Reconstruction of Human Faces from Voice

Submitted in partial fulfillment of the requirements for

the degree of

Doctor of Philosophy

in

Electrical and Computer Engineering

Yandong Wen

B.S., Electronic and Information Engineering, South China University of Technology
M.S., Electronic and Information Engineering, South China University of Technology

Carnegie Mellon University
Pittsburgh, PA

May 2022

Acknowledgements

First and foremost, I would like to thank my advisor Rita Singh (chair of the doctoral committee), and co-advisor Bhiksha Raj, for their immerse support, patience, and encouragement. They have moulded me into the researcher I am today. I could not have asked for better advisors, and I will miss just walking in to their office to talk about anything.

I would like to thank my doctoral committee members: Prof. Bhiksha Raj, Prof. Yaser Sheikh and Prof. Fernando De la Torre. I am indebted for their valuable feedback, constructive suggestions, and technical discussions on my dissertation.

I would like to thank the members of Robust-MLSP research group, Prof. Richard Stern, and colleagues Yang Gao, Wenbo Zhao, Wenbo Liu, Yangyang Xia, Ankit Shah, Mahmoud Al Ismail, Tianyan Zhou, Anurag Kumar, Hira Dhamyal, Shahan Ali Memon, Tyler Vuong, Benjamin Elizalde, Suyoun Kim, Joana Correia, and Mark Lindsey. I had never imagined a more enjoyable graduate school life before I joined CMU.

I appreciate my internship days at Facebook Reality Labs. My thanks go to my mentor Alexander Richard, and colleagues Michael Zollhoefer, Dejan Markovic, and Lele Chen.

I have been fortunate enough to collaborate with many talented researchers: Weiyang Liu, Kaipeng Zhang, Zhiding Yu, Ming Li, Yu Qiao, and Zhifeng Li. Their thoughts, intentionally or unintentionally, consciously or unconsciously, enlighten my mind.

I would like to thank my Pittsburgh family for the wonderful memories: Chenchen Zhu, Yan Xu, Xi Liu, Zhiqian Qiao, Yang Zou, Jianxiao Ge, Yutong Zheng, Ran Tao, Felix Juefei Xu, Ligong Han, Han Zhao, Boyuan Yang, Wen Wang, Zechun Liu, Huilian Qiu, Zhe Gao, Chaoyang Wang, Shuangjia Xue, and Leo.

This thesis is dedicated to those pillars of my life, who have believed in me and supported me: my parents and my sister.

This work has been partly funded by the Defense Science and Technology Agency, Singapore under contract number A025959. Its content does not reflect the position or policy of DSTA and no official endorsement should be inferred.

Abstract

Voices and faces play pivotal roles in our social interactions. Despite their different physical manifestations, voices and faces contain highly similar types of information, including linguistic information (phonemes for voice and viseme for faces), affective state, and identity characteristics (weight, gender, age, etc.). For this reason, the associations between voices and faces have gathered significant research interest in psychology, cognitive science, artificial intelligence, and many other fields.

In this thesis, we attempt to explore the identity associations between voices and faces by developing computational mechanisms for reconstructing faces from voices. More specifically, the task is designed to answer the question: Given an unheard audio clip spoken by an unseen person, can we algorithmically picture a face that has as many associations as possible with the speaker, in terms of identity?

The link between voice and face has been established from many perspectives. Direct relationships include the effect of the underlying skeletal and articulator structure of the face and the tissue covering them, all of which govern the shapes, sizes, and acoustic properties of the vocal tract that produces the voice. Less directly, the same genetic, physical, and environmental influences that affect the development of the face also affect the voice. Given these demonstrable dependencies, it is reasonable to hypothesize that it may be possible to reconstruct faces from voice signals algorithmically. Our hypothesis is that *if any facial parameter influences the speaker's voice, its effects on the voice must be discoverable by a properly designed computational model.*

This thesis presents how we approach the goal of generating faces from voices in three stages. First, we consider the cross-modal matching problem: given a voice recording, one must select the speaker's face from a gallery of face images. To this end, we propose disjoint mapping networks to learn representations of voices and faces in a shared space, such that their representations can be compared to one another. The results of matching empirically demonstrate the possibility of disambiguating faces from the voice. Second, we

address the problem of reconstructing 2D face images from voices. We propose a simple but effective computational framework based on generative adversarial networks (GANs). The generated face images are visually plausible and have identity associations with the true speaker. Last, we investigate the problem of reconstructing 3D facial shapes from voices. We propose an anthropometry-guided framework that identifies which anthropometric measurements (AMs) are predictable from voice, and then reconstructs the 3D facial shapes from those predictable AMs. Compared to baseline methods, our results demonstrate notable improvements, especially in reconstructing the shapes of speakers' noses.

Contents

List of Tables

List of Figures

Chapter 1

Introduction and Background

Voices and faces play pivotal roles in our social interactions. They are special in human neural selectivity [12]. Neurocognitive studies have shown that listening to a voice or viewing a face engages psychological and neural mechanisms that are not engaged by other categories of nonvocal sounds or other object categories [13, 14]. Despite their different physical manifestations, voices and faces contain highly similar types of information, including linguistic information (phonemes for voice and viseme for faces [15]), affective state, and identity characteristics (weight, gender, age, etc.). For this reason, the associations between voices and faces have gathered significant research interest in psychology [16, 17, 18], cognitive science [19, 20], and many other fields.

Researchers in artificial intelligence have also attempted to investigate such audiovisual associations algorithmically. A number of computational models have been developed for animating given faces with the linguistic information and affective state from voice signals [21, 22, 23]. Such animated faces are loosely referred to as "talking faces". The fact that this can be done demonstrates that voices can be accurately associated to the motions and shapes of lip, tongue, teeth, jaw, etc. Research shows that the performance of talking faces can be further improved by leveraging facial landmarks as intermediate representations [24, 25, 26]. Compared to linguistic and affective information, identity associations between voices and faces have received comparatively little scientific attention.

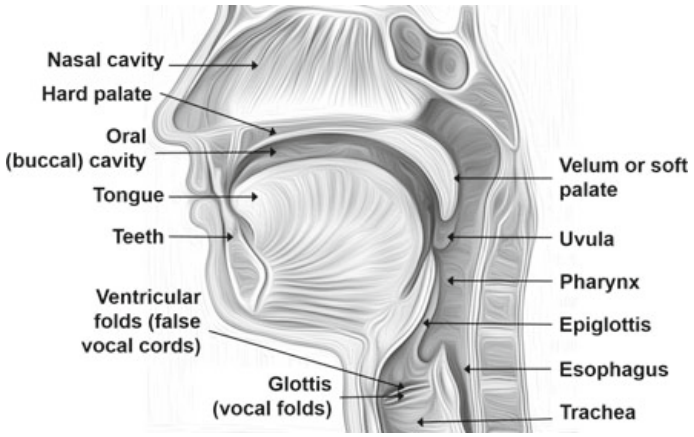


Figure 1.1: Anatomy of the vocal tract.

In this thesis, we attempt to explore the identity associations between voices and faces by developing computational mechanisms for reconstructing faces from voices. More specifically, the task is designed to answer the question: Given an unheard audio clip spoken by an unseen person, can we algorithmically picture a face that has as many associations as possible with the speaker, in terms of identity? It seems magical or impossible, but we as humans do this all the time. When we hear a song, we may imagine what the singer looks like without being aware of it. Or when we talk to someone over telephone, we may form a metal picture of the person we are talking with.

1.1 Motivation

A person's voice is incontrovertibly predictive of their face. Direct relationships stem from the voice production mechanism [27, 28], as shown in Fig. 1.1. The underlying skeletal and articulator structures of the face and the tissue covering them, govern the shapes, sizes, and acoustic properties of the vocal tract that produces the voice. The structures related to the nose affect the nasalence and nasal resonances [29, 30].

Less directly, demographic factors affect the face morphology as well as the voice pitch. For example, aging not only changes the facial appearance (*e.g.* skin texture), but also changes the tissue composition of the vocal cords and the dimensions of the vocal chambers

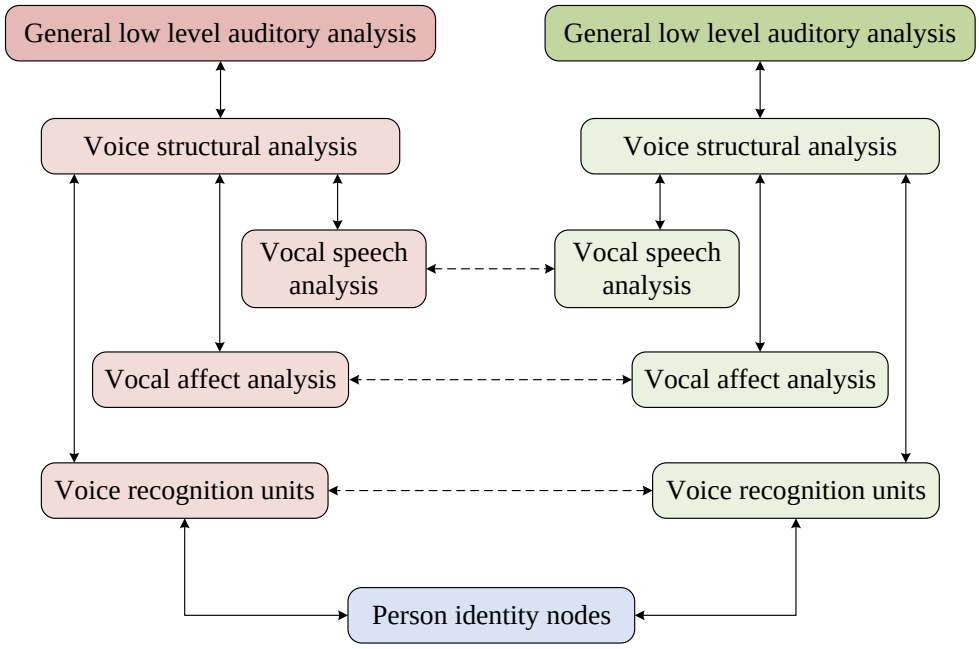


Figure 1.2: A model for voice and face perception. Reproduced from [1]

[31]. Gender also affects voice. Biologically, with higher testosterone levels, males are prone to have a prominent eyebrow ridge, broad chin, small eyes, and thin lips [32], while the vocal folds situated in the larynx also increase in thickness and length, leading to a lower voice pitch [33].

Neurocognitive studies indicate that neuro-cognitive pathways for voices share a common structure with that for faces [34, 1, 35] – the two may follow parallel pathways within a common recognition framework, as shown in Fig. 1.2. In empirical studies humans have demonstrated the ability to associate voices of unknown individuals to pictures of their faces [19, 18]. They have been observed to show improved ability to memorize and recall voices when the pictures of the speaker's face (but not imposter faces) are previously shown to them [36, 37, 38].

Given these demonstrable dependencies, it is reasonable to hypothesize that it may be possible to reconstruct faces from voice signals algorithmically. Our hypothesis is that *if any facial parameter influences the speaker's voice, its effects on the voice must be discoverable by a properly designed computational model.*

1.2 Related Work

In the past few years, much effort has been devoted to the audiovisual association learning. SVHF [39] presents a cross-modal matching setting, where given a voice recording and many face images, one must choose a face image such that the chosen face matches the speaker of the voice recording. A similar setting of matching a face image to many voice recordings is also included. Results show that the correct voice or face can be chosen with more than 80% accuracy in a two-alternative forced-choice setting. Furthermore, experiments also show that the matching accuracy is greatly dependent on demographic factors such as gender and age. This problem is formulated as metric learning in [40, 41]. [42] adopts self-supervised learning for the voice and face matching problem, so that a predictive model can be trained on unlabeled video data.

Recently, Speech2Face [43] has been proposed to generate 2D face images from voice. In this study, the authors have produced average-looking faces and these faces are shown to have consistent ages and facial measurements with the those of real speakers. [44] produce more realistic face images from voices using conditional generative adversarial networks (cGANs) [45, 46] in a closed-set setting. [47] extends the cGANs-based framework with advanced generator architecture, achieving improved quality for the generated images.

1.3 Roadmap

In this thesis, we explore the problem of reconstructing faces from voice and approach the research objective step by step. Specifically, we start with a cross modal selection problem. We develop a disjoint mapping network (DIMNet) to map a voice recording to the face image of the speaker, or vice versa. While matching is not reconstruction, this can be considered as a nonparametric approach for generating faces from voice. The selected face *is* the reconstruction result. Moreover, voice to face matching can be accurately evaluated objectively – we can quantify how accurate the "reconstruction" is. So we believe this is a

good starting point.

However, we also see the limitation of the matching problem, which relies heavily on a large-scale high-quality gallery set. Motivated by this, we work on the problem of 2D face image reconstruction from voice, where we develop a voice to face model based on generative adversarial networks (GANs). For unheard voices spoken by unseen persons, we obtain perceptually plausible results. In a test based on human judgment, we show that the generated faces do have identity associations with the true speaker.

Nonetheless, the generated images include many identity-*un*related factors, like the hair, hat, illumination, background, etc. We cannot control their presence or absence. These limitations motivate us to work on reconstructing 3D facial shapes from voices. The solutions are twofold. First, we propose a self-supervised approach for 3D face reconstruction from videos. This method can then be used to construct an audiovisual dataset, which comprises paired voices and 3D facial shapes. Secondly, we propose a "3D facial shape from voice" (SfV) approach based on anthropometric measurements (AMs). With SfV, we discover many facial AMs that are predictable from voice, and we can reconstruct 3D facial shapes from these measurements. Compared to the baseline methods, our results demonstrate notable improvements, especially in reconstructing the noses of female speakers.

1.4 Applications

This work has a number of potential applications, including those for law enforcement (*e.g.* supporting facial sketches generated by police artists based on eyewitness testimonies with more objective and accurate technologies), forensics, health services, gaming, entertainment, etc. As it becomes more accurate, more uses for it will emerge. Here we introduce two real world applications which have adopted the techniques in this work.

Voice profiling. Voice profiling refers to the process of deducing personal characteristics and bio-relevant parameters from their voices [48, 49], like gender, ethnicity, emotion, height, weights, and so on. Our work enables a new physical entity – the face – to be derived from

Figure 1.3: A live demonstration of voice profiling.

voice. We demonstrated a live system for voice profiling at the World Economic Forum in Tianjin, China from September 18 to 20, 2018. A snapshot of the system is shown in Fig. 1.3. It was tested by nearly a thousand people within a span of 3 days. The system demonstrated made multiple profile deductions from voice, and also recreated the speaker's 3D facial shape in virtual reality. People found it to be surprisingly accurate even in its then nascent state.

Public education. We worked with Wunderman Thompson, a digital marketing and advertising company based in New York with 200 offices in 90 markets and over 20,000 employees, on a project for revealing the "faces behind fraud". This project attempts to create videos, where scammers tell the truth about their scams and reveal how they con people out of money. Previously, such videos were made with the voices and faces from selected voice actors reading a script. With the work we have done in this thesis, we can now use the voices of real scammers to generate their faces. In other words, we can now generate videos that are made using *real scammers' voices* and *the faces generated from them* by our techniques. We present some examples of faces generated from real scammers' voices

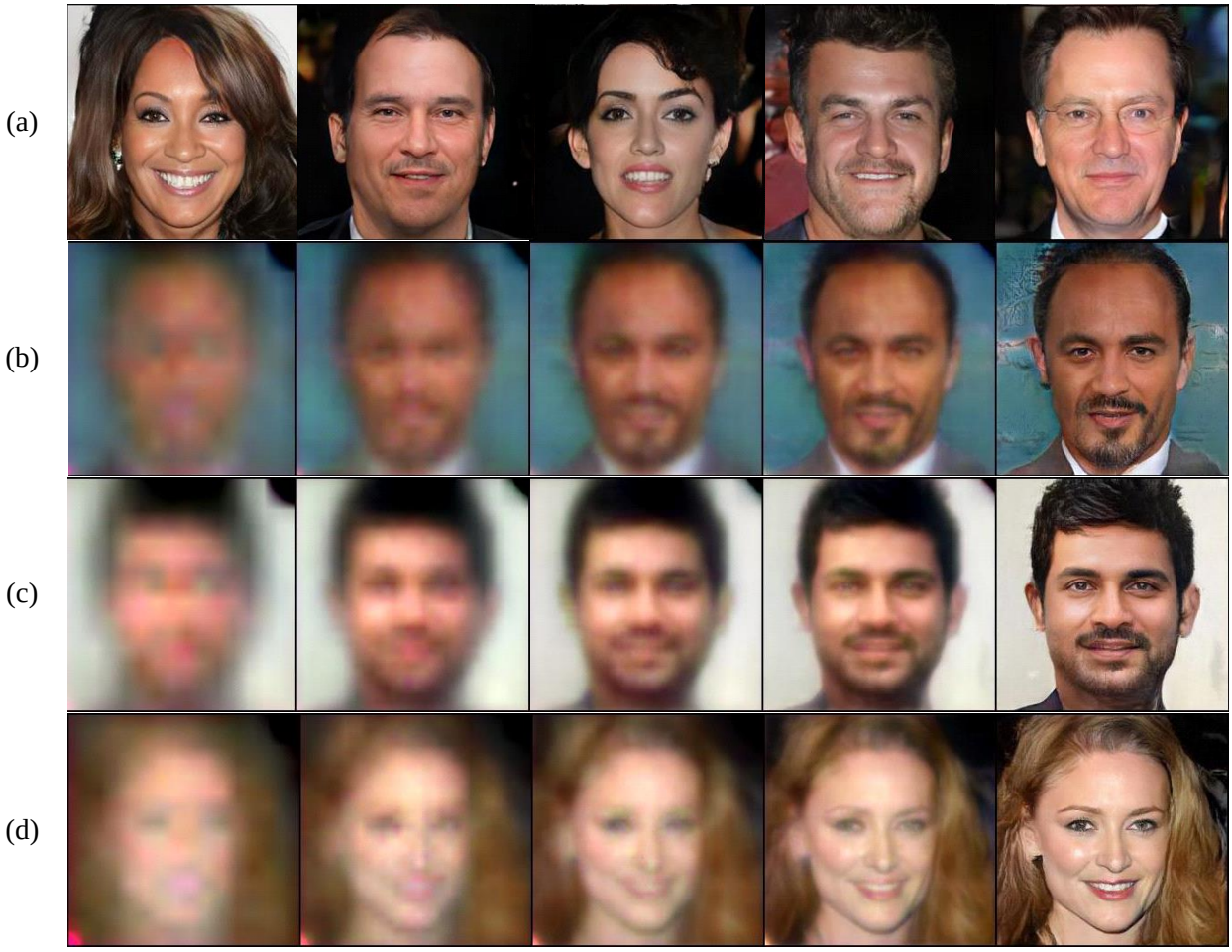


Figure 1.4: (a) 2D face images generated from the voice of real scammers. (b)(c)(d) Snapshots of intermediate images in the process of 2D face generation.

in Fig. 1.4.

1.5 Privacy and Fairness

Privacy. Voices carry a wealth of profile information about the speaker (such age, gender, ethnicity, health status, etc.). Generating faces from voice may further increase the risk of privacy leakage. For this reason, the algorithms that we developed must be used in an ethical way. Facial reconstruction must not be done if the speakers' voice is not completely public or if the speaker it not a public figure and not in agreement with (or has not given consent for) such use of their voice. From a technical perspective, de-identifying voices and

faces [50, 51] is an effective solution to protect the privacy of speakers, and our work can be used for detecting and visualizing the the remaining identity information.

Fairness. It is important for an algorithm to be equitable in its performance for different demographic groups, since biased performance may adversely affect users and even lead to discrimination. In this thesis, where possible, we explicitly take fairness into consideration and perform separate evaluations for different groups of individuals based on the demographic and other factors.

1.6 Thesis Organization

This thesis is organized as follows. In Chapter 2 we introduce a cross modal matching framework for voices and faces and show that it can be used as a nonparametric approach for voice to face generation. In Chapter 3 we present a voice to 2D face image generation model. In Chapter 4 we explore the conditional estimation and apply it to a self-supervised approach for 3D face reconstruction from video. In Chapter 5 we introduce the use of anthropometric measurement, and apply it to 3D facial shape reconstruction from voice. We present our conclusions in Chapter 6.

Chapter 2

Voice and Face Matching

In this chapter, we investigate the associations between voice and face by cross-modal matching. The specific problem we look at is the one wherein we have an existing database of samples of people's voices and images of their faces, and we aim to automatically and accurately determine which voices match to which faces.

2.1 Introduction

A person's face is predictive of their voice. Biologically, the genetic, physical and environmental influences that affect the face also affect the voice.

Humans have been shown to be able to associate voices of unknown individuals to pictures of their faces [19]. Humans also show improved ability to memorize and recall voices when previously exposed to pictures of the speaker's face, but not imposter faces [36, 37, 38]. Cognitively, studies indicate that neuro-cognitive pathways for voices and faces share common structure [34], possibly following parallel pathways within a common recognition framework [1, 35]. The above studies lend credence to the hypothesis that it may be possible to find associations between voices and faces algorithmically as well.

This problem has seen significant research interest, in particular since the recent introduction of the VoxCeleb corpus [3], which comprises collections of video and audio recordings of

a large number of celebrities. The existing approaches [39, 40, 42] have generally attempted to directly relate subjects' voice recordings and their face images, in order to find the correspondences between the two. [39] formulates the mapping as a binary selection task: given a voice recording, one must successfully select the speaker's face from a pair of face images (or the reverse – given a face image, one must correctly select the subject's voice from a pair of voice recordings). They model the mapping as a neural network that is trained through joint presentation of voices and faces to determine if they belong to the same person. In [42, 40], the authors attempt to learn common embeddings (i.e., vector representations) for voices and faces that can be compared to one another to identify associations. The networks that compute the embeddings are also trained through joint presentation of voices and faces, to maximize the similarity of embeddings derived from them if they belong to the same speaker. In all cases, the voice and face are implicitly assumed to directly inform about one another.

In reality, though, it is unclear how much these models capture the direct influence of the voice and face on one another, and how much is explained through implicit capture of higher-level variables such as gender, age, ethnicity etc., which individually predict the two. These higher-level variables, which we will refer to as *covariates*[1] can, in fact, explain much of our ability to match voices to faces (and vice versa) under the previously mentioned "select-from-a-pair" test (where a voice must be used to distinguish the speaker's face from a randomly-chosen imposter). For instance, simply matching the gender of the voice and the face can result in an apparent accuracy of match of up to 75% in a gender-balanced testing setting. Even in a seemingly less constrained "verification" test, where one must only verify if a given voice matches a given face, matching them based on gender alone can result in an equal error rate of 33% [52]. Even matching the voice and the face by age (e.g. matching older-looking faces to older-sounding voices) could result in match accuracy that's significantly better than random.

Previous studies [39, 42] attempt to disambiguate the effect of multiple covariates through

[1]To be clear, these are co*variates*, factors that vary jointly with voice and face, possibly due to some other common causative factors such as genetics, environment, *etc.* They are usually not claimed to be causative factors themselves.

