Towards an Intelligent Agent for the Human Face: Improving the Accuracy, Controllability, and Explainability of 3D Face Reconstruction

William Rowan

Doctor of Philosophy

Computer Science
University of York

November 2024

Abstract

The human face is a critical cue for human interaction, playing an essential role in recognition, communication, and even medical diagnosis. However, 3D face reconstruction from a 2D image is an ill-posed problem, with existing approaches struggling due to natural variation in facial appearance and limited 3D data.

This thesis addresses these limitations by proposing the Intelligent Face Agent (IFA), which envisions a new form of computational interaction with human faces. The IFA accepts multi-modal inputs and offers intuitive, text-driven manipulation and explanation of 3D facial reconstructions. We use this concept to motivate the design and implementation of complementary components in 3D face generation and analysis.

First, we introduce the SynthFace Generator, a novel approach for fast, large-scale 2D-3D paired dataset generation. This method eliminates the need for manual asset creation, producing photorealistic face images with paired 3D shapes. We use this method to create SynthFace, the largest paired 2D-3D dataset for human face shape.

Secondly, we present Text2Face, the first method to enable the direct and complete initialisation of 3D face models from textual descriptions. This enhances the controllability of 3D face reconstruction, offering the opportunity to improve practical applications such as avatar creation.

Finally, we introduce new baselines for evaluating 3D face reconstruction methods. We propose OptiFaces, a novel baseline that assesses the performance achievable by accurately classifying a set of well-distributed reference faces, providing a more meaningful interpretation of reconstruction error. Additionally, we introduce "N Heads Are Better Than One", a new approach for evaluating combinations of existing 3D face reconstruction methods, resulting in a range of robust new baselines for 3D face reconstruction.

The research presented improves the accuracy, controllability, and explainability of 3D face reconstruction, paving the way for broader adoption and application in fields such as healthcare, security, and the creative industries.

Keywords: 3D Morphable Face Models, 3D Face Reconstruction, Multi-modal Representation Learning.

Email: will.rowan@york.ac.uk, wrowan46@gmail.com
Web: http://w-rowan.github.io

Acknowledgements

I would like to express my sincere thanks to my supervisors: Dr. Patrik Huber, Prof. Nick Pears, and Prof. Andrew Keeling. This thesis and my development as a researcher are a testament to their advice and input throughout my studies and research. In particular, I want to acknowledge Nick for introducing me to Computer Vision, as the supervisor of my undergraduate and Master's projects before my PhD. I would further like to acknowledge the UK Engineering and Physical Sciences Research Council who funded my PhD with grant EP/T518025/1.

Most of all, I would like to thank my family, who have supported me throughout my PhD and everything else. I would not have completed this without their support.

Declaration

I declare that this is a presentation of my original work and I am the sole author. This work has not previously been presented for award at this, or any other, university. All sources are acknowledged as References. The pronoun "we" is used to refer to the author and the reader collectively. The research presented in this thesis has resulted in the following publications, corresponding to chapters 3, 4, and 5 respectively:

- Will Rowan, Patrik Huber, Nick Pears, and Andrew Keeling. Fake it without making it: Conditioned face generation for accurate 3D face shape estimation. 2024. **On ArXiV**. [RHPK23a]
- Will Rowan, Patrik Huber, Nick Pears, and Andrew Keeling. Text2face: 3D Morphable Faces From Text. 2023. **In ICLR Tiny Papers**. [RHPK23b]
- Will Rowan, Patrik Huber, Nick Pears, and Andrew Keeling. How Many OptiFaces? A New Evaluation Metric For 3D Face Reconstruction. 2024. **In ICLR Tiny Papers**. [RHPK24a]
- Will Rowan, Patrik Huber, Nick Pears, and Andrew Keeling. N Heads Are Better Than One: Exploring Theoretical Performance Bounds of 3D Face Reconstruction Methods. 2024. **In ECCV Workshop: Foundation Models for 3D Humans**. [RHPK24b]

Signed 11 November 2024, Will Rowan

Contents

List of Figures **xiii**

1 Introduction **1**

1.1 The Intelligent Face Agent 4

1.2 Thesis and Research Questions 6

1.3 Structure 8

1.4 Contributions 10

1.5 Publications 11

2 Related Work **13**

2.1 3D Face Reconstruction 14

2.1.1 The 3D Morphable Model 17

2.1.2 Editing 3DMMs Through Text 19

2.1.3 Facial Reconstruction Fitting Approaches 19

2.2 Evaluation of 3D Face Reconstruction 28

2.2.1 Evaluation Datasets 29

2.2.2 Evaluation Protocols 30

2.3 Controllable Generation of the Human Face 31

2.3.1 Realistic Parameterised Faces 32

2.3.2 Paramaterised 3D Face Datasets for Face-Related Tasks 33

2.4 Critical Analysis of the Literature 34

3 Fast 2D-3D Face Dataset Generation 37

3.1 The SynthFace Generator 40

3.1.1 3D Face Model 42

3.1.2 3D Hair Database 43

3.1.3 Depth Map Generation 44

3.1.4 Conditioned Face Generation 44

3.1.5 Prompt Selection 46

3.1.6 Dataset Demographics 47

3.2 ControlFace for 3D Face Reconstruction 48

3.2.1 Training Data 50

3.2.2 Pre-processing 50

3.2.3 Mapping Network 51

3.2.4 Training Strategy 51

3.3 Experiments and Evaluation 52

3.3.1 The SynthFace Generator 52

3.3.2 Landmark Accuracy of Dataset Generation 54

3.3.3 NoW Benchmark 54

3.3.4 Experimental Setup for ControlFace 55

3.3.5 Discussion and limitations 56

3.4 Summary 58

4 Text-based Initialisation of 3D Face Models 61

4.1 Proposed Method: Text2Face 63

4.1.1 Dataset Generation 63

4.1.2 Mapping Network Architecture 65

4.2 Experiments 66

4.2.1 Qualitative Results 66

4.2.2 Multi-Modal Fitting 70

4.3 Conclusion 71

5 New Baselines for 3D Face Reconstruction 73

5.1 Proposed Method: OptiFaces 75

5.2 OptiFace Optimiser 78

5.2.1 Selecting Representative Faces 78

5.2.2 Idealised Discrete Classifier: Evaluating OptiFaces on a Benchmark 79

5.3 Experiments 81

5.4 Visualisation of OptiFaces 82

5.5 3D Face Reconstruction Model Zoo 86

5.5.1 Combining Existing 3D Face Reconstruction Methods 87

5.5.2 Results 88

5.5.3 Relative Performance: Plot Analysis 89

5.5.4 Decrease in Errors Over Time 93

5.6 Conclusion 94

6 Conclusions & Outlook 97

6.1 Revisiting the Thesis and Research Questions 97

6.2 Personal Reflections 102

6.3 Future Work 103

A SynthFace Supplementary 105

A.1 Textual Appearance Descriptor 105

Bibliography 107

List of Figures

1.1 **An Intelligent Face Agent** Through the design of an Intelligent Face Agent, we envisioned a new way of interacting with the human face, incorporating cues from multiple modalities to guide reconstruction and interpretation. 4

2.1 A typical 3D face reconstruction problem setup. We consider the case of reconstructing a 3D mesh representation of the human face from various input modalities, potentially including text, image(s), and/or partial 3D information. The Processing step can take many forms which we discuss throughout this chapter, including that of a deep regression network that predicts parameters of a 3DMM given an input vector. . 15

2.2 Recovery of 3D shape from RGB pixels is a highly ambiguous problem. From a single image, we show highly different patches of pixels in the presence of self-occlusions (shadows) and fine details such as creases in the skin. We also show how simple factors such as the presence of RGB noise and reduced exposure impact the perceived appearance of the same individual, for which we aim to recover the same underlying 3D shape. 16

2.3 The 2009 Basel Face Model with shape and texture components plus/minus five standard deviations σ [PKA+09]. This variation is indicative of 3D Morphable Models of the human face. 18

2.4 A framework for the classification of facial reconstruction approaches. 20

2.5 Photometric loss E is calculated as the sum of squared differences between the intensity values of the input image $I(x, y)$ and the image generated by the 3DMM $M(x, y)$ at each pixel coordinate (x, y). The operation $\| \cdot \|$ denotes the Euclidean norm, here simplified to the square of the difference, providing a measure to optimise the reconstruction by minimising these differences. 21

2.6 MoFA: Architectural Diagram [TZK+17] 24

2.7 MVF-Net: Architectural Diagram [WBC+19]. 27

3.1 We present SynthFace, the largest available dataset of photorealistic face images and 3D Morphable Model (3DMM) shape parameters. We sample head shape, hair style, custom appearance descriptors, and camera intrinsics to create this dataset of unparalleled diversity and photorealism which requires no real data or crafted assets. 39

3.2 The SynthFace Generator. We sample from a 300-dimensional shape vector, 50-dimensional expression vector, and a 6-dimensional pose vector. We input these to the FLAME decoder to produce a 3D mesh. From this mesh, we extract a depth map that, alongside our textual appearance descriptor, is used as conditioning to generate a photorealistic face. . . . 41

3.3 The SynthFace Dataset includes different poses, perspective projections, and visual identities for each unique 3D shape. Here we present three images within SynthFace, each conditioned on the same 3D shape (including hair) and textual appearance descriptor. 46

3.4 Intermediary prompting result 1 prompt: 'a woman' negative prompt: 'worst quality, normal quality, low quality, low res, blurry, jpeg artifacts, error, sketch , monochrome, geometry' 47

3.5 Intermediary prompting result 2 prompt: 'a woman, profile picture' negative prompt: 'worst quality, normal quality, low quality, low res, blurry, jpeg artifacts, error, sketch , monochrome, geometry' comment: much more realistic with profile picture prompt. 47

3.6 Intermediary prompting result 3 prompt = 'a woman, profile picture, dslr' negative prompt: 'worst quality, normal quality, low quality, low res, blurry, jpeg artifacts, error, sketch , monochrome, geometry' comment: worse - added expression artifact. 48

3.7 The SynthFace Dataset. We present a selection of typical images from the SynthFace dataset. These further demonstrate the diversity and photorealism achieved using our proposed generation method. In the bottom row, we also highlight the most common weaknesses in images generated using Stable Diffusion: chiefly minor issues related to the generation of hair, including issues with colour, positioning, and it being mistaken for another material. 49

3.8 ControlFace training. We train the mapping network within ControlFace on the SynthFace dataset. It is trained to minimise the mesh reconstruction error between a predicted 3D mesh and known 3D mesh for each image in SynthFace. ControlFace at inference is shown outlined. ControlFace accepts an image as input, aligns it, and calculates an ArcFace embedding from this aligned detected face. A mapping network converts this ArcFace embedding to 3DMM parameters. The FLAME decoder generates a full head mesh from these parameters. 50

3.9 Face Synthetics qualitative comparison. Three randomly selected images from Face Synthetics (top) and SynthFace (bottom) are viewed here. Our textual appearance descriptor enables the generation of typical occluders such as glasses. Previously, this required manual asset creation by digital artists as is the case for Face Synthetics. 53

3.10 Example image from SynthFace, showing the projected 2D landmarks from the 3D head mesh on both the depth map and generated image. 55

4.1 Dataset generation, model training, and inference for Text2Face. 64

4.2 The Text2Face Dataset. An image and its corresponding 3D representation, estimated by DECA [FFBB21], are shown. . 65

4.3 Architectural diagram of our Text2Face regressor. 66

4.4 *(l)*: Prompt: “20 year old woman looking at the sky with surprise at UFO overhead". *(r)*: Prompt: “50 year old man looking grumpy". Each sub-figure, from left to right: Shape generated by the text prompt, DALL-E image from the same prompt, textured mesh. 66

4.5 Prompt: “Happy man". 68
4.6 Prompt: “20 year old woman squinting at the sun". 68
4.7 Prompt: “24 year old attractive man". 68
4.8 Prompt: “Photo of a woman screaming". 69
4.9 Prompt: “Photo of a woman with her eyes closed". 69
4.10 Prompt: “50 year old man looking happy after a long day working on the film set". 69
4.11 Prompt: “50 year old woman". 69
4.12 Robert De Niro fit to a 3DMM using Text2Face with the CLIP embedding extracted from each image as input: using the original image (left), a sketch (middle), and an engraving (right). 70

5.1 A visualisation of the first two principal components of FLAME shape for the first 10 OptiFaces calculated using the Headspace dataset. OptiFaces are numbered in the order they are calculated. 83
5.2 A comparison of the OptiFaces computed using the proposed OptiFace Optimiser and a k-means implementation. 84
5.3 OptiFace Heads Visualisation for the first 10 OptiFaces calculated using the OptiFace Optimiser 85
5.4 N Heads Are Better Than One Architecture. We evaluate multiple existing 3D face reconstruction methods (such as MICA, DECA, and TokenFace) for each input image. Our theoretical classifier selects the best reconstruction, enabling us to calculate lower error bounds for combinations of methods and establish new baselines for 3D face reconstruction performance. 86

5.5 Scatter plot of mean errors for MICA and TokenFace across all images in the NoW benchmark in metric reconstruction. The red dashed line represents the line $y = x$. Points below the line indicate images where TokenFace performs better than MICA, while points above the line indicate images where MICA performs better than TokenFace. 90

5.6 Error plots for the NoW benchmark in non-metric reconstruction. We compare errors from two leading approaches and the combination of all methods grouped by their year of publication. 18 methods are considered in total for the 80 meshes included within the test set. 92

5.7 Error plots for the NoW benchmark in metric reconstruction. We compare errors from two leading approaches and the combination of all methods grouped by their year of publication. 14 methods are considered in total. 93

5.8 Error plots for 3D face reconstruction: (a) non-metric reconstruction, and (b) metric reconstruction. Both plots compare MICA alone to combinations of MICA with other methods (DICA, TICA, and FICA), demonstrating potential reductions in error from their combination. 94

Chapter 1

Introduction

The human face is one of the most important cues for human interaction. It can be used to recognise a friend, communicate a feeling, and in certain cases, be used to diagnose medical conditions such as craniosynostosis, where the joints in a baby's skull fuse prematurely, resulting in a misshapen appearance [DPD17]. As humans, we have a learned understanding of the human face, including natural variation in shape and appearance, alongside the impact of changes in perspective, illumination, and occlusions. We recognise friends when they're wearing sunglasses or are partially illuminated, we can visualise them from a few words of description, and we'll still recognise their key visual features from their baby photos to their 80th birthday.

When we take a photo of a friend, our digital camera projects the 3D world to pixels on a screen. In doing so, we compress the dimensionality of the real world into a lower-dimensional representation. In the field of 3D shape estimation, we seek to recover the richer representation of the 3D world from the 2D captured image, recovering a higher-dimensional representation in the process.

Existing Computer Vision approaches for 3D shape estimation struggle in the presence of such natural variation, and they do not offer natural ways to interact with the most human of concepts: our own faces. This thesis proposes methods to do better; to improve the accuracy, controllability, and

explainability of 3D shape estimation of the face. Specifically, 'accuracy' refers to the precision with which the reconstructed models match the actual human faces in shape and detail. 'Controllability' means the ability to manipulate and adjust the reconstructed models easily and intuitively, such as through natural language. 'Explainability' involves the capacity of our methods to provide clear, understandable measures for the outcomes of the reconstruction process. For the purposes of this thesis, we use 3D shape estimation and 3D face reconstruction interchangeably.

Despite its limitations, 3D face shape estimation has proven to be an enabling technology with applications in healthcare [MPS+11], security [BAHS06], and the creative industries [TZS+16]. In this context, our faces are becoming increasingly powerful tools, whether through unlocking our phones or being put directly into a video game after a quick scan using those same devices. However, our computational understanding of the human face remains limited. Despite possessing many sources of knowledge about the human face - whether in natural descriptions, 2D photos, or 3D models - our representations are disconnected and difficult to initialise and interpret.

These limitations are hindering the broader adoption of 3D reconstruction techniques in areas where they could be highly beneficial, such as in 3D analysis for medical diagnosis. Currently, we are unable to interact naturally with the latent representations of these models, for example, through language. This lack of intuitive interaction places a significant knowledge burden on the user, thereby raising the barrier to widespread adoption. Moreover, the scarcity of large-scale 3D face datasets leads to reduced accuracy. This, in turn, diminishes our confidence that our reconstruction methods are both accurate and equitable across different user subgroups, particularly concerning protected characteristics such as race and gender. These challenges impede the full potential of 3D face recon-

struction techniques. In contrast, the development of large-scale 2D image datasets in recent years has revolutionised various medical tasks – including identification of retinopathy, detection of lung cancer, and classification of skin lesions. The research presented in this thesis aims to bring a similar revolution to 3D face reconstruction.

To achieve this, we propose the creation of an Intelligent Face Agent (IFA), a new vision for what we can achieve within the 3D face reconstruction community. This concept offers a new form of computational interaction with the human face which accepts multi-modal inputs (text, images, partial 3D scans) and offers intuitive, text-driven manipulation and explanation of 3D facial reconstructions. Throughout this thesis, we implement aspects of this envisioned Intelligent Face Agent, principally: paired 2D-3D dataset generation, 3D shape initialisation from text, and new baselines for 3D shape reconstruction. In doing so, this thesis increases the accuracy, controllability, and explainability of 3D face reconstruction.

In this introductory chapter, we present:

- Motivation for improving 3D face reconstruction, focusing on the clinical use case of automated prosthesis design.
- An introduction to the Intelligent Face Agent and the intuitions behind its design.
- An overview of how this thesis fulfils core parts of the proposed Intelligent Face Agent.
- The research questions we aim to answer through this thesis.
- The structure of the remaining chapters of this thesis.
- An overview of the novel contributions made to the field of 3D face reconstruction by this thesis.

1.1 The Intelligent Face Agent

We propose the Intelligent Face Agent, as depicted in fig. 1.1, as a conceptual research goal for the 3D face reconstruction research community. Its design is purposefully high-level and representation-agnostic, ensuring it remains focused on the functionality it offers and its potential benefits to key application areas. This design is founded on several core values, which we term: Multiplicity, Integration, and Flexibility. These values are, in turn, based on a number of fundamental intuitions.

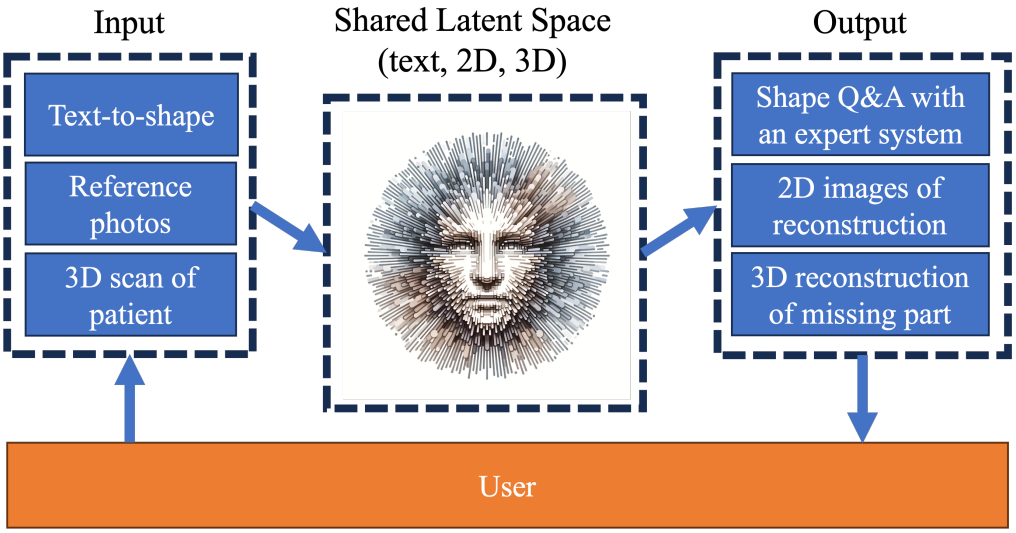


Figure 1.1: **An Intelligent Face Agent** Through the design of an Intelligent Face Agent, we envisioned a new way of interacting with the human face, incorporating cues from multiple modalities to guide reconstruction and interpretation.

Firstly, we assume that modalities of knowledge are inherently multiplicative: being able to connect 2D, 3D, and text-based knowledge of the human face will lead to improvements in tasks for all modalities. The interconnected design of the IFA, through a shared latent space, means that improvements in any one area of reasoning, such as in text-to-3D, will have multiplicative effects on other modalities too.

Secondly, this integrated representation offers opportunities for new

application areas. A single shared latent space for all available knowledge on human face shape and appearance means existing workflows can be simplified. In practice, this means that a face description provided by one individual can produce a reconstruction that can be edited by another individual. This face shape may then be queried using natural language descriptions.

Finally, multiple modalities offers flexibility to both tackle reconstruction problems from all input data modalities available to us and to view outputs in any form. This can allow non-experts to interact with high-dimensional face representations through text before passing a reconstruction to an expert practitioner.

We will illustrate the limitations of existing approaches and importance of the Intelligent Face Agent through the motivating example of prosthesis design. Facial prostheses offer a non-operative form of rehabilitation for patients suffering from facial trauma, offering life-changing physiological and aesthetic benefits. However, maxillofacial surgeons are rare, with two or fewer in 28 countries [MBH+23]. Maxillofacial prosthetists, who often also perform these treatments, also remain in very limited supply and high demand. An ageing worldwide population and commensurate higher incidence of facial cancer, will lead to increasing demand for effective treatment modalities [dCDSB+19].

Automated computer-aided reconstruction will improve access to this life-changing treatment globally. However, current clinical trials using 3D Morphable Models (3DMMs) [JCB+23] for prosthesis design lack fine control and integration of 2D data, such as patient photos. A critical lack of 3D human face data, exacerbated for rarely-occurring events such as facial trauma, limits the potential accuracy of automated solutions. The proposed Intelligent Face Agent would directly address these limitations.

Text-based interactions with 3D shape, as envisioned within the Intelligent Face Agent, will allow clinicians to both initialise and interrogate 3D reconstructions, saving maxillofacial prosthetists valuable time and enabling non-experts to diagnose conditions using expert models. Furthermore, incorporating a generative model across all available modalities will facilitate the creation of customised datasets tailored to specific clinical needs. This approach addresses the challenge of data scarcity, thereby enhancing the accuracy of reconstructions. In combination, these advancements will reduce treatment costs and increase global accessibility. Most importantly, they will enable reconstruction of the patient's prior face, not just a statistically plausible one, which will significantly improve patient outcomes.

This thesis focuses on three sub-tasks of the Intelligent Face Agent: dataset generation, 3D shape initialisation from text, and new baselines for the evaluation of 3D shape reconstruction. In doing so, we present research that improves the accuracy, controllability, and explainability of 3D face reconstruction. These areas are all of critical importance to the implementation of the Intelligent Face Agent and each make novel technical contributions to the research literature.

1.2 Thesis and Research Questions

Having introduced the research context and motivations for this work, our thesis research question can be summarised as follows:

Thesis
Can we connect existing sources of knowledge about the human face to improve the accuracy, controllability, and explainability of 3D face reconstruction?

We decompose this high-level research question into several related research questions which this thesis addresses.

Research Question 1

Can we exploit knowledge about the structure and appearance of the human face, contained within pretrained image generation networks, to generate large-scale datasets for 3D face reconstruction?

Addressed in Chapter 3

Research Question 2

What methods can be employed to increase the diversity of paired training data for 3D face estimation methods?

Addressed in Chapter 3

Research Question 3

Are we able to perform parameterised shape initialisation of faces from natural language descriptions?

Addressed in Chapters 3 and 4

Research Question 4

What new and more interpretable metrics can be developed for evaluating 3D face reconstruction?

Addressed in Chapters 4 and 5

Research Question 5

What theoretical lower bounds can we devise for 3D face reconstruction?

