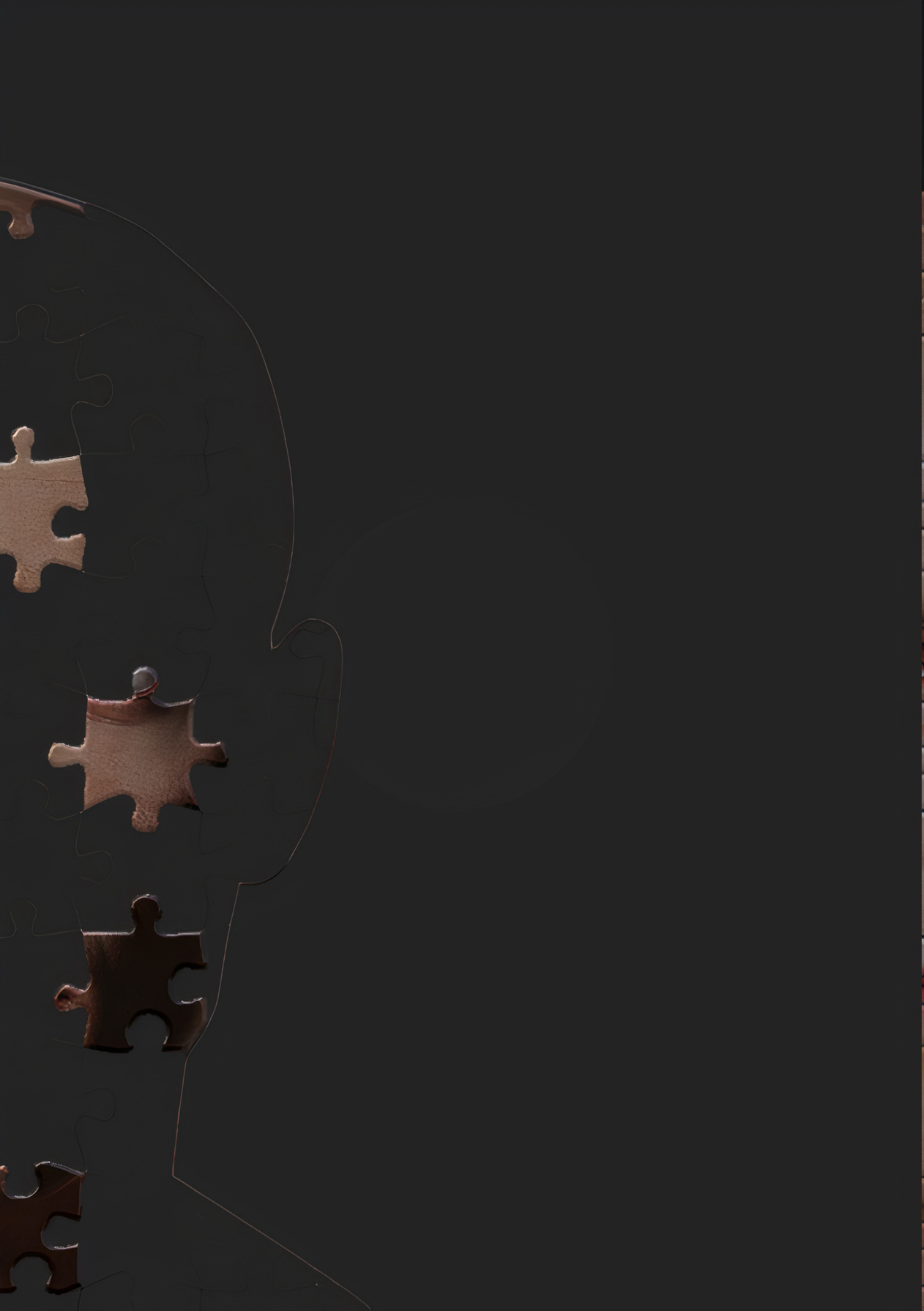
Learning image decomposition for face analysis and synthesis

Ngo, L.M.

Publication date
2021
Document Version
Final published version



Learning Image Decomposition for Face Analysis and Synthesis | Lê Minh Ngô

Learning Image Decomposition for Face Analysis and Synthesis

Lê Minh Ngô

Learning Image Decomposition for Face Analysis and Synthesis

Le Minh Ngo

This book was typeset by the author using LaTeX 2_ε.

Printing: Off Page, Amsterdam

Cover design: USIC Design and Finde Xumara.

ISBN 978-94-93197-81-7

Learning Image Decomposition for Face Analysis and Synthesis

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. K.I.J. Maex
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Aula der Universiteit
op woensdag 27 oktober 2021, te 14.00 uur

door

Le Minh Ngo

geboren te Kyiv

Promotiecommissie

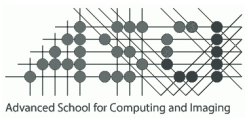
Promotor:	prof. dr. T. Gevers	Universiteit van Amsterdam
Co-promotor:	dr. S. Karaoğlu	Universiteit van Amsterdam
Overige leden:	prof. dr. E. Kanoulas	Universiteit van Amsterdam
	dr. S. Rudinac	Universiteit van Amsterdam
	dr. P. S. M. Mettes	Universiteit van Amsterdam
	prof. dr. E. Eisemann	Technische Universiteit Delft
	prof. dr. A. A. Salah	Universiteit Utrecht

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

This Page Intentionally Left Blank

UNIVERSITEIT VAN AMSTERDAM

The work described in this thesis has been carried out within the graduate school ASCI, dissertation number 422, at the Computer Vision Lab of the University of Amsterdam. The research is supported by 3DUniversum.



It ain’t what you don’t know that gets you into trouble.
It’s what you know for sure that just ain’t so.
— Mark Twain

Kính tặng Bố và Mẹ

Contents

Contents

Chapter 1

Introduction

The human face conveys multiple information cues such as ethnicity, identity, gender, and age. For example, facial wrinkles and sagging skin cues may correspond to the human aging process, while having a beard may indicate the gender of a person. Also certain facial cues may display information about the emotional or intentional state of a person. By analyzing faces, humans are capable to recognize their family, friends, members of their tribes and people they know.

Figure 1.1: El Corazón del Caribe research project. The British Museum. *Source: Newsweek.*

Humans have been interested in depicting faces since ancient times, as far as they could create drawings and sculptures. Face drawings in caves are the earliest known portraits (Fig. 1.1). With the innovation in tools and techniques, artists have discovered more advanced ways to represent the human face. Realistic portraits of important and wealthy people were common in many societies. In particular, Dutch painters from the Dutch

Golden Age are best known for their portrait paintings with highly detailed realism (Fig. 1.2). For many painters, it was the work of their entire life to mature their skills in modeling physics, anatomy, and color interaction on a canvas frame [Hodge, 2019].

Figure 1.2: Frans Hals. The Laughing Cavalier, 1624. Oil on canvas.

Nowadays, consumer cameras have become widely accessible. There is less demand for face paintings since everybody can capture face images and videos by their smartphones. However, the face doesn't become a less important subject. With the availability of computational resources and the advancement of Computer Vision algorithms, it becomes possible to perform different tasks given visual information about faces, captured by cameras, automatically. Insights about physics of color formation, used by the Golden Age painters as a basis in drawing their paintings, are successfully applied to different domains of computer vision, such as multiple view stereo and color correction [Andrew, 2001, Gevers et al., 2012]. In addition, many computer vision applications are inspired by drawing and painting, e.g. generating van Gogh-like paintings from images [Jing et al., 2017].

Computer vision algorithms, which learn from data, has shown a tremendous efficiency with the availability of large face datasets [Yang et al., 2016] together with an increasing amount of computational resources. Many face related tasks can be addressed by discriminative and generative learning models. *Discriminative learning* (analysis) tries to determine its attributes, e.g. location [Han et al., 2021, Ren et al., 2015], age [Dibeklioglu

et al., 2015, Savov et al., 2019], gender, and emotion [Arriaga et al., 2017] given an input human face. While in case of *generative learning* (synthesis), the face appearance is generated from a model given its attributes. Applications include face reenactment [Thies et al., 2016a, Wu et al., 2018a], face swapping [Ngo et al., 2020, Nirkin et al., 2019], aging [Huang et al., 2021a], makeup [Chen et al., 2019], and colorization [Isola et al., 2017].

Deep neural network-based methods have recently shown state-of-the-art results in face analysis and synthesis tasks. However, those tasks remain open research questions due to their complexity and biases in training data and predictive models. A recent line of research is focusing on the design of improved model architectures and the incorporation of *prior domain knowledge* [Hu et al., 2020, Sengupta et al., 2018] to create more robust predictive models which generalize better on unseen data, or require less training data to train on. For example, in the case of a general scene, by learning to decompose an image color into a component that only includes the object color (reflectance) and a component that is entirely dependent on the light source (shading) [Baslamisli et al., 2021] one can focus the scope of analysis solely on the shading component to predict the light source direction. In addition, for faces, prior knowledge about the human face, such as its geometry and surface properties, can be exploited.

The way that prior knowledge has been used in deep neural network-based models can be divided into two categories: (1) the domain knowledge is explicitly modeled in the network architecture, and (2) the domain knowledge is implicitly incorporated via loss functions.

1.1 Explicit Prior Knowledge Modeling

Assuming the face to be a Lambertian surface, a face image color $I(x)$ at location x can be decomposed into two components: one dependent on object intrinsic color (reflectance R), and one dependent on its interaction with the light source (shading S).

$$I(x) = R(x) \times S(x, \mathbf{n}, \mathbf{v}), \tag{1.1}$$

where $S(x, \mathbf{n}, \mathbf{v})$ is dependent on the surface normal $\mathbf{n}$ at position x and the light source direction $\mathbf{v}$. If there are different light sources, shading becomes a function of multiple light source directions. Surface normal $\mathbf{n}$ can be derived further as a variable dependent on object geometry G: $\mathbf{n} = f(G)$. Reflectance $R(x)$ and object geometry G can be further

constrained if additional prior knowledge about an object (face) is known. For example, the relative positions of eyes, nose, and mouth of a human face.

Research in the face domain have been focused on incorporating such kind of prior knowledge to improve model performance. Thus, the image formation model together with a parametric face model have been used to learn face alignment in an unsupervised manner [Koizumi and Smith, 2020] and for 3D face reconstruction [Genova et al., 2018, Tewari et al., 2017].

1.2 Implicit Prior Knowledge Modeling

Many computer vision tasks share common grounds. For example, object detection and semantic segmentation both require knowledge about object texture properties to be learned. Researchers are focusing on a joint exploration of multiple tasks to learn better generalized models with shared representations [Baslamisli et al., 2018, Le et al., 2018]. In the human face domain, different tasks are examined in the context of multi-task learning showing mutual benefits, e.g. age and gender estimation [Levi and Hassner, 2015], face alignment and landmark detection [Dong et al., 2018], face verification and expression recognition [Ming et al., 2019], age and identity estimation [Huang et al., 2021b].

For face synthesis, joint learning of additional attributes like emotion, hair style and skin color [Choi et al., 2018], the use of action units [Pumarola et al., 2018a] is beneficial in the realism of face images. Furthermore, different face synthesis tasks can be used to constrain the learning problem: Choi et al. [2018] learns a single generative model to generate faces by providing desired properties of different semantics (e.g. emotion, hair color); Sengupta et al. [2018] incorporates the image formation model to learn simultaneously the face shape, the reflectance and illuminance; Ngo et al. [2020] unifies facial reenactment and face swap tasks in a single model.

1.3 Research Outline and Questions

This thesis focuses on the improvement of models for face analysis and synthesis tasks by learning face attribute decomposition. When tackling a new computer vision challenge in face analysis and synthesis, in general, deep learning research typically addresses it by using a large datasets with labels.

Therefore, the overall research question of this thesis is as follows. **Can deep learning models benefit from learning to decompose the face representation by explicitly or/and implicitly incorporating domain-specific prior knowledge?**

1.3.1 Pose- and Expression- Robust Age Estimation

Recent methods for age estimation from a single image rely on 2D features and show shortcomings when the face appearance (e.g. head pose, facial expression) changes [Feng et al., 2018, Guo and Wang, 2012, Lu and Tan, 2013]. Due to the large variation of aging patterns [Angulu et al., 2018], it remains a challenging problem in computer vision research.

By incorporating a physics-based image formation prior into the model and by transferring the learned representation into the 3D space, the model may become more robust against 3D transformations such as head-pose and expression changes. Therefore, we address the following research question:

RQ1: How can age estimation models benefit from learning face attribute decomposition using image formation prior?

Chapter 2 focuses on this question by learning an effective representation. We propose an expression-, pose-, illumination-, reflectance-, and geometry-aware deep neural model which reconstructs a 3D face from a single 2D image together with learning visual age features. The idea is to minimize the negative influence of pose and expression variations and to obtain a face representation suited for robust age estimation. Our model is trained to jointly optimize 3D reconstruction and age estimation.

1.3.2 Identity-Unbiased Deception Detection

Recent studies in computer vision are focusing on high-stakes lie detection. Due to the lack of available datasets, there is no research done on both low-stakes and high-stakes deceit detection. Existing deception detection models are prone to identity and environment biases (e.g. skin color, gender, facial geometry, background, and lighting condition) when training on a small dataset. For instance, a model may obtain a high prediction accuracy by observing the skewed gender / identity distribution in positive samples, which is an undesired behavior for deception detection systems.

By incorporating prior knowledge about the face and image formation, and by removing environmental and identity information from the input data, an unbiased representation

for learning deception detection can be obtained. Consequently, the second research question is as follows:

RQ2: How can we use face image formation prior knowledge to remove biases from deception detection models?

To address this question, in Chapter 3, we propose an identity (i.e. facial geometry and skin reflectance) and environment (i.e. illumination) unbiased deceit detection system. Unbiasedness is obtained by conditioning on facial expression and head-pose-related features alone. A facial expression and head-pose feature space are disentangled from other properties by simultaneously learning two separate networks (1) one to predict the identity and environment parameters, and (2) another for temporally related features (i.e. expression and head pose), via image formation prior. A novel Low-Stakes Deceit dataset has been collected. We use the dataset to evaluate the existing automatic high-stakes deceit detection methods on the full spectrum of deceit. To our knowledge, our dataset is the first and only dataset used in low-stake deceit detection.

1.3.3 Self-supervised Face Image Manipulation

In Chapters 2-3, we study the benefit of attribute decomposition on discriminative learning tasks i.e. can generative learning tasks take advantage of it as well?

Manipulation of facial attributes (e.g. expression, pose, and lighting) from a single monocular image is important for different applications, such as video dubbing, augmented reality, and emotion recognition. Many methods require target-specific model training, meaning that the face images of unseen identities cannot be manipulated. Therefore, in this thesis, the third research question is:

RQ3: How can face image decomposition benefit face synthesis tasks?

To address this question, in Chapter 4, we propose a Conditional GAN pipeline conditioned on facial appearances. Appearance modeling is based on 2D-to-3D reconstruction. Facial appearance allows for simultaneously modeling different face attributes in the same feature space flexibly and compactly. By transferring the conditions to the appearance space, the many-to-one mapping problem of FACS (Facial Action Coding System) is circumvented and the method provides the flexibility of a continuous feature space from FACS and facial landmarks. For training, our method requires datasets of video sequences without any labels. During test time, our method manipulates different facial attributes given only one single unseen sample with possibly varying backgrounds and illumination conditions.

1.3.4 Face Reenactment and Swapping

In Chapters 2-4, we study the benefit of domain knowledge prior on various face analysis and synthesis tasks i.e. can we still achieve a better representation of learning without such prior knowledge?

Generating images or videos by manipulating facial attributes (i.e. face reenactment and swapping) has gained a lot of attention in recent years due to their broad range of computer vision and multimedia applications such as video dubbing [Suwajanakorn et al., 2017], gaze correction [Kuster et al., 2012], actor capturing [Kim et al., 2018a, Thies et al., 2016a], and virtual avatar creation [Nagano et al., 2018]. Recent methods show that face swap targeted methods can be used for face reenactment and vice versa. Unfortunately, the visual results on the second task are typically inferior to the first one [Nirkin et al., 2019, Siarohin et al., 2019]. Since those methods are designed for one of the tasks separately, they are not optimal for both.

By observing that face swap and reenactment tasks have many things in common, one can benefit from it by learning those tasks simultaneously. Consequently, the fourth research question is:

RQ4: Can we design a unified face reenactment and swapping without incorporating explicit prior knowledge?

In Chapter 5, we propose a novel pipeline that unifies face swapping and reenactment. A combined approach benefits from the similarities of the two tasks. Learning them together allows for robust face representation and enhances the realism of facial appearance. The proposed algorithm learns an isolated disentangled representation for face attributes without any supervision. Hence, our model can manipulate expression/pose, face identity, and style independently in latent space. We achieve this by directly mapping the disentangled latent representation to the latent space of a pre-trained generator. During inference time, the encoders condition the latent space by source and target face images together with their landmarks and generate the reenacted or swapped face using the pre-trained decoder.

1.4 Materials for Remaining Chapters

This thesis is composed of the following publications:

- **Chapter 2** is based on “Pose and Expression Robust Age Estimation via 3D Face Reconstruction from a Single Image", published in *IEEE/CVF International Conference on Computer Vision 2019, International Workshop on Human Behavior Understanding (ICCVW)* [Savov et al., 2019], by Nedko Savov[I], Lê Minh Ngô[I], Sezer Karaoğlu, Hamdi Dibeklioğlu and Theo Gevers.

  Contribution of authors

  Nedko Savov: all aspects;
  Lê Minh Ngô: all aspects;
  Sezer Karaoğlu: guidance, technical advice, supervision;
  Hamdi Dibeklioğlu: supervision, insights;
  Theo Gevers: supervision and insight.

- **Chapter 3** is based on “Identity Unbiased Deception Detection by 2D-to-3D Face Reconstruction", published in *IEEE/CVF Winter Conference on Applications of Computer Vision 2021 (WACV)* [Ngo et al., 2021], by Lê Minh Ngô, Wei Wang, Burak Mandira, Sezer Karaoğlu, Henri Bouma, Hamdi Dibeklioğlu and Theo Gevers.

  Contribution of authors

  Lê Minh Ngô: all aspects;
  Wei Wang: dataset;
  Burak Mandira: analysis, experiments;
  Sezer Karaoğlu: guidance, technical advice, supervision;
  Henri Bouma: insights;
  Hamdi Dibeklioğlu: technical advice, insights;
  Theo Gevers: supervision and insight.

- **Chapter 4** is based on “Self-supervised Face Image Manipulation by Conditioning GAN on Face Decomposition", published in *IEEE Transaction on Multimedia 2021 (TMM)* [Ngo et al., 2021], by Lê Minh Ngô, Sezer Karaoğlu and Theo Gevers.

  Contribution of authors

  Lê Minh Ngô: all aspects;
  Sezer Karaoğlu: guidance, technical advice, supervision;
  Theo Gevers: supervision and insight.

- **Chapter 5** is based on “Unified Application of Style Transfer for Face Swapping and Reenactment", published in *Asian Conference on Computer Vision 2020*

[I] Indicates equal contributions.

(ACCV) [Ngo et al., 2020] and in *Lecture Notes in Computer Science (LNCS)*, Volume 12626, December 2020, by Lê Minh Ngô[I], Christian aan de Wiel[I], Sezer Karaoğlu and Theo Gevers.

Contribution of authors

Lê Minh Ngô: all aspects;
Christian aan de Wiel: all aspects;
Sezer Karaoğlu: guidance, technical advice, supervision;
Theo Gevers: supervision and insight.

The author has further contributed to the following publications:

- Ipek Ganiyusufoglu, Minh Ngo, Nedko Savov, Sezer Karaoğlu, Theo Gevers, “Spatio-temporal Features for Generalized Detection of Deepfake Videos", under submission to Computer Vision and Image Understanding (CVIU).
- Tim de Haan, Minh Ngo, Sezer Karaoğlu, Theo Gevers, “ Unsupervised Target-Aware Face Blending", under submission to International Conference on Computer Vision (ICCV), 2021.

Chapter 2

Pose- and Expression- Robust Age Estimation

In this chapter, we present a deep learning architecture that exploits 3D face reconstruction to obtain a robust age estimation. To this end, effective representation is learned through an expression-, pose-, illumination-, reflectance-, and geometry-aware deep model reconstructing a 3D face from a single 2D image. The 3D face reconstruction network is combined with an appearance-based age estimation network, where the face reconstruction features are jointly learned with the visual ones. Experiments on large-scale datasets show that our method obtains promising results and outperforms state-of-the-art methods, especially in the presence of strong expressions and large pose variations. Furthermore, cross-dataset experiments show that the proposed method is able to generalize more effectively as opposed to state-of-the-art methods.

2.1 Introduction

The human face is an important source of information. Face properties may reveal different important cues such as emotion, intent, ethnicity, identity, gender, and age. The focus of this chapter is age estimation. Age estimation has many potential applications in

Published in *IEEE/CVF International Conference on Computer Vision, International Workshop on Human Behavior Understanding (ICCVW), 2019* [Savov et al., 2019]

daily life. For instance, in marketing, it can be employed for analyzing which age groups are interested in what kind of products, services, or entertainment. Vending machines of tobacco and alcohol can use age estimation to determine if the user is of legal age.

However, due to the large variation of aging patterns, addressed by Angulu et al. [2018], Geng et al. [2007], age estimation is a challenging task. Existing methods mostly rely on 2D information by exploiting appearance-related features. These features are either handcrafted [Gao and Ai, 2009, Guo et al., 2009, Phillips et al., 2000, Yang and Ai, 2007]) or obtained in a learning manner (e.g. through Convolutional Neural Networks (CNNs) [Hu et al., 2017, Levi and Hassner, 2015, Rothe et al., 2018, Sun et al., 2017, Yang et al., 2015, Zhang et al., 2017a]). Other methods use pose dependent distances between 2D facial landmarks [Farkas, 1994, Kwon et al., 1994] or learn manifolds to directly map 2D images to age.

Methods relying on 2D features have difficulties when the face appearance changes. For instance, a change in expression may introduce disturbing age-related patterns, like wrinkles, and may negatively influence the accuracy of age estimation methods [Guo and Wang, 2012]. Head pose variations that drastically change the facial appearance may also degrade the accuracy of age estimation algorithms [Feng et al., 2018, Lu and Tan, 2013]. These variations cause issues for other visual facial analysis tasks as well, like expression recognition [Rudovic et al., 2013] and landmark detection [Feng et al., 2018]. Robustifying methods for dealing with these variations are extensively explored for face identification [Ding and Tao, 2016]. One subset of solutions attempt to remove the variations from the input image by face frontalization or expression normalization, as a pre-processing step for face identification [Amberg et al., 2008, Zhu et al., 2015]. In that case, any failure from the normalization, being the inability to normalize or the presence of artifacts on the generated image, negatively affects the performance. Such approaches may help to preserve the identity related dominant face features which makes them suitable for identification. However, the reconstructed images lose important high-frequency information such as skin texture detail (i.e. wrinkles) which would reduce age estimation accuracy.

In this work, effective representation is learned through an expression-, pose-, illumination-, reflectance-, and geometry-aware deep model, reconstructing a 3D face from a single 2D image. The goal is to minimize the negative influence of pose and expression variations and to obtain a face representation which is suited for robust age estimation. The proposed model also learns the changes in facial appearance (2D image) through an

appearance subnet. These subnets (2D and 3D) are trained to jointly optimize the 3D reconstruction and age estimation.

The main contributions of this chapter are as follows:

1. To the best of our knowledge, we are the first to exploit 3D face reconstruction and 2D appearance features to jointly model pose and expression robust age estimation through multi-task learning.

2. The proposed multi-task learning model for age estimation achieves state-of-the-art accuracy on the Wiki database, as well as on cross-dataset experiments using UTK and AgeDB.

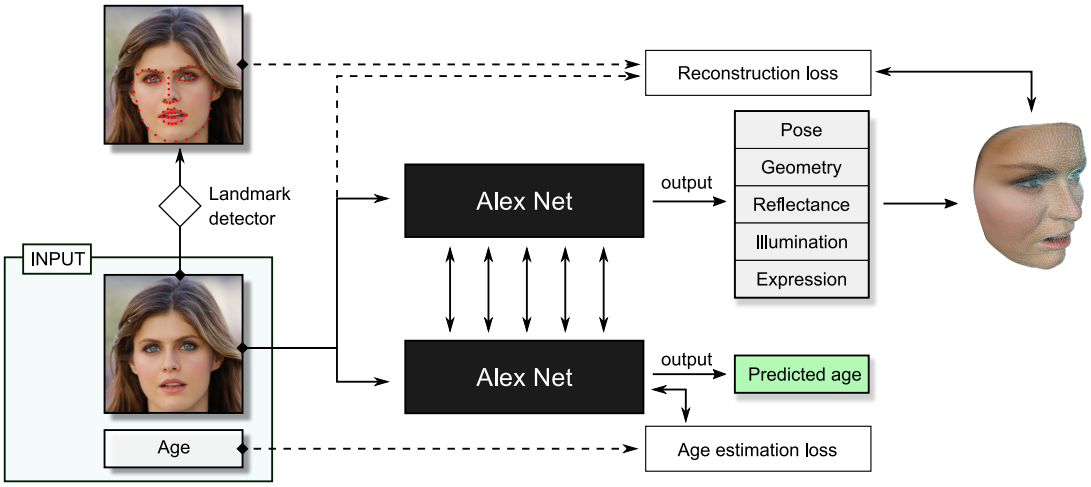


Figure 2.1: Our multi-task learning architecture for combining visual and 3D face reconstruction to perform robust age estimation. Two approaches are explored: (1) with hard parameter sharing - sharing the weights of a single AlexNet CNN, and (2) soft parameter sharing - two AlexNet CNNs have mutually connected layers.

2.2 Related Works

2.2.1 Age Estimation

Until recently, the predominant methods for performing age estimation are based on handcrafted features, focusing on wrinkles, skin texture and 2D shapes such as Local Binary Patterns [Phillips et al., 2000, Yang and Ai, 2007], Bio-Inspired Features [Guo

et al., 2009], and Gabor features [Gao and Ai, 2009]. However, while different hand-crafted features handle some adversarial conditions, none of them are fully robust against expressions, head pose, and illumination variations. More specifically, such approaches are quite sensitive to facial pose since it causes drastic changes in facial appearance.

Convolutional Neural Networks (CNNs) performs better than previous methods for age estimation [Hu et al., 2017, Levi and Hassner, 2015, Rothe et al., 2018, Sun et al., 2017, Yang et al., 2015, Zhang et al., 2017a]. Instead of mapping a full image to a certain age, as in manifold learning, CNNs aim to automatically learn efficient age-related features. Exploiting the benefits of CNNs, Rothe et al. [2018] proposes the Deep Expectation (DEX) algorithm, an age estimator that classifies age and, for more robust predictions, refines the inference prediction with a softmax expectation.

2.2.2 Pose and Expression Robustness

The effect of pose and expression on face analysis tasks is well studied. For instance, to provide pose robustness in face identification, Masi et al. [2019], Napoléon and Alfalou [2017], Paysan et al. [2009], Peng et al. [2017] augment their data by synthesizing face images for varying head poses using statistical 3D face models. In a similar way, Amberg et al. [2008] applies expression neutralization, and Zhu et al. [2015] employs pose normalization before face identification. These approaches may help to preserve the identity related dominant face features which makes them suitable for face identification. On the other hand, reconstructed/synthesized facial images lose important high-frequency details of skin appearance such as wrinkles, which would negatively influence age estimation. Nevertheless, our model is able to simultaneously learn multiple robust features, does not require labels other than age, and it is not influenced by face smoothing on neutralized images. In Lou et al. [2018], age and facial expressions are modeled jointly to achieve expression robustness in age estimation.

2.2.3 Monocular 3D Face Reconstruction

Monocular face reconstruction is the task of decomposing a face into its components (i.e. 3D facial geometry, expression, head pose, skin reflectance, and scene illumination). Computing these components for a single *RGB* image is an ill-posed problem. To this end, methods use statistical 3D models that represent 3D faces with a low dimensional parameter code vector [Blanz and Vetter, 1999, Booth et al., 2018, Gerig et al., 2018,

Paysan et al., 2009]. This code vector contains the encoded face components such as geometry, expression, skin reflectance, and additional parameters depending on the statistical 3D model.