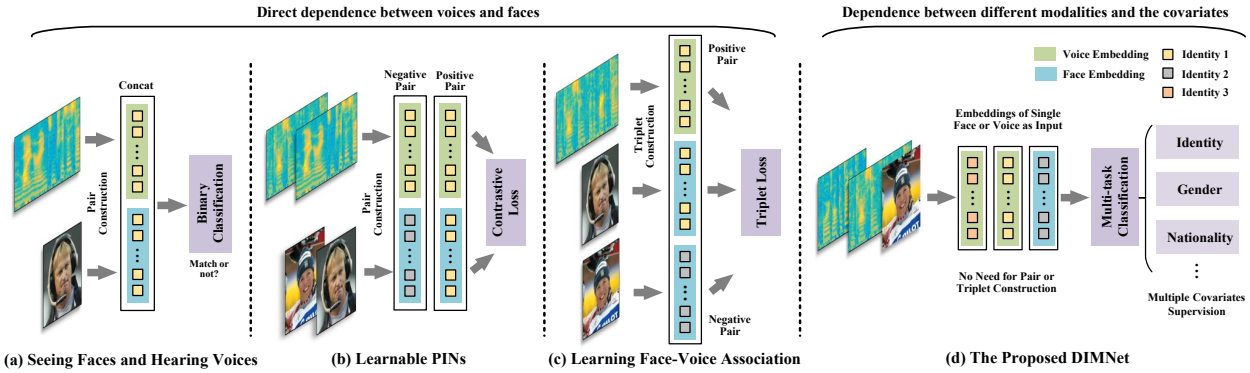


Figure 2.1: Overview of the proposed DIMNet and its comparison to other approaches.

stratified tests that separate the data by covariate value. The results show that at least some of the learned associations are explained by the covariate, indicating that their learning approaches do utilize the covariate information, albeit only implicitly. In this chapter, we propose a novel framework to learn mappings between voices and faces that do not consider any direct dependence between the two, but instead explicitly exploit their individual dependence on the covariates. We define covariate as the identity-sensitive factors that can simultaneously affect voice and face, e.g. nationality, gender, identity (ID), etc. We do not require the *value* these factors take to be the same between the training and test set, since what we are learning is the nature of the covariation with the variable in general, not merely the covariation with the specific values the variable takes in the training set. In contrast to existing methods where supervision is provided through the correspondence of voices and faces, our learning framework, Disjoint Mapping Network (DIMNet), obtains supervision from common covariates, applied *separately* to voices and faces, to learn common embeddings for the two. The comparison between the existing approaches and DIMNets are illustrated in Fig. 2.1 ((a) from [39], (b) from [40], (c) from [42], and (d) from our proposed DIMNets). Compared to other methods, DIMNets present a voice-face embedding framework via multi-task classification and require no pair construction (i.e., both voices and faces can be input sequentially without forming pairs).

DIMNet comprises individual feature learning modules that learn identically-dimensioned features for data from each modality, and a unified input-modality-agnostic classifier that at-

tempts to predict covariates from the learned feature. Data from each modality are presented separately during learning; however the unified classifier forces the feature representations learned from the individual modalities to be comparable. Once trained, the classifier can be removed and the learned feature representations are used to compare data across modalities.

The proposed approach greatly simplifies the learning process and, by considering the modalities individually rather than as coupled pairs, makes much more effective use of the data. Moreover, if multiple covariates are known, they can be simultaneously used for the training through multi-task learning in our framework. As shown in Fig. 2.2, the input training data can be either voice or face, and there is no need for voices and faces to form pairs. Modality switch is to control which embedding network (voice or face) to process the data. While the embeddings are obtained, a multi-task classification network is applied to supervise the learning.

Compared to current methods [39, 40, 42], DIMNets achieve consistently better performance, indicating that direct supervision through covariates is more effective in these settings. We find that of all the covariates, ID provides the strongest supervision. The results obtained from supervision through other covariates also match what may be expected. Our contributions of this chapter are summarized as follows:

- We propose DIMNets, a framework that formulates the problem of cross-modal matching of voices and faces as learning common embeddings for the two through individual supervision from one or more covariates, in contrast to current approaches that attempt to map voices to faces directly. An overview of our framework is given in Fig. 2.2.

- In this framework, we can make full use of multiple kinds of label information (provided by covariates) with a multi-task objective function.

- We achieve the state-of-the-art results on multiple tasks. We are also able to isolate and analyze the effect of the individual covariate on the performance.

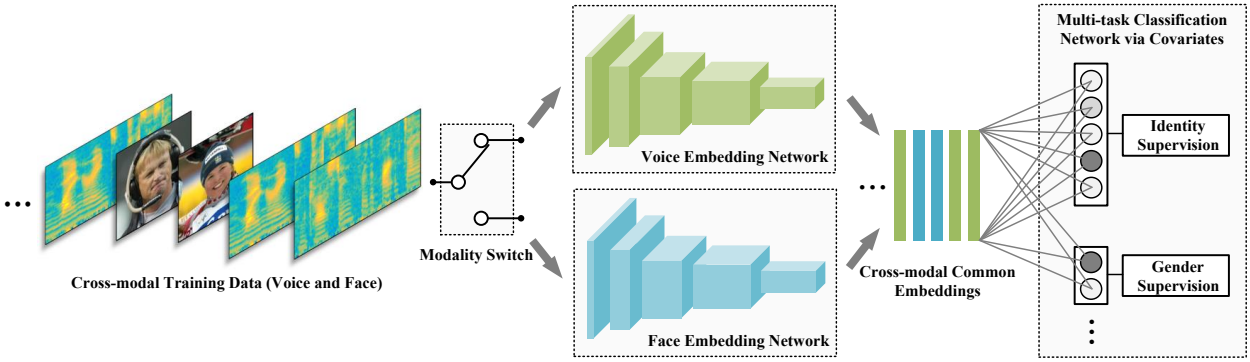


Figure 2.2: Our DIMNet framework.

Moreover, we note that the proposed framework is applicable in any setting where matching of different types of data which have common covariates is required.

2.2 The Proposed Framework

Our goal is to learn common vector representations for both voices and faces, that permit them to be compared to one another. In the following sections we first describe how we learn them from their relationship to common *covariates*. Subsequently, we describe how we will use them for comparison of voices to faces.

2.2.1 Leveraging Covariates to Learn Embeddings

The relationship between voices and faces is largely predicted by *covariates* – factors that individually relate to both the voice and the face. To cite a trivial example, a person's gender relates their voice to their face: male subjects will have male voices and faces, while female subjects will have female voices and faces. More generally, many covariates may be found that relate to both voice and face [53].

Our model attempts to find common representations for both face images and voice recordings by leveraging their relationship to these covariates (rather than to each other). We will do so by attempting to predict covariates from voice and face data in a common embedding space, such that the derived embeddings from the two types of data can be

compared to one another.

Let $\mathcal{V}$ represent a set of voice recordings, and $\mathcal{F}$ represent a set of face images. Let $\mathcal{C}$ be the set of covariates we consider. For the purpose of this chapter, we assume that all covariates are discrete valued (although this is not necessary). Every voice recording in $\mathcal{V}$ and every face in $\mathcal{F}$ can be related to each of the covariates in $\mathcal{C}$. For every covariate $C \in \mathcal{C}$ we represent the value of that covariate for any voice recording v as $C(v)$, and similarly the value of the covariate for any face f as $C(f)$. For example, C could be ID, gender, or nationality. When C is ID, $C(v)$ and $C(f)$ are the ID of voice v and face f, respectively.

Let $F_v(v;\theta_v) : v \mapsto \mathbb{R}^d$ be a *voice embedding* function with parameters θ_v that maps any voice recording v into a d-dimensional vector. Similarly, let $F_f(f;\theta_f)$ be a *face embedding* function that maps any face f into a d-dimensional vector. We aim to learn θ_v and θ_f such that the embeddings of the voice and face for any person are comparable.

For each covariate $C \in \mathcal{C}$ we define a classifier $H_C(x;\phi_C)$ with parameter ϕ_C, which assigns any input $x \in \mathbb{R}^d$ to one of the values taken by C. The classifier $H_C(\cdot)$ is agnostic to which modality its input x was derived from; thus, given an input voice v, it operates on features $F_v(v;\theta_v)$ derived from the voice, whereas given a face f, it operates on $F_f(f;\theta_f)$.

For each v (or f) and each covariate C, we define a loss $L(H_C(F_v(v;\theta_v);\phi_C), C(v))$ between the covariate predicted by $H_C(.)$ and the true value of the covariate for v, $C(v)$. We can now define a *total loss* $\mathcal{L}$ over the set of all voices $\mathcal{V}$ and the set of all faces $\mathcal{F}$, over all covariates as

$$\begin{aligned}\mathcal{L}(\theta_v, \theta_f, \{\phi_C\}) = \sum_{C\in\mathcal{C}} \lambda_C \Bigg(\sum_{v\in\mathcal{V}} L(H_C(F_v(v;\theta_v);\phi_C), C(v)) \\ + \sum_{f\in\mathcal{F}} L(H_C(F_f(f;\theta_f);\phi_C), C(f)) \Bigg)\end{aligned} \tag{2.1}$$

λ_C is the weight for each covariate. In order to learn the parameters of the embedding functions, θ_f and θ_v, we perform the following optimization.

$$\theta_v^*, \theta_f^* = \arg\min_{\theta_v,\theta_f} \min_{\{\phi_C\}} \mathcal{L}(\theta_v, \theta_f, \{\phi_C\}) \tag{2.2}$$

2.2.2 Disjoint Mapping Networks

In DIMNet, we instantiate $F_v(v;\theta_v)$, $F_f(f;\theta_f)$ and $H_C(x;\phi_C)$ as neural networks. Fig. 2.2 shows the network architecture we use to train our embeddings. It comprises three components. The first, labelled *Voice Network* in the figure, represents $F_v(v;\theta_v)$ and is a neural network that extracts d-dimensional embeddings of the voice recordings. The second, labelled *Face Network* in the figure, represents $F_f(f;\theta_f)$ and is a network that extracts d-dimensional embeddings of face recordings. The third component, labelled *Classification Networks* in the figure, is a bank of one or more classification networks, one per covariate considered. Each of the classification networks operates on the d-dimensional features output by the embedding networks to classify one covariate, e.g. gender.

The training data comprise voice recordings and face images. Voice recordings are sent to the voice-embedding network, while face images are sent to the face-embedding network. This switching operation is illustrated by the switch at the input in Fig.2.2. In either case, the output of the embedding network is sent to the covariate classifiers.

As can be seen, at any time the system either operates on a voice, or on a face, i.e. the operations on voices and faces are disjoint. During the learning phase too, the updates of the two networks are disjoint – loss gradients computed when the input is voice only update the voice network, while loss gradients derived from face inputs update the face network, while both contribute to updates of the classification networks.

In our implementation, specifically, $F_v(\cdot)$ is a convolutional neural network that operates on Mel-Spectrographic representations of the speech signal. The output of the final layer is pooled over time to obtain a final d-dimensional representation. $F_f(\cdot)$ is also a convolutional network with a pooled output at the final layer that produces a d-dimensional representation of input images. The classifiers $H_C(\cdot)$ are all simple multi-class logistic-regression classifiers comprising a single softmax layer.

Finally, in keeping with the standard paradigms for training neural network systems, we use the cross-entropy loss to optimize the networks. Also, instead of the optimization in Eq.

2.2, the actual optimization performed is the one below. The difference is inconsequential.

$$\theta_v^*, \theta_f^*, \{\phi_C^*\} = \underset{\theta_v, \theta_f, \{\phi_C\}}{\arg\min}\, \mathcal{L}(\theta_v, \theta_f, \{\phi_C\}) \tag{2.3}$$

2.2.3 Training the DIMNet

All parameters of the network are trained through backpropagation, using stochastic gradient descent. During training, we construct the minibatches with a mixture of speech segments and face images, as the network learns more robust cross-modal features with mixed inputs. Taking voice as an example, we compute the voice embeddings using $F_v(v; \theta_v)$, and obtain the losses using classifiers $H_C(\cdot)$ for all the covariates. We back-propagate the loss gradient to update the voice network as well as the covariate classifiers. The same procedure is also applied to face data: the backpropagated loss gradients are used to update the face network and the covariate classifiers. Thus, the embedding functions are learned using the data from their modalities individually, while the classifiers are learned using data from all modalities.

2.2.4 Using the Embeddings

Once trained, the embedding networks $F_v(v; \theta_v)$ and $F_f(f; \theta_f)$ can be used to extract embeddings from any voice recording or face image. Given a voice recording v and a face image f, we can now compute a similarity between the two through the cosine similarity $S(v, f) = \frac{F_v^\top F_f}{\|F_v\|_2 \|F_f\|_2}$. We can employ this similarity to evaluate the match of any face image to any voice recording. This enables us, for instance, to attempt to rank a collection of faces $f_1, \cdots, f_K$ in order of estimated match to a given voice recording v, according to $S(v, f_i)$, or conversely, to rank a collection of voices $v_1, \cdots, v_K$ according to their match to a face f, on order of decreasing $S(v_i, f)$.

2.3 Experiments

We ran experiments on matching voices to faces, to evaluate the embeddings derived by DIMNets. The details of the experiments are given below.

Datasets. Our experiments were conducted on the Voxceleb [3] and VGGFace [54] datasets. The Voxceleb dataset consists of 153,516 audio segments from 1,251 speakers. Each audio segment is taken from an online video clip with an average duration of 8.2 seconds. For the face dataset, we used a manually filtered version of VGGFace. After face detection, there remain 759,643 images from 2,554 subjects.

We use the intersection of the two datasets, i.e. subjects who figure in both corpora, for our final corpus, which thus includes 1,225 IDs with 667 males and 558 females from 36 nationalities. We use ID, gender and nationality as our covariates, all of which are provided by the datasets. The data are split into train/validation/test sets, following the settings in [39]. Details are shown in Table 2.1.

Table 2.1: Statistics for the data appearing in VoxCeleb and VGGFace.

# of samples	train	validation	test	total
speech segments	112,697	14,160	21,799	148,656
face images	313,593	36,716	58,420	408,729
IDs	924	112	189	1,225
genders	2	2	2	2
nationalities	32	11	18	36
testing instances	-	4,678,897	6,780,750	11,459,647

The visual data used in Section 2.3.4 is densely extracted from the video in VoxCeleb1 dataset at 25/6 fps. It contains 100,000 segmented speaking face-tracks obtained by SyncNet [55], leading to 1,218,575 frames (images). For fair comparison, we follow the train/val/test split strategy from [40] in our experiments. The evaluations are performed based on the provided lists [40], which specify the testing pairs of voices and faces.

Preprocessing. Separated preprocessing pipelines are employed to data from different modalities, *i.e.* audio segments and face images. For audio segments, we use a voice activity detector interface from the WebRTC project to isolate speech-bearing regions of

the recordings. Subsequently, 64-dimensional log Mel-spectrograms are generated, using an analysis window of 25ms, with hop of 10ms between frames. We perform mean and variance normalization of each mel-frequency bin.

For training, we randomly crop out regions of varying lengths of 300 to 800 frames (so the size of the input spectrogram ranges from 300×64 to 800×64 for each mini-batch, around 3 to 8 seconds). For the face data, facial landmarks in all images are detected using MTCNN [56]. The cropped RGB face images of size $128 \times 128 \times 3$ are obtained by similarity transformation. Each pixel in the RGB images is normalized by subtracting 127.5 and then dividing by 127.5. We perform data augmentation by horizontally flipping the images with 50% probability in minibatches (effectively doubling the number of face images).

Training. The detailed network configurations are given in Table 2.2. For the voice network, we use 1D convolutional layers, where the convolution is performed along the axis that corresponds to time. The face network employs 2D convolutional layers. For both, the convolutional layers are followed by batch normalization (BN) [57] and rectified linear unit activations (ReLU) [58]. The numbers within the parentheses represent the size and number of filters, while the subscripts represent the stride and padding. So, for example, $(3, 64)_{/2,1}$ denotes a 1D convolutional layer with 64 filters of size 3, where the stride and padding are 2 and 1 respectively, while $(3 \times 3, 64)_{/2,1}$ represents a 2-D convolutional layer of 64 3×3 filters, with stride 2 and padding 1 in both directions. Note that 924, 2, and 32 are the number of unique values taken by the ID, gender, and nationality covariates, respectively. The final face embedding is obtained by averaging the feature maps from the final layer, i.e. through *average pooling*. The final voice embedding is obtained by averaging the feature maps at the final convolutional layer along the time axis alone. Note that the classification networks are single-layer softmax units with as many outputs as the number of unique values the class can take (2 for gender, 32 for nationalities, and 924 for IDs in our case).

We follow the typical settings of SGD for optimization. Minibatch size is 256. The momentum and weight decay values are 0.9 and 0.001 respectively. To learn the networks from scratch, the learning rate is initialized at 0.1 and divided by 10 after 16K iterations

Table 2.2: CNN architectures: details.

	layer	voice	face
embedding network	Conv	$(3, 256)_{/2,1}$ $\begin{bmatrix}(3, 256)_{/1,1}\\(3, 256)_{/1,1}\end{bmatrix}$	$(3 \times 3, 64)_{/2,1}$ $\begin{bmatrix}(3 \times 3, 64)_{/1,1}\\(3 \times 3, 64)_{/1,1}\end{bmatrix}$
		$(3, 384)_{/2,1}$ $\begin{bmatrix}(3, 384)_{/1,1}\\(3, 384)_{/1,1}\end{bmatrix}$	$(3 \times 3, 128)_{/2,1}$ $\begin{bmatrix}(3 \times 3, 128)_{/1,1}\\(3 \times 3, 128)_{/1,1}\end{bmatrix}$
		$(3, 576)_{/2,1}$ $\begin{bmatrix}(3, 576)_{/1,1}\\(3, 576)_{/1,1}\end{bmatrix}$	$(3 \times 3, 256)_{/2,1}$ $\begin{bmatrix}(3 \times 3, 256)_{/1,1}\\(3 \times 3, 256)_{/1,1}\end{bmatrix}$
		$(3, 864)_{/2,1}$ $\begin{bmatrix}(3, 864)_{/1,1}\\(3, 864)_{/1,1}\end{bmatrix}$	$(3 \times 3, 512)_{/2,1}$ $\begin{bmatrix}(3 \times 3, 512)_{/1,1}\\(3 \times 3, 512)_{/1,1}\end{bmatrix}$
		$(3, 64)_{/2,1}$	$(3 \times 3, 64)_{/2,1}$
	AvgPool	$t \times 1$	$h \times w \times 1$
classification network	FC	64×924, 64×2, 64×32	

and again after 24K iterations. The training is completed at 28K iterations.

Testing. We use the following protocols for evaluation:

- *1:2 Matching.* Here, we are given a probe input from one modality (voice or face), and a gallery of two inputs from the other modality (face or voice), including one that belongs to the same subject as the probe, and another of an "imposter" that does not match the probe. The task is to identify which entry in the gallery matches the probe. We report performance in terms of matching accuracy – namely what fraction of the time we correctly identify the right instance in the gallery.

  To minimize the influence of random selection, we construct as many testing instances as possible through exhaustive enumeration all positive matched pairs (of voice and face). To each pair, we include a randomly drawn imposter in the gallery. We thus have a total of 4,678,897 trials in the validation set, and 6,780,750 trials in the test set.

- *1:N Matching.* This is the same as the 1:2 matching, except that the gallery now

includes $N-1$ imposters. Thus, we must now identify which of the N entries in the gallery matches the probe. Here too results are reported in terms of matching accuracy. We use the same validation and test sets as the 1:2 case, by augmenting each trial with $N-2$ additional imposters. So the number of trials in validation and test sets is the same as earlier.

- *Verification.* We are given two inputs, one a face, and another a voice. The task is to determine if they are matched, i.e. both belong to the same subject. In this problem setting the similarity between the two is compared to a threshold to decide a match. The threshold can be adjusted to trade off *false rejections* (F_R), i.e. wrongly rejecting true matches, with *false alarms* (F_A), i.e. wrongly accepting mismatches. We report results in terms of *equal error rate*, i.e. when $F_R = F_A$. We construct our validation and test sets from those used for the 1:2 matching tests, by separating each trial into two, one comprising a matched pair, and the other a mismatched pair. Thus, our validation and test sets are exactly twice as large as those for the 1:2 test.

- *Retrieval.* The gallery comprises a large number of instances, one or *more* of which might match the probe. The task is to order the gallery such that the entries in the gallery that match the probe lie at the top of the ordering. Here, we report performance in terms of *Mean Average Precision* (MAP) [59]. Here we use the *entire* collection of 58,420 test faces as the gallery for each of our 21,799 test voices, when retrieving faces from voices. For the reverse (retrieving voices from faces), the numbers are reversed.

Each result is obtained by averaging the performances of 5 models, which are individually trained.

Covariates in Training and Testing. We use the three covariates provided in the dataset, namely identity (I), gender (G), and nationality (N) for our experiments. The treatment of covariates differs for training and test.

- *Training.* For training, supervision may be provided by any set of (one two or three)

Table 2.3: Acc. (%) of covariate prediction.

method	gender classification		nationality classification	
	voice	face	voice	face
DIMNet-I	-	-	-	-
DIMNet-G	97.48	99.22	-	-
DIMNet-N	-	-	**74.86**	60.13
DIMNet-IG	**97.70**	**99.42**	-	-
DIMNet-IN	-	-	74.17	60.27
DIMNet-GN	97.59	99.06	74.62	**60.50**
DIMNet-IGN	97.69	99.15	74.37	59.88

covariates. We consider all combinations of covariates, I, G, N, (I,G), (I,N), (G,N) and (I,G,N). Increasing the number of covariates effectively increases the supervision provided to training. All chosen covariates were assigned a weight of 1.0.

- *Testing.* As pointed out in [52], simply recognizing a covariate such as gender can result in seemingly significant matching performance. For instance, just recognizing the subjects' gender from their voice and images can result in a 33% EER for verification, and 25% error in matching for the $1:2$ tests. In order to isolate the effect of covariates on performance hence we also *stratify* our test data by them. Thus we construct 4 testing groups based on the covariates, including the unstratified (U) group, stratified by gender (G), stratified by nationality (N), and stratified by gender and nationality (G, N). In each group the test set itself is separated into multiple strata, such that for all instances within any stratum the covariate values are the same.

2.3.1 Cross-modal Matching

In this section, we report results on the 1:2 and 1:N matching tests. In order to ensure that the embedding networks do indeed leverage on accurate modelling of covariates, we first evaluate the classification accuracy of the classification networks for the covariates themselves. Table 2.3 shows the results.

The rows of the table show the covariates used to supervise the learning. Thus, for instance, the row labelled "DIMNet-I" shows results obtained when the networks have been

trained using ID alone as covariate, the row labelled "DIMNet-G" shows results when supervision is provided by gender, "DIMNet-IG" has been trained using ID and gender, etc.

The columns of the table show the specific covariate being evaluated. Since the identities of subjects in the training and test set do not overlap, we are unable to evaluate the accuracy of ID classification. Note that we can only test the accuracy of the classification network for a covariate if it has been used in the training. Thus, classification accuracy for gender can be evaluated for DIMNet-G, DIMNet-GN and DIMNet-IGN, while that for nationality can be evaluated for DIMNet-N, DIMNet-GN and DIMNet-IGN.

The results in Table 2.3 show that gender is learned very well, and in all cases gender recognition accuracy is quite high. Nationality, on the other hand, is not a well-learned classifier, presumably because the distribution of nationalities in the data set is highly skewed [39], with nearly 65% of all subjects belonging to the USA. It is to be expected therefore that nationality as a covariate will not provide sufficient supervision to learn good embeddings.

1:2 matching. Table 2.4 shows the results for the 1:2 matching tests. In the table, the row labelled "SVHF-Net" gives results obtained with the model of [39].

