ChildNet: Structural Kin-based Facial Synthesis Model with Appearance Control Mechanisms

Martin Pernuš, *Student Member, IEEE,* Simon Dobrišek, *Member, IEEE,* and Vitomir Štruc, *Senior Member, IEEE*

***Abstract*—Kin-based facial synthesis is an increasingly popular topic within the computer vision community, particularly the task of predicting the child appearance based on parental images. Previous work has been limited in terms of model capacity and training data, which consists of tightly cropped and low resolution images, resulting in reduced synthesis performance. We propose ChildNet, a method for kin-based facial synthesis, that alleviates such issues. The proposed method is based on state-of-the-art GAN, GAN inversion encoders and a latent code model, that together work to accurately predict the child image. In order to gain more control over the generated images, we propose an age and gender manipulation module that can optionally refine the child's synthesis results. Our model is capable of generating multiple images per input, while explicitly controlling the variability of the synthesized images. In addition, we introduce a mechanism to control the dominant parental image, allowing us to determine whether the child should inherit more features from either father or mother image. To leverage the high-resolution models, we introduce a new kinship dataset Next Of Kin, which contains 3690 images of high-resolution facial images with high ethnic and age diversity. We evaluate ChildNet against three other kin facial synthesis models on two kinship datasets with extensive experiments. The experiments show superior performance of ChildNet in terms of identity similarity, while exhibiting high perceptual image quality. The source code is publicly available at the following URL: X.**

I. Introduction

Kin-based facial synthesis is a relatively novel computer vision task, where the models are given a single or multiple input facial images, then tasked with synthesizing a facial image that can be convincingly regarded as a real kin-related person. The success of recent image generation models has provided a basis for new models that leverage kinship datasets, traditionally used for kinship recognition tasks. A successful application of such methods could be used in several different domains, such as applications in the entertainment industry or criminal investigations regarding the search of long-missing family members. In our paper we focus on the task of child face synthesis, which is the most popular kin-based facial synthesis subtask.

The kin-based facial synthesis methods have recently made huge strides, owing especially to recent success of unconditional image generation models such as generative adversarial networks (GANs) [1]. The methods make use of the low-dimensional GAN latent space by mapping the input images to a low-dimensional latent code, then using a GAN model as a decoder to generate an image. To obtain a latent code kinship synthesis methods make use of GAN inversion methods that obtain a latent code that roughly corresponds to the input image when processed with GAN. Once a latent code of each of the parents is obtained, they are processed with specialized mathematical operations, usually inspired by some sort of genetic mechanism such as hereditary properties of facial attributes or gene mixing. The result of mathematical operations is a new latent code that, when processed through the GAN decoder, corresponds to a facial image that can be regarded as real kin-related person. The key challenge to further improve performance of existing models is increasing the model capacity, allowing it to generate images that resemble the real children more accurately, while preserving the processing time of a feedforward model. Furthermore, the models lack a proper fine-grained control of the generated image such as control over the variety of generated facial images and control over the dominant parental image.

Fig. 1: **Kin-based facial synthesis.** Our model is capable of a high resolution kin-based facial synthesis, where the synthesized images are visually convincing and exhibit highly similar facial traits to the real child images.

In this paper we present a novel model named ChildNet that further advances the performance of child facial synthesis. Our model addresses the limitations of model expressivity by learning a neural network in the extended latent space of image generation model. The architecture is based on an attention mechanism that selectively attends to parental features and mutation mechanism that slightly offsets the parental-based features. The model is learned end-to-end, avoiding having to specify any kind of hard-coded external knowledge rules in the model implementation. To map images to GAN latent space

we use contemporary GAN inversion encoder-based methods. Furthermore, we design a separate disentanglement module in the extended latent space, allowing us to further refine the generated child image based on provided age and gender information. We design several model mechanisms that can control aspects such as controlling the image variability and specifying the dominant parental image. Based on experiments on two kinship datasets, our model is shown to achieve higher face similarity with real children images than competing methods, while also allowing generation of multiple images per fixed input, controlling the variability of generated images and specifying the dominant parent for the generated images. Figure 1 shows examples of ChildNet synthesized images.

A critical problem of the kin-based facial synthesis methods is a lack of suitable kinship synthesis dataset. Specifically, kin-based facial synthesis models are trained using datasets, where the primary aim is the task of kinship recognition. Therefore, the datasets tend to contain face images with highly variable image quality and contain overly cropped face images. This is in contrast to modern datasets used in training unconditional image generation models [2], [3], which tend to use high-resolution, high quality face images with consistent positioning. Therefore, we present a new kinship dataset Next Of Kin (NOK) that includes 3690 512×512 high quality, high resolution facial images of 553 subjects.

In summary, we make the following main contributions in this paper:

- We present ChildNet, a state-of-the-art model for child generation.
- We introduce an age and gender manipulation mechanism, operating in the extended latent space of StyleGAN.
- We present Next Of Kin dataset, a dataset of high resolution face images with various metadata such as kinship relation, age, ethnicity, gender and emotion.
- Through rigorous experimental quantitative and qualitative evaluation we demonstrate that ChildNet achieves a more accurate child synthesis than competing models.

II. Related work

In this section we present prior work closely related to our paper. We discuss Generative Adversarial Networks (GANs), image editing through the use of GANs, kinship datasets and various methods of kinship face synthesis.

A. Generative Adversarial Networks for Image Generation and Editing

Generative Adversarial Networks (GANs) [1] have become one of the most popular image generation models in the recent years. Since the initial model proposal, GANs were improved in terms of model design and training procedures. DCGAN [4] proposed a convolutional GAN architecture and defined several useful design principles. Karras *et al.* [2] proposed a progressive learning strategy that enabled generation of convincing megapixel resolution images. It was further improved with introduction of StyleGAN [3], the design of which was inspired by the style transfer architecture. Its next iteration, StyleGAN2 [5] modified the design to remove circular artefacts in the generated images. StyleGAN3 [6] proposes a continuous interpretation of network signals to prevent the dependence of the generated image on the absolute pixel coordinates, creating smoother image latent-based interpolations. Considerable progress has also been made in the field of GAN training strategies with various training loss functions and regularization methods [7], [8], [9], [10], [11]. Additional, up-to-date information on GANs can be found in one of the recent survey on this topic [12], [13].

GAN inversion methods are concerned with inverting an image into the latent space of a pre–trained GAN model and performing image editing operations. Abdal *et al.*[14] first proposed an iterative inversion into an extended latent space of StyleGAN, displaying its significant image reconstruction capabilities. The work was further extended in [15], which improved the inversion algorithm and showed additional editing results. To prevent a time–consuming process of embedding an image in a GAN latent space due to its iterative nature, pSp model [16] proposed an image encoder that produced extended latent codes for individual image. E4e model [17] proposed encoder that focused not just on image reconstruction but on preservation of the original latent code distribution, resulting in improved image editing. ReStyle proposed iterative encoder refinement of latent code [18]. HyperStyle [19] proposed a hyper–network that modified the StyleGAN weights based on the input image. Some works proposed other type of latent spaces, such as StyleSpace [20], achieving a more disentangled latent code representation. A survey of GAN inversion methods can be found in [21].

Recently, a considerable number of methods emerged that performed image manipulation based on a pre–trained GAN. Abdal *et al.* [14] performed various image manipulation techniques, such as face morphing, style transfer and expression transfer. The follow-up work [15] demonstrated additional manipulations, such as image crossover, local image editing and inpainting. Shen *et al.* [22] proposed to fit a linear support vector machine on StyleGAN latent space based on face attribute labels of the corresponding images and used the obtained directions to edit images. Some methods focus on supervised discovery of semantic latent directions in the latent space [23], [24]. MaskFaceGAN [25] proposed several constraints during the face latent code optimization for face editing that does not exhibit attribute entanglement issues. Model pSp [16] used encoder model to produce frontalized facial images, conditional image synthesis, inpainting and super-resolution. StyleCLIP [26] performs latent-based image edits based on text inputs.

To achieve fast processing of input images while still retaining high perceptual quality and low distortion we use the E4e encoder [17] as our primary encoder. The encoder is trained to predict latent codes close to distribution of the original latent codes. This makes it less likely for latent operations such as interpolation to introduce artifacts in the final image due to a predicted latent code falling outside the true distribution. We design special latent operations for image editing, specific to the considered task of kin-based facial synthesis such as enforcing a dominant parental image.

B. Kinship Datasets

The basis for training models for kinship synthesis is a dataset with annotation of kinship relations. Over the years of kinship recognition research, several datasets have emerged with varying characteristics in terms of number of images, kin relationships, image quality, type of cropping and additional metadata.

Jiwen Lu *et al.* [27] introduced KinFaceW–I and KinFaceW–II datasets. The facial images were captured under uncontrolled environment with no restriction in terms of pose, occlusion and lighting. J. P. Robinson *et al.* [28] presented the Families in the Wild (FIW) dataset. It consists of 1,000 families with over 10,000 family photos and metadata about 11 types of relations based on gender combinations of parent–child, grandparent–grandchild and siblings relations. The dataset was further extended [29] in order to add data to underrepresented families, extending the image pair dataset count from the initial 400,000 to 650,000. Qin, X *et al.* [30] proposed the TSKinFace database that consists of families of 2589 individuals. Fang *et al.* [31] introduced CornellKin dataset of 150 pairs of public figures and celebrities, along with images of their parents or children. Family–101 dataset [32] was proposed in order to examine the similarity of traits in family members. This dataset consists of 101 families including 607 individuals with 14,816 images. Xia *et al.* [33] proposed UB Kinface dataset to evaluate and analyse algorithms for the task of kinship verification. The images are based on real-world collections of celebrities and politicians. Siblings Database [34] consists of two main subsets: high quality (HQFaces) and low quality (LQFaces) database. HQFaces consists images of 92 sibling pairs that were shot by a professional photographer with uniform background and controlled lighting. LQFaces contains 98 pairs of siblings found over the Internet, where most of the subjects are celebrities.