Addressed in Chapter 5

1.3 Structure

The remainder of this thesis is structured as follows:

Chapter 2, Related Work We introduce related work to support the technical chapters of this thesis. This chapter provides the necessary technical background and identifies existing gaps in knowledge that the subsequent technical chapters aim to address. We focus on parameterised representations of the face, 3D face reconstruction methods, and our existing methods for interacting with and evaluating these reconstructions.

Chapter 3, Paired 2D-3D Dataset Generation While large-scale 2D image datasets have revolutionised medical tasks, 3D applications have been limited due to a lack of paired 2D-3D data. In Chapter 3, we address the challenge posed by the lack of large-scale paired 2D-3D datasets of the human face. We present a dataset generation method for generating customised datasets of paired 2D-3D data for face-related tasks. This generative capability forms a key component of the proposed Intelligent Face Agent, and enhances the potential for analytic capabilities. Using this generation method, we create SynthFace, the largest available dataset of paired 2D-3D faces designed to be balanced by race and gender. Using this newly generated dataset, we developed and trained an accurate 3D face reconstruction method that learns to map a 2D image to its corresponding 3D shape.

Chapter 4, Text2Face: Text-based Initialisation of 3D Face Shape Current 3D representations of the face are difficult to initialise and edit, limiting their potential use by non-experts. For example, when given a verbal or written description of a person, many of us can visualise the shape and appearance of that person. However, computational methods for face reconstruction have been limited to modalities such as photos as input. In Chapter 4, we present the first method for direct and complete generation of

parameterised 3D faces from natural language descriptions. Unlike previous text-to-shape methods, our approach produces a complete 3D Morphable Model (3DMM) representation, enabling further shape adjustments to be made and interoperability with other reconstruction pipelines and parts of the Intelligent Face Agent.

Chapter 5, OptiFaces: A New Baseline for 3D Shape Evaluation 3D face reconstruction is a challenging problem, so much so that the mean face of existing datasets is highly competitive with recent learning-based approaches. In addition, existing evaluation methods and datasets are limited, with an approach achieving state-of-the-art performance on one benchmark often performing poorly on another. In Chapter 5, we introduce two methods for generating universal baselines for 3D face shape estimation: OptiFaces and the Model Zoo. With OptiFaces, we transform the continuous regression task of 3D face shape estimation to one of discrete classification between N face shapes. In doing so, we are able to compare existing methods with the task of accurately classifying N well-separated face shapes. The Model Zoo considers the theoretical lower error bound that can be achieved through combining two or more existing reconstruction methods. Both of these approaches enable intuitive judgements to be made on the degree of facial information learned by a reconstruction approach.

Chapter 6, Conclusion We evaluate our research contributions, as presented in chapters 3-5, against our original research questions. We also consider the progress made towards the Intelligent Face Agent, by ourselves and other researchers, throughout the course of the PhD. We note limitations of our current methods and present opportunities to fully realise the ambitions of the Intelligent Face Agent in future work.

1.4 Contributions

In this thesis, we make the following contributions to the field:

- **Chapter 3**:
  - Design a fast dataset generation method for synthesising paired 2D-3D face datasets with variations in camera pose, appearance, and lighting.
  - Show that this generation method is 20x faster than the existing state-of-the-art method while maintaining photo-realism.
  - Use this generation method to produce SynthFace, the largest available dataset of paired 2D-3D data for 3D face reconstruction.
  - Show that the use of a textual appearance descriptor can improve race and gender balance in generated 3D face datasets.
  - Demonstrate that accurate 3D face estimation can be achieved training exclusively on synthesised 2D-3D data. We show this by training a reconstruction network, ControlFace, on SynthFace and evaluating its performance on the NoW benchmark.
  - Demonstrate the accuracy of our paired 2D-3D data with landmark error evaluation by face region.
- **Chapter 4**:
  - Introduce a method to connect the embedding space of CLIP with the latent space of the FLAME head model, generating a dataset of paired (CLIP, FLAME) embeddings.
  - Train a deep MLP, Text2Face, on this dataset, enabling direct and complete generation of parameterised 3D faces from natural language descriptions.

  - Demonstrate the use of CLIP for multi-modal fitting - showing how we enable 3DMM fitting for the same subject represented as an image, sketch, and engraving.

- **Chapter 5**:

  - Introduce OptiFaces, a universal baseline for 3D face reconstruction. This approach considers the performance that can be achieved if we have a set of N reference faces and a classifier that optimally performs face matching from an input image to the closest of these N faces.
  - Propose a practical implementation of OptiFace calculation using a greedy algorithm that iteratively selects the face that minimises representative error.
  - Mathematically define an idealised discrete classifier as a classifier that always selects the closest OptiFace.
  - Implement a greedy algorithm for OptiFace selection, demonstrating improved face shape distribution in principal component space over the k-means algorithm.
  - Propose the 3D Face Reconstruction Model Zoo, a new way of visualising and understanding the theoretical performance of combining existing reconstruction methods.
  - Compute errors for these new theoretical ensemble methods on the NoW benchmark.

1.5 Publications

The research presented in this thesis has resulted in the following publications, corresponding to chapters 3, 4, and 5 respectively:

- Will Rowan, Patrik Huber, Nick Pears, and Andrew Keeling. Fake it without making it: Conditioned face generation for accurate 3D face shape estimation. 2024. **On ArXiV**. [RHPK23a]
- Will Rowan, Patrik Huber, Nick Pears, and Andrew Keeling. Text2face: 3D Morphable Faces From Text. 2023. **In ICLR Tiny Papers**. [RHPK23b]
- Will Rowan, Patrik Huber, Nick Pears, and Andrew Keeling. How Many OptiFaces? A New Evaluation Metric For 3D Face Reconstruction. 2024. **In ICLR Tiny Papers**. [RHPK24a]
- Will Rowan, Patrik Huber, Nick Pears, and Andrew Keeling. N Heads Are Better Than One: Exploring Theoretical Performance Bounds of 3D Face Reconstruction Methods. 2024. **In ECCV Workshop: Foundation Models for 3D Humans**. [RHPK24b]

Chapter 2

Related Work

In this chapter, we will cover all the work required to both understand the following technical chapters and put them in an appropriate context within the research literature. We will highlight key contributions to the literature in 3D facial modelling, reconstruction, and methods for controllable generation of the human face. More broadly, we will consider research that has improved the accuracy, controllability, and explainability of 3D face reconstruction — each of which ties to the research aims of this thesis.

More specifically, we will outline the following:

- A formulation of the problem of 3D face reconstruction, including its challenges and an overview of existing methods.
- Methods of representing the human face and sources of knowledge about the human face. We focus on the 3D Morphable Model.
- Datasets for 3D face reconstruction, which include those used for training as well as those employed for evaluation in common 3D face reconstruction benchmarks.
- Methods of generating our own datasets for 3D face reconstruction. This involves generating realistic parameterised 3D faces and corresponding 2D images.

In section 2.1, we will formally introduce the problem of 3D face reconstruction, considering its inherent challenges and how existing approaches tackle them. This naturally leads to a discussion of the 3D Morphable Model (3DMM) as both a common representation for 3D faces and as a prior form of knowledge to guide reconstruction. We will discuss further sources of knowledge in the form of Stable Diffusion and shared image-text latent spaces such as CLIP. In section 2.2, we will provide an overview of existing datasets and benchmarks for 3D face reconstruction. This includes how existing methods perform and how errors are calculated on the 3D face.

In section 2.3, we will consider methods for generating controllable representations of the human face, both in 2D and 3D. In particular, we will explore methods of generating large-scale paired 2D-3D face datasets for the training and evaluation of 3D face reconstruction methods. In section 2.4, we will critically analyse the work surveyed, putting it in the context of our proposed Intelligent Face Agent and the following technical chapters.

2.1 3D Face Reconstruction

The accurate reconstruction of 3D face shape from 2D images is a fundamental task in Computer Vision, with a wide range of applications across healthcare, creative, and security industries. In healthcare, 3D face reconstruction enables acupuncture point localisation and visualisation [LHCZ20], prosthesis design [MPS+11], and planning for craniofacial surgery [DPD+22]. In the creative industries, it facilitates real-time facial animation [CWLZ13], personalised avatar design [LSSS18], and live facial re-enactment [TZS+16]. In security, it supports generating 3D models of the face to match police photofits from eyewitness accounts [BAHS06], facial recognition [GMFB+18], and full 3D reconstruction from mugshot

photos [LTL+20]. A more complete overview of applications can be found in survey papers on 3DMMs [EST+20] and monocular 3D face reconstruction [ZTG+18]. It is important to note that as the accuracy, controllability, and explainability of 3D face reconstruction methods improve, the range of potential applications will further increase. Figure 2.1 shows a high-level view of this reconstruction problem: transforming a given input modality to a 3D representation of the face.

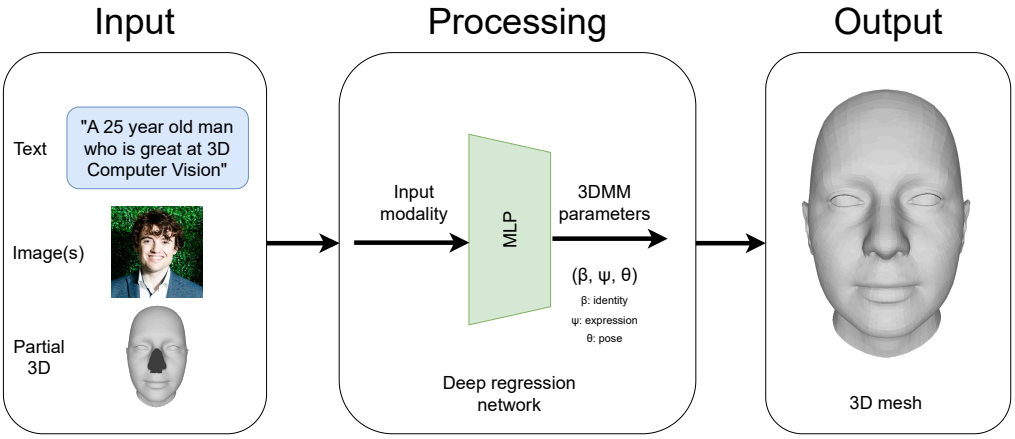


Figure 2.1: A typical 3D face reconstruction problem setup. We consider the case of reconstructing a 3D mesh representation of the human face from various input modalities, potentially including text, image(s), and/or partial 3D information. The Processing step can take many forms which we discuss throughout this chapter, including that of a deep regression network that predicts parameters of a 3DMM given an input vector.

However, estimating 3D face shape from a single image is fundamentally an ill-posed problem. For each RGB value in the image, we must disentangle the geometric shape from its illumination, albedo, and perspective effects. The impact of perspective transformation varies with camera distance, leading to a perspective face shape ambiguity: a 2D image of a face can be explained by many different underlying 3D shapes [BS19]. Accurate reconstruction is further complicated by the wide natural variation in face shape and appearance, including the presence of common occlusions such as hair and glasses. Combined, these factors lead to unresolvable ambiguities.

Crucially, for any of these factors affecting an image of an individual, we want to recover the same underlying 3D face shape. A non-exhaustive set of confounding factors that lead to ambiguity are considered in fig. 2.2.

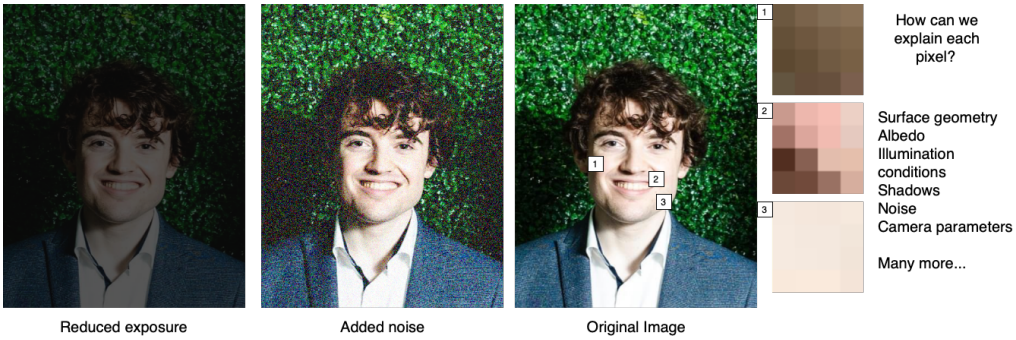


Figure 2.2: Recovery of 3D shape from RGB pixels is a highly ambiguous problem. From a single image, we show highly different patches of pixels in the presence of self-occlusions (shadows) and fine details such as creases in the skin. We also show how simple factors such as the presence of RGB noise and reduced exposure impact the perceived appearance of the same individual, for which we aim to recover the same underlying 3D shape.

Due to the challenging nature of the reconstruction problem, priors have been used to incorporate known information about the appearance and structure of the human face and head [KR18]. A prior is a form of knowledge that we possess about a certain problem before tackling it, which can be used to simplify the problem. For the task of 3D face reconstruction, this has traditionally taken the form of 3D priors of human head shape, with recent approaches considering the use of 2D generative image priors and shared text-image models [ATDN23].

Among 3D face priors, there are two principal categories of prior used to aid reconstruction: model-based and model-free. Model-based priors constrain facial shape to a low-dimensional representation of the face learnt from a number of facial scans. Model-free priors impose generic constraints on the reconstruction, such as spatial and temporal coherence [KR18].We will focus on model-based representations for this thesis, as they offer a strong additional constraint for the challenging task of 3D face reconstruction

from a single image. They are used as both a form of regularisation on reconstruction and as a compact parameterisation for the 3D face shape that we wish to reconstruct.

2.1.1 The 3D Morphable Model

In their seminal work, Blanz and Vetter [BV99] introduce the 3D Morphable Model (3DMM), a learned statistical model of the human face that considers both shape and texture. In this work, they also present an algorithm to reconstruct 3D faces from a single input image. In doing so, they set the foundation for the field of 3D face reconstruction and formalised the intuition that prior knowledge can help us solve ill-posed problems [EST+20].

The 3DMM, as originally proposed, is constructed as a parametric face space learned from a set of 3D scans of individuals. It is constructed by performing dimensionality reduction on a set of training meshes put into dense point-to-point correspondence [BV99]. Using Principal Components Analysis (PCA), a set of 'basis' faces are constructed; a linear combination of these basis faces is used to represent a face using the model. Shape and texture are separated into their own basis states, resulting in a shape vector, S, and a texture vector, T, being defined for each subject. Subsequent work added expression bases to allow the 3DMM to model facial expressions [EST+20].

This dense correspondence allows for a linear combination of faces to produce 'morphs' of the original model [EST+20], thereby allowing for greater interpretation and manipulation of reconstructions. The learned 3D Morphable Model (3DMM) can be utilised as a generative model, enabling novel faces to be generated by sampling from the model's parameter space. Semantically meaningful facial attributes can be mapped to the parameter space of the morphable model by labelling a set of example images with the desired attributes and then computing a weighted sum over these values

[BV99]. Combined, these properties make the 3DMM a powerful tool for facial analysis and generation.

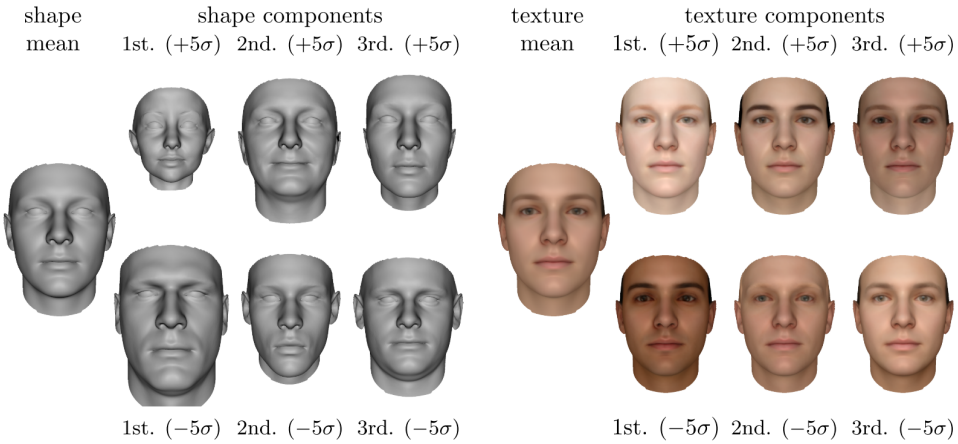


Figure 2.3: The 2009 Basel Face Model with shape and texture components plus/minus five standard deviations σ [PKA+09]. This variation is indicative of 3D Morphable Models of the human face.

A large number of 3D Morphable Models (3DMMs) have become publicly available since their introduction. Popular variants include the Basel Face Model [PKA+09], Liverpool-York Head Model (LYHM) [DPSD20a], and the FLAME head model [LBB+17]. Earlier morphable models, such as the 2009 Basel Face Model, are severely limited by the diversity of the participants used to create them; a predominantly white, young, and Western set of participants limits the applicability of the resulting model. Figure 2.3 shows a visualisation of this model. Booth *et al.*[BRZ+16] use a training dataset that includes much greater diversity in age and race than previous models to create their Large Scale Facial Model (LSFM). This enables the construction of both a global 3DMM and subgroup-specific models based on age, gender, and ethnicity [BRZ+16].

From modelling just the inner region of the human face, newer models such as FLAME [LBB+17] and the LYHM [DPSD20a] now model the complete human head. Models also differentiate themselves by including

texture and expressions, in addition to shape. The Universal Head Model (UHM) [PVO+20] combines two existing 3DMMs, the LYHM and LSFM, to create a new combined face-and-head shape model. They further integrate models of the ears, eyes, teeth, and tongue, which brings us closer to a single, complete 3DMM of the human head.

As representations of human faces, 3DMMs can be evaluated by their compactness, generalisation ability, and specificity [SRN+03]. Compactness measures the variability of the subject training data captured by the model, generalisation refers to the ability to explain faces outside of the training set, and specificity represents the ability to represent only valid faces.

2.1.2 Editing 3DMMs Through Text

CLIP (Contrastive Language-Image Pre-training) [RKH+21a] is a pre-trained visual-textual embedding model. This approach has shown strong zero-shot capabilities on a wide range of Computer Vision tasks and has enabled models to take advantage of this embedding space for downstream tasks, such as text-to-image generation [RPG+21a]. Several approaches have used the expressive power of CLIP to relate text to 3D shape. This includes text-driven generation of stylised meshes [MBOL+22]; general text to shape generation [SCL+22]; and full body 3D avatar creation and animation [HZP+22]. ClipFace [ATDN23] enables text-based texture and expression editing of pre-existing parameterised faces. However, it necessitates a parameterised face as input and maintains a fixed identity. None of these methods consider the generation of a fully parameterised model of the human face, including identity, from text alone.

2.1.3 Facial Reconstruction Fitting Approaches

The task of aligning a 3D face model to a 2D image involves estimating the parameters of the model that best resemble the image. In this con-

text, the specific method used to assess this resemblance distinguishes the various competing approaches. There are two distinct types of fitting methods to learn the parameters of a 3DMM: generative and regression-based approaches [TZK+17]. Generative approaches formulate a non-linear optimisation problem which seeks to minimise the difference between an input image and a representation of the 3D reconstruction, such as a rendering. Intuitively, a set of parameters are initialised for the 3D model and then iteratively updated to better fit the input image. Meanwhile, regression-based approaches directly regress the parameters from the input [ZYY+15].

We can now classify 3D face reconstruction approaches using the following proposed framework shown in fig. 2.4. This framework considers the type of image input, prior knowledge used, and the fitting approach employed. For simplicity, only image input is considered.

Consideration	**Approaches**	
Input	**Monocular:** Single image from a single viewpoint.	**Multi-view:** Multiple images from multiple viewpoints.
Prior	**Model-based:** Facial shape constrained to a low dimensional representation of the face.	**Model-free:** Generic (not specific to faces) constraints on the reconstruction.
Fitting Approach	**Generative:** Fit parameters of a face model by minimising the difference between input image and rendered image of parameterised 3D model.	**Regression-based:** Regress directly to model parameters from an image.

Figure 2.4: A framework for the classification of facial reconstruction approaches.

Generative Approaches: Analysis-by-synthesis

In their proposed algorithm for 3DMM fitting, Blanz and Vetter [BV99] introduced the analysis-by-synthesis framework. Analysis-by-synthesis formulates 3DMM reconstruction as an optimisation problem which minimises the difference between an image and the synthesised 3D model. This difference is quantified by a range of fitting energies, which alone or in combination, form the overall fitting term to be optimised. After initialisation, a new set of 3DMM parameters is iteratively computed from the current set using the specified energy minimisation scheme. Equation (2.1) shows an objective function which utilises a photometric loss.

$$E = \sum_{x,y} \|I(x,y) - M(x,y)\|^2 \tag{2.1}$$

Figure 2.5: Photometric loss E is calculated as the sum of squared differences between the intensity values of the input image $I(x,y)$ and the image generated by the 3DMM $M(x,y)$ at each pixel coordinate (x,y). The operation $\|\cdot\|$ denotes the Euclidean norm, here simplified to the square of the difference, providing a measure to optimise the reconstruction by minimising these differences.

The optimisation-based setting is sensitive to its initial conditions, as iterative optimisation proceeds from this selected initial state[WBC+19]. Furthermore, the optimisation may become trapped in local minima, making it both time-consuming and costly to minimise these errors. These challenges faced within the optimisation-based setting have motivated the development of regression-based approaches, which do not share the same reliance on initial conditions and may be able to avoid these traps.

Regression-Based Approaches

Supervised reconstruction Tran *et al.*[TTHMM17a] use a deep convolutional neural network (CNN) to directly regress the shape and texture

parameters of a 3DMM from a single image. They do this by creating a dataset of surrogate ground truth parameters estimated using multi-image optimisation-based reconstructions. A single-view CNN regressor is subsequently trained to learn this function. A rendering step is not required in this regressor, allowing for faster 3DMM estimation [TTHMM17a]. This demonstrates how advances in generative approaches can also improve regression-based approaches, enabling fast regressors to be trained using their outputs.

Richardson *et al.*[RSK16] generate face geometries directly from a 3DMM, rendering the face as an image under randomised lighting conditions. This results in a dataset of images with known 3DMM parameters; however these images are far from photorealistic. This points to a wider problem in synthesised approaches: a domain gap between synthesised and real data that makes generalisation difficult and task performance poor [KPL$^+$19].

In contrast, Wood *et al.*[WBH$^+$21] render highly realistic 3D face models for landmark localisation, demonstrating that synthesised data can be used to solve real world problems in the wild. Wood *et al.*[WBH$^+$22] build upon this work to train a dense landmark regressor for 702 facial points. A morphable model is fitted to these dense landmarks, leading to state-of-the-art results in 3D face reconstruction.

The success of this approach affirms the potential of network-based methods in advancing 3D shape estimation. However, this approach requires the manual creation of 3D assets with associated time, financial, and computational costs. Furthermore, the rendered images fall short of photorealism which limits their uses for direct 3DMM regression.

Other approaches have considered using the 3D data we have rather than relying on synthesised datasets. Zielonka *et al.*[ZBT22] annotate and unify existing 3D face datasets to enable supervised training of their

MICA (MetrIC fAce) network, achieving state-of-the-art performance on the NoW benchmark [SBFB19]. To do this, they use the facial embedding network, ArcFace [DGXZ19], to extract discriminative identity-bearing features from each input image. This network itself is trained using large 2D face image datasets, demonstrating that additional sources of knowledge can be effectively incorporated within 3D reconstruction pipelines. A mapping network, consisting of an MLP with three fully connected layers, then maps this identity embedding to the parameters of the FLAME head model.

This demonstrates the importance of supervision for reconstruction performance, even when supervised with minimal available data. However, this approach already represents the upper bound for supervised learning using 3D data, unless further data is collected. In combining eight existing datasets, they reach just 2315 individuals; this remains a small dataset for supervised learning techniques. Hence, a generative approach similar to Wood *et al.*[WBH+21] is required to train supervised approaches on larger datasets.