The columns are segregated into two groups, one labelled "voice → face" and the other labelled "face → voice". In the former, the probe is a voice recording, while the gallery comprises faces. In the later the modalities are reversed. Within each group the columns represent the stratification of the test set. "U" represents test sets that are not stratified, and include the various covariates in the same proportion that they occur in the overall test set. The columns labelled "G" and "N" have been stratified by gender and nationality, respectively, while the column "G, N" represents data that have been stratified by both gender and nationality. In the stratified tests, we have ensured that all data within a test instance have the same value for the chosen covariate. Thus, for instance, in a test instance for voice → face in the "G" column, the voice and both faces belong to the same gender. This does not reduce the overall number of test instances, since it only requires ensuring that the gender of the imposter matches that of the probe instance.

We make several observations. First, DIMNet-I performs better than SVHF-Net, im-

Table 2.4: Performance comparison of 1:2 matching for models trained using different sets of covariates.

method	voice → face (ACC %)				face → voice (ACC %)			
	U	G	N	G, N	U	G	N	G, N
SVHF-Net	81.0	63.9	-	-	79.5	63.4	-	-
DIMNet-I	83.5± 0.4	70.9±0.6	82.0±0.5	69.9±0.8	83.5±0.5	71.8±0.6	82.4±0.5	**70.9**±0.8
DIMNet-G	72.9±0.6	50.3±0.7	71.9±0.5	50.2±0.7	72.5±0.5	50.5±0.7	72.2±0.5	50.6±0.7
DIMNet-N	57.5±0.5	55.3±0.7	53.0±0.4	52.0±0.6	56.2±0.4	54.3±0.6	53.9±0.4	52.0±0.6
DIMNet-IG	**84.1**±0.4	**71.3**±0.6	**82.7**±0.6	**70.4**±0.8	**84.0**±0.4	**71.7**±0.6	**83.0**±0.5	70.8±0.5
DIMNet-IN	83.0±0.4	70.0±0.7	81.0±0.6	68.6±0.8	82.9±0.4	70.9±0.6	81.9±0.5	70.2±0.8
DIMNet-GN	75.9±0.4	56.7±0.6	72.9±0.5	53.5±0.7	73.8±0.7	54.9±0.5	72.6±0.5	53.5±0.9
DIMNet-IGN	83.7±0.5	70.8±0.3	81.8±0.5	69.2±0.7	83.6±0.7	71.4±0.5	82.5±0.4	70.5±0.6

proving the accuracies by 2.45%-4.02% for the U group, and 7.01%-8.38% for the G group. It shows that mapping voices and faces to their common covariates is an effective strategy to learn representations for cross-modal matching.

Second, DIMNet-I produces significantly better embeddings that DIMNet-G and DIMNet-N, highlighting the rather unsurprising fact that ID provides the more useful information than the other two covariates. In particular, DIMNet-G respectively achieves 72.90% and 72.47% for voice to face and face to voice matching using only gender as a covariate. This verifies our hypothesis that we can achieve almost 75% matching accuracy by only using the gender. These numbers also agree with the performance expected from the numbers in Table 2.3. As expected, nationality as a covariate does not provide as good supervision as gender. DIMNet-IG is marginally better than DIMNet-I, indicating that gender supervision provides additional support over ID alone.

Third, we note that while DIMNet-I is able to achieve good performance on the dataset stratified by gender, DIMNet-G only achieves random performance. The performance achieved by DIMNet-G on the U dataset is hence completely explained by gender matching. Once again, the numbers match to those in [52].

1:N Matching. We also experiment for $N > 2$. Unlike SVHF-Net [39] that needs to train different models for different N in this setting, we use the same model for different N. The results in Fig. 2.3 shows accuracy as a function of N for various models.

All the results in Fig. 2.3 are consistent with Table 2.4. As expected, the performance

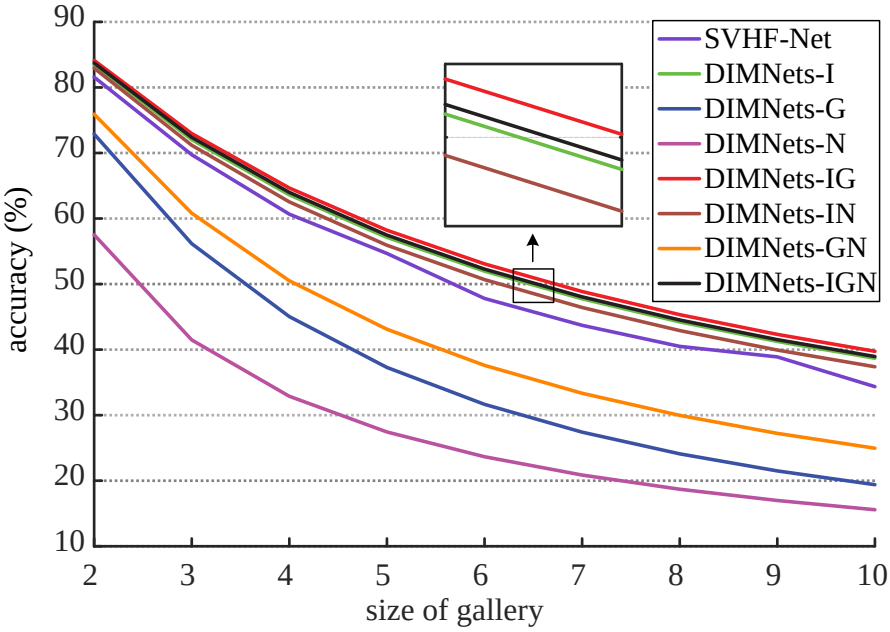


Figure 2.3: Performance of 1:N matching

Table 2.5: Verification results.

method	verification (EER %)			
	U	G	N	G, N
DIMNet-I	25.0±0.2	35.0±0.5	25.9±0.7	35.7±0.9
DIMNet-G	34.9±0.1	49.7±0.2	35.1±0.4	49.7±0.5
DIMNet-N	45.9±0.4	47.0±0.6	47.9±0.8	48.9±1.1
DIMNet-IG	**24.6**±0.2	**34.8**±0.4	**25.5**±0.7	**35.7**±0.8
DIMNet-IN	25.5±0.2	36.2±0.4	27.3±0.7	37.4±0.8
DIMNet-GN	33.3±0.5	46.7±0.2	34.8±0.3	48.1±0.5
DIMNet-IGN	25.0±0.2	35.8±0.4	26.8±0.7	37.3±0.7

of all methods degrades with increasing N. In general, DIMNets that use ID as supervision outperform SVHF-Net by a considerable margin, showing that DIMNets are able to make best use of the ID information. We obtain the best results when both ID and gender are used as supervision covariates. However, the results obtained using only gender information as covariate is much worse, which is also consistent with the analysis in [52].

Table 2.6: Retrieval performance (mAP).

method	voice → face (mAP %)			face → voice (mAP %)		
	ID	gender	nationality	ID	gender	nationality
Random	0.6	52.6	40.7	0.6	52.6	40.7
DIMNet-I	4.3±0.1	89.6±0.4	43.3±0.2	4.2±0.1	88.5±0.4	43.7±0.2
DIMNet-G	1.1±0.1	**97.8**±0.6	41.6±0.2	1.2±0.1	**97.2**±0.6	42.0±0.2
DIMNet-N	1.2±0.1	57.0±0.3	**45.7**±0.7	1.0±0.1	56.9±0.3	**49.3**±0.6
DIMNet-IG	**4.4**±0.1	93.1±0.5	43.2±0.1	**4.2**±0.1	92.2±0.4	43.9±0.2
DIMNet-IN	3.9±0.1	89.7±0.4	44.0±0.7	4.0±0.14	88.4±0.39	45.9±0.66
DIMNet-GN	1.89±0.1	95.9±0.4	45.2±0.6	1.6±0.1	94.00±0.4	48.4±0.5
DIMNet-IGN	4.1±0.1	92.3±0.6	44.1±0.6	4.1±0.1	91.3±0.6	45.8±0.6

2.3.2 Cross-modal Verification

For verification, we need to determine whether an audio segment and a face image are from the same ID or not. We report the equal error rate (EER) for verification in Table 2.5.

In general, DIMNets that use ID as a covariate achieve an EER of about 25%, which is considerably lower than the 33% expected if the verification were based on gender matching alone. The results in Table 2.5 show that using both gender and ID information as covariates can further improve the performance over using ID alone, well validating the superiority of our multi-task learning framework.

Using proper combination of covariates is crucial to the performance. ID is arguably the most effective covariate supervision. More interestingly, nationality is seen to be an ineffective covariate, while gender alone as a covariate produces results that well matches our expectation.

2.3.3 Cross-modal Retrieval

We also perform retrieval experiments using voice or face as query. Table 2.6 lists the mean average precision (mAP) of the retrieval for various models.

The columns in the table represent the covariate being retrieved. Thus, for example, in the "ID" column, the objective is to retrieve gallery items with the same ID as the query, whereas in the "gender" column the objective is to retrieve the same gender.

Table 2.7: AUCs (%) of DIMNets under different testing groups.

	Seen-Heard					Unseen-Unheard				
	U	G	N	A	G, N, A	U	G	N	A	G, N, A
[40]	87.0	74.2	85.9	86.6	74.0	78.5	61.1	77.2	74.9	58.8
DIMNet-I	**95.1**±0.2	**90.8**±0.3	**93.4**±0.2	**95.2**±0.1	**88.9**±0.2	82.5±0.1	71.0±0.3	81.1±0.1	77.7±0.1	**62.8**±0.4
DIMNet-IG	94.7±0.2	89.8±0.2	93.2±0.1	94.8±0.1	87.8±0.2	**83.2**±0.1	**71.2**±0.4	**81.9**±0.2	**78.0**±0.1	**62.8**±0.4

We note that ID-based DIMNets produce the best features for retrieval, with the best performance obtained with DIMNet-IG. Also, as may be expected, the covariates used in training result in the best retrieval of that covariate. Thus, DIMNet-G achieves an mAP of nearly 98% on gender, though on retrieval of ID it is very poor. As in other experiments, nationality remains a poor covariate in general. Compared to gender (2 classes) and nationality (unbalanced 28 classes), retrieving ID is a challenging problem given the large amount of identities (182 classes). The significant and consistent improvements over chance-level results show that the DIMNet models do learn some useful associations between voices and faces.

2.3.4 Comparisons to the current state-of-the-art

We compare DIMNet with the state of the art [40]. The results are reported in Table 2.7. Note that it is fair comparison because the DIMNet models in this section are trained with and evaluated on the same released datasets in [40]. There are two evaluation protocols, including Seen-Heard and Unseen-Unheard scenarios. The identities of the training and testing set have overlaps in Seen-Heard scenario (closed-set), while they are fully disjoint in Unseen-Unheard scenario (open-set). For each scenario, there are 5 testing groups based on the covariates, including the unstratified group (U), group, stratified by gender (G), stratified by nationality (N), stratified by age (A), and stratified by (G, N, A). We compute the area under the curve (AUC) for different testing groups.

It is clear that DIMNets produce better embeddings than [40] for pair-wise verification on both seen-heard and unseen-unheard scenarios. Specifically, DIMNets achieve 8%-15% absolute and 3%-10% absolute improvements on seen-heard and unseen-unheard test set,

respectively. Compared to DIMNet-IG, DIMNet-I performs better on the seen-heard test set while DIMNet-IG is better on the unseen-unheard test set. It implies that introducing useful covariates improves the generalization capability of DIMNet.

2.4 Discussion

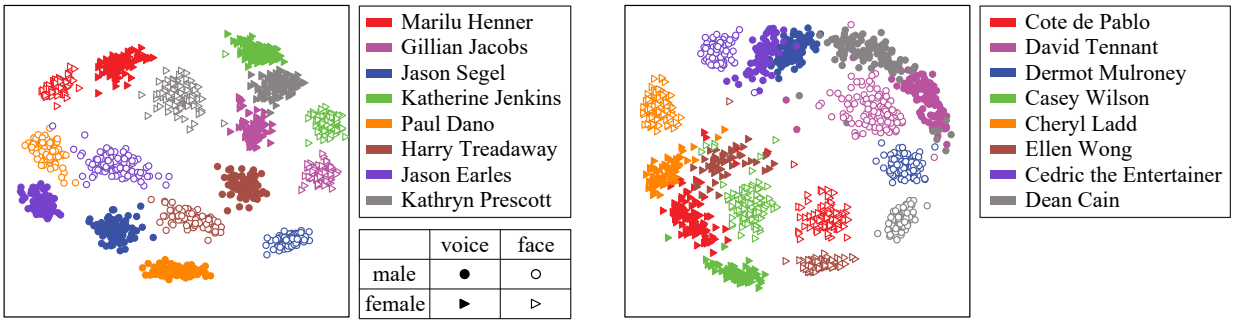
We have proposed that it is possible to learn common embeddings for multi-modal inputs, particularly voices and faces, by mapping them individually to common covariates. In particular, the proposed DIMNet architecture is able to extract embeddings for both modalities that achieves consistently better performance than the methods that directly map faces to voices.

The approach also provides us the ability to tease out the influence of each of the covariates of voice and face data, in determining their relation. The results show that the strongest covariate, not unexpectedly, is ID. The results also indicate that prior results by other researchers who have attempted to directly match voices to faces may perhaps not be learning any direct relation between the two, but implicitly learning about the common covariates, such as ID, gender, etc.

Our experiments also show that although we have achieved possibly the best reported performance on this task, thus far, the performance is not anywhere close to prime-time. In the $1 : N$ matching task, performance degrades rapidly with increasing N, indicating a rather poor degree of true match.

To better understand the problem, we have visualized the learned embeddings from DIMNet-I in Fig. 2.4 to provide more insights. The left panel shows subjects from the training set, while the right panel is from the test set. The visualization method we used is multi-dimensional scaling (MDS) [2], rather than the currently more popular t-SNE [60]. This is because MDS tends to preserve distances and global structure, while t-SNE attempts to retain statistical properties and highlights clusters, but does not preserve distances.

From Fig. 2.4, we immediately notice that the voice and face data for a subject are only



(a) Training Set (b) Testing Set

Figure 2.4: Visualization of voice and face embeddings using multi-dimensional scaling [2].

weakly proximate. While voice and face embeddings for a speaker are generally relatively close to each other, they are often closer to other subjects. Interestingly, the genders separate (even though gender has not been used as a covariate for this particular network), showing that at least *some* of the natural structure of the data is learned. Fig. 2.4 shows embeddings obtained from both training and test data. We can observe similar behaviors in both, showing that the the general characteristics observed are not just the outcome of overfitting to training data. The visualization in Fig. 2.4 also shows that there is still significant room for improvement. For example, it may be possible to force compactness of the distributions of voice and face embeddings through modified loss functions such as the center loss [61] or angular softmax loss [62], or through an appropriately designed loss function that is specific to this task.

Chapter 3

2D Face Reconstruction from Voice

In this chapter, we focus on the problem of reconstructing face images from voices. Specifically, we generate a face image from any given voice recording. The face image is expected to have as many identity associations as possible with the true face of the speaker.

3.1 Introduction

A person's voice is incontrovertibly statistically related to their facial structure. The relationship is, in fact, multi-faceted. *Direct* relationships include the effect of the underlying skeletal and articulator structure of the face and the tissue covering them, all of which govern the shapes, sizes, and acoustic properties of the vocal tract that produces the voice [27, 28]. Less directly, the same genetic, physical and environmental influences that affect the development of the face also affect the voice. Demographic factors, such as gender, age and ethnicity too influence both voice and face (and can in fact be independently inferred from the voice [63, 64] or the face [65]), providing additional links between the two.

Neurocognitive studies have shown that human perception implicitly recognizes the association of faces to voices [1]. Studies indicate that neuro-cognitive pathways for voices share a common structure with that for faces [34] – the two may follow parallel pathways within a common recognition framework [1, 35]. In empirical studies, humans have shown the ability

to associate voices of unknown individuals to pictures of their faces [19]. They are seen to show improved ability to memorize and recall voices when previously shown pictures of the speaker's face, but not imposter faces [36, 37, 38].

On the other hand, reconstructing the face from voice is a challenging, maybe even impossible task for several reasons [66]. First, it is an ill-posed cross-modal problem: Although many face-related factors affect the voice, it may not be possible to entirely disambiguate them from the voice. Even if this were not the case, it is unknown *a priori* exactly what features of the voice encode information about any given facial feature (although one may take guesses [66]). Moreover, the signatures of the different facial characteristics may lie in different spoken sounds; thus, in order to obtain sufficient evidence, the voice recordings must be long enough to have sufficient coverage of sounds to derive all the necessary information. The information containing in a single audio clip may not be sufficient for constructing a face image.

In particular, we aim at addressing this task in an open-set scenario, where reconstruction is performed on unheard and unseen identities, which are never presented during training. Thus the voice-feature to face-feature associations must be learned in a manner that generalizes beyond the set of examples provided in training.

Yet, although *prima facie* the problem seems extremely hard, recent advances in neural network based generative models have shown that they are able to perform similarly challenging generative tasks in a variety of scenarios, when properly structured and trained [67, 68]. In particular, generative adversarial networks (GANs) [45] have demonstrated the ability to learn to generate highly sophisticated imagery, given only signals about the validity of the generated image, rather than detailed supervision of the content of the image itself [46, 69]. We use this ability to learn to generate faces from voices.

For our solution, we propose a simple but effective data-driven framework based on generative adversarial networks (GANs), as illustrated in Fig. 3.1. It includes 4 major components: voice embedding network, generator, discriminator, and classifier. The objective of the network is simple: given a voice recording it must generate a face image that plausibly

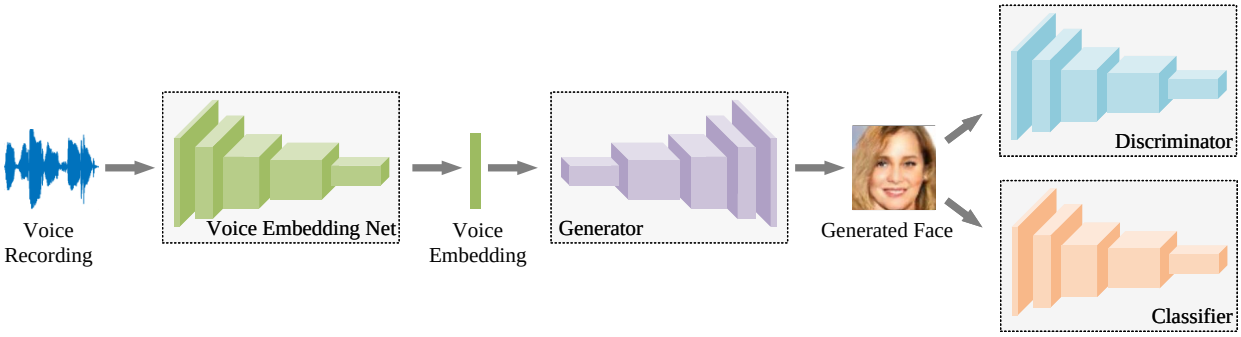


Figure 3.1: The proposed GANs-based framework for generating faces from voices.

belongs to that voice. The voice recording itself is input to the generator network in the form of a voice embedding vector extracted by a voice embedding network. The generator is trained using a pair of discriminators. The first evaluates if the images it generates are realistic face images. The second discriminator (classifier) verifies that the identity of face image output by the generator does indeed match the actual identity of the speaker.

We present both qualitative and quantitative evaluations of the results produced by our model. The qualitative results show that our framework is able to map the voice manifold to face manifold. We can observe many identity associations between the generated faces and the input voices. The generated faces are generally age and gender appropriate, frequently matching the real face of the speaker. Additionally, given non-speech input the outputs become unrealistic, showing that the learned mapping is at least somewhat specific, in that the face manifold it learns are derived primarily from the voice manifold and not elsewhere. In addition, for different speech segments from the same person, the generated faces exhibit reasonable intra-class variation.

We also propose a number of quantitative evaluation metrics to evaluate the output of our network, based on how specific the model is in mapping voices to faces, how well the high-level attributes of the generated face match that of the speaker, and how well the generated faces match the identity (ID) of the speaker itself. For the last metric (ID matching), we leverage the cross-modal matching task [39], wherein, specifically, we need to match a speech segment to one of the two faces, where one is the true face of the speaker, and another is

an "imposter." Our tests reveal that the network is highly specific in generating faces in response to voices, produces quantifiably gender-appropriate faces from voices, and that the matching accuracy is much better than chance, or what may be obtained merely by matching gender. We refer the reader to the experiments section for actual numbers.

Overall, our contributions are summarized as follows:

- We introduce a new task of generating faces from voice in voice profiling. It could be used to explore the relationship between voice and face modalities.
- We propose a simple but effective framework based on generative adversarial networks for this task. Each component in the framework is well motivated.
- We propose to quantitatively evaluate the generated faces by using a cross-modal matching task. Both the qualitative and quantitative results show that our framework is able to generate faces that have identity association with the input voice.

3.2 The Proposed Framework

Before we begin, we first specify some of the notation we will use. We represent voice recordings by the symbol v, using super or subscripts to identify specific recordings. Similarly, we represent face images by the symbol f. We will represent the *identity* of a subject who provides voice or face data as y. We will represent the true identity of (the subject of) voice recording v as y^v and face f as y^f. We represent the function that maps a voice or face recording to its idenitity as $ID()$, i.e. $y^v = ID(v)$ and $y^f = ID(f)$. Additional notation will become apparent as we introduce it.

Our objective is to train a model $F(v;\Theta)$ (with parameter Θ) that takes as input a voice recording v and produces, as output a face image $\hat{f} = F(v;\Theta)$ that belongs to the speaker of v, *i.e.* such that $ID(\hat{f}) = ID(v)$.

We use the framework shown in Figure 2.2 for our model, which decomposes $F(v;\Theta)$ into a sequence of two components, $F_e(v;\theta_e)$ and $F_g(e;\theta_g)$. $F_e(v;\theta_e) : v \rightarrow e$ is a voice embedding

function with parameter θ_e that takes in a voice recording v and outputs an embedding vector e that captures all the salient information in v. $F_g(e;\theta_g) : e \rightarrow f$ is a *generator* function that takes in an embedding vector and generates a face image $\hat{f}$.

We must learn θ_e and θ_g such that $ID(F_g(F_v(v;\theta_e);\theta_g) = ID(v)$.

3.2.1 Training the network

Data. We assume the availability of face and voice data from a set of subjects $\mathcal{Y} = \{y_1, y_2, \cdots, y_k\}$. Correspondingly, we also have a set of voice recordings $\mathcal{V} = \{v_1, v_2, ..., v_N\}$, with identity labels $\mathcal{Y}^v = \{y_1^v, y_2^v, ..., y_N^v\}$ and a set of faces $\mathcal{F} = \{f_1, f_2, ..., f_M\}$ with identity labels $\mathcal{Y}^f = \{y_1^f, y_2^f, ..., y_M^f\}$, such that $y^v \in \mathcal{Y}\ \forall y^v \in \mathcal{Y}^v$ and $y^f \in \mathcal{Y}\ \forall y^f \in \mathcal{Y}^f$. N may not be equal to M.

In addition, we define two sets of labels $\mathcal{R} = \{r_1, r_2, ..., r_M \mid \forall i,\ r_i = 1\}$ and $\hat{\mathcal{R}} = \{\hat{r}_1, \hat{r}_2, ..., \hat{r}_N \mid \forall i,\ \hat{r}_i = 0\}$ corresponding $\mathcal{Y}^f$ and $\mathcal{Y}^v$ respectively. $\mathcal{R}$ is a set of labels that indicates that all faces in $\mathcal{F}$ are "real." $\hat{\mathcal{R}}$ is a set of labels that indicates that any faces generated from any $v \in \mathcal{V}$ are synthetic or "fake."