A common aspect of existing datasets is that the faces either contain tight crops, include low quality images or have insufficient number of samples, which makes it challenging to use them for the task of face synthesis. The Next Of Kin dataset, introduced in this paper, fills an obvious research need and features high-quality face images that make it possible to design and train contemporary generative models capable of producing photo-realistic high-resolution (artificial) face images.

C. Kinship Recognition and Synthesis

Most of the kinship based research is focused on the tasks involving kinship recognition. One of the first methods for classifying parent–child pairs is presented in [31], where low-level features were used, such as average skin region color and histogram of gradients. Xia *et al.* [35] proposed a subspace learning method to transfer the distribution of old parents to younger parents to get a more discriminative parent–child verification problem. This work was extended in [36], where Gabor filters, metric learning and subspace transfer learning were adopted to achieve better verification results. Fang *et al.* [32] presented a method for face reconstruction through linear combination of database parts, which was used to determine which family a given person belongs to. Wang *et al.* [37] proposed a denoising autoencoder based metric learning framework for kinship recognition. The latest Families in the Wild kinship recognition challenge [38] proposes three main challenges: kinship verification, tri-subject verification and search & retrieval of family members for missing children.

In comparison with kinship recognition, there exist fewer methods that deal with the task of kinship face synthesis. Among those, the most popular task is synthesizing the child appearance based on the parental image. DNA-Net [39] proposed a neural network, where the image encoder produced 'gene' features, that were then merged with certain gene selection rule. The produced gene features then served as the input to the decoder that produced the generated face image of the child. Cui *et al.* [40] proposed a model that fused the parental images based on genetic biology knowledge about the hereditary properties about the size of individual facial attributes. StyleDNA [41] proposed an encoder-decoder model that controlled the age and gender of the generated child image, while the gene merging strategy was inspired by the DNANet.

In our work we propose ChildNet, a model that, unlike the above models, has a high expressive power and a structural design. This enables it to define individual operations on the parts of the latent code that correspond to different image characteristics. Furthermore, our design is based on attention and mutation mechanism that is learned end-to-end without the need to manually design the gene–merging strategy.

III. Methodology

A. ChildNet Overview

The goal of a our task is synthesizing a child face image $I_C \in \mathbb{R}^{3\times n\times n}$ given the two parent images $I_F \in \mathbb{R}^{3\times n\times n}$ as father's image and $I_M \in \mathbb{R}^{3\times n\times n}$ as mother's image. We can additionally provide age information $\rho \in \{1, ..., d^\rho\}$ and gender information $\gamma \in \{1, ..., d^\gamma\}$, where d denotes number of elements for each modality, to provide more information to the model. Concretely, the task of ChildNet is to learn a function

$$\psi(I_F, I_M, \rho, \gamma; \theta) = I_C \in \mathbb{R}^{3\times n\times n}, \tag{1}$$

where θ denotes the learnable parameters.

Our model is split into two main submodules: Kinship Module which determines the child latent code based on the parental latent codes and Age & Gender Manipulation Module that optionally modifies the child latent code based on provided age and gender information. After generating the final child latent code, it is forwarded through a decoder model. In our work we use StyleGAN2 [5] trained on FFHQ dataset [3] as our decoder model. In line with contemporary models that edit face images using StyleGAN models [14], [15], [41], we base our mapping model in the extended latent space $\mathcal{W}^+$ that spans multiple 512-dimensional latent vectors, the exact number depending on the model resolution. As we are using 1024×1024 resolution model, out latent space consists of a concatenation of 18 512-dimensional latent vectors.

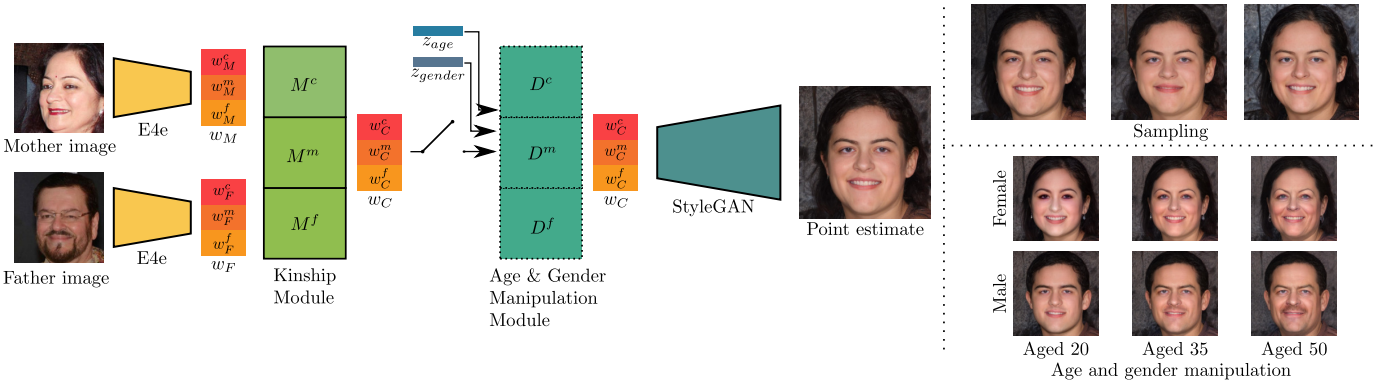


Fig. 2: **A schematic overview of the proposed ChildNet.** The parental images are encoded with E4e model, then processed individually with corresponding Kinship Module layer M^p (coarse, medium or fine). If provided with age and gender information, the latent code is further modified by an Age & Gender Manipulation Module before being fed to StyleGAN decoder. Our proposed model also supports image sampling per fixed image input.

B. Kinship Module

Architecture. The focus of the Kinship Module lies in the latent space of a pre-trained GAN decoder G, training a mapping model that manipulates the parental latent codes in order to produce the optimal child latent code. We denote $(w_F, w_M) \in \mathcal{W}^+$ as father's and mother's latent code, respectively. To map the parental images to $\mathcal{W}^+$ latent space, we use the E4e encoder [17] as $w = E(I)$. The child image is then generated by finding the optimal w_C based on given pair of parental latents (w_F, w_M).

We assume that facial characteristics can be dominantly influenced by one of the parent, however the dominance is allowed to differ with respect to individual facial attributes. For example, face shape could be more influenced by one of the parents, while hair colour and freckles presence could be more influenced by the other parent. We therefore design a structured model that splits the latent codes into coarse, middle and fine parts, as these parts were shown to primarily influence only a subset of facial characteristics [3]. Latent vector parts are processed by a corresponding mapping layer M^p and at the end concatenated together from the predicted parts. The dimensionality of a latent code differs between its parts; the coarse and medium part contain 4 512-dimensional latent vectors each, while the fine part contains 10 512-dimensional latent vectors. Each vector within a single latent part is processed by the module that corresponds to vectors' part.

We furthermore presume that a child's latent code w_C lies somewhere close to the latent space hyperplane, as defined by parental latent codes (w_F, w_M). The module is designed with the aim to determine the latent code that lies on such hyperplane and move from it to a limited degree. Each module part M^p takes parental latent code part as inputs. Each module part has two tasks: first, it attends to individual latent parental components by predicting interpolation coefficients, which form the basis for the predicted child latent code. Second, it predicts mutation components that offset the predicted latent code from the hyperplane, defined by the parental latent codes. By defining the model in a way that determines an individual interpolation and mutation value for each latent code component we are able to restructure the entangled characteristics present in the parental latent codes.

The module part M^p consists of two sublayers with same architecture termed BaseNet; Attention BaseNet M^p_{att} which attends to parental components and Mutation BaseNet M^p_{mut} which offsets the latent code. Given father's and mother's latent code part $(w_F^{p_i}, w_M^{p_i})$, where $p_i \in \{0, ...3\}$ for coarse, $p_i \in \{4, ..., 7\}$ for medium and $p_i \in \{8, ..., 17\}$ for fine part, the child latent code part is defined as:

$$\alpha^{p_i} = M^p_{att}(w_F^{p_i}, w_M^{p_i}) \tag{2}$$

$$\epsilon^{p_i} = M^p_{mut}(w_F^{p_i}, w_M^{p_i}) \tag{3}$$

$$w_C^{p_i} = \alpha^{p_i} \odot w_F^{p_i} + (1 - \alpha^{p_i}) \odot w_M^{p_i} + \epsilon^{p_i}. \tag{4}$$

The BaseNet is defined as a multilayer perceptron. It takes two inputs, denoted as (x_1, x_2) that are first individually processed by separate 5-layered fully connected layers, then concatenated and further processed with another 5-layered fully connected layer, resulting in an output vector y of the same dimensions as the input vector. The BaseNet is visualized in Fig. 3.