Conventional 3D face reconstruction methods employ iterative optimization of an energy function. For instance, Blanz and Vetter [1999] optimizes the parameters by minimizing the error between the reconstructed and original face. Thies et al. [2016a] also uses an iterative approach, yet, it is designed to transfer facial expressions –in videos– between faces. In addition to being computationally expensive, energy minimization approaches have the problem of being reliant on favorable initialization because of typically non-convex functions to optimize. Deep learning methods exist using data augmentation techniques to produce results closer to ground truth fitting [Genova et al., 2018, Kim et al., 2018b]. Some other studies apply an analysis-by-synthesis approach to train the neural network using a physically plausible image formation model [Tewari et al., 2017]. We base our model for extracting pose and expression features on Tewari et al. [2017] with a number of modifications further discussed in this work. To the best of our knowledge, we are the first to use 3D face reconstruction for the age estimation problem.

2.3 Proposed Method

An overview of our method is shown in Fig. 2.1. Given a cropped face image **I**, our AlexNet-based CNN model learns to jointly produce the age prediction $\hat{y}$ (Section 2.3.2) and the 2D-to-3D reconstruction parameterized in a low dimensional latent space **z** (Section 2.3.1).

The appearance and 3D reconstruction features are combined using multi-task learning. Two methods are explored (Section 2.3.3). In the hard parameter sharing approach, a single shared CNN is adopted. The optimized loss is the weighted sum of a 3D fitting loss $\mathscr{L}_{fit}$ and an age estimation class distance loss $\mathscr{L}_{dist}$. In the soft parameter sharing approach, we use two separate CNN backbones and mutually connect multiple of their layers to allow sharing. The loss is the sum of $\mathscr{L}_{fit}$ and $\mathscr{L}_{dist}$. As a backbone, we use AlexNet [Krizhevsky et al., 2012] with a removed last fully connected layer.

2.3.1 Monocular 2D-to-3D Face Reconstruction Subnet

The employed monocular 2D-to-3D face reconstruction (fitting) model jointly decomposes a given 2D face image into its underlying components represented in a low

dimensional code vector $\mathbf{z}$: face rotation $\boldsymbol{\omega} \in \mathbb{SO}3$ and translation $\boldsymbol{\tau} \in \mathbb{R}^3$, face identity $\boldsymbol{\alpha} \in \mathbb{R}^{80}$, face expression $\boldsymbol{\delta} \in \mathbb{R}^{64}$, skin reflecntance $\boldsymbol{\beta} \in \mathbb{R}^{80}$ and illumination $\boldsymbol{\gamma} \in \mathbb{R}^{27}$. A fully connected layer with linear activation is added on top of the AlexNet backbone to infer $\mathbf{z}$.

$$\mathbf{z} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\gamma}, \boldsymbol{\omega}, \boldsymbol{\tau}\}. \tag{2.1}$$

Reflectance and Geometry

The facial geometry $\mathbf{G}(\boldsymbol{\alpha}, \boldsymbol{\delta}) \in \mathbb{R}^{N\times 3}$ and reflectance $\mathbf{L}(\boldsymbol{\beta}) \in \mathbb{R}^{N\times 3}$ are represented as a multilinear PCA model using the Basel Face Model 2017 [Gerig et al., 2018].

$$\mathbf{G}(\boldsymbol{\alpha}, \boldsymbol{\delta}) = \boldsymbol{\mu}_{geom} + \mathbf{E}_{id}[\boldsymbol{\alpha} \cdot \boldsymbol{\sigma}_{id}] + \mathbf{E}_{exp}[\boldsymbol{\delta} \cdot \boldsymbol{\sigma}_{exp}], \tag{2.2}$$

$$\mathbf{L}(\boldsymbol{\beta}) = \boldsymbol{\mu}_{ref} + \mathbf{E}_{ref}[\boldsymbol{\beta} \cdot \boldsymbol{\sigma}_{ref}], \tag{2.3}$$

where $\boldsymbol{\mu}_{geom}$, $\boldsymbol{\mu}_{ref} \in \mathbb{R}^{N\times 3}$ represent the mean neutral geometry and skin reflectance; $\mathbf{E}_{id}$, $\mathbf{E}_{ref} \in \mathbb{R}^{N\times 3\times 80}$, $\mathbf{E}_{exp} \in \mathbb{R}^{N\times 3\times 64}$ correspond to the linear bases of the PCA model together with their standard deviations $\boldsymbol{\sigma}_{id}$, $\boldsymbol{\sigma}_{ref} \in \mathbb{R}^{80}$, $\boldsymbol{\sigma}_{exp} \in \mathbb{R}^{64}$.

Camera Model

We model the face transformation to the camera space by a rigid transformation consisting of rotation $\mathbf{R}(\boldsymbol{\omega}) : \mathbb{R}^3 \rightarrow \mathbb{R}^{3\times 3}$ and translation $\boldsymbol{\tau}$ together with a full perspective transformation Π to obtain vertex coordinates $\mathbf{u}, \mathbf{v} \in \mathbb{R}^N$ on the camera plane.

$$\mathbf{u}, \mathbf{v} = \{u_i, v_i\} = \Pi \circ (\mathbf{R}(\boldsymbol{\omega})\mathbf{G}(\boldsymbol{\alpha}, \boldsymbol{\delta}) + \boldsymbol{\tau}),\ i \in \{1..N\}. \tag{2.4}$$

Illumination Model

We model illumination using the first B = 3 bands of Spherical Harmonics [Müller, 1966] bases $H_b(\mathbf{n}) : \mathbb{R}^{N\times 3} \rightarrow \mathbb{R}^N$ assuming the face surface to be Lambertian with a distant illumination ignoring self-occlusion and cast-shadows. Illumination coefficients are predicted separately for the *RGB* channels. Vertex normals $\mathbf{n}$ are estimated using 1-ring neighborhood. Shaded colour is computed as a Hadamard product between reflectance and shading:

$$\mathbf{C}(\boldsymbol{\beta}, \mathbf{n}, \boldsymbol{\gamma}) = \{c_i\} = \mathbf{L}(\boldsymbol{\beta}) \cdot \sum_{b=1}^{B^2} \gamma_b H_b(\mathbf{n}),\ i \in \{1..N\}. \tag{2.5}$$

Fitting

The energy formulation of Tewari et al. [2017] is used to train the proposed pipeline to predict the code vector $\mathbf{z}$. Our loss consists of a landmark loss $\mathscr{L}_{lan}$, a photometric loss $\mathscr{L}_{photo}$ and a regularization term $\mathscr{L}_{reg}$ balanced using weights λ_{lan} and λ_{photo}.

$$\mathscr{L}_{fit} = \lambda_{lan}\mathscr{L}_{lan} + \lambda_{photo}\mathscr{L}_{photo} + \mathscr{L}_{reg}. \tag{2.6}$$

Photometric Loss

We use the $L_{2,1}$ loss [Ding et al., 2006] to penalize the difference between the predicted per vertex shaded colour (Eq. 2.5) and the ground truth colour at positions $\{\mathbf{u}, \mathbf{v}\}$. The loss is defined for a subset of vertices $\mathscr{V}$ with normals directed toward the camera screen.

$$\mathscr{L}_{photo} = \frac{1}{|\mathscr{V}|}\sum_{i\in\mathscr{V}} \|I(u_i, v_i) - c_i\|_2. \tag{2.7}$$

Landmark Loss

We annotated 48 landmarks with vertex indexes $k_j, j \in \{1..48\}$ on the BFM model and penalize the L_2 difference between ground truth landmarks $\mathbf{l}_j$ and their corresponding prediction $\mathbf{p}_{k_j} = \{u_{k_j}, v_{k_j}\}$ from the 3D model.

$$\mathscr{L}_{lan} = \sum_{j=1}^{48} \|\mathbf{p}_{k_j} - \mathbf{l}_j\|_2^2. \tag{2.8}$$

Regularization

We regularize the model using Tikhonov regularization to enforce the model to predict faces closer to the mean.

$$\mathscr{L}_{reg} = \lambda_{alpha}\sum_{i=1}^{80}\alpha_i^2 + \lambda_{beta}\sum_{i=1}^{80}\beta_i^2 + \lambda_{delta}\sum_{i=1}^{64}\delta_i^2. \tag{2.9}$$

2.3.2 Appearance Subnet

We refer to our age estimation method, that learns visual age features, as the appearance subnet. Our appearance model is derived from Rothe et al. [2018] where the cross-entropy loss is used for resistance to outliers. Following Rothe et al. [2018], for further outlier resistance, we calculate the expectation over the softmax distribution to obtain a prediction $\hat{y}$ during the testing time:

$$\hat{y} = \sum_{i=0}^{M-m} (i+m) \cdot a_i. \tag{2.10}$$

A fully connected layer is added with an output activation vector $\mathbf{a}$ on the backbone. The ground truth age is denoted by y and its one-hot encoding by $\widetilde{y}$. The minimum and maximum age that a model can predict are $m = 0$ and $M = 80$. In contrast to Rothe et al. [2018], a distance term is added to the cross-entropy loss to penalize the probability mass which is different from the correct classes. The modified loss is defined as:

$$\mathscr{L}_{dist} = \lambda_{dist} \sum_{i=0}^{M-m} a_i \cdot d(i,y) - \sum_{i=0}^{M-m} \widetilde{y}_i \cdot \log(a_i), \text{ where } d(i,y) = |(i+m) - y|. \tag{2.11}$$

It is assumed that each class corresponds to one year of age and that the classes are indexed in order of monotonic increase. $d(.)$ is a distance function. Absolute distance is chosen to be used in this work, as it is a natural choice to represent distance and is outlier resistant. λ_{dist} is a constant used to tune the balance between the two terms.

2.3.3 Multi-Task Learning

Both subnets (3D and appearance) have AlexNet as a backbone. This establishes a correspondence between the layers of the two pipelines. The features of the tasks are different. However, as they are processed by the same filter size, the features are at the same scale of detail. The combination of these features is then suitable for processing by both pipelines following the shared layer. In this work, we attempt both hard and soft parameters sharing for multi-task learning.

Hard Parameter Sharing

A single AlexNet is shared for both tasks. The idea is that by joint training, the features are enforced to be suitable to age, and to pose and expression information. The last layer

contains the refined informative features for each task. The loss is a weighted sum of both tasks with weight w:

$$\mathscr{L}_{HPS} = (1-w) \cdot \mathscr{L}_{fit} + w \cdot \mathscr{L}_{dist}. \tag{2.12}$$

Soft Parameter Sharing

The hard parameter model forces the tasks to share all of their CNN features. This may not be optimal. Therefore, we employ a soft parameter sharing technique that can learn which layers to share. In this way, the tasks can produce independent high-level layers. For this, we use Cross-stitch Networks [Misra et al., 2016]. The approach is to have two instances of a backbone, i.e. A and B, one for each task. We choose to mark the 3D reconstruction subnet by A and the appearance subnet by B. So-called *cross-stitch* layers are then inserted in key positions in the deep network. Stitch layers take activations x_i from two layers, one from A and one from B, and blends them together as follows:

$$\begin{bmatrix} \bar{x}_A^i \\ \bar{x}_B^i \end{bmatrix} = \begin{bmatrix} \kappa_{AA}, \kappa_{AB} \\ \kappa_{BA}, \kappa_{BB} \end{bmatrix} \begin{bmatrix} x_A^i \\ x_B^i \end{bmatrix}. \tag{2.13}$$

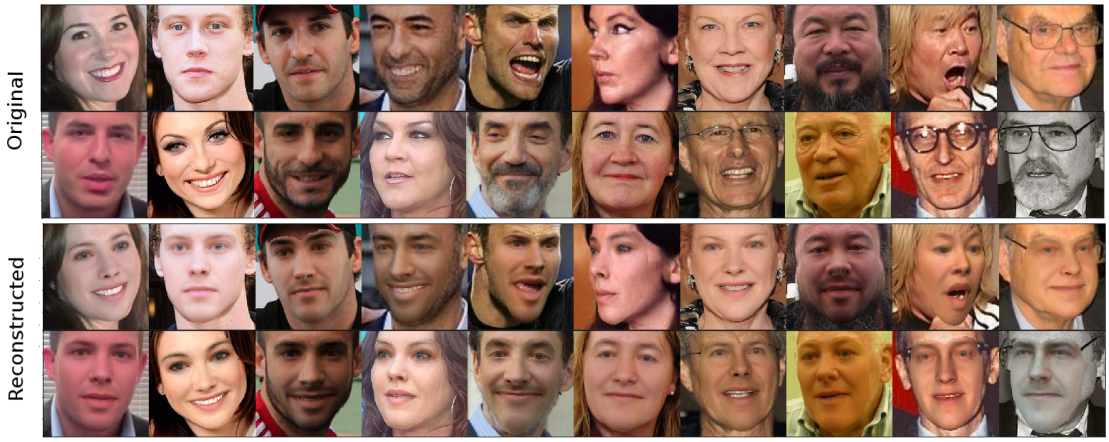


Figure 2.2: 3D reconstructions from our 3D face reconstruction model on samples from the Wiki test set. Shown are the original and the projected on them predicted 3D models.

The κ parameters are trained together with the architecture. They are common for all activations in a pair of layers. Misra et al. [2016] provides information about the positions for the cross-stitch layers inside AlexNet which we use after all max-pooling layers and fully connected layers. The final architecture is shown in Fig. 2.1.

In our implementation, Adam is chosen as an optimizer, but the κ parameters are trained separately with the Adagrad optimizer, to enforce a higher learning rate. This choice is meant to address the κ parameters receiving very small updates because of the magnitude of AlexNet activations, as noted by Misra et al. [2016]. To avoid overfitting, we apply L_2 regularization but only on the age estimation branch, since the 3D reconstruction subnet already has its own regularization term.

2.4 Datasets

The training data is based on the large scale IMDB-Wiki [Rothe et al., 2018] dataset. Different from other datasets, it contains in-the-wild faces with a variety of poses and expressions. Only the Wiki-Cropped subset is used, as it holds more accurate age annotations.

Wiki-Cropped is cleaned by filtering out data crawled from unregulated Wikipedia sandbox and user pages, black-and-white images and photos with undetected face by the dlib face detector [King, 2009]. We keep images labeled below 80 years of age and a maximum of 600 images per age label, to balance the data distribution. Our test set (referred to as Wiki test set) consists of 10% of the cleaned data. To ensure sufficient training data, the test set is distributed as closely as possible to the training set. We alternate between 5 age groups when building a training batch to enforce label diversity. The boundaries of the groups were chosen to be the 20th, 40th, 60th and 80th percentiles of the dataset distribution.

The landmarks are extracted by the dlib face detector [King, 2009] and used for landmark loss $\mathscr{L}_{lan}$ in training.

For cross dataset evaluation, we choose the manually annotated in-the-wild AgeDB dataset [Moschoglou et al., 2017] and the UTKFace dataset [Zhang et al., 2017b].

2.5 Experiments and Results

The success of our approach heavily relies on the success of each subnet, therefore we first demonstrate the qualitative results of our monocular 3D face reconstruction subnet. In Fig. 2.2, original images and their reconstructions can be seen. The reconstructions are visually accurate even under high pose and expression variations.

2.5.1 Evaluating the Appearance Subnet for Age Estimation

In this experiment, we compare the performance of our appearance subnet to two other recent age estimation approaches: Deep Regression Forests [Shen et al., 2018] and SSR-Net [Yang et al., 2018]. Like the proposed appearance baseline, both models are trained on the cleaned Wiki dataset. Training of the appearance subnet is performed with a learning rate of 10^{-5}, Adam optimizer, batch size 5, step learning rate decay and L_2 regularization with weight 0.01. The λ_{dist} parameter of the loss is set to 0.2 which results in close values of the distance loss component and the Cross-Entropy component.

We report on mean absolute error (MAE) between estimated and ground-truth age in Table 2.1. Our appearance subnet outperforms the other methods. We use it as a baseline for further experiments.

Method	MAE
SSR-Net [Yang et al., 2018]	7.33
Deep Regression Forests [Shen et al., 2018]	13.21
Appearance Subnet (Standalone)	**5.86**

Table 2.1: Best MAE test score of different age estimation methods trained on the Wiki dataset. The appearance subnet used as the visual baseline in this work outperforms the other two methods.

2.5.2 Joint Learning of Age Estimation and 3D Face Reconstruction

In this section, we study the performance of the appearance subnet with a joint classification of age estimation and 3D face reconstruction. We show that the performance of age estimation increases by exploiting features learned from the monocular face reconstruction.

In this and subsequent experiments, for soft sharing, we load pre-trained Alexnet weights for the 3D face reconstruction subnet in the joint model. For other cases, we load ImageNet classification pre-trained AlexNet weights. We apply L_2 regularization with weight 10^{-5}, dropout with rate 0.7 on the final layer of age estimation. For comparison, we used the same regularization scheme for the standalone appearance subnet and hard parameter sharing.

After tuning, the MAE score from the hard parameter sharing model (5.74 MAE) marks an age prediction improvement over the independent appearance subnet, as evident in Table 2.4, and shows the benefit from sharing the 3D reconstruction features.

Age estimation weight	MAE
Appearance Subnet	5.86
$w = 0.1$	5.95
$w = 0.3$	**5.74**
$w = 0.5$	5.78
$w = 0.7$	5.83
$w = 0.9$	5.89

Table 2.2: Best MAE test scores of the hard-parameter sharing model after training on Wiki dataset with different weights w for the age estimation loss. The weight of the 3D face reconstruction is $1-w$. Hard parameter sharing outperformed the appearance subnet.

Soft sharing parameters	MAE
$\kappa_{AA} = 0.9, \kappa_{AB} = 0.1$	5.58
$\kappa_{AA} = 0.8, \kappa_{AB} = 0.2$	5.68
$\kappa_{AA} = 0.7, \kappa_{AB} = 0.3$	5.52
$\kappa_{AA} = 0.5, \kappa_{AB} = 0.5$	**5.47**

Table 2.3: Best MAE test scores from tuning soft parameter sharing model's κ parameters on the Wiki dataset.

We obtain better MAE scores with κ parameters that encourage large sharing in the soft parameter sharing model. Table 2.2 gives an overview of the test performance of hard parameter sharing for different choices of the loss weight w. The MAE scores are decreasing with decreasing of the weight for age estimation, which means higher sharing with 3D face reconstruction.

For soft parameter sharing, we assess different choices for the amount of sharing by κ. Our initialization follows the rules $\kappa_{BB} = \kappa_{AA}$ and $\kappa_{AB} = \kappa_{BA} = 1 - \kappa_{AA}$. We chose non-sharing ($\kappa_{AB}$ and κ_{BA}) values from the range $[0.5, 1]$ in order to follow the predetermined rules. If smaller values are chosen, the branches would just switch the CNNs they rely mostly on. The test MAE scores after training on Wiki are shown in Table 2.3. The results show that age estimation benefits from large sharing. Significance of the best

results ($\kappa = 0.5$) is confirmed p-value $2.70 \cdot 10^{-5}$ from t-test after confirming normality with a normality test.

Having outperformed the hard parameter sharing, as shown in Table 2.4, the soft sharing age estimation seems to benefit from the independence of higher layers offered by the soft sharing architecture. As shown in Table 2.4, after 5 repeated training sessions per model, MAE score distributions are narrow and not overlapping. We can conclude that our age prediction is stable. For further experiments, we consider only the much better performing soft parameter sharing model.

Method	Mean ± Std
Appearance Subnet (Standalone)	5.86 ± 0.04
Proposed: HPS (Appearance + 3D Reconstruction Subnets)	5.74 ± 0.04
Proposed: SPS (Appearance + 3D Reconstruction Subnets)	**5.47** ± 0.03

Table 2.4: Mean best MAE test scores and deviations calculated from 5 training sessions on Wiki dataset of the appearance subnet, the proposed soft parameter sharing (SPS) and hard parameter sharing (HPS), combining the Appearance subnet with the 3D Face reconstruction subnet.

2.5.3 Analyzing the Age Estimation Improvements by Pose and Expression

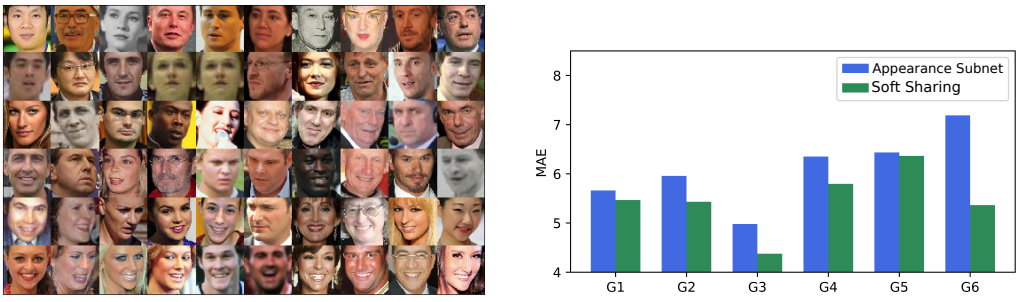
In this experiment, we evaluate the performance of the proposed soft parameter sharing model on varying pose and expression and compare it to the standalone appearance subnet. Each image in the test set is associated with expression (i.e. using predicted expression parameters) and head pose (i.e. using predicted head pose angle). We obtain an expression extremeness metric from the Euclidean norm of the expression vector $\boldsymbol{\delta}$. Our pose extremeness metric is based on the maximum of the exponential coordinates that parameterize a rotation $\boldsymbol{\omega} \in \mathbb{SO}3$. Separately for each of these metrics, we cluster the images into equally balanced groups. For each of the groups, the mean of the MAE differences over all the images falling in the group is computed and plotted to analyze the impact of our model on each challenge.

Fig. 2.3 (a) visualizes the expression strength of each group by showing a number of samples. Fig. 2.3 (b) shows how the MAE changes throughout the groups. The appearance subnet's MAE increases with increasing expressiveness whereas the soft

sharing method always scores better and performs similarly for the different ranges of expressiveness. Therefore, our proposed algorithm is more robust to expression variations. It improves over the appearance subnet the most on the most extreme expressions group (improvement is up to 1.8 MAE).

Fig. 2.4 (a) visualizes the head poses contained in each group. Looking at fig. 2.4 (b), the appearance subnet is much more likely to fail on more extreme poses than the soft sharing model. Moreover, the trend is that increasing the head pose extremeness leads to higher improvement over the appearance subnet. Therefore, the proposed algorithm is more robust to head pose variations. Notably, the improvement is highest for the most extreme head pose variations (1.4 MAE).

Fig. 2.5 further demonstrates the pose and expression robustness of the soft parameter sharing model by visually showing its superior predictions to the appearance subnet on the extreme pose and expression examples.

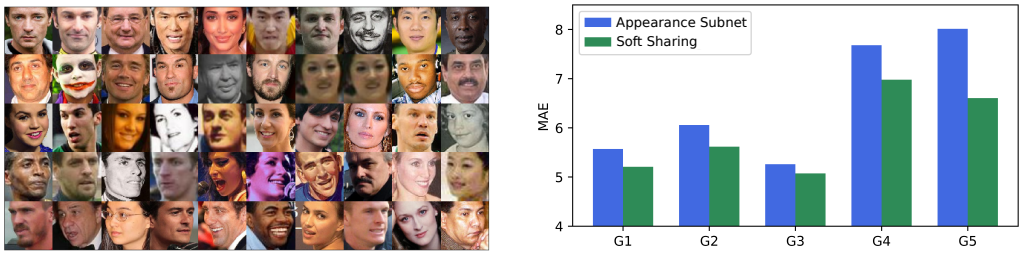


(a) (b)

Figure 2.3: (a) Samples from the expression intensity groups. Each row contains samples from one group. Groups are sorted by increasing metric from top to bottom; (b) The MAE for the soft sharing model and the standalone appearance subnet over the expression extremeness groups. The expression extremeness metric is increasing in the groups from left to right. The results show that the proposed model shows robustness to expression in contrast with the appearance subnet.

2.5.4 Cross-dataset Evaluation

To show if the results extend beyond the dataset used for training, evaluation is done on UTKFace and AgeDB. The expectation is to obtain MAE scores with soft parameter sharing, which are significantly lower than the MAE scores of the standalone appearance subnet. It is not common practice to provide results on cross-dataset evaluation for age



(a) (b)

Figure 2.4: (a) Samples from the rotation groups. Each row contains samples from one group. Groups are sorted by increasing metric from top to bottom. (b) The MAE measures for soft parameter sharing model and appearance subnet over the rotation extremeness groups. The rotation extremeness metric is increasing in the groups from left to right. The results show the robustness of the proposed algorithm to head pose in contrast to the appearance subnet.

Method	UTKFace	AgeDB
Appearance Subnet (Standalone)	9.73	10.27
Proposed: SPS (Appearance + 3D Reconstruction Subnets)	9.54	10.01
t-test p-value	$3.22 \cdot 10^{-9}$	$1.78 \cdot 10^{-8}$

Table 2.5: MAE scores from cross dataset evaluation of the appearance subnet and the soft parameter sharing model (SPS).

estimation since the performance may largely deteriorate. Results are shown in Table 2.5. It can be derived the improvements are significant. It shows the generalizable power of the soft sharing multi-task learning model.

2.6 Conclusion

In this chapter, we have shown that 3D reconstruction features can significantly improve the age estimation performance when jointly learned with appearance features. Our method takes a single 2D image and derives 3D reconstruction features as a new source of pose and facial expression robustness by employing a monocular 3D face reconstruction model. After evaluation, our method has shown to be consistently more robust across variation and improved over the baseline the most with extreme head poses (1.4 MAE) and intensive expressions (1.82 MAE).

Figure 2.5: Age predictions of the Appearance subnet (denoted as Baseline) and the soft parameter sharing model (denoted as SS) on non-frontal and non-neutral faces from the Wiki test set. The improvement of age prediction under extreme pose and expression conditions is visible.

Chapter 3

Identity-Unbiased Deception Detection

Deception is a common phenomenon in society, both in our private and professional lives. However, humans are notoriously bad at accurate deception detection. Based on the literature, human accuracy of distinguishing between lies and truthful statements is 54% on average, in other words, it is slightly better than a random guess. While people do not much care about this issue, in high-stakes situations such as interrogations for series crimes and for evaluating the testimonies in court cases, accurate deception detection methods are highly desirable. To achieve a reliable, covert, and non-invasive deception detection, we propose a novel method that disentangles facial expression and head pose related features using 2D-to-3D face reconstruction technique from a video sequence and uses them to learn characteristics of deceptive behavior. We evaluate the proposed method on the Real-Life Trial (RLT) dataset that contains high-stakes deceits recorded in courtrooms. Our results show that the proposed method (with an accuracy of 68%) improves the state of the art. Besides, a new dataset has been collected, for the first time, for low-stake deceit detection. In addition, we compare high-stake deceit detection methods on the newly collected low-stake deceits.

3.1 Introduction

Deceptive behavior is frequently displayed in daily life, yet, recognition of such behavior or lies is not an easy task for humans. On average, people can correctly classify only

Published in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021* [Ngo et al., 2021]

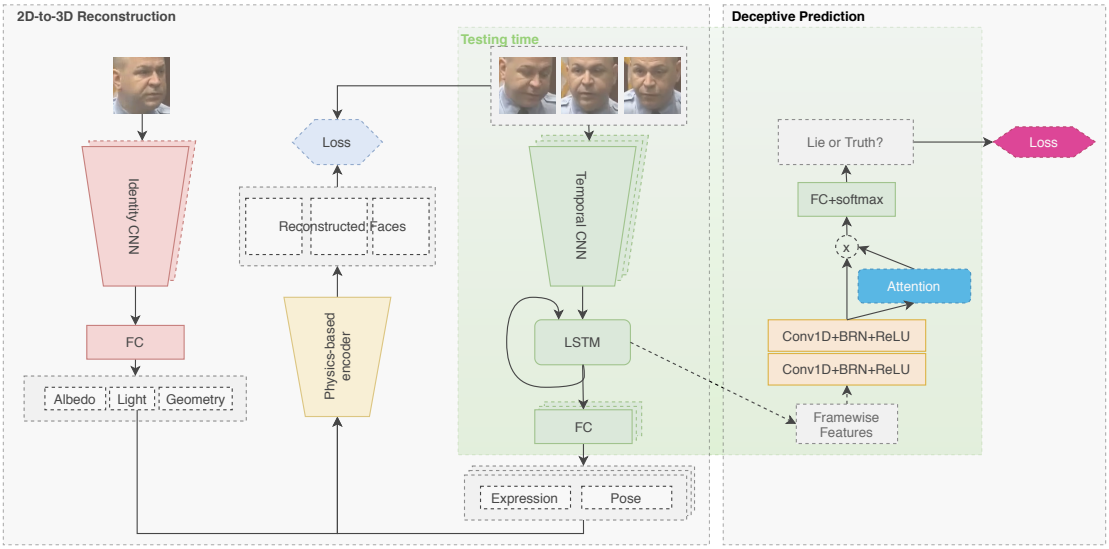


Figure 3.1: Architecture overview. Our proposed method decomposes temporally related features (expression and pose) from identity and environment properties by simultaneously training two CNNs (Identity and Temporal CNNs) to produce two sets of features using 2D-to-3D reconstruction. Features from the Temporal CNN are used for Deceptive Prediction.