GAN framework. In training the model, we impose two supervision signals. First, the output $\hat{f}$ of the generator in response to any actual voice input v must be a realistic face image. Second, it must belong to the same identity as the voice, *i.e.* $ID(\hat{f}) = f^v$. As explained in Section 1, we will use a GAN framework to train $F_e(.;\theta_e)$ and $F_g(.;\theta_g)$. This will require the definition of *adversary* that provide losses that can be used to learn the model parameters.

We define an adversarial objective. First, the *discriminator* F_d determines if any input image (f or $\hat{f}$) is a genuine picture of a face, or one generated by the generator, *i.e.* assigns any face image (f or $\hat{f}$) to its real/fake label (r or $\hat{r}$). The loss function for F_d is defined as $L_d(F_d(f), r)$ (or $L_d(F_d(\hat{f}), \hat{r})$). Second, *classifier* F_c learns to assign any real face image f to its identity label y^f. Accordingly, the loss function for F_c is $L_c(F_c(f), y^f)$.

Last, the generator F_g takes in a voice recording v and attempts to generate any face

image $\hat{f}$ that can be classified to real label r and identity label y^v by F_d and F_g, respectively. The corresponding loss function for F_g is $L_d(F_d(F_g(v), r) + L_c(F_c(F_g(v)), y^v)$.

In our implementation, we instantiate $F_e(v;\theta_e)$, $F_g(e;\theta_g)$, $F_d(f;\theta_d)$ and $F_c(f;\theta_c)$ as convolutional neural networks, shown in Fig. 2.2. F_e is the component labeled as *Voice Embedding Network*. v is the Mel-Spectrographic representations of speech signal. The output of the final convolutional layer is pooled over time, leading to a q-dimensional vector e. F_g is labeled as *Generator*. f and $\hat{f}$ are RGB images with the same resolution of $w \times h$. F_d and F_c are labeled as *Discriminator* and *Classifier*, respectively. The loss functions L_d and L_c of these two components are the cross-entropy loss.

Training the network. The training data comprise a set of voice recordings $\mathcal{V}$ and a set of face images $\mathcal{F}$. From the voice recordings in $\mathcal{V}$ we could obtain the corresponding generated face images $\hat{\mathcal{F}} = \{\hat{f} = F_g(F_e(v)) \mid \forall v \in \mathcal{V}\}$.

The framework is trained in an adversarial manner. To simplify, we use a pretrained voice embedding network $F_e(v;\theta_e)$ from a speaker recognition task, and freeze the parameter θ_e when training our framework. F_c is trained to maximize $\sum_{i=1}^{M} L_c(F_c(f_i), y_i)$ with fixed θ_e, θ_g and θ_d. Similarly, F_d is trained to maximize $\sum_{i=1}^{M} L_d(F_d(f_i), r_i) + \sum_{i=1}^{N} L_d(F_d(\hat{f}_i), \hat{r}_i)$ with fixed θ_e, θ_g and θ_c. The F_g is trained to maximize $\sum_{i=1}^{N} L_d(F_d(\hat{f}_i), r_i) + L_c(F_c(\hat{f}_i), y_i^v)$ with θ_e, θ_d and θ_c fixed, where $\hat{f}_i = F_g(F_e(v_i))$. The training pipeline is summarized in Algorithm 1.

Once trained, $F_d(f;\theta_d)$ and $F_c(f;\theta_c)$ can be removed. Only $F_d(f;\theta_d)$ and $F_c(f;\theta_c)$ are used for face generation from voice during the inference phase. It is worth noting that the targeted scenario is *open-set*. In our evaluations, the model is required to work on previously unseen and unheard identities, *i.e.* $y^v \in \mathcal{Y}$ in the training phase, while $y^v \notin \mathcal{Y}$ in the testing phase.

Algorithm 1 The training algorithm of the proposed framework

Input: A set of voice recordings with identity label $(\mathcal{V}, \mathcal{Y}^v)$. A set of labeled face images with identity label $(\mathcal{F}, \mathcal{Y}^f)$. A voice embedding network $F_e(v;\theta_e)$ trained on $\mathcal{V}$ with speaker recognition task. θ_e is fixed during the training. Randomly initialized θ_g, θ_d, and θ_c

Output: The parameters θ_g.

1: **while** not converged **do**
2: Randomly sample a minibatch of n voice recordings $\{v_1, v_2, ..., v_n\}$ from $\mathcal{V}$
3: Randomly sample a minibatch of m face images $\{f_1, f_2, ..., f_m\}$ from $\mathcal{F}$
4: Update the discriminator $F_d(f;\theta_d)$ by ascending the gradient
$$\nabla_{\theta_d}\big(\sum_{i=1}^{n}\log(1-F_d(\hat{f}_i)) + \sum_{i=1}^{m}\log F_d(f_i)\big)$$
5: Update the classifier $F_c(f;\theta_c)$ by ascending the gradient ($a[i]$ indicates the i-th element of vector a)
$$\nabla_{\theta_c}\big(\sum_{i=1}^{m}\log F_c(f_i)[y_i^f]\big)$$
6: Update the generator $F_g(f;\theta_c)$ by ascending the gradient
$$\nabla_{\theta_g}\big(\sum_{i=1}^{n}\log F_c(F_g(F_e(v_i)))[y_i^v] + \sum_{i=1}^{m}\log F_d(F_g(F_e(v_i)))\big)$$
7: **end while**

3.3 Experiments

In our experiments, the voice recordings are from the `Voxceleb` [39] dataset and the face images are from the manually filtered version of `VGGFace` [54] dataset. Both datasets have identity labels. We use the intersection of the two datasets with the common identities, leading to 149,354 voice recordings and 139,572 frontal face images of 1,225 subjects. We follow the train/validation/test split in [39]. The details are shown in Table 3.1.

Separated data pre-processing pipelines are employed to audio segments and face images. For audio segments, we use a voice activity detector interface from the WebRTC project to isolate speech-bearing regions of the recordings. Subsequently, we extract 64-dimensional log Mel-spectrograms using an analysis window of 25ms, with a hop of 10ms between frames. We perform mean and variance normalization of each mel-frequency bin. We randomly crop an audio clips around 3 to 8 seconds for training, but use the entire recording for testing. For the face data, facial landmarks in all images are detected using [70]. The cropped RGB face images of size $3 \times 64 \times 64$ are obtained by similarity transformation. Each pixel in the RGB images is normalized by subtracting 127.5 and then dividing by 127.5.

Training. The network architecture is given in Table 3.2. The parameters in the con-

Table 3.1: Statistics of the datasets used in our experiments

	train	validation	test	total
# of speech segments	113,322	14,182	21,850	149,354
# of face images	106,584	12,533	20,455	139572
# of subjects	924	112	189	1,225

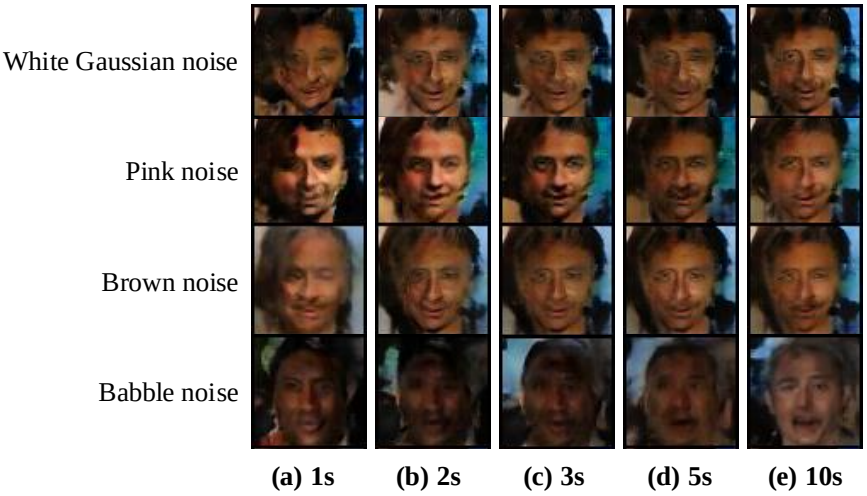


Figure 3.2: The generated face images from noise input. (a)-(e) are 1, 2, 3, 5, and 10 seconds, respectively.

volutional layers of discriminator and classifier are shared in our experiments. For the voice embedding network, we use 1D convolutional layers. Conv $3_{/2,1}$ denotes 1D convoluitonal layer with kernel size of 3, where the stride and padding are 2 and 1, respectively. Each convolutional layer is followed by a Batch Normalization (BN) [57] layer and Rectified Linear Units (ReLU) [58]. The output shape is shown accordingly, where $t_{i+1} = \lceil (t_i - 1)/2 \rceil + 1$. The final outputs are pooled over time, yielding a 64-dimensional embedding. We use 2D deconvolutional layers with ReLU for the generator and 2D convolutional layers with Leaky ReLU (LReLU) for the discriminator and classifier. The final output is given by fully connected (FC) layer. We basically follow the hyperparameter setting in [71]. We used the Adam optimizer [72] with learning rate of 0.0002. β_1 and β_2 are 0.5 and 0.999, respectively. We use a minibatch size of 128 samples. The training is completed at 100K iterations.

Table 3.2: CNN architectures: details.

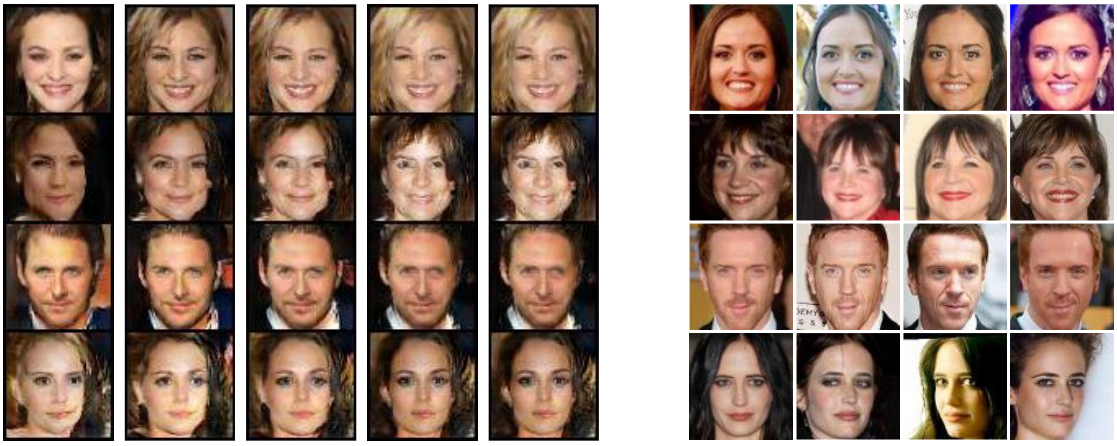
Voice Embedding Network			Generator		
Layer	Act.	Output shape	Layer	Act.	Output shape
Input	-	$64 \times t_0$	Input	-	$64 \times 1 \times 1$
Conv $3_{/2,1}$	BN + ReLU	$256 \times t_1$	Deconv $4 \times 4_{/1,0}$	ReLU	$1024 \times 4 \times 4$
Conv $3_{/2,1}$	BN + ReLU	$384 \times t_2$	Deconv $3 \times 3_{/2,1}$	ReLU	$512 \times 8 \times 8$
Conv $3_{/2,1}$	BN + ReLU	$576 \times t_3$	Deconv $3 \times 3_{/2,1}$	ReLU	$256 \times 16 \times 16$
Conv $3_{/2,1}$	BN + ReLU	$864 \times t_4$	Deconv $3 \times 3_{/2,1}$	ReLU	$128 \times 32 \times 32$
Conv $3_{/2,1}$	BN + ReLU	$64 \times t_5$	Deconv $3 \times 3_{/2,1}$	ReLU	$64 \times 64 \times 64$
AvePool $1 \times t_5$	-	64×1	Deconv $1 \times 1_{/1,0}$	-	$3 \times 64 \times 64$
Discriminator			Classifier		
Layer	Act.	Output shape	Layer	Act.	Output shape
Input	-	$3 \times 64 \times 64$	Input	-	$3 \times 64 \times 64$
Conv $1 \times 1_{/1,0}$	LReLU	$32 \times 64 \times 64$	Conv $1 \times 1_{/1,0}$	LReLU	$32 \times 64 \times 64$
Conv $3 \times 3_{/2,1}$	LReLU	$64 \times 32 \times 32$	Conv $3 \times 3_{/2,1}$	LReLU	$64 \times 32 \times 32$
Conv $3 \times 3_{/2,1}$	LReLU	$128 \times 16 \times 16$	Conv $3 \times 3_{/2,1}$	LReLU	$128 \times 16 \times 16$
Conv $3 \times 3_{/2,1}$	LReLU	$256 \times 8 \times 8$	Conv $3 \times 3_{/2,1}$	LReLU	$256 \times 8 \times 8$
Conv $3 \times 3_{/2,1}$	LReLU	$512 \times 4 \times 4$	Conv $3 \times 3_{/2,1}$	LReLU	$512 \times 4 \times 4$
Conv $4 \times 4_{/1,0}$	LReLU	$64 \times 1 \times 1$	Conv $4 \times 4_{/1,0}$	LReLU	$64 \times 1 \times 1$
FC 64×1	Sigmoid	1	FC $64 \times k$	Softmax	k

3.3.1 Qualitative Results

As a first experiment, we compared the outputs of the network in response to various noise signals to outputs obtained from actual speech recordings. Figure 3.2 shows outputs generated for four different types of noise. Each row shows the generated faces using one of the four noise audio segments with different durations.

We evaluated noise segments of different durations (1, 2, 3, 5, and 10 seconds) to observe how the generated faces change with the duration. The generated images are seen to be blurry, unrecognizable and generally alike, since there is no identity information in noise. With longer noise recordings, the results do not improve. Similar results are obtained over a variety of noises.

On the other hand, when we use regular speech recordings with the aforementioned durations as inputs, outputs tend to be realistic faces, as seen in Figure 3.3. The results indicate that while the generator does learn to produce face-like images, actual faces are produced chiefly in response to actual voice. We infer that while the generator has learned



(a) 1s (b) 2s (c) 3s (d) 5s (e) 10s (f) Reference images

Figure 3.3: (a)-(e) The generated face images from regular speech recordings with different durations. (f) the corresponding reference face images.

to map the speech manifold to the manifold of faces, it maps other inputs outside of this manifold.

Figure 3.3 also enables us to subjectively evaluate the actual output of the network. These four speakers (from top to bottom) are Danica McKellar, Cindy Williams, Damian Lewis, and Eva Green. These results are typical and not cherry-picked for presentation (several of our reconstructions match the actual speaker closely, but we have chosen not to selectively present those to avoid misrepresenting the actual performance of the system).

The generated images are on the left, while the reference images (the actual faces of the speakers) are on the right. To reduce the perceptual bias and better illustration, we show multiple reference face images for each speaker. Although the generated and the reference face indicate different persons, the identity information of these two are matched in some sense (like gender, ethnicity, etc.). With longer speech segments, the generated faces gradually converge to faces associatable with the speaker. Figure 3.4 shows additional examples demonstrating that the synthesized images are generally age- and gender-matched with the speaker. For each group, images on the left are the generated images and images on the right are the references.

In the next experiment, we select 7 different speech recordings of each speaker and generate the faces from the entire recordings. The results are shown in Figure 3.5 (reference

(a) (b) (c)

Figure 3.4: The generated faces from (a) old voices, (b) male voices, (c) female voices.

(a) Generated face images **(b) Reference images**

Figure 3.5: Faces generated from different segments of speech from the same speaker, and the reference faces. Each row shows the results for the same speaker.

images are also provided for comparison). Once again, the results are typical and not cherrypicked. We believe that the images in each row exhibit reasonable variations of the same person (except the fourth image in the first row), indicating that our model is able to build a mapping between the speech group to face group, thus retaining the identity of the face across speech segments from the same speaker.

3.3.2 Quantitative Results

We attempt to quantitatively distinguish between the faces generated in response to noise from those generated from voice using the discriminator $F_d()$ itself. Note that $F_d()$ is biased, and has explicitly been trained to tag synthesized faces as fake. The mean and standard

deviation (obtained from 1000 samples) of its output value in response to actual voices are 0.09 and 0.13 respectively, while for (an identical number of) noise inputs they are 0.06 and 0.07 respectively. Even a biased discriminator is clearly able to distinguish between faces generated from voices, and those obtained from noise.

As a first test, we attempted to ID (classify) faces derived from test recordings of the speakers in the training set. On a set of 4620 recordings, from 924 identities, we achieved a top-1 accuracy of 61.7% and a top-5 accuracy of 82.3%, showing that the voice-based face reconstructions *do* actually faithfully capture the identities of *known* subjects.

For unknown subjects, we ran a gender classifier on 21,850 generated faces (from the speech segments in test set). The classifier was trained on the face images in the training set using the network architecture of the discriminator, with an accuracy of 98.97% on real face images. The classifier obtained a 96.45% accuracy in matching the gender of the generated faces to the known gender of the speaker, showing that the generated faces are almost always of the correct gender.

Finally, we evaluated our model by leveraging the task of voice to face matching. Here, we are given a voice recording, a face image of the true speaker, and a face image of an imposter. We must match the voice to the face of the true speaker. Ideally, the probe voice could be replaced by the generated face image if they carry the same identity information. So the voice to face matching problem reduces to a typical face verification or face recognition problem. The resulting matching accuracy could be used to quantitatively evaluate the association between the speech segment and the generated face.

We construct the testing instances (a probe voice recording, a true face image, and an "imposter" face) using data in the testing set, leading to 2,353,560 trials. We also compute the matching accuracy on about 50k trials constructed from a small part of the training set to see how well the model fit to the training data. We also perform stratified experiments based on gender where we select the imposter face with the same gender as the true face. In this case, gender information cannot be used for matching anymore, leading to a more fair test.

The results are shown in Table 3.3. Our results are given by replacing the probe voice embeddings by the embeddings of the generated face. The high accuracies obtained on the training set for the unstratified and gender stratified tests (96.83% and 93.98% respectively) show that generated faces do carry correct identity information for the training set. The results on the test set on unstratified and gender stratified tests (76.07% and 59.69% respectively) are better than those in DIMNets-G [73], indicating that our model learns more associations than gender. The large drop compared to the results on the training set shows however that considerable room remains to improve generalizability in the model.

Table 3.3: The voice to face matching accuracies.

	unstratified group (ACC. %) (training set / testing set)	stratified group by gender (ACC. %) (training set / testing set)
DIMNets-I [73]	- / 83.45	- / 70.91
DIMNets-G [73]	- / 72.90	- / 50.32
ours	96.83 / 76.07	93.98 / 59.69

3.4 Discussion

The proposed GAN-based framework is seen to achieve reasonable reconstruction results. The generated faces have identity associations with the true speaker. There remains considerable room for improvement. Firstly, there are obvious issues with the GAN-based output: The produced faces have features such as hair that are presumably not predicted by voice, but simply obtained from their co-occurrence with other features. The model may be more appropriately learned through data cleaning that removes obviously unrelated aspects of the facial image, such as hair and background. The proposed model is vanilla in many ways. For instance [74, 75] describe several explicit correspondences between speech and face features, e.g. different phonetic units are known to relate to different facial features. We are investigating models that explicitly consider these issues.

Chapter 4

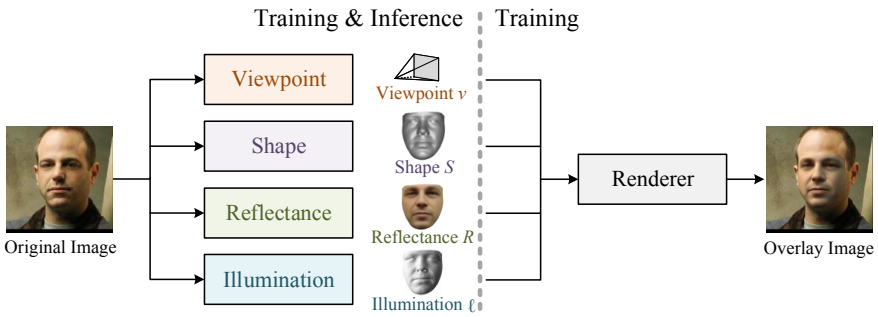
3D Face Reconstruction from Video

In this chapter, we propose a self-supervised approach for reconstructing 3D faces from video. This method is used for collecting a 3D audiovisual dataset from video data, where we can extract paired 3D faces and voice recordings for subsequent research.

4.1 Introduction

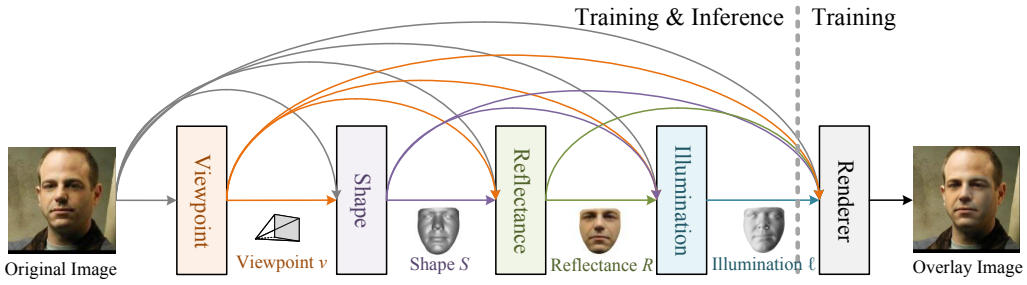
Reconstructing 3D faces from single-view 2D images has been a longstanding problem in computer vision. The common approach represents the 3D face as a combination of its *shape*, as represented by the 3D coordinates of a number of points on its surface called vertices, and its *texture*, as represented by the reflectances of red, green and blue at these vertices [76]. The problem then becomes learning a regression model between the 2D images, and vertices and their reflectances.

The regression itself may be learned using training data where both, the 2D images and the corresponding 3D parameters are available. However, these data are scarce, and even the ones that are available generally only have shape information [77, 78, 79]; the ones that do have other parameters are usually captured in a controlled environment [80] or are synthetic [81], which is not representative of real-world images. Consequently, there is great interest in self-supervised learning methods, which learn the regression model from natural in-the-wild



$$P(S,R,v,\ell|I) = P(S|I)\,P(R|I)\,P(v|I)\,P(\ell|I)$$

(a) Most existing methods assume conditional independence.



$$P(S,R,v,\ell|I) = P(v|I)\,P(S|I,v)\,P(R|I,v,S)\,P(\ell|I,v,S,R)$$

(b) The proposed CEST makes no assumption on dependency.

Figure 4.1: Conventional 3D face reconstruction and our CEST framework. The dotted lines separate the modules used for inference from those used for training.

2D images or videos, without explicit access to 3D training data [5, 82].

The problem is complicated by the fact that the actual image formation depends not only on the shape and texture of the face, but also on the illumination (the intensity and direction of the incident light), and other factors such as the viewpoint (incorporating the orientation of the face and the position of the camera), etc. Thus, the learned regression model must also account for these factors. To this end, the general approach is one where shape, reflectance, illumination and viewpoint parameters are all extracted from the 2D image. The regression model that extracts these facial parameters are learned through *self-supervision*: the extracted facial parameters are recombined to render the original 2D image, and the model parameters are learned to minimize the reconstruction error.

