Temporal Non-Volume Preserving Approach to Facial Age-Progression and Age-Invariant Face Recognition

Chi Nhan Duong [1,2], Kha Gia Quach [1,2], Khoa Luu [2] , T. Hoang Ngan Le [2] and Marios Savvides [2]
[1] Computer Science and Software Engineering, Concordia University, Montréal, Québec, Canada
[2] CyLab Biometrics Center and the Department of Electrical and Computer Engineering,
Carnegie Mellon University, Pittsburgh, PA, USA
{chinhand, kquach, kluu, thihoanl}@andrew.cmu.edu, msavvid@ri.cmu.edu

Abstract

Modeling the long-term facial aging process is extremely challenging due to the presence of large and non-linear variations during the face development stages. In order to efficiently address the problem, this work first decomposes the aging process into multiple short-term stages. Then, a novel generative probabilistic model, named Temporal Non-Volume Preserving (TNVP) transformation, is presented to model the facial aging process at each stage. Unlike Generative Adversarial Networks (GANs), which requires an empirical balance threshold, and Restricted Boltzmann Machines (RBM), an intractable model, our proposed TNVP approach guarantees a tractable density function, exact inference and evaluation for embedding the feature transformations between faces in consecutive stages. Our model shows its advantages not only in capturing the non-linear age related variance in each stage but also producing a smooth synthesis in age progression across faces. Our approach can model any face in the wild provided with only four basic landmark points. Moreover, the structure can be transformed into a deep convolutional network while keeping the advantages of probabilistic models with tractable log-likelihood density estimation. Our method is evaluated in both terms of synthesizing age-progressed faces and cross-age face verification and consistently shows the state-of-the-art results in various face aging databases, i.e. FG-NET, MORPH, AginG Faces in the Wild (AGFW), and Cross-Age Celebrity Dataset (CACD). A large-scale face verification on Megaface challenge 1 is also performed to further show the advantages of our proposed approach.

1. Introduction

Face age progression is known as the problem of aesthetically predicting individual faces at different ages. Origin from finding the missing children, age progression has shown its potential in many applications varied from wanted fugitives, cross-age face verification, security system to other cosmetic studies against aging. Aesthetically synthesizing faces of a subject at different development stages is a very challenging task. Human aging is complicated and differs from one individual to the next. Both *intrinsic factors* such as heredity, gender, and ethnicity, and *extrinsic factors*, i.e. environment and living styles, jointly contribute to this process and create large aging variations between individuals. As illustrated in Figure 1, given a face of a subject at the age of 34 [17], a set of closely related family faces has to be provided to a forensic artist as references to generate multiple outputs of his faces at 40s, 50s, 60s, and 70s.

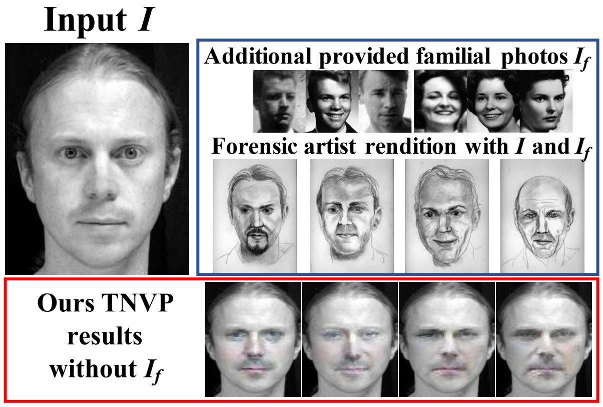

Input I
Additional provided familial photos I_f
Forensic artist rendition with I and I_f
Ours TNVP results without I_f

Figure 1: An illustration of age progression from forensic artist and our TNVP model. Given an input I of a subject at 34 years old [17], a forensic artist rendered his age-progressed faces at 40s, 50s, 60s and 70s by reference to his familial photos I_f. Without using I_f, our TNVP can aesthetically produce his age-progressed faces.

In recent years, automatic age progression has become a prominent topic and attracted considerable interest from the computer vision community. The conventional meth-

Table 1: Comparing the properties between our TNVP approach and other age progression methods, where ✗ represents *unknown* or *not directly applicable* properties. Deep learning (DL), Dictionary (DICT), Prototype (PROTO), AGing pattErn Subspace (AGES), Composition (COMP), Probabilistic Graphical Models (PGM), Log-likelihood (LL), Adversarial (ADV)

	Our TNVP	TRBM[15]	RNN[26]	acGAN[2]	HFA[28]	CDL[22]	IAAP[10]	HAGES[25]	AOG[23]
Model Type	DL	DL	DL	DL	DICT	DICT	PROTO	AGES	COMP
Architecture	PGM+CNN	PGM	CNN	CNN	Bases	Bases	✗	✗	Graph
Loss Function	LL	LL	ℓ_2	ADV+ℓ_2	LL+ℓ_0	$\ell_2 + \ell_1$	✗	ℓ_2	✗
Tractable	✓	✗	✓	✓	✓	✓	✗	✓	✗
Non-Linearity	✓	✓	✓	✓	✗	✗	✗	✗	✗

ods [6, 12, 16, 23] simulated face aging by adopting parametric linear models such as Active Appearance Models (AAMs) and 3D Morphable Models (3DMM) to interpret the face geometry and appearance before combining with physical rules or anthropology prior knowledge. Some other approaches [3, 10, 20] predefined some prototypes and transferred the difference between them to produce age-progressed face images. However, since face aging is a non-linear process, these linear models have lots of difficulties and the quality of their synthesized results is still limited. Recently, deep learning based models [15, 26] have also come into place and produced more plausible results. In [26], Recurrent Neural Networks (RNN) are used to model the intermediate states between two consecutive age groups for better aging transition. However, it still has the limitations of producing blurry results by the use of a fixed reconstruction loss function, i.e. ℓ_2-norm. Meanwhile, with the advantages of graphical models, the Temporal Restricted Boltzmann Machines (TRBM) has shown its potential in the age progression task [15]. However, its partition function is intractable and needs some approximations during training.

Contributions of this work: This paper presents a novel generative probabilistic model, named Temporal Non-Volume Preserving (TNVP) transformation, for age progression. This approach enjoys the strengths of both probabilistic graphical models to produce better synthesis quality by avoiding the regular reconstruction loss function, and deep residual networks (ResNet) [8] to improve the highly non-linear feature generation. The proposed TNVP guarantees a *tractable* log-likelihood density estimation, *exact* inference and evaluation for embedding the feature transformations between faces in consecutive age groups.