47% of lies and 61% of truthful statements [Bond Jr and DePaulo, 2006].

Therefore, reliable methods for deception detection is an important need specifically for high-stakes situations such as court cases, and suspect/witness interrogations for further investigations and low-stakes situations to improve our daily communications. However, the ubiquitous polygraph, the most commonly known deception detection mechanism is unreliable [Fiedler et al., 2002].

Invasive approaches such as PET (positron emission tomography) and fMRI (functional magnetic resonance imaging) based methods perform better but they are neither fully reliable nor practical in deception detection where compactness or portability is required. Besides, the invasive nature of such mechanisms leaves them to be easily tricked by skilled deceivers [Fiedler et al., 2002]. Hence, deception detection requires non-invasive and covert methods for accurate detection. The difficulty in non-invasive deception detection lies in the weakness of external cues, since a large volume of work indicates that deceits are barely evident in behaviour [Hartwig and Bond Jr, 2014].

Recent developments in computer vision, along with the availability of deceptive behavior videos, have increased the research interest on deceit detection from visual patterns. The driving mechanism behind this ambition is the (subconscious) leakage of behavioral

cues to deception [Hartwig and Bond Jr, 2014]. These cues are often weak, very fast, or subjective, making them hard to interpret by humans. Recent studies on automated deception detection [Morales et al., 2017] exploits different behavioral modalities such as facial actions/expressions, head pose/movement, gaze, hand gestures, and even vocal features in the analysis [Abouelenien et al., 2014, Morales et al., 2017]. In contrast, our work focuses solely on temporally coherent disentangled facial cues.

High-level visual features used in the literature [Morales et al., 2017] such as facial action units are prone to errors due to challenging environmental conditions (i.e. illumination, viewpoint, occlusion, etc.). Thus, features extracted under challenging conditions can be unreliable. In this work, to cope with such issues, we propose to exploit 2D-to-3D face reconstruction to obtain an effective low-level representation for more reliable deception detection. 2D-to-3D face reconstruction aims at decomposing a face image into its components such as 3D facial geometry, expression, skin reflectance, head pose, and illumination parameters. Expression and head pose components are expected to carry important information for deceit detection [Lakhani and Taylor, 2003].

Although a successful decomposition has been a backbone for many face-related computer vision tasks (e.g. face recognition, emotional expression recognition, head pose estimation, etc.), this work is the first one that exploits face reconstruction for deceit detection. To this end, we propose an identity (i.e. facial geometry and skin reflectance) and environment (i.e. illumination) unbiased deceit detection system. Unbiasedness is achieved by conditioning on facial expression and head-pose related features alone. Facial expression and head-pose feature space are disentangled from other properties by simultaneously learning two separate networks, one to predict the identity and environment parameters and another for temporally related features (i.e. expression and head pose). Our results show that the proposed novel method for deception detection improves the state of the art high-stakes deceit detection, as well as it provides comparable results with the methods which make use of manually annotated facial attributes (e.g. facial actions/expressions, gaze, and head movement).

All prior automatic methods have been focusing on high-stakes deceit detection. There is no study available for automatic low-stakes deceit detection also because there is no low-stakes deceit detection dataset available. In our work, a novel Low-Stakes Deceit dataset has been collected with 624 high-res recordings of 312 subjects. To the best of our knowledge, the Low-Stakes Deceit dataset is the first and the only dataset available for low-stakes deceit detection. Besides, we use the dataset also to evaluate the existing automatic high-stakes deceit detection methods on the full spectrum of deceit.

To summarize, our contribution is four-fold:

- A novel method is proposed for deception detection on videos. The proposed method disentangles head pose and facial expression from facial identity (i.e. skin reflectance and 3D facial geometry) and illumination, using 2D-to-3D face reconstruction.
- The Real-Life Trial dataset has been cleaned and state-of-the-art high-stakes deceit detection methods have been re-evaluated using Leave-One-Person-Out (LOPO) validation.
- The proposed method outperforms the existing state-of-the-art and outperforms professional experts on the high-stakes deceit detection task.
- A new Low-Stakes Deceit (LSD) dataset is introduced. To our knowledge, it's the first visual dataset for low-stakes deceit detection. For the first time, we create a benchmark for state-of-the-art automatic high-stake deceit detection methods on low-stake deceit detection. The dataset will allow further research to be done on low-stake deceit detection.

3.2 Related Works

3.2.1 Deception Detection

At the basis of deception detection through nonverbal cues stands the leakage hypothesis, which states that –if the stakes of a lie are high enough– involuntary, subconscious cues of deceit will emerge from a liar [Hartwig and Bond Jr, 2014]. One can divide observable cues into physiological cues, body language cues, and facial cues. One of the problems with intangible constructs such as deceit is that these cues range from highly objective ones (vocal pitch) to highly subjective measurements (facial pleasantness). Hence, this section aims to provide an overview of objective, non-verbal cues that are relevant to the scope of using visual features for deception detection.

Concerning facial cues, a multitude of signals have been identified to correlate with deceit, such as lip pressing [Burgoon et al., 2017], smiling and pupil dilation, and facial rigidity [Pentland et al., 2017]. However, the studies often find contradictory results [Bouma et al., 2016, Vrij et al., 2019]. Besides, performance is highly dependent

on the data used for training and validation, with some datasets being noticeably easier than others [Wu et al., 2017a]. Secondly, the circumstances under which the lies were elicited are influential: multiple studies indicate that deceptive cues increase in magnitude with increased cognitive load [Vrij et al., 2017]. Hence, the final application and training data should have comparable cognitive load during data recording.

Micro-expressions pose another viable source of information [Yan and Chen, 2018], even though other studies have shown that only a small amount of people exhibit micro-expressions when lying [DesJardins and Hodges, 2015]. Facial action units (AUs) are also found to be informative for deceit detection [Morales et al., 2017].

One of the most recent methods of automated deceit detection is proposed by Morales et al. [2017]. This method fuses information from audio-visual modalities, where visual features in the form of 408 cues, including gaze, orientation, and FACS information, are extracted using OpenFace [Baltrušaitis et al., 2016] and later fused with verbal and acoustic features. Fusion occurs through a concatenation of statistical functional vectors, after which random forests and decision trees are used for deception classification. Differently, Pérez-Rosas et al. [2015] presents a baseline method for their introduced Real-Life Trial dataset, which models manually coded visual features such as expression, head movement, and hand gestures together with speech transcriptions using random forests and decision trees.

In literature, deceit is typically categorized into high-stakes (hold severe consequences for the liar) and low-stakes (simple lies that individuals get away with most often). All prior automated deceit detection methods, to our knowledge, have been focusing on the high-stakes deceit detection problem. Thus, there was no research has been done on low-stakes deceit detection. Low-stakes deceit detection is considered more challenging than high-stakes deceit detection since people in high-stakes situations are expected to behave more nervously [Lakhani and Taylor, 2003]. In more than 30 human behavior studies on low-stakes and high-stakes deceit conducted by other researchers an average accuracy of 55% has been achieved by *professional experts* on low-stakes in comparison with 67% for high-stakes [O'Sullivan et al., 2009].

3.2.2 Monocular Face Reconstruction

The decomposition of image components requires inverting the complex real-world image formation process. The reconstruction by inverting image formation is an ill-posed problem because an infinite number of combinations can produce the same 2D

image [Blanz and Vetter, 1999]. In general, we can categorize face reconstruction methods into two groups, namely, iterative [Blanz and Vetter, 1999, Garrido et al., 2013, Thies et al., 2015, 2016a] and deep learning based [Tewari et al., 2017]. Iterative approaches try to optimize parameters by minimizing the error between projected (reconstructed face) and the original image in an iterative (analysis-by-synthesis) manner [Blanz and Vetter, 1999]. The energy functions are mostly non-convex. The good fitting can only be obtained by close initialization to the global optimum, which is only possible with some level of control during image capture. Since these approaches are computationally expensive they are not preferred in this work.

Deep learning based methods, to reconstruct a face from a single monocular image, typically uses either data augmentation techniques to regress prediction to be close to the ground truth [Genova et al., 2018, Kim et al., 2018b, Sengupta et al., 2018] or applies the similar analysis-by-synthesis approach to train the neural network using a physically plausible image formation model [Deng et al., 2019, Genova et al., 2018, Koizumi and Smith, 2020, Tewari et al., 2017]. These methods produce sufficient reconstruction quality for certain tasks, however, they sacrifice details in order to be tractable for challenging, unconstrained images. Since such methods cannot avoid expression information to be leaked in 3D facial geometry, it is likely that there is an information loss while capturing expression. To reliably capture facial movements, the separation of 3D facial geometry and expression components are quite important.

Some works have been proposed to overcome such issues by using RGB videos instead of single monocular images [Garrido et al., 2013, Thies et al., 2015, 2016a]. However, these works are based on the iterative optimization approach (requires energy minimization for new input data). Convolutional Neural Network (CNN) architectures are recently explored for video-based dense real-time face reconstruction. In this work, we present a novel identity-aware, dense, and real-time face reconstruction CNN pipeline which receives RGB videos as input. Unlike previous monocular reconstruction methods, our method disentangles identity-related features (i.e. 3D facial geometry and reflectance) and illumination from temporally dependent parameters (i.e. expression and head pose) by simultaneously learning two CNNs for those sets of parameters using 2D-to-3D reconstruction. Disentanglement of temporally dependent features is important for deception detection since it allows our method to be unbiased towards subject identity and recorded environment.

3.3 Proposed Method

Our goal is to predict if a talking person is lying based on visual input i.e. face image. A sequence of RGB face images $\{\mathbf{I}_i\} \in \mathbb{R}^{W \times W \times 3}$ is passed to the Convolutional Neural Network (CNN) backbone to predict head-pose and facial-expression related features. Expression and head-pose are disentangled from other properties using 2D-to-3D reconstruction, which simultaneously learns latent face attributes together with environmental conditions. Constraining prediction on expression and head-pose alone allows us to be unbiased from facial identity and environment conditions which are irrelevant for deceit detection. Prior psychology studies have shown expression and pose-related behaviors such as eye contact, facial twitching, pauses, stuttering, and hesitance to be indicative of lie detection [Lakhani and Taylor, 2003]. Temporal features (i.e. expression and head pose) are used further in the second CNN to produce the final deceit detection. An overview of our method is shown in Fig. 3.1.

3.3.1 Modeling Deceptive Behaviour

We model lie detection as a Multiple Instance Learning problem [Ilse et al., 2018]. Given features extracted from video frames, our model assigns a single label (lie/truth) for the entire video. For a video annotated as a lie, we assume that there is at least one sub-sequence, where the person shows a deceptive behavior. For a video annotated as a truth, we assume that everything in the video is a truth. Thus, any sequence of frames which contains a lie sub-sequence is labeled as a lie. Given expression and pose related frame-wise features our deception prediction model extracts local temporal features $\mathbf{h_t} \in \mathbb{R}^{T \times C}$ using two layers of 1D-convolutions over the temporal dimension, where T is a sequence length, and C is the number of filters. Attention block *Att* weighs features based on their usefulness for the final task. Final linear layer *fc* with sigmoid is used to produce final prediction $\mathbf{y}$.

$$\mathbf{y} = \sigma\left(\text{fc}\left[\sum_t softmax(Att(\mathbf{h_t})) \cdot \mathbf{h_t}\right]\right). \tag{3.1}$$

3.3.2 Expression and Pose Features Disentanglement

A 2D face image $\mathbf{I}_i$ can be described using latent parameters $\mathscr{P} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\gamma}, \boldsymbol{\omega}, \mathbf{t}\} \in \mathbb{R}^{257}$ from which the original face can be reconstructed. We use CNN to predict those

parameters. $\boldsymbol{\alpha} = \{\alpha_i\}$, $\boldsymbol{\beta} = \{\beta_i\} \in \mathbb{R}^{80}$ and $\boldsymbol{\delta} = \{\delta_i\} \in \mathbb{R}^{64}$ are parameters correspond to 3D face geometry, albedo and expression; $\boldsymbol{\gamma} \in \mathbb{R}^{9\times3}$ describes scene illumination; $\boldsymbol{\omega} \in \mathbb{SO}(3)$ and $\mathbf{t} \in \mathbb{R}^3$ describe face rotation and translation.

Our model consists of two CNN backbones. The first component, **Identity CNN**, is used to predict identity and environment related parameters (identity geometry $\boldsymbol{\alpha}$, albedo $\boldsymbol{\beta}$ and lighting condition $\boldsymbol{\gamma}$). Face image is passed to the MobileNetV2 backbone [Sandler et al., 2018]. Its last layer is replaced by a fully connected layer with linear activation to predict $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ parameters. The second component, **Temporal CNN**, is used to predict face expression $\boldsymbol{\delta}$ and object transformations $\mathbf{w}, \mathbf{t}$ based on a sequence of RGB face images $\{\mathbf{I}_i\} \in \mathbb{R}^{W\times W\times3}$. MobileNetV2 backbone is followed by a recurrent layer LSTM and a fully connected layer with linear activation to predict $\boldsymbol{\delta}, \mathbf{w}, \mathbf{t}$.

We use LSTM to capture temporal relations between video frames and as an expression and pose-related feature space for the Deceptive Prediction network.

Disentanglement of expression and pose feature space from other properties is achieved by simultaneously learning all latent parameters $\mathscr{P}$ using 2D-to-3D reconstruction via *Physics-based encoder*. Disentanglement is an important property for our deception framework since it allows it to be unbiased towards identity and environment properties by assuming that lie cues are dependent on temporally related properties (expression, pose) only.

3.3.3 2D-to-3D Reconstruction

Albedo and Geometry

3D face geometry and albedo are parametrized using a multi-linear PCA model [Gerig et al., 2018]. Face geometry is represented as a point cloud $\mathbf{X}$ in the Euclidean space with the corresponding albedo attributes $\mathbf{B} \in \mathbb{R}^{N\times3}$.

$$\mathbf{X} = \mathbf{A}_{\mathbf{geom}} + \mathbf{P}_{\mathbf{id}}[\boldsymbol{\alpha} \cdot \boldsymbol{\sigma}_{id}] + \mathbf{P}_{\mathbf{exp}}[\boldsymbol{\delta} \cdot \boldsymbol{\sigma}_{exp}], \tag{3.2}$$

$$\mathbf{B} = \mathbf{A}_{\mathbf{alb}} + \mathbf{P}_{\mathbf{alb}}[\boldsymbol{\beta} \cdot \boldsymbol{\sigma}_{alb}], \tag{3.3}$$

where $\mathbf{A}_{\mathbf{geom}}, \mathbf{A}_{\mathbf{alb}} \in \mathbb{R}^{N\times3}$ are the mean face geometry and skin albedo; $\mathbf{P}_{\mathbf{id}}, \mathbf{P}_{\mathbf{alb}} \in \mathbb{R}^{N\times3\times80}, \mathbf{P}_{\mathbf{exp}} \in \mathbb{R}^{N\times3\times64}$ are principal components of PCA models for face identity, albedo and expression respectively; together with their standard deviations $\boldsymbol{\sigma}_{id}$, $\boldsymbol{\sigma}_{alb} \in \mathbb{R}^{80}$, $\boldsymbol{\sigma}_{exp} \in \mathbb{R}^{64}$.

Face Transformation

We model face movement in the scene using 6DoF transformation $\mathbf{T}$. Rotation matrix $\mathbf{R}(\mathbf{w}) : \mathbb{R}^3 \rightarrow \mathbb{R}^{3\times3}$ is represented in $\boldsymbol{\omega} \in \mathbb{R}^3 \in \mathbb{SO}(3)$, and translation $\mathbf{t} \in \mathbb{R}^3$ in x, y, z directions.

Illumination Model

Illumination changes are modeled using the first 3 bands of spherical harmonics basis function $\mathbf{H}_j$ assuming face is a Lambertian surface [Thies et al., 2016b]. The intensity of the i-th vertex c_i is defined as a product of vertex reflectance b_i and a shading component.

$$c_i = b_i \sum_{j=1}^{3^2} \gamma_j \mathbf{H}_j \Big(\mathbf{R}(\boldsymbol{\omega})\mathbf{n}_i \Big), i \in 1..N, \tag{3.4}$$

where $\mathbf{n}_i$ is a vertex normal of the i-th vertex. We define illumination parameters γ_j separately for each RGB channels, and consequently have 27 parameters in total. Vertex normal is estimated based on 1-ring triangle neighbors. Triangle topology is known from the face morphable model.

Projection Model

An obtained 3D point cloud $\mathbf{X}$ is mapped into a 2D plane by applying a rigid transformation $\mathbf{T}$ and perspective transformation $\mathbf{\Pi}$ which is a product of projection $\mathbf{V}$ and viewport $\mathbf{P} \in \mathbb{R}^{4\times4}$ matrices:

$$\begin{bmatrix} \hat{x} \\ \hat{y} \\ \hat{z} \\ \hat{d} \end{bmatrix} = \underbrace{\begin{bmatrix}\mathbf{V}\end{bmatrix} \times \begin{bmatrix}\mathbf{P}\end{bmatrix}}_{\mathbf{\Pi}} \times \underbrace{\begin{bmatrix} \mathbf{R}(\boldsymbol{\omega}) & \mathbf{t} \\ \mathbf{0} & 1 \end{bmatrix}}_{\mathbf{T}} \times \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}. \tag{3.5}$$

$\hat{u}, \hat{v}$ coordinates, and depth can be obtained by division by the homogeneous coordinate $\hat{d}$. The focal length is assumed to be fixed and principal points to be in the middle of the projection screen. $\hat{u}, \hat{v}$ together with vertex color c are used for producing the final reconstructed face.

Figure 3.2: Sample video frames from the RLT dataset. The dataset contains videos of trials under different lighting conditions, pose, with multiple people in the scene. Some of the videos are heavily occluded and don't contain visible facial features.

3.3.4 Training Losses

We use cross-entropy loss between ground-truth labels $\mathbf{y_{gt}} \in \{0, 1\}$ and predictions $\mathbf{y} \in [0, 1]$ to train our Deceptive Prediction pipeline.

$$\mathscr{L}_{dec} = \mathbf{y_{gt}} \cdot \log \mathbf{y} + (1 - \mathbf{y_{gt}}) \cdot \log(1 - \mathbf{y}). \tag{3.6}$$

For 2D-to-3D reconstruction we employ the energy minimization strategy of Tewari et al. [2017]. In total our loss consists of 3 main components: landmark loss E_{land}, vertex-wise photometric loss E_{vert} and regularization term E_{reg}.

$$\mathscr{L} = w_{land} E_{land} + w_{vert} E_{vert} + E_{reg}. \tag{3.7}$$

L_2 difference between landmark projections p from a predicted 3D face model and ground truth landmark l_j are used. In total, we use $|\mathscr{F}| = 48$ landmarks for optimization covering eyebrows, eye corners, nose, mouth, and chin.

$$E_{land} = \frac{1}{|\mathscr{F}|} \sum_{j \in \mathscr{F}} \| p_{k_j} - l_j \|_2^2, \tag{3.8}$$

where we define k_j as an annotated vertex index of the j-th landmark on the 3D model.

We define photometric loss as a $L_{2,1}$ difference [Ding et al., 2006] between vertex intensity color and its corresponded color from the original image. To find an intensity

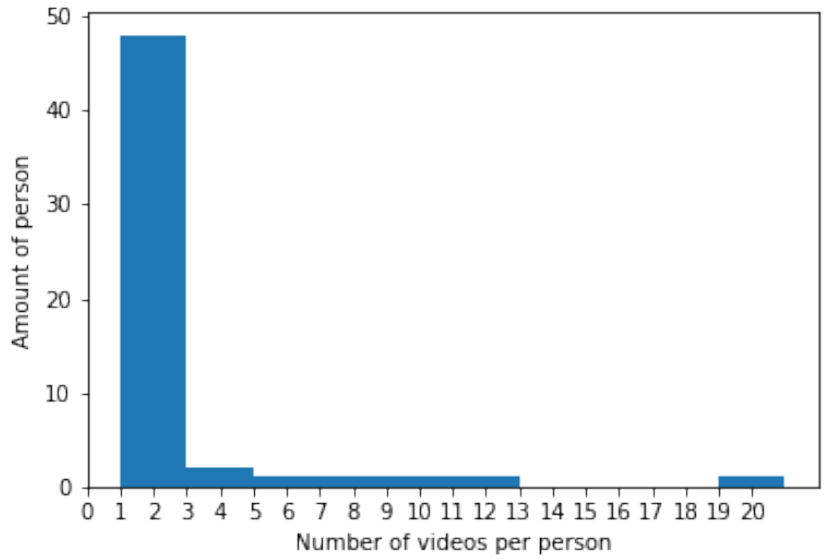


Figure 3.3: Distribution of videos per person in the RLT dataset. RLT dataset is imbalanced, with a few identities with a large number of videos.

color on image space we perform interpolation. We filter out vertices which contribute to the photometric loss based on normal direction, $|\mathcal{V}|$ is the number of vertices.

$$E_{vert} = \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \| \mathbf{c}_i - \mathbf{X}_{\hat{u}_i, \hat{v}_i} \|_2. \tag{3.9}$$

We use Tikhonov regularization [Thies et al., 2016b] to enforce parameters to be in the plausible range.

$$E_{reg} = w_\alpha \sum_{i=1}^{80} \alpha_i^2 + w_\beta \sum_{i=1}^{80} \beta_i^2 + w_\delta \sum_{i=1}^{64} \delta_i^2. \tag{3.10}$$

3.4 Datasets

Real-Life Trial Dataset

We employ the Real-Life Trial dataset [Pérez-Rosas et al., 2015] which contains 121 videos from real-life high-stakes scenarios that are publicly available. See Fig. 3.2 for visual samples from dataset. It has 61 deceptive and 60 truthful trial clips of 21 female and 35 male subjects whose ages vary between 16 and 60. The average duration of videos is about 28 seconds. When constructing the dataset, Pérez-Rosas et al. [2015]

enforce some visual constrains for videos such as the defendant or witness and his or her face should be identified during most of the footage.

Nonetheless, the video quality is noisy: the defendant's face is not always clearly visible in the video, the defendant and witnesses may appear both in the scene. Previous works, which rely on the confidence of the face detector alone, extract visual features from both defendant and witnesses for deceptive prediction. In addition, the dataset is unbalanced: the amount of videos per identity differentiates significantly (Fig. 3.3). One performing K-fold validation might include videos of the same person both in testing and training split, and hence achieving high accuracy. Consequently, for each video in the dataset, we have manually annotated all witnesses and removed them from the video sequence. If multiple faces appear in the scene, we remove all faces except the defendant. 5 videos without faces in the scene / occluded faces have been removed which leaves 116 videos for LOPO validation.

Low-Stakes Deceit (LSD) Dataset

We have collected a new dataset of low-stakes deceit which contains 624 high-res recordings of 143 males and 169 females interviewees under a controlled environment. Data collection was carried out as a part of Science Live, the innovative research programme of Science Center NEMO[I]. To our knowledge, our dataset is the first visual dataset available for studying low-stakes deceit in the literature. The age of participants varies between 7 to 72 years (Fig. 3.6). Among them, 209 participants speak Dutch and 103 participants speak English. Participants are facing the camera frontally and answer the interviewer's questions. The environmental conditions (e.g. illumination, background) are remained the same during whole recording sessions.

The interviewees are asked to describe two abstract scenes (Fig. 3.4): one on the visual card provided to the interviewee beforehand, and another which s/he did not see in advance. We define the first description as truth and the second as a lie. As a result, we have collected 2 recordings for every 312 identities with positive and negative labels. *Since the experiment setting doesn't imply a punishment for the contrived answer, the collected recordings can be used to study low-stakes lie.* Samples of our novel LSD dataset are shown in Fig. 3.5. We asked interviewees to judge peer recordings and used this information to measure the human accuracy on this dataset.

[I]Science Center NEMO, Amsterdam, http://www.e-nemo.nl.

Figure 3.4: Example of abstract scenes provided to the interviewees during the LSD dataset collection process.

Figure 3.5: Sample video frames from our newly collected LSD dataset. The dataset contains video of the similar lighting conditions, pose with a single person in the scene.

3.5 Implementation Details

We train our 2D-to-3D face reconstruction network for 200K iterations on 300VW [Chrysos et al., 2018] and CelebA datasets [Liu et al., 2015] using a batch size of 5 and Adam optimizer [Kingma and Ba, 2014] with learning rate of 10^{-5}. Loss weights are set to be $w_{vert} = 1.92$, $w_{land} = 0.0019$, $w_{\alpha} = 2.9 \times 10^{-5}$, $w_{\beta} = 4.93 \times 10^{-8}$, $w_{\delta} = 2.32 \times 10^{-5}$.

For training the Deceptive Prediction network we use RLT dataset for high-stakes lies and our newly collected LSD dataset for low-stakes lies. Models are trained on the batch

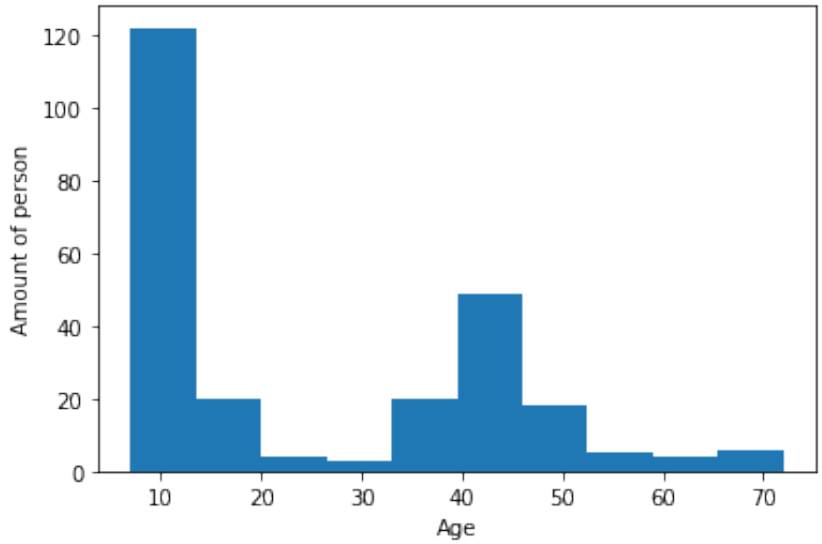


Figure 3.6: Age distribution of interviewees in our novel LSD dataset. Distribution is heavily tilted to 10 since it has been collected in the Science Museum popular among children.

size of 8 for 100 epochs. Early stopping is performed based on the validation score. We use Adam optimizer with a learning rate of 10^{-3}.

300VW contains video sequences with annotated 68 landmarks for each frame. We crop faces based on a bounding box on ground truth landmarks with 10% expansion. We process CelebA using dlib [King, 2009] for face detection and FAN [Bulat and Tzimiropoulos, 2017] for landmark detection. In total, we have collected 94K images from 300VW coming from 49 videos and 200K images from CelebA. Images from RLT and LSD datasets are processed in the same manner.

For each video sequence of 300VW we randomly select a cropped face as an input for the Identity-CNN. We randomly sample a sequence of 3 crop faces with a random step size from 1 to 5 frames as an input for the Temporal-CNN. For CelebA we assume that we have a 1-frame video sequence for each image. Images are randomly flipped to augment the dataset size. We train the model alternating CelebA and 300VW batches.

MobileNetV2 backbones are pretrained using ImageNet. We add offset to the 0-th band SH coefficient and z-translation to make sure the initial 3D face model has a plausible initial illumination and is centered in the middle of the screen. Basel Face Model 2017 [Gerig et al., 2018] is used for 3D face geometry, albedo and expression.

Type	Model	Feature	Accuracy	Precision	Recall
Manual	Pérez-Rosas et al. [2015] DT*	Hand-labeled features	0.67	0.64	0.74
	Pérez-Rosas et al. [2015] RF*	Hand-labeled features	0.71	0.70	0.70
Automatic	Morales et al. [2017] DT*	OpenFace features	0.50	0.48	0.38
	Morales et al. [2017] RF*	OpenFace features	0.56	0.57	0.40
	Hara et al. [2018]	CNN features	0.59	0.57	0.63
	Wu et al. [2017b] RF	Motion Features	0.54	0.55	0.42
	Zhao et al. [2017]	CNN features	0.47	0.44	0.26
	Ours	LSTM features	0.68	0.66	0.72

Table 3.1: State-of-the-art comparison on the high-stakes lies task using RLT dataset (*: only facial features are used).