The solution, however, remains ambiguous because a 2D image may be obtained from different combinations of shape, texture, illumination and viewpoint. To ensure that the self-

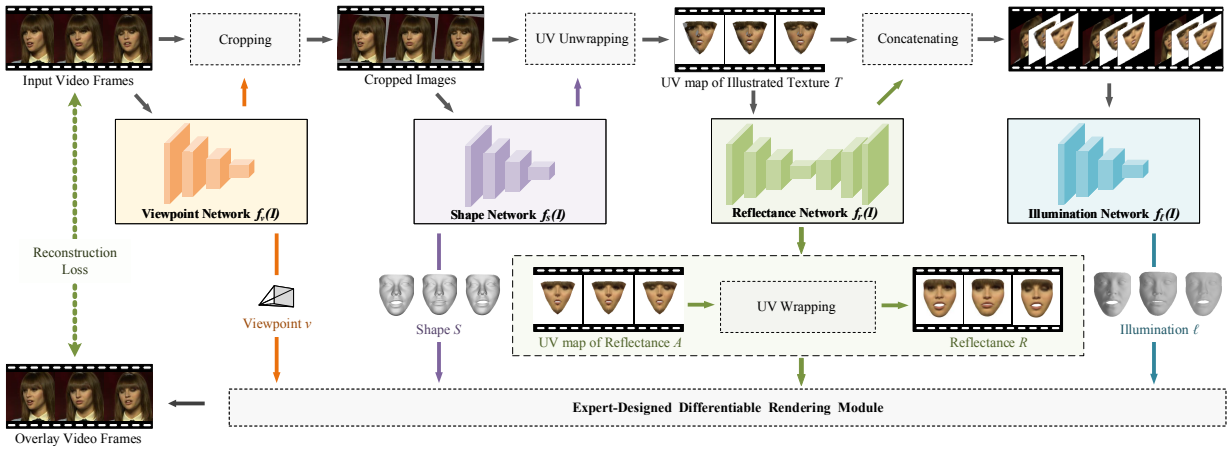


Figure 4.2: The overall training pipeline of the proposed CEST framework.

supervision provides meaningful disentanglement, the manner in which the facial parameters are recombined to reconstruct the 2D image are based on the actual physics of image formation [5, 82, 81]. To further reduce potential ambiguities, regularizations are necessary. Reflectance *symmetry* has been proposed as a regularizer [83, 4, 84], wherein the reflectance of a face image and its mirror reflection are assumed to be identical. Smoothness has also been employed to regularize the shape and reflectance [82, 4]. Additional regularization may be obtained by considering correspondences between multiple images of the same face [85, 7], particularly when they are obtained under near identical conditions such as the sequence of images from a video. The approach in [7] has considered reflectance *consistency*, where reflectances of all image frames in a video clip are assumed to be similar.

In all of these prior works, the target parameters, namely the shape, reflectance, illumination and viewpoint parameters are all *individually* estimated, without considering their direct influences on one another, although they are jointly optimized. In effect, at inference time they assume that the estimate of, *e.g.* the reflectance, is conditionally independent of the estimated shape or viewpoint, given the original 2D image. The coupling among the four is only considered during (self-supervised) training, where they must all combine to faithfully recreate the input 2D image [86, 87, 6, 83, 7]. This is illustrated in Fig. 4.1(a).

In reality, 2D images are reduced-dimensional projections, and thus imperfect representations of the full three-dimensional structure of the face, and the aspects of reflectance and

illumination imprinted in them are not independent of the underlying shape of the object or the viewpoint they were captured from. Therefore, the captured 2D image represents a joint interaction among viewpoint, shape, reflectance and illumination. Consequently, the statistical estimates of any of these four factors may not, in fact, be truly conditionally independent of one another given only the 2D image (although, given the entire 3D model they might have been). Thus, modelling all of these variables as being conditionally independent effectively represents a lost opportunity since, by predicting them individually, the constraints they impose on one another are ignored. Optimization-based approaches [88, 85, 89] do attempt to capture the dependence by iteratively estimating shape and reflectance from one another. However, these methods require correspondence information of the image sequence in a video and suffer from costly inference.

In this chapter, we propose a novel learning-based framework based on conditional estimation (CEST). CEST explicitly considers the statistical dependency of the various 3D facial parameters (shape, viewpoint, reflectance and illumination) upon one another, when derived from single 2D image. The specific form of the dependencies adopted in this chapter is shown in Fig. 4.1(b). We note that the CEST framework is very general and allows us to consider any other dependency structures. Our paper serves as one of the many potential choices that work well in practice. To this end, we present a specific, and intuitive, solution in CEST, where the viewpoint, facial shape, facial reflectance, and illumination are predicted *sequentially* and *conditionally*. In this context, the prediction of facial shape is conditioned on the input image and the derived viewpoint; the prediction of facial reflectance is conditioned on the input image, derived viewpoint and facial shape; and so forth.

As before, learning remains self-supervised, through comparison of re-rendered 2D images obtained with the estimated 3D face parameters to the original images. As additional regularizers, we also employ reflectance symmetry constraints [83, 4, 84], and reflectance consistency constraints (across frames in a short video clip) [7]. These are included in the form of cross-frame reconstruction error terms, the number of which increases quadratically with the number of video frames considered together for self-supervision. To address the

dramatically increased number of reconstruction terms, we propose a stochastic optimization strategy to improve training efficiency.

We present ablation studies and comparisons to state-of-the-art methods [5, 83, 7] to evaluate CEST. We show that CEST produces better reflectance and structured illumination, leading to more realistic rendered faces with fine facial details, compared to all other tested methods. It also achieves better shape estimation accuracy on AFLW2000-3D [90] and MICC [91] datasets than current state-of-the-art self-supervised and *fully supervised* approaches. Overall, our contributions can be summarized as follows:

- We propose CEST, a conditional estimation framework for 3D face reconstruction that explicitly considers the statistical dependencies among 3D face parameters.
- We propose a specific design for the decomposition of conditional estimation, where the viewpoint, shape, reflectance, and the illumination are derived sequentially.
- We propose a stochastic optimization strategy to efficiently incorporate reflectance symmetry and consistency constraints into CEST. As the number of video frames increase, the computational complexity of CEST is increased linearly, rather than quadratically.

4.2 Related Work

Monocular 3D face reconstruction by self-supervised learning. Many research studies published recently aim to learn 3D facial parameters from a single image in a self-supervised manner. In [6], the authors propose a coarse-to-fine framework to improve the details in reconstructed 3D faces. Ayush *et al.* [5] present a model-based deep convolutional face autoencoder (MoFA) to fit a 3DMM to shape, reflectance, and illuminance. InverseFaceNet [92] trains a direct regression model on a synthetic training corpus that is generated by self-supervised bootstrapping. SfSNet [81] combines labeled synthetic and unlabeled real-world images in learning, and produces accurate depth map, and reflectance and shade

disentanglement. To better characterize facial details, 3DMM is generalized to a nonlinear model in [82, 83]. [93] uses mesh convolutions for 3D faces, leading to a light-weight model with competitive performance. [11] incorporates the multi-view consistency from geometry, pixel, and depth as constraints.

However, these approaches generally do not consider correspondences across frames in a video. FML [7] is the first self-supervised framework that incorporates video clues in training. The shape and reflectance for each video frame are approximated by averaging the shapes and reflectances in a video clip. However, models trained on the averaged representations may not work well for a single image if the number of multi-frame images is large, due to the large gap between averaged and isolated images. On the contrary, CEST uses representations from single images. More importantly, it uses conditional estimation for predicting the facial parameters, and does not assume conditional independence between them, an often unrealistic assumption employed in the previously mentioned approaches.

Optimization-based 3D face reconstruction. [85] proposes to fit a template model to photo-collections by updating the viewpoint, geometry, lighting, and texture iteratively. [89] fits a face model to detected 3D landmarks, and refines the texture and geometry details. [86] learns facial subspaces for identity and expression variations with a parametric shape prior. [94] considers 3D face reconstruction as a global variational energy minimization problem, and estimates dense low-rank 3D shapes for video frames.

While these approaches can be considered conditional estimation, they focus on deriving 3D facial parameters from video, and are not relevant to the problem of deriving them from single-frame images, the problem addressed in our work. For CEST, video clips are viewed as consistent collections of images used to better learn the model.

4.3 The Proposed Framework

In this work, we adopt a common practice from 3D Morphable Model (3DMM) [76], which represents a 3D face as a combination of shape and reflectance. The shape comprises a

collection of vertices $\boldsymbol{S} = [\boldsymbol{S}(1); \boldsymbol{S}(2); ...; \boldsymbol{S}(K)] \in \mathbb{R}^{K\times 3}$, where K is the number of vertices and $\boldsymbol{S}(i) = [\boldsymbol{S}(i,1), \boldsymbol{S}(i,2), \boldsymbol{S}(i,3)]$ denotes the xyz coordinates in the Cartesian coordinate system. The typology for $\boldsymbol{S}$ is consistent for different faces. The reflectance comprises a collection of pixel values $\boldsymbol{R} = [\boldsymbol{R}(1); \boldsymbol{R}(2); ...; \boldsymbol{R}(K)] \in \mathbb{R}^{K\times 3}$. Each row $\boldsymbol{R}(i) = [\boldsymbol{R}(i,1), \boldsymbol{R}(i,2), \boldsymbol{R}(i,3)]$ comprises the pixel values (*ie*, RGB) at position $\boldsymbol{S}(i)$.

4.3.1 Framework Overview

The problem of 3D face reconstruction from a 2D image is that of obtaining estimates of the shape $\boldsymbol{S}$, reflectance $\boldsymbol{R}$, viewpoint $\boldsymbol{v}$ and illumination $\boldsymbol{\ell}$, given an input image $\boldsymbol{I}$. Statistically, we aim to estimate the most likely values for these variables, given the input image:

$$\hat{\boldsymbol{S}}, \hat{\boldsymbol{R}}, \hat{\boldsymbol{v}}, \hat{\boldsymbol{\ell}} = \arg\max_{\boldsymbol{S},\boldsymbol{R},\boldsymbol{v},\boldsymbol{\ell}} P(\boldsymbol{S}, \boldsymbol{R}, \boldsymbol{v}, \boldsymbol{\ell}|\boldsymbol{I}) \tag{4.1}$$

The challenges of the aforementioned estimation are twofold: first $P(\boldsymbol{S}, \boldsymbol{R}, \boldsymbol{v}, \boldsymbol{\ell}|\boldsymbol{I})$ must be modelled, and second, $\arg\max_{\boldsymbol{S},\boldsymbol{R},\boldsymbol{v},\boldsymbol{\ell}} P(\boldsymbol{S}, \boldsymbol{R}, \boldsymbol{v}, \boldsymbol{\ell}|\boldsymbol{I})$ must be computed.

Modelling $P(\boldsymbol{S}, \boldsymbol{R}, \boldsymbol{v}, \boldsymbol{\ell}|\boldsymbol{I})$ directly is a challenging problem, and the problem must be factored down. Prior approaches [82, 5, 93] have decomposed this problem by assuming that shape, reflectance, viewpoint and illumination are all conditionally independent, given the image. We formulate this decomposition as $P(\boldsymbol{S}, \boldsymbol{R}, \boldsymbol{v}, \boldsymbol{\ell}|\boldsymbol{I}) = P(\boldsymbol{S}|\boldsymbol{I})P(\boldsymbol{R}|\boldsymbol{I})P(\boldsymbol{v}|\boldsymbol{I})P(\boldsymbol{\ell}|\boldsymbol{I})$. This leads to simplified estimates where each of the variables can be independently estimated, i.e. $\hat{\boldsymbol{S}} = \arg\max_{\boldsymbol{S}} P(\boldsymbol{S}|\boldsymbol{I})$, $\hat{\boldsymbol{R}} = \arg\max_{\boldsymbol{R}} P(\boldsymbol{R}|\boldsymbol{I})$, etc. As we have discussed earlier, the conditional independence assumption is questionable, since the conditioning variable, $\boldsymbol{I}$, is a lower-dimensional projection of the 3D face that entangles the four variables.

In CEST we explicitly model the conditional dependence, as shown in Fig. 4.1(b). Specifically we decompose the joint probability as

$$\begin{aligned} & P(\boldsymbol{S}, \boldsymbol{R}, \boldsymbol{v}, \boldsymbol{\ell}|\boldsymbol{I}) \\ = \; & P(\boldsymbol{v}|\boldsymbol{I})P(\boldsymbol{S}|\boldsymbol{I}, \boldsymbol{v})P(\boldsymbol{R}|\boldsymbol{I}, \boldsymbol{v}, \boldsymbol{S})P(\boldsymbol{\ell}|\boldsymbol{I}, \boldsymbol{v}, \boldsymbol{S}, \boldsymbol{R}) \end{aligned} \tag{4.2}$$

Coupling the variables in this manner results in a complication: even factored as above, maximizing the joint probability with respect to $\boldsymbol{S}$, $\boldsymbol{R}$, $\boldsymbol{v}$, and $\boldsymbol{\ell}$ must be jointly performed,

since the variables are coupled. We approximate it instead with the following sequential estimate, based on the sequential decomposition above:

$$\begin{aligned} \hat{\boldsymbol{v}} &= \arg\max_{\boldsymbol{v}} P(\boldsymbol{v}|\boldsymbol{I}) & \hat{\boldsymbol{S}} &= \arg\max_{\boldsymbol{S}} P(\boldsymbol{S}|\boldsymbol{I}, \hat{\boldsymbol{v}}) \\ \hat{\boldsymbol{R}} &= \arg\max_{\boldsymbol{R}} P(\boldsymbol{R}|\boldsymbol{I}, \hat{\boldsymbol{v}}, \hat{\boldsymbol{S}}) & \hat{\boldsymbol{\ell}} &= \arg\max_{\boldsymbol{\ell}} P(\boldsymbol{\ell}|\boldsymbol{I}, \hat{\boldsymbol{v}}, \hat{\boldsymbol{S}}, \hat{\boldsymbol{R}}) \end{aligned} \tag{4.3}$$

The second challenge is that of actually computing the arg max operations in Equation 3. Rather than attempting to model the probability distributions explicitly and maximizing them, we will, instead, model the estimators in Equation 3 as parametric functions:

$$\begin{aligned} \hat{\boldsymbol{v}} &= f_v(\boldsymbol{I}; \theta_v) & \hat{\boldsymbol{S}} &= f_s(\boldsymbol{I}, \hat{v}; \theta_s) \\ \hat{\boldsymbol{R}} &= f_r(\boldsymbol{I}, \hat{v}, \hat{\boldsymbol{S}}; \theta_r) & \hat{\boldsymbol{\ell}} &= f_\ell(\boldsymbol{I}, \hat{\boldsymbol{v}}, \hat{\boldsymbol{S}}, \hat{\boldsymbol{R}}; \theta_{\boldsymbol{\ell}}) \end{aligned} \tag{4.4}$$

The problem of *learning* to estimate the 3D facial parameters thus effectively reduces to that of estimating the parameters θ_v, θ_s, θ_r and θ_l.

Using the common approach, we formulate the learning process for these parameters through an autoencoder. $f_v()$, $f_s()$, $f_r()$ and $f_\ell()$ are, together, viewed as the learnable encoder in the autoencoder, which estimate $\boldsymbol{v}$, $\boldsymbol{S}$, $\boldsymbol{R}$ and $\boldsymbol{\ell}$ respectively. The decoder is a deterministic differentiable *renderer* $\mathcal{R}()$ with no learnable parameters, which reconstructs the original input $\boldsymbol{I}$ from the values derived by the encoder as $\hat{\boldsymbol{I}} = \mathcal{R}(\boldsymbol{S}, \boldsymbol{R}, \boldsymbol{v}, \boldsymbol{\ell})$. The parameters of the encoder are learned to minimize the error between $\hat{\boldsymbol{I}}$ and $\boldsymbol{I}$.

4.3.2 Facial Parameters Inference

Viewpoint. We first predict the viewpoint parameters from the given image, using a function $f_v(\boldsymbol{I}; \boldsymbol{\theta}_v) : \boldsymbol{I} \rightarrow \boldsymbol{v} \in \mathbb{R}^7$. Here $\boldsymbol{v}$ is used to parameterize the weak perspective transformation [95], including 3D spatial rotation (SO(3)), the translation (xyz coordinates), and the scaling factor.

Shape. The prediction of shape is conditioned on the given image $\boldsymbol{I}$ and the predicted $\boldsymbol{v}$. Since the same face captured with different viewpoints should be correspond to the same facial shape, it is beneficial to exclude as much viewpoint information from the image

$\boldsymbol{I}$ as possible before the shape prediction. With the predicted $\boldsymbol{v}$, we can align the image to its canonical view in 2D plane, as shown in Fig. 4.2. The viewpoint $\boldsymbol{v}$ comprises the scale factor $\boldsymbol{v}_1$, 3D spatial rotation parameters $[\boldsymbol{v}_2, \boldsymbol{v}_3, \boldsymbol{v}_4]$, and 3D translation parameters $[\boldsymbol{v}_5, \boldsymbol{v}_6, \boldsymbol{v}_7]$. The original image $\boldsymbol{I}$ is cropped to its canonical view in 2D plane with viewpoint v. The cropping is given by $(\boldsymbol{I} \circ \boldsymbol{v})(x', y') = \boldsymbol{I}(x, y)$, where the transformation from (x', y') to (x, y) is formulated in the following.

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \exp(\boldsymbol{v}_1) \cdot \cos \boldsymbol{v}_4 & \exp(\boldsymbol{v}_1) \cdot \sin \boldsymbol{v}_4 & \boldsymbol{v}_5 \\ -\exp(\boldsymbol{v}_1) \cdot \sin \boldsymbol{v}_4 & \exp(\boldsymbol{v}_1) \cdot \cos \boldsymbol{v}_4 & \boldsymbol{v}_6 \end{bmatrix} \begin{bmatrix} x' \\ y' \end{bmatrix} \tag{4.5}$$

Bilinear interpolation is used if x or y is not an integer.

The cropped image is denoted by $\boldsymbol{I} \circ \boldsymbol{v}$. A function $f_s(\boldsymbol{I} \circ \boldsymbol{v}; \boldsymbol{\theta}_s) : \boldsymbol{I} \circ \boldsymbol{v} \rightarrow \boldsymbol{\alpha} \in \mathbb{R}^{228 \times 1}$ with learnable parameter $\boldsymbol{\theta}_s$ is constructed to predict the shape coefficients $\boldsymbol{\alpha}$. The shape coefficients $\boldsymbol{\alpha}$ are defined by a statistical model of 3D facial shape:

$$\vec{\boldsymbol{S}} = \bar{\boldsymbol{S}} + \boldsymbol{U}\boldsymbol{\alpha}, \tag{4.6}$$

where $\vec{\boldsymbol{S}} \in \mathbb{R}^{3K \times 1}$ is the vectorized $\boldsymbol{S}$, and $\bar{\boldsymbol{S}} \in \mathbb{R}^{3K \times 1}$ is the mean shape. $\boldsymbol{U} \in \mathbb{R}^{3K \times 228}$ is the PCA basis from Basel Face Model (BFM) [96] and 3DFFA [90] for identity and expression variation, respectively. $\bar{\boldsymbol{S}}$ and $\boldsymbol{U}$ are fixed during the training and testing of CEST. With the predicted $\boldsymbol{\alpha}$, the shape $\boldsymbol{S}$ can be obtained using equation 4.6.

Reflectance. Previous approaches usually predict the reflectance coefficients in a predefined model [5, 4], unwrapped UV map of reflectance [82, 83, 80, 97], or graph representation of the reflectance [98, 93] from the image directly. In CEST, we adopt the UV map representation for reflectance. However, the prediction of the reflectance is conditioned not only on the given image $\boldsymbol{I}$, but also on the predicted viewpoint $\boldsymbol{v}$ and shape $\boldsymbol{S}$.

The process is illustrated in Fig. 4.2. We first compute the image-coordinate facial shape $\boldsymbol{Q} \in \mathbb{R}^{K \times 2}$ by projecting the world-coordinate facial shape $\boldsymbol{S}$ with viewpoint $\boldsymbol{v}$ using weak perspective transformation. The 3D spatial rotation is represented by a rotation vector $\boldsymbol{w} = [\boldsymbol{v}_2; \boldsymbol{v}_3; \boldsymbol{v}_4] \in \mathbb{R}^{3 \times 1}$: the unit vector $\boldsymbol{u} = \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}$ is the axis of rotation, and the magnitude $\phi = \|\boldsymbol{w}\|_2$ is the rotation angle. The weak perspective transformation is used to project the

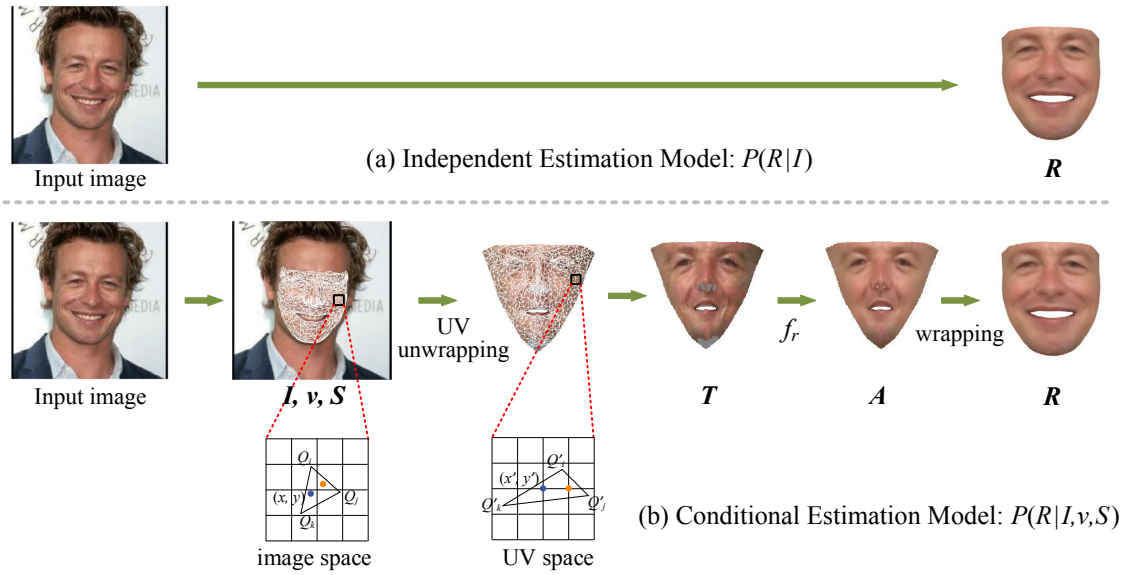


Figure 4.3: Illustration of generating the UV map of the illuminated texture.

world-coordinate facial shape $\boldsymbol{S}$ to image-coordinate $\boldsymbol{Q}$, as formulated in

$$\begin{bmatrix} \boldsymbol{Q}(i,1) \\ \boldsymbol{Q}(i,2) \\ \boldsymbol{Q}(i,3) \end{bmatrix} = \exp(\boldsymbol{v}_1) \cdot \left(\left(\boldsymbol{w}\boldsymbol{w}^\intercal + (\cos\phi) \cdot (1 - \boldsymbol{w}\boldsymbol{w}^\intercal) \right) \begin{bmatrix} \boldsymbol{S}(i,1) \\ \boldsymbol{S}(i,2) \\ \boldsymbol{S}(i,3) \end{bmatrix} + (\sin\phi) \cdot \boldsymbol{w} \times \begin{bmatrix} \boldsymbol{S}(i,1) \\ \boldsymbol{S}(i,2) \\ \boldsymbol{S}(i,3) \end{bmatrix} \right) + \begin{bmatrix} \boldsymbol{v}_5 \\ \boldsymbol{v}_6 \\ \boldsymbol{v}_7 \end{bmatrix}. \quad (4.7)$$

Next, we construct an intermediate representation, *i.e.* UV map of the illuminated texture $\boldsymbol{T}$ [95], which is obtained by unwrapping the given image $\boldsymbol{I}$ based on the predicted face shape $\boldsymbol{Q}$. Subsequently, the UV map of reflectance $\boldsymbol{A}$ is predicted from the illuminated texture $\boldsymbol{T}$ by a reflectance function $f_r(\boldsymbol{T}; \boldsymbol{\theta}_r)$. The reflectance $\boldsymbol{R}$ can be recovered from $\boldsymbol{A}$ by UV wrapping.

