Automatic Face Aging in Videos via Deep Reinforcement Learning

Chi Nhan Duong [1], Khoa Luu [2], Kha Gia Quach [1], Nghia Nguyen [2],
Eric Patterson [3], Tien D. Bui [1], Ngan Le [4]
[1] Computer Science and Software Engineering, Concordia University, Canada
[2] Computer Science and Computer Engineering, University of Arkansas, USA
[3] School of Computing, Clemson University, USA
[4] Electrical and Computer Engineering, Carnegie Mellon University, USA
[1]{dcnhan, kquach}@ieee.org, bui@encs.concordia.ca, [2]{khoaluu, nhnguyen}@uark.edu,
[3]ekp@clemson.edu, [4]thihoanl@andrew.cmu.edu

Abstract

This paper presents a novel approach to synthesize automatically age-progressed facial images in video sequences using Deep Reinforcement Learning. The proposed method models facial structures and the longitudinal face-aging process of given subjects coherently across video frames. The approach is optimized using a long-term reward, Reinforcement Learning function with deep feature extraction from Deep Convolutional Neural Network. Unlike previous age-progression methods that are only able to synthesize an aged likeness of a face from a single input image, the proposed approach is capable of age-progressing facial likenesses in videos with consistently synthesized facial features across frames. In addition, the deep reinforcement learning method guarantees preservation of the visual identity of input faces after age-progression. Results on videos of our new collected aging face AGFW-v2 database demonstrate the advantages of the proposed solution in terms of both quality of age-progressed faces, temporal smoothness, and cross-age face verification.

1. Introduction

Age-related facial technologies generally address the two areas of age estimation [4, 21, 20, 23, 8, 22] and age progression [6, 29, 46, 30, 41, 35]. The face age-estimation problem is defined as building computer software that has the ability to recognize the ages of individuals in a given photograph. Comparatively, the face age-progression problem necessitates the more complex capability to predict the future facial likeness of people appearing in images [19]. Aside from the innate curiosity of individuals, research of face aging has its origins in cases of missing persons and wanted fugitives, in either case law enforcement desires plausible age-progressed images to facilitate searches. Accurate face aging also provides benefits for numerous practical applications such as age-invariant face recognition [44, 43, 17]. There have been numerous anthropological, forensic, computer-aided, and computer-automated approaches to facial age-progression. However, the results from previous methods for synthesizing aged faces that represent accurate physical processes involved in human aging are still far from perfect. This is especially so in age-progressing videos of faces, due to the usual challenges for

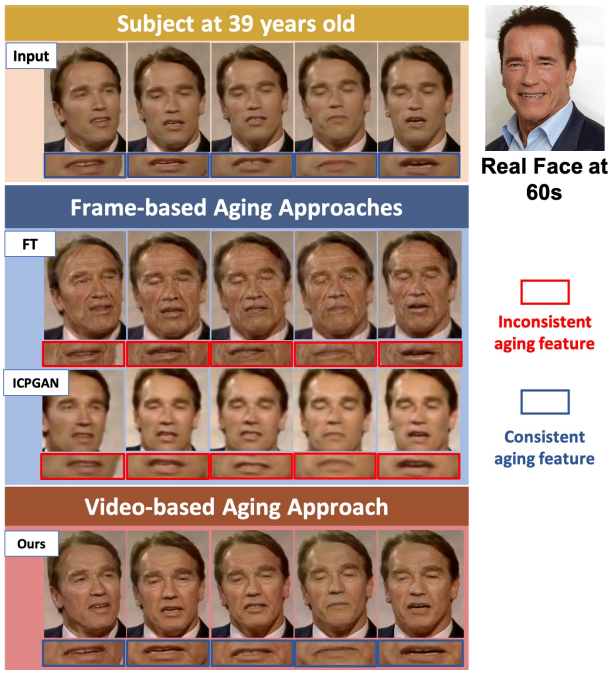

Figure 1: Given an input video, while frame-based approaches produce inconsistent aging features, our video-based method ensures consistency among video frames.

Table 1: The comparison of the properties between our video-based approach and other age progression methods.

	Ours	ICPGAN [41]	TNVP [9]	CAAE [46]	RFA [40]	TRBM [7]
Modality	**Video-based**	Image-based	Image-based	Image-based	Image-based	Image-based
Temporal Consistency	**Yes**	No	No	No	No	No
Aging Mechanism	**One-shot**	One-shot	Multiple-shot	One-shot	One-shot	Multiple-shot
Architecture	**DL + RL**	DL	DL	DL	DL	DL
Tractability	✓	✓	✓	✓	✓	✗

face processing involving pose, illumination, and environment variation as well as differences between video frames.

There have been two key research directions in age progression for both conventional computer-vision approaches and recent deep-learning methods – *one-shot synthesis* and *multiple-shot synthesis*. Both approaches have used facial image databases with longitudinal sample photos of individuals, where the techniques attempt to discover aging patterns demonstrated over individuals or the population represented. In one-shot synthesis approaches, a new face at the target age is directly synthesized via inferring the relationships between training images and their corresponding age labels then applying them to generate the aged likeness. These prototyping methods [2, 15, 33] often classify training images in facial image databases into age groups according to labels. Then the average faces, or mean faces, are computed to represent the key presentation or archetype of their groups. The variation between the input age and the target age archetypes is complimented to the input image to synthesize the age-progressed faces at the requested age. In a similar way, Generative Adversarial Networks (GANs) [46, 41] methods present the relationship between semantic representation of input faces and age labels by constructing a deep neural network generator. It is then combined with the target age labels to synthesize output results.

Meanwhile, in multiple-shot synthesis, the longitudinal aging process is decomposed into multiple steps of aging effects [9, 7, 35, 40, 45]. These methods build on the facial aging transformation between two consecutive age groups. Finally, the progressed faces from one age group to the next are synthesized step-by-step until they reach the target age. These methods can model the long-term sequence of face aging using this strategy. However, these methods still have drawbacks due to the limitations of long-term aging not being well represented nor balanced in face databases.