In our framework, the long-term face aging is first considered as a composition of short-term stages. Then TNVP models are constructed to capture the facial aging features transforming between two successive age groups. By incorporating the design of ResNet [8] in the structure, our TNVP is able to efficiently capture the non-linear facial aging feature related variance. In addition, it can be robustly employed on face images in-the-wild without strict alignments or any complicated preprocessing steps. Finally, the connections between latent variables can act as "memory" and contribute to produce a smooth age progression while preserving the identity throughout the transitions.

In summary, the novelties of our approach are three-fold. **(1)** We propose a novel generative probabilistic models with tractable density function to capture the non-linear age variances. **(2)** The aging transformation can be effectively modeled using our TNVP. Similar to other probabilistic models, our TNVP is more advanced in term of embedding the complex aging process. **(3)** Unlike previous aging approaches that suffer from a burdensome preprocessing to produce the dense correspondence between faces, our model is able to synthesize realistic faces given any input face in the wild. Table 1 compares the properties between our TNVP approach and other age progression methods.

2. Related Work

This section reviews various age progression approaches which can be divided into four groups: *prototyping*, *modeling*, *reconstructing*, and *deep learning-based approaches*.

Prototyping approaches use the age prototypes to synthesize new face images. The average faces of people in the same age group are used as the prototypes [20]. The input image can be transformed into the age-progressed face by adding the differences between the prototypes of two age groups [3]. Kemelmacher-Shlizerman et al. [10] proposed to construct sharper average prototype faces from a large-scale set of images in combining with subspace alignment and illumination normalization.

Modeling-based approaches represent facial shape and appearance via a set of parameters and model facial aging process via aging functions. Lanitis et al. [12] and Pattersons et al. [16] proposed to use AAMs parameters together with four aging functions for both general and specific aging processes. Luu et al. [14] incorporated common facial features of siblings and parents to age progression. Geng et al. [6] proposed an AGing pattErn Subspace (AGES) approach to construct a subspace for *aging patterns* as a chronological sequence of face images. Later, Tsai et al. [25] improved the stability of AGES by adding subject's characteristics clue. Suo et al. [23, 24] modeled a face us-

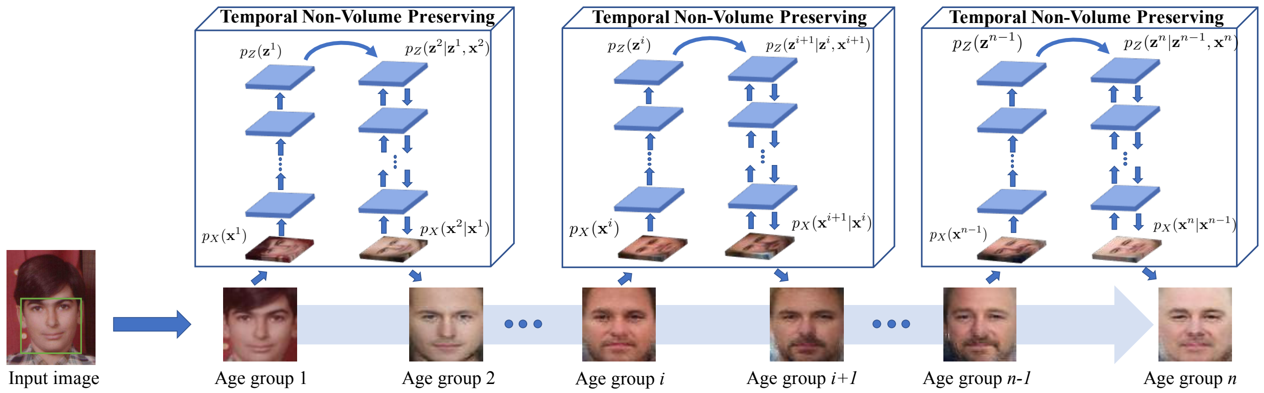
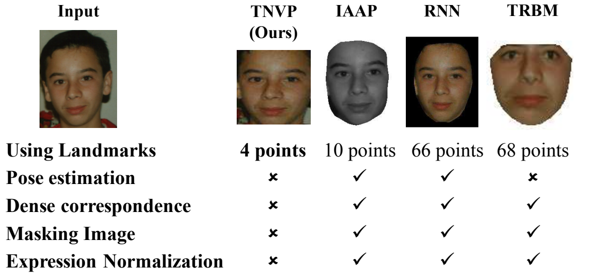

Temporal Non-Volume Preserving
$p_Z(\mathbf{z}^1)$
$p_Z(\mathbf{z}^2|\mathbf{z}^1, \mathbf{x}^2)$
$p_X(\mathbf{x}^1)$
$p_X(\mathbf{x}^2|\mathbf{x}^1)$
Temporal Non-Volume Preserving
$p_Z(\mathbf{z}^i)$
$p_Z(\mathbf{z}^{i+1}|\mathbf{z}^i, \mathbf{x}^{i+1})$
$p_X(\mathbf{x}^i)$
$p_X(\mathbf{x}^{i+1}|\mathbf{x}^i)$
Temporal Non-Volume Preserving
$p_Z(\mathbf{z}^{n-1})$
$p_Z(\mathbf{z}^n|\mathbf{z}^{n-1}, \mathbf{x}^n)$
$p_X(\mathbf{x}^{n-1})$
$p_X(\mathbf{x}^n|\mathbf{x}^{n-1})$
Input image
Age group 1
Age group 2
Age group *i*
Age group *i+1*
Age group *n-1*
Age group *n*

Figure 2: The proposed TNVP based age progression framework. The long-term face aging is decomposed into multiple short-term stages. Then given a face in age group i, our TNVP model is applied to synthesize face in the next age group. Each side of our TNVP is designed as a deep ResNet network to efficiently capture the non-linear facial aging features.



	TNVP (Ours)	IAAP	RNN	TRBM
Using Landmarks	**4 points**	10 points	66 points	68 points
Pose estimation	×	✓	✓	×
Dense correspondence	×	✓	✓	✓
Masking Image	×	✓	✓	✓
Expression Normalization	×	✓	✓	✓

Figure 3: Comparisons between the preprocessing processes of our approach and other aging approaches: IAAP [10], RNN based [26], and TRBM based [15] models. Our preprocessing is easy to run, less dependent on the landmarking tools, and efficiently deals with in-the-wild faces. ✓represents "included in the preprocessing steps".

ing a three-layer And-Or Graph (AOG) of smaller parts, i.e. eyes, nose, mouth, etc. and learned the aging process for each part by applying a Markov chain.