Model	Feature	Accuracy (EN/NL)	Precision (EN/NL)	Recall (EN/NL)
Human	Visual + Audio	0.516	-	-
Morales et al. [2017] DT	OpenFace features	0.55 / 0.52	0.55 / 0.52	0.57 / 0.53
Morales et al. [2017] RF	OpenFace features	0.55 / 0.50	0.54 / 0.50	0.57 / 0.45
Hara et al. [2018]	CNN features	0.53 / 0.54	0.53 / 0.55	0.53 / 0.52
Zhao et al. [2017]	CNN features	0.47 / 0.51	0.47 / 0.52	0.51 / 0.44
Ours	LSTM features	0.54 / 0.52	0.53 / 0.52	0.64 / 0.65

Table 3.2: State-of-the-art comparison on the low-stakes lies task using LSD dataset.

3.6 Experiments and Results

In this section, we provide the details and results of conducted experiments. We start with a comparison with other methods on the high-stakes lies task. Next, we evaluate how methods designed for the low-stakes lies task performs in the low-stakes settings. Last, we provide additional analysis of age and gender effects. We considered lie as *positive* and truth as *negative* throughout the experiments when calculating accuracy, precision, and recall.

3.6.1 Baselines

In this section, we describe baselines for our experiments.

Morales et al. [2017] is tested with a decision tree (DT) and random forest (RF) classifiers with default parameters as in the papers. OpenFace [Baltrušaitis et al., 2016] is used to extract facial features in default output (i.e. basics, gaze, pose, 2D and 3D facial landmark locations, rigid and non-rigid shape parameters, action units) and apply some statistical metrics (max, min, mean, median, std, kurtosis, skewness, etc.) to create one feature vector per video.

Pérez-Rosas et al. [2015], which is the basis for Morales et al. [2017], is also implemented with a decision tree (DT) and random forest (RF) classifiers with default parameters as mentioned in their papers. They use manually labeled features. Since our system focuses only on facial features, we excluded hand-related features from their experimental setup to obtain comparable results.

3D-ResNext [Hara et al., 2018] is pretrained on Kinetics dataset [Kay et al., 2017] and finetuned starting from the third block. During training, a random temporal sampling of 30 frames is used. In inference, we use a non-overlapping sliding window of size 30 and take the mean scores of windows as the final score per video.

Time-CNN [Fawaz et al., 2019, Zhao et al., 2017] is a CNN for time series classification. This method reveals time series patterns through 1D convolutions on the temporal vector of each feature dimension.

DARE [Wu et al., 2017b] is a multimodal deception method. For our experiments we use a model with motion features only provided by authors.

3.6.2 High-stakes Deceit

We perform a comparison with other methods on the high-stakes deceit settings using the Real-Life Trial dataset. Results are reported in the Table 3.1. Leave-one-person-out (LOPO) validation is used to solve the dataset's flaw: the imbalanced amount of videos per subject (Fig. 3.3) which causes one subject to appear in both training and test splits when using K-Fold or leave-one-out validation. Subjects who have either too few (1) or too many videos (20% of the remaining videos) are always kept in the training set. 15% to 20% of the remaining videos are randomly separated as the validation. We try to get as much balanced as possible training and validation splits in terms of classes. To have a balanced training set, we randomly downsampled the majority class in terms of quantity to have an equal number of instances from each class.

Morales et al. [2017] mentioned 71.07% and 73.55% accuracy results for their visual model with DT and RF classifiers, respectively. However, they obtained these figures

erroneously by applying *leave-one-out* validation which causes subject overlaps between the test and train dataset. In this experiment, the results of both Morales et al. [2017] and Pérez-Rosas et al. [2015] are reported under LOPO settings instead.

The last row of Table 3.1 shows the performance of our proposed deception detection method. Our method performs on par with manual deceit methods that rely on hand-labeled features and achieves the best performance among automatic methods. Note that hand-labeled features are not possible in a real-life scenario. A significant improvement over other automatic facial feature extraction based methods shows that our method can extract more reliable facial features under challenging conditions since RLT dataset consists of varying illumination conditions and subjects are recorded under various viewing angles at various distances to the camera.

3.6.3 Low-stakes Deceit

We compare methods, which are designed for high-stakes settings, on the low-stakes deceit detection task using our newly collected LSD dataset. To our knowledge, we are the first to evaluate automatic deception detection methods on low-stakes deceit detection. Methods are evaluated separately on subsets with Dutch and English speakers. We applied X-Fold validation and made sure the same subject didn't occur simultaneously in training/validation/testing splits. Results are reported in the Table 3.2.

Automatic methods in general works as well as human evaluators (51.6% accuracy) on our benchmark, in spite of using visual-only features versus visual and audio for humans. In the case of our method, it's constrained to facial expression and pose related properties alone. This constraint prevents the model from biases toward subject identity and environment condition (an important property for deceit detection systems), however, simultaneously creates more challenges for deceptive behavior prediction. In addition, our dataset is collected under controlled settings (e.g. subjects are frontally facing the camera, subjects are sitting at a certain distance from the camera, faces are well lit). Such a controlled setting eases the problem of reliable facial feature extraction which explains why all automatic facial feature extraction based deceit detection methods achieve similar accuracy (54%) in our dataset.

Low-stakes deceit detection is a very challenging problem since people in low-stakes situations tend to behave less nervous, and hence showing less behavioral changes. In more than 20 human behavior studies on low-stakes deceit conducted by other researchers an average accuracy of 55% has been achieved by *professional experts* in comparison

with high-stakes deceit studies with an average accuracy of 67% [O'Sullivan et al., 2009]. Thus, our deceit detection method performs with similar accuracy to that of professional experts in the low-stakes deceit detection task on our benchmark.

3.6.4 Influence of Age

Since our LSD dataset provides age labels, we have clustered results into age classes to evaluate the correlation between age and deceit detection accuracy (Table 3.3). We separated samples into 3 categories: children, young adult, middle age, and above. We have observed higher accuracy on lie detection for children in comparison to adults in the English language split. This might be explained by children being more expressive with their expression. However, results require further research for a definitive conclusion.

Age	Lang.	Accuracy	Precision	Recall	# samples
< 18	EN	0.56	0.56	0.58	62
	NL	0.51	0.51	0.64	228
$\geq 18, < 45$	EN	0.53	0.52	0.66	106
	NL	0.53	0.52	0.70	134
≥ 45	EN	0.50	0.50	0.64	28
	NL	0.54	0.53	0.57	56

Table 3.3: Clustering low-stakes results by age.

3.6.5 Influence of Gender

We investigate the effect of gender on RLT and our LSD datasets. RLT dataset has been manually annotated with gender labels. The results are summarized in Table 3.4. High precision and recall values of females may suggest that the feature extraction of males is more challenging and has high variation. However, this can also be related to the number of samples as we have female subjects almost as twice as males subjects in the RLT dataset. For the low-stakes settings, we have observed better accuracy on the female split for English speakers with less conclusive results for Dutch.

Dataset	Gender	Accuracy	Precision	Recall	# samples
RLT	Male	0.65	0.38	0.50	46
	Female	0.70	0.76	0.77	70
LS EN	Male	0.53	0.53	0.58	106
	Female	0.55	0.54	0.69	98
LS NL	Male	0.52	0.51	0.67	176
	Female	0.52	0.52	0.64	242

Table 3.4: Gender-specific deceit detection results on the RLT and LSD datasets.

3.7 Conclusion

We have presented a novel method for deception detection based on reliable facial expression and head pose related features. Those properties have been disentangled using a 2D-to-3D face reconstruction technique which simultaneously learns (a) face identity, environment parameters, and (b) facial expression and head pose using separate convolutional neural networks, and hence achieves their separation. Our pipeline models deceit detection as a Multiple Instance Learning problem conditioned on reconstruction features. It's real-time and (with an accuracy of 68%) improves the state-of-the-art as well as providing on par results with the use of manually coded facial attributes (71%) in the high-stakes deception detection on the challenging RLT dataset. We have collected a new low-stake deceit detection dataset. To our knowledge, we are the first to evaluate automatic visual-based high-stake deceit detection methods on low-stakes deceit detection tasks. In the low-stakes lies deception detection task it has achieved results on par with professional experts however there is still room for improvement. We hope that the newly collected dataset will allow further research to be done on low-stake deceit detection.

Chapter 4

Self-supervised Face Image Manipulation

We present a novel architecture for manipulating facial expressions, head poses, and lighting conditions from a single monocular image. Recent methods based on Generative Adversarial Networks show promising results in expression manipulation. However, the variation is either defined by a limited number of classes or not well suitable for explicit manipulation of different attributes such as pose and lighting conditions. Besides, state-of-the-art methods are mostly focused on frontal faces.

Therefore, in this paper, a new Generative Adversarial Network architecture is proposed by explicitly conditioning on the appearance image space which is the product of direct manipulation of facial expressions, light and pose conditions of the face model in 3D space. In addition, the method only requires video sequences for training. Therefore, it is self-supervised. Unlike other face manipulation methods, the proposed method does not require target specific training. Large scale experiments show that our method outperforms state-of-the-art methods for different scenarios.

4.1 Introduction

Facial attribute (e.g. expression, pose, and lighting) manipulation from a single monocular image is important for different applications, such as video dubbing, augmented reality,

Published in *IEEE Transaction on Multimedia (TMM), 2021* [Ngo et al., 2021]

Figure 4.1: Manipulation (i.e. expression, pose, and illumination) of a single face image. We propose a novel GAN pipeline conditioned on facial appearance. Given a single image (the first column in this figure), the proposed method generates faces with new expressions, poses, and illumination conditions. The proposed method is self-supervised and does not require target specific training.

and emotion recognition. Based on recent developments of Generative Adversarial Networks (GAN's), current state-of-the-art methods in conditional image synthesis are able to generate realistically-looking images.

A pioneering method is StarGAN [Choi et al., 2018] producing realistic images by conditioning on a set of defined discrete attributes. In general, the set of attributes is limited as it requires annotation of the training data. Moreover, each attribute is concatenated to the input image as a separate channel resulting in a linear increase in the number of input parameters for the first layer. Recent modifications provide more flexibility by introducing conditioning over continuous variables such as landmarks and body poses [Pumarola et al., 2018b, Sanchez and Valstar, 2020].

GANimation [Pumarola et al., 2018a] achieves remarkable performance in manipulating

facial expressions by conditioning a GAN by the Facial Action Coding System (FACS). Because the use of FACS is a many-to-one mapping, different combinations of action units may lead to the same facial appearance. This makes face-to-face mapping difficult to learn. Moreover, FACS can only model facial expressions ignoring other attributes such as pose and illumination conditions.

Another line of work using GAN's, involves the manipulation of facial attributes in video sequences [Kim et al., 2018a, Wu et al., 2018a]. Those methods require a target-specific model training. Hence, face images of unseen examples cannot be manipulated. A separate model should be trained for each specific video which is a limiting factor in usability.

In this work, we propose a novel GAN pipeline conditioned on facial appearances. Appearance modeling is based on 2D-to-3D reconstruction. Facial appearance allows for simultaneously modeling different face attributes (See Fig. 4.1 for an illustration) in the same feature space in a flexible and compact manner. By transferring the conditioning to the appearance space, the many-to-one mapping problem of FACS is circumvented and the method provides the flexibility of a continuous feature space from FACS and landmarks. For training, our method requires datasets of video sequences [Chrysos et al., 2018, Gross et al., 2010, Wolf et al., 2011] without any label/GT. During test time, our method manipulates different facial attributes given only one single unseen sample with possibly varying backgrounds and illumination conditions. Choi et al. [2018] and Pumarola et al. [2018a] perform conditioning in attribute and FACs space respectively without direct 2D image correspondences.

The main contributions of the chapter are as follows:

- We introduce a pipeline for self-supervised face manipulation. Only a single unseen monocular image is used as input. No target-specific training is required.
- The proposed method uses image formation (i.e. face image decomposition) as the basis of GAN conditioning. Therefore, it manipulates different face image components (expression, pose, and light) based on a compact appearance representation.
- Deep insights (with numerical experiments) are provided for the potential application of our novel pipeline.

4.2 Related Works

4.2.1 Generative Adversarial Networks

Generative Adversarial Networks (GANs) have shown to produce results with high visual realism on the task of image synthesis. They aim to approximate the target data distribution by training alternately two models, one trying to generate samples close to the real distribution and another trying to distinguish the generated from real samples [Goodfellow et al., 2014]. Many of the follow-up methods are focusing on stabilization of the training process of GANs by introducing modification on training losses (Wasserstein GAN [Mahajan et al., 2018], Geometric GAN [Lim and Ye, 2017], cycle consistency [Zhu et al., 2017]), normalization techniques (spectral normalization [Miyato et al., 2018], layer normalization [Ba et al., 2016]) and GANs architecture (by introducing multiple decoders for a generator [Zhou et al., 2019] or adversarial approximator [Xu et al., 2019]).

4.2.2 Image-to-image Translation

Methods based on conditional GANs [Isola et al., 2017, Mirza and Osindero, 2014] makes modeling on conditional distribution on an input data possible, and hence can be used for image-to-image translation tasks including style transfer [Zhu et al., 2017], scene generation [Wang et al., 2018], intrinsic image decomposition [Lettry et al., 2018] and inpainting [Yu et al., 2018]. Pix2Pix [Isola et al., 2017] combines an encoder-decoder architecture with adversarial training to predict a target image given a source image based on pixel correspondences. To overcome the limitation of pixel-to-pixel correspondence, Zhu et al. [2017] introduces a cycle consistency loss which allows the model to be trained without image pairs. To manipulate an input image using attributes, StarGAN [Choi et al., 2018] proposes to concatenate attributes as additional channels. As attributes are discrete, they are limited in coping with semantically meaningful interpolation. To this end, GANimation [Pumarola et al., 2018a] extends the StarGAN architecture to include continuous attributes. Conditioning on attributes in StarGAN and GANimation results in a linear complexity: the number of attributes is equal to the additional channels to be concatenated to the input. Besides, the GANimation model focuses on modeling a single type of continuous attributes (i.e. action units). In contrast, by using the face decomposition space, our proposed method models different types of attributes, such as

face identity, expression, pose, and lighting condition. This enables us to compress the representation of attributes into a 3 channel-image (Fig. 4.2).

4.2.3 Face Image Manipulation

Face image manipulation focuses on editing different face attributes used in various applications such as virtual makeup [Chen et al., 2019], face enhancement [He et al., 2019], beautification [Diamant et al., 2019], aging [Liu et al., 2019] and relighting [Han et al., 2020]. Chen et al. [2019] condition a generative network on an input face image and a reference non-makeup domain face. Their Glow architecture tries to disentangle make-up features from non-makeup features and forces a prediction to remove makeup features. He et al. [2019] condition the decoder of a generative network on discrete space attributes, similar to Liu et al. [2019] which condition their model on discrete age and gender attributes. Han et al. [2020] use light source classes to condition their generative model to produce images with certain lighting conditions. Diamant et al. [2019] condition a model on a continuous attribute representing the level of beauty of the desired face. In the case of He et al. [2019] and Han et al. [2020], additional attribute classification loss is used together with discriminator to improve the adversarial signal, while Liu et al. [2019] come up with a modification on discriminator based on wavelet packet transformation. In many cases, Perceptual loss is used to encourage similarity between predicted and generated images [Chen et al., 2019, Diamant et al., 2019, Liu et al., 2019].

Input to the face image manipulation algorithm can be either a single monocular image [Choi et al., 2018, Pumarola et al., 2018a], an image sequence of the same person [Kim et al., 2018a, Thies et al., 2016b, Wu et al., 2018a] or a 3D texture [Nagano et al., 2018]. Early approaches addressed the problem of face image manipulation by warping using landmarks [Averbuch-Elor et al., 2017], intrinsic image decomposition [Li et al., 2018], and face priors [Blanz and Vetter, 1999]. Most of the recent methods are GAN-based [Diamant et al., 2019, Han et al., 2020, He et al., 2019, Kim et al., 2018b, Liu et al., 2019, Pumarola et al., 2018a, Wu et al., 2018a]. However, those GAN methods are constrained by the discrete number of classes for manipulation [Choi et al., 2018, Han et al., 2020, He et al., 2019, Liu et al., 2019], same head poses and lighting conditions [Nagano et al., 2018, Pumarola et al., 2018a], or they require a model to be trained specifically for each target image/video [Kim et al., 2018a, Wu et al., 2018a]. In contrast, our method (1) only requires a single monocular image during test time, (2) no target-specific training

is required, (3) can change simultaneously different face image components such as expression, pose and illumination direction.

Recently, GAN-based face reenactment methods are proposed [Nirkin et al., 2019, Siarohin et al., 2019, Zakharov et al., 2020] capable of manipulating both the head pose and facial expressions. However, for large pose variations, these methods may produce unrealistic results [Nirkin et al., 2019, Siarohin et al., 2019], or may fail to preserve the image background [Zakharov et al., 2020] and the identity of the face [Nirkin et al., 2019]. In contrast, our model maintains the face identity and image background while keeping the facial expressions consistent.

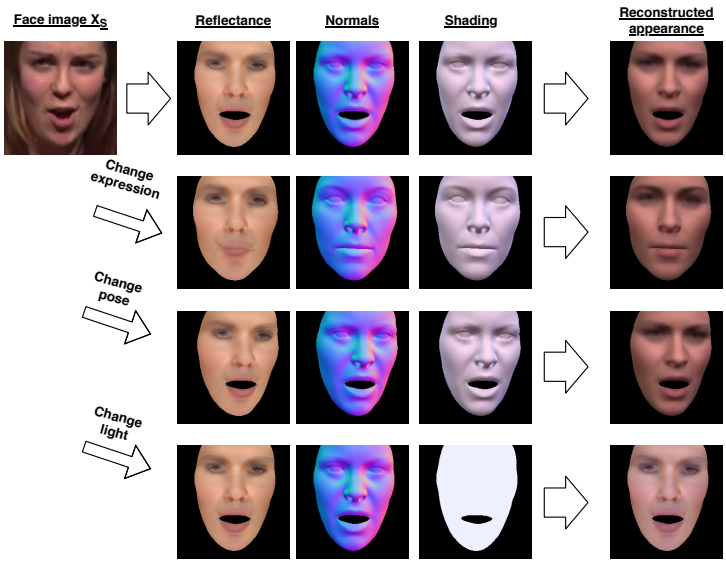


Figure 4.2: Appearance modeling via 2D-to-3D reconstruction. Using a single representation, appearance conditioning allows for different facial attributes to be changed simultaneously such as facial expressions, head poses, and lighting conditions.

4.3 Proposed Method

An overview of our method is shown in Fig. 4.3. Given a cropped source face image $\mathbf{X}_S \in \mathbb{R}^{H\times W\times 3}$ and a target appearance image $\mathbf{Y}_T$, the generative model G produces a face image $\hat{\mathbf{X}}_S \in \mathbb{R}^{H\times W\times 3}$ with a new expression, pose and lighting condition described by $\mathbf{Y}_T$, but preserves the face identity and background constrained by $\mathbf{X}_S$ (Section 4.3.1).

The appearance image representation is obtained using a 2D-to-3D reconstruction of the face image $\mathbf{X}_S$, followed by applying expression, light and pose in latent space $\mathbf{x}$ and projecting it back onto the (original) image plane (see Fig. 4.2 and Section 4.3.2).

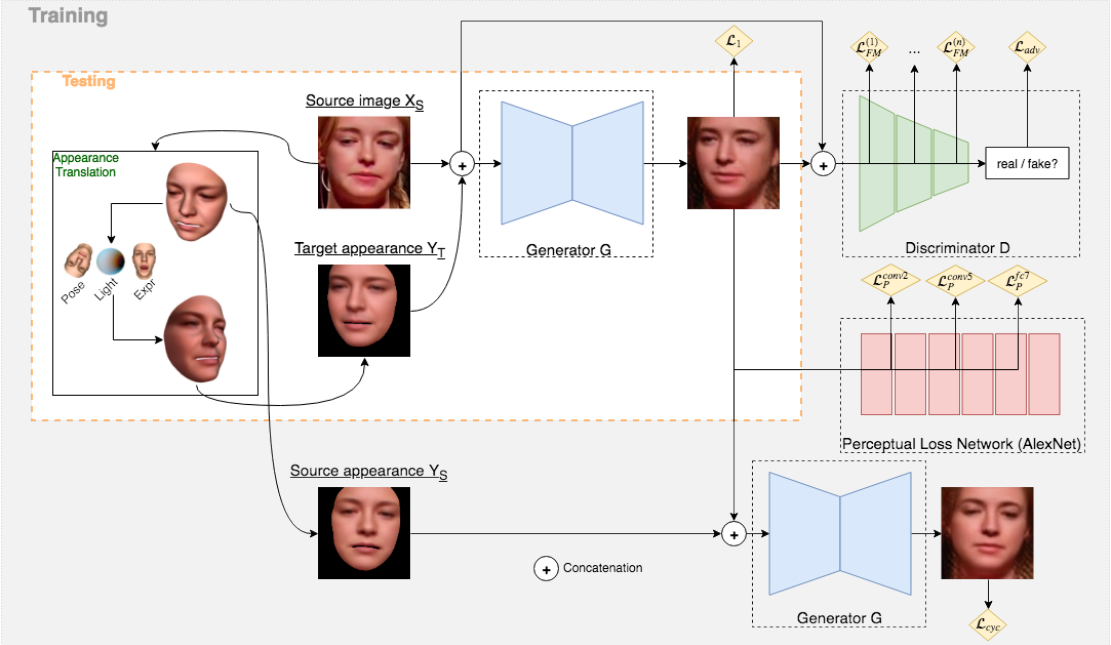


Figure 4.3: Our proposed method generates realistically-looking face images with different facial expressions, lighting conditions, and poses. Generative model G predicts a face with background $\hat{\mathbf{X}}_\mathbf{S}$ conditioned on the source image $\mathbf{X}_S$ and the reconstructed target appearance $\mathbf{Y}_T$. Architectural blocks with the same color have shared weights.

We train our model by optimizing alternately the generator G and the discriminator D in a two-player minimax game (Eq. 4.1). G tries to fool D with fake examples $\hat{\mathbf{X}}_S$, and D tries to differentiate the generated examples $\hat{\mathbf{X}}_S$ from real images $\mathbf{X}_S$ (Section 4.3.3).

$$G* = arg \min_G \max_D \mathscr{L}(G, D). \tag{4.1}$$

4.3.1 Network Architecture

Generator

Our architecture is inspired by the network design of Conditional GAN's for the image-to-image translation task [Choi et al., 2018, Isola et al., 2017, Pumarola et al., 2018a, Zhu et al., 2017]. For the generatator, a ResNet-based encoder-decoder is used to produce two outputs, i.e. an attention mask $\mathbf{A} = G_A(\mathbf{X}_S, \mathbf{Y}_T) \in [0, 1]$ together with the color transformation $\mathbf{C} = G_C(\mathbf{X}_S, \mathbf{Y}_T)$. The final prediction is obtained by interpolating $\mathbf{C}$ and

$\mathbf{X}_S$ using $\mathbf{A}$.

$$\hat{\mathbf{X}}_S = (1 - \mathbf{A}) \circ \mathbf{C} + \mathbf{A} \circ \mathbf{X}_S. \tag{4.2}$$

The attention mechanism enforces the network to directly learn a residual over the appearance image by taking features from the source image into consideration.

Discriminator

For the discriminator D, PatchGAN [Zhu et al., 2017] is used to predict real/fake for overlapping local patches of size 70×70. We use spectral normalization on convolution weights [Miyato et al., 2018] to stabilize the training. In contrast to previous methods, the generated image $\hat{\mathbf{X}}_S$ is conditioned not only on itself, but also on the target appearance $\mathbf{Y}_T$ and original face $\mathbf{X}_S$. Hence, our discriminator enforces that the generated examples are photo-realistic and that the results are consistent with respect to the expected properties of the target representation without the need for an auxiliary classification task as in StarGAN [Choi et al., 2018] or GANimation [Pumarola et al., 2018a].

4.3.2 Appearance Translation using 2D-to-3D Reconstruction and Image Formation

Given a 2D image of a human face, a 3D model is estimated. We parameterize a 3D model of M vertices using a semantic code vector $\mathbf{x} \in \mathbb{R}^{257}$.

$$\mathbf{x} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}, \mathbf{w}, \mathbf{t}\}, \tag{4.3}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^{80}$ corresponds to a neutral face shape, $\boldsymbol{\beta} \in \mathbb{R}^{80}$ denotes skin reflectance, $\boldsymbol{\gamma} \in \mathbb{R}^{27}$ relates to the lighting condition, $\mathbf{w} \in \mathbb{R}^3$ and $\mathbf{t} \in \mathbb{R}^3$ are the face rotation in $\mathbb{SO}3$ space and translation respectively. We follow the approach of Tewari et al. [2018] and use the AlexNet [Krizhevsky et al., 2012] backbone to produce the code vector $\mathbf{x}$. However, other methods for 3D face reconstruction from a monocular image can be used as well [Kim et al., 2018b, Tewari et al., 2018, Thies et al., 2016b]. Representing a 2D image as a compact latent vector allows us to efficiently manipulate different attributes of the appearance conditioning. A sample appearance translation is shown in Fig. 4.2.

Reflectance and Geometry

We constraint the facial geometry $\mathbf{S}(\boldsymbol{\alpha}, \boldsymbol{\delta}) \in \mathbb{R}^{M\times 3}$ and reflectance $\mathbf{R}(\boldsymbol{\beta}) \in \mathbb{R}^{M\times 3}$ by a multilinear PCA model using the Basel Face Model 2017 [Gerig et al., 2018]:

$$\begin{aligned} \mathbf{B}(\boldsymbol{\alpha}, \boldsymbol{\delta}) &= \boldsymbol{\mu}_{neut} + \boldsymbol{\alpha}\boldsymbol{\sigma}_{neut}\mathbf{E}_{neut} + \boldsymbol{\mu}_{exp} + \boldsymbol{\delta}\boldsymbol{\sigma}_{exp}\mathbf{E}_{exp} \\ \mathbf{L}(\boldsymbol{\beta}) &= \boldsymbol{\mu}_{ref} + \boldsymbol{\beta}\boldsymbol{\sigma}_{ref}\mathbf{E}_{ref}, \end{aligned} \tag{4.4}$$

where $\boldsymbol{\mu}_{neut}, \boldsymbol{\mu}_{exp}, \boldsymbol{\mu}_{ref} \in \mathbb{R}^{M\times 3}$ represent the mean neutral identity geometry, expression and skin reflectance respectively, $\mathbf{E}_{neut}, \mathbf{E}_{ref} \in \mathbb{R}^{M\times 3\times 80}$, $\mathbf{E}_{exp} \in \mathbb{R}^{M\times 3\times 64}$ correspond to linear bases of the PCA model together with their standard deviations $\boldsymbol{\sigma}_{neut}, \boldsymbol{\sigma}_{ref} \in \mathbb{R}^{80}$, $\boldsymbol{\sigma}_{exp} \in \mathbb{R}^{64}$. Expression (blend shape) basis covers more than 99% of the original PCA model.

Camera Model

We model the face transformation to the camera space using a rigid transformation consisting of rotation parameters $\mathbf{R}(\mathbf{w}) : \mathbb{R}^3 \rightarrow \mathbb{R}^{3\times 3}$ and translation $\mathbf{t}$ along x, y, z axes together with a full perspective transformation $\Pi : \mathbb{R}^3 \rightarrow \mathbb{R}^2$. Since the camera intrinsics of images in the wild are unknown, we fix the field of view and the principal point to be 0.5 and $\{\frac{W}{2}, \frac{H}{2}\}$ given an image of size $W \times H$ forcing the model to compensate the intrinsics by the pose parameters.

$$\mathbf{u}, \mathbf{v} = \Pi \circ (\mathbf{R}(\mathbf{w})\mathbf{B}(\boldsymbol{\alpha}, \boldsymbol{\delta}) + \mathbf{t}), \tag{4.5}$$

where $\mathbf{u}, \mathbf{v} \in \mathbb{R}^M$ are u, v coordinates of mesh vertices on the camera plane and $\circ$ is the Hadamard product.

Vertex Normals

Vertex normals $\mathbf{N} = \{\mathbf{n}_i\} \in \mathbb{R}^{M\times 3}$ are estimated using a 1-ring neighbour triangles. We are using a triangle topology provided by the morphable model.