Existing age-progression methods all similarly suffer from problems in both directions. Firstly, they only work on single input images. Supposing there is a need to synthesize aging faces presented in a captured video, these methods usually have to split the input video into separate frames and synthesize every face in each frame *independently* which may often present *inconsistencies* between synthesized faces. Since face images for each frame are synthesized separately, the aging patterns of generated faces of the same subject are also likely not coherent. Furthermore, most aging methods are unable to produce *high-resolution* images of age progression, important for features such as fine lines that develop fairly early in the aging process. This may be especially true in the latent based methods [15, 9, 7, 35, 40, 45].

Contributions of this work: This paper presents a deep Reinforcement Learning (RL) approach to Video Age Progression to guarantee the consistency of aging patterns in synthesized faces captured in videos. In this approach, the age-transformation embedding is modeled as the optimal selection using Convolutional Neural Network (CNN) features under a RL framework. Rather than applying the image-based age progression to each video frame independently as in previous methods, the proposed approach has the capability of exploiting the temporal relationship between two consecutive frames of the video. This property facilitates maintaining consistency of aging information embedded into each frame. In the proposed structure, not only can a *smoother synthesis* be produced across frames in videos, but also the *visual fidelity* of aging data, i.e. all images of a subject in different or the same age, is preserved for better age transformations. To the best of our knowledge, our framework is one of the first face aging approaches in videos. Finally, this work contributes a new large-scale face-aging database[1] to support future studies related to automated face age-progression and age estimation in both images and videos.

2. Related work

This section provides an overview of recent approaches for age progression; *these methods primarily use still images*. The approaches generally fall into one of four groups, i.e. modeling, reconstruction, prototyping, and deep learning-based approaches.

Modeling-based approaches aim at modeling both shape and texture of facial images using parameterization method, then learning to change these parameters via an aging function. Active Appearance Models (AAMs) have been used with four aging functions in [16, 28] to model linearly both the general and the specific aging processes. Familial facial cues were combined with AAM-based techniques in [24, 29]. [30] incorporated an AAM reconstruction method to the synthesis process for a higher photographic fidelity of aging. An AGing pattErn Subspace (AGES) [14] was proposed to construct a subspace for aging patterns as a chrono-

[1] https://face-aging.github.io/RL-VAP/

Table 2: The properties of our collected AGFW-v2 in comparison with other aging databases. For AGFW-v2 video set, the images of the subjects in old age are also collected for reference in terms of subject's appearance changing.

Database	# Images	# Subjects	Label type	Image type	Subject type	Type
MORPH - Album 1 [31]	1,690	628	Years old	Mugshot	Non-famous	Image DB
MORPH - Album 2 [31]	55,134	13,000	Years old	Mugshot	Non-famous	Image DB
FG-NET [10]	1,002	82	Years old	In-the-wild	Non-famous	Image DB
AdienceFaces [18]	26,580	2,984	Age groups	In-the-wild	Non-famous	Image DB
CACD [3]	163,446	2,000	Years old	In-the-wild	Celebrities	Image DB
IMDB-WIKI [32]	52,3051	20,284	Years old	In-the-wild	Celebrities	Image DB
AgeDB [27]	16,488	568	Years old	In-the-wild	Celebrities	Image DB
AGFW [7]	18,685	14,185	Age groups	In-the-wild/Mugshot	Non-famous	Image DB
AGFW-v2 (Image)	**36,299**	**27,688**	**Age groups**	**In-the-wild/Mugshot**	**Non-famous**	**Image DB**
AGFW-v2 (Video)	**20,000**	**100**	**Years old**	**Interview/Movie-style**	**Celebrities**	**Video DB**

logical sequence of face images. In [38], AGES was enhanced with guidance faces consisting the subject's characteristics for more stable results. Three-layer And-Or Graph (AOG) [37, 36] was used to model a face as a combination of smaller parts, i.e. eyes, nose, mouth, etc. Then a Markov chain was employed to learn the aging process for each part.

In *reconstruction-based* approaches, an aging basis is unified in each group to model aging faces. Person-specific and age-specific factors were independently represented by sparse-representation hidden factor analysis (HFA) [45]. Aging dictionaries (CDL) [35] were proposed to model personalized aging patterns by attempting to preserve distinct facial features of an individual through the aging process.

Prototyping-based approaches employed proto-typical facial images in a method to synthesize faces. The average face of each age group is used as the representative image for that group, and these are named the "age prototypes" [33]. Then, by computing the differences between the prototypes of two age groups, an input face can be progressed to the target age through image-based manipulation [2]. In [15], high quality average prototypes constructed from a large-scale dataset were employed in conjunction with the subspace alignment and illumination normalization.

Recently, *Deep learning-based approaches* have yielded promising results in facial age progression. Temporal and Spatial Restricted Boltzmann Machines (TRBM) were introduced in [7] to represent the non-linear aging process, with geometry constraints, and to model a sequence of reference faces as well as wrinkles of adult faces. A Recurrent Neural Network (RNN) with two-layer Gated Recurrent Unit (GRU) was employed to approximate aging sequences [40]. Also, the structure of Conditional Adversarial Autoencoder (CAAE) was applied to synthesize aged images in [1]. Identity-Preserved Conditional Generative Adversarial Networks (IPCGANs) [41] brought the structure of Conditional GANs with perceptual loss into place for synthesis process. A novel generative probabilistic model, called Temporal Non-Volume Preserving (TNVP) transformation [9] was proposed to model a long-term facial aging as a sequence of short-term stages.

3. Data Collection

The quality of age representation in a face database is one of the most important features affecting the aging learning process and could include such considerations as the number of longitudinal face-image samples per subject, the number of subjects, the range and distribution of age samples overall, and the population representation presented in the database. Previous public databases used for age estimation or progression systems have been very limited in the total number of images, the number of images per subject, or the longitudinal separation of the samples of subjects in the database, i.e. FG-NET [10], MORPH [31], AgeDB [27]. Some recent ones may be of larger scale but have noise within the age labels, i.e. CACD [3], IMDB-WIKI [32]. In this work we introduce an extension of Aging Faces in the Wild (AGFW-v2) in terms of both *image and video* collections. Table 2 presents the properties of our collected AGFW-v2 in comparison with others.