Reconstructing-based methods reconstruct the aging face from the combination of an aging basis in each group. Shu et al. [22] proposed to build aging coupled dictionaries (CDL) to represent personalized aging pattern by preserving personalized facial features. Yang et al. [28] modeled person-specific and age-specific factors separately via sparse representation hidden factor analysis (HFA).

Recently, *deep learning-based approaches* are being developed to exploit the power of deep learning methods. Duong et al. [15] employed TRBM to model the non-linear aging process with geometry constraints and spatial RBMs to model a sequence of reference faces and wrinkles of adult faces. Wang et al. [26] modeled aging sequences using a RNN with two-layer gated recurrent unit (GRU). Conditional Generative Adversarial Networks (cGAN) is also applied to synthesize aged images in [2].

3. Our Proposed Method

The proposed TNVP age-progression architecture consists of three main steps. (1) Preprocessing; (2) Face variation modeling via mapping functions; and (3) Aging transformation embedding. With the structure of the mapping function, our TNVP model is tractable and highly non-linear. It is optimized using a log-likelihood objective function that produces sharper age-progressed faces compared to the regular ℓ_2-norm based reconstruction models. Figure 2 illustrates our TNVP-based age progression architecture.

3.1. Preprocessing

Figure 3 compares our preprocessing step with other recent age progression approaches, including Illumination Aware Age Progression (IAAP) [10], RNN based [26], and TRBM based Age Progression [15] models. In those approaches, burdensome face normalization steps are applied to obtain the dense correspondence between faces. The use of a large number of landmark points makes them highly dependent on the stability of landmarking methods that are challenged in the wild conditions. Moreover, masking the faces with a predefined template requires a separate shape adjustment for each age group in later steps.

In our method, given an image, the facial region is simply detected and aligned according to fixed positions of four landmark points, i.e. two eyes and two mouth corners. By avoiding complicated preprocessing steps, our proposed architecture has two advantages. Firstly, a small number of landmark points, i.e. only four points, leverages the dependency to the quality of any landmarking method. Therefore, it helps to increase the robustness of the system. Secondly, parts of the image background are still included, and thus it implicitly embeds the shape information during the modeling process. From the experimental results, one can easily notice the change of the face shape when moving from one

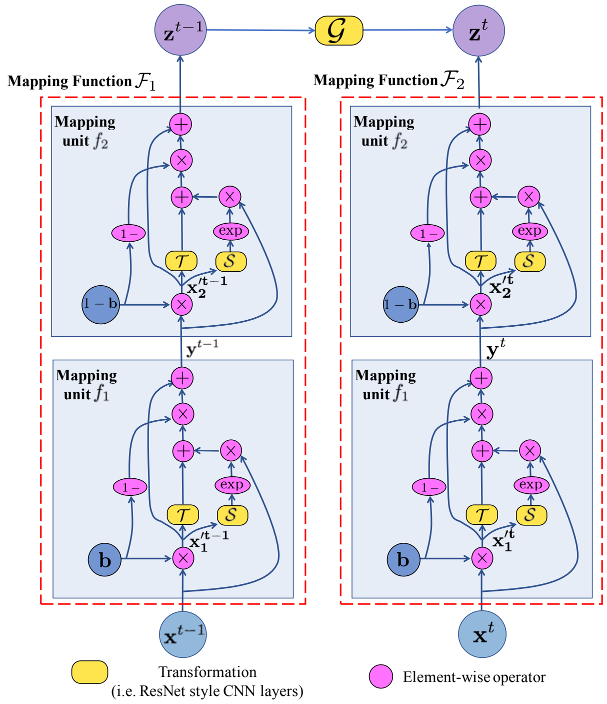


Figure 4: Our proposed TNVP structure with two mapping units. Both transformations $\mathcal{S}$ and $\mathcal{T}$ can be easily formulated as compositions of CNN layers.

age group to the next.

3.2. Face Aging Modeling

Let $\mathcal{I} \subset \mathbb{R}^D$ be the image domain and $\{\mathbf{x}^t, \mathbf{x}^{t-1}\} \in \mathcal{I}$ be observed variables encoding the texture of face images at age group t and $t-1$, respectively. In order to embed the aging transformation between these faces, we first define a bijection mapping function from the image space $\mathcal{I}$ to a latent space $\mathcal{Z}$ and then model the relationship between these latent variables. Formally, let $\mathcal{F} : \mathcal{I} \to \mathcal{Z}$ define a bijection from an observed variable $\mathbf{x}$ to its corresponding latent variable $\mathbf{z}$ and $\mathcal{G} : \mathcal{Z} \to \mathcal{Z}$ be an aging transformation function modeling the relationships between variables in latent space. As illustrated in Figure 4, the relationships between variables are defined as in Eqn. (1).

$$\begin{aligned} \mathbf{z}^{t-1} &= \mathcal{F}_1(\mathbf{x}^{t-1}; \theta_1) \\ \mathbf{z}^{t} &= \mathcal{H}(\mathbf{z}^{t-1}, \mathbf{x}^t; \theta_2, \theta_3) \\ &= \mathcal{G}(\mathbf{z}^{t-1}; \theta_3) + \mathcal{F}_2(\mathbf{x}^t; \theta_2) \end{aligned} \tag{1}$$

where $\mathcal{F}_1, \mathcal{F}_2$ define the bijections of $\mathbf{x}^{t-1}$ and $\mathbf{x}^t$ to their latent variables, respectively. $\mathcal{H}$ denotes the summation of $\mathcal{G}(\mathbf{z}^{t-1}; \theta_3)$ and $\mathcal{F}_2(\mathbf{x}^t; \theta_2)$. $\theta = \{\theta_1, \theta_2, \theta_3\}$ present the parameters of functions $\mathcal{F}_1, \mathcal{F}_2$ and $\mathcal{G}$, respectively. Indeed, given a face image in age group $t-1$, the probability density function can be formulated as in Eqn. (2).

