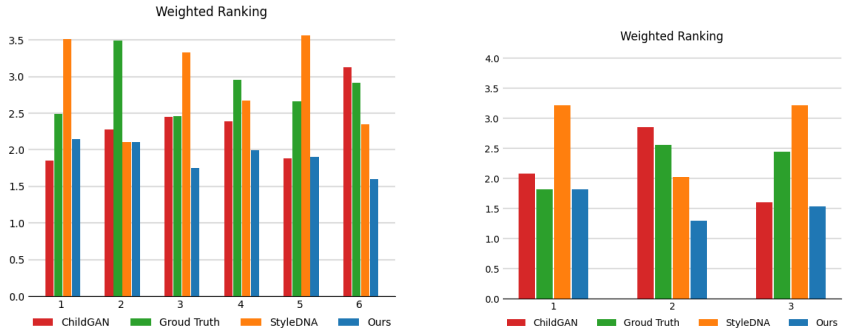
C.2 Sample Questions and Response Details

Figure 7a is a visual illustration of a sample question in our questionnaire. Participants are given a pair of parental faces and four possible child faces (one is the ground truth face, and the other three are synthesized ones.). Then, they are asked to rank A, B, C, and D in the order of resemblance to the parent images.



(a) (b)

Fig. 7: (a) A sample question in our survey for the resemblance of the synthesized child faces to the reference face of the parents. (b) A sample question in our survey for naturalness and photorealism of images.

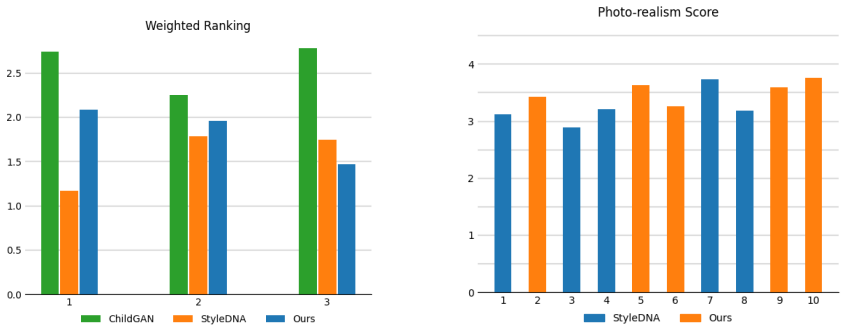


(a) (b)

Fig. 8: (a) The weighted ranks in questions about parent resemblance from the first session. (b) The weighted ranks in questions about parent resemblance from the second session.

Figure 8a is the weighted rank of images in each question about resemblance from the first session of our survey, and Figure 8b is from the second session; We use the mean opinion score for the second session, which ranges from 1 to 5 (*i.e.*, higher values are better.). The ranks in the second session are converted from the opinion scores for consistent comparisons. The final results in the subjective

evaluation section are obtained by averaging the weighted ranks of the questions of each session respectively. In addition, for naturalness and photorealism, Figure 9a is the weighted rank of images from the first session, and Figure 9b is the average score of each question from the second session.



(a) (b)

Fig. 9: (a) The weighted ranks in questions about naturalness and photorealism from the first session. (b) The average opinion score of each question about naturalness and photorealism from the second session. The opinion score ranges from 1 to 5, and a higher value is better.