3.1. Image dataset

AGFW [7] was first introduced with 18,685 images with individual ages sampled ranging from 10 to 64 years old. Based on the collection criteria of AGFW, a double-sized database was desired. Compared to other age-related databases, *most of the subjects in AGFW-v2 are not public figures and less likely to have significant make-up or facial modifications*, helping embed accurate aging effects during the learning process. In particular, AGFW-v2 is mainly collected from three sources. Firstly, we adopt a search engine using different keywords, e.g. male at 20 years old, etc. Most images come from the daily life of non-famous subjects. Besides the images, all publicly available meta-data related to the subject's age are also collected. The second part comes from mugshot images that are accessible from public domain. These are passport-style photos with ages reported by service agencies. Finally, we also include the Productive Aging Laboratory (PAL) database [26]. In total, AGFW-v2 consists of 36,299 images divided into 11 age groups with a span of five years.

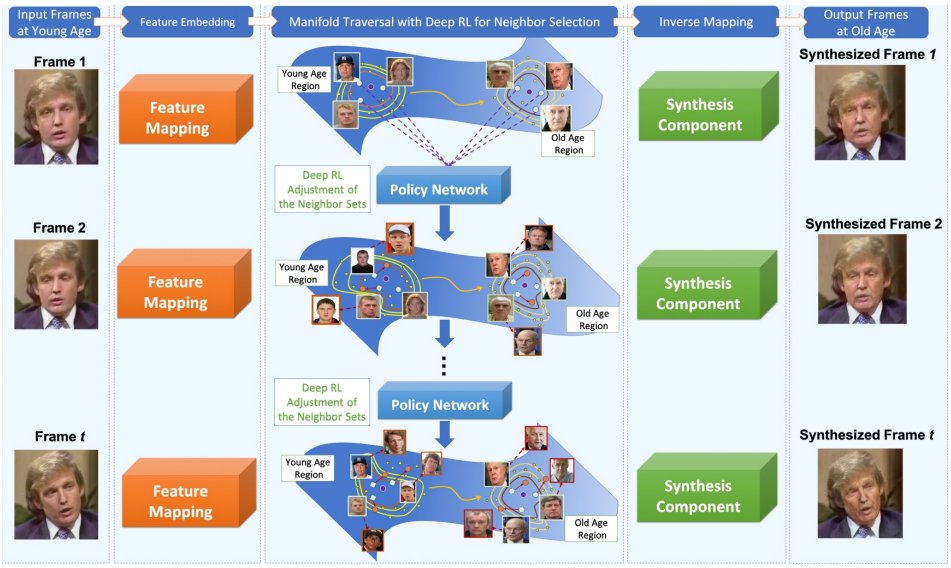

Figure 2: The structure of the face aging framework in video. **Best viewed in color and 2× zoom in.**

3.2. Video dataset

Along with still photographs, we also collected a video dataset for temporal aging evaluations with 100 videos of celebrities. Each video clip consists of 200 frames. In particular, searching based on the individuals' names during collection efforts, their interview, presentation, or movie sessions were selected such that only one face, in a clear manner, is presented in the frame. Age annotations were estimated using the year of the interview session versus the year of birth of the individual. Furthermore, in order to provide a reference for subject's appearance in old age, the face images of these individuals at the current age are also collected and provided as meta-data for the subjects' videos.

4. Video-based Facial Aging

In the simplest approach, age progression of a sequence may be achieved by independently employing image-based aging techniques on each frame of a video. However, treating single frames independently may result in inconsistency of the final aged-progressed likeness in the video, i.e. some synthesized features such as wrinkles appear differently across consecutive video frames as illustrated in Fig. 1. Therefore, rather than considering a video as a set of independent frames, this method exploits the temporal relationship between frames of the input video to maintain visually cohesive age information for each frame. The aging algorithm is formulated as the sequential decision-making process from a goal-oriented agent while interacting with the temporal visual environment. At time sample, the agent integrates related information of the current and previous frames then modifies action accordingly. The agent receives a scalar reward at each time-step with the goal of maximizing the total long-term aggregate of rewards, emphasizing effective utilization of temporal observations in computing the aging transformation employed on the current frame.

Formally, given an input video, let $\mathcal{I} \in \mathbb{R}^d$ be the image domain and $\mathbf{X}^t = \{\mathbf{x}_y^t, \mathbf{x}_o^t\}$ be an image pair at time-step t consisting of the t-th frame $\mathbf{x}_y^t \in \mathcal{I}$ of the video at young age and the synthesized face $\mathbf{x}_o^t \in \mathcal{I}$ at old age. The goal is to learn a synthesis function $\mathcal{G}$ that maps $\mathbf{x}_y^t$ to $\mathbf{x}_o^t$ as.

$$\mathbf{x}_o^t = \mathcal{G}(\mathbf{x}_y^t)|\mathbf{X}^{1:t-1} \tag{1}$$

The conditional term indicates the temporal constraint needs to be considered during the synthesis process. To learn $\mathcal{G}$ effectively, we decompose $\mathcal{G}$ into sub-functions as.

$$\mathcal{G} = \mathcal{F}_1 \circ \mathcal{M} \circ \mathcal{F}_2 \tag{2}$$

where $\mathcal{F}_1 : \mathbf{x}_y^t \mapsto \mathcal{F}_1(\mathbf{x}_y^t)$ maps the young face image $\mathbf{x}_y^t$ to its representation in feature domain; $\mathcal{M} : (\mathcal{F}_1(\mathbf{x}_y^t); \mathbf{X}^{1:t-1}) \mapsto \mathcal{F}_1(\mathbf{x}_o^t)$ defines the traversing function in feature domain; and $\mathcal{F}_2 : \mathcal{F}_1(\mathbf{x}_o^t) \mapsto \mathbf{x}_o^t$ is the mapping from feature domain back to image domain.

Based on this decomposition, the architecture of our proposed framework (see Fig. 2) consists of three main processing steps: (1) Feature embedding; (2) Manifold traversal; and (3) Synthesizing final images from updated features. In the second step, a Deep RL based framework is proposed to guarantee the consistency between video frames in terms of aging changes during synthesis process.