The basic idea for computing the $\boldsymbol{T}$ is illustrated in Fig. 4.3. For each $\boldsymbol{T}(x', y')$ (the pixel values at position (x', y')), we trace its corresponding position (x, y) in $\boldsymbol{I}$. The illuminated texture can be simply obtained by $\boldsymbol{T}(x', y') = \boldsymbol{I}(x, y)$, where bilinear interpolation is used for inferring the pixel values of $\boldsymbol{I}$ at position (x, y) if x or y is not an integer. The computation of (x, y) is as follows. First, the canonical face shape $\bar{\boldsymbol{S}}$ is mapped to the UV space by cylinder unwrapping. We determine the triangle enclosing the point (x', y') on a grid based on the vertex connectivity, which is provided by the 3DMM. The triangle is represented by its three vertices $\boldsymbol{Q}'(i)$, $\boldsymbol{Q}'(j)$, and $\boldsymbol{Q}'(k)$. Since the topology of the facial shape in image space and UV space are the same, the vertices in these two space have one-to-one correspondence. We could

easily get the corresponding vertices $\boldsymbol{Q}(i)$, $\boldsymbol{Q}(j)$, and $\boldsymbol{Q}(k)$. Now the position (x, y) can be computed by $x = \kappa_1\boldsymbol{Q}(i,1)+\kappa_2\boldsymbol{Q}(j,1)+\kappa_3\boldsymbol{Q}(k,1)$ and $y = \kappa_1\boldsymbol{Q}(i,2)+\kappa_2\boldsymbol{Q}(j,2)+\kappa_3\boldsymbol{Q}(k,2)$, where the κs are the coefficients computed by $\boldsymbol{Q}'(i)$, $\boldsymbol{Q}'(j)$, $\boldsymbol{Q}'(k)$, and (x', y') in barycentric coordinate system [99]. Given the vertices of a triangle $(\boldsymbol{Q}(i), \boldsymbol{Q}(j), \boldsymbol{Q}(k))$ and its enclosing grid point (x, y) on image. The barycentric coefficients can be computed by

$$\begin{gathered}
\boldsymbol{d}_i = \begin{bmatrix} \boldsymbol{Q}(j,1) - \boldsymbol{Q}(i,1) \\ \boldsymbol{Q}(j,2) - \boldsymbol{Q}(i,2) \end{bmatrix}, \quad \boldsymbol{d}_j = \begin{bmatrix} \boldsymbol{Q}(k,1) - \boldsymbol{Q}(i,1) \\ \boldsymbol{Q}(k,2) - \boldsymbol{Q}(i,2) \end{bmatrix}, \quad \boldsymbol{d}_k = \begin{bmatrix} x - \boldsymbol{Q}(i,1) \\ y - \boldsymbol{Q}(i,2) \end{bmatrix}, \\
d_{ii} = \boldsymbol{d}_i^\intercal \boldsymbol{d}_i, \quad d_{jj} = \boldsymbol{d}_j^\intercal \boldsymbol{d}_j, \quad d_{ij} = \boldsymbol{d}_i^\intercal \boldsymbol{d}_j, \quad d_{ki} = \boldsymbol{d}_k^\intercal \boldsymbol{d}_i, \quad d_{kj} = \boldsymbol{d}_k^\intercal \boldsymbol{d}_i, \\
\kappa_2 = \frac{d_{jj}d_{ki} - d_{ij}d_{kj}}{d_{ii}d_{jj} - d_{ij}d_{ij}}, \quad \kappa_3 = \frac{d_{ii}d_{kj} - d_{ij}d_{ki}}{d_{ii}d_{jj} - d_{ij}d_{ij}}, \quad \kappa_1 = 1 - \kappa_2 - \kappa_3.
\end{gathered} \tag{4.8}$$

The barycenteric coefficients κ_1, κ_2, and κ_3 are in the range of $[0, 1]$ if the grid point (x, y) is in the triangle. For the invisible triangles (caused by self-occlusion), we simply ignore them.

With the illuminated texture $\boldsymbol{T}$, the UV map of the reflectance $\boldsymbol{A}$ can be produced by a function $f_r(\boldsymbol{T}; \boldsymbol{\theta}_r)$, where $\boldsymbol{\theta}_r$ is the learnable parameters. It is worth noting that the input $(\boldsymbol{T})$ and output $(\boldsymbol{A})$ of f_r are spatially aligned in UV space, so the learning process can be greatly facilitated. Subsequently, the reflectance $\boldsymbol{R}$ is obtained by a wrapping function $\boldsymbol{R} = \Psi(\boldsymbol{A})$ [95], which has no learnable parameters. The wrapping function $\Psi : \boldsymbol{A} \in \mathbb{R}^{256\times256\times3} \rightarrow \boldsymbol{R} \in \mathbb{R}^{K\times3}$ is defined as $\boldsymbol{R}(i) = \boldsymbol{A}(\boldsymbol{U}(i,1), \boldsymbol{U}(i,2))$, where i is the index for the vertices of a 3D face. $\boldsymbol{R}(i)$ and $\boldsymbol{A}(\boldsymbol{U}(i,1), \boldsymbol{U}(i,2))$ are 3-dimensional vectors. $\boldsymbol{U} \in \mathbb{R}^{K\times2}$ is the coordinates of shape in UV space from 3DMM [76]. Again, bilinear interpolation is used if $\boldsymbol{U}(i,1)$ or $\boldsymbol{U}(i,2)$ is not an integer.

Illumination. Following the previous studies [87, 83], we assume the distant smooth illumination and purely *Lambertian* surface properties [100]. Spherical Harmonics (SH) [101] are employed to approximate the incident radiance at a surface. We use 3 SH bands, leading to 9 SH coefficients. The illumination function is defined as $f_{\boldsymbol{\ell}}(\boldsymbol{I}, \boldsymbol{T}, \boldsymbol{A}; \boldsymbol{\theta}_{\boldsymbol{\ell}}) : (\boldsymbol{I}, \boldsymbol{T}, \boldsymbol{A}) \rightarrow \boldsymbol{\ell} \in \mathbb{R}^{9\times1}$, which takes the given image, illuminated texture map and UV map of reflectance as input, and produces the illumination parameters.

So far, the 3D face model parameters $\boldsymbol{R}$, $\boldsymbol{S}$, $\boldsymbol{v}$, and $\boldsymbol{\ell}$ are predicted, and we are able

to recombine them and render the image by the expert-designed rendering module, $\hat{\boldsymbol{I}} = \mathcal{R}(\boldsymbol{S}, \boldsymbol{R}, \boldsymbol{v}, \boldsymbol{\ell})$.

4.3.3 Objectives for Self-Supervised Learning

The functions f_s, f_r, f_v, and f_ℓ are modelled by convolutional neural networks (CNNs) with learnable parameters $\boldsymbol{\theta}_s$, $\boldsymbol{\theta}_r$, $\boldsymbol{\theta}_v$, and $\boldsymbol{\theta}_\ell$, respectively. Since all the learning modules and expert-designed renderer are differentiable, the proposed framework is end-to-end trainable. The learning objective is to minimize the differences between the original image $\boldsymbol{I}$ and the rendered image $\hat{\boldsymbol{I}}$. Following the practices in previous work, the learning objective does not include the pixels in nonface region, *e.g.* hair, sunglasses, scarf, etc. We identify if a pixel belongs to face or nonface region by a face segmentation network f_{seg}, which is trained on CelebAMask-HQ dataset [102] with the segmentation labels provided in the dataset. Once trained, f_{seg} is fixed during the training and testing of CEST. We denote the effective face region as a mask $\boldsymbol{M}$, so the pixel at position (x, y) is included in reconstruction if $\boldsymbol{M}(i, j) = 1$, and excluded if $\boldsymbol{M}(i, j) = 0$. The photometric loss can be written as

$$\begin{aligned} \mathcal{L}_{ph} &= \mathcal{E}(\boldsymbol{I}, \boldsymbol{S}, \boldsymbol{R}, \boldsymbol{v}, \boldsymbol{\ell}, \boldsymbol{M}) \\ &= \|\boldsymbol{M} \otimes \boldsymbol{I} - \boldsymbol{M} \otimes \hat{\boldsymbol{I}}\|_1 \\ &= \|\boldsymbol{M} \otimes \boldsymbol{I} - \boldsymbol{M} \otimes \mathcal{R}(\boldsymbol{S}, \boldsymbol{R}, \boldsymbol{v}, \boldsymbol{\ell})\|_1, \end{aligned} \tag{4.9}$$

where $\|\cdot\|_1$ measure the ℓ_1 distance and $\otimes$ denotes the element-wise multiplication. However, if we simply optimize $\mathcal{L}_{ph}$, CEST will learn a degraded solution, where the reflectance $\boldsymbol{A}$ simply copies the pixel values from $\boldsymbol{T}$, and $\boldsymbol{\ell}$ yields an isotropic radiator, radiating the same intensity of radiation in all directions. In this case, CEST does not learn semantically disentangled facial parameters, but leads to a perfect reconstruction for $\hat{\boldsymbol{I}}$.

To avoid this, we adopt the symmetry and consistency constraints for reflectance. The facial reflectance is assumed to be horizontally symmetric and consistent in a video clip. Suppose $\boldsymbol{I}_i$ and $\boldsymbol{I}_j$ are two face images from the same video clip. One of the possible solutions is to add the regularization terms $\|\boldsymbol{R}_i - \boldsymbol{R}_i^{\bowtie}\|$, $\|\boldsymbol{R}_j - \boldsymbol{R}_j^{\bowtie}\|$, and $\|\boldsymbol{R}_i - \boldsymbol{R}_j\|$ to the

learning objective, where $\boldsymbol{R}_i^{\bowtie}$ and $\boldsymbol{R}_j^{\bowtie}$ are the horizontally flipped versions of $\boldsymbol{R}_i$ and $\boldsymbol{R}_j$. However, it is difficult to tune loss weights to balance the reconstruction and regularization terms. Instead, we adopt an alternative solution by constructing additional reconstruction terms as constraints [84]. The learning objective for reconstructing $\boldsymbol{I}_i$ and $\boldsymbol{I}_j$ can be written as

$$\begin{aligned}\mathcal{L}_{ph} = &\mathcal{E}(\boldsymbol{I}_i, \boldsymbol{S}_i, \boldsymbol{R}_i, \boldsymbol{v}_i, \boldsymbol{\ell}_i, \boldsymbol{M}_i) + \mathcal{E}(\boldsymbol{I}_j, \boldsymbol{S}_j, \boldsymbol{R}_j, \boldsymbol{v}_j, \boldsymbol{\ell}_j, \boldsymbol{M}_j) \\ &+ \mathcal{E}(\boldsymbol{I}_i, \boldsymbol{S}_i, \boldsymbol{R}_j, \boldsymbol{v}_i, \boldsymbol{\ell}_i, \boldsymbol{M}_i) + \mathcal{E}(\boldsymbol{I}_j, \boldsymbol{S}_j, \boldsymbol{R}_i, \boldsymbol{v}_j, \boldsymbol{\ell}_j, \boldsymbol{M}_j) \\ &+ \mathcal{E}(\boldsymbol{I}_i, \boldsymbol{S}_i, \boldsymbol{R}_i^{\bowtie}, \boldsymbol{v}_i, \boldsymbol{\ell}_i, \boldsymbol{M}_i) + \mathcal{E}(\boldsymbol{I}_j, \boldsymbol{S}_j, \boldsymbol{R}_j^{\bowtie}, \boldsymbol{v}_j, \boldsymbol{\ell}_j, \boldsymbol{M}_j) \\ &+ \mathcal{E}(\boldsymbol{I}_i, \boldsymbol{S}_i, \boldsymbol{R}_j^{\bowtie}, \boldsymbol{v}_i, \boldsymbol{\ell}_i, \boldsymbol{M}_i) + \mathcal{E}(\boldsymbol{I}_j, \boldsymbol{S}_j, \boldsymbol{R}_i^{\bowtie}, \boldsymbol{v}_j, \boldsymbol{\ell}_j, \boldsymbol{M}_j)\end{aligned} \tag{4.10}$$

Stochastic optimization. As can be seen, the number of reconstruction terms is increased dramatically. From n frames of the same video, $2n^2$ reconstruction terms can be constructed. This is not scalable. To address this problem, we propose to optimize the learning objective in a stochastic way. For each training iteration, only a subset of the reconstruction terms are optimized. Specifically, a set of video frames $\{\boldsymbol{I}_1, \boldsymbol{I}_2, ..., \boldsymbol{I}_N\}$ are randomly sampled from different videos. The frames are grouped by videos, labeled as $\boldsymbol{\xi} = \{\xi_1, \xi_2, ..., \xi_N\}$. For any $\boldsymbol{I}_i$, instead of enumerating all the possible reflectances and obtaining numerous reconstruction terms, we randomly select some other frame from the same video, denoted as $\boldsymbol{I}_j$ (under the condition of $\xi_j = \xi_i$), and use $\boldsymbol{R}_j$ and $\boldsymbol{R}_i^{\bowtie}$ to construct two reconstruction terms for $\boldsymbol{I}_i$. With this strategy, the number of reconstruction terms is reduced from $O(n^2)$ to $O(n)$. Formally, the learning objective can be written as

$$\mathcal{L}_{ph} = \frac{1}{N} \sum_{i=1, \xi_j = \xi_i}^{N} \big(\mathcal{E}(\boldsymbol{I}_i, \boldsymbol{S}_i, \boldsymbol{R}_j, \boldsymbol{v}_i, \boldsymbol{\ell}_i, \boldsymbol{M}_i) + \mathcal{E}(\boldsymbol{I}_i, \boldsymbol{S}_i, \boldsymbol{R}_i^{\bowtie}, \boldsymbol{v}_i, \boldsymbol{\ell}_i, \boldsymbol{M}_i)\big). \tag{4.11}$$

To stabilize the training of CEST, we use 2D key points via $\mathcal{L}_{kp} = \frac{1}{NN_{kp}} \sum_{i=1}^{N} \sum_{j=1}^{N_{kp}} \|\boldsymbol{Q}_i(k_j) - \boldsymbol{q}_i(j)\|_1$ where $\boldsymbol{q}(j)$ is the set of detected 2D key points on image, and k_j is the index of the vertex associating to the 2D key point. We also regularize the energies of shape coefficients with $\mathcal{L}_{rg} = \frac{1}{N} \sum_{i=1}^{N} \|\boldsymbol{\alpha}_i\|_2^2$. An off-the-shelf landmark detector [70] is used to produce $N_{kp} = 68$ key points for a detected face. The total loss consists of the following

terms:

$$\mathcal{L} = \mathcal{L}_{ph} + \lambda_1 \mathcal{L}_{kp} + \lambda_2 \mathcal{L}_{rg} \tag{4.12}$$

where λ_1 and λ_2 are hyperparameters.

4.4 Experiments

We qualitatively and quantitatively evaluate CEST with ablation experiments and comparisons to state-of-the-art methods [5, 42, 7, 103]. In ablation experiments, we compare CEST to the independent version of CEST (IEST) where facial parameters are estimated in a uncoupled way, and other variants trained with different constraints. Qualitative results include the predicted shape, reflectance, illumination, reconstructed face, etc. We also show the relighted faces, which are obtained by illuminating reflectances with different illuminations. Quantitative results evaluate the qualities of the predicted shape and rendered face. The metrics we used are normalized mean error (NME) [104] and photometric error for shape and rendered face, respectively. NME is defined as the average per-vertex Euclidean distance between the predicted and targeted point clouds normalized by the outer 3D interocular distance. Photometric error is the mean absolute errors between pixel values in the original images and reconstruction images.

4.4.1 Experimental Settings

For fair comparison, we train two separate CEST models with VoxCeleb1 [3] and 300W-LP [90] respectively. VoxCeleb1 is a video dataset collected from the Internet. The videos of speakers are captured in different in-the-wild scenarios. A subset of 4,727 videos of 267 persons are used in the training, leading to 6,279,609 video frames. The faces in video frames are cropped to the size of 256×256 based on the detected facial key points using [70]. 300W-LP is a synthetic image dataset, containing 122,450 images provided with dense

Table 4.1: CNN architectures for viewpoint, illumination, and shape networks: details.

Viewpoint & Illumination Network			Shape Network		
Layer	Act.	Output shape	Layer	Act.	Output shape
Input	-	$256 \times 256 \times 3$	Input	-	$256 \times 256 \times 3$
Conv $4 \times 4_{/2,1}$	BN + ReLU	$128 \times 128 \times 32$	Conv $4 \times 4_{/2,1}$	BN + ReLU	$128 \times 128 \times 64$
Conv $4 \times 4_{/2,1}$	BN + ReLU	$64 \times 64 \times 32$	Conv $4 \times 4_{/2,1}$	BN + ReLU	$64 \times 64 \times 64$
Conv $4 \times 4_{/2,1}$	BN + ReLU	$32 \times 32 \times 64$	Conv $4 \times 4_{/2,1}$	BN + ReLU	$32 \times 32 \times 128$
Conv $4 \times 4_{/2,1}$	BN + ReLU	$16 \times 16 \times 64$	Conv $4 \times 4_{/2,1}$	BN + ReLU	$16 \times 16 \times 128$
$\begin{bmatrix}\text{Conv3} \times 3_{/1,1} \\ \text{Conv3} \times 3_{/1,1}\end{bmatrix}$	BN + ReLU BN + ReLU	$16 \times 16 \times 64$ $16 \times 16 \times 64$	$\begin{bmatrix}\text{Conv3} \times 3_{/1,1} \\ \text{Conv3} \times 3_{/1,1}\end{bmatrix}$	BN + ReLU BN + ReLU	$16 \times 16 \times 128$ $16 \times 16 \times 128$
Conv $4 \times 4_{/2,1}$	BN + ReLU	$8 \times 8 \times 128$	Conv $4 \times 4_{/2,1}$	BN + ReLU	$8 \times 8 \times 256$
$\begin{bmatrix}\text{Conv3} \times 3_{/1,1} \\ \text{Conv3} \times 3_{/1,1}\end{bmatrix}$	BN + ReLU BN + ReLU	$8 \times 8 \times 128$ $8 \times 8 \times 128$	$\begin{bmatrix}\text{Conv3} \times 3_{/1,1} \\ \text{Conv3} \times 3_{/1,1}\end{bmatrix}$	BN + ReLU BN + ReLU	$8 \times 8 \times 256$ $8 \times 8 \times 256$
Conv $4 \times 4_{/2,1}$	BN + ReLU	$4 \times 4 \times 128$	Conv $4 \times 4_{/2,1}$	BN + ReLU	$4 \times 4 \times 256$
Conv $4 \times 4_{/2,1}$	-	$1 \times 1 \times d$	Conv $4 \times 4_{/2,1}$	-	$1 \times 1 \times 228$

landmarks. Since we focus on self-supervised learning, we only use a sparse set of 68 sparse landmarks as a regularization in training.

Network Architecture. We use standard encoder networks for viewpoint, shape and illumination predictions, and a network similar to U-Net [105] for reflectance prediction. The detailed configurations are given in Table 4.1. Parameter d is 7 for viewpoint network f_v and 9 for illumination network f_ℓ. Conv $3_{/2,1}$ denotes convoluitonal layer with kernel size of 3, where the stride and padding are 2 and 1, respectively. Each convolutional layer is followed by a Batch Normalization (BN) [57] layer and Rectified Linear Units (ReLU). Bilinear interpolation is adopted for the upsampling operation. Specifically, in Table 4.1, the layers in brackets are residual blocks. In Table 4.2, we use shortcut to connect the feature maps of encoder and decoder, but different from U-Net, we use addition rather than concatenation to integrate information in the feature maps. For those encoder output shapes in brackets (*e.g.* "[128×128×64]"), the feature map will be added as a shortcut to the decoder feature map (also with the same brackets).

Training. For the training with VoxCeleb1, the minibatch consists of 128 video frames from 32 clips. For each video clip, we randomly selected 4 video frames. The training is completed at 50K iterations. For the training with 300W-LP, the minibatch consisted 128

Table 4.2: CNN architectures for reflectance networks: details. The layers in the decoder (from input to output) are listed from bottom to top.

Reflectance Network					
U-Net Encoder (↓)			U-Net Decoder (↑)		
Encoder Layer	Act.	Output shape	Decoder Layer	Act.	Output shape
Input	-	$256 \times 256 \times 3$	Output	-	$256 \times 256 \times 3$
-	-	-	Conv $3 \times 3_{/1,1}$	Tanh	$256 \times 256 \times 3$
-	-	-	Conv $3 \times 3_{/1,1}$	BN + ReLU	$256 \times 256 \times 3$
Conv $4 \times 4_{/2,1}$	BN + ReLU	$128 \times 128 \times 64$	Upsample (2×)	-	$256 \times 256 \times 64$
Conv $3 \times 3_{/1,1}$	BN + ReLU	$[128 \times 128 \times 64]$	Conv $3 \times 3_{/1,1}$	BN + ReLU	$[128 \times 128 \times 64]$
-	-	-	Conv $3 \times 3_{/1,1}$	BN + ReLU	$128 \times 128 \times 64$
Conv $4 \times 4_{/2,1}$	BN + ReLU	$64 \times 64 \times 64$	Upsample (2×)	-	$128 \times 128 \times 64$
Conv $3 \times 3_{/1,1}$	BN + ReLU	$[64 \times 64 \times 64]$	Conv $3 \times 3_{/1,1}$	BN + ReLU	$[64 \times 64 \times 64]$
-	-	-	Conv $3 \times 3_{/1,1}$	BN + ReLU	$64 \times 64 \times 64$
Conv $4 \times 4_{/2,1}$	BN + ReLU	$32 \times 32 \times 128$	Upsample (2×)	-	$64 \times 64 \times 128$
Conv $3 \times 3_{/1,1}$	BN + ReLU	$[32 \times 32 \times 128]$	Conv $3 \times 3_{/1,1}$	BN + ReLU	$[32 \times 32 \times 128]$
-	-	-	Conv $3 \times 3_{/1,1}$	BN + ReLU	$32 \times 32 \times 128$
Conv $4 \times 4_{/2,1}$	BN + ReLU	$16 \times 16 \times 128$	Upsample (2×)	-	$32 \times 32 \times 128$
Conv $3 \times 3_{/1,1}$	BN + ReLU	$[16 \times 16 \times 128]$	Conv $3 \times 3_{/1,1}$	BN + ReLU	$[16 \times 16 \times 128]$
-	-	-	Conv $3 \times 3_{/1,1}$	BN + ReLU	$16 \times 16 \times 128$
Conv $4 \times 4_{/2,1}$	BN + ReLU	$8 \times 8 \times 256$	Upsample (2×)	-	$16 \times 16 \times 256$
Conv $3 \times 3_{/1,1}$	BN + ReLU	$[8 \times 8 \times 256]$	Conv $3 \times 3_{/1,1}$	BN + ReLU	$[8 \times 8 \times 256]$
Conv $4 \times 4_{/2,1}$	BN + ReLU	$4 \times 4 \times 256$	Conv $3 \times 3_{/1,1}$	BN + ReLU	$8 \times 8 \times 256$
Conv $3 \times 3_{/1,1}$	BN + ReLU	$4 \times 4 \times 256$	Upsample (2×)	-	$8 \times 8 \times 256$

randomly selected images, and the total iteration is 20K. For both models, we used Adam [72] optimizer with learning rate of 0.001. λ_1 and λ_2 are 1 and 0.1 unless stated otherwise.