Loss function. Our loss function contains three main terms. The identity loss term ensures the generated child image and the ground truth child image look visually similar. We achieve this using ArcFace recognition model [42] R, maximizing the cosine similarity between model embeddings:

$$\mathcal{L}_{id} = 1 - \cos(R(G(w_C)), R(I_C)), \tag{5}$$

where I_C denotes the ground truth child image.

Throughout our experiments, we found that triplet loss in the latent space helps with the model performance. We set the anchor child latent to w_C, a positive one is defined as $w_C^+ = E(I_C^+)$ and a negative one as $w_C^- = E(I_C^-)$, where I_C^+ denotes the ground truth child image and I_C^- denotes a child image which does not belong to ground truth child's parents. The loss term is then defined as

$$\mathcal{L}_{tri} = \max(||w_C - w_C^+||_2^2 - ||w_C - w_C^-||_2^2 + \delta, 0), \tag{6}$$

where δ is an experimentally defined margin value.

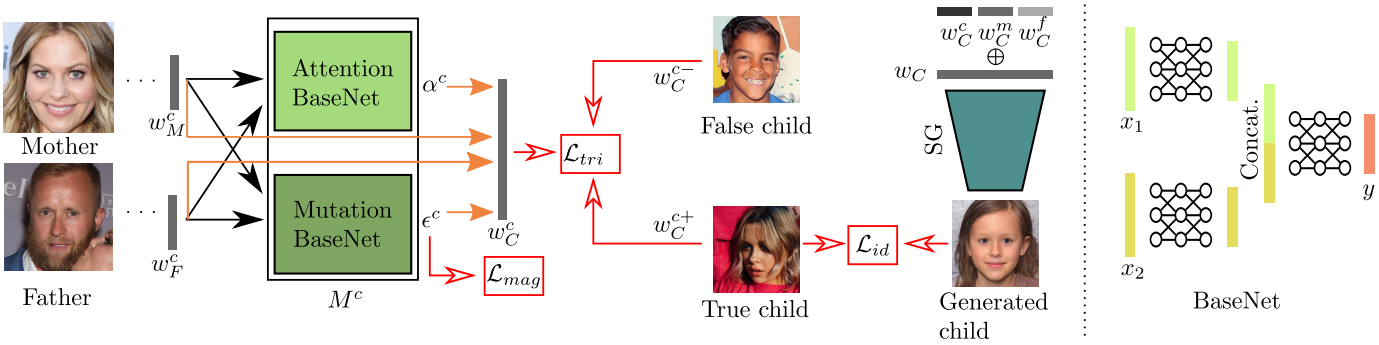


Fig. 3: **Detailed architecture and training scheme of Kinship Module**. The left side of the figure presents the architecture and training scheme of the Kinship Module coarse part M^c. The other model parts M^m and M^f follow the same general outline. The parental latent codes (w_M^c, w_F^c) get processed with Attention BaseNet, which provides interpolation coefficients α^c, and with Mutation BaseNet, which provides residual coefficients ϵ^c. This forms the basis for child latent code w_C^c. It is concatenated with results of other sublayers (w_C^m, w_C^f) to form a complete extended latent vector w_C with which we generate the child's image. During training we employ the identity $\mathcal{L}_{id}$, triplet $\mathcal{L}_{tri}$ and magnitude $\mathcal{L}_{mag}$ loss terms. The right side presents a basic premise of a BaseNet. Based on the input vectors (x_1, x_2) it processes those with a series of fully connected layers, then concatenating the produced features and processing them with another series of fully connected layers.

Finally, to ensure that the mutated child latent code stays relatively close to the hyperplane, defined by the parent latent codes, we define a loss term that limits the magnitude of latent code residual ϵ. The loss term is defined as

$$\mathcal{L}_{mag} = ||\epsilon||_2^2 \tag{7}$$

The final loss function is defined as a weighted sum of the described loss terms:

$$\mathcal{L} = \lambda_{id}\mathcal{L}_{id} + \lambda_{tri}\mathcal{L}_{tri} + \lambda_{mag}\mathcal{L}_{mag}, \tag{8}$$

where each λ denotes a scalar weight of the corresponding loss term.

C. Age and Gender Manipulation

In line with contemporary child generation approaches, we propose an age and gender manipulation mechanism, allowing a finer control of the generated child image. The model training is based on the idea of cycle consistency [43].

Based on the input face image along with arbitrary gender and age information, our task is to generate a face image that expresses the desired age and gender semantics, while preserving the identity information as much as possible. Given the age value ρ and gender value γ, our disentanglement model D defines the following mapping:

$$w' = D(w, \rho, \gamma). \tag{9}$$

Architecture. The Age & Gender Manipulation Module D has a similar architecture as a Kinship Module. It is again split into three parts, where each part is a single BaseNet. When training the model we also make use of E4e encoder and StyleGAN decoder. The E4e encoder maps the original face image I_{orig} into w vector. The D module takes w vector, age value $\rho \in \{0, ..., 9\}$ and gender value $\gamma \in \{0, 1\}$ as an input. The age and gender input interact as an embedding index to their learnable embedding vectors $z_{age}^{\rho} \in \mathbb{R}^d$ and $z_{gender}^{\gamma} \in \mathbb{R}^d$, where d denotes the dimension of the embedding. The gender and age embeddings are concatenated and repeated along the w layer dimension. The mapping network then processes w and and concatenated age & gender embedding with a BaseNet model, producing a residual vector ϵ. The concatenated residual vector ϵ is added to the w embedding: $w' = w + \epsilon$ and fed through StyleGAN decoder, producing the final image. The architecture of the model is shown in Figure 4.

Loss function. Our data consists of images $I \in \mathbb{R}^{3\times n\times n}$ and its ground truth annotation of age $\rho_{orig} \in \{0, 1, ..., 9\}$ as well as gender $\gamma_{orig} \in \{0, 1\}$. During training we generate three types of images: reconstruction image I_{rec}, target image I_{tar} and cycle image I_{cyc} as well as their corresponding residuals:

$$I_{rec}, \epsilon_{rec} \leftarrow D(I, \rho_{orig}, \gamma_{orig}) \tag{10}$$
$$I_{tar}, \epsilon_{tar} \leftarrow D(I, \rho_{rand}, \gamma_{rand}) \tag{11}$$
$$I_{cyc}, \epsilon_{cyc} \leftarrow D(I_{tar}, \rho_{orig}, \gamma_{orig}), \tag{12}$$

where ρ_{orig} and γ_{orig} denote original age and gender values, while ρ_{rand} and γ_{rand} denote uniformly randomly sampled age and gender values.

In order to achieve the target age and gender attribute semantics on the target image I_{tar}, we use pre–trained age model A and gender model G. We define age loss $\mathcal{L}_{age}$ and gender loss $\mathcal{L}_{gender}$ as cross-entropy loss:

$$\mathcal{L}_{age} = -\sum_{c=1}^{m_a} y_c^a \log A_c(I_{tar}) \tag{13}$$
$$\mathcal{L}_{gender} = -\sum_{c=1}^{m_g} y_c^g \log G_c(I_{tar}) \tag{14}$$

where m denotes the number of classes, y denotes the ground truth and A_c, G_c denote the age and gender model prediction for the c-th class, respectively.

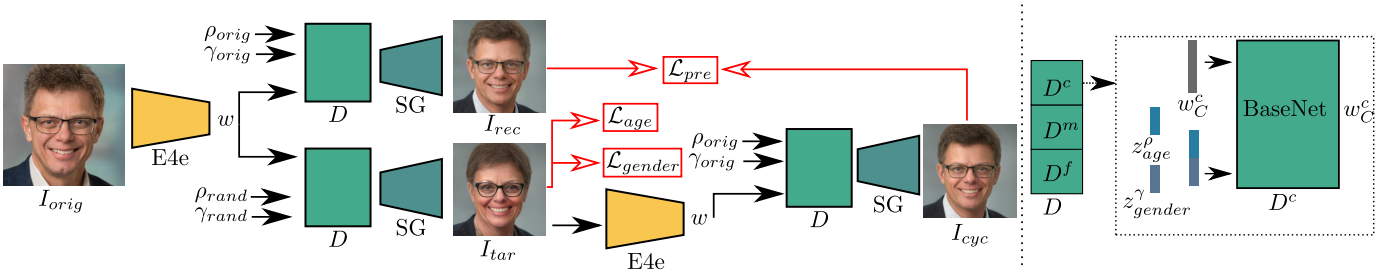


Fig. 4: **A schematic overview of Age & Gender Manipulation Module**. The left side of the figure presents the Age & Gender Manipulation Module training setup. Supplying the ground truth information about age and gender we obtain reconstructed image I_{rec}, while supplying randomized information produces target image I_{tar}. The correct semantics on the target image are obtained using age $\mathcal{L}_{age}$ and gender $\mathcal{L}_{gender}$ losses. Using the original age and gender information, the target image I_{tar} is transformed into the cycle image I_{cyc}. Preservation loss is applied to both reconstructed I_{rec} and cycle I_{cyc} images. The right side of the figure presents a detailed input to a coarse part of the module. The other parts of the module have the same setup with their corresponding latent code parts.

To prevent adversarial-like solutions, we again resort to magnitude loss:

$$\mathcal{L}_{mag} = ||\epsilon_{rec} + \epsilon_{tar} + \epsilon_{cyc}||_2^2 \tag{15}$$

Finally, in order to preserve the cycle consistency, we apply preservation loss on reconstruction and cycle images:

$$\mathcal{L}_{recon} = \text{LPIPS}(I_1, I_2) + ||I_1 - I_2||_2^2 \tag{16}$$

$$\mathcal{L}_{pre} = \mathcal{L}_{recon}(I_{orig}, I_{rec}) + \mathcal{L}_{recon}(I_{orig}, I_{cyc}), \tag{17}$$

where LPIPS denotes perceptual loss [44].