4.1. Feature Embedding

The first step of our framework is to learn an embedding function $\mathcal{F}_1$ to map $\mathbf{x}_y^t$ into its latent representation $\mathcal{F}_1(\mathbf{x}_y^t)$. Although there could be various choices for $\mathcal{F}_1$, to produce high quality synthesized images in later steps, the

chosen structure for $\mathcal{F}_1$ should produce a feature representation with two main properties: (1) *linearly separable* and (2) *detail preserving*. On one hand, with the former property, transforming the facial likeness from one age group to another age group can be represented as the problem of linearly traversing along the direction of a single vector in feature domain. On the other hand, the latter property guarantees a certain detail to be preserved and produce high quality results. In our framework, CNN structure is used for $\mathcal{F}_1$. It is worth noting that there remain some compromises regarding the choice of deep layers used for the representation such that both properties are satisfied. *Linear separability* is preferred in deeper layers further along the linearization process while *details of a face* are usually embedded in more shallow layers [25]. As an effective choice in several image-modification tasks [12, 13], we adopt the normalized VGG-19[2] and use the concatenation of three layers $\{conv3\_1, conv4\_1, conv5\_1\}$ as the feature embedding.

4.2. Manifold Traversing

Given the embedding $\mathcal{F}_1(\mathbf{x}_y^t)$, the age progression process can be interpreted as the linear traversal from the younger age region of $\mathcal{F}_1(\mathbf{x}_y^t)$ toward the older age region of $\mathcal{F}_1(\mathbf{x}_o^t)$ within the deep-feature domain. Then the Manifold Traversing function $\mathcal{M}$ can be written as in Eqn (3).

$$\begin{aligned}\mathcal{F}_1(\mathbf{x}_o^t) &= \mathcal{M}(\mathcal{F}_1(\mathbf{x}_y^t); \mathbf{X}^{1:t-1}) \\ &= \mathcal{F}_1(\mathbf{x}_y^t) + \alpha\Delta^{\mathbf{x}^t|\mathbf{X}^{1:t-1}}\end{aligned} \tag{3}$$

where α denotes the user-defined combination factor, and $\Delta^{\mathbf{x}^t|\mathbf{X}^{1:t-1}}$ encodes the amount of aging information needed to reach the older age region for the frame $\mathbf{x}_y^t$ conditional on the information of previous frames.

4.2.1 Learning from Neighbors

In order to compute $\Delta^{\mathbf{x}^t|\mathbf{X}^{1:t-1}}$ containing only aging effects without the presence of other factors, i.e. identity, pose, etc., we exploit the relationship in terms of the aging changes between the nearest neighbors of $\mathbf{x}_y^t$ in the two age groups. In particular, given $\mathbf{x}_y^t$, we construct two neighbor sets $\mathcal{N}_y^t$ and $\mathcal{N}_o^t$ that contain K nearest neighbors of $\mathbf{x}_y^t$ in the young and old age groups, respectively. Then $\Delta^{\mathbf{x}^t|\mathbf{X}^{1:t-1}} = \Delta^{\mathbf{x}^t|\mathbf{X}^{1:t-1}}_{\mathcal{A}(\cdot,\mathbf{x}_y^t)}$ is estimated by:

$$\Delta^{\mathbf{x}^t|\mathbf{X}^{1:t-1}} = \frac{1}{K}\left[\sum_{\mathbf{x}\in\mathcal{N}_o^t}\mathcal{F}_1(\mathcal{A}(\mathbf{x},\mathbf{x}_y^t)) - \sum_{\mathbf{x}\in\mathcal{N}_y^t}\mathcal{F}_1(\mathcal{A}(\mathbf{x},\mathbf{x}_y^t))\right]$$

where $\mathcal{A}(\mathbf{x},\mathbf{x}_y^t)$ denotes a face-alignment operator that positions the face in $\mathbf{x}$ with respect to the face location in $\mathbf{x}_y^t$. Since only the nearest neighbors of $\mathbf{x}_y^t$ are considered in the two sets, conditions apart from age difference should be sufficiently similar between the two sets and subtracted away in $\Delta^{\mathbf{x}^t|\mathbf{X}^{1:t-1}}$. Moreover, the averaging operator also helps to ignore identity-related factors, and, therefore, emphasizing age-related changes as the main source of difference to be encoded in $\Delta^{\mathbf{x}^t|\mathbf{X}^{1:t-1}}$. The remaining question is how to choose the appropriate neighbor sets such that the aging changes provided by $\Delta^{\mathbf{x}^t|\mathbf{X}^{1:t-1}}$ and $\Delta^{\mathbf{x}^{t-1}|\mathbf{X}^{1:t-2}}$ are consistent. In the next section, a Deep RL based framework is proposed for selecting appropriate candidates for these sets.

4.2.2 Deep RL for Neighbor Selection

A straightforward technique of choosing the neighbor sets for $\mathbf{x}_y^t$ in young and old age is to select faces that are close to $\mathbf{x}_y^t$ based on some *closeness criteria* such as distance in feature domain, or number of matched attributes. However, since these criteria are not frame-interdependent, they are unable to maintain visually cohesive age information across video frames. Therefore, we propose to exploit the relationship presented in the image pair $\{\mathbf{x}_y^t, \mathbf{x}_y^{t-1}\}$ and the neighbor sets of $\mathbf{x}_y^{t-1}$ as an additional guidance for the selection process. Then an RL based framework is proposed and formulated as a sequential decision-making process with the goal of maximizing the temporal reward estimated by the consistency between the neighbor sets of $\mathbf{x}_y^t$ and $\mathbf{x}_y^{t-1}$.

Specifically, given two input frames $\{\mathbf{x}_y^t, \mathbf{x}_y^{t-1}\}$ and two neighbor sets $\{\mathcal{N}_y^{t-1}, \mathcal{N}_o^{t-1}\}$ of $\mathbf{x}_y^{t-1}$, the agent of a policy network will iteratively analyze the role of each neighbor of $\mathbf{x}_y^{t-1}$ in both young and old age in combination with the relationship between $\mathcal{F}_1(\mathbf{x}_y^t)$ and $\mathcal{F}_1(\mathbf{x}_y^{t-1})$ to determine new suitable neighbors for $\{\mathcal{N}_y^t, \mathcal{N}_o^t\}$ of $\mathbf{x}_y^t$. A new neighbor is considered appropriate when it is sufficiently similar to $\mathbf{x}_y^t$ and maintains aging consistency between two frames. Each time a new neighbor is selected, the neighbor sets of $\mathbf{x}_y^t$ are updated and received a reward based on estimating the similarity of embedded aging information between two frames. As a result, the agent can iteratively explore an optimal route for selecting neighbors to maximize the long-term reward. Fig. 3 illustrates the process of selecting neighbors for age-transformation relationship.