Illumination Model

Shading $\mathbf{S} = \{s_i^{(j)}\} \in \mathbb{R}^{M\times 3}, j \in \{r, g, b\}$ is modelled using the first $B = 3$ bands of Spherical Harmonics [Ramamoorthi and Hanrahan, 2001] basis functions $\mathbf{H}_b : \mathbb{R}^3 \rightarrow \mathbb{R}$ separately per *RGB* image channels assuming the face to be a Lambertian surface with distant illumination while ignoring self-occlusion and cast shadows.

$$\mathbf{s}_i^{(j)} = \sum_{b=1}^{B^2} \gamma_b^{(j)} \mathbf{H}_b(\mathbf{n}_i). \tag{4.6}$$

The final per-vertex color intensity $\mathbf{I} \in \mathbb{R}^{M\times 3}$ is computed as an element-wise product between the shading $\mathbf{S}$ and reflectance $\mathbf{L}$ components.

Appearance Image Formation

Given a triangle mesh topology and u, v coordinates, we render the vertex attributes using z-buffer based rendering $\mathbf{F} : \mathbb{R}^{M\times 3} \rightarrow \mathbb{R}^{W\times H\times 3}$ to produce the face reconstruction $\bar{\mathbf{I}}$ from vertex-wise representations $\mathbf{I}$. Now, our physics-based image formation model converts a semantic code vector $\mathbf{x}$ estimated from a monocular image $\mathbf{X}_S$ into a reconstructed appearance image representation $\mathbf{Y}_S$. Different appearances can be obtained by changing the parameters of $\mathbf{x}$ (i.e. expression $\boldsymbol{\delta}$, light $\boldsymbol{\gamma}$ and pose $\mathbf{w}$).

4.3.3 Training Losses

Our training objective combines the adversarial loss $\mathscr{L}_{adv}$ with feature matching $\mathscr{L}_{FM}$, perceptual $\mathscr{L}_P$, reconstruction $\mathscr{L}_1$ and cycle consistency $\mathscr{L}_{cyc}$ losses. Parameters λ_{FM}, λ_P, λ_1, λ_{cyc} regulate their contributions.

$$\min_G \Big(\big(\max_D \mathscr{L}_{adv}(G, D) \big) + \lambda_1 \mathscr{L}_1(G) + \lambda_{FM} \mathscr{L}_{FM}(G) + \lambda_{cyc} \mathscr{L}_{cyc}(G) + \lambda_P \mathscr{L}_P(G) \Big). \tag{4.7}$$

Adversarial Loss

The adversarial loss enables to produce sharp photo-realistic results. The standard formulation of GAN's is trained by minimizing the Jensen-Shannon divergence [Goodfellow et al., 2014, Isola et al., 2017] between the real and generated image distributions. In our experiments, we are using a non-saturated version of the adversarial loss[I] which has

[I] We replace $\mathbb{E}_{\mathbf{X}_S \sim p(\mathbf{X}_S)}$ by $\mathbb{E}_\Delta$ in the formulas.

shown to provide comparable results to later modifications such as Wasserstein GAN and Geometric GAN [Kurach et al., 2018].

$$\mathscr{L}_{adv} = \mathbb{E}_{\Delta}\big[\log(1 - D(G(\mathbf{X}_S, \mathbf{Y}_T)))\big] + \mathbb{E}_{\Delta}\big[\log D(\mathbf{X}_T)\big], \tag{4.8}$$

where $\mathbf{X}_T$ is the target ground truth image.

Reconstruction Loss and Cycle Consistency

G is not able to produce plausible face images with the adversarial loss alone. Therefore, we force the generator output $\hat{\mathbf{X}}_\mathbf{S} = G(\mathbf{X}_S, \mathbf{Y}_T)$ to be close to the ground truth target image $\mathbf{X}_T$ using L1 norm.

$$\mathscr{L}_1 = \mathbb{E}_{\Delta}\big[|G(\mathbf{X}_S, \mathbf{Y}_T) - \mathbf{X}_T|\big]. \tag{4.9}$$

In addition, the generator is encouraged to produce the same original source image $\mathbf{X}_S$ given $\hat{\mathbf{X}}_\mathbf{S}$ and appearance $\mathbf{Y}_S$ of $\mathbf{X}_S$.

$$\mathscr{L}_{cyc} = \mathbb{E}_{\Delta}\big[|G(\hat{\mathbf{X}}_S, \mathbf{Y}_S) - \mathbf{X}_T|\big]. \tag{4.10}$$

Perceptual Loss

We further regularize training by the perceptual loss [Johnson et al., 2016] encouraging high level style features $F(\mathbf{X})$ of the pretrained on ImageNet Deep Neural Network (in our case AlexNet [Krizhevsky et al., 2012]) from $\hat{\mathbf{X}}_S$ to be similar to the one extracted from $\mathbf{X}_T$.

$$\mathscr{L}_P = \mathbb{E}_{\Delta}\,|F(\hat{\mathbf{X}}_S) - F(\mathbf{X}_T)|. \tag{4.11}$$

Mean Feature Matching Loss

Finally, we employ a form of feature matching loss [Wang et al., 2018] to penalize the difference between mean discriminator features given the generated image and mean discriminator features given the real image.

$$\mathscr{L}_{FM} = |\mathbb{E}_{\Delta}\big[\frac{1}{K} D(G(\mathbf{X}_S, \mathbf{Y}_T))\big] - \mathbb{E}_{\Delta}\big[\frac{1}{K} D(\mathbf{X}_T)\big]|. \tag{4.12}$$

$\mathscr{L}_{adv}$ in combination with $\mathscr{L}_1$, $\mathscr{L}_{FM}$, $\mathscr{L}_{cyc}$ and $\mathscr{L}_P$ provides the best results in our experiments.

4.4 Experimental Setup

4.4.1 Datasets

For training, our GAN pipeline uses a combination of YoutubeFaces [Wolf et al., 2011], Multi-PIE [Gross et al., 2010] and 300VW [Chrysos et al., 2018] datasets. In total, we collect about 500K images which are split into 3279 chunks. We make sure that different people are not in the same chunk. However, one person can occur in multiple chunks.

We crop the face boxes based on landmark boxes with 10% extension from its boundary. Ground truth landmarks are available in 300VW and face boxes are available in Youtube-Faces. When ground truth is not available, FAN [Bulat and Tzimiropoulos, 2017] is used when considering landmarks. The dlib CNN-based face detector [King, 2009] is used to detect face bounding boxes. Landmark information is not used during training, except for face box cropping.

In one epoch, for each image in the training set, we uniformly sample a target image from the same chunk. We increase the dataset variability by randomly flipping horizontally image pairs and randomly cropping the target image.

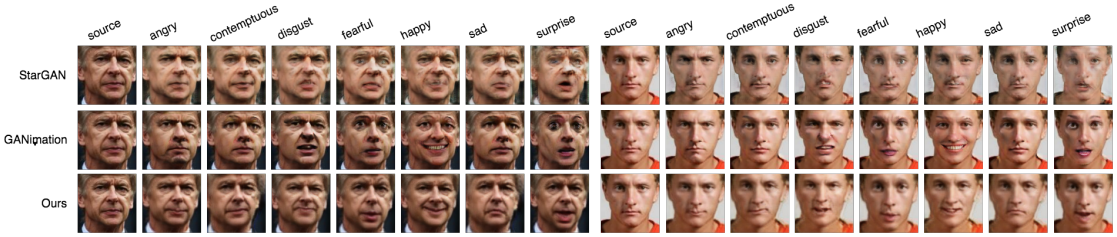


Figure 4.4: Qualitative comparison to state-of-the-art methods. We are performing cross-dataset comparison on CelebA [Liu et al., 2015] applying RaFD [Langner et al., 2010] expressions on a source image. It can be seen that for both StarGAN and GANimation, the best approaches in the literature for emotion manipulation, that they lose identity information, while our method preserves facial identity. Also, some of the results of StarGAN and GANimation show strong artifacts.

4.4.2 Training Setup

Our implementation is in Tensorflow. We train our model using Adam optimizer [Kingma and Ba, 2014] with $\beta_1 = 0.001$, $\beta_2 = 0.9$ and a learning rate 0.0001 for 40 epochs. The

learning rate is decayed linearly to 0 in the last 10 epochs. Input images have size of 128×128 and are fed to the network in batches of 32.

Our discriminator weights use spectral normalization [Miyato et al., 2018]. Generator weights are updated once per discriminator weight update. We use $\lambda_1 = 10$, $\lambda_{cyc} = 100$, $\lambda_p = 1$, $\lambda_{fm} = 1$ for the training loss.

4.5 Experiments and Results

4.5.1 Ablation Study

We conduct an ablation study to analyze the influence of each loss component. We used video frames from the test split of 300VW [Chrysos et al., 2018] consisting of 113K images. We randomly select a subset of 10K images. For each video frame in the validation split, we randomly sample a target image from the same sequence. The trained model is used to apply the appearance of the target image on the source. Average FID score [Heusel et al., 2017] and its standard deviation are reported in table 4.1 for different configurations. We train each model for 30 epochs. All configurations are trained using adversarial and reconstruction losses. Removing the reconstruction loss during the training leads to mode collapse. Results show that each loss component improves the quality of the final results. The most significant contribution is obtained by the perceptual loss (40 FID) and cycle consistency (86 FID). Feature matching loss contributes the third (24 FID).

$\mathscr{L}_{L1}$	$\mathscr{L}_{cyc}$	$\mathscr{L}_{FM}$	$\mathscr{L}_{P}$	FID Score ↓
✓	✓	✓	✓	**147.02 ± 6.65**
✓	✓		✓	171.29 ± 7.43
✓	✓	✓		211.44 ± 10.68
✓	✓			222.49 ± 9.46
✓				308.04 ± 8.01

Table 4.1: Ablation study performed on the 300VW test split. We report average and std FID scores for each configuration.

4.5.2 Expression Manipulation: Comparison to State-of-the-Art

In this experiment, we qualitatively compare our method against the state-of-the-art baseline methods, StarGAN [Choi et al., 2018], and GANimation [Pumarola et al., 2018a]. For a fair comparison, we compare discrete emotion categories which are represented in the RafD [Langner et al., 2010] dataset on the independent CelebA dataset [Liu et al., 2015], since StarGAN requires training on the annotated RafD dataset. We train both GANimation and StarGAN using the provided source code, default parameters on RafD and EmotioNet dataset respectively. For GANimation, we extract the target Action Unit's using OpenFace [Baltrusaitis et al., 2015]. For our method, we perform 2D-to-3D reconstruction to extract the desired target expression. Results are reported in Fig. 4.4. As can be observed in Fig. 4.4, the proposed method not only preserves source identity better but also do not have artifacts like other methods. More visuals can be found in the supplementary material.

4.5.3 Head Pose Manipulation

In this experiment, we qualitatively evaluate the effectiveness of the proposed method for changing the head pose of a given face image. We systematically (in each direction 30 degrees with an interval of 10 degrees) apply yaw rotation to the source images. We use the model trained on the 300VW dataset. Test images are used from Helen dataset. The results are shown in Fig. 4.5.

Fig. 4.5 clearly shows that the proposed method can robustly apply head pose manipulation to face images. Although most of the profile views are occluded for the source images, the proposed method realistically generates profile views. In addition, the proposed method uses scene illumination of the source while generating the target appearance. Therefore, generated faces have consistent and smooth illumination changes.

4.5.4 Light Direction Manipulation

In this experiment, we qualitatively evaluate the effectiveness of the proposed method for changing light directions for a given face image. We use the Multi-PIE [Gross et al., 2010] dataset for extracting illumination direction. CelebA [Liu et al., 2015] is used as the source image. For visualization, 3 dominant light directions are selected (i.e. left, frontal, and right). The extracted light directions are transferred to the target appearance

Figure 4.5: Head pose manipulation. Qualitative results on the Helen dataset. Images in the middle column are the source images. 4th, 5th, and 6th column images are 10°, 20°, and 30°rotated images respectively. 1st, 2nd, and 3rd column images are -30°, -20°, and -10°rotated images respectively.

which is used to condition the source image. Since the proposed method models light using the Lambertian assumption (SH based light prediction), shading is formed by the dominant light source direction. The results are shown in Fig. 4.6.

The results show that the light direction of the target robustly applied to source images. The side of the face, which is lit by the dominant scene illumination, has lighter image pixels whereas the other side has darker image pixels. It is clear that the topmost source image has a strong shading line on the right side of the face (due to directional scene light). After applying the target light direction, the strong shading line disappears. Instead, there is a smooth illumination change that appears on the face. This is due to the fact that the proposed method assumes Lambertian reflection.

4.5.5 Quantitative Comparison

We provide additional numerical comparisons of our proposed method to related GAN-based face reenactment methods, which can do both expressions and pose manipulation, using the 300VW test split (Table 4.2). 30 videos are selected from the Category 1 subset, which contains talking faces with variations in expressions and head poses. To obtain a

Method	Reconstruction Type	Identity Preservation ↑	Expression correctness ↓
Nirkin et al. [2019], ICCV'19	N/A	0.474	17.426
Siarohin et al. [2019], NeurIPS'19	N/A	**0.827**	24.493
Zakharov et al. [2020], ECCV'20	N/A	0.557	26.378
Ours	Tewari et al. [2017]	0.574	**8.594**
Ours	Deng et al. [2019]	0.671	13.228

Table 4.2: Quantitative comparison on the 300VW test split. We evaluated identity preservation and expression correctness for each method.

balanced dataset, we select 100 frames from each video (each 5th frame). Expression and head pose from selected videos are transferred to 3 randomly selected face identities from other videos. Identity frames are selected to be the most frontal face of the video based on the head pose prediction computed by a pretrained Hopenet [Ruiz et al., 2018]. In total, we evaluate our method on 9100 reenacted images coming from 90 video pairs.

Identity preservation is compared using cosine similarity from the latent space of VGG-Face2 features [Cao et al., 2018] between generated and target face images (higher is better). Expression correctness is compared using the mean L1 distance of facial landmarks (in pixels, image resized to 256) using a pretrained FAN detector [Bulat and Tzimiropoulos, 2017] between source and generated images (lower is better).

In terms of identity preservation, on our benchmark First Order Motion method [Siarohin et al., 2019] performs the best, followed by our method and Bi-layer [Zakharov et al., 2020]. First Order Motion [Siarohin et al., 2019] exploits dense optical flow and warping features capturing better the target image color distribution. Bi-layer [Zakharov et al., 2020] under-performs our method on identity preservation because it focuses on the face region. It produces better low-frequency details on the skin region but fails on the background and hair. FSGAN [Nirkin et al., 2019] is designed for face swapping and has limited capacity on the face reenactment task.

In terms of expression preservation, our method outperforms the baselines, followed by FSGAN and First Order Motion. First Order Motion fails on large head poses due to misalignments caused by optical flow features. Expressions generated by Bi-layer are dependent on the accuracy of the landmarks. Our method is more robust since it's conditioned on the full face representation.

Conditioning on Deep3DReconstruction provides better identity preservation. MoFA provides better expression preservation. Those observations correspond with the quality of the 2D-to-3D reconstruction results provided by reconstruction methods.

Figure 4.6: Light direction manipulation qualitative results on the CelebA dataset. The source images are represented in the left-most column. The target light direction is presented in the topmost right. The light direction is extracted from the target image using SH coefficients. Then, the extracted light direction is applied to the target image appearance.

Deep3DReconstruction uses additional a Perceptual loss for training which improves the quality of the identity. It also uses the full 3D face model for the reconstruction loss which may influence expression preservation in comparison to MoFA which uses a cropped 3D face region.

In conclusion, our method outperforms GAN-based face reenactment methods in terms of facial expression transfer errors (8.594 vs 17.426 for FSGAN) and provides comparable results on identity preservation.

4.6 Limitations

In spite of the promising results in comparison to state-of-the-art methods for face image manipulation, our method still has few limitations. The quality of generated images are dependent on the quality of 2D-to-3D reconstruction and consequently is prone to the

same errors of 2D-to-3D reconstruction algorithms such as occlusions and extreme poses (for which even face detection algorithm may fail), and dependency on the quality of face tracker (Fig. 4.7). However such problematic cases are also a problem for FACS extraction. In addition, due to a limitation of the morphable face model, expressiveness of expression basis are constrained by the PCA basis.

Figure 4.7: Failure cases of our method. (1) - heavily occluded faces, (2) - extreme facial expressions, (3) - occlusions not modeled by 2D-to-3D fitting and attention map, (4) and (5) - extreme head poses.

4.7 Conclusion

This chapter proposes a self-supervised method to manipulate a single monocular face image by conditioning GAN on face decomposition using appearance transfer. Our conditioning has shown to be a more flexible representation in comparison to previous GAN-based methods that use discrete classes, landmarks, or action units. Thus, conditioning on the appearance image allows us to manipulate head pose, scene illumination, and facial expression using a single conditioning space.

The qualitative results on 300VW dataset show that the proposed method is outperforming GAN-based state-of-the-art face reenactment methods, which can do both expression and head pose manipulation, in terms of expression correctness, and provides competitive results in term of identity preservation. We show that our method is agnostic to face decomposition methods and works with any 2D-to-3D reconstruction method which allows pose, expression, and light manipulation.

Chapter 5

Face Reenactment and Swapping

Face reenactment and face swap have gained a lot of attention due to their broad range of applications in computer vision. Although both tasks share similar objectives (e.g. manipulating expression and pose), existing methods do not explore the benefits of combining these two tasks.

In this chapter, we introduce a unified end-to-end pipeline for face swapping and reenactment. We propose a novel approach for isolated disentangled representation learning of specific visual attributes in an unsupervised manner. A combination of the proposed training losses allows us to synthesize results in a one-shot manner. The proposed method does not require subject-specific training.

We compare our method against state-of-the-art methods for multiple public datasets of different complexities. The proposed method outperforms other SOTA methods in terms of realistic-looking face images.

5.1 Introduction

Generating images or videos by manipulating facial attributes (i.e. face reenactment and swapping) has gained a lot of attention in recent years due to their broad range of computer vision and multimedia applications such as video dubbing [Suwajanakorn

Published in *Asian Conference on Computer Vision, 2020* [Ngo et al., 2020]

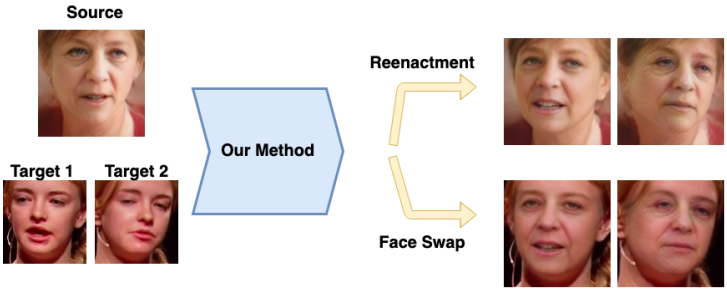


Figure 5.1: Our algorithm takes source and target images and produces reenacted and swapped face results using a single unified pipeline.

et al., 2017], gaze correction [Kuster et al., 2012], actor capturing [Kim et al., 2018a, Thies et al., 2016a], and virtual avatar creation [Nagano et al., 2018].

Face reenactment [Siarohin et al., 2019, Thies et al., 2016a] aims to manipulate facial attributes such as expression, pose or gaze of a video or a single image, whereas face swap [Nirkin et al., 2019, 2018] tries to seamlessly replace a face from a source image with a target face while maintaining the realism of the facial appearance. To perform such transfer, face swap techniques manipulate face attributes such as expression, pose, and identity. Although the face attribute manipulation for both face reenactment and face swap is similar, they have never been considered in a unified pipeline. To this end, in this work, we propose a single unified model for both face swapping and reenactment tasks allowing the model to produce a more robust face representation and exploiting the constraints from the two tasks to improve the realism of facial appearances.

Before the introduction of deep neural networks, face reenactment and swapping are typically solved by 3D modeling [Dale et al., 2011, Garrido et al., 2014, Nirkin et al., 2018, Thies et al., 2015, 2016a]. The 3D face image is transformed into a 3D representation, where latent parameters of the 3D representation are manipulated and projected back in a 2D space. Although those methods produce results with high realism, they are not able to generalize well on unseen data. Hence, for each target face the model parameters have to be tuned.

Current generative models make it feasible to synthesize realistic-looking images [Choi et al., 2018, Karras et al., 2019a]. Consequently, recent research is focused on improving the quality of the *face image generation* process [Choi et al., 2018, Karras et al., 2019a, Pumarola et al., 2018a] using generative models. Only a few methods explore the direction of using generative models for face reenactment or face swapping. Although these tasks share similarities, previous methods only focus on solving one of the two tasks independently and are supervised [Korshunova et al., 2016, Li et al., 2019, Wu

et al., 2018a]. Recently methods show that face swap targeted methods can be used for face reenactment and vice versa. Unfortunately, the visual results on the second task are typically inferior to the first one [Nirkin et al., 2019, Siarohin et al., 2019]. Since those methods are designed for one of the tasks separately, they are not optimal for both. In contrast to existing methods, we integrate both tasks into one combined model. To our knowledge, our method is the first unsupervised method designed to perform both tasks in a unified end-to-end manner.

In this work, we propose a novel pipeline that unifies face swapping and reenactment (Fig. 5.1). A combined approach benefits from the similarities of the two tasks. Learning them together allows for robust face representation and enhances the realism of facial appearance. The proposed algorithm learns an isolated disentangled representation for face attributes without any supervision. Hence, our model can manipulate expression/-pose, face identity, and style independently in latent space. We achieve this by directly mapping the disentangled latent representation to the latent space of a pre-trained generator. During inference time, the encoders condition the latent space by source and target face images together with their landmarks and generate the reenacted or swapped face using the pre-trained decoder. Prediction is done in a one-shot manner (i.e. only a single image of a person is required). The model's training loss incorporates contextual and identity losses to preserve the face identity, regardless of the source face. As a result, our model obtains visually more appealing results in cross-gender face swapping compared to the baselines.

We evaluate our method on multiple datasets of various complexities: 300VW with videos of talking people [Chrysos et al., 2015], and UvA-NEMO with spontaneous and fake smiles in a controlled environment [Dibeklioğlu et al., 2012]. Experiments demonstrate that our method (on average) performs better on face reenactment and face swapping tasks than existing state-of-the-art methods focusing only on a single task.

To summarize, our contribution is four-fold:

- A novel method is proposed to perform face swapping and reenactment tasks in a joint manner. To our knowledge, our method is the first method to jointly perform the two tasks in a unified end-to-end architecture.
- The proposed method is *subject agnostic*: it does not require subject-specific training.
- A novel approach is proposed to learn an isolated disentangled representation for single visual attributes (i.e. the expression/pose, identity, and style) by using

a pre-trained generator with a disentangled latent space. This allows for a full control over the face manipulation process in an unsupervised manner. Hence, our approach does not require ground truth data for expression/pose, identity, and style learning of reenactment outputs.

- A combination of training losses allows us to synthesize results in a one-shot manner and to outperform competitive methods in cross-gender face manipulation.

5.2 Related Works

5.2.1 Generative Models

Generative models based on Generative Adversarial Networks (GANs) are advantageous for the task of image synthesis [Nguyen et al., 2016, Radford et al., 2015]. However, until recently, those models can be considered as black boxes with latent representations which are hard to interpret. In addition, the realism of the generated results, in particular for face image synthesis [Choi et al., 2018, Pumarola et al., 2018a], is limited (with artifacts in identity preservation).

Recently, StyleGAN [Karras et al., 2019a] introduces a novel way to condition the latent code through an affine transformation, corresponding to a specific style [Karras et al., 2019b], by using Adaptive Instance Normalization (AdaIN) [Huang and Belongie, 2017]. AdaIN allows the model to generate images with more realistic face appearance compared to previous methods [Karras et al., 2017]. Furthermore, the aforementioned architecture modifications, combined with a revised training approach [Karras et al., 2019a,b], enable the separation of high-level and stochastic attributes making the latent representation easier to interpret. Hence, the face attributes of a generated image can be changed accordingly by manipulating the latent representation (i.e. disentanglement property). Recent methods integrate StyleGAN into different applications as a pre-trained network for face enhancement and animation [Gabbay and Hoshen, 2019]. The state-of-the-art StyleGAN2 [Karras et al., 2019b] enhances the architecture of StyleGAN by redesigning normalization flow and by applying the same network topology for low and high resolution. Image2StyleGAN [Abdal et al., 2019] proposes a method to map an existing image to the latent representation of StyleGAN by iteratively optimizing a latent code to minimize the loss function. Mapping an image to latent space enables a user to change specific image attributes provided by the StyleGAN latent space.

However, this method has a drawback in terms of efficiency and generalization: each new image is optimized separately until convergence to obtain a corresponding latent space limiting the applicability of the method for real-time applications. In contrast, our novel isolated disentangled representation learning method solves this problem by introducing encoders that learn to map the desired facial attributes to the corresponding changes in the latent representation. By constraining the mapping by encoders and by using a specific unsupervised training procedure, our approach manipulates the latent space in such a way that it is able to mix disentangled expression/pose, identity and style attributes in a robust manner.

5.2.2 Face Reenactment

Face reenactment focuses on changing attributes of the face image while keeping the face identity the same. Prior methods focus on different facial attributes like expression [Choi et al., 2018, Pumarola et al., 2018a], skin color [Choi et al., 2018], lighting [Zhou et al., 2019] and pose, or a combination of those [Thies et al., 2016a]. These methods are mostly used in applications such as virtual avatar or puppeteering, targeting high realistic-looking faces but ignoring background preservation [Nagano et al., 2018]. Other approaches focus more on video dubbing and deepfake generation, preserving the realism for both the foreground and background of the scene [Kim et al., 2018a, Pumarola et al., 2018a, Thies et al., 2016a]. Attribute conditioning is modeled by using different modalities like facial landmarks [Sanchez and Valstar, 2018], action units [Pumarola et al., 2018a] and 3D morphable models [Kim et al., 2018a] for pose and/or expression, and spherical harmonics for lighting [Zhou et al., 2019]. Some methods simplify attribute inference by conditioning directly on the face image. In contrast, our method uses a face image to condition identity and style together with facial landmarks for pose and expression.

Several methods perform face reenactment by manipulating the latent space [Abdal et al., 2019, Fu et al., 2019, Shen et al., 2020]. Abdal et al. [2019] compute the relative difference between two images by calculating the differences in their latent spaces and applied to the source latent code afterwards. Fu et al. [2019] propose an approach capable of reenacting faces by using encoders to compute the representations of pose, expression, style and identity in the latent space.

In combination with the use of a pre-trained generator, we aim to condition the generator in a one-shot manner during both training and testing time. InterFaceGAN [Shen et al.,

2020] on the other hand, does use a pre-trained generator, but computes latent codes based on attribute scores (e.g. smile, glasses, gender etc.) making it a supervised method. Since our approach does not require ground-truth labels or attribute scores, we have full control over the face manipulation process using only a minimal amount of training data.

Methods that focus on the quality of images and identity/background preservation are typically target-specific [Kim et al., 2018a]. Hence, the model is trained for a particular scene with a single face identity. Other non-target, one-shot methods [Choi et al., 2018, Pumarola et al., 2018a] produce decent results, but they fail in producing consistent face identities between images of the same person (video sequence) [Choi et al., 2018, Siarohin et al., 2019]. Our method is also a one-shot method. However, in contrast to previous methods, the aim of our method is to produce realistic-looking faces with identity-preservation by exploiting the disentanglement property of the pre-trained model.

5.2.3 Face Swapping

Face swapping aims to change the facial identity but to keep other face attributes constant. Applications range from face identity obfuscation [Bitouk et al., 2008] to recreation [Kemelmacher-Shlizerman, 2016] and entertainment [Nirkin et al., 2018]. Recent methods obtain realistic results by using GANs [Korshunova et al., 2016, Li et al., 2019, Nirkin et al., 2019] conditioning identity attributes using either a face image or its facial landmarks. Besides, face segmentation is usually required to position a generated face on the original face [Nirkin et al., 2019, 2018].

Most face reenactment and swapping approaches rely on the use of generative adversarial networks [Abdal et al., 2019, Choi et al., 2018, Fu et al., 2019, Pumarola et al., 2018a, Zakharov et al., 2019, Zhu et al., 2017]. A major drawback of the aforementioned methods is their training process, the interpolation quality and lack of disentanglement.

Despite the similarity between face reenactment and face swapping tasks, there are no methods, to the best of our knowledge, which successfully unify these tasks. Siarohin et al. [2019] mainly focuses on the problem of face reenactment, but shows inferior results on the task of face swapping. Nirkin et al. [2019] shows the opposite. This approach is mainly focused on face swapping, but the results on face reenactment lack realistic-looking appearance. Moreover, those methods are complex and multi-staged. Thus, Nirkin et al. [2018] proposes four separate GANs for reenactment, segmentation, inpainting and blending. Siarohin et al. [2019] uses a separate motion network to extract dense optical flow and requires an extra segmentation network for face swapping. In

contrast, to our knowledge, we are the first method to unify face reenactment and face swap in one single unified pipeline.