The final loss term is defined as a weighted sum of the described loss terms:

$$\mathcal{L} = \lambda_{age}\mathcal{L}_{age} + \lambda_{gender}\mathcal{L}_{gender} + \lambda_{mag}\mathcal{L}_{mag} + \lambda_{pre}\mathcal{L}_{pre}, \tag{18}$$

where each λ denotes a scalar weight of the corresponding loss term.

IV. Next Of Kin Dataset

A. Motivation

As shown in the related work section, while early datasets were limited in size both in terms of subjects as well kin relations (e.g., mother–son/daughter, father–son/daughter), more recent ones contain larger numbers of families with kin relationships that go beyond just two generations. Nevertheless, as shown in Table I and Fig. 5, existing datasets are still designed mostly for studying automatic kinship recognition techniques and typically do not contain data that allows for the development and training of high-quality image synthesis models. In order to train a high quality model, we introduce the Next Of Kin dataset that features high resolution face images, suitable for the task of kinship synthesis.

B. Data Collection

The dataset collection is based on the family relation data of the Families in the Wild dataset. Images are scraped from the web and manually filtered to ensure the high image quality. Manual filtering removed any images with watermarks, occluded faces, low resolution images and images with bad visual quality. We select 512×512 as the optimal image resolution and specify minimal acceptable resolution as 256×256. We search for additional high-resolution images of the subjects with low image count. The metadata was manually reviewed and corrected for the gathered subjects.

C. Data Processing

The image preprocessing steps are closely related to the ones used in construction of CelebAHQ dataset [2]. The first step consists of generating facial landmarks and cropping the facial image with proper orientation. We choose RetinaFace [45] as our main model for generating the face landmarks. All landmarks are visually inspected and manually corrected for any irregularities. Based on corrected landmarks, a face image is rotated and cropped. Finally, we add a corresponding $\mathcal{W}+$ latent code for each image by projecting each image to StyleGAN2 FFHQ model. We add gender and ethnicity information for every subject as well as age and emotion information for every image.

D. Dataset Characteristics

The proposed dataset consists of $3,690$ high resolution images of 553 subjects. The average number of images per subject is 6.67 ± 4.48 with a total of $127,719$ positive triplet combinations. We provide official training, validation and testing split, where we took special care as to avoid any overlap of child images between the splits. The distribution of the metadata statistics in terms of gender, ethnicity, age and emotion are shown in Fig. 6.

To evaluate the image quality of our proposed dataset we evaluate the face image quality of datasets. The evaluation is based on an face image quality assessment model CR-FIQA [46], which estimates an image quality score per facial image. The model predicts scores for images of cropped faces, therefore we only reports the results for datasets with such crops. We furthermore exclude results for Siblings HQ, where the model performs poorly due to domain gap between its training (in-the-wild facial images) and Siblings HQ images (professional images with uniform green background). The

Dataset	#Images	#Subjects	Kin Relations	Image Resolution	Characteristics	Aim	Quality Score
KinFaceW-I	1,066	1066	P-C	64 × 64	Tight crops	Kin. recognition	1.503
KinFaceW-II	2,000	2000	P-C	64 × 64	Tight crops	Kin. recognition	1.770
Families in the Wild (FIW)	30,725	10,676	P-C, GP-GC, Sib	124 × 108	Tight crops	Kin. recognition	1.906
TSKinface	2,589	2,589	P-C	64 × 64	Grayscale	Kin. recognition	1.988
Family101	14,816	607	P-C	150 × 120	Tree-like labels	Kin. recognition	2.069
UB Kinface	600	400	P-C	Variable (383 × 306)	Two generations	Kin. recognition	/
Siblings HQ	948	184	Sib	4256 × 2832	Uniform background	Kin. recognition	/
Siblings LQ	196	196	Sib	Variable (466 × 395)	Low image count	Kin. recognition	/
CornellKin	300	150	P-C	100 × 100	Low image count	Kin. recognition	2.139
Next Of Kin (NOK, ours)	3,690	553	P-C	512 × 512	High-resolution	Kin. synthesis	**2.236**

TABLE I: **Overview of kinship datasets reviewed in this paper and in comparison to the introduced NOK dataset.** Abbreviations describing *Subjects* stand for parent-child (P-C), grandparent-grandchild (GP-GC) and siblings (Sib). For datasets that contain images of variable resolution, we report the median resolution in parenthesis. The last column reports the average quality image score as assessed by an off-the-shelf model.

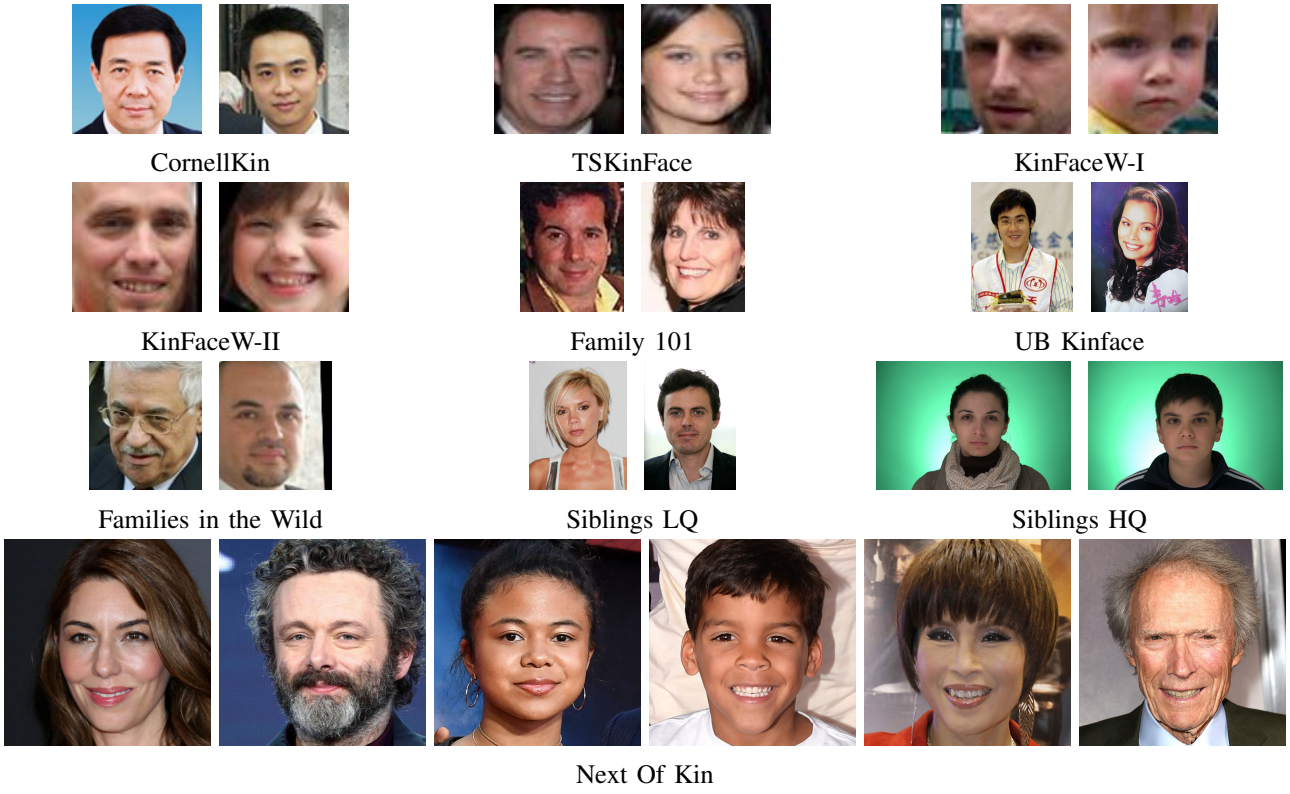


Fig. 5: **Visual comparison of images from kinship datasets.** Our proposed dataset, Next Of Kin, contains high-resolution, high fidelity images of subjects of varied ethnic groups. The images contain constant face positioning, which is especially useful for various tasks of image synthesis.

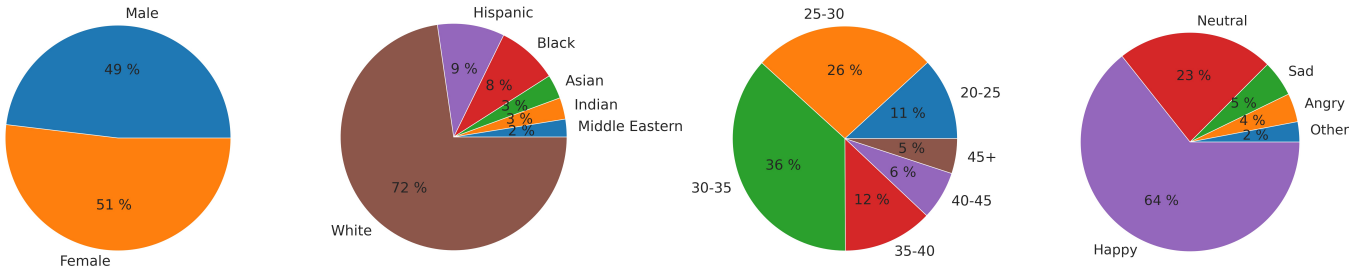
images are resized by the smaller axis and center-cropped to 112 × 112. The results are shown in Table I.