State: The state at i-th step $\mathbf{s}_i^t = \left[\mathbf{x}_y^t, \mathbf{x}_y^{t-1}, \mathbf{z}_i^{t-1}, (\mathcal{N}^t)_i, \bar{\mathcal{N}}^t, \mathbf{M}_i\right]$ is defined as a composition of six components: (1) the *current frame* $\mathbf{x}_y^t$; (2) the *previous frame* $\mathbf{x}_y^{t-1}$; (3) the *current considered neighbor* $\mathbf{z}_i^{t-1}$ of $\mathbf{x}_y^{t-1}$, i.e. either in young and old age groups; (4) the *current construction of the two neighbor sets* $(\mathcal{N}^t)_i = \{(\mathcal{N}_y^t)_i, (\mathcal{N}_o^t)_i\}$ of $\mathbf{x}_y^t$ until step i; (5) the *extended neighbor sets* $\bar{\mathcal{N}}^t = \{\bar{\mathcal{N}}_y^t, \bar{\mathcal{N}}_o^t\}$ consisting of N neighbors, i.e. $N > K$, of $\mathbf{x}_y^t$ for each age group. and (6) a *binary mask* $\mathbf{M}_i$ indicating which samples in $\bar{\mathcal{N}}^t$ are already chosen in previous steps. Notice that in the initial state $\mathbf{s}_0^t$, the two neighbor sets $\{(\mathcal{N}_y^t)_0, (\mathcal{N}_o^t)_0\}$ are

[2]This network is trained on ImageNet for better latent space.

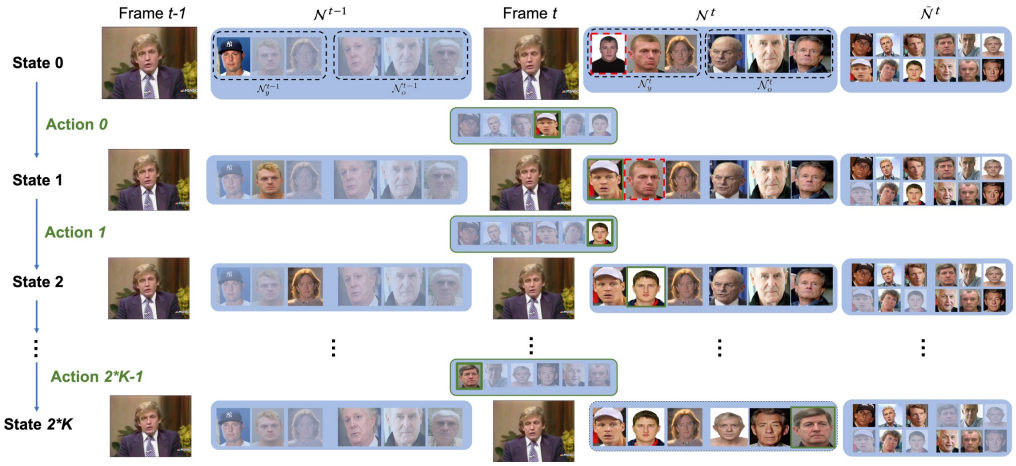

Figure 3: The process of selecting neighbors for age-transformation relationship. **Best viewed in color and 2× zoom in.**

initialized using the K nearest neighbors of $\mathbf{x}_y^t$ of the two age groups, respectively. Two measurement criteria are considered for finding the nearest neighbors: *the number of matched facial attributes*, e.g gender, expressions, etc.; and *the cosine distance between two feature embedding vectors*. All values of the mask $\mathbf{M}_i$ are set to 1 in $\mathbf{s}_0^t$.

Action: Using the information from the chosen neighbor $\mathbf{z}_i^{t-1}$ of $\mathbf{x}_y^{t-1}$, and the relationship of $\{\mathbf{x}_y^t, \mathbf{x}_y^{t-1}\}$, an action a_i^t is defined as selecting the new neighbor for the current frame such that with this new sample added to the neighbor sets of the current frame, the aging-synthesis features between $\mathbf{x}_y^t$ and $\mathbf{x}_y^{t-1}$ are more consistent. Notice that since not all samples in the database are sufficiently similar to $\mathbf{x}_y^t$, we restrict the action space by selecting among N nearest neighbors of $\mathbf{x}_y^t$. In our configuration, $N = n * K$ where n and K are set to 4 and 100, respectively.

Policy Network: At each time step i, the policy network first encodes the information provided in state $\mathbf{s}_i^t$ as

$$\begin{aligned} \mathbf{u}_i^t &= \left[\delta_{\mathcal{F}_1}^{\text{pool5}}(\mathbf{x}_y^t, \mathbf{x}_y^{t-1}), \mathcal{F}_1^{\text{pool5}}(\mathbf{z}_i^{t-1})\right] \\ \mathbf{v}_i^t &= \left[d\left((\mathcal{N}^t)_i, \mathbf{x}_y^t\right), d\left(\bar{\mathcal{N}}^t, \mathbf{x}_y^t\right), \mathbf{M}_i\right] \end{aligned} \tag{4}$$

where $\mathcal{F}_1^{\text{pool5}}$ is the embedding function as presented in Sec. 4.1, but the *pool5* layer is used as the representation; $\delta_{\mathcal{F}_1}^{\text{pool5}}(\mathbf{x}_y^t, \mathbf{x}_y^{t-1}) = \mathcal{F}_1^{\text{pool5}}(\mathbf{x}_y^t) - \mathcal{F}_1^{\text{pool5}}(\mathbf{x}_y^{t-1})$ embeds the relationship of $\mathbf{x}_y^t$ and $\mathbf{x}_y^{t-1}$ in the feature domain. $d\left((\mathcal{N}^t)_i, \mathbf{x}_y^t\right)$ is the operator that maps all samples in $(\mathcal{N}^t)_i$ to their representation in the form of cosine distance to $\mathbf{x}_y^t$. The last layer of the policy network is reformulated as $P(\mathbf{z}_i^t = \mathbf{x}_j | \mathbf{s}_i^t) = e^{c_i^j} / \sum_k c_i^k$, where $\mathbf{c}_i = \mathbf{M}_i \odot \left(\mathbf{W}\mathbf{h}_i^t + \mathbf{b}\right)$ and $\mathbf{h}_i^t = \mathcal{F}_\pi\left(\mathbf{u}_i^t, \mathbf{v}_i^t.\theta_\pi\right)$; $\{\mathbf{W}, \mathbf{b}\}$ are weight and bias of the hidden-to-output connections. Since $\mathbf{h}_i^t$ consists of the features of the sample picked for neighbors of $\mathbf{x}_y^{t-1}$ and the temporal relationship between $\mathbf{x}_y^{t-1}$ and $\mathbf{x}_y^t$, it directly encodes the information of *how the face changes* and *what aging information* from the previous frame has been used. This process helps the agent evaluate its choice to confirm the optimal candidate of $\mathbf{x}_y^t$ to construct the neighbor sets.