$$\begin{aligned} p_{X^t}(\mathbf{x}^t|\mathbf{x}^{t-1}; \theta) &= p_{X^t}(\mathbf{x}^t|\mathbf{z}^{t-1}; \theta) \\ &= p_{Z^t}(\mathbf{z}^t|\mathbf{z}^{t-1}; \theta) \left| \frac{\partial \mathcal{H}(\mathbf{z}^{t-1}, \mathbf{x}^t; \theta_2, \theta_3)}{\partial \mathbf{x}^t} \right| \\ &= p_{Z^t}(\mathbf{z}^t|\mathbf{z}^{t-1}; \theta) \left| \frac{\partial \mathcal{F}_2(\mathbf{x}^t; \theta_2)}{\partial \mathbf{x}^t} \right| \end{aligned} \tag{2}$$

where $p_{X^t}(\mathbf{x}^t|\mathbf{x}^{t-1}; \theta)$ and $p_{Z^t}(\mathbf{z}^t|\mathbf{z}^{t-1}; \theta)$ are the distribution of $\mathbf{x}^t$ conditional on $\mathbf{x}^{t-1}$ and the distribution of $\mathbf{z}^t$ conditional on $\mathbf{z}^{t-1}$, respectively. In Eqn. (2), the second equality is obtained using the change of variable formula. $\frac{\partial \mathcal{F}_2(\mathbf{x}^t; \theta_2)}{\partial \mathbf{x}^t}$ is the Jacobian. Using this formulation, instead of estimating the density of a sample $\mathbf{x}^t$ conditional on $\mathbf{x}^{t-1}$ directly in the complicated high-dimensional space $\mathcal{I}$, the assigned task can be accomplished by computing the density of its corresponding latent point $\mathbf{z}^t$ given $\mathbf{z}^{t-1}$ associated with the Jacobian determinant $\left| \frac{\partial \mathcal{F}_2(\mathbf{x}^t; \theta_2)}{\partial \mathbf{x}^t} \right|$. There are some recent efforts to achieve the tractable inference process via approximations [11] or specific functional forms [5, 7, 13]. Section 3.3 introduces a non-linear bijection function that enables the exact and tractable mapping from the image space $\mathcal{I}$ to a latent space $\mathcal{Z}$ where the density of its latent variables can be computed exactly and efficiently. As a result, the density evaluation of the whole model becomes exact and tractable.

3.3. Mapping function as CNN layers

In general, a bijection function between two high-dimensional domains, i.e. image and latent spaces, usually produces a large Jacobian matrix and is expensive for its determinant computation. In order to enable the tractable property for $\mathcal{F}$ with lower computational cost, this section introduces a non-linear mapping unit structure that maps variables from image space to intermediate latent spaces where the density can be computed exactly and efficiently. Then the bijection mapping function $\mathcal{F}$ is formulated as a composition of mapping units. With this structure, $\mathcal{F}$ can be efficiently set up as a deep convolutional network and enjoys the strengths of both deep networks and probabilistic models with tractable log-likelihood density estimation.

3.3.1 Mapping unit

Given an input $\mathbf{x}$, a unit $f : \mathbf{x} \to \mathbf{y}$ defines a mapping from $\mathbf{x}$ to an intermediate latent state $\mathbf{y}$ as in Eqn. (3).

$$\mathbf{y} = \mathbf{x}' + (1 - \mathbf{b}) \odot \left[\mathbf{x} \odot \exp(\mathcal{S}(\mathbf{x}')) + \mathcal{T}(\mathbf{x}') \right] \tag{3}$$

where $\mathbf{x}' = \mathbf{b} \odot \mathbf{x}$; $\odot$ denotes the Hadamard product; $\mathbf{b} = [1, \cdots, 1, 0, \cdots, 0]$ is a binary mask where the first d elements of $\mathbf{b}$ is set to one and the rest is zero; $\mathcal{S}$ and $\mathcal{T}$ represent the scale and the translation functions, respectively.

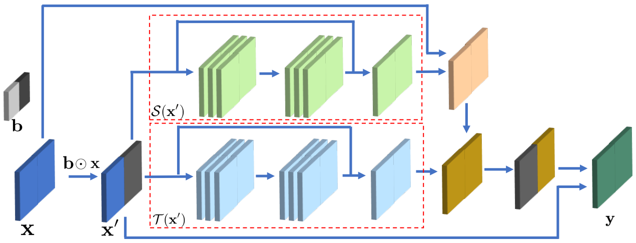


Figure 5: An illustration of mapping unit f whose transformations $\mathcal{S}$ and $\mathcal{T}$ are represented with 1-residual-block CNN network.

The Jacobian of this transformation unit is given by

$$\begin{aligned}\frac{\partial f}{\partial \mathbf{x}} &= \begin{bmatrix} \frac{\partial \mathbf{y}_{1:d}}{\partial \mathbf{x}_{1:d}} & \frac{\partial \mathbf{y}_{1:d}}{\partial \mathbf{x}_{d+1:D}} \\ \frac{\partial \mathbf{y}_{d+1:D}}{\partial \mathbf{x}_{1:d}} & \frac{\partial \mathbf{y}_{d+1:D}}{\partial \mathbf{x}_{d+1:D}} \end{bmatrix} \\ &= \begin{bmatrix} \mathbb{I}_d & 0 \\ \frac{\partial \mathbf{y}_{d+1:D}}{\partial \mathbf{x}_{1:d}} & \operatorname{diag}\left(\exp(\mathcal{S}(\mathbf{x}_{1:d}))\right) \end{bmatrix}\end{aligned} \tag{4}$$

where $\operatorname{diag}\left(\exp(\mathcal{S}(\mathbf{x}_{1:d}))\right)$ is the diagonal matrix such that $\exp(\mathcal{S}(\mathbf{x}_{1:d}))$ is their diagonal elements. This form of $\frac{\partial f}{\partial \mathbf{x}}$ provides two nice properties for the mapping unit f. Firstly, since the Jacobian matrix $\frac{\partial f}{\partial \mathbf{x}}$ is triangular, its determinant can be efficiently computed as,

$$\left|\frac{\partial f}{\partial \mathbf{x}}\right| = \prod_j \exp(s_j) = \exp\left(\sum_j s_j\right) \tag{5}$$

where $\mathbf{s} = \mathcal{S}(\mathbf{x}_{1:d})$. This property also introduces the tractable feature for f. Secondly, the Jacobians of the two functions $\mathcal{S}$ and $\mathcal{T}$ are not required in the computation of $\left|\frac{\partial f}{\partial \mathbf{x}}\right|$. Therefore, any non-linear function can be chosen for $\mathcal{S}$ and $\mathcal{T}$. From this property, the functions $\mathcal{S}$ and $\mathcal{T}$ are set up as a composition of CNN layers in ResNet (i.e. residual networks) [8] style with skip connections. This way, high level features can be extracted during the mapping process and improve the generative capability of the proposed model. Figure 5 illustrates the structure of a mapping unit f. The inverse function $f^{-1}: \mathbf{y} \rightarrow \mathbf{x}$ is also derived as

$$\mathbf{x} = \mathbf{y}' + (1 - \mathbf{b}) \odot \left[(\mathbf{y} - \mathcal{T}(\mathbf{y}')) \odot \exp(-\mathcal{S}(\mathbf{y}'))\right] \tag{6}$$

where $\mathbf{y}' = \mathbf{b} \odot \mathbf{y}$.