V. Experimental Setup

A. Datasets and Experimental Splits

Kinship Modelling. Two datasets are used to compare our proposed method with existing methods. The first one is Families in the Wild, which is also the main training dataset of contemporary models, while the second one is our proposed Next Of Kin dataset. For training on Families in the Wild we crop the original images with looser crops than the proposed dataset crops. We use the official training and validation splits for both datasets. The testing procedure on Families in the Wild is based on the official triplet parent-child data, but we expand it by using up to five images per person and creating all possible triplet image combinations to get more testing data. Furthermore, we remove out images whose diagonal is less than 90 pixels long. In the end, our testing split consists of 15,815 triplet images for FIW and 12,265 triplet images for NOK.

Age & Gender Manipulation. We use FFHQ Aging dataset [47] that provides several annotations of the original FFHQ dataset [3], including gender and age information. The



(a) Gender distribution (b) Ethnicity distribution (c) Age distribution (d) Emotion distribution

Fig. 6: **Next Of Kin subject and image characteristics in form of attributes distribution.** Gender and ethnicity distributions are calculated over persons included in the dataset. Age and emotion distributions are calculated over all dataset images.

ages are split into 10 age ranges $\{[0,2],[3,6],...,[70,120]\}$. We split the dataset into training and validation split, training split containing the first 90% of the data and validation split containing the rest.

B. Implementation Details and Methods

Kinship Module. We train our proposed model on FIW and NOK dataset using the official training–validation splits. The batch size is set to 16. We use Adam optimization algorithm [48] with learning rate $\eta = 3 \cdot 10^{-4}$. We identify weighting factors experimentally and are set to $\lambda_{id} = 1, \lambda_{tri} = 10^{-3}, \lambda_{mag} = 1$. The triplet margin value was set to $\delta = 0.1$.

ChildNet is evaluated in comparison with multiple competing models, i.e., DNANet [39], HeredityGAN [40] and StyleDNA [41]. We implement DNANet and HeredityGAN ourselves, while StyleDNA is based on the official release code. We modify the HeredityGAN model so as to provide age manipulation capability along with the original gender manipulation. The individual implementation details can be found in Appendix section.

Age & Gender Manipulation. Our age model A and gender model G are based on ResNet50 model [49]. We use cross-entropy loss, Adam optimization with learning rate $\eta = 3 \cdot 10^{-4}$ and a batch size of 64. When training the Age & Gender Manipulation Module we set the weighting factors to $\lambda_{age} = 1, \lambda_{gender} = 3, \lambda_{mag} = 0.3, \lambda_{pre} = 30$. The weighting factors were set based on visual analysis of the reconstructed, target and cycle images throughout initial experiments. We use a batch size of 2 and use no dropout or batch normalization. We use the Adam optimization algorithm with learning rate set to $\eta = 10^{-3}$. The dimensionality of the embeddings is set to $d = 128$. The model is trained for $46,500$ iterations.

VI. Results and Discussion

A. Comparison to State-Of-The-Art

Visual Analysis We first compare the visual results of models for the task of unconditional child synthesis, i.e. without additional age and gender information. We compare the model results in Fig. 7. We can observe that DNANet results follow the learned distribution with tightly cropped faces. The image quality of the faces is hindered as it does not use state-of-the-art decoder models like other methods. HeredityGAN generates high quality images, however the faces do not exhibit younger expected age of a child as it does not use a deaging mechanism. StyleDNA's results are convincing, but the variability of the results is relatively low to other methods and the face expression remains relatively constant with respect to input parental images. Our model ChildNet achieves the most visually pleasing results as it generates images of young child images that could be regarded as real child images.

Identity Similarity. To evaluate the quality of image identities produced by ChildNet, we adopt two different face recognition models, namely FaceNet [50] and ArcFace [42] models. We use these models to extract the face embeddings between the generated child and the real child, then evaluate the cosine similarity between the two embeddings. Unlike other approaches that chose to evaluate their performance on 100 positive and 100 negative child pairs, we propose a different evaluation strategy. We avoid evaluation method that is based on negative child pairs due to stochastic results caused by random negative child sampling. Furthermore, we argue that similarity to the real child (positive pair) is more important than the dissimilarity between the generated child and a random child (negative pair). Therefore, our evaluation is based on the cosine similarities of positive pairs, i.e. similarities between the generated and real child images. In comparison with other evaluation approaches, our testing procedure uses a much larger number of samples ($15,815$ triplets for FIW dataset and $12,265$ for NOK dataset).

The evaluation is done by constructing a histogram of predicted cosine similarities and a histogram of perfect cosine similarities (value 1 with 100 % confidence). We measure the distance between the predicted and perfect histograms with the earth mover's distance. We also report the mean of the cosine similarities. The histogram results for FIW and NOK database are shown in Figure 8 and the quantitative results are shown in Table II.

Perceptual Similarity. In this experiment we focus on measuring the image perceptual quality between the synthesized and real images. Higher image quality corresponds with less distortions and image artefacts in the image. We measure the perceptual quality using the the LPIPS metric [44] with AlexNet backend [51], which was shown to agree well with

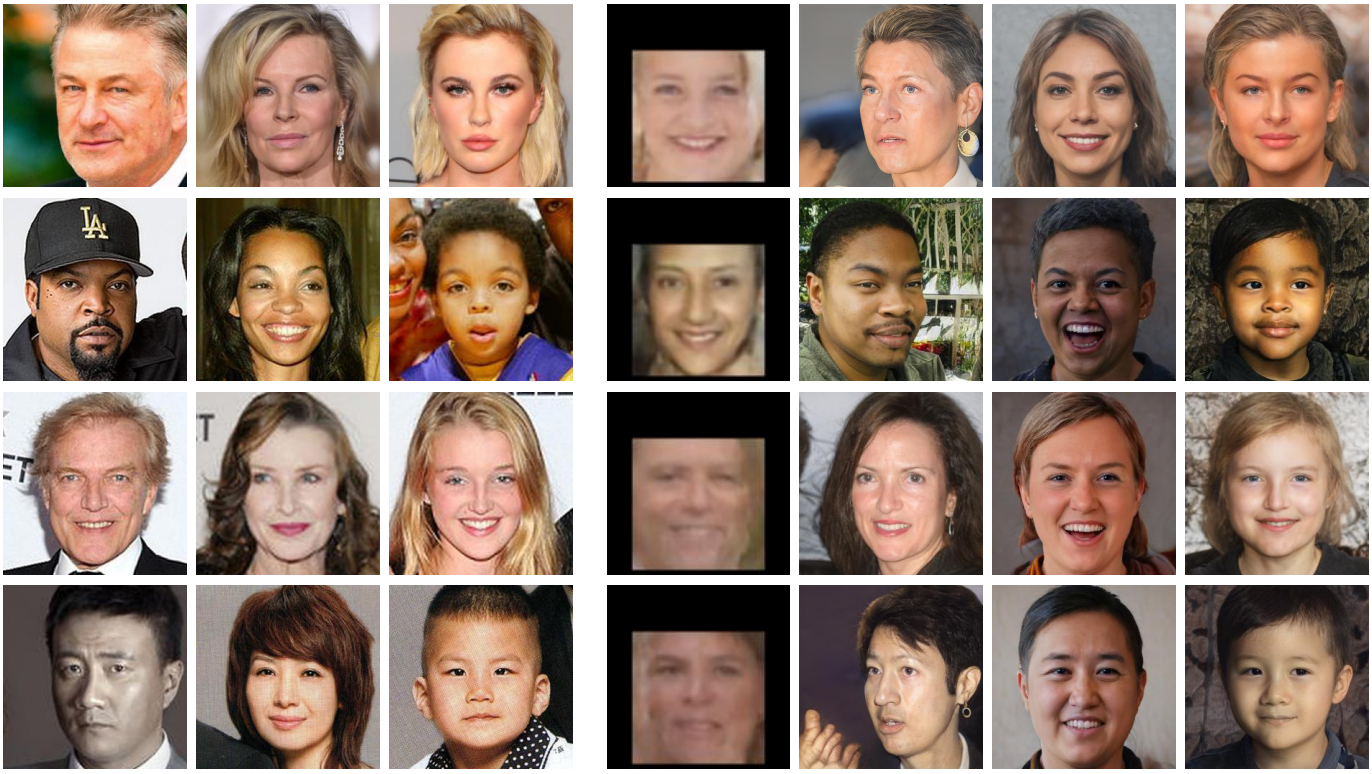


Fig. 7: **Visual comparison of model results.**. First three columns show father, mother and the real child image, respectively. The next four columns show the results for each tested model.

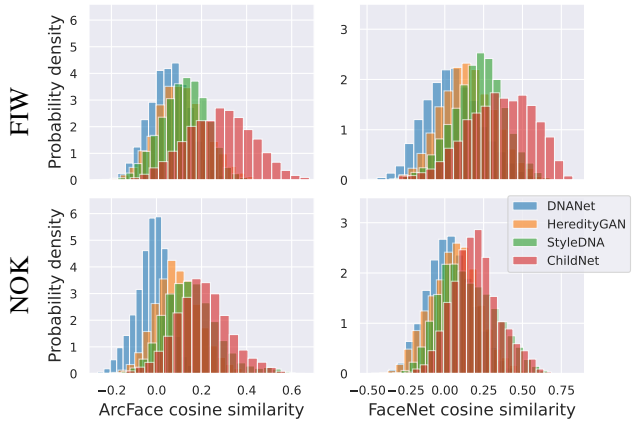


Fig. 8: **Comparison of verification scores.** The histograms are calculated based cosine embedding similarities, as obtained by ArcFace and FaceNet face recognition models.