Other significant works in this field include exploring a hybrid loss function for weakly-supervised learning [DYX+19], generating surrogate ground truth data via multi-image 3DMM fitting using joint optimisation [LZZ+18], and learning an image-to-image translation network using known depth and feature maps generated from a 3DMM [SRK17].

Self-supervision Supervised approaches for 3D face reconstruction are limited by a lack of 3D data; capturing 3D data is costly and time-consuming, often making large-scale 3D datasets infeasible. This limitation has led to the widespread use of self-supervised approaches [TZK+17, TZG+18, WBC+19, SSL+20, CWW+20, FFBB21, DBB22, FRPP+23, RGP+24]. In the self-supervised setting, the model uses the data itself, such as images of

the human face, to generate its own supervisory signals without requiring known 3D data. However, many of these approaches perform poorly in metric reconstruction [SBFB19]. This observation has prompted efforts to generate our own paired 2D-3D datasets, as covered in section 2.3.

Tewari *et al.* [TZK+17] combine previous generative and regression-based approaches in their model-based deep convolutional autoencoder: MoFA. Through using an end-to-end encoder-decoder architecture with a fully differentiable image formation layer, the network is trained in an unsupervised setting solely on 2D images. An image is taken as input, converted into a semantically meaningful code vector by the encoder, and then decoded into a 3D model by the image formation layer within the decoder. End-to-end training is achieved using backpropagation through the network; the loss includes a sparse landmark loss, dense photometric alignment, and a regularisation term. Figure 2.6 shows the architectural diagram of this approach.

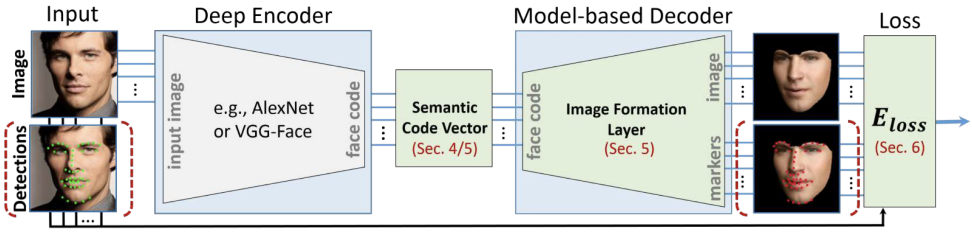


Figure 2.6: MoFA: Architectural Diagram [TZK+17]

Other approaches have since built upon the self-supervised techniques introduced in MoFA. Genova *et al.*[GCM+18] perform unsupervised 3DMM regression with differentiable rendering. The novelty lies in the use of an identity loss based on identity descriptors extracted using FaceNet [SKP15], which is designed to produce an output vector that isolates identity, learning invariance to factors such as expression, pose, and illumination. Hence, the network aims to preserve the identity of the individual during training,

leading to more recognisable reconstructions.

Chen *et al.*[CWW+20] propose a self-supervised framework that includes both a high-level perceptual loss and pixel-wise photometric loss for multi-level supervision. They also incorporate an image-to-image translation network to predict a displacement map in UV-space, in order to model finer facial details. Meanwhile, RingNet [SBFB19] enforces a shape-consistency loss for images of the same subject in order to learn to disentangle identity from expression.

Shang *et al.*[SSL+20] use multi-view geometric and photometric consistency in their MGCNet to produce view-consistent shapes from a single input image. This builds upon the work of [WBC+19], but for a single image input, using view synthesis with defined co-visible maps to provide a multi-view constraint for a single image. Lin *et al.*[LYSZ20] combine the regression network from [DYX+19] with three graph convolutional networks to produce high-fidelity textures from a single image. These networks refine the initial coarse texture generated in the space of the 3DMM with details from the input image.

DECA [FFBB21] uses multiple images of the same person during training to learn to disentangle person-specific details of the face from expression-dependent features, such as wrinkles. EMOCA [DBB22] introduces a deep perceptual emotion consistency loss, which is used to enable more expressiveness in the reconstruction than DECA. Similarly, SPECTRE [FRPP+23] adds a lipreading perceptual loss to DECA for improved mouth movements across videos. Zhang *et al.*[ZCL+23] employ a Vision Transformer [DBK+20] to encode independent facial components from the image as tokens. Their method, TokenFace, employs a hybrid training strategy by performing supervised training on a unified collection of existing 3D face datasets and self-supervised training on large image datasets. This approach

achieves state-of-the-art results on the NoW benchmark for both metric and non-metric reconstruction.

Multi-image Reconstruction Deng *et al.*[DYX+19] consider how to best combine multiple monocular reconstructions of a single individual. In doing so, they train two separate deep neural networks: R-Net and C-Net. R-Net performs monocular reconstruction using a hybrid loss function for weakly-supervised learning. The outputs from R-Net can then be combined in a way that reflects the available information to be gained from each input image, as each image varies in its usefulness for learning each 3DMM parameter. Hence, C-Net produces confidence scores for each identity-bearing 3DMM parameter for an image. This enables multi-image reconstruction and introduces a powerful concept: using deep neural networks to directly estimate the potential contribution of parts of the input for 3D face reconstruction.

Wu *et al.*[WBC+19] introduced MVF-Net, the first multi-view, regression-based approach to estimating 3DMM parameters. Unlike prior regression-based methods [PB16, DYX+19], which indirectly performed multi-image fitting by aggregating monocular reconstructions, MVF-Net employs an end-to-end trainable CNN to directly regress 3DMM parameters from multiple views. This multi-view setting introduces additional constraints that resolve ambiguities, enforced by a novel self-supervised view alignment loss. This method also incorporates a differentiable dense optical flow estimator to quantify and minimise view alignment errors during training.

Furthermore, a photometric reprojection error is utilised to ensure that the estimated model accurately explains the input images across all views. Textures are sampled from one view using the predicted 3D model and camera parameters, and the model is then rendered to another view to compute the loss between the rendered and observed views, as depicted in

fig. 2.7. This approach achieved state-of-the-art performance on the MICC dataset, surpassing previous regression-based methods at the time of its publication [TZK+17, TTHMM17a, DYX+19].

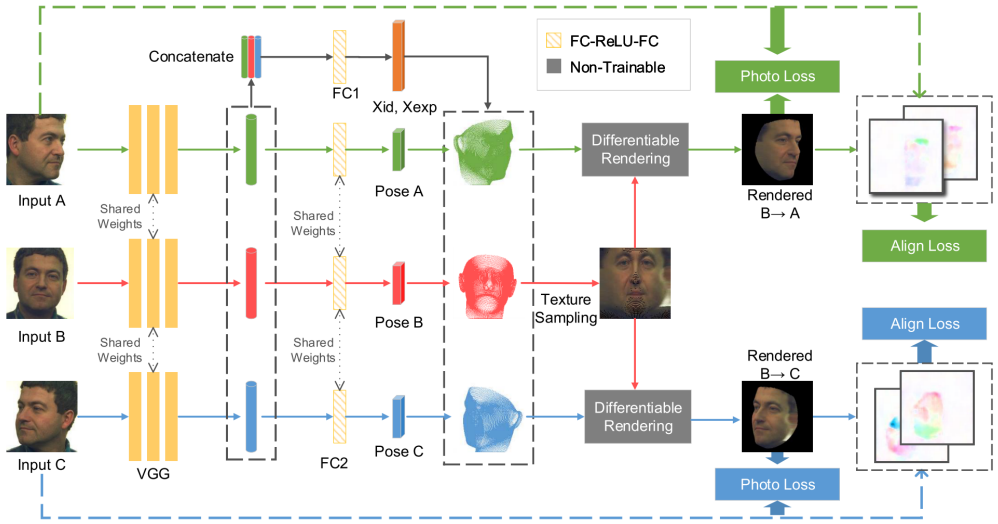


Figure 2. An overview of the proposed model.

Figure 2.7: MVF-Net: Architectural Diagram [WBC+19].

Amberg *et al.*[ABF+07] consider multi-image analysis by synthesis, using model-based stereo to reconstruct high-quality surfaces. This technique relies on stereo information, thus removing the need to estimate lighting and albedo parameters and directly estimating shape instead. A 3DMM prior serves as a regularisation term to guide the fitting process. The objective function combines landmark, silhouette, and colour difference terms. Notably, the silhouette term provides a lighting-invariant cue for reconstruction, achieved by simultaneously fitting a visible contour to the input images. Together, these terms provide superior reconstruction performance, significantly enhancing the state-of-the-art at the time. These multi-image methods are of great importance to single-image reconstruction due to their potential to generate datasets of pseudo ground-truth data that can be used to train new methods.

Optimising Identity Vectors. The loss function used for supervised

3D reconstruction requires careful consideration. Tran *et al.*[TTHMM17b] introduce an asymmetric Euclidean loss for minimising errors between predicted and actual parameter vectors; this decouples over-estimation errors from under-estimation errors. A standard Euclidean loss favours estimates close to zero due to 3DMM parameters following a multivariate Gaussian distribution centred at zero by construction. They report more realistic face reconstructions using their asymmetric Euclidean loss.

However, these losses minimise distance in the vector space of 3DMM parameters rather than minimising reconstruction error directly. Richardson *et al.*[RSK16] directly calculate the Mean Squared Error (MSE) between generated 3D mesh representations. This ensures the loss takes into account how the parameter values affect the reconstructed geometry. Zielonka *et al.*[ZBT22] also employ a mesh-based loss, but they introduce a region-dependent weight mask to weigh the facial region much more heavily than the rest of the head. We aim for accurate 3D face shape estimation, so we will optimise directly in the 3D space using a mesh loss.

2.2 Evaluation of 3D Face Reconstruction

The development of robust 3D face reconstruction methods relies on the availability of high quality datasets for both training and evaluation. However, dataset creation and curation is particularly challenging for 3D tasks due to the significant financial and time-based costs of 3D capture. This section considers existing methods for 3D face shape evaluation, including both datasets and evaluation protocols.

2.2.1 Evaluation Datasets

MICC Evaluation Dataset

The Florence 3D/MICC dataset has been commonly used to evaluate the state of the art in facial reconstruction [BDBM11]. For each of the 53 participants enlisted, at least three 3D scans are captured using a 3D scanner, covering their front, right profile, and left profile. Three videos were also taken of each subject: one indoors with a cooperative subject and two further videos of an uncooperative subject, one indoors and one outdoors.

However, the results derived from using this dataset offer limited information on how the methods may perform on other datasets, populations, or tasks. The faces in each video are in low resolution and in a limited number of controlled settings. Furthermore, the individuals included are predominantly young men, reducing the ability to generalise any results calculated using this dataset. On both counts, it lacks much of the natural variation present in the real world.

NoW Benchmark

Sanyal *et al.*[SBFB19] introduced the NoW (Not quite-in-the-wild) dataset to address the limitations of existing evaluation datasets, particularly in terms of viewing angle, illumination, and common occlusions. It contains high-quality images and 3D head scans of 100 subjects. Images were captured using an iPhone X, and 3D scans were captured using a 3dMD scanner. The individuals are divided into validation and test sets, consisting of 20 and 80 identities, respectively. For each individual, the dataset includes images under different poses, occlusions, and expressions. This represents a clear improvement in the number and variability of situations

captured compared to the Florence3D dataset. The NoW benchmark has become the standard for evaluating both metric and non-metric 3D face reconstruction from 2D images. In metric reconstruction, real-world measures are computed, whereas non-metric reconstruction focuses on qualitative aspects such as visual similarity and perceptual accuracy.

REALY Benchmark

The REALY benchmark [CZR+22] introduces region-specific evaluation for the nose, mouth, forehead, and cheek. The benchmark consists of 100 scans taken from the Headspace dataset [DPSD20a], alongside corresponding multi-view images of the subjects. Region-wise shape alignment is introduced to the evaluation procedure to provide more accurate, bidirectional correspondences between reconstructions and the ground truth, which aims to lead to more accurate measures of error.

2.2.2 Evaluation Protocols

Typical Evaluation Setup Several reconstruction techniques use a similar approach to evaluate their methods using the Florence 3D dataset [WBC+19, TZK+17, GCM+18, DYX+19]. Firstly, the ground truth model is cropped to 95 mm around the tip of the nose. An iterative closest point (ICP) algorithm is then used to align the estimated model to the ground truth model. This aligns the estimated and known meshes to the same local coordinate system. Finally, the point-to-plane distances between the two models are calculated and reported. A similar approach is used to evaluate methods on the NoW benchmark, with an additional rigid alignment step performed based on the scan-to-mesh distance following an initial landmark-based alignment.

Domain-specific Evaluation Other surrogate evaluation functions have been proposed, aiming to reduce the reliance on large, expensive 3D evaluation datasets and instead directly consider the problem domain for reconstruction. Genova *et al.*[GCM+18] evaluate their approach against other methods using a facial recognition similarity metric. This metric is computed using the cosine similarity between the input image and the rendering of the estimated model, as output by a pre-trained face recognition network. The higher the similarity, the better the likeness of the reconstructions to the individuals. Crucially, this method does not require any 3D ground truth data. It is clear that similar approaches may be used to provide a domain-specific evaluation.

However, we currently lack approaches that offer informative baselines across existing 3D face reconstruction methods. Existing methods can only be compared by their performance relative to other existing methods, which also limits the adoption of new datasets.

2.3 Controllable Generation of the Human Face

Despite the widespread success of supervised learning across Computer Vision tasks, it has been severely limited in 3D face-related tasks due to a lack of training data. In this context, supervised learning involves the use of paired 2D-to-3D data, whether real or synthetic, which formally comprises a set of face images and their corresponding 3D model representations [SBFB19]. We have seen in section 2.1 that existing methods either utilise all available 3D data [ZBT22, RGP+24] or synthesise their own [WBH+22]. These methods represent the current state-of-the-art performance in 3D face reconstruction.

However, real datasets do not scale and lack diversity, having reached their natural limit until further data is collected, while existing synthesised

datasets lack photorealism and require manual asset creation. Hence, in this section, we consider methods to generate realistic parameterised faces and in doing so examine potential techniques to generate our own datasets for 3D face-related tasks.

2.3.1 Realistic Parameterised Faces

Automating the tedious manual work behind photorealistic face generation remains an open challenge and long-term goal of 3D face representations [EST+20]. 3D Morphable Models (3DMMs) provide parametric control but generate unrealistic images; Generative Adversarial Networks (GANs) generate photorealistic images but lack explicit control [GGU+20]. Combining the parametric control of a 3DMM with the expressive power of generative image models for faces has the potential to create large-scale datasets for supervised 3D face reconstruction.

Recent work has sought to harness the best of both worlds. StyleRig [TEB+20] was the first approach to offer explicit control over a pretrained StyleGAN through a 3DMM, allowing for parametric editing of generated images. Building upon this, Ghosh *et al.*[GGU+20] condition StyleGAN2 [KLA+20a] on rendered FLAME [LBB+17] geometry and photometric details to add parametric control to GAN-based face generation, facilitating full control over the image generation process. Sun *et al.*[SWZ+23] propose a NeRF-based 3D face synthesis network which enforces similarity with a mesh generated by a 3DMM. However, in all these cases, the resulting images fall short of photorealism.

In the field of image synthesis, probabilistic diffusion models now represent the state-of-the-art, surpassing the capabilities of GANs in image sample quality [DN21]. These models, which have developed significantly since their proposal [SDWMG15], have been further improved by concurrent advances in transformer-based architectures [VSP+17] and text-image embedding

spaces [RPG+21b]. Publicly available text-image embedding spaces such as CLIP [RKH+21b] have further diversified and enhanced these models [RDN+22].

Stable Diffusion is a powerful text-to-image diffusion model, synthesising high resolution images from textual prompts using a Latent Diffusion architecture [RBL+22]. ControlNet [ZRA23], a HyperNetwork that influences the weights of a larger paired network [HDL16], enables a variety of input modalities to be used to condition the output of Stable Diffusion. Implementations include depth maps, user sketches, and normal map conditioning networks, among others.

2.3.2 Paramaterised 3D Face Datasets for Face-Related Tasks

Three-dimensional face datasets for face-related tasks rely on real or synthetic data. Richardson *et al.*[RSK16] generate face geometries directly from a 3D Morphable Model (3DMM), rendering the face as an image under randomised lighting conditions. This results in a dataset of images with known 3DMM parameters; however these images are far from photorealistic. This points to a wider problem in synthesised approaches: a domain gap between synthesised and real data that makes generalisation difficult and task performance poor [KPL+19].

In contrast, Wood *et al.*[WBH+21] combine a generative parametric 3D face model with a library of hand-crafted assets to create their Face Synthetics dataset. They use this model and assets to render 100K face images. However, this approach requires the manual creation of 3D assets with associated time, financial, and computational costs. There are 162 types of eyebrow modelled alone. Furthermore, the rendered images fall short of photorealism which limits their uses for direct 3DMM regression.

Other approaches have considered using the real 3D data we have rather than relying on synthesised datasets. Zielonka *et al.*[ZBT22] unify existing

3D face datasets. However, this approach already represents the upper bound for supervised learning using 3D data, unless further data is collected. In combining 8 existing datasets, they reach just 2315 individuals; this remains a small dataset for supervised learning techniques. Hence, a generative approach similar to Wood *et al.*[WBH+21] is required for unconstrained dataset generation.

A lack of training data is most acute for applications that rely on events that are rarely occurring, such as facial trauma. This is the case in maxillofacial prosthesis design where clinicians want to reconstruct a missing region of the face following an accident or surgical intervention. The facial areas of such patients are not modelled in standard datasets for 3D face shape estimation. An ongoing clinical trial is comparing digitally manufactured prostheses with conventional manufacture [JCB+23]. 3DMMs will be used in the digital arm of this trial for facial completion.

2.4 Critical Analysis of the Literature

Since the introduction of the 3D Morphable Model by Blanz and Vetter in 1999 [BV99], we have developed a range of computational methods to improve modelling of the human face, enabling both more accurate representations of the face and more accurate reconstruction. However, 3D face reconstruction remains an inherently under-constrained problem, particularly when reconstructing from a single image. We require new methods to increase the accuracy, controllability, and explainability of 3D face reconstruction.

Deep learning has provided us with the ability to learn complex non-linear functions that map the space of images to a plausible space of corresponding 3D reconstructions. However, a lack of available paired 2D-3D training data led the community to adopt self-supervised approaches. These approaches

enabled us to leverage the wealth of knowledge contained in existing face image datasets during network training. The advent of photorealistic 2D image generation models now presents a new opportunity: to integrate this wealth of knowledge about the human face with our existing parameterised 3D models of the human face, such as the 3DMM.

In doing so, we have the opportunity to generate large-scale customised datasets for the supervised training of 3D face reconstruction networks. Crucially, this will allow us to better model the variability present in the real world. This approach includes the capability to create datasets that are balanced by race and gender. Additionally, it enables the modelling of rare cases, such as facial injuries, for which we have limited data.

In recent years, shared image-text latent spaces, such as CLIP, have revolutionised both Computer Vision and Natural Language Processing. We have also seen the potential of these techniques to revolutionise how we interact with both new and existing representations of the human face. When correctly utilised, CLIP offers the opportunity to initialise and edit faces directly with text. This is a key feature of the Intelligent Face Agent we envisioned in section 1.1.

These new models will also require new methods of evaluation. Our current evaluation datasets are limited in the number of subjects they include and the variation they capture. There is an opportunity to enhance our understanding of how our existing datasets test our reconstruction methods and to develop new datasets and metrics for evaluating 3D face reconstruction. These evaluation methods could also consider other forms of input, such as text, which are increasingly being used as cues for reconstruction. In cases where reconstruction is more ambiguous, new approaches are necessary. From these new methods of evaluation, we will improve our understanding of existing methods and be able to demonstrate and enhance

real-world performance.

From this overview of the literature, it is clear that there exists a unique opportunity to improve the way we perform 3D face reconstruction. These areas of research are mutually beneficial, as improvements in evaluation will enhance real-world performance, and improved generation methods will enable new methods of interaction, such as text-based initialisation and refinement. In the following technical chapters, we will build upon these opportunities identified within the literature to enhance the accuracy, controllability, and explainability of 3D face reconstruction.

Chapter 3

Fast 2D-3D Face Dataset Generation

Research Question 1

Can we exploit knowledge about the structure and appearance of the human face, contained within pretrained image generation networks, to generate large-scale datasets for 3D face reconstruction?

Research Question 2

What methods can be employed to increase the diversity of paired training data for 3D face estimation methods?

Dataset creation and curation presents a challenging problem in Computer Vision, particularly for tasks involving 3D faces. In section 2.4, we identified the opportunity for us to improve reconstruction by integrating our existing 2D knowledge of the human face, as contained within image generation models, with existing parameterised models of the 3D face.

In this chapter, we present an end-to-end dataset generation methodology for producing photorealistic 2D images paired with known 3D face shapes, as parameterised by a 3DMM. In doing so, we fulfill part of the core functionality of the Intelligent Face Agent envisioned in section 1.2, enabling customised dataset generation for face-related tasks. We employ this method

to create SynthFace, the largest available dataset of paired 2D-3D face data, which is designed to be balanced by race and gender. This chapter illustrates a novel method to leverage existing knowledge of the human face for generating large-scale datasets for 3D face reconstruction.

The creation of large-scale datasets of paired 2D-3D will help us train methods to resolve some of the inherent ambiguities present in 3D face reconstruction from a single image. So far, creating large-scale 3D face datasets has been infeasible due to the costs and time associated with 3D capture. One approach to reconstruction in this data-limited environment is to unify existing 3D face datasets [ZBT22]. However, images from these 2D-3D datasets suffer from a lack of diversity and already represent an upper bound on dataset size until further 3D data is collected. Meanwhile, existing synthetic 3D face datasets are not photorealistic, require handcrafted assets, and are computationally expensive to produce. Additionally, work on face-related tasks has raised the issue of unfair models [BG18], which exhibit greater bias when tested on unbalanced data [KJ21]. This underscores the importance of collecting diverse, large-scale datasets.

In this chapter, we address the problem of generating parameterised 3D geometry of the human face with diverse corresponding face images. We describe how to combine a generative parametric 3D face model with a library of 3D hairstyles and textual appearance descriptors to render photorealistic training images.

This is achieved by conditioning Stable Diffusion on both our combined head and hair models alongside our textual appearance descriptors. Using a textual appearance descriptor allows users to customise their dataset as requirements change. For example, we may want to analyse the face shape of older people wearing face masks. With existing synthetic datasets, this would require the creation of new assets which would be costly and

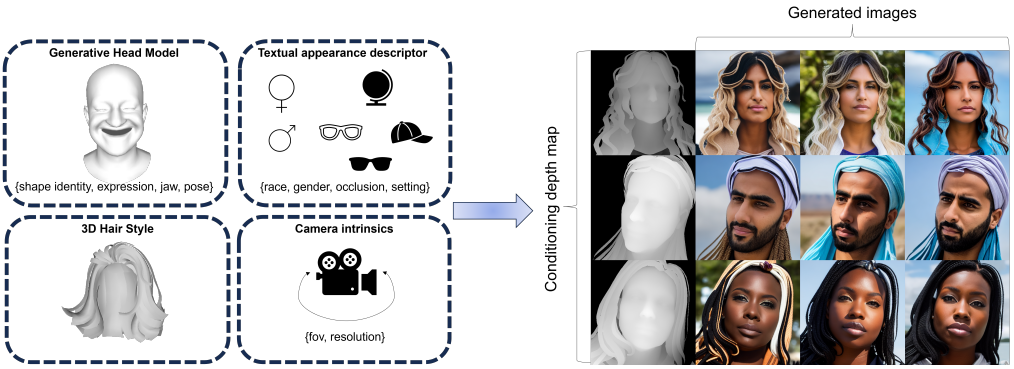


Figure 3.1: We present SynthFace, the largest available dataset of photorealistic face images and 3D Morphable Model (3DMM) shape parameters. We sample head shape, hair style, custom appearance descriptors, and camera intrinsics to create this dataset of unparalleled diversity and photorealism which requires no real data or crafted assets.

time-consuming. However, with our approach, a user can update the textual appearance descriptor and see the results in seconds, with a full dataset available in hours. This approach is also highly modular with each part modifiable for the requirements of a task. This enables unparalleled image diversity and unconstrained 2D-3D dataset generation. With diverse, synthesised datasets, it is possible to solve problems for rare cases without requiring any real data. Our approach is shown in fig. 3.1.