3.3.2 Mapping function

The bijection mapping function $\mathcal{F}$ is formulated by composing a sequence of mapping units $\{f_1, f_2, \cdots, f_n\}$.

$$\mathcal{F} = f_1 \circ f_2 \circ \cdots \circ f_n \tag{7}$$

The computation of the Jacobian of $\mathcal{F}$ is no more difficult than its units and still remains tractable.

$$\frac{\partial \mathcal{F}}{\partial \mathbf{x}} = \frac{\partial f_1}{\partial \mathbf{x}} \cdot \frac{\partial f_2}{\partial f_1} \cdots \frac{\partial f_n}{\partial f_{n-1}} \tag{8}$$

Similarly, the derivations of its determinant and inverse are

$$\begin{aligned}\left|\frac{\partial \mathcal{F}}{\partial \mathbf{x}}\right| &= \left|\frac{\partial f_1}{\partial \mathbf{x}}\right| \cdot \left|\frac{\partial f_2}{\partial f_1}\right| \cdots \left|\frac{\partial f_n}{\partial f_{n-1}}\right| \\ \mathcal{F}^{-1} &= (f_1 \circ f_2 \circ \cdots \circ f_n)^{-1} = f_1^{-1} \circ f_2^{-1} \circ \cdots \circ f_n^{-1}\end{aligned} \tag{9}$$

Since each mapping unit leaves part of its input unchanged (i.e. due to the zero-part of the mask $\mathbf{b}$), we alternatively change the binary mask $\mathbf{b}$ to $1 - \mathbf{b}$ in the sequence so that every component of $\mathbf{x}$ can be jointed through the mapping process. As mentioned in the previous section, since each mapping unit is set up as a composition of CNN layers, the bijection $\mathcal{F}$ with the form of Eqn. (7) becomes a deep convolutional networks that maps its observed variable $\mathbf{x}$ in $\mathcal{I}$ to a latent variable $\mathbf{z}$ in $\mathcal{Z}$.

3.4. The aging transform embedding

In the previous section, we present the invertible mapping function $\mathcal{F}$ between a data distribution p_X and a latent distribution p_Z. In general, p_Z can be chosen as a prior probability distribution such that it is simple to compute and its latent variable z is easily sampled. In our system, a Gaussian distribution is chosen for p_Z, but notice that our proposed model can still work well with any other prior distributions. Since the connections between $\mathbf{z}^{t-1}$ and $\mathbf{z}^t$ embed the relationship between variables of different Gaussian distributions, we further assume that their joint distribution is a Gaussian. From Eqn. (1) and Figure 4, the latent variable $\mathbf{z}^t$ is computed from two sources: (1) the mapping from observed variable $\mathbf{x}^t$ defined by $\mathcal{F}_2(\mathbf{x}^t; \theta_2)$ and (2) the aging transformation from $\mathbf{z}^{t-1}$ defined by $\mathcal{G}(\mathbf{z}^{t-1}; \theta_3)$. The transformation $\mathcal{G}$ between $\mathbf{z}^{t-1}$ and $\mathbf{z}^t$ is formulated as,

$$\mathcal{G}(\mathbf{z}^{t-1}; \theta_3) = \mathbf{W}\mathbf{z}^{t-1} + \mathbf{b}_{\mathcal{G}} \tag{10}$$

where $\theta_3 = \{\mathbf{W}, \mathbf{b}_{\mathcal{G}}\}$ represents the connecting weights and bias of latent-to-latent interactions. Then the joint distribution $p_{Z^t,Z^{t-1}}(\mathbf{z}^t, \mathbf{z}^{t-1})$ can be computed as follows.

$$\begin{aligned}\mathbf{z}^{t-1} &\sim \mathcal{N}(0, \mathbb{I}) \\ \mathcal{F}_2(\mathbf{x}^t, \theta_2) = \bar{\mathbf{z}}^t &\sim \mathcal{N}(0, \mathbb{I}) \\ p_{Z^t,Z^{t-1}}(\mathbf{z}^t, \mathbf{z}^{t-1}; \theta) &\sim \mathcal{N}\left(\begin{bmatrix}\mathbf{b}_{\mathcal{G}} \\ 0\end{bmatrix}, \begin{bmatrix}\mathbf{W}^T\mathbf{W} + \mathbb{I} & \mathbf{W} \\ \mathbf{W} & \mathbb{I}\end{bmatrix}\right)\end{aligned} \tag{11}$$

3.5. Model Learning

The parameters $\theta = \{\theta_1, \theta_2, \theta_3\}$ of the model are optimized to maximize the log-likelihood:

$$\theta_1^*, \theta_2^*, \theta_3^* = \arg \max_{\theta_1,\theta_2,\theta_3} \log p_{X^t}(\mathbf{x}^t | \mathbf{x}^{t-1}; \theta_1, \theta_2, \theta_3) \tag{12}$$

From Eqn. (2), the log-likelihood can be computed as

$$\begin{aligned}\log p_{X^t}(\mathbf{x}^t | \mathbf{x}^{t-1}; \theta) &= \log p_{Z^t}(\mathbf{z}^t | \mathbf{z}^{t-1}, \theta) + \log \left|\frac{\partial \mathcal{F}_2(\mathbf{x}^t; \theta_2)}{\partial \mathbf{x}^t}\right| \\ &= \log p_{Z^t,Z^{t-1}}(\mathbf{z}^t, \mathbf{z}^{t-1}; \theta) \\ &\quad - \log p_{Z^{t-1}}(\mathbf{z}^{t-1}; \theta_1) + \log \left|\frac{\partial \mathcal{F}_2(\mathbf{x}^t; \theta_2)}{\partial \mathbf{x}^t}\right|\end{aligned}$$