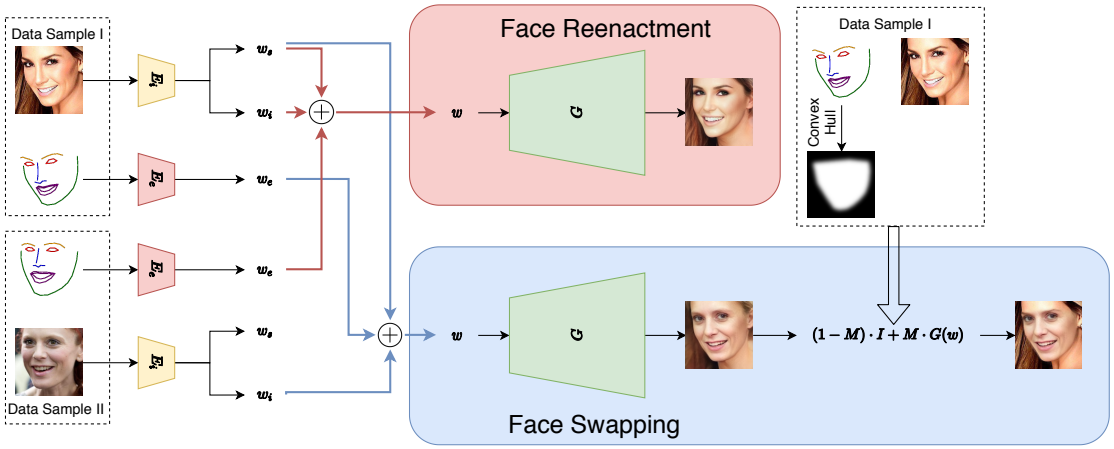


Figure 5.2: Our architecture combines Face Swapping and Reenactment into a single unified pipeline with the help of our novel isolated disentangled representation learning algorithm.

5.3 Proposed Method

An overview of our method is shown in Fig. 5.2. Our goal is to produce a face image $\hat{\mathbf{x}}$ while predicting identity attributes w_i, style attributes w_s and pose/facial expression attributes w_{pe} from a given face image and its landmarks. We propose a novel isolated disentangled representation learning algorithm to separate w_i, w_s and w_{pe}. Using the proposed algorithm, attributes of the source and target images can be manipulated in the latent space via mixing using linear addition, since changing one attribute doesn't influence another due to their isolation. For the `face swapping` task, w_i is taken from the source image, while other attributes are taken from the target image. For the `face reenactment` task, w_{pe} and w_s are taken from the source image, keeping identity w_i from the target.

5.3.1 Disentanglement Property and Vector Computations

Our encoders are trained to compute a latent code in the latent space $w \in \mathcal{W}^+$ of a pre-trained generator. Since the latent space is disentangled, face attributes can be

manipulated by using vector arithmetics in $\mathcal{W}^+$ [Abdal et al., 2019]. For example, given an image N_A and its latent code w_1 (person A with a neutral expression), N_B and its latent code w_2 (person B with a neutral expression) and another image S_B with a latent code w_3 (person B smiling), it's possible to generate an image of a person A smiling by conditioning the generator G on a latent code $G\big(w_1 + (w_3 - w_2)\big)$.

Our method uses that principle by predicting isolated latent codes for style w_s, identity w_i and pose/expression w_{pe} based on the input image and its corresponding landmarks, assuming those latent codes to be with a disentanglement property. Final latent code can be constructed via linear addition of the three isolated components:

$$w = \mu_G + w_i + w_s + w_{pe}, \tag{5.1}$$

where μ_G is the mean of the generator's latent space $\mathcal{W}^+$ with disentanglement property [Abdal et al., 2019, Karras et al., 2019b].

Since w is constructed from the latent codes w_s, w_i, and w_{pe}, full control is obtained for changing the style, identity, pose and expression of the resulting image I, by exploiting the high-quality images produced by the pre-trained generator. Note that our method allows for subject-agnostic face manipulation executed in a one-shot fashion during inference.

5.3.2 Architecture

The source face and the target face (together with its facial landmarks) are used as inputs to the two separate encoders E_i and E_{pe} respectively. These encoders approximate a latent code for face style w_s, identity w_i and pose/expression w_{pe}. The network latent space is manipulated using encoder outputs to obtain either face swap by swapping the identity latent code or face reenactment by swapping the pose latent code. All latent codes are combined into the final latent code w. Then, w is fed to a decoder G to produce the final visual result. In the case of face swapping, a face mask M is generated by using the convex hull of the landmarks [Yang and Lim, 2019].

Encoders

Our architecture contains two different types of encoders: (1) the identity encoder E_i, and (2) the pose encoder E_{pe}. These encoders predict a latent code $w \in \mathcal{W}^+$ corresponding to either the identity, style, or pose of the input image.

For the design of the architecture, we base our encoders on the encoder of Pix2Pix [Isola et al., 2017]. To map the input images and landmarks to their corresponding latent codes, we add n separate fully connected blocks to the architecture, where n is the first dimension of the extended latent space. This fully connected blocks consist of 2 fully connected layers. E_i contains 2 of these fully connected block sets, for style (w_s) and identity (w_i) respectively.

Identity encoder $E_i(\mathbf{x})$ takes an input image $\mathbf{x}$ and estimates the identity latent code $w_i \in \mathcal{W}^+$ and style latent code $w_s \in \mathcal{W}^+$. Latent code w_i is trained to contain only pose- and expression-invariant identity features of the person.

Pose encoder $E_{pe}(\mathbf{x}_s)$ uses the facial landmarks of $\mathbf{x}$ denoted by $\mathbf{x}_s$ as an input. $E_{pe}(\mathbf{x}_s)$ predicts a latent code $w_{pe} \in \mathcal{W}^+$ containing both the pose and expression of $\mathbf{x}_s$. The landmarks are represented as *RGB* images of landmark boundaries [Zakharov et al., 2019].

$$w_i^{\mathbf{x}}, w_s^{\mathbf{x}} = E_i(\mathbf{x}), \quad w_{pe}^{\mathbf{x}} = E_{pe}(\mathbf{x}_s). \tag{5.2}$$

Decoder

Generator $G(w)$ is a pre-trained network with fixed weights. It takes a latent code $w \in \mathcal{W}^+$ as an input. Here $\mathcal{W}^+$ is the latent space of $G(w)$. $G(w)$ generates an image $\hat{\mathbf{x}}$ corresponding to latent code w. In this work, the StyleGANv2 architecture is used. However, other models with similar disentanglement properties and continuous latent spaces can be used instead.

5.3.3 Face Reenactment and Swapping

The reconstructed original face is defined as a function $G(w)$ over its identity $w_i^{\mathbf{x}}$, style $w_s^{\mathbf{x}}$ and expression/pose $w_{pe}^{\mathbf{x}}$ parameters:

$$\hat{\mathbf{x}} = G(\mu_G + w_i^{\mathbf{x}} + w_s^{\mathbf{x}} + w_{pe}^{\mathbf{x}}). \tag{5.3}$$

Face Reenactment

Faces are reenacted by changing the expression and pose parameters $w_{pe}^{\mathbf{y}}$ to the pose/-expression shown in the target image $\mathbf{y}$ and keeping other parameters identical $w_i^{\mathbf{x}}$ and

$w_s^{\mathbf{x}}$. Since $w_i^{\mathbf{x}}$, $w_s^{\mathbf{x}}$ and $w_{pe}^{\mathbf{x}}$ parameters are separated, the resulting image $\dot{\mathbf{x}}$ is defined as a function of their sum:

$$\dot{\mathbf{x}} = G(\mu_G + w_i^{\mathbf{x}} + w_s^{\mathbf{x}} + w_{pe}^{\mathbf{y}}). \tag{5.4}$$

Face Swapping

Face swapping is performed by keeping the $w_s^{\mathbf{x}}$ and $w_l^{\mathbf{x}}$ parameters unchanged and to modify the identity latent code to $w_i^{\mathbf{y}}$:

$$\tilde{\mathbf{x}} = G(\mu_G + w_i^{\mathbf{y}} + w_s^{\mathbf{x}} + w_{pe}^{\mathbf{x}}). \tag{5.5}$$

To swap faces, a facial mask $\mathbf{M}$ is obtained by computing a convex hull of the landmarks and to add a Gaussian blur [Yang and Lim, 2019]. The final swapped face is generated by interpolation:

$$(1 - \mathbf{M}) \cdot \mathbf{x} + \mathbf{M} \cdot \tilde{\mathbf{x}}. \tag{5.6}$$

5.3.4 Losses

The objective function, to train our unified face swapping/reenactment architecture, consists of 5 terms: reconstruction loss $\mathscr{L}_{MSE}$, perceptual loss $\mathscr{L}_{per}$, landmark loss $\mathscr{L}_L$, identity losses for the aligned reconstructed image $\mathscr{L}_{id}^a$ and for the unaligned identity-swapped/reenacted image $\mathscr{L}_{id}^u$. Those terms are weighted using hyperparameters λ_i, $\mathrm{i} \in \{1..5\}$.

$$\mathscr{L} = \lambda_1 \mathscr{L}_{MSE} + \lambda_2 \mathscr{L}_{per} + \lambda_3 \mathscr{L}_L + \lambda_4 \mathscr{L}_{id}^a + \lambda_5 \mathscr{L}_{id}^u. \tag{5.7}$$

Reconstruction and Perceptual Losses

We compute the mean squared error between input and predicted images as a reconstruction loss for efficient color embedding. $\mathscr{L}_{MSE}$ is calculated for the reconstructed image $\hat{\mathbf{x}}$ and the identity-swapped image $\tilde{\mathbf{x}}$. This loss function mainly helps to isolate w_s ensuring a proper color embedding. To capture finer features, the LPIPS distance is

used [Banerjee et al., 2018, Karras et al., 2019b, Zhang et al., 2018]. L_{per} is taken as the reconstruction loss and is calculated only for the reconstructed image $\hat{\mathbf{x}}$.

$$\mathscr{L}_{MSE} = \| \hat{\mathbf{x}} - \mathbf{x} \|_2^2 + \| \tilde{\mathbf{x}} - \mathbf{x} \|_2^2, \tag{5.8}$$

$$\mathscr{L}_{per} = \text{LPIPS}(\hat{\mathbf{x}}, \mathbf{x}). \tag{5.9}$$

Landmark Loss

The landmark loss term is used to isolate pose and expression from identity and style. A pre-trained facial landmark extraction network ψ [Bulat and Tzimiropoulos, 2017] is taken to extract the landmark heatmaps from an image $\mathbf{x}$. The loss function attempts to minimize the L_2 distance between the extracted heatmaps of the facial landmarks of the source image $\mathbf{x}$ and the target image $\mathbf{y}$, while keeping the latent code for identity and style identical. Landmarks do contain identity (e.g. eye and mouth shape). This means that landmark loss adds an identity bias to the resulting image.

We separate the heatmap sets into two different sets, the expression landmarks ψ_E and the jaw landmarks ψ_J. Parameters γ_1, γ_2 adjust the importance of these landmark sets respectively.

$$\mathscr{L}_L = \gamma_1 \| \psi(\dot{\mathbf{x}})_E - \psi(\mathbf{y})_E \|_2^2 + \gamma_2 \| \psi(\dot{\mathbf{x}})_J - \psi(\mathbf{y})_J \|_2^2 . \tag{5.10}$$

Identity Loss

The identity loss [Fu et al., 2019, Hu et al., 2018] isolates identity in a separate latent code w_i. The layer activations are used of a pre-trained identity recognition network Φ [Wu et al., 2018b]. For our purpose, we use activations $l \in L$ of two specific convolution layers and the last two fully connected layers.

The identity loss is applied to the convolution layers by calculating the contextual loss [Mechrez et al., 2018] $\mathscr{L}_{id}^a$ over these layers. Note that this will only work for images with the same pose (the reconstructed image), since the convolutions do not capture rotations properly.

$$\mathscr{L}_{id}^a = \sum_{l \in L} \| \text{CX}(\Phi(\hat{\mathbf{x}}), \phi(\mathbf{x}) \|_2^2 . \tag{5.11}$$

To ensure correct identity in the reenacted frames, a loss function is required to detect the identity of a face independent of the pose. A mean squared error is calculated for

the activations of the fully connected layers of Φ. During training, faces are reenacted with random landmarks from the dataset making our approach more robust to landmark biases.

$$\mathcal{L}_{id}^{u} = \sum_{l \in L} \left(\| \Phi(\dot{\mathbf{x}}, l) - \Phi(\mathbf{x}, l) \|_2^2 + \| \Phi(\tilde{\mathbf{x}}, l) - \Phi(\mathbf{y}, l) \|_2^2 \right). \tag{5.12}$$

5.3.5 Training Details

We trained both our method and the pre-trained generator on the subset of 183K images from the CelebA face dataset [Liu et al., 2015]. Faces are detected using the Dlib [King, 2009]. Face bounding boxes are computed based on an expanded by 10% bounding boxes over facial landmarks [Bulat and Tzimiropoulos, 2017] and resized to 128×128. Parameters of the network were optimized using the Adam optimizer with a learning rate of 10^{-5} for 100 epochs, batch size = 4. In our experimental setup, we used $\lambda_1, \lambda_2 = 5$, $\lambda_3 = 1$, $\lambda_4, \lambda_5 = 0.05$, $\gamma_1 = 1$ and $\gamma_2 = 50$, since it yielded the best results.

We use StyleGANv2 in our experiments. For StyleGANv2 latent code manipulation, we use the extended latent space $w \in \mathcal{W}^+$, which predicts a different latent code for every level of a pre-trained generator. Using $\mathcal{W}^+$ allows for a better embedding of an image, but is also possible to cope with images that do not have a latent embedding.

5.4 Experiments and Results

In this section, we evaluate the qualitative and quantitative performance of our proposed method and compare it to the state-of-the-art. We perform an ablation study to analyze the influence of the loss components in section 5.4.1. Results on latent space interpolation are discussed in section 5.4.2. Comparison to state-of-the-art in face swap and reenactment are provided in section 5.4.3. For all experiments, a cross-dataset evaluation is conducted for our method and baselines.

5.4.1 Ablation Study

An ablation study is conducted for the loss components to assess their influence on the face swapping and reenactment tasks on the 300VW dataset [Chrysos et al., 2015]. This dataset contains 114 high-quality videos of talking people. The dataset is preprocessed

Figure 5.3: Ablation Study. Face swap and reenactment results of our method trained with different loss configurations. Our full model results are shown in the last row.

	Face reenactment				Face swap			
Metric	C1	C2	C3	C4	C1	C2	C3	C4
(a)	2.4 ±0.08	2.02 ±0.08	1.96 ±0.1	**2.69 ±0.16**	**2.68 ±0.13**	2.59 ±0.12	2.63 ±0.12	2.56 ±0.14
(b)	1.57	1.57	1.48	**1.46**	1.54	1.50	1.51	**1.48**
(c)	7.12 ±0.22	7.47 ±0.21	6.84 ±0.2	**5.31 ±0.23**	5.44 ±0.21	5.01 ±0.21	5.27 ±0.2	**4.4 ±0.21**
(d)	N/A	N/A	N/A	N/A	**0.42**	0.49	0.52	0.51
(e)	N/A	N/A	N/A	N/A	1.84 ±0.18	1.78 ±0.16	2.08 ±0.16	**1.45 ±0.15**

Table 5.1: Quantitative ablation study evaluation on 300VW dataset. Reported metrics are (a) Inception Score, (b) FID source vs generated, (c) KID source vs generated, (d) FID target vs generated and (e) KID target vs generated.

by cropping faces based on the given (ground truth) landmark bounding boxes with 10% extension to each direction.

The qualitative results of our method trained with 4 different loss configurations are shown in Fig. 5.3:

C1 - $\mathscr{L}$ without contextual loss $\mathscr{L}_{id}^{a}$ and identity loss $\mathscr{L}_{id}^{u}$;

C2 - $\mathscr{L}$ without $\mathscr{L}_{id}^{u}$;

C3 - $\mathscr{L}$ without $\mathscr{L}_{id}^{a}$;

C4 - our final model with $\mathscr{L}$.

Configurations with other losses being disabled produce significantly degenerated visual results. Consequently, they are crucial for our method.

Contextual loss $\mathscr{L}_{id}^{a}$ supports identity preservation of the source image both in reenactment and face swapping tasks (C2 vs C1). However, it has difficulty with the pose and expression preservation of the target image. Thus, expression and pose are influenced by the content of the source face.

Identity loss $\mathscr{L}_{id}^{u}$ is beneficial for expression/pose isolation and visual sharpness. However, it has difficulties in identity preservation (C3 vs C1). Besides, for the face reenactment task, the reenacted shape of the source person is morphed by target images. It can be seen that the source rounded face becomes oval (C3: Face Reenactment, columns 1, 2). A trade-off result is obtained by combining $\mathscr{L}_{id}^{a}$ and $\mathscr{L}_{id}^{u}$ together (C4 vs C1).

For quantitative evaluation, different metrics are computed which are commonly used in image synthesis evaluation and shown in Table 5.1. Inception Score [Salimans et al., 2016] uses pre-trained on ImageNet Inception Network to compute the KL divergence between conditional and marginal label distributions over generated data (higher - better). Frechet-Inception distance [Heusel et al., 2017] computes Wasserstein-2 distance between distributions of real and generated samples in the Inception Net feature space (lower - better). Kernel-Inception distance [Bińkowski et al., 2018] measures dissimilarity between distributions of real and generated samples (lower - better).

Since the generated results of our method are unaligned in term of face attributes, FID and KID metrics are used only as an indicator of how our face identity is similar to the real data distribution. In case of face reenactment, the identity should be as close as possible to source face image. In case of face swap, we want a generated face to capture both properties of source and target image. Consequently, for face swap generated images, the FID and KID metrics are reported both in comparison with the source and target image data distributions. Source and target subjects are randomly selected from the 300VW dataset. Evaluation is performed on a sample of 10K generated images.

In the task of face reenactment, the evaluation metrics support our qualitative experimental results: our method with combined contextual and identity losses generates visual results with identity closer to the source face image distribution (C1 vs C2, 1.57 vs 1.46 FID). In the case of face swap, it can be observed that the distribution of generated images is closer to the distribution of target face images (C4 1.48 vs 0.51 FID). With the introduction of the additional regularization into our model, visual results start to capture more and more properties from the source image (C1 vs C4, 1.54 vs 1.48 FID).

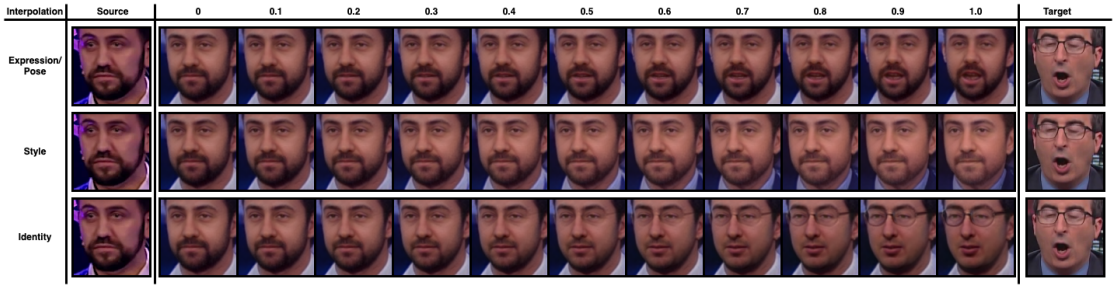


Figure 5.4: Interpolation of the latent space. Row 1: expression and pose interpolation. Row 2: Style interpolation. Row 3: identity interpolation. The last column represents a target expression/pose, style, or identity respectively. The results show that our novel disentangled representation learning algorithm can robustly isolates face attributes so that we can manipulate each attribute independently.

5.4.2 Latent Space Interpolation

In this section, our method is analyzed to interpolate over different face attribute dimensions. Given a source image, its face attributes are gradually changed where expression/-pose, style or identity are modified to become closer to the target face image. Qualitative results on the 300VW dataset are shown in Fig. 5.4. The 300VW dataset is preprocessed in the same way as described in the section 5.4.1. The first column shows the source image. Our algorithm changes gradually an attribute dimension to become closer to the target image of the last column.

Given a source w_1 and target attribute w_2, our model generates meaningful face images conditioned on the interpolated latent code $\alpha w_1 + (1-\alpha)w_2$. Note that in the case of style, a costume of John Oliver gradually starts to appear, while in the case of identity, we can observe the disappearance of beard, an emergence of his glasses and eyebrows. Despite the challenges given by cross-dataset evaluation, our model preserves attributes dimensions on challenging cases with face accessory and occlusion. Image2StyleGAN [Abdal et al., 2019] show the capability to map face attributes into the latent space of StyleGAN. However, the latent space of expression/pose, identity and style are not fully disentangled. For example, it's not possible to manipulate expression/pose property separately without influencing identity or style. In contrast, our mapping to the latent space provides more flexibility.

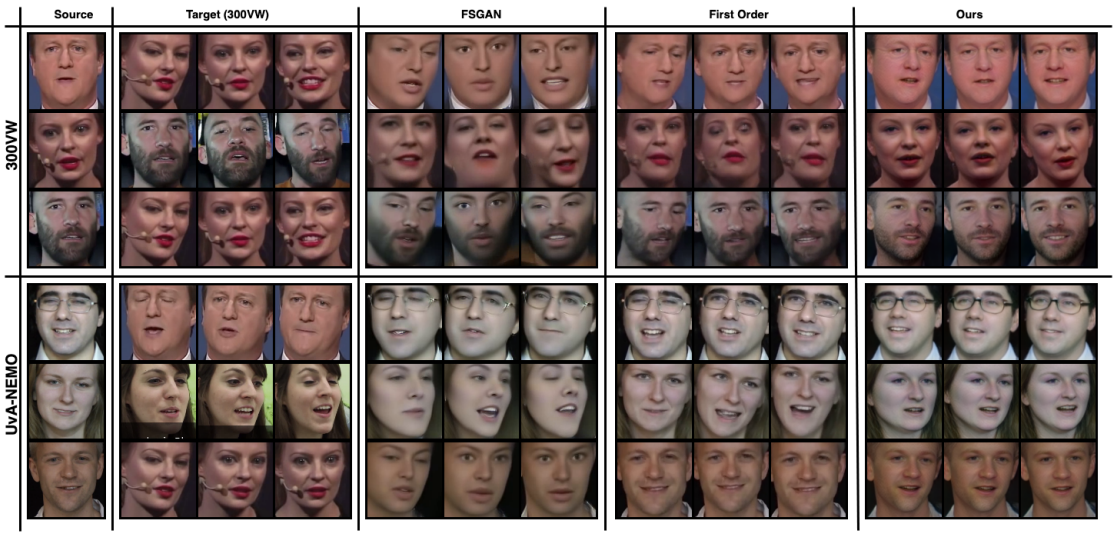


Figure 5.5: Qualitative comparison of face reenactment results on 300VW and UvA-NEMO datasets. Pose and expression from target images (second column) are applied on the source image (first column). Faces are produced by the baseline methods, FSGAN and First Order Motion Model, and predictions provided by our novel unified pipeline.

5.4.3 Face Swap and Reenactment State-of-the-Art Comparison

Qualitative Evaluation

We evaluate qualitatively our method on the face swapping and face reenactment tasks. We perform cross-dataset evaluation of our method with results produced by FSGAN [Nirkin et al., 2019] and First Order Motion Model [Siarohin et al., 2019] on the 300VW [Chrysos et al., 2015] and UvA-NEMO [Dibeklioğlu et al., 2012] datasets. These methods are selected because they are state-of-the-art which can do both face swap and reenactment. For our purpose, the available pre-trained model is used provided by authors of FSGAN and First Order Motion. For fairness of comparison, we use models trained on a different dataset from UvA-NEMO and 300VW. The datasets are preprocessed by cropping faces based on landmark bounding boxes with 10% extension to each direction. For 300VW, the provided ground truth landmarks are used. For UvA-NEMO, the landmarks are extracted by using FAN [Bulat and Tzimiropoulos, 2017].

In the first experiment, we qualitatively compare our method with the state-of-the-art for the face reenactment task. The visual comparison is shown in the Fig. 5.5. For the First Order Motion model, its pre-trained model is used with *absolute motion* for both face reenactment and face swap experiments, since only the absolute motion mode is

Figure 5.6: Qualitative comparison of face swapping results on the 300VW and UvA-NEMO datasets. First column: source image from which identity properties are taken. Second column: target images, on which those properties are applied. Faces produced by the baseline methods, FSGAN [Nirkin et al., 2019] and First Order Motion Model [Siarohin et al., 2019], and predictions provided by our novel unified pipeline.

capable of computing face swaps. Our method shows comparable quality of reenactment results to First Order Motion and outperforms FSGAN in terms of identity preservation. Besides, since our latent space is constrained by the pre-trained generator, it's less prone to produce artifacts not inherent to a human face (First Order Motion, second row, middle image, eyes). However, this constraint has also a drawback in terms of facial accessories it's capable of modeling (the disappearance of a microphone in the second row). Note that, since First Order Motion is focused on the face reenactment task, it produces better results than the FSGAN model.

In the second experiment, we qualitatively compare our method with state-of-the-art in the context of face swapping. The visual comparison is shown in the Fig. 5.6. For the face swapping task, GAN based methods may fail in cross-gender face swapping due to the difference between gender appearance and shape. We show that our method produces realistic-looking results both for male-to-female (rows 1, 3, 6) and female-to-male swapping (row 2) compared to competitive methods: First Order Motion keeps the lipstick color of a target face (row 3), FSGAN loses the identity of the source image (rows 1, 3, 6). Note that, since FSGAN is focused on the face swapping task, it produces better results than the First Order Motion model.

Quantitative Evaluation

We provide additional quantitative evaluations on 300VW to verify preservation of identity/expression/pose w.r.t. SOTA and to motivate the benefit of joint learning (Table 5.2). We compare identity preservation using cosine similarity between latent space of VGG-Face2 features [Cao et al., 2018]. Headpose correctness is compared using absolute distance in degrees of yaw-pitch-roll predicted from a pretrained Hopenet [Ruiz et al., 2018]. Expression correctness is compared using mean absolute distance of facial landmarks (in pixels, image resized to 256) using a pretrained FAN detector [Bulat and Tzimiropoulos, 2017]. Our method outperforms SOTA in the swapping task on 3 benchmarks. On the reenactment task Firt Order Motion performs better on identity and headpose preservation however, on average, our method outperforms SOTA.

	Identity ↑			Headpose ↓			Expression ↓		
	First Order	FSGAN	Ours	First Order	FSGAN	Ours	First Order	FSGAN	Ours
Reenactment	**0.578**	0.461	0.517	**2.811**	4.268	3.364	4.883	51.56	**3.983**
Swap	0.308	0.317	**0.412**	2.628	2.823	**2.113**	3.902	2.554	**3.072**
Avg	0.443	0.389	**0.464**	**2.719**	3.546	2.739	4.393	27.057	**3.528**

Table 5.2: Quantitative evaluation on 300VW.

5.5 Limitations

Despite promising results presented in this work, our method has several limitations. First, the expressiveness of the generated facial expressions is dependent on its presence in the training dataset and the quality of face landmarks provided by the landmark detector. Second, our model does not explicitly model occlusion and consequently relies on a pre-trained generator to have a capacity of modeling occlusions, such as accessories or makeup. Finally, both landmark plots and source images contain a bias in terms of identity, pose and expression.

5.6 Conclusion

In this work, we proposed a novel approach for isolated disentangled representation learning combined with an end-to-end method capable of performing both face reenactment and swapping. To our knowledge, our method is the first approach that is designed to solve both objectives in a unified pipeline.

We showed that our method is trained in an unsupervised way to achieve equally good visual results on both tasks. In addition, it's capable of producing results in a one-shot manner during inference time. The qualitative results on multiple public datasets show that the proposed method is outperforming SOTA methods which can perform both face reenactment and swap.

Chapter 6

Summary and Conclusion

Below are individual chapter summary followed by the thesis conclusion.

6.1 Summary

6.1.1 Pose- and Expression- Robust Age Estimation

We show that by incorporating prior knowledge about face image formation one can significantly improve the age estimation performance. Our method takes a single 2D image and derives 3D reconstruction features as a new source of pose and facial expression robustness by employing a monocular 3D face reconstruction model.

Experiments show our method to be consistently more robust across expression and pose variation and improved the baseline the most with extreme head poses (1.4 MAE) and intensive expressions (1.82 MAE).