The output of the policy network is an $N+1$-dimension vector $\mathbf{p}$ indicating the probabilities of all available actions $P(\mathbf{z}_i^t = \mathbf{x}_j | \mathbf{s}_i^t), j = 1..N$ where each entry indicates the probability of selecting sample $\mathbf{x}_j$ for step i. It is noticed that the $N+1$-th value of $\mathbf{p}$ indicates an action that there is no need to update the neighbor sets in this step. During training, an action a_i^t is taken by stochastically sampling from this probability distribution. During testing, the one with highest probability is chosen for synthesizing process.

State transition: After decision of action a_i^t in state $\mathbf{s}_i^t$ has been made, the next state $\mathbf{s}_{i+1}^t$ can be obtained via the state-transition function $\mathbf{s}_{i+1}^t = Transition(\mathbf{s}_i^t, a_i^t)$ where $\mathbf{z}_i^{t-1}$ is updated to the next unconsidered sample $\mathbf{z}_{i+1}^{t-1}$ in neighbor sets of $\mathbf{x}_y^{t-1}$. Then the neighbor that is least similar to $\mathbf{x}_y^t$ in the corresponding sets of $\mathbf{z}_i^{t-1}$ is replaced by $\mathbf{x}_j$ according to the action a_i^t. The *terminate state* is reached when all the samples of $\mathcal{N}_y^{t-1}, \mathcal{N}_o^{t-1}$ are considered.

Reward: During training, the agent will receive a reward signal r_i^t from the environment after executing an action a_i^t at step i. In our proposed framework, the reward is chosen to measure aging consistency between video frames as.

$$r_i^t = \frac{1}{\| \Delta_{i,\mathcal{A}(\cdot,\mathbf{x}_y^t)}^{\mathbf{x}^t|\mathbf{X}^{1:t-1}} - \Delta_{\mathcal{A}(\cdot,\mathbf{x}_y^t)}^{\mathbf{x}^{t-1}|\mathbf{X}^{1:t-2}} \| + \epsilon} \tag{5}$$

Notice that in this formulation, we align all neighbors of both previous and current frames to $\mathbf{x}_y^t$. Since the same alignment operator $\mathcal{A}(\cdot, \mathbf{x}_y^t)$ on all neighbor sets of both previous and current frames is used, the effect of alignment factors, i.e. poses, expressions, location of the faces, etc., can be minimized in r_i^t. Therefore, r_i^t reflects only the difference in aging information embedded into $\mathbf{x}_y^t$ and $\mathbf{x}_y^{t-1}$.

Model Learning: The training objective is to maximize the sum of the reward signals: $R = \sum_i r_i^t$. We optimize the recurrent policy network with the REINFORCE algorithm [42] guided by the reward given at each time step.

Figure 4: **Age Progression Results.** For each subject, the two rows shows the input frames at the young age, and the age-progressed faces at 60-years old, respectively.

4.3. Synthesizing from Features

After the neighbor sets of $\mathbf{x}_y^t$ are selected, the $\Delta^{\mathbf{x}^t|\mathbf{X}^{1:t-1}}$ can be computed as presented in Sec. 4.2.1 and the embedding of $\mathbf{x}_y^t$ in old age region $\mathcal{F}_1(\mathbf{x}_o^t)$ is estimated via Eqn. (3). In the final stage, $\mathcal{F}_1(\mathbf{x}_o^t)$ can then be mapped back into the image domain $\mathcal{I}$ via $\mathcal{F}_2$ which can be achieved by the optimization shown in Eqn. (6) [25].

$$\mathbf{x}_o^{t*} = \arg\min_{\mathbf{x}} \frac{1}{2} \parallel \mathcal{F}_1(\mathbf{x}_o^t) - \mathcal{F}_1(\mathbf{x}) \parallel_2^2 + \lambda_{V^\beta} R_{V^\beta}(\mathbf{x}) \quad (6)$$

where R_{V^β} represents the Total Variation regularizer encouraging smooth transitions between pixel values.

5. Experimental Results

5.1. Databases

The proposed approach is trained and evaluated using training and testing databases that are not overlapped. Particularly, the neighbor sets are constructed using a large-scale database composing face images from our collected **AGFW-v2** and **LFW-GOOGLE** [39]. Then Policy network is trained using videos from **300-VW** [34]. Finally, the video set from AGFW-v2 is used for evaluation.

LFW-GOOGLE [39]: includes 44,697 high resolution images collected using the names of 5,512 celebrities. This database does not have age annotation. To obtain the age label, we employ the age estimator in [32] for initial labels which are manually corrected as needed after estimation.

300-VW [34]: includes 218595 frames from 114 videos. Similar to the video set of AGFW-v2, the videos are movie or presentation sessions containing one face per frame.

Figure 5: **Age Progression Results.** Given different frames of a subject, our approach can consistently synthesized the faces of that subject at different age groups.

5.2. Implementation Details

Data Setting. In order to construct the neighbor sets for an input frames in young and old ages, images from AGFW-v2 and LFW-GOOGLE are combined and divided into 11 age groups from 10 to 65 with the age span of five years.

Model Structure and Training. For the policy network, we employ a neural network with two hidden layers of 4096 and 2048 hidden units, respectively. Rectified Linear Unit (ReLU) activation is adopted for each hidden layer. The videos from 300-VW are used to train the policy network.