Unlike existing methods, our approach requires no manual asset creation and offers a 20x speedup over the state-of-the-art in dataset generation. We use this method to generate SynthFace, the largest available dataset of 3D Morphable Model (3DMM) parameters and photorealistic face images. Using SynthFace, we train a method for 3D face reconstruction, demonstrating that conditioned dataset generation is competitive with approaches using real-world data and producing a strong baseline for zero-shot dataset training.

We summarise the contributions of this chapter as follows:

- We introduce an end-to-end dataset generation methodology for producing paired 2D and 3D face data that removes the burden of asset

creation and produces unparalleled diversity and photorealism. It offers a greater than 20x speedup in generation time over the current state-of-the-art.

- We present the SynthFace dataset, complete with 180K photorealistic face images and corresponding 3DMM parameters. This dataset is divided into four categories: expressions, neutral, selfies, and occlusions.
- We demonstrate improved gender diversity through the inclusion of a textual appearance descriptor which conditions image generation. We further present a scheme which is designed to balance the dataset by race.
- We verify the effectiveness of our zero-shot dataset generation approach for the challenging task of monocular 3DMM parameter regression without using any real data - the first method to do so on the NoW benchmark.

In Section 3.1, we introduce the SynthFace Generator, our method for generating paired 2D-3D face data which we use to generate the SynthFace dataset. This includes details on variation encoded within the dataset, including shape, appearance, and perspective changes. In Section 3.2, we introduce ControlFace, a deep neural network trained on this new SynthFace dataset. In Section 3.3, we compare our SynthFace generator against the state-of-the-art Face Synthetics [WBH+21] dataset. In addition, we evaluate ControlFace on the NoW Benchmark, demonstrating competitive performance with the state-of-the-art without requiring any real data.

3.1 The SynthFace Generator

Synthetic dataset generation approaches for 3D face-related tasks are able to produce diverse, realistic training images. However, these images are not

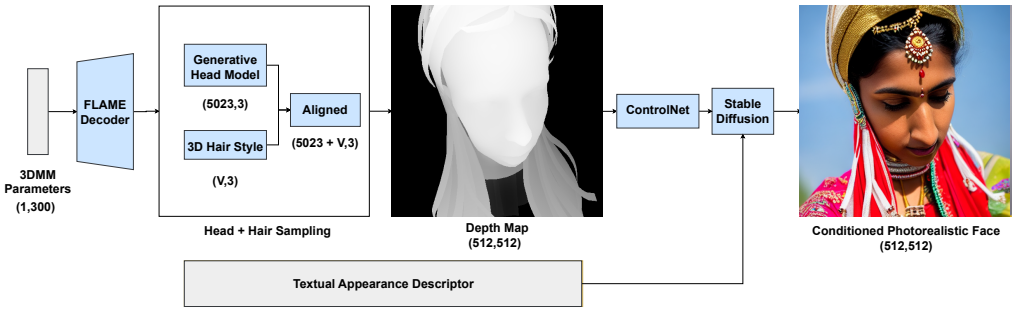


Figure 3.2: The SynthFace Generator. We sample from a 300-dimensional shape vector, 50-dimensional expression vector, and a 6-dimensional pose vector. We input these to the FLAME decoder to produce a 3D mesh. From this mesh, we extract a depth map that, alongside our textual appearance descriptor, is used as conditioning to generate a photorealistic face.

photorealistic, require extensive asset creation, and are computationally expensive to render. To address this, we use conditioned Stable Diffusion [RBL+22] to produce photorealistic images without requiring asset creation or slow rendering pipelines. We generate a comprehensive dataset for 3D face-related tasks, comprising 180K photorealistic faces with 20K distinct 3D facial shapes. We call this the *SynthFace Dataset*. We design the dataset to be split into four distinct categories: neutral, expressions, selfies, and occlusions. This aims to increase the real-world usability of SynthFace for a range of face-related tasks.

We begin by sampling a head from a popular generative 3D head model. We then select a hair style from an existing 3D hair style collection and align it to the head. We independently sample from a range of facial appearance attributes to create a textural appearance descriptor. This descriptor and a depth rendering of our combined 3D head and hair are used to condition a Stable Diffusion model for photorealistic training image generation. The SynthFace Generator is shown in fig. 3.2. We render all 180K $(512, 512)$ resolution images in 30 hours utilising 10 GTX 1080 GPUs, which demonstrates an order of magnitude lower resource requirement compared to similar work [WBH+21]. This section describes how we achieve

this without manual asset creation.

3.1.1 3D Face Model

We use the FLAME head model [LBB+17] as a generative model for face shape. FLAME is a linear 3DMM with both identity and expression parameters. Linear blend skinning (LBS) and pose-determined corrective blendshapes are used to model the neck, jaw, and eyeballs around joints. This results in a head model containing N = 5023 vertices and K = 4 joints. FLAME takes coefficients for shape $\vec{\beta} \in \mathbb{R}^{|\beta|}$, pose $\vec{\theta} \in \mathbb{R}^{|\theta|}$, and expression $\vec{\psi} \in \mathbb{R}^{|\psi|}$. These are modelled as vertex displacements from a template mesh $\overline{\mathbf{T}}$, with $M(\vec{\beta}, \vec{\theta}, \vec{\psi})$ returning N vertices when supplied with these shape, pose, and expression parameters. $T_P(\vec{\beta}, \vec{\theta}, \vec{\psi})$ is the template mesh with vertex offsets for shape, pose, and expression applied. A skinning function W rotates the vertices of T_P around joints $J \in \mathbb{R}^{3K}$. This is linearly smoothed by blendweights $\mathcal{W} \in \mathbb{R}^{K \times N}$. The model is formally defined as:

$$M(\vec{\beta}, \vec{\theta}, \vec{\psi}) = W(T_P(\vec{\beta}, \vec{\theta}, \vec{\psi}), \mathbf{J}(\vec{\beta}), \vec{\theta}, \mathcal{W}) \tag{1}$$

where

$$T_P(\vec{\beta}, \vec{\theta}, \vec{\psi}) = \overline{\mathbf{T}} + B_S(\vec{\beta}; S) + B_P(\vec{\theta}; P) + B_E(\vec{\psi}; E). \tag{2}$$

Due to different face shapes requiring different joint locations, joints are defined as a function of $\vec{\beta}$. Equation (2) includes shape, pose, and expression blendshapes, denoted by $B_S(\vec{\beta}; S)$, $B_P(\vec{\theta}; P)$, and $B_E(\vec{\psi}; E)$ respectively. We use eq. (1) to generate a complete 3D mesh of the head from these coefficients. This approach enables us to create an arbitrary number of human head shapes with expression and pose, each compactly represented by a set of 3DMM parameters.

3.1.2 3D Hair Database

We extend the existing FLAME head model with 300 3D hair styles. We use the hair styles selected by Zhang *et al.*[ZZ18], from the 3D Hairs in the Wild (3DHW) dataset [CSW+16], to train HairGAN. We compute a linear transformation to map these hair models to the reference FLAME head mesh in metric space. At generation time, an additional linear transformation is calculated to align the chosen hair with the generated head shape.

First, we compute a transformation that maps all hair samples from the database, which are already aligned in their own coordinate space, to the mean face of the FLAME head model. This ensures that each hair style is correctly attached to the mean face. However, when we initialise the FLAME head model with different shape parameters, its size changes from this mean face. Therefore, we require additional transformations to ensure the correct scale and translation. Without this, the same hair style would look small on a wide head and very large on a much narrower head and suffer from misalignment with the top of the head. To correct for this, we calculate the vector between the top of the mean head and the generated head, performing a translation of the hair by this vector. The top of both heads are found by searching for the maximum y coordinate of all vertices.

Additionally, we calculate the scaling in both height and width from the mean face to the generated head, and apply this as a scaling factor to the hair in both x and y directions. We determine these scale factors by finding the difference between maximal and minimal vertex positions in x and y for both faces, and then calculating the ratio of $generatedhead/meanhead$. We find this process is sufficient for consistently stable fitting of, as shown by inspection of mapped samples and images generated from these combined meshes.

3.1.3 Depth Map Generation

In building SynthFace, we use all 300 FLAME shape parameters (β), 50 expression parameters (ψ), and 3 pose parameters (θ) specifying global rotation. We sample identity parameters, $\vec{\beta}$, individually from a Gaussian distribution with mean 0 and s.d. 1.0. This enables a wide variation of face shape within our dataset. Expression coefficients, $\vec{\psi}$, are sampled from a Gaussian distribution with mean 0 and s.d. 0.6. We sample pitch, yaw, and roll from a Gaussian distribution with mean 0 and s.d. of 10°, 20°, and 5° respectively.

The SynthFace dataset is split into four distinct categories: neutral, expressions, selfies, and occlusions. This is designed to ensure the resulting dataset covers a wide range of human appearance, including the effects of changes in perspective from selfies and common occluders such as sunglasses. Each of these has 5K unique identities and 45K generated images. For neutral images, we set the expression coefficients to 0.

We use a perspective camera with a 72.4° field of view. We vary the distance between the camera and subject from 80 to 150 world units for selfies and 150 to 500 world units for all other categories. This leads to perspective projection effects which model real-world image changes within our dataset.

3.1.4 Conditioned Face Generation

We use the depth version of *ControlNet* [ZRA23] to modulate the output of *Stable Diffusion 1.5* [RBL+22]. We do not fine-tune *ControlNet* - this demonstrates its generalisability and is a strong positive indicator for its performance in real-world scenarios. It takes a depth map and textual prompts (positive and negative prompts) as input to produce an image. We produce 3 images per prompt. The inference procedure is set to run

for 15 steps. We use customised prompts for race, gender, and the three main types of occlusions. This results in the following prompt template: "{*occlusion*}, {*race*} {*gender*}, profile picture, dslr". The full negative prompt and example positive prompts are included in A.1. A quarter of all images in the SynthFace dataset model occlusions. This is split equally between glasses, sunglasses, and hoodies. The prompts for these are given in eq. (3.1). All images, including those under occlusion, are split equally by race and gender as defined in eq. (3.2) and eq. (3.3) respectively.

$$\text{occlusions} = \left\{ \begin{array}{l} \text{glasses, sunglasses,} \\ \text{hoodie} \end{array} \right\} \tag{3.1}$$

$$\text{race} = \left\{ \begin{array}{l} \text{White, Black, Indian,} \\ \text{East Asian, Southeast Asian,} \\ \text{Middle Eastern, Latino} \end{array} \right\} \tag{3.2}$$

$$\text{gender} = \left\{ \text{woman, man} \right\} \tag{3.3}$$

In contrast to other 3D face datasets, we include a large number of different identities for the same face shape. An identity here is an individual recognisable person in 2D image space; a shape is the 3D mesh as parameterised by the 3DMM. We produce 9 images per distinct 3D shape, each capturing a different visual identity, but with the same underlying 3D shape. Figure 3.3 shows how even when we keep the appearance descriptor fixed, different identities are included within SynthFace for the same shape. We believe we are the first to incorporate this approach into a dataset for 3D face shape estimation by design. Hence, by including multiple identities for the same underlying 3D shape, SynthFace offers unique opportunities to learn to disentangle the underlying shape from the perceived identity.

Figure 3.3: The SynthFace Dataset includes different poses, perspective projections, and visual identities for each unique 3D shape. Here we present three images within SynthFace, each conditioned on the same 3D shape (including hair) and textual appearance descriptor.

3.1.5 Prompt Selection

Prompting enables image generation models, and other vision-language models, to have their output informed by text-based guidance. Prompt selection concerns the choice of a prompt for a desired effect. In our case, this is to produce diverse and lifelike images of humans. We have two tools to achieve this: positive and negative prompts. Positive prompts specify a direction for the generation model to take, here a direction in latent space to step towards as part of the de-noising process during diffusion, which can be towards specific styles and semantic specifications. Negative prompts take the generation process in the other direction, making the resulting images less closely resemble the contents of a negative prompt. Combined, this enables experimental tuning of a generation model, with the continual refinement of these prompts being able to improve the results obtained by

the model.

A systematic method was undertaken to iteratively refine our prompts to generate realistic human faces. This process involved starting with a single text prompt, 'studio portrait', and iteratively adding single phrases, both to positive and negative prompts, to build an improved prompt. The impact of these additional phrases was qualitatively evaluated in each case. Only phrases that produced more visually lifelike outputs were retained with a preference for compactness - if a prompt phrase has negligible impact on visual appearance, it is removed.

A selection of this process is presented in figs. 3.4 to 3.6.

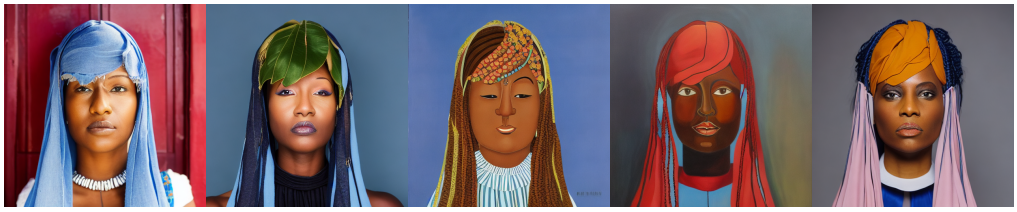


Figure 3.4: Intermediary prompting result 1
prompt: 'a woman'
negative prompt: 'worst quality, normal quality, low quality, low res, blurry, jpeg artifacts, error, sketch , monochrome, geometry'

Figure 3.5: Intermediary prompting result 2
prompt: 'a woman, profile picture'
negative prompt: 'worst quality, normal quality, low quality, low res, blurry, jpeg artifacts, error, sketch , monochrome, geometry'
comment: much more realistic with profile picture prompt.

3.1.6 Dataset Demographics

We use *FaceLib* [Ayo19] to estimate age and gender information from all generated faces. With the use of our textual appearance descriptor,

Figure 3.6: Intermediary prompting result 3
prompt = 'a woman, profile picture, dslr'
negative prompt: 'worst quality, normal quality, low quality, low res, blurry, jpeg artifacts, error, sketch , monochrome, geometry'
comment: worse - added expression artifact.

our SynthFace dataset is estimated to be 58.2% male and 41.8% female; this binary is reductive but useful as a diagnostic. Without our textual appearance descriptor, we evaluate the generated dataset to be 83.1% male and 16.9% female. This dramatic improvement in gender balance demonstrates the power of our descriptors and our method for generating customised datasets. This result concurs with our own inspection of results, which noted considerably improved gender diversity and strong prompt adherence when using the textual appearance descriptor. Figure 3.7 shows further images from the final SynthFace dataset, including cases of minor artefacts.

3.2 ControlFace for 3D Face Reconstruction

We introduce *ControlFace*, a deep neural network trained on our new SynthFace dataset. This network aims to disentangle shape from identity and perspective through supervised training on a large dataset that contains multiple identities and multiple observed views for the same shape. This takes direct advantage of the benefits of our generation pipeline.

ControlFace accepts an image as input and outputs a shape vector $x \in \mathbb{R}^{300}$ for the FLAME decoder to produce a reconstructed face with neutral expression. All architectures, training, and evaluation are implemented

Figure 3.7: The SynthFace Dataset. We present a selection of typical images from the SynthFace dataset. These further demonstrate the diversity and photorealism achieved using our proposed generation method. In the bottom row, we also highlight the most common weaknesses in images generated using Stable Diffusion: chiefly minor issues related to the generation of hair, including issues with colour, positioning, and it being mistaken for another material.

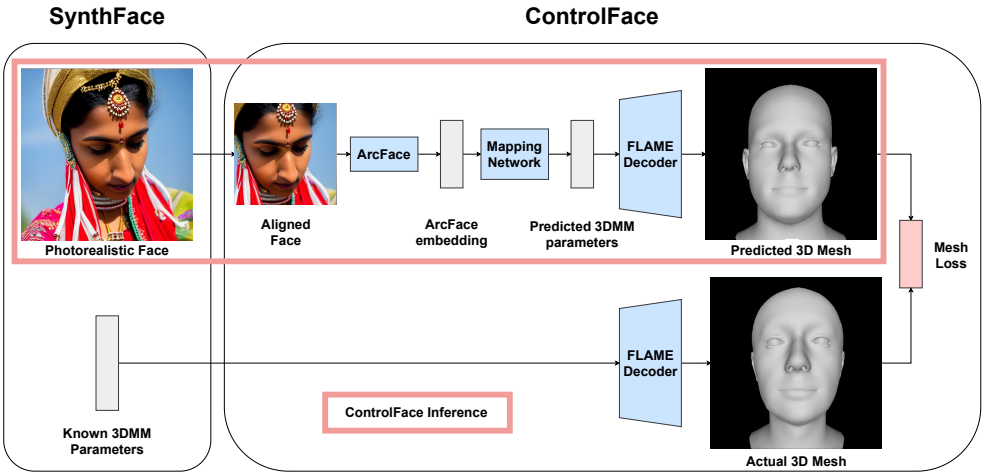


Figure 3.8: ControlFace training. We train the mapping network within ControlFace on the SynthFace dataset. It is trained to minimise the mesh reconstruction error between a predicted 3D mesh and known 3D mesh for each image in SynthFace. ControlFace at inference is shown outlined. ControlFace accepts an image as input, aligns it, and calculates an ArcFace embedding from this aligned detected face. A mapping network converts this ArcFace embedding to 3DMM parameters. The FLAME decoder generates a full head mesh from these parameters.

using PyTorch [PGM+19]. Figure 3.8 shows the training process in full, including the inference pipeline used during model deployment.

3.2.1 Training Data

We use the entirety of the SynthFace dataset as our training data. SynthFace contains 180K images of 20K unique shape identities. A unique shape identity is defined as a unique set of 3DMM parameters. For each of these unique shape identities, we have three different views of the subject and three photorealistic images of varying appearance for each of these views.

3.2.2 Pre-processing

First, faces are detected in each image using *RetinaFace* [DGV+20]. This provides a bounding box used to crop each image and warp it to a frontal pose. Next, we use the pretrained ArcFace network [DGXZ19] as a feature extractor for face description. ArcFace's 512-dimensional output embedding

is used as input for a mapping network.

3.2.3 Mapping Network

We use the same mapping network architecture as presented by Zielonka *et al.*[ZBT22]. This network consists of three fully-connected layers followed by a linear output layer. Weights are randomly initialised and we train this network to regress a shape vector $y \in \mathbb{R}^{300}$ from an ArcFace embedding vector $x \in \mathbb{R}^{512}$. This vector contains coefficients for all 300 identity bases in the FLAME head model.

3.2.4 Training Strategy

We split the SynthFace dataset into training and validation sets, following an 85/15 split. We train our mapping network on the training set and select the best performing model based on the validation loss; we use early stopping with a patience of 10 to achieve this and run for 100 epochs.

We use the AdamW optimiser for optimisation with learning rate $\eta = 1 \times 10^{-5}$ and weight decay $\lambda = 2 \times 10^{-4}$. We use the same optimisation strategy and masked mesh loss function as Zielonka *et al.*[ZBT22]:

$$L = \sum_{(I,G)} |\kappa_{\text{mask}}(G_{\text{3DMM}}(M(\text{ArcFace}(I))) - G)|, \tag{3.4}$$

which puts emphasis on inner facial regions in reconstruction. κ_{mask} is a region-dependent weight mask with values: 150 for the face region, 1 for the back of the head, and 0.1 for the eyes and ears. This loss is calculated for all pairs of input images, I, and known meshes, G, within SynthFace. $G_{\text{3DMM}}(M(\text{ArcFace}(I))$ is the predicted mesh after the image is passed through ArcFace, the mapping network M, and then the FLAME decoder G_{3DMM}.

3.3 Experiments and Evaluation

We evaluate our generation method for several key criteria, including generation time, photorealism, and CO_2 footprint during generation. The CO_2 footprint for each generation method was estimated using the Machine Learning Impact calculator presented in [LLSD19]. We then evaluate our trained network for ControlFace for the task of 3D face reconstruction on the NoW benchmark. Ours is the only method here that utilises no real 2D or 3D data for supervised training.

3.3.1 The SynthFace Generator

Comparison	Face Synthetics [WBH+21]	SynthFace (ours)
GPU hours	7200	**300**
CO_2 (t)	1.37	**0.023**
Photorealistic	✗	✓
No Handcrafted assets	✗	✓
Dataset Size	100K	**180K**
Resolution	**512x512**	**512x512**

Table 3.1: Comparison of SynthFace with the state-of-the art synthetic dataset generation method. Each generation method is evaluated based on generation time, resulting CO_2 emissions, photorealism, the requirement for manual asset creation, and dataset size. Bold indicates better.

The results in Table 3.1 clearly demonstrate the efficiency and effectiveness of our SynthFace Generator in comparison to the state-of-the-art Face Synthetics method [WBH+21]. Our method offers an over 20x speedup. This efficiency is further reflected in the drastic reduction of CO_2 emissions, where SynthFace is predocted to produce only 0.023 tons compared to 1.37 tons by Face Synthetics.

Figure 3.9: Face Synthetics qualitative comparison. Three randomly selected images from Face Synthetics (top) and SynthFace (bottom) are viewed here. Our textual appearance descriptor enables the generation of typical occluders such as glasses. Previously, this required manual asset creation by digital artists as is the case for Face Synthetics.

Moreover, our method achieves photorealism without requiring any real 2D or 3D data for supervised training, a clear advancement over existing methods. The ability to generate photorealistic images without relying on real assets underscores the potential of SynthFace in various applications, particularly where data privacy is a concern. Additionally, the larger dataset size of 180K images compared to 100K by Face Synthetics, while still requiring far lower computational resources, showcases the scalability of our approach. A qualitative comparison with Face Synthetics is shown in fig. 3.9. SynthFace can be seen to offer qualitatively superior results with increased face realism.

3.3.2 Landmark Accuracy of Dataset Generation

To demonstrate the effectiveness of our dataset generation method for producing 2D-3D consistent data, we compare detected facial landmarks on the generated images with known landmarks from the underlying 3D shape. First, we run a landmark detector [BT17] on a sample of 5,000 images from SynthFace, storing all 68 facial landmarks. We then calculate the median absolute error between detected and known landmarks, normalising by the inter-ocular distance (IOD) of the detected landmarks for scale invariance. Table 3.2 shows the resulting errors, including different regions of the face.

Region	IOD error
Nose (ours)	4.99%
Eyes (ours)	8.72%
Mouth (ours)	8.18%
All (ours)	9.64%

Table 3.2: Landmark consistency for 5,000 randomly sampled images from SynthFace

BlazeFace [BKV+19], a popular landmark detector, exhibits a median IOD error of 10.4% when applied by its authors to a similar facial landmarking task. The strong landmark consistency exhibited by SynthFace demonstrates the usefulness of our proposed dataset for further downstream tasks and the efficacy of our dataset generation method. Figure 3.10 shows a visualisation of the known 3D landmarks from our generative head model projected onto both the depth map and generated image.

3.3.3 NoW Benchmark

We will now test our proposed reconstruction method against the *NoW* benchmark [SBFB19]. The NoW benchmark consists of 2054 images for 100 identities. It has become the standard benchmark for evaluating 3D

Figure 3.10: Example image from SynthFace, showing the projected 2D landmarks from the 3D head mesh on both the depth map and generated image.

face reconstruction from 2D images. These are split into validation and test sets consisting of 20 and 80 identities respectively. For each individual, the dataset includes images under different poses, occlusions, and expressions. We use the publicly available validation set of NoW for evaluation.