6.1.2 Identity-Unbiased Deception Detection

We show that face image formation prior can be used to disentangle identity and environment-related features from the face input data. The newly proposed deception detection method is based on reliable facial expression and head pose-related features. We achieve separation of those properties by simultaneously learning two separate CNNs using 2D-to-3D face reconstruction: (a) one CNN for face identity and environment

parameters, and (b) another CNN for facial expression and head pose. Our pipeline predicts a single label given a frame sequence and models deceit detection as Multiple Instance Learning problems conditioned on reconstruction features.

Prior works have been focusing on high-stakes deceit detection mainly because of the lack of publicly available datasets for low-stake deceit. A new Low-Stakes Deceit (LSD) dataset has been collected to address this issue. To our knowledge, we are the first to evaluate automatic visual-based high-stake deceit detection methods on low-stakes deceit detection tasks.

Experiments show our method to improve the state-of-the-art as well as providing on par results with the use of manually coded facial attributes (71%) in the high-stakes deception detection on the challenging RLT dataset. In the low-stakes lies deception detection task it has achieved results on par with professional experts however there is still room for improvement.

6.1.3 Self-supervised Face Image Manipulation

In prior chapters, the benefit of attribute decomposition using face image formation prior has been shown for discriminative learning tasks. In this chapter, we explore further the benefit of attribute decomposition for generative learning tasks.

We propose a self-supervised method to manipulate a single monocular face image by conditioning GAN on face decomposition using appearance transfer. Our conditioning has shown to be a more flexible representation in comparison to previous GAN-based methods that use discrete classes, landmarks, or action units. Thus, conditioning on the appearance image allows us to manipulate head pose, scene illumination, and facial expression using a single conditioning space.

Experiments show that our proposed method provides competitive results in terms of identity preservation, and outperforms, in terms of expression correctness, GAN-based state-of-the-art face reenactment methods, which can do both expression and head pose manipulation. Our method is agnostic to face decomposition methods and works with any 2D-to-3D reconstruction method which allows pose, expression, and light manipulation.

6.1.4 Face Reenactment and Swapping

In prior chapters, we study the benefit of face image formation prior knowledge in various face analysis and synthesis tasks. In this chapter, we explore the possibility to learn attribute decomposition without such prior knowledge.

We propose a method for isolated disentangled representation learning combined with an end-to-end method capable of performing both face reenactment and swapping. To our knowledge, this is the first approach that is designed to solve both objectives in a unified pipeline in an unsupervised manner. It achieves that by mapping the disentangled latent representation to the latent space of a pre-trained generator with disentanglement property. During the test time, our method requires source and target images together with their facial landmarks to predict reenacted or swapped results.

Experiments show the proposed method to achieve equally good visual results on reenactment and swapping tasks. In addition, it's capable of producing results in a one-shot manner during inference time. The qualitative results on multiple public datasets show that the proposed method outperforms state-of-the-art methods, which can perform both face reenactment and swap.

6.2 Conclusion

This thesis is focused on learning face attribute decomposition to improve face analysis and synthesis. Given an image of a face, prior knowledge about its interaction with the environment, light source, and how the face may geometrically change, are considered. By incorporating such kind of prior knowledge, one can achieve a better, more robust predictive models which generalize well on unseen data. In the case of the infeasibility of modeling a prior, a commonality between face analysis or synthesis tasks can be considered to learn a joint attribute representation conditioned on constraints from those tasks.

Prior knowledge about face can help to achieve better discriminative learning models. In this thesis, we show how domain knowledge about face can be explicitly incorporated into age estimation and deception detection models. A similar idea can be adapted for generative learning models for the monocular face image manipulation task. In the absence / incapability of explicit modeling, implicit modeling of prior knowledge via multi-task learning can be used.

In Chapter 2 we answer **the first research question** and show how deep learning-based age estimation can be robustified by jointly learning 2D-to-3D face reconstruction and age estimation tasks. **The second research question** concerns the identity biases in the existing deception detection methods. We answer it in Chapter 3 by proposing a deep learning-based deception detection method, which is capable of the disentanglement

of the identity and environment-related features from input data. To address **the third research question** we propose in Chapter 4 a self-supervised conditional GAN-based method that is capable of face image manipulation given face decomposition. **The fourth research question** concerns the feasibility of learning face reenactment and swapping tasks without explicit prior. We address the question in Chapter 5 by proposing an unsupervised model which is capable of solving both objectives by mapping the disentangled latent representation to the latent space of a pre-trained generator with disentanglement property.

In conclusion, decomposition of the face representation using prior knowledge benefits deep learning models by making them generalize better on unseen data. Our research support this conclusion in age estimation, deception detection, and manipulation of attributes of a face image. However, further research is required to evaluate our hypothesis in other face-related tasks, such as emotion and kinship recognition.

Appendix A

Self-supervised Face Image Manipulation

A.1 Implementation Details

This supplemental material provides more results and evaluation and implementation details of our novel pipeline for self-supervised face image manipulation. Fig. A.2 shows generator and discriminator architectures of our pipeline. Source image $\mathbf{X}_S$ is concatenated with target appearance $\mathbf{Y}_T$ to form an input (6 channels) of the generative model. For the discriminator, the input is 9 channels image consist of source image $\mathbf{X}_S$, target appearance $\mathbf{Y}_T$, and target image $\mathbf{X}_T$ / fake target image $\hat{\mathbf{X}}_S$.
We use convolutions followed by instance normalization and leaky relu ($\alpha = 0.2$) for generator. For discriminator spectral normalization [Miyato et al., 2018] for convolution together with leaky relu are used. Note that amount of filters in our generator/discriminator architectures is 2 times less than the one used in competitive methods StarGAN [Choi et al., 2018] and GANimation [Pumarola et al., 2018a].

A.2 Additional Results

Fig. A.3 shows additional cross-dataset qualitative comparison to state-of-the-art methods GANimation [Pumarola et al., 2018a] and StarGAN [Choi et al., 2018] under RaFD expression settings. Next we provide additional qualitative results of manipulation of pose (Fig. A.1), face expression (Fig. A.4) and light (Fig. A.5). Finally, simultaneous change of head pose and expression produced by our pipeline is shown in the Figure A.6.

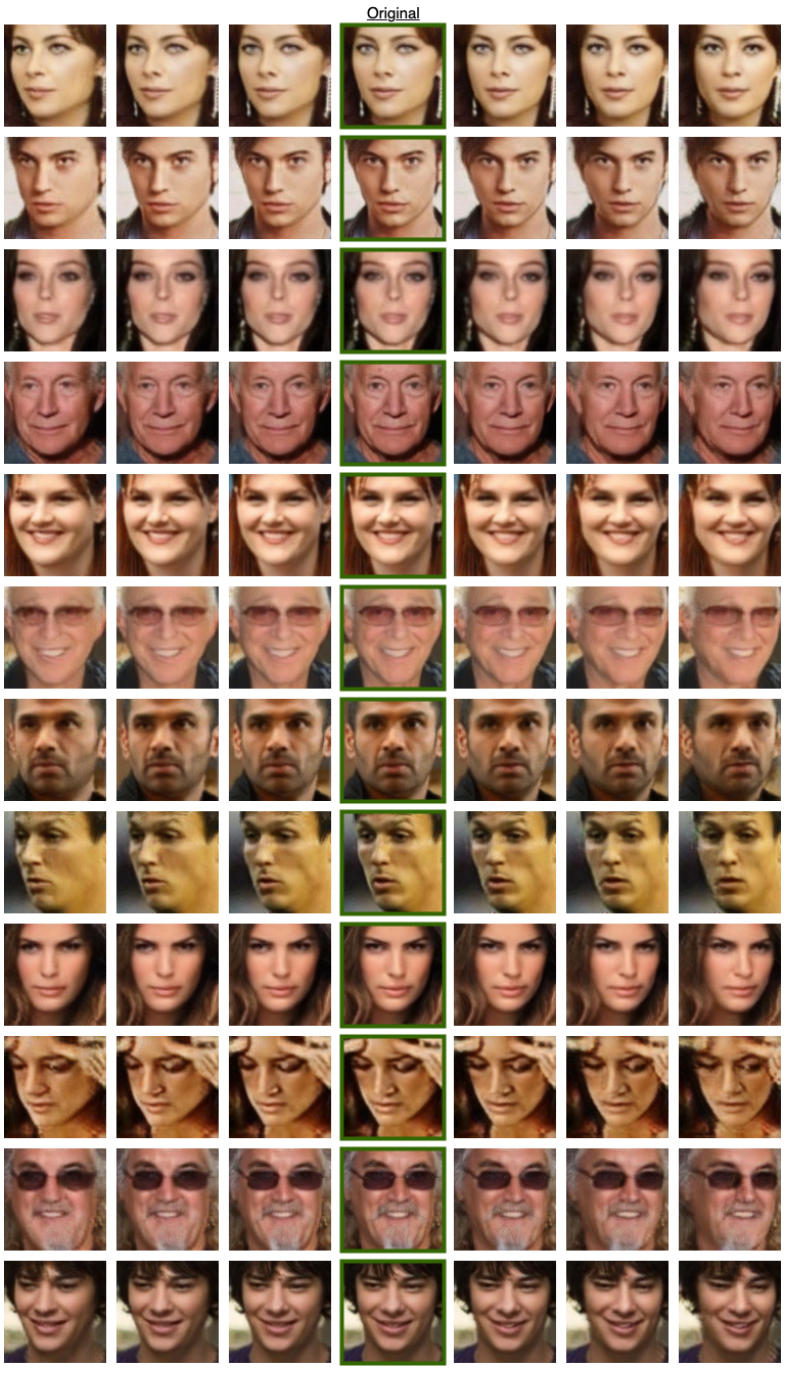


Figure A.1: Additional qualitative results for pose manipulation of a single face image from CelebA.

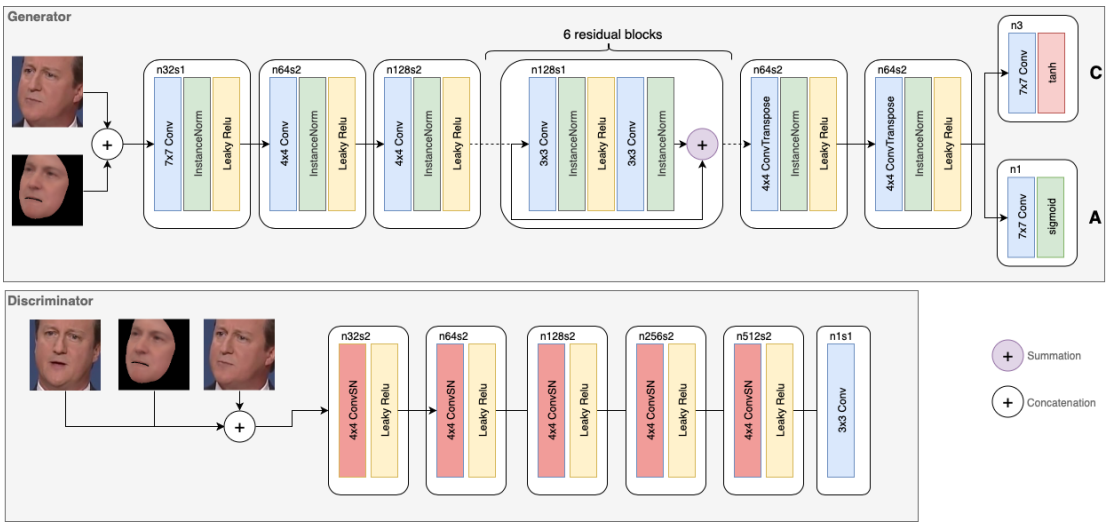


Figure A.2: Generator and discriminator architectures of our pipeline. As ConvSN we denote convolution layer with spectral normalization [Miyato et al., 2018]. Note that amount of filters in our generator is 2 times less than the one used in competitive methods StarGAN and GANimation [Choi et al., 2018, Pumarola et al., 2018a].

Figure A.3: Additional results for qualitative comparison to state-of-the-art methods GANimation [Pumarola et al., 2018a] and StarGAN [Choi et al., 2018]. We are performing cross-dataset comparison on CelebA [Liu et al., 2015] applying RaFD [Langner et al., 2010] expression on a source image.

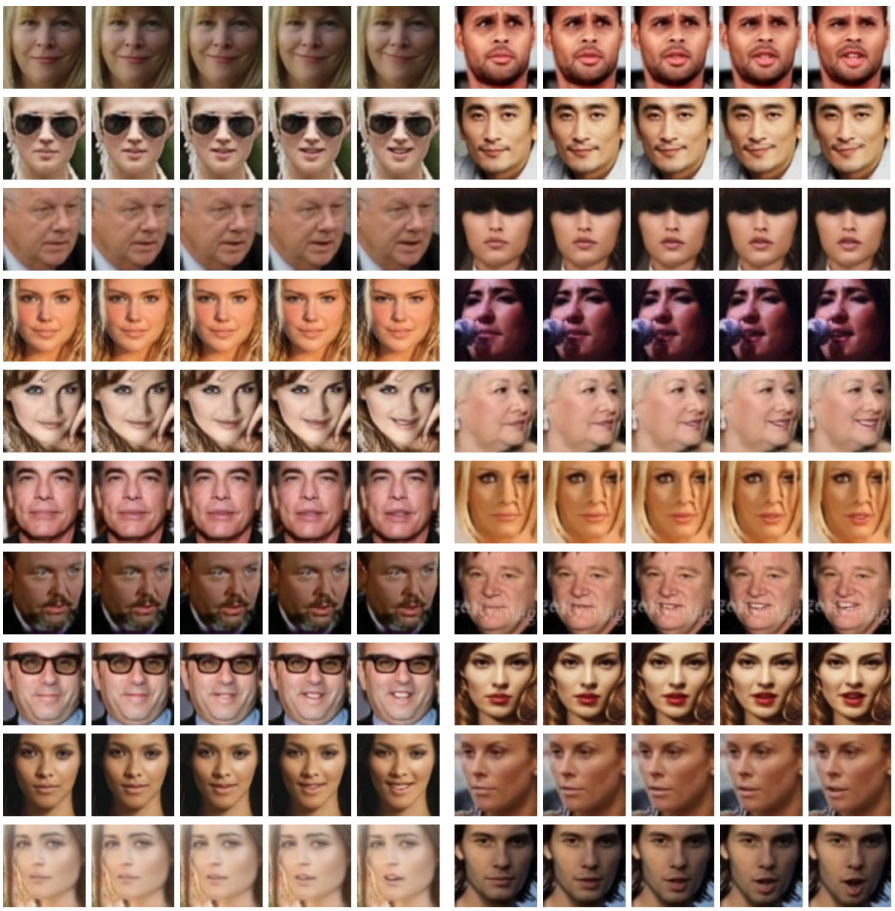


Figure A.4: Additional qualitative results for expression manipulation of a single face image from CelebA.

Figure A.5: Additional qualitative results for light manipulation of a single face image. Target light direction (first column) from the Multi-PIE [Gross et al., 2010] dataset is applied on images from the CelebA [Liu et al., 2015] dataset.

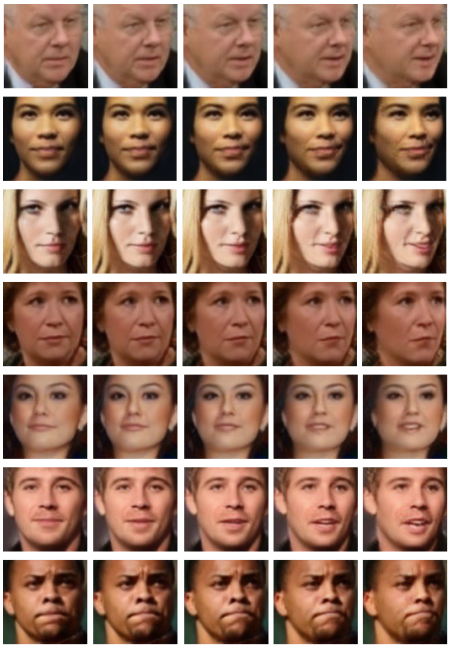


Figure A.6: Additional qualitative results for expression and pose manipulation of a single face image from CelebA.

Samenvatting

Hoofdstuk 2

We tonen aan dat de nauwkeurigheid voor de taak van leeftijdsschatting aanzienlijk verbeterd kan worden door het integreren van a priori kennis over het in beeld brengen van gezichten. Onze methode neemt één 2D-beeld en leidt hieruit 3D-reconstructiefeatures af, als een nieuwe bron van robuustheid voor pose en gezichtsuitdrukking, door gebruik te maken van een monoculair 3D-gezichtsreconstructiemodel.

De experimenten laten zien dat onze method consist robuuster is voor gezichtsuitdrukkking en variatie in pose, en overtreft de baseline het meest voor extreme poses van het hoofd (1,4 MAE) en intensieve gezichtsuitdrukkingen (1,82 MAE).

Hoofdstuk 3

We laten zien dat a priori kennis over het in beeld brengen van gezichten gebruikt kan worden om onderscheid te maken tussen de identiteit van het gezicht en omgevingsgerelateerde features, uit de ingevoerde data van gezichten. De nieuw ontwikkelde fraudedetectiemethode is gebaseerd op betrouwbare gezichtsuitdrukkings- en posegerelateerde features. Het lukt ons de genoemde twee eigenschappen uit elkaar te houden, door tegelijkertijd twee CNNs te trainen met behulp van 2D-naar-3D gezichtsreconstructie: (a) één CNN voor de parameters van identiteit van het gezicht en omgevingsfactoren, en (b) één CNN voor gezichtsuitdrukking en pose van het hoofd. Onze pipeline voorspelt één label, gegeven een reeks frames, en modelleert fraudedetectie als een Multiple Instance Learning-probleem, geconditioneerd op reconstructiefeatures.

Eerder werk heeft zicht vooral gericht op fraudedetectie waarbij de inzet (mogelijke positieve of negatieve consequenties) hoog is, vooral door een gebrek aan openbaar

toegankelijke datasets voor misleiding waarbij de inzet laag is. Wij hebben een lage-inzetfraudedataset (Low-Stakes Deceit, LSD) verzameld om aan dit probleem tegemoet te komen. Naar ons weten zijn wij de eersten die automatische visuele hoge-inzetfraudedetectiemethoden evalueert op lage-inzetfraudedetectietaken.
Experimenten laten zien dat onze methode beter presteert dan de state-of-the-art en daarnaast resultaten laat zien die gelijkwaardig zijn aan het gebruik van handmatig gecodeerde gezichtseigenschappen voor hoge-inzetfraudedetectie op de uitdagende RLT dataset (71%). Op de lage-inzetfraudedetectietaak behaalde onze methode vergelijkbare resultaten als professionele experts, hoewel er nog ruimte is voor verbetering.

Hoofdstuk 4

In eerdere hoofdstukken is het voordeel aagetoond van de decompositie van attributen met behulp van a priori kennis over het in beeld brengen van gezichten voor discriminatieve leertaken. In dit hoofdstuk bespreken we het potentieel van decompositie van attributen voor generatieve leertaken.
We introduceren een zelf-gesuperviseerde methode voor het manipuleren van één monoculair beeld van een gezicht, door het conditioneren van een GAN op gezichtsdecompositie, door gebruik te maken van overdracht van uiterlijk. Onze conditionering blijkt een meer flexibele representatie in vergelijking met eerdere GAN-gebaseerde methoden die discrete klassen, oriëntatiepunten of actie-eenheden gebruiken. Het conditioneren op het beeld van het uiterlijk maakt het mogelijk om gezichtspose, belichting van de scène en gezichtsuitdrukking te manipuleren met behulp van één conditionerende ruimte.
Uit de experimenten blijkt dat de methode die wij voorstellen competitieve resultaten laat zien in termen van behoud van identiteit, en overtreft zelfs, in termen van de correctheid van de gezichtsuitdrukking, GAN-gebaseerde state-of-the-art gezichtsoverdrachtsmethoden, die zowel gezichtsuidrukking als pose van het hoofd kunnen manipuleren. Onze methode is onafhankelijk van de gebruikte gezichtsdecompositiemethode en werkt met elke 2D-naar-3D reconstructiemethode die pose, uitdrukking en lichtmanipulatie ondersteunt.

Hoofdstuk 5

In eerdere hoofdstukken onderzochten we het voordeel van a priori kennis over het in beeld brengen van gezichten voor meerdere gezichtsanalyse en -synthesetaken. In dit

hoofdstuk onderzoeken we de mogelijkheid om de decompositie van attributen te leren zonder a priori kennis.

We introduceren een methode voor het leren van geïsoleerde ontwarde representaties gecombineerd met een end-to-end methode die zowel de overdracht van pose/uitdrukkingen als het verwisselen van gezichten mogelijk maakt. Naar ons weten is dit de eerste methode die is ontwikkeld om beide doelen binnen één gezamenlijke pipeline uit te voeren, middels een ongesuperviseerd paradigma. Dit wordt bereikt door een functie op te stellen van de ontwarde latente representatie naar de latente ruimte van een gepretrainde generator met ontwarringseigenschap. Tijdens de evaluatiefase kan onze methode, op basis van bron- en doelafbeeldingen en hun oriëntatiepunten, overgedragen of verwisselde resultaten voorspellen.

Onze experimenten laten zien dat de voorgestelde methode goede visuele resultaten bereikt op de overdrachts- en verwisseltaken. Bovendien is het model in staat one-shot resultaten te produceren tijdens de evaluatiefase. De kwalitatieve resultaten op verschillende openbare datasets tonen dat de voorgestelde methode state-of-the-art methoden, die zowel overdracht als verwisseling kunnen uitvoeren, overtreft.

Bibliography

Abdal, R., Qin, Y., and Wonka, P. (2019). Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4432–4441.

Abouelenien, M., Pérez-Rosas, V., Mihalcea, R., and Burzo, M. (2014). Deception detection using a multimodal approach. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 58–65. ACM.

Amberg, B., Knothe, R., and Vetter, T. (2008). Expression invariant 3d face recognition with a morphable model. In *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 1–6.

Andrew, A. M. (2001). Multiple view geometry in computer vision. *Kybernetes*.

Angulu, R., Tapamo, J. R., and Adewumi, A. O. (2018). Age estimation via face images: a survey. *EURASIP Journal on Image and Video Processing*, 2018(1):42.

Arriaga, O., Valdenegro-Toro, M., and Plöger, P. (2017). Real-time convolutional neural networks for emotion and gender classification. *CoRR*, abs/1710.07557.

Averbuch-Elor, H., Cohen-Or, D., Kopf, J., and Cohen, M. F. (2017). Bringing portraits to life. *ACM Transactions on Graphics (Proceeding of SIGGRAPH Asia 2017)*, 36(6):196.

Ba, J. L., Kiros, J. R., and Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*.

Baltrusaitis, T., Mahmoud, M., and Robinson, P. (2015). Cross-dataset learning and person-specific normalisation for automatic action unit detection. *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 06:1–6.

Baltrušaitis, T., Robinson, P., and Morency, L.-P. (2016). Openface: an open source facial behavior analysis toolkit. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–10. IEEE.

Banerjee, S., Scheirer, W. J., Bowyer, K. W., and Flynn, P. J. (2018). On hallucinating context and background pixels from a face mask using multi-scale gans. *CoRR*, abs/1811.07104.

Baslamisli, A. S., Das, P., Le, H. A., Karaoglu, S., and Gevers, T. (2021). Shadingnet: Image intrinsics by fine-grained shading decomposition. *International Journal of Computer Vision*.

Baslamisli, A. S., Groenestege, T. T., Das, P., Le, H. A., Karaoglu, S., and Gevers, T. (2018). Joint learning of intrinsic images and semantic segmentation. In *European Conference on Computer Vision*.

Bitouk, D., Kumar, N., Dhillon, S., Belhumeur, P., and Nayar, S. K. (2008). Face swapping: Automatically replacing faces in photographs. *ACM Trans. Graph.*, 27(3):1–8.

Bińkowski, M., Sutherland, D. J., Arbel, M., and Gretton, A. (2018). Demystifying mmd gans.

Blanz, V. and Vetter, T. (1999). A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '99, pages 187–194, New York, NY, USA. ACM Press/Addison-Wesley Publishing Co.

Bond Jr, C. F. and DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and social psychology Review*, 10(3):214–234.

Booth, J., Roussos, A., Ponniah, A., Dunaway, D., and Zafeiriou, S. (2018). Large scale 3d morphable models. *International Journal of Computer Vision*, 126(2-4):233–254.

Bouma, H., Burghouts, G., den Hollander, R., Van Der Zee, S., Baan, J., ten Hove, J.-M., van Diepen, S., van den Haak, P., and van Rest, J. (2016). Measuring cues for stand-off deception detection based on full-body nonverbal features in body-worn cameras. In *Optics and Photonics for Counterterrorism, Crime Fighting, and Defence XII*, volume 9995, page 99950N. International Society for Optics and Photonics.

Bulat, A. and Tzimiropoulos, G. (2017). How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*.

Burgoon, J. K., Magnenat-Thalmann, N., Pantic, M., and Vinciarelli, A. (2017). *Social signal processing*. Cambridge University Press.

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., and Zisserman, A. (2018). Vggface2: A dataset for recognising faces across pose and age. In *International Conference on*

Automatic Face and Gesture Recognition.

Chen, H.-J., Hui, K.-M., Wang, S.-Y., Tsao, L.-W., Shuai, H.-H., and Cheng, W.-H. (2019). Beautyglow: On-demand makeup transfer framework with reversible generative network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10042–10050.

Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., and Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797.

Chrysos, G., Antonakos, E., Zafeiriou, S., and Snape, P. (2015). Offline deformable face tracking in arbitrary videos. In *Proceedings of IEEE International Conference on Computer Vision Workshops, 300 Videos in the Wild (300-VW): Facial Landmark Tracking in-the-Wild Challenge & Workshop*, pages 1–9, Santiago, Chile. IEEE.

Chrysos, G. G., Antonakos, E., Snape, P., Asthana, A., and Zafeiriou, S. (2018). A comprehensive performance evaluation of deformable face tracking in-the-wild. *International Journal of Computer Vision*, 126(2-4):198–232.

Dale, K., Sunkavalli, K., Johnson, M. K., Vlasic, D., Matusik, W., and Pfister, H. (2011). Video face replacement. *ACM Trans. Graph.*, 30(6):1–130.

Deng, Y., Yang, J., Xu, S., Chen, D., Jia, Y., and Tong, X. (2019). Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *IEEE Computer Vision and Pattern Recognition Workshops*.

DesJardins, N. M. L. and Hodges, S. D. (2015). Reading between the lies: Empathic accuracy and deception detection. *Social Psychological and Personality Science*, 6(7):781–787.

Diamant, N., Zadok, D., Baskin, C., Schwartz, E., and Bronstein, A. M. (2019). Beholder-gan: Generation and beautification of facial images with conditioning on their beauty level. *arXiv preprint arXiv:1902.02593*.

Dibeklioglu, H., Alnajar, F., Salah, A. A., and Gevers, T. (2015). Combining facial dynamics with appearance for age estimation. *IEEE Transactions on Image Processing*.

Dibeklioğlu, H., Salah, A. A., and Gevers, T. (2012). Are you really smiling at me? spontaneous versus posed enjoyment smiles. In *European Conference on Computer Vision*, pages 525–538. Springer.

Ding, C. and Tao, D. (2016). A comprehensive survey on pose-invariant face recognition.

ACM Transactions on Intelligent Systems and Technology, 7(3):37.

Ding, C., Zhou, D., He, X., and Zha, H. (2006). R1-pca: Rotational invariant l1-norm principal component analysis for robust subspace factorization. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 281–288, New York, NY, USA. ACM.

Dong, X., Yu, S.-I., Weng, X., Wei, S.-E., Yang, Y., and Sheikh, Y. (2018). Supervision-by-Registration: An unsupervised approach to improve the precision of facial landmark detectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 360–368.

Farkas, L. G. (1994). *Anthropometry of the Head and Face*. Raven Pr.

Fawaz, H. I., Forestier, G., Weber, J., Idoumghar, L., and Muller, P.-A. (2019). Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, 33(4):917–963.

Feng, Z.-H., Kittler, J., Awais, M., Huber, P., and Wu, X.-J. (2018). Wing loss for robust facial landmark localisation with convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2235–2245.

Fiedler, K., Schmid, J., and Stahl, T. (2002). What is the current truth about polygraph lie detection? *Basic and Applied Social Psychology*, 24(4):313–324.

Fu, C., Hu, Y., Wu, X., Wang, G., Zhang, Q., and He, R. (2019). High fidelity face manipulation with extreme pose and expression. *arXiv preprint arXiv:1903.12003*.

Gabbay, A. and Hoshen, Y. (2019). Style generator inversion for image enhancement and animation. *CoRR*, abs/1906.11880.