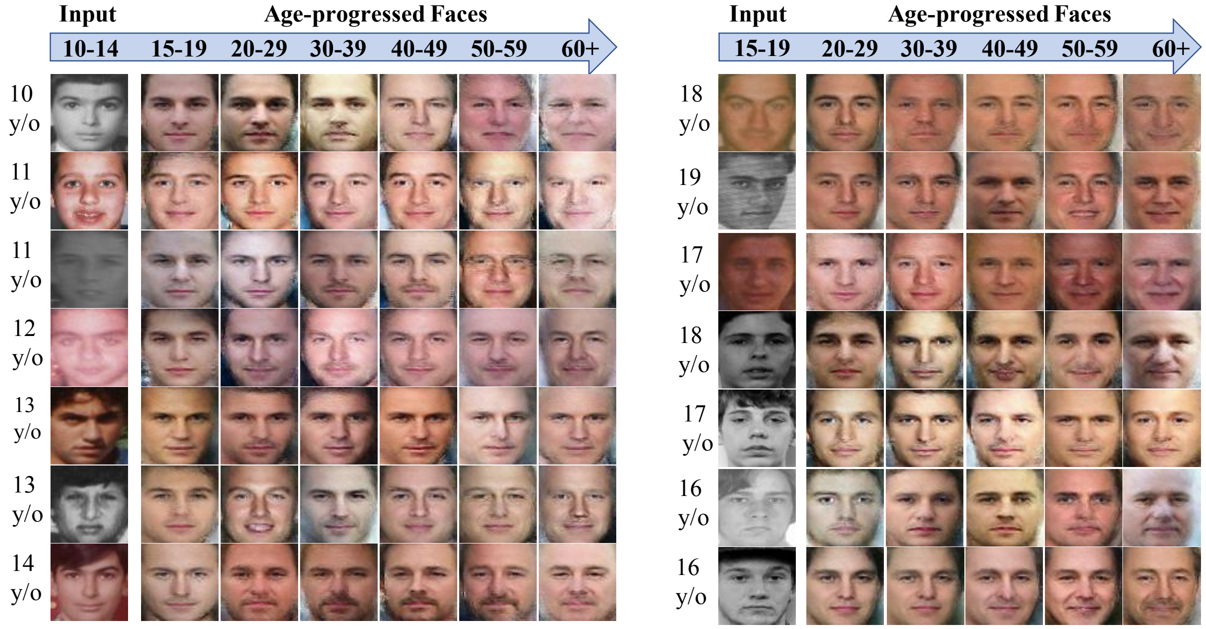


Figure 6: Age Progression Results against FG-NET and MORPH. Given input images, plausible age-progressed faces in different age ranges are automatically synthesized. **Best viewed in color.**

where the first two terms are the two density functions and can be computed using Eqn. (11) while the third term (i.e. the determinant) is obtained using Eqns. (9) and (5). Then the Stochastic Gradient Descent (SGD) algorithm is applied to optimize parameter values.

3.6. Model Properties

Tractability and Invertibility: With the specific structure of the bijection $\mathcal{F}$, our proposed graphical model has the capability of modeling arbitrary complex data distributions while keeping the inference process tractable. Furthermore, from Eqns. (6) and (9), the mapping function is invertible. Therefore, both inference (i.e. mapping from image to latent space) and generation (i.e. from latent to image space) are exact and efficient.

Flexibility: as presented in Section 3.3.1, our proposed model introduces the freedom of choosing the functions $\mathcal{S}$ and $\mathcal{T}$ for their structures. Therefore, different types of deep learning models can be easily exploited to further improve the generative capability of the proposed TNVP. In addition, from Eqn. (3), the binary mask $\mathbf{b}$ also provides the flexibility for our model if we consider this as a template during the mapping process. Several masks can be used in different levels of mapping units to fully exploit the structure of the data distribution of the image domain $\mathcal{I}$.

Although our TNVP shares some similar features with RBM and its family such as TRBM, the log-likelihood estimation of TNVP is tractable while that in RBM is intractable and requires some approximations during training process. Compared to other methods, our TNVP also shows its advantages in high-quality synthesized faces (by avoiding the ℓ_2 reconstruction error as in *Variational Autoencoder*) and efficient training process (i.e. avoid maintaining a good balance between generator and discriminator as in case of GANs).

4. Experimental Results

4.1. Databases

We train our TNVP system using AginG Faces in the Wild (AGFW) [15] and a subset of the Cross-Age Celebrity Dataset (CACD) [4]. Two other public aging databases, i.e. FG-NET [1] and MORPH [19], are used for testing.

AginG Faces in the Wild (AGFW): consists of 18,685 images that covers faces from 10 to 64 years old. On average, after dividing into 11 age groups with the span of 5 years, each group contains 1700 images.

Cross-Age Celebrity Dataset (CACD) is a large-scale dataset with 163446 images of 2000 celebrities. The age range is from 14 to 62 years old.

FG-NET is a common aging database that consists of 1002 images of 82 subjects and has the age range from 0 to 69. Each face is manually annotated with 68 landmarks.

MORPH includes two albums, i.e. MORPH-I and MORPH-II. The former consists of 1690 images of 515 subjects and the latter provides a set of 55134 photos from 13000 subjects. We use MORPH-I for our experiments.

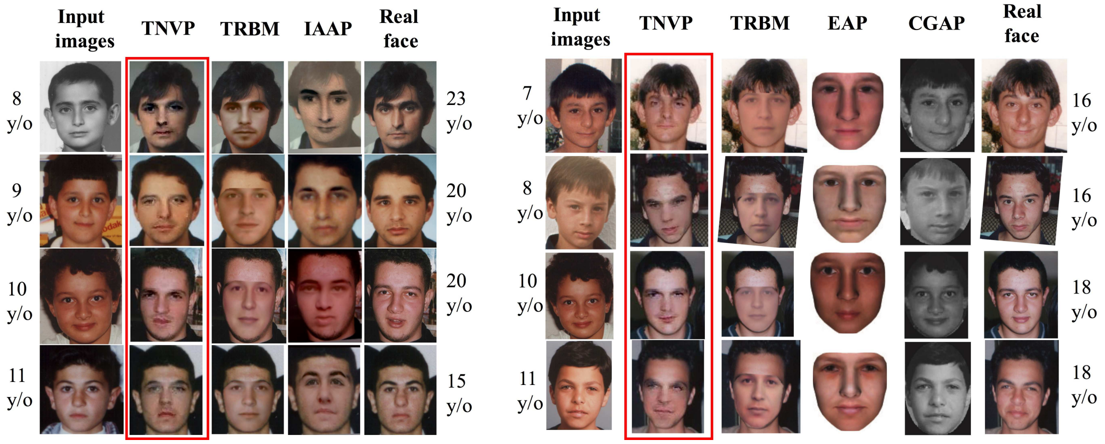


Figure 7: Comparisons between our TNVP against other approaches: IAAP [10], TRBM-based [15], Exemplar based (EAP) [21], and Craniofacial Growth (CGAP) [18] models. **Best viewed in color.**