3.3.4 Experimental Setup for ControlFace

First, a rigid alignment of the predicted meshes to the scans is performed using key facial landmarks. Then the scan-to-mesh distance between the predicted mesh and scan is performed for each vertex. The mean, median, and standard deviations of these distances are computed across all images within the validation set with millimetre error reported. It is important to note that due to scaling, these results should not be interpreted as true metric errors. Table 3.3 shows a comparison of our ControlFace approach with current state-of-the-art methods. All methods presented use supervised or self-supervised learning.

Method	Med.	Mean	Std	Train
Deep3D [DYX+19]	1.286	1.864	2.361	✗ ✗ ✓
DECA (detail)	1.19	1.469	1.249	✗ ✗ ✓
DECA [FFBB21]	1.178	1.464	1.253	✗ ✗ ✓
AlbedoGAN (detail)	0.95	1.173	0.987	✗ ✓ ✓
MICA [ZBT22]	0.913	1.130	0.948	✓ ✗ ✗
AlbedoGAN [RGP+24]	0.903	1.122	0.957	✗ ✓ ✓
ControlFace (ours)	1.241	1.565	1.328	✓ ✓ ✓

Table 3.3: Reconstruction error (mm) on the validation set of the NoW benchmark [SBFB19] in non-metrical reconstruction. Comparison results are presented from [RGP+24]. The final column includes ticks and crosses which indicate whether the method meets specified criteria. The first element indicates whether supervised training between images and 3DMM parameters is employed. The second and third elements indicate the use of synthetic data: first for 2D images and then for 3D meshes.

3.3.5 Discussion and limitations

SynthFace Generator. The qualitative results, as illustrated in fig. 3.9, demonstrate the diverse, photorealistic faces we generate with our method. This improves the state-of-the art in parametric dataset generation without requiring any asset creation. Our approach also produces an over 20x speedup in generation time, reducing estimated CO2 consumption by nearly 60x. Using a pre-trained diffusion model removes the computationally expensive rendering step and allows us to prompt for diverse attributes using our textual appearance descriptor.

We verify the shape consistency of our generated 2D images to the underlying 3D mesh by comparing projection errors against those of a popular face landmark detector. A key benefit of our method is the ability to generate a dataset with diverse poses and appearances. However, landmark detection is likely to perform worse under such conditions, making shape verification a more challenging task for our generated dataset. Further work

should investigate these effects further, including errors grouped by pose.

ControlFace. Our method, ControlFace, is to the best of our knowledge the first supervised approach, trained using no real data or manual asset creation, to be evaluated on the NoW benchmark. We achieve this by introducing a novel method for large dataset generation for 3D face shape estimation. We demonstrate that accurate 3D face shape estimation is possible in this zero-shot dataset setting, offering the first baseline approach for this setting.

However, while strong, ControlFace does not achieve state-of-the-art performance. Future work could consider training on both SynthFace and a real captured 3D dataset to determine whether this could improve performance further.

Unconstrained dataset generation. Our generation methodology is easily extensible. A longer generation time can lead to a larger dataset and improvements in 2D and 3D generative model capabilities can directly feed into future work. We believe this will enable future methods trained on zero-shot generation datasets to close the performance gap with methods such as MICA and AlbedoGAN. Datasets for specific use cases, be that large pose variations or expressions, can be created by updating parameters in our generation code.

In unifying existing 3D face datasets, MICA reaches a natural limit in supervised learning on existing data sources. This is where the opportunity for synthesised approaches such as SynthFace lies. SynthFace can scale beyond this natural limit in real paired data.

Application to modelling rare cases. Our work in unconstrained dataset generation allows for the modelling of rarer clinical cases; for example, in the case of orbital (eye) defects. A lack of 3D data is most acute for tasks which include rarely occurring events such as facial trauma.

Orbital reconstruction could benefit from a dataset designed for the task using synthesised data. Current landmark-based methods struggle in the presence of asymmetrical facial defects. Our method can enable learning-based approaches in the absence of sufficiently-large real datasets.

Ethical considerations. Generative models like Stable Diffusion require extensive datasets for training which typically rely on publicly available data. Consequently, there's a likelihood that individuals' data has been used without their explicit consent. This raises clear ethical and legal concerns, particularly for models deployed in the real world. However, diffusion also presents a clear advantage. Diffusion, as used in our method, enables privacy-preserving dataset creation. This is particularly useful in the creation of large customised datasets for medical tasks where opportunities for dataset collection are limited. In doing so, there is a large positive upside to this work.

We have shown our dataset generation method beats the existing state-of-the-art in significantly lower CO2 emissions during generation. This is a clear benefit in using our method over the existing state-of-the-art. We encourage the further adoption of this comparison as a key metric for the performance of Computer Vision techniques.

3.4 Summary

We have addressed a key challenge in dataset generation for 3D face-related tasks, removing the requirement for manual asset creation and enabling photorealistic face generation with paired 3D shape. The resulting dataset, SynthFace, is the largest dataset of its kind and offers unique opportunities to disentangle shape from identity for accurate 3D face reconstruction. We demonstrate the promise of our zero-shot dataset generation method for this task, producing a strong initial baseline result. The complete

SynthFace dataset and generation code will be made publicly available, offering a new method for customised dataset generation. In doing so, this approach fulfills a core part of the generative capabilities envisioned within the proposed Intelligent Face Agent, presenting an approach to improve both the controllability and accuracy of 3D face reconstruction.

Addressing research question one, we clearly demonstrate that using the proposed SynthFace Generator, we are able to exploit the structure and appearance of the human face to generate large-scale datasets for 3D face reconstruction. This is shown in both the qualitative results presented from the generated SynthFace dataset and further demonstrated in strong quantitative results using ControlFace, a 3D face reconstruction network trained exclusively on the SynthFace dataset. We further address research question two by demonstrating a much improved gender balance when the textual appearance descriptor within the SynthFace Generator prompts for gender in the resulting image. In doing so, we increase the proportion of individuals estimated to be female within the SynthFace dataset from 16.9% to 41.8%.

Unlike previous generation methods, ours is easily extensible, computationally inexpensive, and produces photorealistic face images. It is over 20x faster than the relevant state-of-the-art for synthetic dataset generation. It further allows us to address race and gender bias in existing Computer Vision datasets, providing a method to balanced by race and gender using a textual appearance descriptor. We see this approach to combining 2D and 3D face knowledge for dataset generation to hold great potential, particularly as existing 3D face datasets reach their limit for supervised learning.

We expect improvements in image generation, 3D face models, and conditioning networks to all improve the accuracy of this method for downstream

tasks such as 3D face reconstruction; our work provides a strong baseline for the first zero-shot dataset approach. We believe this work will form the basis of a number of exciting future developments in this domain.

In the next chapter, we will consider how we can further improve the controllability of 3D face reconstruction by enabling the direct initialisation of the 3D head shape, as parameterised by a 3DMM, from textual descriptions. This will address a crucial part of the functionality required to realise the proposed Intelligent Face Agent. Furthermore, it will build upon the foundations laid in this chapter by proposing a novel method to integrate existing textual and visual information about the human face to achieve this goal.

Chapter 4

Text-based Initialisation of 3D Face Models

Research Question 3

Are we able to perform parameterised shape initialisation of faces from natural language descriptions?

Generative 3D shape models, such as 3D Morphable Models (3DMMs) [BV99], are useful statistical priors with which to explain 2D/3D images of 3D objects, such as human faces, by reconstructing their shape and texture. However, 3DMMs are difficult to initialise and edit, limiting their potential use by non-experts. In section 1.2, we identified prosthesis design as one application area which would benefit from new input modalities. Whereas current reconstruction methods rely on photo or video, we as humans often communicate our ideas using language, from which we are able to visualise key features as a mental image. Extending 3DMMs to perform initialisation from textual descriptions will enable us to widen the potential use cases of 3D face reconstruction.

In this chapter, we present a method for direct and complete generation of parameterised 3D faces from natural language descriptions. To our knowledge, at the time of conception and implementation, this was the first *text to fully parameterised (including identity) 3D face model* approach. This

fulfills the text-to-shape functionality of the proposed Intelligent Face Agent outlined in Chapter 1. Furthermore, there are a multitude of generative applications for 3DMMs which this work directly improves. For example, [BAHS06] generate police photofits from eyewitness accounts, but they rely on manual manipulation of 3DMM parameters. Instead, we propose a method which allows descriptive text to be directly mapped to the latent space of the 3DMM. In doing so, we enable photofits to be initialised directly from witnesses' textual descriptions or a sketch.

We also introduce new possibilities for the initialisation and refinement of 3D shapes for prosthesis design. In a recent study, Jablonski *et al.*[JMS+24] tested the use of 3DMMs for nasal prosthesis design, formulating the problem as one of shape completion given the rest of the face. They found the 3DMM method to outperform traditional CAD and manual scuplting methods for this task, without requiring expert maxillofacial clinicians [JMS+24]. Our method could further improve upon this approach, both in initialisation and text-based editing of the 3DMM used, to enable accurate and automated prosthesis design.

As identified in section 2.3, previous methods such as ClipFace [ATDN23] enable text-based editing of 3DMMs, but they do not consider the generation of a fully parameterised model of the human face, including identity, from text alone. In this work, for the first time, we bring together CLIP and 3DMMs to enable direct text to 3DMM generation, including identity. To do this, we propose a novel pipeline that enables us train a deep MLP, Text2Face, to map from CLIP embedding space to the space of 3DMMs. We then show the strong qualitative results of this conceptually simple approach.

We summarise the contributions of this chapter as follows:

- We introduce a novel method, Text2Face, which enables direct ini-

tialisation of 3DMM shape (identity and expression) from textual descriptions.

- We present a range of qualitative results demonstrating the power of this approach for shape initialisation.
- We demonstrate the use of our proposed method, Text2Face, to enable multi-modal fitting.

4.1 Proposed Method: Text2Face

Our method enables us to generate a fully parameterised 3D model of the human head, including identity, expression, and a detail map, from a single textual description. To do this, we first generate a dataset comprising pairs of CLIP embeddings and 3DMM parameters. Examples of the process to generate these paired embeddings is detailed in fig. 4.1. We then propose and train a mapping network, Text2Face, to map from these CLIP embeddings to their respective 3DMM parameters in the space of the FLAME head model.

4.1.1 Dataset Generation

To train a method to generate a full 3D head model from textual descriptions, we rely on a key intuition: shared text-image latent spaces, such as CLIP, allow us to bootstrap existing methods to generate new multi-modal datasets. We can then use these datasets to learn to directly regress from one modality to another. Here, we use this intuition to train reconstruction methods that enable us to directly regress parameterised 3D shape from text.

Face reconstruction from an image is a well-studied problem with many available methods. Consider a reconstruction method R. We can use R to recover a set of 3DMM parameters, $\mathbf{p}$, from a given image set $\mathbf{I}$. Using CLIP, we extract a set of CLIP embeddings $\mathbf{e}_{\text{CLIP}}$ from the image set $\mathbf{I}$.

In doing so, we have constructed a dataset of paired CLIP embeddings $\mathbf{e}_{\text{CLIP}}$ and their related 3DMM parameters $\mathbf{p}$. Following this, we can train a reconstruction method R' to map from CLIP embeddings $\mathbf{e}_{\text{CLIP}}$ to 3DMM parameters $\mathbf{p}$, thereby enabling text-to-3D generation.

We construct our image set $\mathbf{I}$ by synthesising 50,000 adult faces using StyleGAN2 [KLA+20b], selecting images estimated to be older than 18 using py-agender [But18]. A CLIP embedding $\mathbf{e}_{\text{CLIP}}$ is extracted from each image in $\mathbf{I}$ using the ViT-L/14-336px vision transformer model [RKH+21a].

We further estimate identity, pose, expression, and a detailed displacement map δ, for each image in FLAME model space [LBB+17] using DECA [FFBB21], a state-of-the-art method for monocular 3D face reconstruction. DECA is a self-supervised reconstruction method that has learned to recover and reconstruct a 3D head model (FLAME) and detailed facial geometry from a single image. This displacement map δ, introduced within DECA, enables mid-frequency details such as wrinkles to be represented on the FLAME mesh. An example face image $\mathbf{I}_{\text{example}}$, generated from StyleGAN2, and its corresponding 3D shape $\mathbf{p}_{\text{example}}$, estimated by DECA, are shown in fig. 4.2. Our full dataset generation pipeline is shown in fig. 4.1.

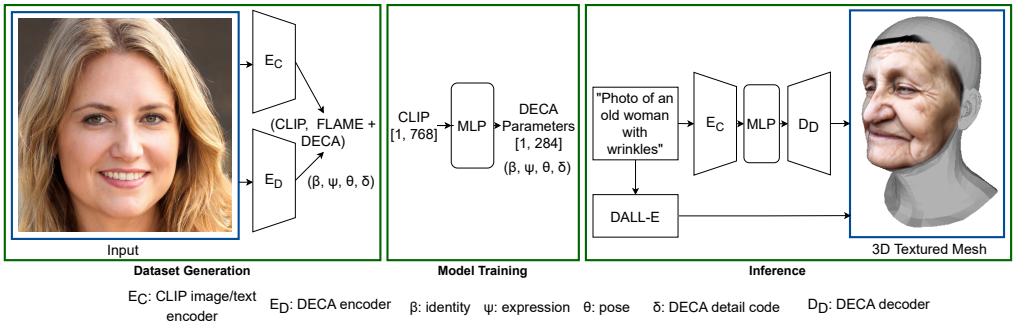
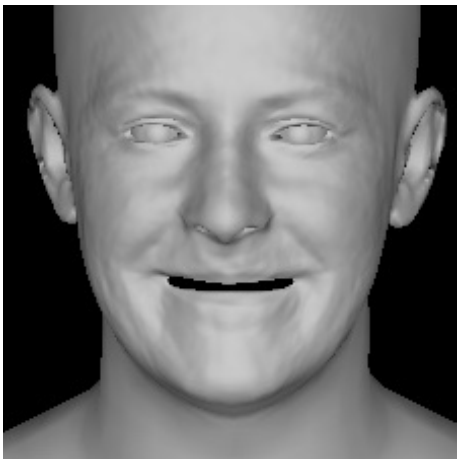


Figure 4.1: Dataset generation, model training, and inference for Text2Face.

(a) Image produced by StyleGAN2.



(b) Corresponding 3D mesh generated by DECA.

Figure 4.2: The Text2Face Dataset. An image and its corresponding 3D representation, estimated by DECA [FFBB21], are shown.

4.1.2 Mapping Network Architecture

We train a deep MLP, Text2Face, on this dataset to map the CLIP embedding space to the FLAME parameter space. At inference time, we take advantage of CLIP's interchangeable text-image latent space by using CLIP's text encoder on a sequence of input text. We supply this CLIP embedding as input to Text2Face which generates parameters for a fully parameterised 3D face. These are passed to the DECA decoder [FFBB21] to produce the 3D mesh. Hence, training Text2Face on embeddings extracted solely from images enables inference for both text and images. We further use DALL-E [RPG+21a] to generate an image for each text prompt. We extract a texture from this image using detected facial landmarks and paste it on to the mesh using the procedure supplied by Feng *et al.*[FFBB21]. The overall architecture is shown in Figure 4.3.

We use the Adam optimiser [KB14] with a learning rate of 1e-3 and a batch size of 64. We train for 100 epochs, using early stopping with a patience of 10.

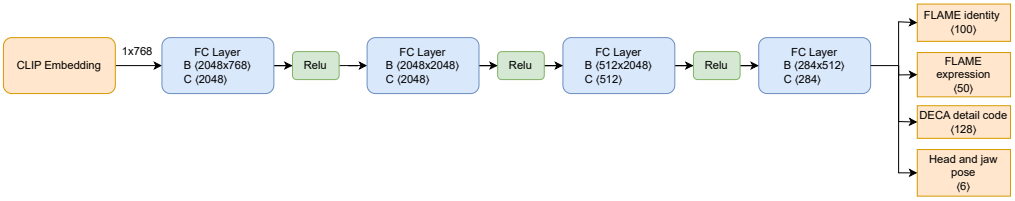


Figure 4.3: Architectural diagram of our Text2Face regressor.

4.2 Experiments

Figure 4.1 shows the resulting textured 3D mesh from the text prompt: “Photo of an old woman with wrinkles". Figure 4.4 shows detailed textureless meshes generated by the specified text prompts. The image generated by DALL-E from this same prompt, along with the subsequent textured mesh, are also shown. The pose and scale are estimated from the generated image using detected landmarks, and then applied to both meshes for visual comparison. Hence, the 3D mesh is aligned to the image as shown.

Figure 4.4: *(l)*: Prompt: “20 year old woman looking at the sky with surprise at UFO overhead". *(r)*: Prompt: “50 year old man looking grumpy". Each sub-figure, from left to right: Shape generated by the text prompt, DALL-E image from the same prompt, textured mesh.

4.2.1 Qualitative Results

In the following figures 4.5 to 4.11, we show three images. The first image shows the 3D mesh constructed from the 3DMM parameters regressed by Text2Face from the specified text prompt. The second image is generated by DALL-E [RPG+21a] from the same text prompt. The final image shows this generated image mapped onto the generated mesh as texture, again using the procedure supplied by Feng *et al.*[FFBB21]. The pose for the

meshes displayed here are estimated from the image generated by DALL-E; this allows for the texture to be fully displayed as shown. This full pipeline is implemented to enable a textured 3D mesh to be generated directly from a single text prompt.

The identity, expression, and detail code are all regressed directly by Text2Face from the text prompt. The texture is illustrative but presents one limitation: the predicted shape from Text2Face and the predicted texture from DALL-E are separate processes which lead to uncorrelated outputs. For example, given this pipeline, it is reasonable for the texture of 'happy man' to have eyes open while the mesh has closed eyes. Future work should consider combining these two generation steps into a single forward pass of a model.

Figure 4.5: Prompt: “Happy man".

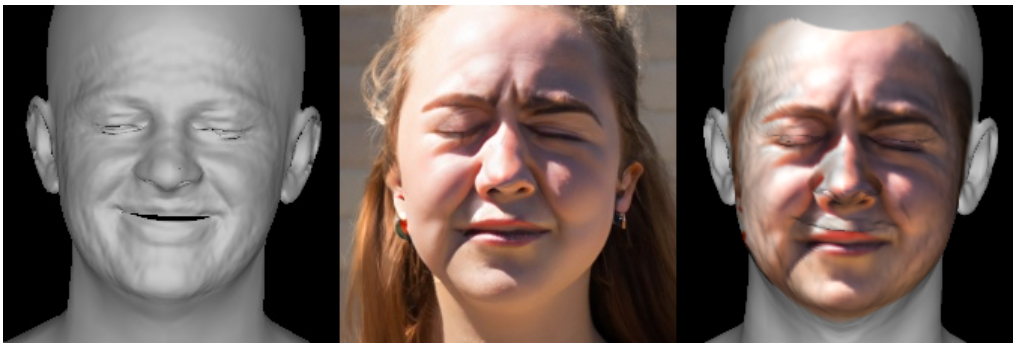


Figure 4.6: Prompt: “20 year old woman squinting at the sun".

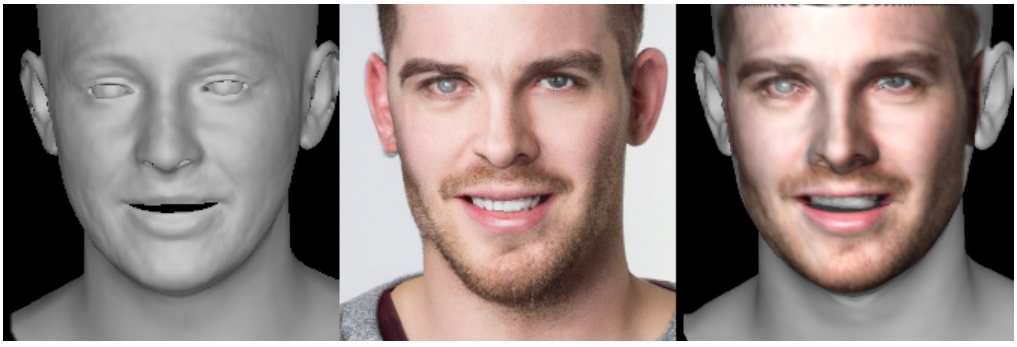


Figure 4.7: Prompt: “24 year old attractive man".

Figure 4.8: Prompt: “Photo of a woman screaming".

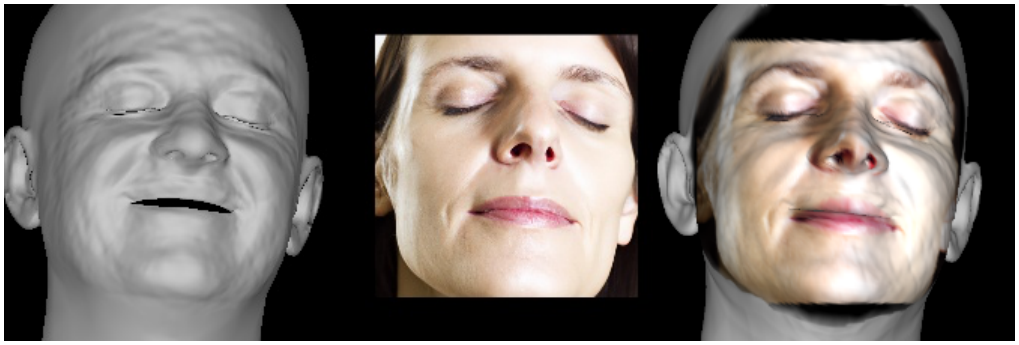


Figure 4.9: Prompt: “Photo of a woman with her eyes closed".

Figure 4.10: Prompt: “50 year old man looking happy after a long day working on the film set".

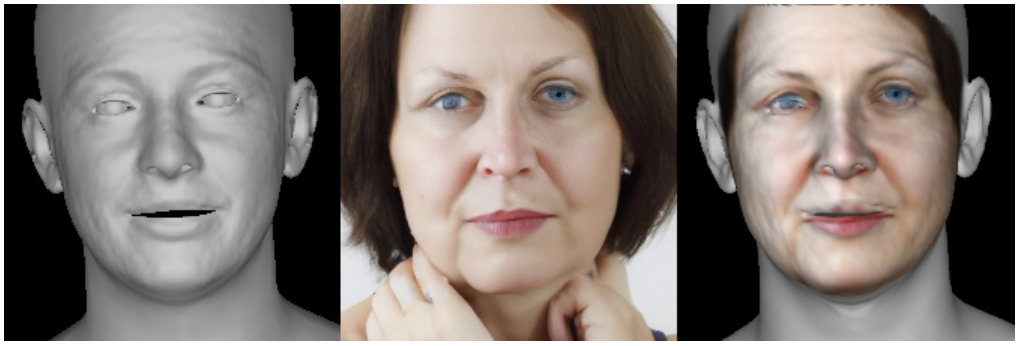


Figure 4.11: Prompt: “50 year old woman".

4.2.2 Multi-Modal Fitting

Here we further demonstrate the ability of Text2Face to enable multi-modal input fitting to a 3DMM. To do this, we consider three related images. We take an image of Robert De Niro and create sketch and sculpture versions of the image. We extract the CLIP embedding from each image using the ViT-L/14-336px vision transformer model [RKH+21a] and pass this as input to Text2Face which regresses the 3DMM parameters, including identity, expression, and a personal detail code.

The result is shown in Figure 4.12. We estimate face pose from the original image of De Niro, giving all meshes this same pose to enable direct comparison. We also show texture mapping results for all generated meshes, using the first image to generate this texture.