Gao, F. and Ai, H. (2009). Face age classification on consumer images with gabor feature and fuzzy lda method. In *International Conference on Biometrics*, pages 132–141.

Garrido, P., Valgaerts, L., Rehmsen, O., Thormaehlen, T., Perez, P., and Theobalt, C. (2014). Automatic face reenactment. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '14, page 4217–4224, USA. IEEE Computer Society.

Garrido, P., Valgaerts, L., Wu, C., and Theobalt, C. (2013). Reconstructing detailed dynamic face geometry from monocular video. *ACM Trans. Graph.*, 32(6):158–1.

Geng, X., Zhou, Z.-H., and Smith-Miles, K. (2007). Automatic age estimation based on facial aging patterns. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29(12):2234–2240.

Genova, K., Cole, F., Maschinot, A., Sarna, A., Vlasic, D., and Freeman, W. T. (2018). Unsupervised training for 3d morphable model regression. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Gerig, T., Morel-Forster, A., Blumer, C., Egger, B., Luthi, M., Schönborn, S., and Vetter, T. (2018). Morphable face models-an open framework. In *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 75–82.

Gevers, T., Gijsenij, A., Van de Weijer, J., and Geusebroek, J.-M. (2012). *Color in computer vision: fundamentals and applications*, volume 23. John Wiley & Sons.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 2672–2680, Cambridge, MA, USA. MIT Press.

Gross, R., Matthews, I., Cohn, J., Kanade, T., and Baker, S. (2010). Multi-pie. *Image Vision Comput.*, 28(5):807–813.

Guo, G., Mu, G., Fu, Y., and Huang, T. S. (2009). Human age estimation using bio-inspired features. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 112–119.

Guo, G. and Wang, X. (2012). A study on human age estimation under facial expression changes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2547–2553.

Han, J., Karaoglu, S., Le, H.-A., and Gevers, T. (2021). Object features and face detection performance: Analyses with 3d-rendered synthetic data. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9959–9966.

Han, X., Yang, H., Xing, G., and Liu, Y. (2020). Asymmetric joint gans for normalizing face illumination from a single image. *IEEE Transactions on Multimedia*, 22(6):1619–1633.

Hara, K., Kataoka, H., and Satoh, Y. (2018). Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6546–6555.

Hartwig, M. and Bond Jr, C. F. (2014). Lie detection from multiple cues: A meta-analysis. *Applied Cognitive Psychology*, 28(5):661–676.

He, Z., Zuo, W., Kan, M., Shan, S., and Chen, X. (2019). Attgan: Facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing*,

28(11):5464–5478.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, pages 6629–6640.

Hodge, S. (2019). *Painting Masterclass: Creative Techniques of 100 Great Artists*. Thames & Hudson, London.

Hu, X., Ren, W., LaMaster, J., Cao, X., Li, X., Li, Z., Menze, B. H., and Liu, W. (2020). Face super-resolution guided by 3d facial priors. *CoRR*, abs/2007.09454.

Hu, Y., Wu, X., Yu, B., He, R., and Sun, Z. (2018). Pose-guided photorealistic face rotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8398–8406.

Hu, Z., Wen, Y., Wang, J., Wang, M., Hong, R., and Yan, S. (2017). Facial age estimation with age difference. *IEEE Transactions on Image Processing*, 26(7):3087–3097.

Huang, X. and Belongie, S. J. (2017). Arbitrary style transfer in real-time with adaptive instance normalization. *CoRR*, abs/1703.06868.

Huang, Z., Chen, S., Zhang, J., and Shan, H. (2021a). Pfa-gan: Progressive face aging with generative adversarial network. *IEEE Transactions on Information Forensics and Security*, 16:2031–2045.

Huang, Z., Zhang, J., and Shan, H. (2021b). When age-invariant face recognition meets face age synthesis: A multi-task learning framework. In *CVPR*.

Ilse, M., Tomczak, J. M., and Welling, M. (2018). Attention-based deep multiple instance learning. *arXiv preprint arXiv:1802.04712*.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. *CVPR*.

Jing, Y., Yang, Y., Feng, Z., Ye, J., and Song, M. (2017). Neural style transfer: A review. *CoRR*, abs/1705.04058.

Johnson, J., Alahi, A., and Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. (2017). Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*.

Karras, T., Laine, S., and Aila, T. (2019a). A style-based generator architecture for

generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. (2019b). Analyzing and improving the image quality of StyleGAN. *CoRR*, abs/1912.04958.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., and Zisserman, A. (2017). The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Kemelmacher-Shlizerman, I. (2016). Transfiguring portraits. *ACM Trans. Graph.*, 35(4).

Kim, H., Garrido, P., Tewari, A., Xu, W., Thies, J., Nießner, N., Pérez, P., Richardt, C., Zollhöfer, M., and Theobalt, C. (2018a). Deep video portraits. *ACM Transactions on Graphics 2018 (TOG)*, 37:1–14.

Kim, H., Zollhöfer, M., Tewari, A., Thies, J., Richardt, C., and Theobalt, C. (2018b). InverseFaceNet: Deep monocular inverse face rendering. In *IEEE Conference on Computer Vision and Pattern Recognition*.

King, D. E. (2009). Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10(60):1755–1758.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Koizumi, T. and Smith, W. A. P. (2020). "Look ma, no landmarks!" - unsupervised, model-based dense face alignment. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part II*, volume 12347 of *Lecture Notes in Computer Science*, pages 690–706. Springer.

Korshunova, I., Shi, W., Dambre, J., and Theis, L. (2016). Fast face-swap using convolutional neural networks. *CoRR*, abs/1611.09577.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105.

Kurach, K., Lucic, M., Zhai, X., Michalski, M., and Gelly, S. (2018). The GAN landscape: Losses, architectures, regularization, and normalization. *CoRR*, abs/1807.04720.

Kuster, C., Popa, T., Bazin, J.-C., Gotsman, C., and Gross, M. (2012). Gaze correction for home video conferencing. *ACM Trans. Graph.*, 31(6).

Kwon, Y. H. et al. (1994). Age classification from facial images. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 762–767.

Lakhani, M. and Taylor, R. (2003). Beliefs about the cues to deception in high- and low-stake situations. *Psychology, Crime & Law*, 9(4):357–368.

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., Hawk, S. T., and van Knippenberg, A. (2010). Presentation and validation of the Radboud Faces Database. *Cognition & Emotion*, 24(8):1377–1388.

Le, H. A., Baslamisli, A. S., Mensink, T., and Gevers, T. (2018). Three for one and one for three: Flow, segmentation, and surface normals. In *British Machine Vision Conference*.

Lettry, L., Vanhoey, K., and Gool, L. V. (2018). Darn: A deep adversarial residual network for intrinsic image decomposition. *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1359–1367.

Levi, G. and Hassner, T. (2015). Age and gender classification using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 34–42.

Li, C., Zhou, K., Wu, H.-T., and Lin, S. (2018). Physically-based simulation of cosmetics via intrinsic image decomposition with facial priors. *IEEE transactions on pattern analysis and machine intelligence*, 41(6):1455–1469.

Li, L., Bao, J., Yang, H., Chen, D., and Wen, F. (2019). Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv preprint arXiv:1912.13457*.

Lim, J. H. and Ye, J. C. (2017). Geometric gan. *arXiv preprint arXiv:1705.02894*.

Liu, Y., xia Li, Q., Sun, Z., and Tan, T. (2019). A3gan: An attribute-aware attentive generative adversarial network for face aging. *ArXiv*, abs/1911.06531.

Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). Deep learning face attributes in the wild. In *International Conference on Computer Vision*.

Lou, Z., Alnajar, F., Alvarez, J. M., Hu, N., and Gevers, T. (2018). Expression-invariant age estimation using structured learning. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 40(2):365–375.

Lu, J. and Tan, Y.-P. (2013). Ordinary preserving manifold analysis for human age and head pose estimation. *IEEE Trans. on Human-Machine Systems*, 43(2):249–258.

Mahajan, D., Agrawal, N., Keerthi, S. S., Sellamanickam, S., and Bottou, L. (2018). An efficient distributed learning algorithm based on effective local functional approximations. *Journal of Machine Learning Research*, 19:74:1–74:37.

Masi, I., Chang, F.-J., Choi, J., Harel, S., Kim, J., Kim, K., Leksut, J., Rawls, S., Wu,

Y., Hassner, T., et al. (2019). Learning pose-aware models for pose-invariant face recognition in the wild. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 41(2):379–393.

Mechrez, R., Talmi, I., and Zelnik-Manor, L. (2018). The contextual loss for image transformation with non-aligned data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 768–783.

Ming, Z., Xia, J., Luqman, M. M., Burie, J., and Zhao, K. (2019). Dynamic multi-task learning for face recognition with facial expression. *CoRR*, abs/1911.03281.

Mirza, M. and Osindero, S. (2014). Conditional generative adversarial nets. *CoRR*, abs/1411.1784.

Misra, I., Shrivastava, A., Gupta, A., and Hebert, M. (2016). Cross-stitch networks for multi-task learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3994–4003.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. (2018). Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*.

Morales, M. R., Scherer, S., and Levitan, R. (2017). Openmm: An open-source multimodal feature extraction tool. In *Proc. Interspeech 2017*, pages 3354–3358.

Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I., and Zafeiriou, S. (2017). Agedb: the first manually collected, in-the-wild age database. In *Conference on Computer Vision and Pattern Recognition Workshops*, volume 2 (3), page 5.

Müller, C. (1966). Spherical harmonics, volume 17 of lecture notes in mathematics.

Nagano, K., Seo, J., Xing, J., Wei, L., Li, Z., Saito, S., Agarwal, A., Fursund, J., and Li, H. (2018). Pagan: Real-time avatars using dynamic textures. *ACM Trans. Graph.*, 37(6).

Napoléon, T. and Alfalou, A. (2017). Pose invariant face recognition: 3d model from single photo. *Optics and Lasers in Engineering*, 89:150–161.

Ngo, L. M., de Wiel, C. a., Karaoglu, S., and Gevers, T. (2020). Unified application of style transfer for face swapping and reenactment. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*.

Ngo, L. M., Wang, W., Mandira, B., Karaoglu, S., Bouma, H., Dibeklioglu, H., and Gevers, T. (2021). Identity unbiased deception detection by 2d-to-3d face reconstruction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer*

Vision (WACV), pages 145–154.

Ngo, M., Karaoglu, S., and Gevers, T. (2021). Self-supervised face image manipulation by conditioning gan on face decomposition. *IEEE Transactions on Multimedia*, pages 1–1.

Nguyen, A., Yosinski, J., Bengio, Y., Dosovitskiy, A., and Clune, J. (2016). Plug & play generative networks: Conditional iterative generation of images in latent space. *CoRR*, abs/1612.00005.

Nirkin, Y., Keller, Y., and Hassner, T. (2019). FSGAN: Subject Agnostic Face Swapping and Reenactment. In *Proceedings of the IEEE international conference on computer vision*, pages 7184–7193.

Nirkin, Y., Masi, I., Tran, A. T., Hassner, T., and Medioni, G. (2018). On face segmentation, face swapping, and face perception. In *IEEE Conference on Automatic Face and Gesture Recognition*.

O'Sullivan, M., Frank, M. G., Hurley, C. M., and Tiwana, J. (2009). Police lie detection accuracy: The effect of lie scenario. *Law and Human Behavior*, 33(6):530.

Paysan, P., Knothe, R., Amberg, B., Romdhani, S., and Vetter, T. (2009). A 3d face model for pose and illumination invariant face recognition. In *IEEE International Conference on Advanced Video and Signal based Surveillance*, pages 296–301.

Peng, X., Yu, X., Sohn, K., Metaxas, D. N., and Chandraker, M. (2017). Reconstruction-based disentanglement for pose-invariant face recognition. In *IEEE International Conference on Computer Vision*, pages 1623–1632.

Pentland, S. J., Twyman, N. W., Burgoon, J. K., Nunamaker Jr, J. F., and Diller, C. B. (2017). A video-based screening system for automated risk assessment using nuanced facial features. *Journal of Management Information Systems*, 34(4):970–993.

Pérez-Rosas, V., Abouelenien, M., Mihalcea, R., and Burzo, M. (2015). Deception detection using real-life trial data. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, pages 59–66, New York, NY, USA. ACM.

Phillips, P. J., Moon, H., Rizvi, S. A., and Rauss, P. J. (2000). The feret evaluation methodology for face-recognition algorithms. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(10):1090–1104.

Pumarola, A., Agudo, A., Martinez, A. M., Sanfeliu, A., and Moreno-Noguer, F. (2018a). Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings*

of the European Conference on Computer Vision (ECCV), pages 818–833.

Pumarola, A., Agudo, A., Sanfeliu, A., and Moreno-Noguer, F. (2018b). Unsupervised Person Image Synthesis in Arbitrary Poses. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8620–8628.

Radford, A., Metz, L., and Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434.

Ramamoorthi, R. and Hanrahan, P. (2001). A signal-processing framework for inverse rendering. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '01, pages 117–128, New York, NY, USA. ACM.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Rothe, R., Timofte, R., and Van Gool, L. (2018). Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, 126(2-4):144–157.

Rudovic, O., Pantic, M., and Patras, I. (2013). Coupled gaussian processes for pose-invariant facial expression recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 35(6):1357–1369.

Ruiz, N., Chong, E., and Rehg, J. M. (2018). Fine-grained head pose estimation without keypoints. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Salimans, T., Goodfellow, I. J., Zaremba, W., Cheung, V., Radford, A., and Chen, X. (2016). Improved techniques for training gans. *CoRR*, abs/1606.03498.

Sanchez, E. and Valstar, M. (2020). A recurrent cycle consistency loss for progressive face-to-face synthesis. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) (FG)*, pages 109–116, Los Alamitos, CA, USA. IEEE Computer Society.

Sanchez, E. and Valstar, M. F. (2018). Triple consistency loss for pairing distributions in gan-based face synthesis. *CoRR*, abs/1811.03492.

Sandler, M., Howard, A. G., Zhu, M., Zhmoginov, A., and Chen, L. (2018). Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. *CoRR*, abs/1801.04381.

Savov, N., Ngo, M., Karaoglu, S., Dibeklioglu, H., and Gevers, T. (2019). Pose and expression robust age estimation via 3d face reconstruction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*.

Sengupta, S., Kanazawa, A., Castillo, C. D., and Jacobs, D. W. (2018). Sfsnet: Learning shape, reflectance and illuminance of faces 'in the wild'. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shen, W., Guo, Y., Wang, Y., Zhao, K., Wang, B., and Yuille, A. L. (2018). Deep regression forests for age estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2304–2313.

Shen, Y., Gu, J., Tang, X., and Zhou, B. (2020). Interpreting the latent space of gans for semantic face editing. In *CVPR*.

Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., and Sebe, N. (2019). First order motion model for image animation. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Sun, L., Qiu, S., Li, Q., Liu, H., and Zhou, M. (2017). Age estimation via pose-invariant 3d face alignment feature in 3 streams of cnn. In *Pacific Rim Conference on Multimedia*, pages 172–183.

Suwajanakorn, S., Seitz, S. M., and Kemelmacher-Shlizerman, I. (2017). Synthesizing obama: Learning lip sync from audio. *ACM Trans. Graph.*, 36(4).

Tewari, A., Zollhöfer, M., Garrido, P., Bernard, F., Kim, H., Pérez, P., and Theobalt, C. (2018). Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 2549–2559. IEEE Computer Society.

Tewari, A., Zollöfer, M., Kim, H., Garrido, P., Bernard, F., Perez, P., and Christian, T. (2017). MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. In *The IEEE International Conference on Computer Vision (ICCV)*.

Thies, J., Zollhöfer, M., Nießner, M., Valgaerts, L., Stamminger, M., and Theobalt, C. (2015). Real-time expression transfer for facial reenactment. *ACM Transactions on Graphics (TOG)*, 34(6).

Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., and Nießner, M. (2016a).

Face2face: Real-time face capture and reenactment of rgb videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2387–2395.

Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., and Nießner, M. (2016b). Face2Face: Real-Time Face Capture and Reenactment of RGB Videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395.

Vrij, A., Fisher, R. P., and Blank, H. (2017). A cognitive approach to lie detection: A meta-analysis. *Legal and Criminological Psychology*, 22(1):1–21.

Vrij, A., Hartwig, M., and Granhag, P. A. (2019). Reading lies: Nonverbal communication and deception. *Annual Review of Psychology*, 70(1):295–317. PMID: 30609913.

Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., and Catanzaro, B. (2018). High-resolution image synthesis and semantic manipulation with conditional gans. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8798–8807.

Wolf, L., Hassner, T., and Maoz, I. (2011). Face recognition in unconstrained videos with matched background similarity. In *CVPR*, pages 529–534. IEEE Computer Society.

Wu, P., Liu, H., Xu, C., Gao, Y., Li, Z., and Zhang, X. (2017a). How do you smile? towards a comprehensive smile analysis system. *Neurocomputing*, 235:245–254.

Wu, W., Zhang, Y., Li, C., Qian, C., and Loy, C. C. (2018a). Reenactgan: Learning to reenact faces via boundary transfer. In *ECCV*.

Wu, X., He, R., Sun, Z., and Tan, T. (2018b). A light cnn for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11):2884–2896.

Wu, Z., Singh, B., Davis, L. S., and Subrahmanian, V. S. (2017b). Deception detection in videos. *CoRR*, abs/1712.04415.

Xu, W., Keshmiri, S., and Wang, G. (2019). Adversarially approximated autoencoder for image generation and manipulation. *IEEE Transactions on Multimedia*, 21(9):2387–2396.

Yan, W.-J. and Chen, Y.-H. (2018). Measuring dynamic micro-expressions via feature extraction methods. *Journal of Computational Science*, 25:318–326.

Yang, C. and Lim, S.-N. (2019). Unconstrained facial expression transfer using style-based generator. *arXiv preprint arXiv:1912.06253*.

Yang, S., Luo, P., Loy, C. C., and Tang, X. (2016). Wider face: A face detection

benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yang, T.-Y., Huang, Y.-H., Lin, Y.-Y., Hsiu, P.-C., and Chuang, Y.-Y. (2018). Ssr-net: A compact soft stagewise regression network for age estimation. In *International Joint Conference on Artificial Intelligence*, pages 1078–1084.

Yang, X., Gao, B.-B., Xing, C., Huo, Z.-W., Wei, X.-S., Zhou, Y., Wu, J., and Geng, X. (2015). Deep label distribution learning for apparent age estimation. In *IEEE International Conference on Computer Vision Workshops*, pages 102–108.

Yang, Z. and Ai, H. (2007). Demographic classification with local binary patterns. In *International Conference on Biometrics*, pages 464–473.

Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., and Huang, T. S. (2018). Generative image inpainting with contextual attention. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5505–5514.

Zakharov, E., Ivakhnenko, A., Shysheya, A., and Lempitsky, V. (2020). Fast bi-layer neural synthesis of one-shot realistic head avatars. In *European Conference of Computer vision (ECCV)*, pages 524–540.

Zakharov, E., Shysheya, A., Burkov, E., and Lempitsky, V. (2019). Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9459–9468.

Zhang, K., Gao, C., Guo, L., Sun, M., Yuan, X., Han, T. X., Zhao, Z., and Li, B. (2017a). Age group and gender estimation in the wild with deep ror architecture. *IEEE Access*, 5:22492–22503.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*.

Zhang, Z., Song, Y., and Qi, H. (2017b). Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Zhao, B., Lu, H., Chen, S., Liu, J., and Wu, D. (2017). Convolutional neural networks for time series classification. *Journal of Systems Engineering and Electronics*, 28(1):162–169.

Zhou, H., Hadap, S., Sunkavalli, K., and Jacobs, D. W. (2019). Deep single-image portrait relighting. In *The IEEE International Conference on Computer Vision (ICCV)*.

Zhou, Y., Jiang, R., Wu, X., He, J., Weng, S., and Peng, Q. (2019). Branchgan: Unsupervised mutual image-to-image transfer with a single encoder and dual decoders.

IEEE Transactions on Multimedia, 21(12):3136–3149.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232.

Zhu, X., Lei, Z., Yan, J., Yi, D., and Li, S. Z. (2015). High-fidelity pose and expression normalization for face recognition in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 787–796.

Epilogue

So remember this principle when something threatens to cause you pain:
the thing itself was no misfortune at all;
to endure it and prevail is great good fortune.
— Marcus Aurelius

Human life can be compared to a fairy tale. An imperfect hero, who out of curiosity and boredom, decides to go on an adventure to explore the world. On his journey, he is helped by other heroes. He overcomes his fears and limitations, fights injustice, and gets positive rewards. The story I am writing is full of luck, an unconventional starting location, and amazing people that I have met and who supported me along the way, to whom I am forever thankful.

If somebody asked me 5 years ago what my career plan was, as I was looking at significantly brighter PhDs around me, I would have said that being a researcher was not on the list. But looking back, I am very grateful that my path led me in that direction. Like one might accidentally learn some French by reading Leo Tolstoy, I have gained a lot of unexpected and exciting experiences during my PhD.

I would like to thank my promoter Prof. Theo Gevers, who put trust in me and provided me with the opportunity to pursue the academic career. I am very grateful to be taught by you during the master's program. Thank you for your guidance, knowledge, and support during my time in the Netherlands. I have learned a lot from you as a person. To not leave the game if there is still a chance to win. To see the forest for the trees. Thank you for providing me the freedom and flexibility to balance my academic and non-academic life.

Many thanks to my co-promoter Dr. Sezer Karaoğlu. Thank you for showing the importance of helping other scholars, as I have been helped by others. Thank you for being a friend, for your constant support and guidance, and for challenging my points of view. From you, I have learned to accept more my failures, to go with the flow, and

recognize the fact that many things in life are not under my control. Also, thank you for organizing BBQs! The best BBQs are definitely Turkish!

I would like to thank Prof. Peter van Emde Boas, Prof. Albert Salah, Prof. Elmar Eisemann, Prof. Evangelos Kanoulas, Dr. Stevan Rudinac, and Dr. Pascal Mettes, who agreed to be a part of my defense committee despite their busy schedules, and for reviewing my thesis. In addition, I would like to thank Evangelos and Stevan who taught me Information Retrieval and Information Visualization during the masters program. During these courses, I have acquired knowledge, which has remained relevant until today. The hard-working attitude of Stevan was inspirational. When I was hanging around the midnight third floor during my first year, I have frequently been entertained by the wonderful violin music of Shuai Liao and Stevan, who were out for coffee.

I would like to thank Dr. An Le and Riaan Zoetmulder who kindly agreed to be my paranymphs. I would also thank Riaan for getting me into the gym, and for giving Jonathan Israel's book for light reading. I thank An for valuable conversations in the cantine. I would like to thank Peter Dekker for helping with the Dutch translation of the thesis summary. In addition, I have learned a lot from Peter about Dutch culture. Furthermore, I would like to thank Elena Ponomareva, Bohdan Rybak, and Dr. An Le who spent their valuable time proofreading the thesis draft. Many thanks to Dr. Anıl Baslamisli, Partha Das, and Dr. An Le who helped with proofreading paper submissions. I would like to thank Finde Xumara for helping me fix the cover page and the thesis layout. I am so lucky to have a UI/UX genius sitting next to me at the office! Finally, I would like to thank my co-authors Christian aan de Wiel, Nedko Savov, Burak Mandira, Wei Wang, and Dr. Hamdi Dibeklioğlu without whom the publications in this thesis would not have existed at all.

Many thanks to my colleagues at 3DUniversum and UvA I had the chance to talk to or collaborate with, these colleagues are; Dr. Anıl Baslamisli (your SoundCloud tracks are amazing, Qt Sessions Radio Show - the best deep house music ever!), Dr. Yang Liu (your hot pot is delicious!), Sjoerd Dijkstra (thank you for showing me homebrew and git pipeline!), Morris Franken (you're a inkscape-gimp-blender wizard, Morris!), Giuseppe Cilli, Masoumeh Bakhtiariziabari, Rick Groenendijk, Dr. Hanan ElNaghy, Gjorgji Strezoski (many thanks for helping with a poster!), Emiel Hoogeboom, Dr. Shuai Liao, Dr. Noureldien Hussein (I have learned a lot from you on the train to Den Haag!), Dr. Berkay Kicanaoğlu (thank you for TAing me CV1 labs!), Mert Kilickaya, Tushar Nimbhorkar, Shaojie Jiang (thank you for being a kind housemate!), William Thong (Japanese food was nice indeed!), Yunlu Chen (I am amazed with your math skill!),

Zhiwei Ai, Ozzy Ülger, Wei Zeng, Partha Das (thank you for sharing your experience about India!), Wei Wang, Jian Han, Yahui Zhang, Dr. Deepak Gupta, Kirill Gavrilyuk (the only person with whom I could practice Russian at the university), Jiaojiao Zhao, Shuo Chen, Zhiwei Ai, Kien Nguyen. It was a fun experience running the Computer Vision 2 lab sessions for 4 years together with many of you! Collaboration with Jian and Anıl didn't lead to the final product, nevertheless, it was a valuable experience! Many Thanks to the senior Computer Vision lab staff; Dr. Leo Dorst, Dr. Arnoud Visser, Dr. Shaodi You, Dr Dennis Koelma, Dr. Thomas Mensink, Dr. Emrah Bostan, and Dr. Jan-Mark Geusebroek. Many thanks to Dr. Maarten van Someren for his kindness, support, and guidance through the difficulties of the first master's years at UvA.

Special thanks to Dennis and Morris, without whom all servers would have died, and to Leo, who is an inspiration on what a real researcher should be like. Many thanks to Leo for his books, which he left in the CV lab, where I had an opportunity to read them.

I thank the master students, who I had the privilege to supervise during my PhD time. I have learned a lot from you. I would like to thank Axel Bremer, Mattijs Blankesteijn, Tim de Haan, Christian aan de Wiel, Ipek Ganiyusufoğlu, and Filip Pandža.

I am very thankful to people, with whom I enjoyed the meaningful discussion, who challenged my worldview, and made me a better person. I thank the wonderful teachers in maths and physics, which I have had at the university and school.

Many thanks to Dmitry Kostyanoy, Elena Ponomareva, Vera Kostyanaya, Phuong Vu, Lam Quach, Nguyen Tran, Mai Nguyen, Anna Kravets, Dennis Hutten, who were generous enough to host me during my time away from research. Thank you for your hospitality, delicious food, and company. I thank Arthur Bražinskas, Dang Vu, Tho Nguyen, Hien Le, Andriy Bychkovskyi, Mai Anh Nguyen, Thien Dong, Duy Ly, Yen Le and Anna Kornetska for inspiring conversations and their contribution to my understanding of the world. I was very fortunate to have a chance to collaborate closely with Peter, Riaan, Arthur, and Elena. Thank you for being reliable teammates and covering my back! I thank ex-PhD scholars, Dr. Thang Pham, Dr. Duong Vu, Dung Chu, and Dr. Cuong Dinh, for sharing their wisdom in pursuing PhD.

I thank Dr. Michael Zollhöfer, Dr. Justus Thies, Ayush Tewari, and Dr. Lucas Theis for answering questions with regards to their papers.

I thank Dr. Andrew Ng, who introduced me to Artificial Intelligence. I thank Dr. Kostyantyn Kharchenko, who introduced me to Computer Vision. I thank Prof. Thomas Vetter and Dr. Marcel Lüthi, with whom I was fortunate to talk and have lunches in Basel. Prof. Thomas Vetter was not only a great teacher but also a great tour guide.

Epilogue

I thank all the authors whose books I have read, they are my teachers too. I would like to thank Prof. Christopher Bishop, Prof. Niklaus Wirth, Dr. Brian Kernighan, Dr. Dennis Ritchie, Larry Wall, and Rob Pike. I thank Orwell, Remarque, Tolstoy, Gogol, de Saint-Exupéry, Puzo, Doyle, Rowling, and many others I couldn't name in this acknowledgment. I thank them for their knowledge and wisdom, for their heroes and villains, for their tragedies and happy endings.

I thank Anders Hejlsberg. His Turbo Pascal 5.0 introduced me to programming, and to contributors of Fedora, who created a reliable operating system, which robustly worked for 4 years despite not receiving any "dnf update".

I would like to thank my sister Phuong. You got into a better uni (I am sorry, UvA), I am really proud of you! Most importantly, I would like to thank my mom and dad, who constantly reminded me of the existence of other more important things in life outside of my research and work.

Finally, I thank all of my friends, my extended family members, from whom I received constant support. I thank all people who taught me, showed me my limitations, and enriched my worldview during my lifetime.

Thank you, дякую, спасибо, cảm ơn, dank je voor je hulp. I am very grateful to you for your help and support.

Haarlem, 12 March 2021 Minh Ngô