4.4.2 Ablation Experiments

The results of ablation study are shown in Fig. 4.4. We first present the original and reconstructed image (overlay) for comparison, following by the reflectance, illuminated texture, facial shape (geometry), and illumination in canonical view. The results are (a) CEST with two constraints; (b) uncoupled CEST with two constraints; (c) CEST with only reflectance consistency constraint; (d) CEST with reflectance symmetry constraint (the number of video frames is 1); (e) CEST with no constraint on reflectance; (f) reflectance consistency is applied to videos, not video clips.

CEST and IEST. IEST is trained with the same settings as CEST, except the facial parameters are estimated independently from image during training and testing. The

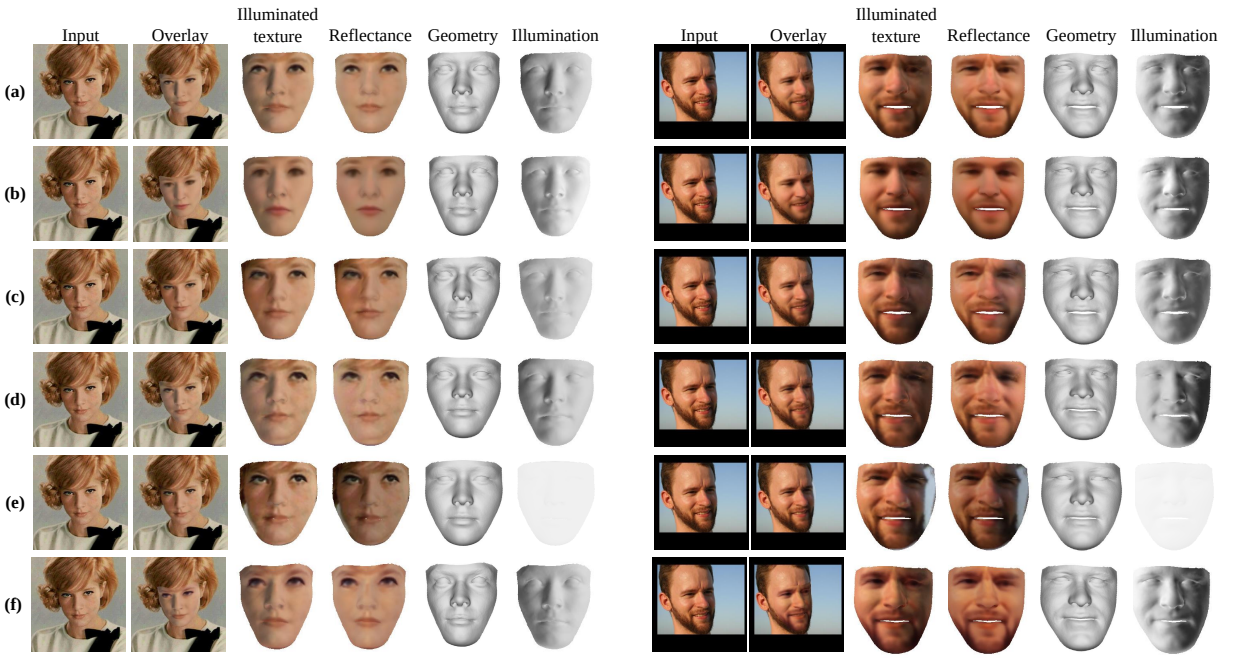


Figure 4.4: Ablation studies with different constraints.

results are shown in Fig. 4.4 (a) and (b), respectively. We can see that CEST produces realistic overlay, disentangled reflectance and illumination, and geometry with personal characteristics and expressions. Compared to CEST, IEST achieves reasonable results, but the reflectances are not as detailed as those from CEST, resulting in inferior overlays and illuminated textures. It validates our hypothesis that the coupled estimation can better formulate the problem and facilitate the learning.

Reflectance symmetry and consistency constraints. We train multiple variants of CEST with only symmetry constraint, only consistency constraints, and without the two constraints, and show their results in Fig. 4.4 (c), (d), and (e), respectively. Compared (a) and (c) we observe that the reflectance symmetry constraint leads to better reflectance and illumination separation. This is because the horizontally flipped video frames can provide more illumination variations to the training set, enabling CEST to learn to model different illuminations properly. On the other hand, if the reflectance consistency in video clip is not used, the decomposition of reflectance and illumination is not performed well. Some illumination remains around the eyes region in the reflectance (see the right hand side of

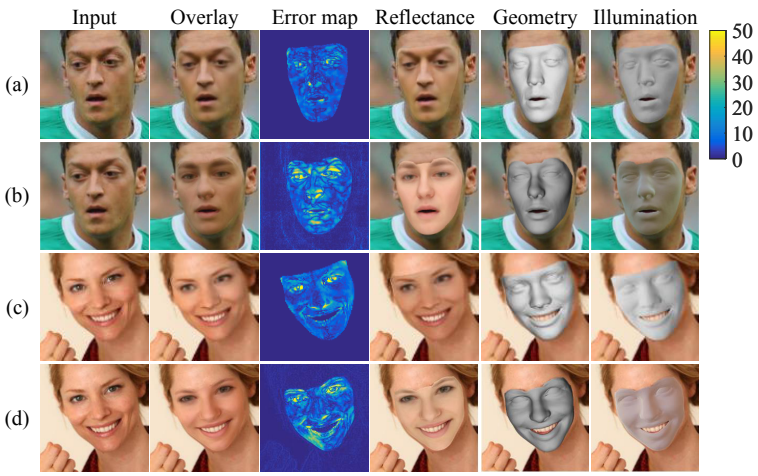


Figure 4.5: Comparisons with MoFA. (a) and (c) are results from CEST. (b) and (d) are results from MoFA.

the Fig. 4.4 (d)). Lastly, if we do not use any constraints on reflectance, CEST learns the degraded solution (Fig. 4.4(e)), where the reflectance simply copies the pixel values from the image, and illumination is an isotropic radiator, radiating the same intensity of radiation in all directions. Moreover, we note that the degraded solution also affects the learned facial shape, which has less personal characteristics in Fig. 4.4 (e). Fig. 4.4 (f) shows the results from CEST trained with reflectance consistency across video. The performance is comparable to those from CEST trained with default setting (reflectance consistency across video clip). It shows that consistency constraint can be generalized to longer videos if the recording environments are not changed dramatically.

4.4.3 Qualitative Results

In this section, we compare CEST to most relevant state-of-art methods with qualitative results.

Comparison to MoFA [5]. MoFA is a fully model-based framework. Its representation power is limited by the linear 3DMM model. In addition, all facial parameters from MoFA are independently predicted from the original image. On the contrary, we use a model-free method for reflectance, and the whole inference process is based on coupled estimation. We

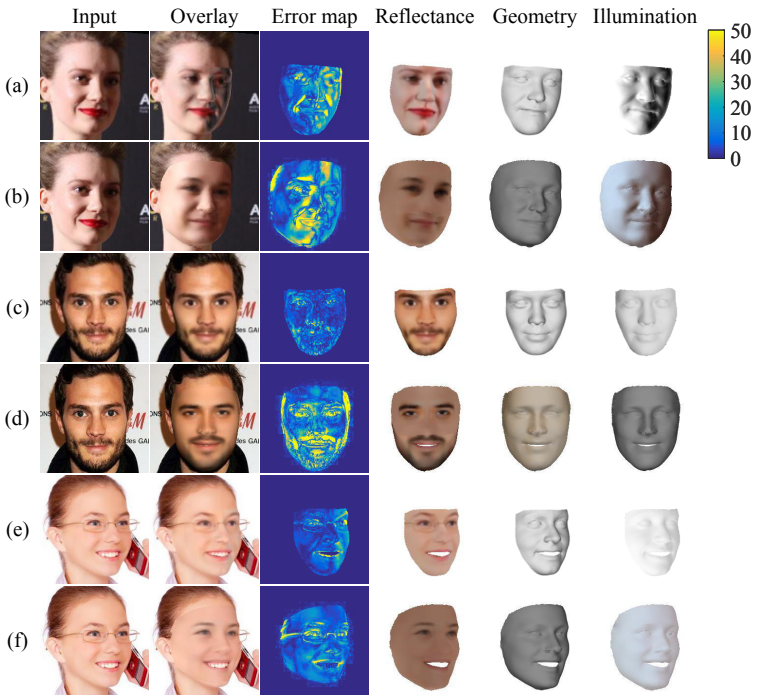


Figure 4.6: Comparisons with nonlinear 3DMM. (a), (c), and (e) are results from CEST. (b), (d), and (f) are results from N3DMM.

visualize the overlay, reflectance, geometry, illumination , as well as the errors between input and rendered image (overlays) in Fig. 4.5. As can be observed, results from MoFA suffer from out-of-subspace reflectance variations. Compared to MoFA, we obtain comparable shape, but significantly better reflectance, illumination, and rendered face by capturing more details.

Comparison to N3DMM [83]. N3DMM generalizes 3DMM model to a nonlinear space and improves the quality of rendered faces. However, N3DMM also infers the reflectance from the input image only, and uses too many heuristic constraints, e.g. reflectance constancy, shape smoothness, supervised pretraining, etc. So their models can only capture low-frequency variations on reflectance. For example, in Fig. 4.6 (b) the lip stick is missing in the reflectance, and the skin colors in reflectances are almost identical for different persons. These limitations lead to higher reconstruction error. In contrast, our results produce realistic reconstruction, with more accurate reflectance and illumination, as well as lower reconstructed error (Fig. 4.6).

Comparison to FML [7]. FML properly incorporates video clues in training and can

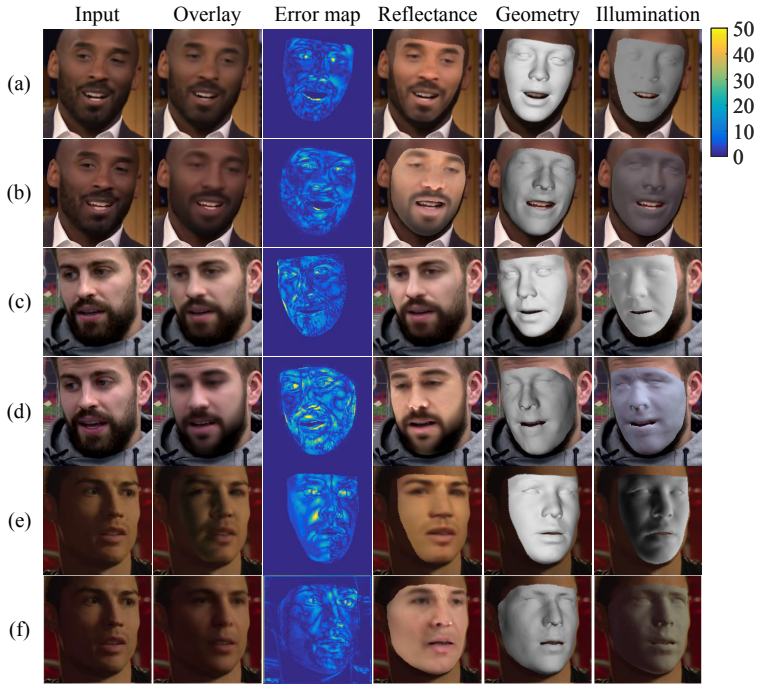


Figure 4.7: Comparisons with FML. (a), (c), and (e) are results from CEST. (b), (d), and (f) are results from FML. Images are from the video frames in VoxCeleb1 dataset [3]

render realistic faces. However, its reconstructed reflectances are prone to an average skin color. In comparison, CEST yields more accurate skin color (see Fig. 4.7 (a), (c), and (e)) by incorporating the learned shape and viewpoint in the estimation of reflectance. Qualitative results clearly show that our results have more reasonable disentanglement between reflectance and illumination. They also contribute to better visual quality of rendered faces. Notably, there are considerable differences in the eye and nose regions from the overlay in Fig. 4.7.

Relighting. Since CEST predicts the reflectances of faces, they can be easily re-lighted with different lighting conditions. Fig. 4.8 shows the re-lit faces in canonical view. In particular, the last two target faces are under harsh lighting, which also examines the illumination removal ability of CEST. The re-lit results again validate that CEST is capable of estimating well-disentangled facial parameters and capturing the reflectance and illumination variations in real-world face images.

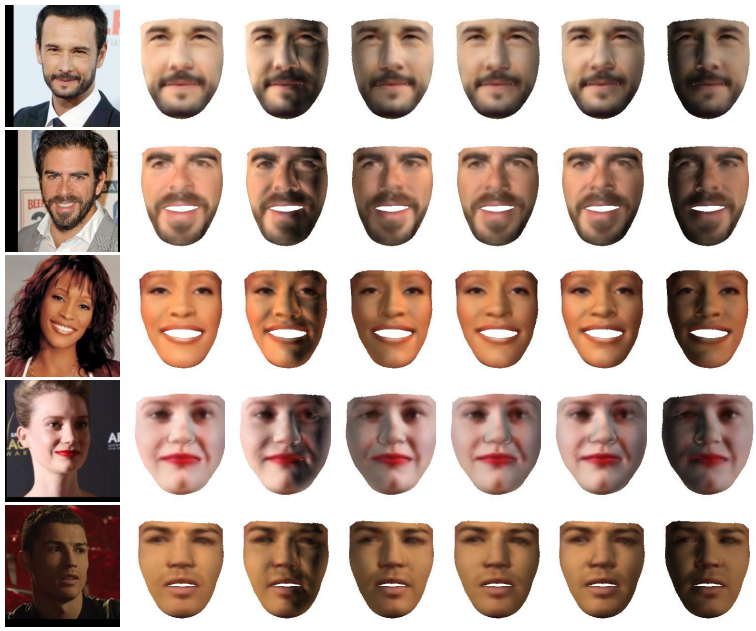
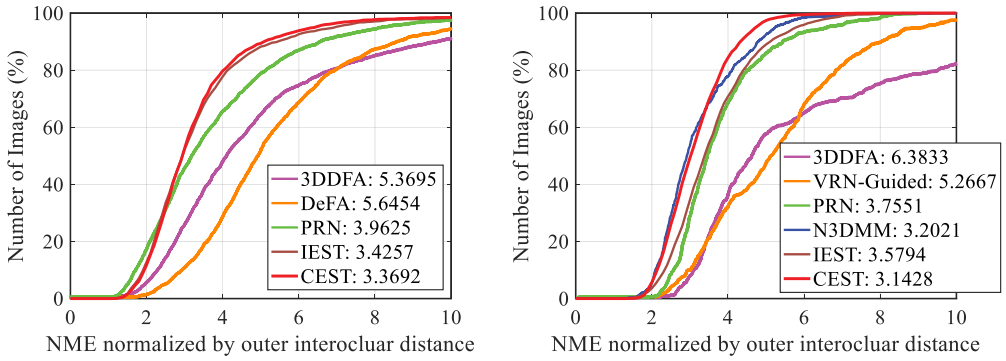


Figure 4.8: Lighting transfer results.

4.4.4 Quantitative Results

We first perform quantitative evaluations on the AFLW2000-3D dataset, including 2,000 unconstrained face images with large pose variations. The ground truth of AFLW2000-3D is given by the results from 3DMM fitting, which may be somewhat noisy. The second evaluation is on MICC Florence 3D Face dataset, which consists of high-resolution 3D scans from 53 subjects. We follow the practices in [104] to render 2,550 testing images using the provided 3D scans. Each subject is rendered in 20 difference poses using a pitch of -15, 20 or 25 degrees and a yaw of -80, -40, 0, 40 or 80 degrees.

In order to compare with previous work, NME is computed based on a set of 19,618 vertices defined by [104] in their evaluation. The point correspondences are determined by the iterative closest point (ICP) algorithm [106]. We compute the cumulative errors distribution (CED) curves and compare it to current prevailing methods such as 3DDFA [90], DeFA [107], and PRN [103] on AFLW2000-3D. For MICC, we compare CEST to 3DDFA [90], VRN [104], and PRN [103]. The results are given in Fig. 4.9. CEST achieves 3.37 and 3.14 NME on AFLW2000-3D and MICC datasets, respectively. More interestingly, our method



(a) CED Curves on AFLW2000-3D dataset (b) CED Curves on MICC dataset

Figure 4.9: CED curves on AFLW2000-3D and MICC datasets. For example, a point at (4, 63) means that 63% of images have NME less than 4.

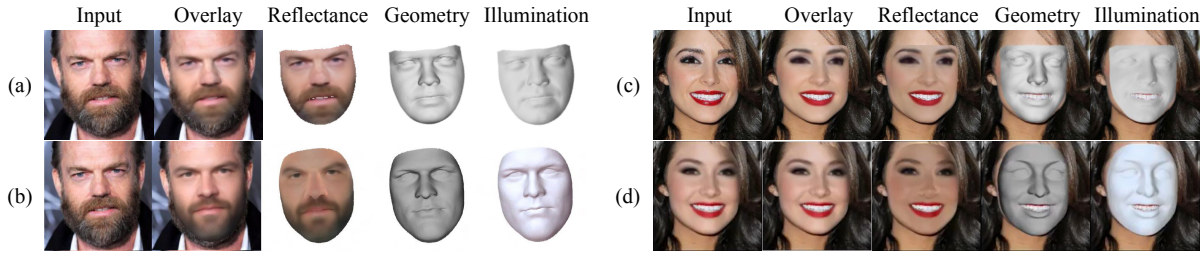


Figure 4.10: Comparisons to [4]. (a) and (c) Results from CEST. (b) and (d) Results from [4].

performs better than the *fully supervised* techniques for shape estimation, e.g. 3DDFA (5.37 on AFLW2000-3D and 6.38 on MICC) and PRN (3.96 on AFLW2000-3D and 3.76 on MICC). Additionally, our method can also estimate facial reflectance and illumination, while both 3DDFA and PRN can not. Compared to N3DMM on MICC dataset, CEST achieves slightly lower NME (3.14 vs. 3.20). Notably, N3DMM uses dense landmarks for supervised pretraining while CEST only uses the 68 sparse landmarks.

Qualitative comparisons. we show more comparisons to the state-of-art methods [8, 9, 6, 108]. Since there is no publicly available implementations for these methods, we compare to the results presented in their papers.

Overall, CEST produces more stable and reasonable geometries, detailed reflectances, and realistic reconstructions of the 3D faces. As shown in Fig. 4.10 (a) (b), Fig. 4.11, Fig. 4.12, and Fig. 4.13, the facial shapes predicted by CEST are more accurate in facial

expressions and lip closure. In addition, the predicted reflectances show more personal characteristics, but less remaining illumination, as illustrated in Fig. 4.12. Lastly, CEST yields faithful 3D reconstructions, capturing more details than the other methods (see Fig 4.11).

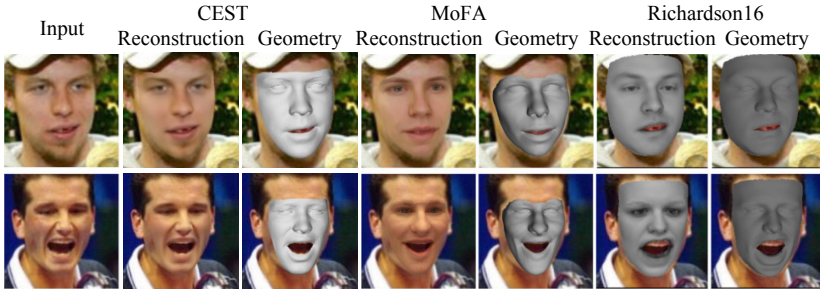


Figure 4.11: Comparisons to MoFA [5] and [6]. Our estimated shapes show more accurate expressions.

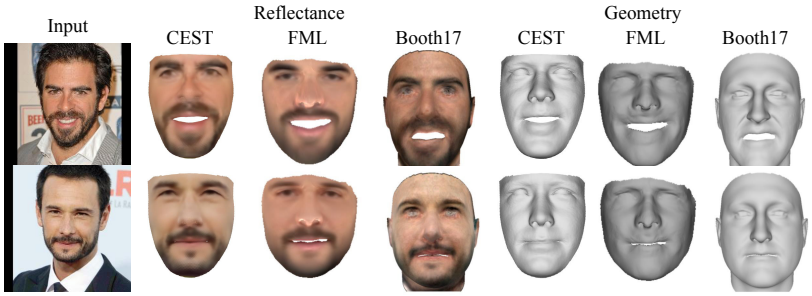


Figure 4.12: Comparing CEST to FML [7] and [8].

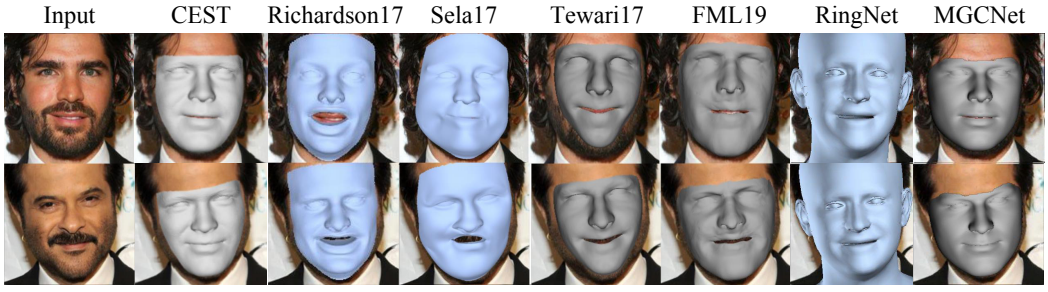


Figure 4.13: Comparing the estimated shapes from CEST to those from [6], [9], [5], [7], [10], and [11] (from left to right). Our estimated shapes are more stable and accurate.

Challenging Cases. We present some examples with dark skin in Fig. 4.14. Although most people in the training set (VoxCeleb) are Caucasian, CEST still produces reasonable

illumination and albedo for these examples. One limitation is that the reconstruction of the non-lambertian surface is inaccurate, e.g. eyes with unusual gaze directions.

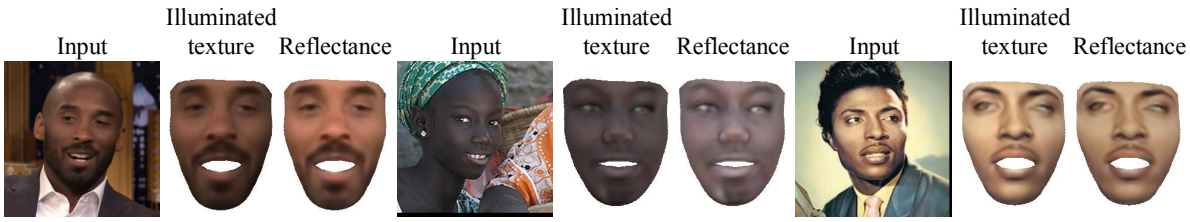


Figure 4.14: Some challenging examples.

Photometric Error. We compare CEST, IEST, FML [7] and Garrido [109] on overlay face reconstruction. To measure the quality of the overlay images, we compute the average photometric error (R,G,B pixel values are from 0 to 255) between the input face image and the overlay face image. We experiment on 1,000 images in CelebA dataset [110]. Table 4.3 shows that the conditional estimation is beneficial for reconstructing the 3D face, and the proposed CEST outperforms existing methods by a large margin.