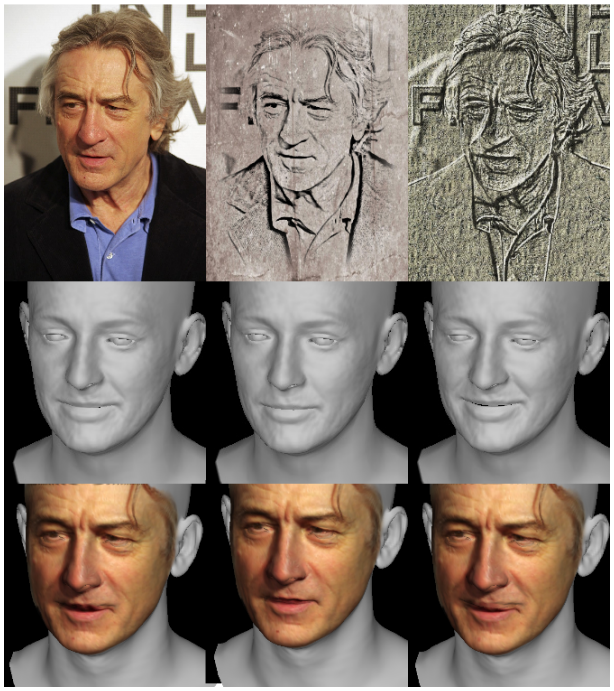


Figure 4.12: Robert De Niro fit to a 3DMM using Text2Face with the CLIP embedding extracted from each image as input: using the original image (left), a sketch (middle), and an engraving (right).

4.3 Conclusion

We have presented the first method to generate fully parameterised 3D face shape from textual descriptions. This capability not only enhances the practical applications of photofit specifications and avatar creation but also expands the utility of 3DMMs across various domains. By leveraging the power of CLIP embeddings, we have successfully bridged the gap between textual data and 3D face shape initialisation and modelling for the first time, providing a framework that can adapt to diverse input modalities and generate highly detailed facial reconstructions.

These text-to-shape capabilities fulfil a core objective of the Intelligent Face Agent proposed in Chapter 1, improving the controllability of 3D face reconstruction. We have also demonstrated that the multi-modal capabilities of CLIP can be inherited using our reconstruction pipeline, opening up a new field of research in multi-modal 3DMM fitting. Furthermore, the output representation of the FLAME head model will enable widespread use of this method for downstream tasks. Future work should consider inherited gender and racial biases from CLIP [AKC+21], their impact on 3D face generation, and how this can be minimised.

In the next chapter, we will consider how we can improve the explainability of 3D face reconstruction by introducing further intuitive methods to understand error. We introduce OptiFaces, a universal theoretical lower error bound from N face shapes; We will then consider theoretical lower error bounds from existing reconstruction methods before showing we can use the insights developed using these methods to improve state-of-the-art performance in 3D face reconstruction.

Chapter 5

New Baselines for 3D Face Reconstruction

Research Question 4

What new and more interpretable metrics can be developed for evaluating 3D face reconstruction?

Research Question 5

What theoretical lower bounds can we devise for 3D face reconstruction?

The accurate reconstruction of 3D facial shape and size from 2D images is a fundamental task in Computer Vision, with applications ranging from animation to facial recognition systems. However, current approaches often do little better than the mean face of existing 3D face models [SBFB19] in terms of 3D reconstruction error. In fact, the mean face from the FLAME head model [LBB+17] outperforms all pre-2019 approaches on the NoW benchmark [SBFB19]. This demonstrates how little information these previous methods are able to extract from an input image beyond a generic prior on face shape and provides an informative and intuitive baseline to beat.

We extend this idea from a single reference face (i.e. the mean face) to a

collection of N well-separated reference faces, which we call OptiFaces. We then consider the performance that can be achieved if we have a classifier that optimally performs face matching from an input image to the closest of these N faces. This enables intuitive judgements to be made about the degree of facial information learned by existing approaches. For example, if the error for 5 OptiFaces is state-of-the-art, one only needs to be able to consistently perform accurate face matching to 5 well-separated face shapes to achieve this performance. For face reconstruction benchmarks, where 3D shape is known for a given image, OptiFaces provide a new set of baselines to beat which makes quantitive performance more interpretable and helps us better distinguish between existing methods.

In this chapter, we provide a formal definition of OptiFaces and a method for calculating a set of OptiFaces from a 3D face dataset. We apply this method to calculate OptiFaces from the Headspace dataset [DPSD20b], and test these on the NoW validation set [SBFB19], providing a strong set of baselines for a state-of-the-art monocular 3D face reconstruction benchmark.

We then introduce the 3D Face Reconstruction Model Zoo, addressing the problem of combining existing 3D face reconstruction methods to improve reconstruction accuracy. We calculate theoretical lower error bounds for these combined methods on the NoW benchmark. In doing so, we provide a new, intuitive way of viewing reconstruction performance. For example, we can compare all pre-2023 methods combined versus the current state-of-the-art method on the NoW benchmark. In this case, TokenFace [ZCL+23] wins, demonstrating a significant improvement in performance that surpasses combinations of prior methods.

Both OptiFaces and the Model Zoo provide new, concise ways of evaluating and discussing 3D face reconstruction performance through the

calculation of theoretical lower bounds from empirical methods. Combined, they offer the research community a new set of tools to understand and analyse 3D reconstruction performance over time.

We summarise the contributions of this chapter as follows:

- We introduce OptiFaces, a universal baseline for 3D face reconstruction which considers the performance that can be achieved by optimally performing face-matching from an input image to a well-separated set of representative 3D face shapes.
- We propose and implement a greedy algorithm, the OptiFace Optimiser, to iteratively select a representative set of faces from a face database to minimise representative error.
- We mathematically define an idealised discrete classifier which we implement to calculate OptiFace errors on the NoW benchmark.
- We propose the 3D Face Reconstruction Model Zoo: a method to calculate theoretical lower error bounds that could be achieved by combining existing reconstruction methods. This offers a new way to understand the performance of existing methods and the potential benefits of combining their outputs.
- We present theoretical lower error bounds for combinations of existing approaches, including sets of models grouped by publication date. We then analyse and discuss these results.

5.1 Proposed Method: OptiFaces

We formulate the problem of finding OptiFaces as follows. Given a set of N face meshes, each represented by V vertices in a 3D space, our goal is to find a subset of X representative faces that minimise the total reconstruction

error. This error is quantified by the mean squared error (MSE) between the vertices of the N faces and the closest face within the the selected subset of X faces. This error metric is chosen as it is easy to calculate for any set of faces in an unsupervised manner. Formally, the problem is defined as follows:

Let $F = \{f_1, f_2, ..., f_N\}$ be the set of all face meshes in correspondence, where each face mesh f_i is represented by its vertices $V_i \in \mathbb{R}^{V \times 3}$ where V is the number of vertices and $v_{i,k}$ is the kth vertex in f_i.

The error function between two face meshes f_i and f_j may be defined as:

$$E(f_i, f_j) = \frac{1}{V} \sum_{k=1}^{V} \|v_{i,k} - v_{j,k}\|^2$$

The objective is then formulated as follows:

- Find a subset $S \subseteq F$ such that $|S| = X$ and $S_i \subseteq S_i + 1$ (greedy set member allocation).
- This subset S should minimise the total reconstruction error across all faces in F.
- The total error is expressed as:

$$\underset{S \subseteq F, |S|=X}{\operatorname{argmin}} \sum_{f_i \in F} \min_{f_j \in S} E(f_i, f_j)$$

We define an **idealised discrete classifier** as a classifier that always selects the nearest OptiFace. The idealised discrete classifier, denoted as $C(f_i, S)$, selects the face mesh f_j from subset S that minimises the error function E for a given face mesh f_i. Formally, the classifier is defined as:

$$C(f_i, S) = \underset{f_j \in S}{\operatorname{argmin}} E(f_i, f_j)$$

By calculating a representative set of faces from one dataset and then selecting the closest face for a given image using our idealised classifier, we establish new error baselines for existing reconstruction benchmarks. We

introduce a simple greedy algorithm to compute the set of representative faces in section 5.2 and implement an idealised discrete classifier for the NoW benchmark, which identifies the face with the lowest calculated error for a given image. Algorithm 1 and Algorithm 2 present pseudo-code for OptiFace calculation and selection using the idealised discrete classifier, respectively. We evaluate this on the NoW benchmark for 1, 5, and 10 OptiFaces, establishing strong new new baselines for 3D face reconstruction. Section 5.4 includes visualisations of the computed OptiFaces as meshes, within PCA space, and in comparison with those calculated using the k-means algorithm.

The k-means algorithm is an unsupervised clustering algorithm that groups inputs based on their similarity in feature space [IEA+23]. Initially, data points are randomly assigned to clusters, and at each step, cluster membership is updated by calculating the squared Euclidean distance between each point and the cluster centroids. Each data point is then reassigned to the nearest cluster. In our case, each data point represents a face in FLAME PCA space, with the centroids of each cluster corresponding to the k OptiFaces. We set the number of clusters, k, to be equal to the number of OptiFaces, and run k-means until convergence.

While k-means is quick to compute and effective for clustering based on feature similarity, it does not inherently guarantee the selection of the most diverse or representative faces from each cluster. In addition, it is highly dependent on the initialisation, is not deterministic in its selection, and requires full recomputation each time we wish to add another OptiFace. To address this, we employ a greedy algorithm, which allows us to sequentially select faces that deterministically minimise our error metric, using the set of O OptiFaces to select the $O + 1 - th$ OptiFace. The greedy approach ensures that the chosen faces are real, well-distributed faces across the

feature space, potentially providing a better representation of the variations in facial geometry than k-means alone, which can result in suboptimal selections if cluster centroids are too close to one another.

5.2 OptiFace Optimiser

To calculate our error metric for N OptiFaces, we must first calculate a set of OptiFaces from a set of face shapes as described in section 5.2.1, and then evaluate these representative faces on a face reconstruction benchmark, as outlined in section 5.3. The following section provides additional details on the practical implementation of these steps.

5.2.1 Selecting Representative Faces

We employ a greedy algorithm for iteratively building a set of OptiFaces. This algorithm assumes all input faces follow the same topology, a condition we meet by using a database of faces already registered to the FLAME head model. This alignment may also be completed as a pre-processing step, enabling this method to be applied to any set of faces. The pseudo-code for our proposed OptiFace Optimiser is presented in Algorithm 1.

We apply the OptiFace Optimiser to the Headspace dataset [DPSD20b]. The Headspace dataset is chosen because it is one of the largest available collections of 3D face shapes, featuring a wide range of age variations and a strong gender balance. This ensures that we are able to select representative faces that cover the wide natural variety of human face shape. On a single GTX 1080, the computation times for this algorithm are 1 minute for 1 OptiFace, 33 minutes for 5 OptiFaces, and 94 minutes for 10 OptiFaces.

Algorithm 1 Greedy algorithm for OptiFace calculation.

Require: Faces F, Desired number of OptiFaces O

Ensure: Subset of selected OptiFaces S, Indices of selected OptiFaces I

```
Calculate the mean face mean_face from F
Initialise S with the face closest to mean_face
Initialise I with the index of this face
while |S| < O do
    Set min_error to infinity
    for each face f in F not in S do
        Temporarily add f to S
        Calculate reconstruction error E for F given S
        if E < min_error then
            Update min_error to E
            Set best_face to f
        end if
    end for
    Add best_face to S
    Add the index of best_face to I
end while
return S, I
```

5.2.2 Idealised Discrete Classifier: Evaluating OptiFaces on a Benchmark

Now that we have a set of representative faces, we can use an implementation of our idealised discrete classifier to evaluate the performance of a method that always selects the optimal OptiFace. This algorithm is designed to evaluate the effectiveness of a set of OptiFaces in face reconstruction tasks, particularly within the context of benchmarks. It processes a dataset consisting of image pairs alongside their corresponding ground truth 3D

meshes, aiming to determine the OptiFace that most closely approximates each ground truth mesh. Psuedo-code for our discrete idealised classifier is presented in Algorithm 2.

We perform evaluation on the NoW benchmark [SBFB19], a standard benchmark for 3D face reconstruction from a single image. We selected it for its diversity in individuals, environments, and capture settings—ranging from neutral and expressions to selfie and occlusions. These uncontrolled settings more accurately represent real-world conditions, making NoW a challenging benchmark where our shape-only baselines enable us to better distinguish existing approaches.

NoW considers the evaluation of both metric and non-metric reconstruction. Metric evaluation assesses the error in real-world units, with the goal of methods being to recover the actual scale of the face as well as finer facial details. In both cases, the predicted mesh is rigidly aligned (rotation and translation) to the scan using a set of predicted landmarks, followed by further alignment using a scan-to-mesh distance, which is calculated for every vertex in the scan [SBFB19]. In non-metric evaluation, the initial alignment also includes scaling. We focus on non-metric evaluation in 5.1, as this type of error highlights the finer facial details recovered, rather than the coarse scale estimated by each method.

The summary statistics computed by the algorithm provide insights into the performance of efficient face-matching for facial reconstruction on the NoW benchmark. These performance profiles then serve as strong baselines with which to evaluate and compare against existing face reconstruction methods.

Algorithm 2 Benchmark-Based Idealised Classifier for OptiFace Selection

Require: Benchmark dataset D containing pairs of images and ground truth meshes, Set of representative OptiFaces S

Ensure: Summary statistics of errors for each image in D with the best matching OptiFace

```
function BenchmarkClassifier(D, S)
    Initialise an empty list errors to store errors for each image
    for each pair (image, gt_mesh) in D do
        Initialise min_error to infinity
        Initialise selected_face to null
        for each OptiFace f_j in S do
            Calculate error E between gt_mesh and f_j
            if E < min_error then
                Update min_error to E
                Update selected_face to f_j
            end if
        end for
        Add min_error to the list errors
    end for
    Compute summary statistics from errors
    return Summary statistics
end function
```

5.3 Experiments

We use the FLAME-registered faces [ZBT22] of the Headspace dataset [DPSD20b] as a database of 1,211 face shapes. We calculate OptiFaces for this using the algorithm outlined in 1. We then implement an idealised discrete classifier for the NoW validation set, using this to compute errors

for 1, 5, and 10 OptiFaces. The results are presented in table 5.1 alongside comparison 3D face reconstruction methods.

Method	Med.	Mean	Std.
Deep3D [DYX+19]	1.286	1.864	2.361
DECA (detail)	1.19	1.469	1.249
DECA [FFBB21]	1.178	1.464	1.253
AlbedoGAN (detail)	0.95	1.173	0.987
MICA [ZBT22]	0.913	1.130	0.948
AlbedoGAN [RGP+24]	0.903	1.122	0.957
1 OptiFace	1.527	1.859	1.524
5 OptiFaces	1.144	1.436	1.231
10 OptiFaces	1.125	1.416	1.226

Table 5.1: Reconstruction error (mm) on the validation set of the NoW benchmark [SBFB19] in non-metrical reconstruction. Comparison results are presented from [RGP+24].

With just 5 OptiFaces, we beat all pre-2022 face reconstruction methods. More precisely, all methods before that period are outperformed by an approach that can select the optimal face from just 5 unique face shapes. Headspace and the NoW benchmark are independent datasets, collected in different parts of the world. This suggests that OptiFaces generalise well across different distributions of face shape, encompassing demographic and geographical variations.

5.4 Visualisation of OptiFaces

In figure 5.1, we present a visualisation of the first 10 OptiFaces calculated using the OptiFace Optimiser and evaluated on the NoW benchmark. We show the first two principal components of shape from the FLAME head model, alongside all other faces in the Headspace dataset. This effectively

demonstrates the extensive coverage our OptiFaces provide for the two principal components of face shape that capture the most variation, out of a total of 300. The OptiFaces are numbered in the order they are calculated.

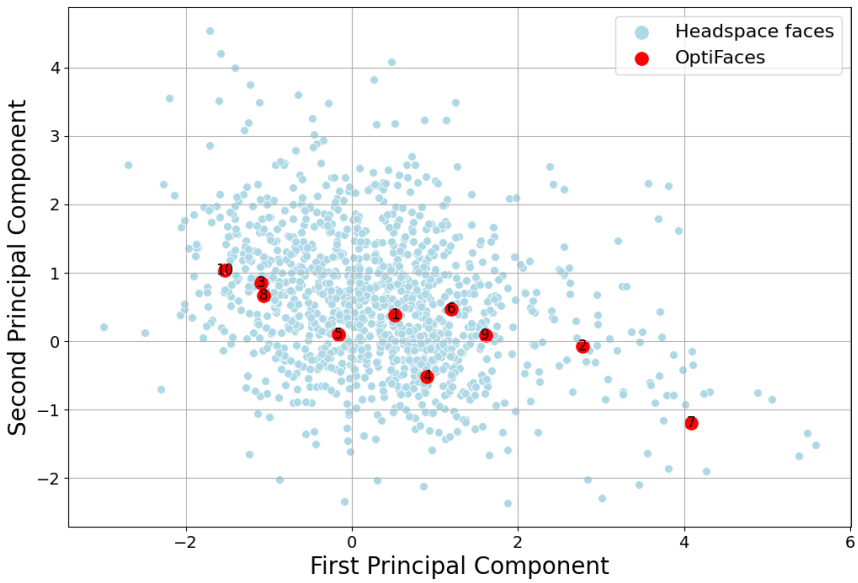


Figure 5.1: A visualisation of the first two principal components of FLAME shape for the first 10 OptiFaces calculated using the Headspace dataset. OptiFaces are numbered in the order they are calculated.

In figure 5.2, we compare the OptiFaces calculated using the k-means algorithm with those derived from the OptiFace Optimiser. It appears that the OptiFace Optimiser selects better distributed faces from the Headspace dataset for the first two principal components of face shape, thereby covering a wider range of potential face shapes. To perform this analysis, we applied the k-means algorithm to all Headspace faces within the PCA space of the FLAME head model. We used the implementation from scikit-learn [PVG+11], setting a fixed seed of 10 and running for 100 iterations. The centroids obtained from this process were then selected as our OptiFaces.

Figure 5.3 displays the ordered meshes of all 10 OptiFaces, which are also

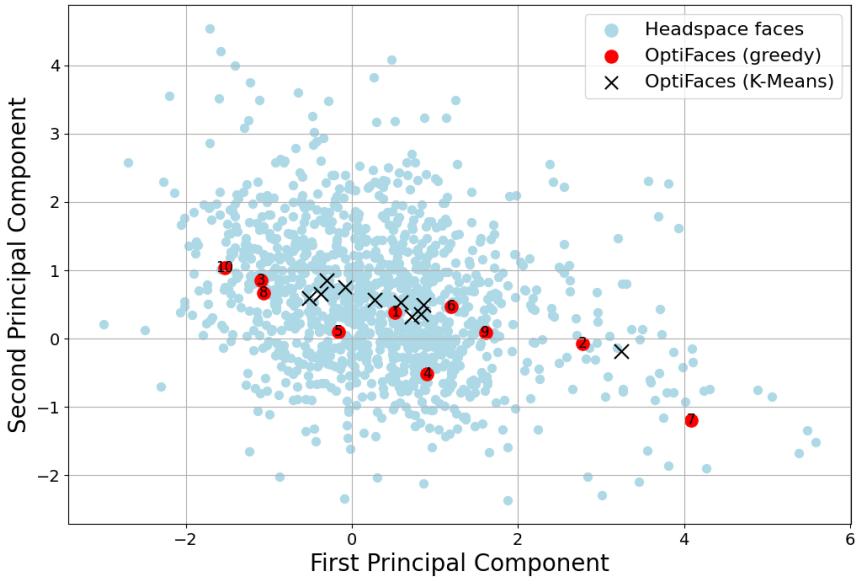
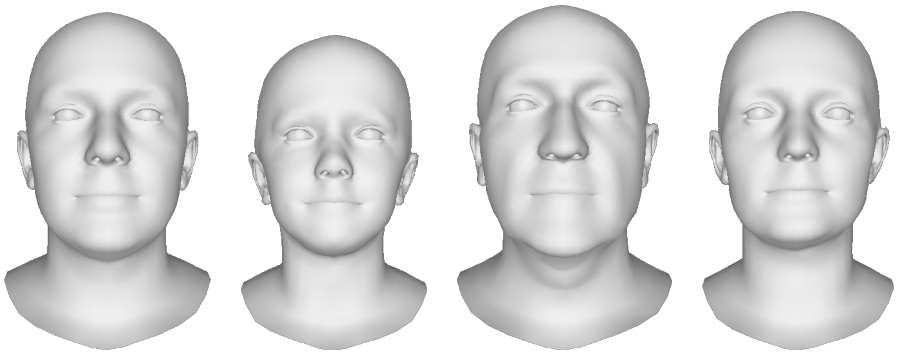
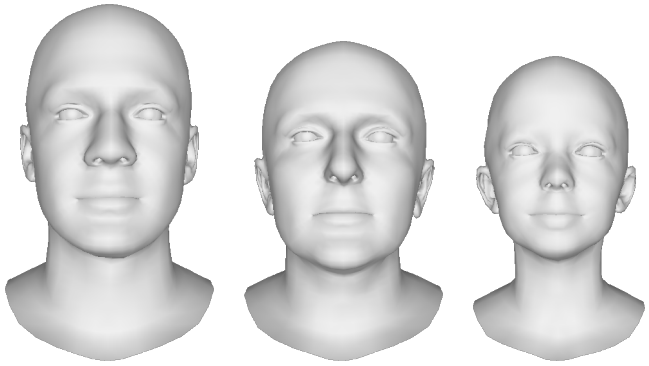


Figure 5.2: A comparison of the OptiFaces computed using the proposed OptiFace Optimiser and a k-means implementation.

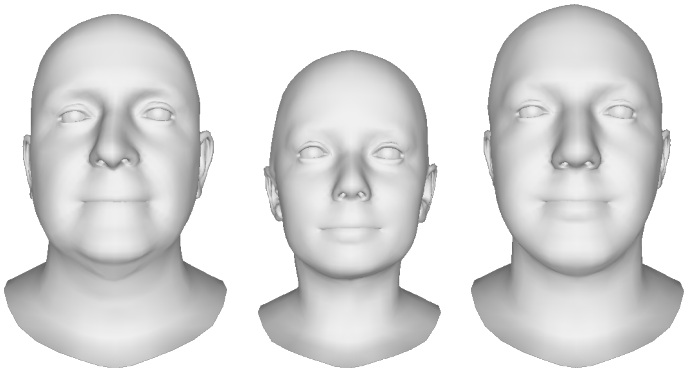
illustrated in figures 5.1 and 5.2 and evaluated in table 5.2. All images are rendered to a common scale using MeshLab [CCC+08] with a field of view (FOV) of 60. The ordering of the meshes demonstrates how the OptiFace Optimiser selects representative faces to minimise reconstruction error.



(a) 1st OptiFace (b) 2nd OptiFace (c) 3rd OptiFace (d) 4th OptiFace



(e) 5th OptiFace (f) 6th OptiFace (g) 7th OptiFace



(h) 8th OptiFace (i) 9th OptiFace (j) 10th OptiFace

Figure 5.3: OptiFace Heads Visualisation for the first 10 OptiFaces calculated using the OptiFace Optimiser

5.5 3D Face Reconstruction Model Zoo

It is a common saying that 'two heads are better than one'. This expression captures the intuition that two people considering the same problem are often better than one. Our 3D Face Reconstruction Model Zoo extends this idea to 'N heads are better than one' for 3D face reconstruction. This literal interpretation follows naturally from our work in OptiFaces. We propose that having N reconstruction methods, each producing a single reconstruction for a given input image, offers the opportunity to develop a challenging set of new baselines for 3D face reconstruction. Selecting the optimal reconstruction from these methods is analogous to choosing the best reconstruction network for a given input image. In this section, we simulate this optimal selection using an idealised discrete classifier, as shown in fig. 5.4, and report our findings.

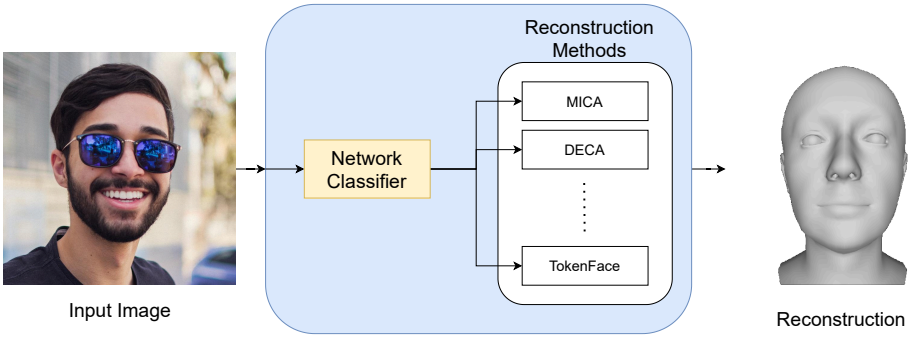


Figure 5.4: N Heads Are Better Than One Architecture. We evaluate multiple existing 3D face reconstruction methods (such as MICA, DECA, and TokenFace) for each input image. Our theoretical classifier selects the best reconstruction, enabling us to calculate lower error bounds for combinations of methods and establish new baselines for 3D face reconstruction performance.