TABLE II: **Face verification results.** We show results in terms of earth mover's distance and histogram mean between the predicted and ideal histograms of ArcFace cosine similarities

	Earth mover's distance ↓			
	Families In the Wild		Next Of Kin	
	ArcFace	FaceNet	ArcFace	FaceNet
DNANet	0.942	0.947	1.017	0.966
HeredityGAN	0.887	0.851	0.906	0.936
StyleDNA	0.877	0.783	0.847	0.853
ChildNet	**0.720**	**0.645**	**0.798**	**0.785**
	Histogram mean ↑			
	Families In the Wild		Next Of Kin	
	ArcFace	FaceNet	ArcFace	FaceNet
DNANet	0.058	0.053	−0.017	0.034
HeredityGAN	0.113	0.149	0.094	0.064
StyleDNA	0.123	0.217	0.153	0.147
ChildNet	**0.280**	**0.355**	**0.205**	**0.197**

TABLE III: **Perceptual quality results**. Average LPIPS distance between generated child images and real child images. Higher perceptual quality corresponds to lower LPIPS distance.

	Families In the Wild	Next Of Kin
DNANet	0.796	0.838
HeredityGAN	0.555	0.670
StyleDNA	**0.545**	0.633
ChildNet	0.546	**0.624**

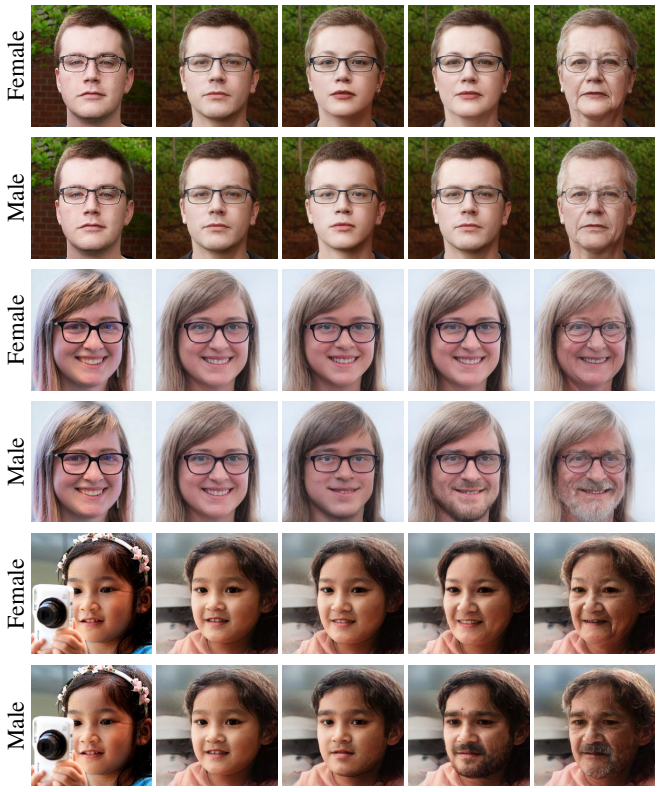


Fig. 9: **Age & gender manipulation results on FFHQ Aging dataset.** The first column shows the original images. E4e maps images into latent vectors, whose GAN decoder reconstructions are displayed in second column. The last three columns show the results in different age groups, where individual rows represent different gender.

human judgement. The results are shown in Table III. We can observe that the perceptual differences between methods that use StyleGAN as decoder are fairly small. The best results on Families In the Wild is achieved by StyleDNA model, with ChildNet being a close contender. Childnet achieves the best results on Next Of Kin dataset. We presume that the use of a non-extended latent space results in images with higher perceptual quality, albeit at the cost of lower model expression.

Age and Gender Manipulation. Here, we are testing the quality of the Age & Gender Manipulation Module. We first show several examples of manipulation on FFHQ Aging

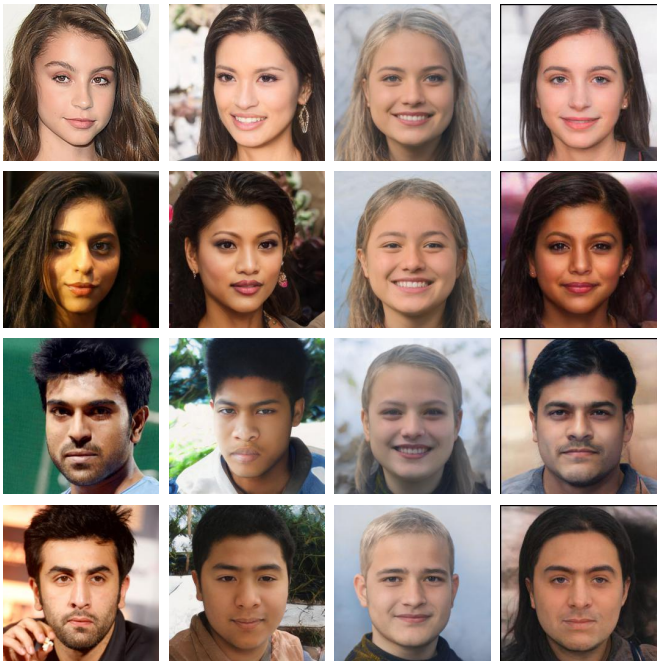


Fig. 10: **Visual examples of age and gender manipulation on NOK dataset.** The models are presented with age and gender information along the usual parental images when synthesizing the child image.

TABLE IV: **Face verification results with applied Age & Gender Module.** The results are presented for Families In the Wild and Next Of Kin datasets. ChildNet achieves the best results when compared with competing models.

Earth mover's distance↓		
	Families In the Wild	Next Of Kin
DNANet	0.913	0.982
HeredityGAN	0.855	0.854
StyleDNA	0.763	0.839
ChildNet	**0.701**	**0.779**
Histogram mean ↑		
	Families In the Wild	Next Of Kin
DNANet	0.087	0.018
HeredityGAN	0.145	0.146
StyleDNA	0.237	0.161
ChildNet	**0.299**	**0.221**

TABLE V: **Ablation study of ChildNet loss terms**. The study is done with respect to earth mover's distance and histogram mean. We use ArcFace model to calculate similarities.

	$\mathcal{L}_{id}=0$		$\mathcal{L}_{id}=1$	
Earth mover's distance ↓				
	$\mathcal{L}_t=0$	$\mathcal{L}_t=10^{-3}$	$\mathcal{L}_t=0$	$\mathcal{L}_t=10^{-3}$
$\mathcal{L}_{mag}=0$	0.848	0.847	0.794	0.810
$\mathcal{L}_{mag}=1$	0.846	0.842	0.793	**0.791**
Histogram mean ↑				
	$\mathcal{L}_{id}=0$		$\mathcal{L}_{id}=1$	
	$\mathcal{L}_t=0$	$\mathcal{L}_t=10^{-3}$	$\mathcal{L}_t=0$	$\mathcal{L}_t=10^{-3}$
$\mathcal{L}_{mag}=0$	0.152	0.153	0.206	0.190
$\mathcal{L}_{mag}=1$	0.155	0.158	0.207	**0.209**

TABLE VI: **Ablation study of ChildNet architecture** The study is done with respect to earth-mover's distance and histogram mean. We use ArcFace model to calculate similarities.

	Unstructured		Structured	
Earth mover's distance ↓				
	Scalars	Vectors	Scalars	Vectors
No Mut. BaseNet	0.833	0.802	0.832	0.803
With Mut. BaseNet	0.847	0.798	0.829	**0.791**
Histogram mean ↑				
	Unstructured		Structured	
	Scalars	Vectors	Scalars	Vectors
No Mut. BaseNet	0.167	0.198	0.168	0.197
With Mut. BaseNet	0.153	0.202	0.171	**0.209**

dataset, as shown in Fig. 9, where the input images are mapped into latent code by E4e encoder. The E4e reconstructions are shown in the second column. We can observe that the model is able to convincingly change the facial appearance to match the specified gender and age group.

Next, we evaluate the identity similarity when the models are presented with parental images as well as age and gender information of the child. We test the models on Next of Kin dataset, which already provides the required information, and on Families in the Wild, which only provides gender information. To get the age information, we use our trained age model A. We measure the cosine similarities with ArcFace recognition model. The results are shown in Table IV . Example visual results for NOK dataset are shown in Fig. 10, where we compare methods with similar visual quality.

B. ChildNet Characteristics

ChildNet Ablation Study. We ablate the loss terms and architecture of ChildNet model with respect to identity similarity as measured by ArcFace model. The results are reported in terms of earth-mover distance and average cosine similarity. In order to speed up training convergence, all the ablation tests are done on decreased model resolution of 256×256 with an increased learning rate of $\eta = 1 \cdot 10^{-3}$.

We analyze the effect of loss terms in the final loss function of Kinship Module in Eq. 8 to the verification similarity of the generated child images, as measured by the ArcFace model. Towards this end ChildNet model was trained on Next Of Kin dataset with different combinations of loss terms.