Computational time. Processing time of the synthesized process depends on the resolution of the input video frames. It roughly takes from 40 seconds per 240×240 frame or 4.5 minutes per video frame with the resolution of 900×700. We evaluate on a system using an Intel i7-6700 CPU@3.4GHz with an NVIDIA GeForce TITAN X GPU.

5.3. Age Progression

This section demonstrates the validity of the approach for robustly and consistently synthesizing age-progressed faces across consecutive frames of input videos.

Age Progression in frontal and off-angle faces. Figs. 4 and 5 illustrate our age-progression results across frames from AGFW-v2 videos that contain both frontal and off-angle faces. From these results, one can see that even in case of *frontal faces* (i.e. the major changes between frames come from facial expressions and movements of the mouth and lips), or *off-angle faces* (i.e. more challenging due to the pose effects in the combination of other varia-

Figure 6: **Comparisons between age progression approaches**. For each subject, the top row shows frames in the video at a younger age. The next three rows are our results, TNVP [9] and Face Transformer [11], respectively.

tions), our proposed method is able to robustly synthesize aging faces. Wrinkles of soft-tissue areas (i.e. under the subject's eyes; around the cheeks and mouth) are coherent robust between consecutive synthesized frames. We also compare our methods against Temporal Non-volume Preserving (TNVP) approach [9] and Face Transformer (FT) [11] in Fig. 6. These results further show the advantages of our model when both TNVP and FT are unable to ensure the consistencies between frames, and may result in different age-progressed face for each input. Meanwhile, in our results, the temporal information is efficiently exploited. This emphasizes the crucial role of the learned policy network.

Aging consistency. Table 3 compares the aging consistency between different approaches. For the ***consistency measurement***, we adopt the average inverted reward r^{-1} of all frames for each synthesis video. Furthermore, to validate the ***temporal smoothness***, we firstly compute the optical flow, i.e. an estimation of image displacements, between frames of each video to estimate changes in pixels through time. Then we evaluate the differences (ℓ_2-norm) between the flows of the original versus synthesis videos. From these results, one can see that policy network has consistently and robustly shown its role on maintaining an appropriate aging amount embedded to each frame, and, therefore, producing smoother synthesis across frames in the output videos.

5.4. Video Age-Invariant Face Recognition

The effectiveness of our proposed approach is also validated in terms the performance gain for cross-age face verification. With the present of RL approach, not only is the consistency guaranteed, but also are the improvements made in both matching accuracy and matching score deviation. We adapt one of the state-of-the-art deep face matching models in [5] for this experiment. We set up the face verification as follows. For all videos with the subject's age labels in the video set of AGFW-v2, the proposed approach

Table 3: Comparison results in terms of consistency and temporal smoothness (*smaller value indicates better consistency*); and matching accuracy (*higher value is better*).

Method	Aging Consistency	Temporal Smoothness	Matching Accuracy
Original Frames	–	–	60.61%
FT [11]	378.88	85.26	67.5%
TNVP [9]	409.45	87.01	71.57%
IPCGANs [41]	355.91	81.45	73.17%
Ours(Without RL)	346.25	75.7	78.06%
Ours(With RL)	**245.64**	**61.80**	**83.67%**

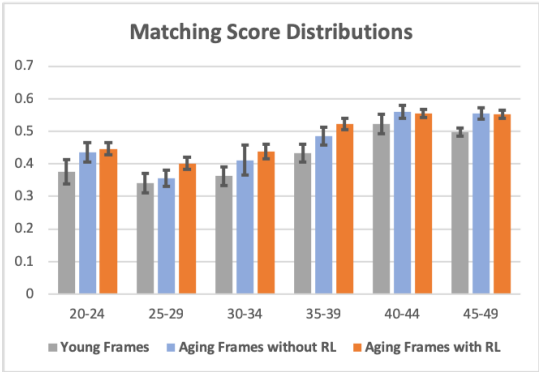

Figure 7: The distributions of the matching scores (of each age group) between frames of original and age-progressed videos against real faces of the subjects at the current age.

is employed to synthesize all video frames to the current ages of the corresponding subjects in the videos. Then each frame of the age-progressed videos is matched against the real face images of the subjects at the current age. The matching scores distributions between original (young) and aged frames are presented in Fig. 7. Compared to the original frames, our age-progressed faces produce higher matching scores and, therefore, improve the matching performance over original frames. Moreover, with the consistency during aging process, the score deviation is maintained to be low. This also helps to improve the overall performance further. The matching accuracy among different approaches is also compared in Table 3 to emphasize the advantages of our proposed model.

6. Conclusions

This work has presented a novel Deep RL based approach for age progression in videos. The model inherits the strengths of both recent advances of deep networks and reinforcement learning techniques to synthesize aged faces of given subjects both plausibly and coherently across video frames. Our method can generate age-progressed facial likenesses in videos with consistently aging features across frames. Moreover, our method guarantees preservation of the subject's visual identity after synthesized aging effects.

References

[1] Grigory Antipov, Moez Baccouche, and Jean-Luc Dugelay. Face aging with conditional generative adversarial networks. *arXiv preprint arXiv:1702.01983*, 2017.

[2] D Michael Burt and David 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.

[3] Bor-Chun Chen, Chu-Song Chen, and Winston H. Hsu. Cross-age reference coding for age-invariant face recognition and retrieval. In *ECCV*, 2014.

[4] C. Chen, W. Yang, Y. Wang, K. Ricanek, and K. Luu. Facial feature fusion and model selection for age estimation. In *Conf. on Automatic Face and Gesture Recognition (FG)*, pages 1–7. IEEE, 2011.

[5] Jiankang Deng, Jia Guo, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *arXiv preprint arXiv:1801.07698*, 2018.

[6] Chi Nhan Duong, Khoa Luu, Kha Gia Quach, and Tien D Bui. Beyond principal components: Deep boltzmann machines for face modeling. In *CVPR*, pages 4786–4794. IEEE, 2015.

[7] Chi Nhan Duong, Khoa Luu, Kha Gia Quach, and Tien D. Bui. Longitudinal face modeling via temporal deep restricted boltzmann machines. In *CVPR*, June 2016.

[8] Chi Nhan Duong, Kha Gia Quach, Khoa Luu, Hoai Bac Le, and Karl Ricanek Jr. Fine tuning age estimation with global and local facial features. In *Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–7. IEEE, 2011.