The idea of combining 3D face reconstruction outputs has predominantly been applied to practically combining reconstructions of the same individual from multiple viewpoints to improve reconstruction accuracy. In these cases, the same network is used for reconstruction, with the 3DMM parameters

averaged.

To better understand the performance of current methods, we consider the optimal theoretical performance that could be achieved by always being able to optimally combine two or more existing reconstruction networks. We use the idealised discrete classifier, defined in 5.2.2, to achieve this by always selecting the reconstruction with the known lowest error for the input image. In doing so, we create the **3D Face Reconstruction Model Zoo**, a tool that can be used to combine any subset of existing approaches on leading benchmarks.

The Model Zoo improves our understanding of how networks can have complementary error profiles and helps us better understand the practical ramifications of their combination. We calculate performance values for combinations of networks on the test set of the NoW benchmark for both metric and non-metric reconstruction. This enables us to calculate an empirical theoretical lower error bound of combining any subset of reconstruction methods.

5.5.1 Combining Existing 3D Face Reconstruction Methods

To calculate the theoretical lower error bounds of combining existing methods, we adapt the idealised discrete classifier defined in 5.2.2. We use this classifier to always selects the optimal reconstruction network for a given input image. The idealised discrete classifier, $C(I, R)$, selects the reconstruction network R_i from set R that minimises the error function E for an input image I. Formally, the classifier is defined as:

$$C(I, R) = \underset{R_i \in R}{\operatorname{argmin}} E(I, R_i)$$

Where:

- $C(I, R)$ is the classifier that selects the optimal reconstruction network.

- I is the given input image.
- R is the set of reconstruction networks.
- R_i is a reconstruction network in set R.
- $E(I, R_i)$ is the error obtained by applying reconstruction network R_i to input image I.

This approach provides a lower bound on the error rate achievable by any practical method selection strategy, serving as a benchmark for evaluating real-world implementations and guiding future research. It is important to note that we only consider the selection of existing methods from their final outputs. In practice, combining existing methods at an earlier stage, such as at the input or at the feature-level, could result in performance uplifts with fundamentally different and perhaps improved profiles to those shown in this chapter.

5.5.2 Results

In table 5.2, we present results for the optimal combination of several existing methods. This includes the combination of all methods grouped by year. For example, '2023 & earlier' includes the combination of 18 existing networks. All methods with results presented by 1.5.2024 on the public NoW benchmark are considered, except for the network 3DFFA-V2 due to missing error data for several images. These time-grouped results are further visualised as error plots in fig. 5.6 and fig. 5.7.

We include a variety of results calculated by combining two existing networks. The names of these methods are derived from their parent methods. For example, DICA is a combination of DECA and MICA while TOCUS is a combination of TokenFace and FOCUS.

Our results demonstrate that combining multiple 3D face reconstruction methods can significantly reduce the reconstruction error. For instance,

TICA, which combines TokenFace and MICA, shows improved performance in metric reconstruction, although it offers only minimal improvements for non-metric reconstruction. This suggests that MICA is stronger than TokenFace for certain types of images in metric reconstruction where scale is the important factor but provides very limited benefit in recovering precise facial shapes, which is required in non-metric reconstruction. These insights from our results provide a starting point for future work to investigate the cases where this holds true.

These findings demonstrate the potential benefits of developing hybrid approaches that leverage the strengths of multiple existing methods to achieve more accurate 3D face reconstructions. Additionally, our results show significant improvements in the aggregate performance of methods over time and indicate where future performance gains can be made. These aspects are further explored in the following sections.

5.5.3 Relative Performance: Plot Analysis

The scatter plot in 5.5 illustrates the mean errors for MICA and TokenFace across all images in the NoW benchmark. Each point in the plot is a paired error value for each image. The red dashed line represents $y = x$, where both methods have equal error. Points below this line indicate that TokenFace has a lower error for those images, while points above the line indicate that MICA has a lower error. This plot provides insight into the relative performance of these two methods on a per-image basis and demonstrates the potential for combining existing reconstruction methods.

For methods most suited to combining, we would expect to see a scatter plot where points are distributed evenly around the $y = x$ line, demonstrating complementary performance. Intuitively, we can interpret this as: if one method is bad the other is good and vice-versa in large values and in equal proportion. For methods least suited to combining, we would see a plot

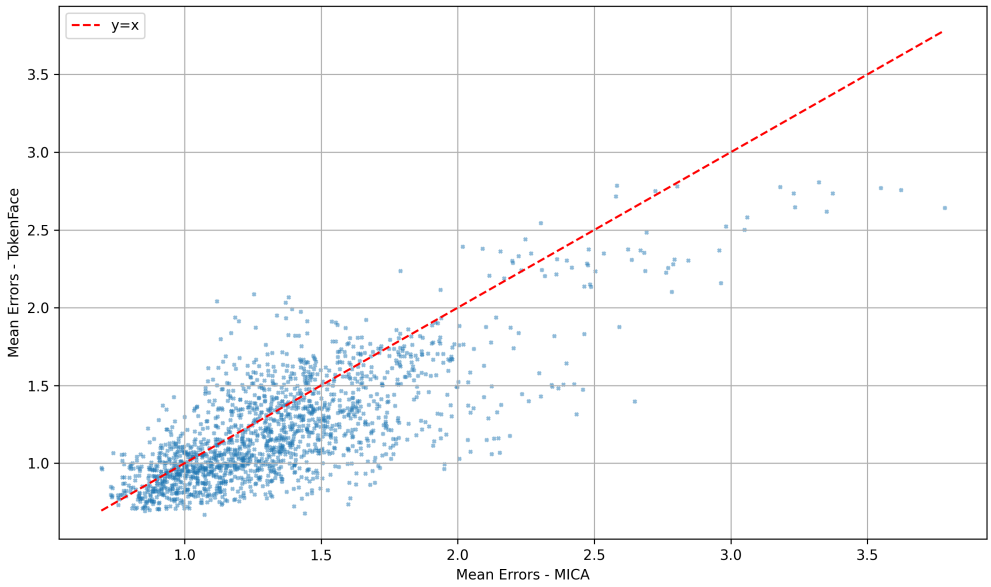


Figure 5.5: Scatter plot of mean errors for MICA and TokenFace across all images in the NoW benchmark in metric reconstruction. The red dashed line represents the line $y = x$. Points below the line indicate images where TokenFace performs better than MICA, while points above the line indicate images where MICA performs better than TokenFace.

where points cluster predominantly on one side of the line, indicating that one method consistently outperforms the other. In this case, the inferior method is unlikely to contribute much to overall performance.

This plot shows a significant number of points both below and above the line, indicating that neither method consistently outperforms the other across all images. This suggests that combining MICA and TokenFace could leverage their complementary strengths, reducing the overall reconstruction error. We find this holds true with TICA offering modest improvements over TokenFace alone, as shown in Table 5.2.

Method	Non-Metric			Metric (mm)		
	Median	**Mean**	**Std**	**Median**	**Mean**	**Std**
FLAME mean [LBB+17]	1.21	1.53	1.31	1.49	1.92	1.68
PRNet [FWS+18]	1.50	1.98	1.88	–	–	–
Deng *et al.** [DYX+19]	1.23	1.54	1.29	2.26	2.90	2.51
Deng *et al..*** [DYX+19]	1.11	1.41	1.21	1.62	2.21	2.08
RingNet [SBFB19]	1.21	1.53	1.31	1.50	1.98	1.77
MGCNet [SSL+20]	1.31	1.87	2.63	1.70	2.47	3.02
UMDFA [KS20]	1.52	1.89	1.57	2.31	2.97	2.57
Dib *et al..* [DTA+21]	1.26	1.57	1.31	1.59	2.12	1.93
DECA [FFBB21]	1.09	1.38	1.18	1.35	1.80	1.64
MICA [ZBT22]	0.90	1.11	0.92	1.08	1.37	1.17
FOCUS [LMFV+23]	1.04	1.30	1.10	1.41	1.85	1.70
TokenFace [ZCL+23]	0.76	0.95	0.82	0.97	1.24	1.07
DICA	0.87	1.09	0.92	1.04	1.34	1.17
TICA	0.75	0.94	0.82	0.93	1.20	1.05
FICA	0.86	1.08	0.92	1.04	1.33	1.17
TECA	0.75	0.95	0.83	0.95	1.22	1.07
TOCUS	0.75	0.95	0.83	0.95	1.21	1.07
2018 & earlier	1.42	1.85	1.73	3.91	4.84	4.02
2019 & earlier	0.99	1.27	1.12	1.37	1.86	1.72
2020 & earlier	0.97	1.26	1.19	1.26	1.72	1.65
2021 & earlier	0.94	1.23	1.15	1.14	1.51	1.39
2022 & earlier	0.80	1.02	0.93	0.93	1.20	1.06
2023 & earlier	0.72	0.93	0.84	0.86	1.12	0.99

Table 5.2: Results for both metric and non-metric reconstruction on the test set of the NoW benchmark. We compile results for existing methods and compare them with theoretical lower bounds for combinations of methods, both by date of publication and for novel combinations of existing methods such as DICA which is the lower error bound for combining MICA and DECA. For Deng *et al.** denotes the Tensorflow implementation while ** denotes the PyTorch version.

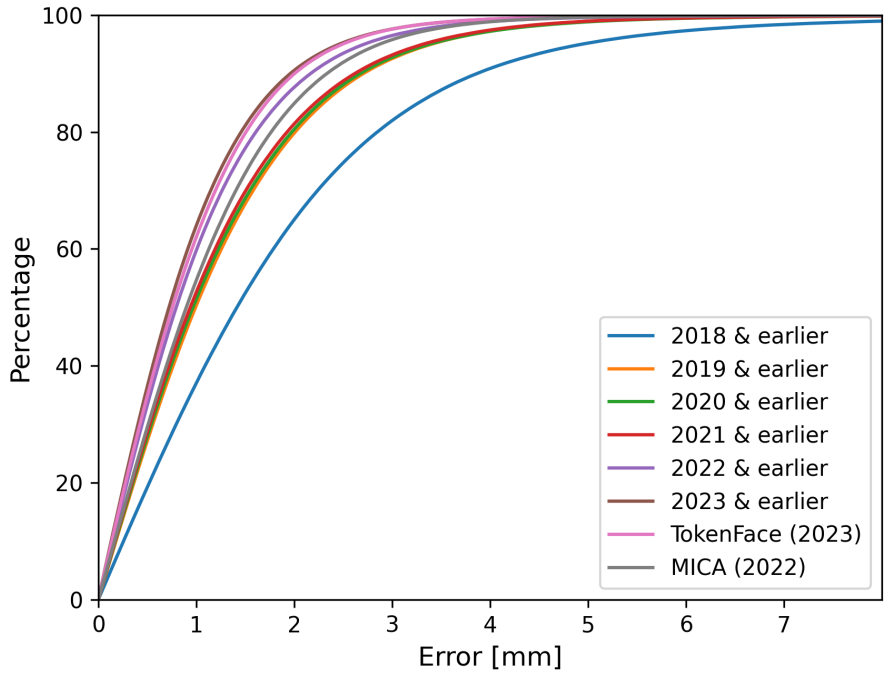


Figure 5.6: Error plots for the NoW benchmark in non-metric reconstruction. We compare errors from two leading approaches and the combination of all methods grouped by their year of publication. 18 methods are considered in total for the 80 meshes included within the test set.

Figures 5.6 and 5.7 illustrate the difference in performance for methods combined according to their date of publication. Each figure displays error curves with each curve representing the cumulative error distribution for methods published up to a given year. These figures clearly show a consistent and dramatic reduction in error over time, demonstrating improvements in 3D face reconstruction accuracy as newer methods are introduced. They further illustrate the substantial improvements made in metric reconstruction from 2017 with substantial progress being made each year while non-metric reconstruction errors appear to be more incremental in recent years.

Figures 5.8a and 5.8b show the performance of MICA combined with

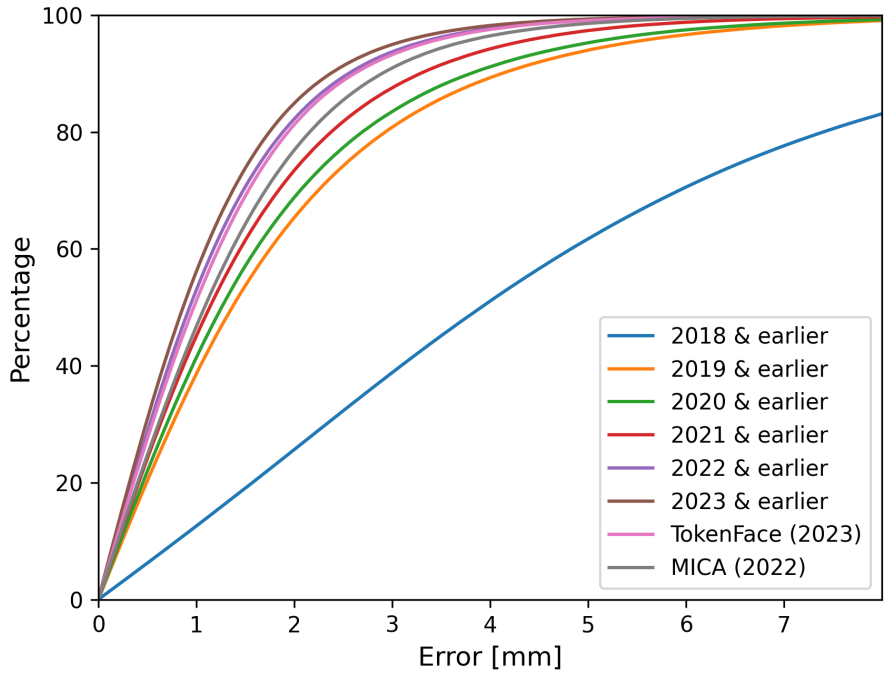
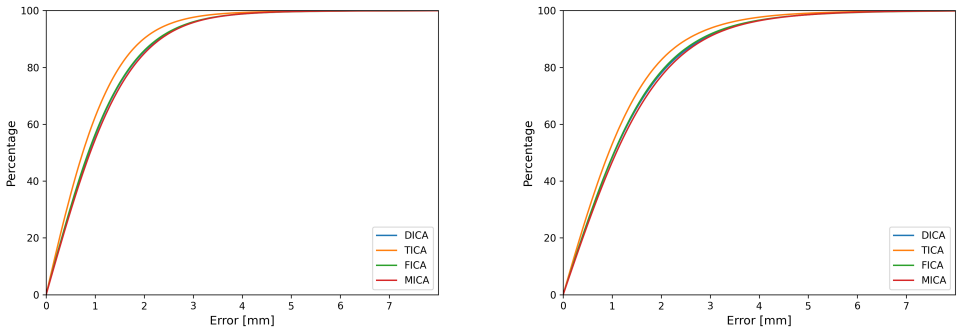


Figure 5.7: Error plots for the NoW benchmark in metric reconstruction. We compare errors from two leading approaches and the combination of all methods grouped by their year of publication. 14 methods are considered in total.

our newly-created variants: DICA, TICA, and FICA, for non-metric and metric reconstructions, respectively. These plots indicate that combining these methods yields lower errors compared to using MICA alone.

5.5.4 Decrease in Errors Over Time

In table 5.3, we expand upon the plots of progress shown in figures 5.6 and 5.7 to quantitatively assess error reductions over time. We do this for both metric and non-metric errors in 3D face reconstruction on the NoW benchmark. As before, we consider the errors for combined methods given by the year. For example, '2022 to 2023' refers to the performance difference between theoretically combining the performance of all methods released prior to 2022 with all methods released before 2023.



(a) (b)

Figure 5.8: Error plots for 3D face reconstruction: (a) non-metric reconstruction, and (b) metric reconstruction. Both plots compare MICA alone to combinations of MICA with other methods (DICA, TICA, and FICA), demonstrating potential reductions in error from their combination.

We observe improvements in both metric and non-metric reconstruction over time; however, we note that metric reconstruction has been the subject of the largest relative improvement since the introduction of the NoW benchmark. A cumulative drop in errors of 77.96% is reported for metric reconstruction, compared to 48.72% for non-metric reconstruction between 2018 and 2023.

It is important to note that several methods only report results for non-metric reconstruction as they are designed to recover facial details but not facial scale. This means that this is not a like-for-like comparison on the level of individual methods but a comparison in the aggregate of methods available at certain points in time.

5.6 Conclusion

We have presented OptiFaces, a novel baseline for 3D reconstruction techniques that selects face meshes that optimally minimise reconstruction error. We have demonstrated that a small number of OptiFaces offer a valuable baseline on the NoW benchmark and a novel way to evaluate reconstruc-

Year	Median (%)	Mean (%)	Std (%)	Cumulative (%)
Non-Metric Errors				
2018 to 2019	-30.28	-31.35	-35.26	-30.28
2019 to 2020	-2.02	-0.79	6.25	-31.66
2020 to 2021	-3.09	-2.38	-3.36	-33.77
2021 to 2022	-14.89	-17.07	-19.13	-43.91
2022 to 2023	-10.00	-8.82	-9.68	-48.72
Metric Errors				
2018 to 2019	-64.96	-61.57	-57.21	-64.96
2019 to 2020	-8.03	-7.53	-4.07	-67.48
2020 to 2021	-9.52	-12.21	-15.76	-70.81
2021 to 2022	-18.42	-20.53	-23.74	-76.21
2022 to 2023	-7.53	-6.67	-6.60	-77.96

Table 5.3: Percentage change in errors over time for non-metric and metric errors. Cumulative reductions, from 2018 to 2023, are calculated for the median errors reported. Here 'X to Y' refers to the performance difference between theoretically combining the performance of all methods released prior to X with all methods released before Y.

tion performance by connecting it to the associated classification problem. OptiFaces are easy to compute and provide a novel set of dataset-specific metrics for 3D face reconstruction, giving a more meaningful interpretation of 3D shape reconstruction error.

In addition, through calculating the theoretical performance uplift of combining networks, we have identified the gap between current state-of-the-art methods and the optimal theoretical performance achieved through combining existing methods. This gap reveals the potential for improvement in 3D face reconstruction techniques. The marked improvements over time, particularly in metric reconstruction, also demonstrate the progress that has been made within the field to achieve more accurate reconstruction.

The two approaches presented in this chapter, OptiFaces and the Model Zoo, offer complementary perspectives on the performance of existing reconstruction methods. They both consider theoretical performance through the use of an idealised discrete classifier to offer new baselines that are explainable and readily communicable, improving our understanding of existing methods. In doing so, they fulfil a core objective of the Intelligent Face Agent proposed in chapter 1.1 to develop new, more intuitive ways of evaluating and discussing 3D face reconstruction performance.

They also offer something rare in this field: a target to hit on existing benchmarks. Currently, the only targets that exist are single existing methods. If you beat those, where do you aim next? This work offers a suite of new targets to hit. Together, these approaches contribute to a deeper understanding of the strengths and limitations of existing 3D face reconstruction techniques and pave the way for future advancements in the field.

In the next chapter, we will revisit the original research questions of this thesis and consider how the past three technical chapters have addressed these questions. We will also reflect on the experience as a researcher in this field. Finally, we will discuss limitations of the work presented in this thesis and propose future work which will help further realise the aims of the proposed Intelligent Face Agent.

Chapter 6

Conclusions & Outlook

At the beginning of this thesis, we outlined a vision for a new method of interaction with 3D face representations as conceptualised in the Intelligent Face Agent (IFA). This vision was supported by a thesis research question and several subordinate research questions, each tied to core functionality of the proposed IFA.

In this chapter, we:

- Revisit the original research questions of this thesis in the context of our contributions, as presented in our technical chapters.
- Reflect on the current state of 3D face reconstruction methods.
- Discuss limitations of the work presented in this thesis and propose further work that will enable us to fully realise the promise of Intelligent Face Agents.

6.1 Revisiting the Thesis and Research Questions

This thesis has been organised around addressing the following research question:

Thesis
Can we connect existing sources of knowledge about the human face to improve the accuracy, controllability, and explainability of 3D face reconstruction?

The work presented in the three technical chapters of this thesis demonstrates that we can improve the accuracy, controllability, and explainability of 3D face reconstruction by connecting existing sources of knowledge about the human face. In Chapter 3, we introduce a method, the SynthFace Generator, for fast paired 2D-3D dataset generation. This is achieved by connecting a known source of 3D information about the human face, the FLAME head model, to the 2D information contained within Stable Diffusion 1.5 to generate SynthFace. The integration of text prompting within this dataset generation approach improves gender balance within SynthFace, enhancing the controllability of dataset generation methods. We then demonstrate that competitive performance can be achieved through training a regression network, ControlFace, using this dataset.

In Chapter 4, we presented the first method for directly generating fully parameterised 3D heads from textual descriptions. To achieve this, we connected the image-text latent space of CLIP to the latent space of the FLAME head model, enabling initialisation from text in addition to images. This improved the controllability of 3D face reconstruction.

Finally, in Chapter 5, we improve the explainability of 3D face reconstruction by proposing OptiFaces, a new method for generating baselines for existing reconstruction benchmarks by considering the theoretical optimal performance of the corresponding classification problem for N well-separated face shapes.

This high-level research question for the thesis is then decomposed into

several related research questions. We will now use these research questions to highlight key contributions made by this thesis.

Research Question 1

Can we exploit knowledge about the structure and appearance of the human face, contained within pretrained image generation networks, to generate large-scale datasets for 3D face reconstruction?

Addressed in Chapter 3

In Chapter 3, we successfully exploited knowledge about the structure and appearance of the human face, contained within pretrained image generation networks, to generate large-scale datasets for 3D face reconstruction. The SynthFace Generator connects the FLAME head model, a known source of 3D facial structure information, with the 2D image generation capabilities of Stable Diffusion 1.5. This approach removes the need for manual asset creation, enabling the generation of photorealistic face images paired with 3D shapes. SynthFace, the resulting dataset, is the largest of its kind and offers unique opportunities to disentangle shape from identity, improving both the accuracy and the controllability of 3D face reconstruction models. The effectiveness of this method was demonstrated through strong qualitative results and competitive quantitative results using ControlFace, a network trained exclusively on the SynthFace dataset.

Research Question 2

What methods can be employed to increase the diversity of paired training data for 3D face estimation methods?

Addressed in Chapter 3

Chapter 3 addresses this research question by showing that the integ-

ration of text prompts within the SynthFace Generator can significantly enhance the diversity of the generated dataset. By using textual appearance descriptors, we can guide the generation process to include a more balanced representation of genders, increasing the proportion of individuals estimated to be female in the dataset from 16.9% to 41.8%. We are also able to use text-based conditioning to balance by race. In addition, our method enables a wide variety of viewing angles, head shapes, and occlusions to be specified by a change in parameter values to the SynthFace Generator. In doing so, our method enables the creation of more diverse and representative datasets, which are crucial for training robust and fair 3D face reconstruction models.

Research Question 3
Are we able to perform parameterised shape initialisation of faces from natural language descriptions?
Addressed in Chapters 3 and 4

Chapters 3 and 4 demonstrate that it is indeed possible to perform parameterised shape initialisation of faces from natural language descriptions. In Chapter 4, we introduced a method that connects the image-text latent space of CLIP with the FLAME head model, enabling the generation of fully parameterised 3D heads from textual descriptions. This method enhances the controllability of 3D face reconstruction by allowing initialisation from text inputs, thus expanding the utility of 3D Morphable Models (3DMMs). Chapter 3 supports this by showing how these text-based initialisations can be integrated into the dataset generation process, providing a more versatile approach to creating training data.