The results are shown in Table V. We can observe that a randomly initialized network already provides a decent result as the initialized parameter weights tend to average out the parental latent codes. The lack of a magnitude loss does not reflect in the quantitative results, however the results showed a considerable amount of visual artefacts, as shown in Fig. 11 due to the effect of adversarial model convergence. Adding a triplet boost helps achieve a slightly better result without impairing the visual quality.

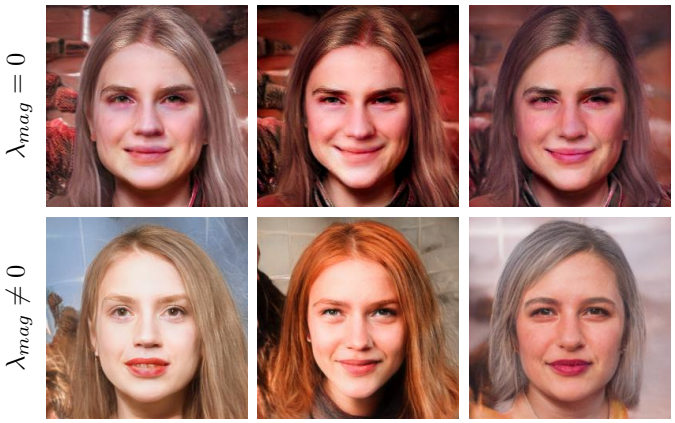


Fig. 11: **Visual results of ChildNet without λ_{mag} loss term.** Here we compare ChildNet synthesized images when the magnitude of epsilon vector ϵ is not regularized through magnitude loss term $\mathcal{L}_{mag}$, which corresponds to zeroing the loss weight $\lambda_{mag} = 0$. We can observe that the synthesized images exhibit artefacts due to adversarial properties of the model convergence.

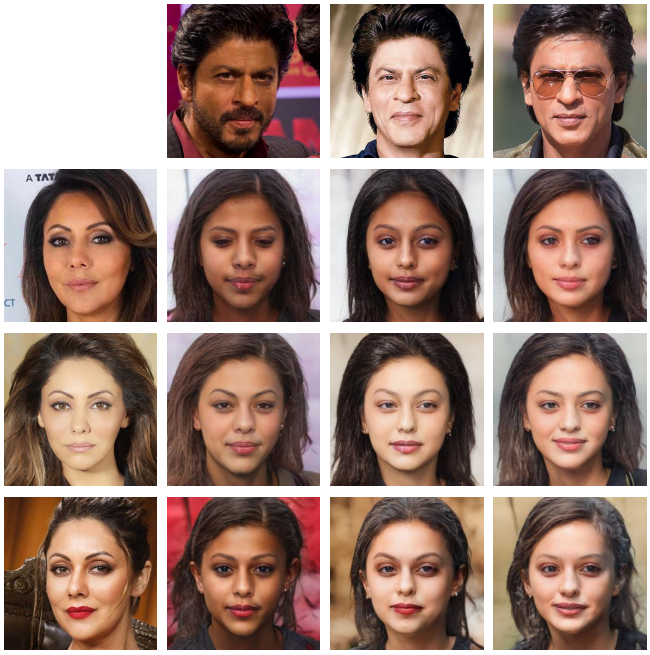


Fig. 12: **Comparison of ChildNet results given different real parent input images.** The changes in identity given different parent images are relatively small - child images all have long hair and similar face structure.

We further analyze the effects of ChildNet architecture. We analyze the results with respect to the following architecture settings:

- Presence of Mutation BaseNet, i.e. whether a latent offset is truly required,
- Structured model vs unstructured model, i.e. whether a separation of model on coarse, medium and fine part contributes to results,
- The formulation of BaseNet output, i.e. whether a scalar output (equivalent to a vector with repeated elements) can achieve the same results as vector prediction.

The results are shown in Table VI. We can observe the largest result degradation when the BaseNet model outputs a single value instead of a vector. We speculate the reason behind improved results is in the superior ability to disentangle the factors of variation in the GAN latent space when using a vector formulation. The effect of a structured model is most prominent when used in combination with Mutation BaseNet, implying the latter's strong influence in our proposed model.

Using Different Images of Same Parents. In this experiment we visualize the ChildNet output when presented with images, containing the same parent identity. The aim of the experiment is analyzing the synthesized image appearance change when presented with different inputs. An example result is presented in Fig. 12. The results show that the input image indeed affects the synthesized image, slightly changing characteristics such as skin colour, hair colour as well as face shape to a certain degree. The input images also strongly affect the face background.

Child Sampling and Image Variation Control. In this experiment we analyze the mechanism for image variation given fixed parental images. We propose to use the dropout mechanism [52] to ensure the child appearance variation. Typically, dropout mechanism is only used for training, randomly zeroing out elements to prevent co-adaption of neurons, while during the testing phase no elements are zeroed-out and the elements are appropriately scaled. We use dropout mechanism in test mode for point estimate image synthesis, which is used for most experiments. However, dropout in training mode can also be applied on trained network to provide image variation in generated child images. We show example results in Fig. 13, where the dropout rate is set to default training value $p_d = 0.5$. The variations mostly affect the stochastic aspects of faces such as face expression or exact hair placement while most basic facial attributes remain the same as in original image.

Fig. 13: **Illustration of image variety given fixed input using dropout mechanism.** The dropout value is set to the default one (as used during training stage) and is set to $p_d = 0.5$

Fig. 14: **Effect of modifying the dropout value.** By modifying the dropout percentage p_d we can adjust the image variation of the synthesized images. Higher values correspond to higher

Furthermore, we can vary the dropout rate, denoted as p_d, to control the amount of variation present in synthesized images. Note that the network was originally trained with dropout rate $p_d = 0.5$. As shown in Fig. 14, lowering the dropout rate results in reduced amount of images variation while the higher dropout rate increases the images variation.

Analyzing the Interpolation Coefficients. In this experiment, we are analyzing the distribution of the interpolation coefficients α as predicted by the Kinship Manipulation Module. A distribution of α values close to 0.5 would indicate that the model tends to synthesize an average face between the parents. The opposite case would indicate that the model is able to find certain parental latent components that it attends

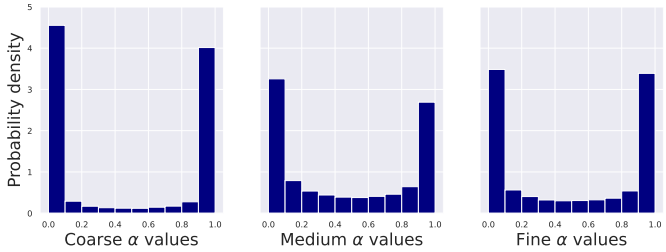


Fig. 15: **Analysis of interpolation coefficients.** This figure shows histograms of α interpolation coefficients for coarse, medium and fine part of ChildNet model, as obtained on the test split of the NOK dataset. The high probability density in the range tails indicate that the model is able to selectively attend to specific latent components of the selected parent.

to with a large weight.

We run the experiment on the test split of the NOK dataset, gathering the interpolation coefficients per every part (coarse, medium and fine). A histogram of the resulting coefficients is shown in Fig. 15. Interestingly, we find that the interpolation coefficients are largely distributed near their extreme values 0 and 1, most prominently at the coarse layer, which affects the most basic face appearance. The results suggest that the model is able to find certain parental latent components that should be taken almost exclusively from one of the parents.

Controlling the Dominant Parent. We implement a mechanism with which we can control the child appearance to look more like one of the parents. This allows us to modify the interpolation coefficients α^p in Eq. 2 by introducing a scalar $\delta_d \in \mathbb{R}^{[-1,1]}$ to be weighted more towards either of the input parental images. We can achieve a more mother-like appearance in the case of negative values and a more father-like appearance in the case of positive values. The updated interpolation coefficients are calculated as:

$$\alpha \leftarrow \begin{cases} \text{lerp}(\alpha, \mathbf{1}; |\delta_d|) & \text{if } \delta_d >= 0 \\ \text{lerp}(\alpha, \mathbf{0}; |\delta_d|) & \text{if } \delta_d < 0 \end{cases}, \tag{19}$$

where α is a concatenation of α^p parts, $\mathbf{1}$ and $\mathbf{0}$ denote vectors of ones and zeros with the same size as α and lerp denotes linear interpolation. Visual examples are presented in Fig. 16. The results show that we our mechanism indeed allows us to provide a mother-like or father-like appearance to synthesized image.

Results on Parents with Different Ethnicity. In this experiments, we are visually analyzing the synthesized child images, when the model is presented with parental images of different ethnicity. According to genetic studies [53], the skin colour of a child tends to follow an average skin colour of differently-coloured parents. Ideally, a trained kin synthesis model would also incorporate such genetic law. We select an image of person with African, Indian, Latino and White ethnicity and feed it to our ChildNet model. The results are shown in Fig. 17. From the synthesized images we can observe that ChildNet indeed generates faces with a skin color that is close to a blend between the two parents or one that is close to a single parent.

VII. Conclusion

In this paper we introduce ChildNet, a novel model for child synthesis. It consists of two main modules; Kinship Module and Age & Gender Manipulation Module. The Kinship Module is based on parental latent code manipulation through structural attention and mutation modules. With Age & Gender Manipulation Module we are able to specify the age and gender of the generated child image. Experiments show state-of-the-art performance with respect to verification similarity and strong performance with respect to image perceptual quality. We further show the versatility of the model with respect to generating multiple images, controlling the image variability and controlling the dominant parent. Lastly, due to limitations of existing kinship datasets for the task of child image synthesis, we introduce a novel kinship dataset Next Of Kin, which contains high resolution face images and metadata of kinship relations, age, gender, ethnicity and emotion.