4.2. Implementation details

To train our TNVP model, we first select a subset of 572 celebrities from CACD as in the training protocol of [15]. All images of these subjects are then classified into 11 age groups ranging from 10 to 65 with the age span of 5 years. Next, the aging sequences for each subject are constructed by collecting and combining all image pairs that cover two successive age groups of that subject. This process results in 6437 training sequences. All training images from these sequences and the AGFW dataset are then preprocessed as presented in Section 3.1. After that, a two-step training process is applied to train our TNVP age progression model. In the first step, using faces from AGFW, all mapping functions (i.e. $\mathcal{F}_1, \mathcal{F}_2$) are pretrained to obtain the capability of face interpretation and high-level feature extraction. Then, our TNVP model is employed to learn the aging transformation between faces presented in the face sequences.

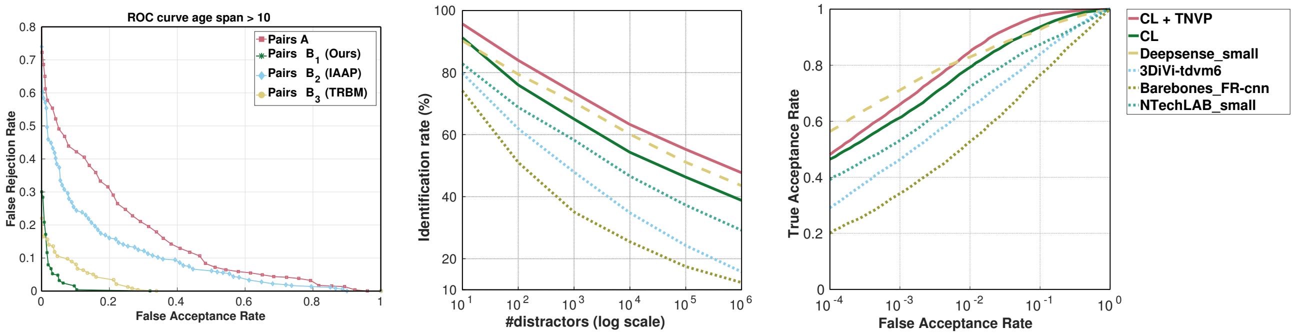
For the model configuration, the number of units for each mapping function is set to 10. In each mapping unit f_i, two Residual Networks with rectifier non-linearity and skip connections are set up for the two transformations $\mathcal{S}$ and $\mathcal{T}$. Each of them contains 2 residual blocks with 32 feature maps. The convolutional filter size is set to 3×3. The training time for TNVP model is 18.75 hours using a machine of Core i7-6700 @3.4GHz CPU, 64.00 GB RAM and a single NVIDIA GTX Titan X GPU and TensorFlow environment. The training batch size is 64.

4.3. Age Progression

After training, our TNVP age progression system is applied to all faces over 10 years old from FG-NET and MORPH. As illustrated in Figure 6, given input faces at different ages, our TNVP is able to synthesize realistic age-progressed faces in different age ranges. Notice that none of the images in FG-NET or MORPH is presented in the training data. From these results, one can easily see that our TNVP not only efficiently embed the specific aging information of each age group to the input faces but also robustly handles in-the-wild variations such as expressions, illumination, and poses. Particularly, beards and wrinkles naturally appear in the age-progressed faces around the ages of 30-49 and over 50, respectively. The face shape is also implicitly handled in our model and changes according to different individuals and age groups. Moreover, by avoiding the ℓ_2 reconstruction loss and taking the advantages of maximizing log-likelihood, sharper synthesized results with aging details are produced using our proposed model. We compare our synthesized results with other recent age progression works whose results are publicly available such as IAAP [10], TRBM-based model [15] in Figure 7. The real faces of the subjects at target ages are provided for reference. Other approaches, i.e. Exemplar based Age Progression (EAP) [21] and Craniofacial Growth (CGAP) model [18], are included for further comparisons. Notice that since our TNVP model is trained using the faces ranging from 10 to 64 years old, we choose the ones with ages close to 10 during the comparison. These results again show the advantages of our TNVP model in terms of efficiently handling the non-linear variations and aging embedding.

4.4. Age-Invariant face verification

This experiment validates the effectiveness of our TNVP model by showing the performance gain for cross-age face verification using our age-progressed faces. In both testing protocols, i.e. small-scale with images pairs from FG-NET and large-scale benchmark on Megaface Challenge 1, we show that our aged faces can provide significant improvements on top of the face matching model without re-training on cross-age databases. We employ one of the state-of-the-



(a) ROC curves of FGNET pairs

(b) CMC curves on MegaFace

(c) ROC curves on MegaFace

Figure 8: From left to right: (a) ROC curves of face verification from 1052 pairs synthesized from different age progression methods; (b) ROC and (c) CMC curves of different face matching methods and the improvement of CL method using our age-progressed faces (under the protocol of MegaFace challenge 1).

art deep face matching model [27], i.e. Center Loss (CL).

Under the *small-scale protocol*, in FG-NET database, we randomly pick 1052 image pairs with the age gap larger than 10 years of either the same or different person. This set is denoted as **A** consisting of a positive list of 526 image pairs of the same person and a negative list of 526 image pairs of two different subjects. From each image pair of set **A**, using the face with younger age, we synthesize an age-progressed face image at the age of the older one using our proposed TNVP model. This forms a new matching pair, i.e. the aged face vs. the original face at older age. Applying this process for all pairs of set **A**, we obtain a new set denoted as set $\mathbf{B_1}$. To compare with IAAP [25] and TRBM [15] methods, we also construct two other sets of image pairs similarly and denote them as set $\mathbf{B_2}$ and $\mathbf{B_3}$, respectively. Then, the False Rejection Rate-False Acceptance Rate (FRR-FAR) is computed and plotted under the Receiver Operating Characteristic (ROC) curves for all methods (Fig. 8a). Our method achieves an improvement of **30**% on matching performance over the original pair (set **A**) while IAAP and TRBM slightly increase the rates.