Research Question 4

What new and more interpretable metrics can be developed for evaluating 3D face reconstruction?

Addressed in Chapter 5

In Chapter 5, we introduced Optifaces, a metric that offers a novel way to evaluate reconstruction performance by linking it to the associated classification problem. This metric is concise and easily communicable, providing a clear and straightforward assessment of 3D face reconstruction methods.

Research Question 5

What theoretical lower bounds can we devise for 3D face reconstruction?

Addressed in Chapter 5

Chapter 5 addresses this research question by introducing OptiFaces, a novel baseline that provides theoretical lower bounds for 3D face reconstruction. OptiFaces selects face meshes that optimally minimise reconstruction error. This method offers a new set of dataset-specific metrics for 3D face reconstruction, giving a meaningful interpretation of reconstruction error. By considering the theoretical optimal performance for well-separated face shapes, OptiFaces provides a robust baseline against which current and future 3D face reconstruction methods can be evaluated.

Chapter 5 also considers the theoretical performance which may be achieved by combining existing reconstruction networks. Using the NoW Model Zoo, we construct a variety of methods, such as DICA which is a combination of MICA and DECA, in order to better understand current performance and provide new baselines for the future. We find that

TokenFace, the state-of-the-art method on the NoW benchmark beats all pre-2023 methods combined in non-metric reconstruction, demonstrating a significant improvement in the field with the introduction of this approach.

Both of these methods broaden our understanding of the limits of current techniques and offer informed guidance for future research directions.

6.2 Personal Reflections

The period in which this research has been conducted has coincided with a rapid new wave of AI research and public awareness of our work. This becomes clear when you consider that the majority of the ideas and methods discussed in this thesis are based on works that didn't exist at the start of my PhD in October 2020. CLIP, which is the foundation for Text2Face presented in Chapter 4, was introduced in January 2021. Stable Diffusion [RBL+22], which is used in Chapter 3, was introduced in December 2021, while ControlNet [ZRA23] was proposed in February 2023.

The rapid pace of research publications in this field and the wider interest in AI from the general public have led to a unique set of pressures on researchers. There is a constant need to stay updated with the latest developments and to adapt quickly to new methodologies and tools. I believe novel ways of connecting existing tools, such as in Chapters 3 and 4, and new methods for evaluation, as proposed in Chapter 5, are important contributions at times of rapid iteration like these. They enable us to experience the benefits of these new tools and ideas while offering a realistic view of their performance in key tasks such as 3D face reconstruction.

The proliferation of tools, both in deployment and free public access on the web and through mobile apps, increases the responsibility on us as researchers. This is a period where public engagement is vital, and dialogue needs to be two-way. As a community, we can learn as much, if not more,

from other disciplines and the general public than they can learn about Computer Vision and AI from us. All deployed systems exist and operate in a context that is often far removed from a Computer Science building and those working inside. For my part, I have written extensively throughout my PhD for student newspapers, satirical magazines, and pursued a wide range of interests in arts and culture. I have come to understand that people are often fascinated by the possibilities of this work but rightfully sceptical about many proposed applications.

With advancement seemingly so rapid—image generation is an excellent example, with blocky 32x32 generations [VdOKE+16] now reaching photorealistic 1080p generation [PEL+23] in just a few years—it would be easy to assume this will continue like Moore's law for the number of transistors on a chip. However, compute is finite, our architectures are likely sub-optimal, and the future is less predictable at times of fast change than at times of steady progress. The tools we already have can be used for good, and we should consider what exactly we mean by progress. Measures such as performance-per-watt [SGM19] offer a chance to prioritise efficiency, both for the environment and for the neatness of efficient methods. If progress stopped tomorrow, as a field, we would still have many busy years ahead.

6.3 Future Work

In this thesis, we have made significant research contributions to three core sub-tasks of the Intelligent Face Agent: dataset generation, 3D shape initialisation from text, and new baselines for the evaluation of 3D shape reconstruction. However, to fully realise the potential of the proposed Agent, further work should consider integrating all modalities (text, 2D, 3D) into a shared latent space and training models in an end-to-end manner.

Recent work, such as 4M [MBK+23], offers a promising approach for any-to-any generation for related Computer Vision tasks. In this work, Mizrahi *et al.*[MBK+23] unify the representation space of many different yet complementary vision modalities. They achieve this by mapping these modalities to tokens and training a single transformer encoder-decoder to perform cross-modal prediction from a single masked modality. Future work could consider developing masked multimodal any-to-any methods specifically for faces. By focusing on a single object category, we could integrate grounded 3D representations such as the FLAME head model to enable iterative reconstruction and editing across all modalities.

While we have considered new baselines for 3D shape error, there are many evaluation questions raised by the multimodal nature of this work. The original diagram of the proposed IFA (Figure 1.1) includes text as an output modality with the suggested use of an expert system for Q & A with generated output shapes. This would necessitate new methods of evaluation and extensive user testing to ensure this functionality is both accurate and reliable.

Intelligent Face Agents are designed to be practical implementations that are useful in the real-world. An early example presented in this thesis considered the case of automated prosthesis design. There is traditionally a large gap between a published research paper and a product but further evaluation metrics and clinical studies of these methods would move us further in that direction. This is important as the research presented here offers incredible opportunities to democratise access to life-changing treatment options such as facial prostheses.

Appendix A

SynthFace Supplementary

A.1 Textual Appearance Descriptor

The textual appearance descriptor is composed of a positive prompt and a negative prompt. These are passed to ControlNet 1.1 to condition the output of Stable Diffusion 1.5.

For occlusions:

- Positive prompt: {age_prompt} {race_prompt} {gender_prompt} {occlusion_prompt}, {outdoor_prompt}, profile picture, dslr

For neutral, expressions, and selfies:

- Positive prompt: {age_prompt} {race_prompt} {gender_prompt}, {outdoor_prompt}, profile picture, dslr

The negative prompt remains the same for all classes of generated images.

- Negative prompt: worst quality, normal quality, low quality, low res, blurry, text, watermark, logo, banner, extra digits, cropped, jpeg artifacts, signature, username, error, sketch, duplicate, ugly, monochrome, horror, geometry, mutation, disgusting

Bibliography

[ABF+07] Brian Amberg, Andrew Blake, Andrew Fitzgibbon, Sami Romdhani, and Thomas Vetter. Reconstructing high quality face-surfaces using model based stereo. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007.

[AKC+21] Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. Evaluating clip: towards characterization of broader capabilities and downstream implications. *arXiv preprint arXiv:2108.02818*, 2021.

[ATDN23] Shivangi Aneja, Justus Thies, Angela Dai, and Matthias Nießner. Clipface: Text-guided editing of textured 3d morphable models. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023.

[Ayo19] Sajjad Ayoubi. Facelib, 2019.

[BAHS06] Volker Blanz, Irene Albrecht, Jörg Haber, and H-P Seidel. Creating face models from vague mental images. In *Computer Graphics Forum*, volume 25, pages 645–654. Wiley Online Library, 2006.

[BDBM11] Andrew D Bagdanov, Alberto Del Bimbo, and Iacopo Masi.

The florence 2d/3d hybrid face dataset. In *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding*, pages 79–80, 2011.

[BG18] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91, 2018.

[BKV+19] Valentin Bazarevsky, Yury Kartynnik, Andrey Vakunov, Karthik Raveendran, and Matthias Grundmann. Blazeface: Sub-millisecond neural face detection on mobile gpus. *arXiv preprint arXiv:1907.05047*, 2019.

[BRZ+16] James Booth, Anastasios Roussos, Stefanos Zafeiriou, Allan Ponniah, and David Dunaway. A 3d morphable model learnt from 10,000 faces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5543–5552, 2016.

[BS19] Anil Bas and William AP Smith. What does 2d geometric information really tell us about 3d face shape? *International Journal of Computer Vision*, 127:1455–1473, 2019.

[BT17] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE international conference on computer vision*, pages 1021–1030, 2017.

[But18] Michael Butlitsky. py-agender, 2018.

[BV99] Volker Blanz and Thomas Vetter. A morphable model for

the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999.

[CCC+08] Paolo Cignoni, Marco Callieri, Massimiliano Corsini, Matteo Dellepiane, Fabio Ganovelli, Guido Ranzuglia, et al. Meshlab: an open-source mesh processing tool. In *Eurographics Italian chapter conference*, volume 2008, pages 129–136. Salerno, Italy, 2008.

[CSW+16] Menglei Chai, Tianjia Shao, Hongzhi Wu, Yanlin Weng, and Kun Zhou. Autohair: Fully automatic hair modeling from a single image. *ACM Transactions on Graphics*, 35(4), 2016.

[CWLZ13] Chen Cao, Yanlin Weng, Stephen Lin, and Kun Zhou. 3d shape regression for real-time facial animation. *ACM Transactions on Graphics (TOG)*, 32(4):1–10, 2013.

[CWW+20] Yajing Chen, Fanzi Wu, Zeyu Wang, Yibing Song, Yonggen Ling, and Linchao Bao. Self-supervised learning of detailed 3d face reconstruction. *IEEE Transactions on Image Processing*, 29:8696–8705, 2020.

[CZR+22] Zenghao Chai, Haoxian Zhang, Jing Ren, Di Kang, Zhengzhuo Xu, Xuefei Zhe, Chun Yuan, and Linchao Bao. Realy: Rethinking the evaluation of 3d face reconstruction. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII*, pages 74–92. Springer, 2022.

[DBB22] Radek Daněček, Michael J Black, and Timo Bolkart. Emoca: Emotion driven monocular face capture and animation. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 20311–20322, 2022.

[DBK+20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[dCDSB+19] Fernanda Pereira de Caxias, Daniela Micheline Dos Santos, Lisiane Cristina Bannwart, Clovis Lamartine de Moraes Melo Neto, and Marcelo Coelho Goiato. Classification, history, and future prospects of maxillofacial prosthesis. *International journal of dentistry*, 2019, 2019.

[DGV+20] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5203–5212, 2020.

[DGXZ19] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

[DN21] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[DPD17] Hang Dai, Nick Pears, and Christian Duncan. A 2d morphable model of craniofacial profile and its application to craniosynostosis. In *Medical Image Understanding and Analysis: 21st Annual Conference, MIUA 2017, Edinburgh, UK, July 11–13, 2017, Proceedings 21*, pages 731–742. Springer, 2017.

[DPD+22] Christian Duncan, Nick E Pears, Hang Dai, Will AP Smith, and Paul O'Higgins. Applications of 3d photography in craniofacial surgery. *Journal of Pediatric Neurosciences*, 17(Suppl 1):S21–S28, 2022.

[DPSD20a] Hang Dai, Nick Pears, William Smith, and Christian Duncan. Statistical modeling of craniofacial shape and texture. *International Journal of Computer Vision*, 128(2):547–571, 2020.

[DPSD20b] Hang Dai, Nick Pears, William Smith, and Christian Duncan. Statistical modeling of craniofacial shape and texture. *International Journal of Computer Vision*, 128(2):547–571, 2020.

[DTA+21] Abdallah Dib, Cedric Thebault, Junghyun Ahn, Philippe-Henri Gosselin, Christian Theobalt, and Louis Chevallier. Towards high fidelity monocular face reconstruction with rich reflectance using self-supervised learning and ray tracing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12819–12829, 2021.

[DYX+19] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set.

In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2019.

[EST+20] Bernhard Egger, William AP Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, et al. 3d morphable face models—past, present, and future. *ACM Transactions on Graphics (TOG)*, 39(5):1–38, 2020.

[FFBB21] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4):1–13, 2021.

[FRPP+23] Panagiotis P Filntisis, George Retsinas, Foivos Paraperas-Papantoniou, Athanasios Katsamanis, Anastasios Roussos, and Petros Maragos. Spectre: Visual speech-informed perceptual 3d facial expression reconstruction from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5744–5754, 2023.

[FWS+18] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 534–551, 2018.

[GCM+18] Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T Freeman. Unsupervised training for 3d morphable model regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8377–8386, 2018.

[GGU+20] Partha Ghosh, Pravir Singh Gupta, Roy Uziel, Anurag Ranjan, Michael J Black, and Timo Bolkart. Gif: Generative interpretable faces. In *2020 International Conference on 3D Vision (3DV)*, pages 868–878. IEEE, 2020.

[GMFB+18] Thomas Gerig, Andreas Morel-Forster, Clemens Blumer, Bernhard Egger, Marcel Luthi, Sandro Schönborn, and Thomas Vetter. Morphable face models-an open framework. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 75–82. IEEE, 2018.

[HDL16] David Ha, Andrew Dai, and Quoc V Le. Hypernetworks. *arXiv preprint arXiv:1609.09106*, 2016.

[HZP+22] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *arXiv preprint arXiv:2205.08535*, 2022.

[IEA+23] Abiodun M Ikotun, Absalom E Ezugwu, Laith Abualigah, Belal Abuhaija, and Jia Heming. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, 622:178–210, 2023.

[JCB+23] Rachael Y Jablonski, Trevor J Coward, Paul Bartlett, Andrew J Keeling, Chris Bojke, Sue H Pavitt, and Brian R Nattress. Improving facial prosthesis construction with contactless scanning and digital workflow (impressed): study protocol for a feasibility crossover randomised controlled trial of digital versus conventional manufacture of facial

prostheses in patients with orbital or nasal facial defects. *Pilot and Feasibility Studies*, 9(1):110, 2023.

[JMS+24] Rachael Y Jablonski, Taran Malhotra, Daniel Shaw, Trevor J Coward, Farag Shuweihdi, Chris Bojke, Sue H Pavitt, Brian R Nattress, and Andrew J Keeling. Comparison of trueness and repeatability of facial prosthesis design using a 3d morphable model approach, traditional computer-aided design methods, and conventional manual sculpting techniques. *The Journal of Prosthetic Dentistry*, 2024.

[KB14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[KJ21] Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1548–1558, 2021.

[KLA+20a] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

[KLA+20b] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

[KPL+19] Amlan Kar, Aayush Prakash, Ming-Yu Liu, Eric Cameracci, Justin Yuan, Matt Rusiniak, David Acuna, Antonio Torralba, and Sanja Fidler. Meta-sim: Learning to generate synthetic datasets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4551–4560, 2019.

[KR18] Mohammad Rami Koujan and Anastasios Roussos. Combining dense nonrigid structure from motion and 3d morphable models for monocular 4d face reconstruction. In *Proceedings of the 15th ACM SIGGRAPH European Conference on Visual Media Production*, pages 1–9, 2018.

[KS20] Tatsuro Koizumi and William AP Smith. "look ma, no landmarks!"–unsupervised, model-based dense face alignment. In *European Conference on Computer Vision*, pages 690–706. Springer, 2020.

[LBB+17] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017.

[LHCZ20] Kun-Chan Lan, Min-Chun Hu, Yi-Zhang Chen, and Jun-Xiang Zhang. The application of 3d morphable model (3dmm) for real-time visualization of acupoints on a smartphone. *IEEE Sensors Journal*, 21(3):3289–3300, 2020.

[LLSD19] Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*, 2019.

[LMFV+23] Chunlu Li, Andreas Morel-Forster, Thomas Vetter, Bernhard Egger, and Adam Kortylewski. Robust model-based face

reconstruction through weakly-supervised outlier segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 372–381, 2023.

[LSSS18] Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. Deep appearance models for face rendering. *ACM Transactions on Graphics (ToG)*, 37(4):1–13, 2018.

[LTL+20] Jie Liang, Huan Tu, Feng Liu, Qijun Zhao, and Anil K Jain. 3d face reconstruction from mugshots: Application to arbitrary view face recognition. *Neurocomputing*, 410:12–27, 2020.

[LYSZ20] Jiangke Lin, Yi Yuan, Tianjia Shao, and Kun Zhou. Towards high-fidelity 3d face reconstruction from in-the-wild images using graph convolutional networks. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 5891–5900, 2020.

[LZZ+18] Feng Liu, Ronghang Zhu, Dan Zeng, Qijun Zhao, and Xiaoming Liu. Disentangling features in 3d face shapes for joint face reconstruction and recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5216–5225, 2018.

[MBH+23] CY Ma, NA Beck, MZ Hockaday, CJ Niedziela, CA Ritchie, JA Harris, E Roudnitsky, PKR Guntaka, SY Yeh, J Middleton, et al. The global distribution of oral and maxillofacial surgeons: a mixed-methods study. *International Journal of Oral and Maxillofacial Surgery*, 2023.

[MBK+23] David Mizrahi, Roman Bachmann, Oğuzhan Fatih Kar, Teresa Yeo, Mingfei Gao, Afshin Dehghan, and Amir Zamir. 4M: Massively multimodal masked modeling. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[MBOL+22] Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. Text2mesh: Text-driven neural stylization for meshes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13492–13502, 2022.

[MPS+11] AA Mueller, Pascal Paysan, Ralf Schumacher, H-F Zeilhofer, B-I Berg-Boerner, Juerg Maurer, Thomas Vetter, Erik Schkommodau, Philipp Juergens, and Katja Schwenzer-Zimmerer. Missing facial parts computed by a morphable model and transferred directly to a polyamide laser-sintered prosthesis: an innovation study. *British journal of oral and maxillofacial surgery*, 49(8):e67–e71, 2011.

[PB16] Marcel Piotraschke and Volker Blanz. Automated 3d face reconstruction from multiple images using quality measures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3418–3427, 2016.

[PEL+23] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[PGM+19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer,

James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[PKA+09] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *2009 sixth IEEE international conference on advanced video and signal based surveillance*, pages 296–301. Ieee, 2009.

[PVG+11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[PVO+20] Stylianos Ploumpis, Evangelos Ververas, Eimear O'Sullivan, Stylianos Moschoglou, Haoyang Wang, Nick Pears, William AP Smith, Baris Gecer, and Stefanos Zafeiriou. Towards a complete 3d morphable model of the human head. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4142–4160, 2020.

[RBL+22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[RDN+22] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu,

and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[RGP+24] Aashish Rai, Hiresh Gupta, Ayush Pandey, Francisco Vicente Carrasco, Shingo Jason Takagi, Amaury Aubel, Daeil Kim, Aayush Prakash, and Fernando De la Torre. Towards realistic generative 3d face models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3738–3748, 2024.

[RHPK23a] Will Rowan, Patrik Huber, Nick Pears, and Andrew Keeling. Fake it without making it: Conditioned face generation for accurate 3d face shape estimation. *arXiv preprint arXiv:2307.13639*, 2023.

[RHPK23b] Will Rowan, Patrik Huber, Nick Pears, and Andrew Keeling. Text2face: 3d morphable faces from text. *The First Tiny Papers Track at ICLR 2024*, 2023.

[RHPK24a] Will Rowan, Patrik Huber, Nick Pears, and Andrew Keeling. How many optifaces? a new evaluation metric for 3d face reconstruction. In *The Second Tiny Papers Track at ICLR 2024*, 2024.

[RHPK24b] William Rowan, Patrik Huber, NE Pears, and Andrew Keeling. N heads are better than one: Exploring theoretical performance bounds of 3d face reconstruction methods. In *European Conference on Computer Vision Workshop (ECCVw) 2024: Foundation Models for 3D Humans*. Springer Science+ Business Media, 2024.

[RKH+21a] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[RKH+21b] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[RPG+21a] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[RPG+21b] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[RSK16] Elad Richardson, Matan Sela, and Ron Kimmel. 3d face reconstruction by learning from synthetic data. In *2016 fourth international conference on 3D vision (3DV)*, pages 460–469. IEEE, 2016.

[SBFB19] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J Black. Learning to regress 3d face shape and expression from an image without 3d supervision. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7763–7772, 2019.

[SCL+22] Aditya Sanghi, Hang Chu, Joseph G Lambourne, Ye Wang, Chin-Yi Cheng, Marco Fumero, and Kamal Rahimi Malekshan. Clip-forge: Towards zero-shot text-to-shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18603–18613, 2022.

[SDWMG15] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

[SGM19] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in nlp. *arXiv preprint arXiv:1906.02243*, 2019.

[SKP15] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[SRK17] Matan Sela, Elad Richardson, and Ron Kimmel. Unrestricted facial geometry reconstruction using image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1576–1585, 2017.

[SRN+03] Martin A Styner, Kumar T Rajamani, Lutz-Peter Nolte, Gabriel Zsemlye, Gábor Székely, Christopher J Taylor, and Rhodri H Davies. Evaluation of 3d correspondence meth-

ods for model building. In *Information Processing in Medical Imaging: 18th International Conference, IPMI 2003, Ambleside, UK, July 20-25, 2003. Proceedings 18*, pages 63–75. Springer, 2003.

[SSL+20] Jiaxiang Shang, Tianwei Shen, Shiwei Li, Lei Zhou, Mingmin Zhen, Tian Fang, and Long Quan. Self-supervised monocular 3d face reconstruction by occlusion-aware multi-view geometry consistency. In *European Conference on Computer Vision*, pages 53–70. Springer, 2020.

[SWZ+23] Keqiang Sun, Shangzhe Wu, Ning Zhang, Zhaoyang Huang, Quan Wang, and Hongsheng Li. Cgof++: Controllable 3d face synthesis with conditional generative occupancy fields. *IEEE transactions on pattern analysis and machine intelligence*, 2023.

[TEB+20] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. Stylerig: Rigging stylegan for 3d control over portrait images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6142–6151, 2020.

[TTHMM17a] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gérard Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5163–5172, 2017.

[TTHMM17b] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gérard Medioni. Regressing robust and discriminative 3d morphable

models with a very deep neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5163–5172, 2017.

[TZG+18] Ayush Tewari, Michael Zollhöfer, Pablo Garrido, Florian Bernard, Hyeongwoo Kim, Patrick Pérez, and Christian Theobalt. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2549–2559, 2018.

[TZK+17] Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 1274–1283, 2017.

[TZS+16] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 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, 2016.

[VdOKE+16] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29, 2016.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia

Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[WBC+19] Fanzi Wu, Linchao Bao, Yajing Chen, Yonggen Ling, Yibing Song, Songnan Li, King Ngi Ngan, and Wei Liu. Mvf-net: Multi-view 3d face morphable model regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 959–968, 2019.

[WBH+21] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J Cashman, and Jamie Shotton. Fake it till you make it: face analysis in the wild using synthetic data alone. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3681–3691, 2021.

[WBH+22] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Matthew Johnson, Jingjing Shen, Nikola Milosavljević, Daniel Wilde, Stephan Garbin, Toby Sharp, Ivan Stojiljković, et al. 3d face reconstruction with dense landmarks. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII*, pages 160–177. Springer, 2022.

[ZBT22] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *European conference on computer vision*, pages 250–269. Springer, 2022.

[ZCL+23] Tianke Zhang, Xuangeng Chu, Yunfei Liu, Lijian Lin, Zhendong Yang, Zhengzhuo Xu, Chengkun Cao, Fei Yu, Changyin Zhou, Chun Yuan, et al. Accurate 3d face reconstruction with facial component tokens. In *Proceedings of the IEEE/CVF*

International Conference on Computer Vision, pages 9033–9042, 2023.

[ZRA23] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. pages 3836–3847, 2023.

[ZTG+18] Michael Zollhöfer, Justus Thies, Pablo Garrido, Derek Bradley, Thabo Beeler, Patrick Pérez, Marc Stamminger, Matthias Nießner, and Christian Theobalt. State of the art on monocular 3d face reconstruction, tracking, and applications. In *Computer graphics forum*, volume 37, pages 523–550. Wiley Online Library, 2018.

[ZYY+15] Xiangyu Zhu, Junjie Yan, Dong Yi, Zhen Lei, and Stan Z Li. Discriminative 3d morphable model fitting. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 1, pages 1–8. IEEE, 2015.

[ZZ18] Meng Zhang and Youyi Zheng. Hair-gans: Recovering 3d hair structure from a single image. *arXiv preprint arXiv:1811.06229*, 2018.