[9] Chi Nhan Duong, Kha Gia Quach, Khoa Luu, Ngan Le, and Marios Savvides. Temporal non-volume preserving approach to facial age-progression and age-invariant face recognition. In *The IEEE Int'l Conference on Computer Vision (ICCV)*, Oct 2017.

[10] FG-NET. Fg-net aging database. In *http://www.fgnet.rsunit.com*.

[11] FT. Face transformer (ft) demo. In *http://cherry.dcs.aber.ac.uk/transformer/*.

[12] Leon Gatys, Alexander S Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 262–270, 2015.

[13] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

[14] Xin Geng, Zhi-Hua Zhou, and Kate Smith-Miles. Automatic age estimation based on facial aging patterns. *PAMI*, 29(12):2234–2240, 2007.

[15] Ira Kemelmacher-Shlizerman, Supasorn Suwajanakorn, and Steven M Seitz. Illumination-aware age progression. In *CVPR*, pages 3334–3341. IEEE, 2014.

[16] Andreas Lanitis, Chris J Taylor, and Timothy F Cootes. Toward automatic simulation of aging effects on face images. *PAMI*, 24(4):442–455, 2002.

[17] H. N. Le, K. Seshadri, K. Luu, and M. Savvides. Facial aging and asymmetry decomposition based approaches to identication of twins. *Journal of Pattern Recognition*, 48:3843–3856, 2015.

[18] Gil Levi and Tal Hassner. Age and gender classification using convolutional neural networks. In *CVPRW*, 2015.

[19] K. Luu. Computer approaches for face aging problems. In *The 23th Canadian Conference on Artificial Intelligence (CAI)*. Ottawa, Canada, 2010.

[20] K. Luu, T. D. Bui, K. Ricanek Jr., and C. Y. Suen. Age estimation using active appearance models and support vector machine regression. In *Intl. Conference on Biometrics: Theory, Applications and Systems (BTAS)*. IEEE, 2009.

[21] K. Luu, T. D. Bui, and C. Y. Suen. Kernel spectral regression of perceived age from hybrid facial features. In *Conf. on Automatic Face and Gesture Recognition (FG)*, pages 1–7. IEEE, 2011.

[22] K. Luu, K. Ricanek Jr., T. D. Bui, and C. Y. Suen. The familial face database: A longitudinal study of family-based growth and development on face recognition. In *Robust Biometrics: Understanding Science Technology (ROBUST)*. IEEE, 2008.

[23] K. Luu, K. Seshadri, M. Savvides, T. D. Bui, and C. Y. Suen. Contourlet appearance model for facial age estimation. In *Intl. Joint Conf. on Biometrics (IJCB)*, pages 1–7. IEEE, 2011.

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

[25] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *CVPR*, pages 5188–5196, 2015.

[26] M. Minear and D. C. Park. A life span database of adult facial stimuli. *Behavior Research Methods, Instruments, Computers*, pages 630–633, 2004.

[27] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *Proceedings of IEEE Intl Conf. on Computer Vision and Pattern Recognition (CVPR-W 2017)*, Honolulu, Hawaii, June 2017.

[28] Eric 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.

[29] Eric Patterson, Amrutha Sethuram, Midori Albert, and Karl Ricanek. Comparison of synthetic face aging to age progression by forensic sketch artist. In *IASTED Int'l Conference on Visualization, Imaging, and Image Processing, Palma de Mallorca, Spain*, 2007.

[30] Eric Patterson, Amrutha Sethuram, and Karl Ricanek. An improved rendering technique for active-appearance-model-based automated age progression. In *Proceedings of ACM SIGGRAPH 2013: SIGGRAPH Posters*, 2013.

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

[32] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *Int'l Journal of Computer Vision (IJCV)*, July 2016.

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

[34] Jie Shen, Stefanos Zafeiriou, Grigoris G Chrysos, Jean Kossaifi, Georgios Tzimiropoulos, and Maja Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *ICCVW*, pages 50–58, 2015.

[35] Xiangbo Shu, Jinhui Tang, Hanjiang Lai, Luoqi Liu, and Shuicheng Yan. Personalized age progression with aging dictionary. In *ICCV*, December 2015.

[36] Jinli Suo, Xilin Chen, Shiguang Shan, Wen Gao, and Qionghai Dai. A concatenational graph evolution aging model. *PAMI*, 34(11):2083–2096, 2012.

[37] Jinli Suo, Song-Chun Zhu, Shiguang Shan, and Xilin Chen. A compositional and dynamic model for face aging. *PAMI*, 32(3):385–401, 2010.

[38] Ming-Han Tsai, Yen-Kai Liao, and I-Chen Lin. Human face aging with guided prediction and detail synthesis. *Multimedia tools and applications*, 72(1):801–824, 2014.

[39] Paul Upchurch, Jacob Gardner, Kavita Bala, Robert Pless, Noah Snavely, and Kilian Weinberger. Deep feature interpolation for image content changes. *arXiv preprint arXiv:1611.05507*, 2016.

[40] Wei Wang, Zhen Cui, Yan Yan, Jiashi Feng, Shuicheng Yan, Xiangbo Shu, and Nicu Sebe. Recurrent face aging. In *CVPR*, pages 2378–2386, 2016.

[41] Z. Wang, W. Luo X. Tang, and S. Gao. Face aging with identity-preserved conditional generative adversarial networks. In *CVPR*, 2018.

[42] Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. In *Machine Learning*, pages 229–256, 1992.

[43] F. Xu, K. Luu, and M. Savvides. Spartans: Single-sample periocular-based alignment-robust recognition technique applied to non-frontal scenarios. *Trans. on Image Processing (TIP)*, 24:4780–4795, 2015.

[44] J. Xu, K. Luu, M. Savvides, T. D. Bui, and C. Y. Suen. Investigating age invariant face recognition based on periocular biometrics. In *Intl. Joint Conf. on Biometrics (IJCB)*. IEEE, 2011.

[45] Hongyu Yang, Di Huang, Yunhong Wang, Heng Wang, and Yuanyan Tang. Face aging effect simulation using hidden factor analysis joint sparse representation. *TIP*, 25(6):2493–2507, 2016.

[46] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *CVPR*, July 2017.