In addition, our model is also experimented on the *large-scale Megaface* [9] challenge 1 with FGNET test set. Practical face recognition models should achieve high performance against having gallery set of millions of distractors and probe set of people at various ages. In this testing, 4 billion pairs are generated between the probe and gallery sets where the gallery includes one million distractors. Thus, only improvements on Rank-1 identification rate with one million distractors and verification rate at low FAR are meaningful [9]. Fig. 8b shows Rank-1 identification rates as the number of distractors increasing and the rates with one million distractors are shown in Table 2. We compute the TAR-FAR and show ROC curves[1] in Fig. 8c. The model from DeepSense achieves the best performance under the cross-age training set while the CL model [29] trained solely on CASIA WebFace dataset having < 0.49M images without cross-age information. From these results, we show that face matching models can directly benefit from our TNVP model to improve their robustness against aging effects. Particularly, by using our age-progressed images without re-training, the CL model [29] not only obtains **10**% improvements but also outperforms other models trained with a small training set as shown in Table 2.

Table 2: Rank-1 Identification Accuracy with one million Distractors (MegaFace Challenge 1 - FGNET). Protocol “small” means ≤0.5M images trained. “Cross-age” means trained with cross-age faces.

Methods	Protocol	Cross-age	Accuracy
Barebones_FR	Small	Y	7.136 %
3DiVi	Small	Y	15.78 %
NTechLAB	Small	Y	29.168 %
DeepSense	Small	Y	43.54 %
CL [27]	Small	**N**	38.79%
CL + TNVP	Small	**N**	**47.72**%

5. Conclusion

This paper has presented a novel generative probabilistic model with a tractable density function for age progression. The model inherits the strengths of both probabilistic graphical model and recent advances of ResNet. The nonlinear age-related variance and the aging transformation between age groups are efficiently captured. Given the log-likelihood objective function, high-quality age-progressed faces can be produced. In addition to a simple preprocessing step, geometric constraints are implicitly embedded during the learning process. The evaluations in both quality of synthesized faces and cross-age verification showed the robustness of our TNVP.

[1] The results of other methods are provided in MegaFace website.

References

[1] *FG-NET Aging Database*. http://www.fgnet.rsunit.com.

[2] G. Antipov, M. Baccouche, and J.-L. Dugelay. Face aging with conditional generative adversarial networks. *arXiv preprint arXiv:1702.01983*, 2017.

[3] D. M. Burt and D. I. Perrett. Perception of age in adult caucasian male faces: Computer graphic manipulation of shape and colour information. *Proceedings of the Royal Society of London B: Biological Sciences*, 259(1355):137–143, 1995.

[4] B.-C. Chen, C.-S. Chen, and W. H. Hsu. Cross-age reference coding for age-invariant face recognition and retrieval. In *ECCV*, 2014.

[5] L. Dinh, J. Sohl-Dickstein, and S. Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

[6] X. Geng, Z.-H. Zhou, and K. Smith-Miles. Automatic age estimation based on facial aging patterns. *PAMI*, 29(12):2234–2240, 2007.

[7] M. Germain, K. Gregor, I. Murray, and H. Larochelle. Made: Masked autoencoder for distribution estimation. In *ICML*, pages 881–889, 2015.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, June 2016.

[9] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *CVPR*, 2016.

[10] I. Kemelmacher-Shlizerman, S. Suwajanakorn, and S. M. Seitz. Illumination-aware age progression. In *CVPR*, pages 3334–3341. IEEE, 2014.

[11] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[12] A. Lanitis, C. J. Taylor, and T. F. Cootes. Toward automatic simulation of aging effects on face images. *PAMI*, 24(4):442–455, 2002.

[13] H. Larochelle and I. Murray. The neural autoregressive distribution estimator. In *AISTATS*, volume 1, page 2, 2011.

[14] K. Luu, C. Suen, T. Bui, and J. K. Ricanek. Automatic child-face age-progression based on heritability factors of familial faces. In *BIdS*, pages 1–6. IEEE, 2009.

[15] C. Nhan Duong, K. Luu, K. Gia Quach, and T. D. Bui. Longitudinal face modeling via temporal deep restricted boltzmann machines. In *CVPR*, June 2016.

[16] E. Patterson, K. Ricanek, M. Albert, and E. Boone. Automatic representation of adult aging in facial images. In *Proc. IASTED Intl Conf. Visualization, Imaging, and Image Processing*, pages 171–176, 2006.

[17] E. Patterson, A. Sethuram, M. Albert, and K. Ricanek. Comparison of synthetic face aging to age progression by forensic sketch artist. In *IASTED International Conference on Visualization, Imaging, and Image Processing, Palma de Mallorca, Spain*, 2007.

[18] N. Ramanathan and R. Chellappa. Modeling age progression in young faces. In *CVPR*, volume 1, pages 387–394. IEEE, 2006.

[19] K. Ricanek Jr and T. Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *FGR 2006.*, pages 341–345. IEEE, 2006.

[20] D. Rowland, D. Perrett, et al. Manipulating facial appearance through shape and color. *Computer Graphics and Applications, IEEE*, 15(5):70–76, 1995.

[21] C.-T. Shen, W.-H. Lu, S.-W. Shih, and H.-Y. M. Liao. Exemplar-based age progression prediction in children faces. In *ISM*, pages 123–128. IEEE, 2011.

[22] X. Shu, J. Tang, H. Lai, L. Liu, and S. Yan. Personalized age progression with aging dictionary. In *ICCV*, December 2015.

[23] J. Suo, X. Chen, S. Shan, W. Gao, and Q. Dai. A concatenational graph evolution aging model. *PAMI*, 34(11):2083–2096, 2012.

[24] J. Suo, S.-C. Zhu, S. Shan, and X. Chen. A compositional and dynamic model for face aging. *PAMI*, 32(3):385–401, 2010.

[25] M.-H. Tsai, Y.-K. Liao, and I.-C. Lin. Human face aging with guided prediction and detail synthesis. *Multimedia tools and applications*, 72(1):801–824, 2014.

[26] W. Wang, Z. Cui, Y. Yan, J. Feng, S. Yan, X. Shu, and N. Sebe. Recurrent face aging. In *CVPR*, pages 2378–2386, 2016.

[27] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, pages 499–515. Springer, 2016.

[28] H. Yang, D. Huang, Y. Wang, H. Wang, and Y. Tang. Face aging effect simulation using hidden factor analysis joint sparse representation. *TIP*, 25(6):2493–2507, 2016.

[29] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.