References

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems (NIPS)*, 2014.
[2] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," in *International Conference on Learning Representations (ICLR)*, 2018.
[3] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4401–4410.
[4] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *International Conference on Learning Representations (ICLR)*, 2016.
[5] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8110–8119.
[6] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila, "Alias-free generative adversarial networks," *Advances in Neural Information Processing Systems*, vol. 34, 2021.
[7] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *International Conference on Computer Vision (ICCV)*, 2017, pp. 2794–2802.
[8] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International Conference on Machine Learning (ICML)*, 2017, pp. 214–223.
[9] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *Advances in Neural Information Processing Systems (NIPS)*, 2017, pp. 5767–5777.
[10] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," in *International Conference on Learning Representations (ICLR)*, 2018.
[11] L. Mescheder, A. Geiger, and S. Nowozin, "Which training methods for gans do actually converge?" in *International Conference on Machine learning (ICML)*, 2018, pp. 3481–3490.
[12] Z. Wang, Q. She, and T. E. Ward, "Generative adversarial networks in computer vision: A survey and taxonomy," *ACM Computing Surveys*, vol. 54, no. 2, pp. 1–38, 2021.
[13] D. Saxena and J. Cao, "Generative adversarial networks (gans) challenges, solutions, and future directions," *ACM Computing Surveys*, vol. 54, no. 3, pp. 1–42, 2021.
[14] R. Abdal, Y. Qin, and P. Wonka, "Image2stylegan: How to embed images into the stylegan latent space?" in *International Conference on Computer Vision (ICCV)*, 2019, pp. 4431–4440.
[15] ——, "Image2stylegan++: How to edit the embedded images?" in *Computer Vision and Pattern Recognition (CVPR)*, 2020.
[16] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. 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*, 2021, pp. 2287–2296.
[17] O. Tov, Y. Alaluf, Y. Nitzan, O. Patashnik, and D. Cohen-Or, "Designing an encoder for stylegan image manipulation," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 4, pp. 1–14, 2021.

Fig. 16: **Illustration of dominant parent mechanism.** We can control the dominant parental image with scalar $\delta_d \in \mathbb{R}^{[-1,1]}$, where the negative values result in a more mother-like appearance, while the positive values result in a more father-like appearance.

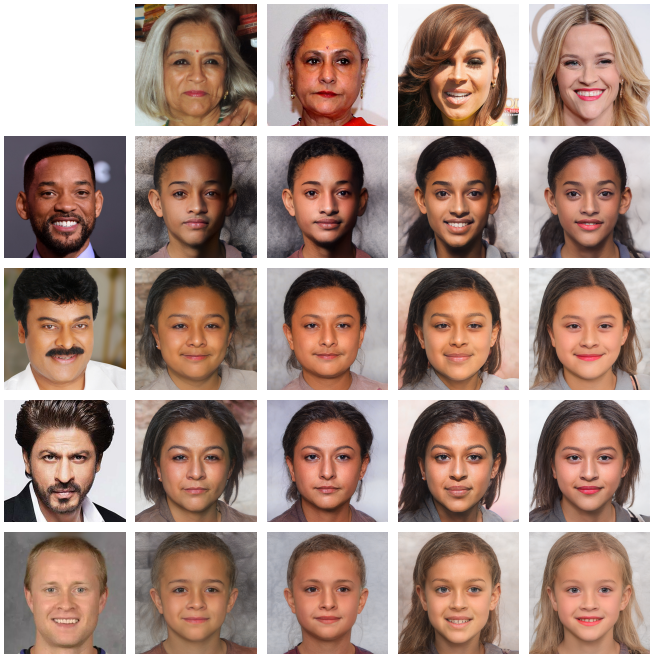


Fig. 17: **Model predictions on parents with different ethnicity.** We can observe that the ChildNet model synthesies child images, which exhibit the ethnic group, which can be reliably associated with one or both of the parents. The test is performed on NOK dataset.

[18] Y. Alaluf, O. Patashnik, and D. Cohen-Or, "Restyle: A residual-based stylegan encoder via iterative refinement," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6711–6720.

[19] Y. Alaluf, O. Tov, R. Mokady, R. Gal, and A. H. Bermano, "Hyperstyle: Stylegan inversion with hypernetworks for real image editing," *arXiv preprint arXiv:2111.15666*, 2021.

[20] Z. Wu, D. Lischinski, and E. Shechtman, "Stylespace analysis: Disentangled controls for stylegan image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 863–12 872.

[21] W. Xia, Y. Zhang, Y. Yang, J.-H. Xue, B. Zhou, and M.-H. Yang, "Gan inversion: A survey," *arXiv preprint arXiv:2101.05278*, 2021.

[22] Y. Shen, C. Yang, X. Tang, and B. Zhou, "Interfacegan: Interpreting the disentangled face representation learned by gans," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[23] Y. Shen and B. Zhou, "Closed-form factorization of latent semantics in gans," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1532–1540.

[24] E. Härkönen, A. Hertzmann, J. Lehtinen, and S. Paris, "Ganspace: Discovering interpretable gan controls," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9841–9850, 2020.

[25] M. Pernuš, V. Štruc, and S. Dobrišek, "High resolution face editing with masked gan latent code optimization," *arXiv preprint arXiv:2103.11135*, 2021.

[26] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski, "Styleclip: Text-driven manipulation of stylegan imagery," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2085–2094.

[27] J. Lu, X. Zhou, Y.-P. Tan, Y. Shang, and J. Zhou, "Neighborhood repulsed metric learning for kinship verification," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 2, pp. 331–345, 2013.

[28] J. P. Robinson, M. Shao, Y. Wu, and Y. Fu, "Families in the wild (fiw) large-scale kinship image database and benchmarks," in *Proceedings of*

the 24th ACM international conference on Multimedia, 2016, pp. 242–246.

[29] J. P. Robinson, M. Shao, Y. Wu, H. Liu, T. Gillis, and Y. Fu, "Visual kinship recognition of families in the wild," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 40, no. 11, pp. 2624–2637, 2018.

[30] X. Qin, X. Tan, and S. Chen, "Tri-subject kinship verification: Understanding the core of a family," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1855–1867, 2015.

[31] R. Fang, K. D. Tang, N. Snavely, and T. Chen, "Towards computational models of kinship verification," in *2010 IEEE International conference on image processing*. IEEE, 2010, pp. 1577–1580.

[32] R. Fang, A. C. Gallagher, T. Chen, and A. Loui, "Kinship classification by modeling facial feature heredity," in *2013 IEEE International Conference on Image Processing*. IEEE, 2013, pp. 2983–2987.

[33] S. Xia, M. Shao, J. Luo, and Y. Fu, "Understanding kin relationships in a photo," *IEEE Transactions on Multimedia*, vol. 14, no. 4, pp. 1046–1056, 2012.

[34] A. Vieira, Tiago F.and Bottino, A. Laurentini, and M. De Simone, "Detecting siblings in image pairs," *The Visual Computer*, vol. 30, no. 12, pp. 1333–1345, Dec 2014.

[35] S. Xia, M. Shao, and Y. Fu, "Kinship verification through transfer learning," in *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

[36] M. Shao, S. Xia, and Y. Fu, "Genealogical face recognition based on ub kinface database," in *CVPR 2011 WORKSHOPS*. IEEE, 2011, pp. 60–65.

[37] S. Wang, J. P. Robinson, and Y. Fu, "Kinship verification on families in the wild with marginalized denoising metric learning," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, 2017, pp. 216–221.

[38] J. P. Robinson, Y. Yin, Z. Khan, M. Shao, S. Xia, M. Stopa, S. Timoner, M. A. Turk, R. Chellappa, and Y. Fu, "Recognizing families in the wild (rfiw): The 4th edition," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE, 2020, pp. 857–862.

[39] P. Gao, S. Xia, J. Robinson, J. Zhang, C. Xia, M. Shao, and Y. Fu, "What will your child look like? dna-net: Age and gender aware kin face synthesizer," *arXiv preprint arXiv:1911.07014*, 2019.

[40] X. Cui, W. Zhou, Y. Hu, W. Wang, and H. Li, "Heredity-aware child face image generation with latent space disentanglement," *arXiv preprint arXiv:2108.11080*, 2021.

[41] C.-H. Lin, H.-C. Chen, L.-C. Cheng, S.-C. Hsu, J.-C. Chen, and C.-Y. Wang, "Styledna: A high-fidelity age and gender aware kinship face synthesizer," in *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. IEEE, 2021, pp. 1–8.

[42] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[43] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[44] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[45] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-shot multi-level face localisation in the wild," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5202–5211.

[46] F. Boutros, M. Fang, M. Klemt, B. Fu, and N. Damer, "Cr-fiqa: Face image quality assessment by learning sample relative classifiability," *arXiv preprint arXiv:2112.06592*, 2021.

[47] R. Or-El, S. Sengupta, O. Fried, E. Shechtman, and I. Kemelmacher-Shlizerman, "Lifespan age transformation synthesis," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[48] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2014.

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[50] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[51] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[52] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[53] G. A. Harrison, "The measurement and inheritance of skin colour in man," *The Eugenics Review*, vol. 49, no. 2, p. 73, 1957.