Table 4.3: Photometric errors obtained by different methods.

Method	**CEST**	IEST	FML [7]	Garrido16 [109]
Photometric Error	**10.74**	13.76	20.65	21.95

4.5 Discussion

We have proposed a conditional estimation framework, called CEST, for 3D face reconstruction from single-view images. CEST addresses the reconstruction problem with a more general formulation, which does not assume conditional independence. We have also proposed a specific decomposition for the conditional probability of different 3D facial parameters. Together with the reflectance symmetry and consistency constraints, CEST can be trained efficiently with video datasets. Both qualitative and quantitative results prove that the conditional estimation is useful. CEST is able to produce high quality and well-disentangled facial parameters for single-view images.

The proposed CEST can be improved from many aspects. Firstly, more accurate and unambiguous facial parameters can be obtained by exploring the temporal information in video. Second, the performance of shape estimation can be boosted by a more advanced morphable model, which also benefits the subsequent estimations of other facial parameters. Moreover, adding perceptual loss could also be an effective way to improve the visual quality of the facial parameters.

Chapter 5

3D Facial Shape Reconstruction from Voice

So far, we are able to generate visually plausible 2D face images from voice, and the generated faces do have common demographic attributes with the true speakers. However, image representation has several inevitable limitations. First, there are many voice-unrelated aspects in the reconstructed images, such as hair, glasses, make-up, hat, etc. The varying viewpoints, illuminations, and background also introduce unexpected variations to the reconstruction, making the results even less explainable. Second, it is nontrivial to compare the generated and real faces objectively. The most commonly used metrics for image reconstruction are Fréchet inception distance (FID) [111] and inception score (IS) [112]. These metrics compare the differences of generated and real images, rather than the faces, so they can not serve as objective metrics to quantify how close the generated faces are to the ground truths. Motivated by the limitations of image representations, we propose to reconstruct 3D facial shapes from voice.

5.1 Introduction

General 3D facial shape is represented by the 3D coordinates of a number of points on its surface called vertices [76]. We can visually perceive the facial characteristics from these vertices, like nose, eyes, mouth, jaw, etc. This representation inherently excludes the identity-unrelated factors like expressions[1], hairs, glasses, illumination, background, etc. More importantly, since the topology of 3D facial shape is predefined and consistent across different faces, we can easily measure the reconstruction accuracy with distances between the predicted vertices and their ground truths.

However, reconstructing 3D facial shapes from voice is a challenging problem. The exact association between voice and 3D facial shape is still unknown. The information in voices may only encode some key characteristics of the 3D facial shape, and we do not know which they are. Even if this were not the case, which voice features have higher correspondence to reconstruct the 3D face remains an open question. On the other hand, there is no large-scale publicly available dataset with paired voice recordings and accurate 3D facial shapes. Learning or discovering audiovisual associations is impractical by a purely data-driven method, which further complicates the problem. Suppose we directly apply voice-to-vertex regression on a mixed-gender dataset (including male and female samples). In that case, the learned regressor takes a shortcut [113] – disambiguating 3D facial shape by the gender information in voice, as shown on the left of Fig. 5.1(b). Subsequently, we perform two follow-up experiments, where we perform regressions on male and female subsets, respectively. The slight improvement on the mixed-gender dataset immediately disappears. These results clearly show the challenges of learning audiovisual associations beyond gender.

Given these challenges, we take an alternative view on the reconstruction problem. Instead of directly reconstructing the entire 3D facial shapes from voice, we consider its prerequisite: what features in 3D facial shape can be estimated from voice? By explicitly exploring

[1]Note that the 3D facial shapes refer to those of neutral expressions since this thesis focuses on the identity-related associations between voices and faces, rather than the instantaneous expression information.

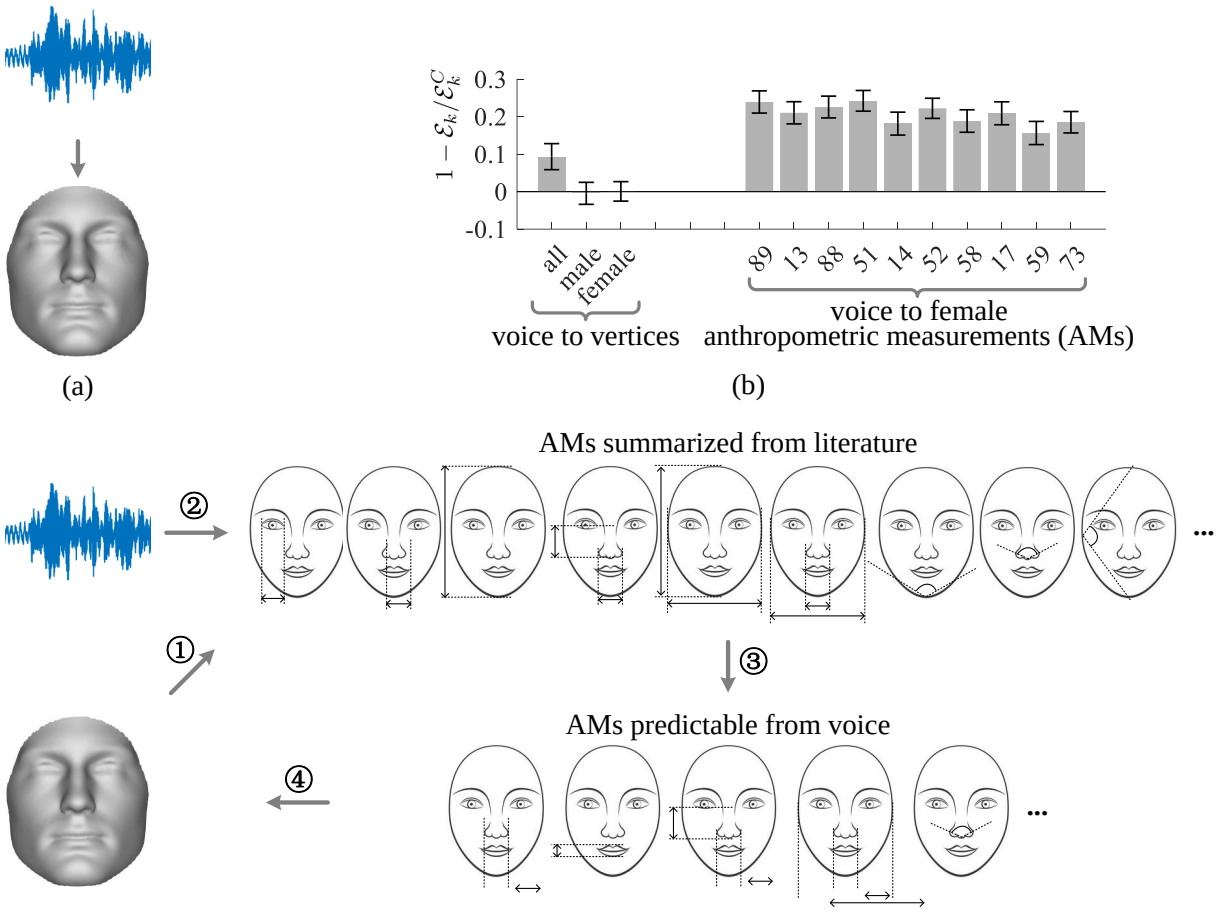


Figure 5.1: (a)(c) Comparison of the voice-to-vertex regression and the proposed SfV. (b) Improvements of voice-to-vertex regression (left) and SfV (right). 0 indicates the chance level performance.

the features that are predictable from voice and can be used in the reconstruction, we can effectively prevent the data-driven model from taking unintended shortcuts and improve the generalization.

In this chapter, we propose an anthropometry-guided framework for reconstructing 3D facial shapes from voice. The method, called shape from voice (SfV), is motivated by the voice production mechanism [75], in which studies have shown the anthropometric measurements (AMs) like the dimensions of nasal cavities [114] or cranium [115, 116] directly

influence the speaker's voices. SfV first identifies which AMs are predictable from voice, and reconstruct the 3D facial shape from those predictable AMs. The predictable AMs act as intermediate variables that bridge the gap between the voice and 3D facial shape. In addition, the AMs are more robust than 3D coordinates since they are invariant to any rigid transformation, which is neither predictable from voice nor of interest to this work.

The proposed SfV is illustrated in Fig. 5.1(c), including four key steps. First, we summarize and compute a number of AMs that are potentially correlated with voice production from the anthropometry literature [117, 118, 119, 120, 121]. Second, we predict the AMs from voice by training estimators with uncertainty learning. The reliability of the estimators can be significantly enhanced with the ability to predict the uncertainty [122]. Third, we select the AMs predictable from voice by hypothesis testing. The null hypothesis is made for each AM and states the AM is unpredictable from voice. We can successfully reject the corresponding null hypothesis if any AM estimation is better than chance on a held-out validation set with statistical significance. We present several predictable AMs on the right of the Fig. 5.1 (b). Last, we reconstruct the 3D facial shapes by a fitting process [76] based on the predictable AMs. This is done by adjusting a set of coefficients in low-dimensional space, such that the differences between the AMs of the generated 3D facial shape and the predicted AMs are minimized. Intuitively, if there are more predictable AMs spanning different locations of a face, the reconstruction can be more indistinguishable.

In the experiments stratified by gender, we discover a number of female AMs that are predictable from voice. We visualize the predictable AMs and find that most of them are located around noses. Moreover, the accuracy of the AM estimation is greatly improved by filtering out a proportion of voice samples with high uncertainties. The further improvements suggest that more AMs are discovered to be predictable from voice, including a few male AMs. With the discovered AMs, we achieve visible improvements in the reconstructed 3D facial shapes, especially the noses of the female speakers.

5.2 The Proposed Framework

Our objective is to reconstruct any speaker's 3D facial shape from their voice recordings. In this section, we first formulate our problem and introduce some necessary notations. Subsequently, we describe the proposed SfV method in detail, followed by discussions on the results.

Our problem is formulated as follows. We are given a set of paired voice recordings and 3D facial shapes $\{(\boldsymbol{v}_1, \boldsymbol{f}_1), (\boldsymbol{v}_2, \boldsymbol{f}_2), ...\}$ from different people, where $\boldsymbol{v}_i$ is a voice recording spoken by the i-th person and $\boldsymbol{f}_i$ is a 3D facial shape scanned from the speaker of $\boldsymbol{v}_i$. Our goal is to reconstruct the 3D facial shape $\boldsymbol{f}$ of any speaker from their voice recording $\boldsymbol{v}$. To do so, we leverage a set of AMs $\{\boldsymbol{m}^{(1)}, \boldsymbol{m}^{(2)}, ..., \boldsymbol{m}^{(K)}\}$ that are computed from $\boldsymbol{f}$, where K is a positive integer and $\boldsymbol{m}^{(k)}$ ($k \in [1, K]$) represents the k-th AM. Accordingly, the overall dataset is denoted as $\mathcal{D} = \{(\boldsymbol{v}_1, \boldsymbol{f}_1, \boldsymbol{m}_1), (\boldsymbol{v}_2, \boldsymbol{f}_2, \boldsymbol{m}_2), ...\}$, where each triplet consists of a voice recording, a scanned 3D facial shape, and K AMs from the same person.

SfV addresses the problem of reconstructing 3D facial shapes from voice by a 4-step pipeline: (i) compute the ground truth AMs from 3D facial shapes; (ii) estimate AMs from voices; (iii) identify which AMs are predictable from voice; (iv) reconstruct 3D facial shapes from the predictable AMs. Unlike regular regression, SfV does not assume the dependency between the input (voice) and output (AM) of the estimators. For this reason, we construct an additional validation set for empirically validating the dependency. Specifically, we split the dataset $\mathcal{D}$ into a training set $\mathcal{D}_t$ for estimator learning, a validation set $\mathcal{D}_{v_1}$ for estimator selection, a validation set $\mathcal{D}_{v_2}$ for AM selection, and an evaluation set (or testing set) $\mathcal{D}_e$ for evaluating the reconstructed 3D facial shapes, among which there is no overlapping.

5.2.1 Training AM estimators

There is a large body of literature on anthropometry. Many AMs of human faces are discovered and have been shown to associate with voice production [117, 118, 119, 120, 121]. We summarize the most commonly used AMs in Fig. 5.2 and Table 5.1. The chosen AMs are

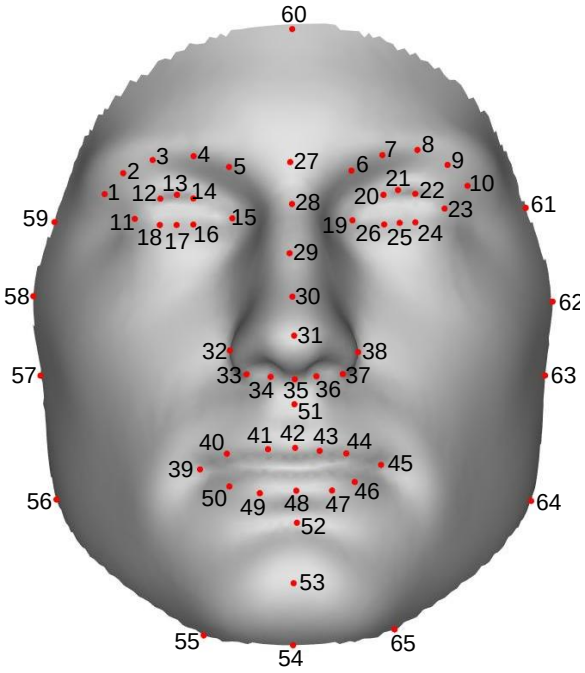


Figure 5.2: The selected landmarks.

Table 5.1: The summarized AMs.

distance					proportion			angle	
11-15	19-23	15-19	11-23	5-6	32-38 / 28-31	33-37 / 28-31	58-62 / 28-31	57-58-59	56-57-58
1-10	3-8	59-61	58-62	57-63	57-63 / 28-31	56-64 / 28-31	55-65 / 28-31	55-56-57	54-55-56
56-64	55-65	32-38	33-37	34-36	39-45 / 28-31	32-38 / 60-54	33-37 / 60-54	65-54-55	64-65-54
49-47	41-43	55-54	54-65	39-45	58-62 / 60-54	57-63 / 60-54	56-64 / 60-54	63-64-65	62-63-64
49-47	40-44	42-48	13-17	21-25	55-65 / 60-54	39-45 / 60-54	59-61 / 58-62	61-62-63	32-30-38
60-28	60-31	60-35	60-51	60-52	58-62 / 57-63	57-63 / 56-64	56-64 / 55-65	32-31-38	28-31-35
60-54	28-31	27-51	31-35	31-52	58-62 / 32-38	57-63 / 32-38	56-64 / 32-38	52-53-54	28-31-32
31-53	31-54	35-51	42-48	52-53	59-61 / 32-38	55-65 / 32-38	39-45 / 32-38	38-31-28	38-51-32
53-54	51-54	55-56	56-57	57-58	60-54 / 28-31	52-53 / 53-54	51-52 / 51-54	64-54-56	30-31-35
58-59	61-62	62-63	63-64	64-65	28-31 / 28-54				

categorized as distance, proportion, and angles. For example, "11-15" denotes the distance between the 11-th and 15-th landmarks, and "32-38 / 28-31" denotes the proportion of two distances. For angle "57-58-59", the arms are the line segments of 57-58 and 59-58, and the vertex is the 58-th landmark. These features are more robust than 3D coordinate representations since the variations caused by spatial misalignment are completely eliminated. As a comparison, 3D coordinates encode the position, orientation, and shape information of the 3D face, of which the first two might introduce unexpected noises.

In SfV, we use convolutional neural networks (CNNs) to estimate the AM from voice. Let

$F_k(\boldsymbol{v};\theta_k) : \boldsymbol{v} \mapsto \mathbb{R}$ be an estimator that maps any voice recording $\boldsymbol{v}$ into the k-th predicted AM[2], where θ_k is the learnable CNN parameters. This is a regular regression problem, so the learning objective for the k-th AM is

$$\theta_k^* = \arg\min_{\theta_k} \frac{1}{|\mathcal{D}_t|} \sum_{(\boldsymbol{v},\boldsymbol{m}^{(k)})\in\mathcal{D}_t} (F_k(\boldsymbol{v};\theta_k) - \boldsymbol{m}^{(k)})^2, \tag{5.1}$$

where $|\mathcal{D}_t|$ is the number of the triplets (voice, face, and AMs) in dataset $\mathcal{D}_t$. By incorporating uncertainty into the estimator learning, the prediction becomes a random variable, rather than a singe value. Following [122], Gaussian distribution is employed to the prediction. Now estimator $F_k(\boldsymbol{v};\theta_k)$ maps $\boldsymbol{v}$ into the mean of the i-th predicted AM. Similarly, we define an uncertainty estimator $G_k(\boldsymbol{v};\phi_k) : \boldsymbol{v} \to \mathbb{R}^+ \cup \{0\}$ that $\boldsymbol{v}$ into the variance of the k-th predicted AM. Again, ϕ_k is the learnable CNN parameters. The predicted AM and its ground truth become $\mathcal{N}(F_k(\boldsymbol{v}), G_k(\boldsymbol{v}))$ and $\mathcal{N}(\boldsymbol{m}^{(k)}, 0)$, respectively [122]. Given two random variables, a more reasonable learning objective is to minimize their KL divergence [122].

$$\{\theta_k^*, \phi_k^*\} = \arg\min_{\theta_k,\phi_k} \frac{1}{|\mathcal{D}_t|} \sum_{(\boldsymbol{v},\boldsymbol{m}^{(k)})\in\mathcal{D}_t} \frac{(F_k(\boldsymbol{v};\theta_k) - \boldsymbol{m}^{(k)})^2}{G_k(\boldsymbol{v};\phi_k)} + \ln G_k(\boldsymbol{v};\phi_k), \tag{5.2}$$

As can be observed from Eqn. 5.2, for a fixed $(F_k(\boldsymbol{v};\theta_k) - \boldsymbol{m}^{(k)})^2$, there is an optimal variance $G_k(\boldsymbol{v};\phi_k) = (F_k(\boldsymbol{v};\theta_k) - \boldsymbol{m}^{(k)})^2$ such that the loss function is minimized. So the uncertainty estimator G_k is learned to produce a small variance if the prediction error is small, and vice versa. On the contrary, a smaller variance indicates that the predicted AM is more likely to yield small prediction error, *i.e.* closed to the ground truth. Therefore, we can choose to trust the predicted AMs when the predicted variances are small, and defer the voice recordings to human experts otherwise. An extreme case is $G_k(\boldsymbol{v}) \equiv 1$, where the uncertainty learning model (Eqn. 5.2) degrades to the regular regression model (Eqn. 5.1).

Fusion. In practice, a long voice recording $\boldsymbol{v}$ is fed into CNNs (F_k and G_k) in the form of multiple short segment inputs $\{\boldsymbol{v}^{(1)}, \boldsymbol{v}^{(2)}, ..., \boldsymbol{v}^{(L)}\}$. So we obtain a sequence of means and variances for the predicted AM. In the training phase, we compute the loss for each

[2]While each estimator is assumed to predict one AM here, they can be easily extended to predict multiple AMs.

segment individually and average them as the training loss. While in the testing phase, the predicted AM and its uncertainty are given by aggregating the predictions from all voice segments. Assuming the short segments from a long voice recording are class-conditionally independent, the formulations of aggregation are

$$\begin{aligned} \hat{\boldsymbol{m}}^{(k)} &= \sum_{l=1}^{L} \frac{\boldsymbol{w}^{(k)}}{G_k(\boldsymbol{v}^{(l)})} \cdot F_k(\boldsymbol{v}^{(l)}), \\ \frac{1}{\boldsymbol{w}^{(k)}} &= \sum_{l=1}^{L} \frac{1}{G_k(\boldsymbol{v}^{(l)})}, \end{aligned} \tag{5.3}$$

where $\hat{\boldsymbol{m}}^{(k)}$ is the aggregated mean and also the predicted k-th AM. However, the aggregated variance $\boldsymbol{w}^{(k)}$ is not used as the uncertainty of the predicted k-th AM. Since the conditional independence assumption does not always hold in cases such as noises, silences, the computed aggregated variance will be biased by the number of voice segments in the long recording. So we calibrate the uncertainty as $\hat{\boldsymbol{w}}^{(k)} = L \cdot \boldsymbol{w}^{(k)}$.

5.2.2 Identifying predictable AMs

We have collected a number of AMs and trained estimators for predicting them. However, only few of the AMs are actually predictable from voice, which we had anticipated while designing the task. To identify those AMs, we use hypothesis testing to them individually. Formally, we can write the null and alternative hypotheses for the k-th AM as

$$\begin{aligned} H_0 &: \text{the AM } \boldsymbol{m}^{(k)} \text{ is NOT predictable from voice} \\ H_1 &: \text{the AM } \boldsymbol{m}^{(k)} \text{ is predictable from voice} \end{aligned}$$

In order to reject H_0, we only need to find an counterexample to show that voice is indeed useful in predicting AM $\boldsymbol{m}^{(k)}$. An effective example is to compare the estimators with and without the voice input. If there exists a learned estimator $F_k(\boldsymbol{v})$ performing better than the chance-level estimator C_k without using voice input and the results are statistically significant, we can successfully reject H_0 and accept H_1. Here the chance-level estimator for the k-th AM is a constant $C_k = \frac{1}{|\mathcal{D}_t|} \sum_{\boldsymbol{m}^{(k)} \in \mathcal{D}_t} \boldsymbol{m}^{(k)}$, which is the mean $\boldsymbol{m}^{(k)}$ of the training

set $\mathcal{D}_t$. So the null and alternative hypotheses can be rewritten as

$$H_0 : \mu(\mathcal{E}_k/\mathcal{E}_k^C) \geq 1$$
$$H_1 : \mu(\mathcal{E}_k/\mathcal{E}_k^C) < 1,$$

where $\mathcal{E}_k$ and $\mathcal{E}_k^C$ are the mean square errors of estimators with and without voice inputs on validation set $\mathcal{D}_{v_2}$, respectively. The formulations of $\mathcal{E}_k$ and $\mathcal{E}_k^C$ are given below.

$$\begin{aligned} \mathcal{E}_k &= \frac{1}{|\mathcal{D}_{v_2}|} \sum_{\boldsymbol{m}^{(k)} \in \mathcal{D}_{v_2}} (\hat{\boldsymbol{m}}^{(k)} - \boldsymbol{m}^{(k)})^2, \\ \mathcal{E}_k^C &= \frac{1}{|\mathcal{D}_{v_2}|} \sum_{\boldsymbol{m}^{(k)} \in \mathcal{D}_{v_2}} (C_k - \boldsymbol{m}^{(k)})^2. \end{aligned} \tag{5.4}$$

To ensure the statistical significance, we perform multiple experiments, where the $\mathcal{D}_t$, $\mathcal{D}_{v_1}$, and $\mathcal{D}_{v_2}$ are obtained by different random splits. Since the true variance of $\mathcal{E}_k/\mathcal{E}_k^C$ is unknown, the type of hypothesis testing is one-sided paired-sample t-test. The upper bound of the confidence interval (CI) is given by

$$\begin{aligned} CI_l &= \mu(\mathcal{E}_k/\mathcal{E}_k^C) - t_{1-\alpha,\nu} \cdot \frac{\sigma(\mathcal{E}_k/\mathcal{E}_k^C)}{\sqrt{N}} \\ CI_u &= \mu(\mathcal{E}_k/\mathcal{E}_k^C) + t_{1-\alpha,\nu} \cdot \frac{\sigma(\mathcal{E}_k/\mathcal{E}_k^C)}{\sqrt{N}} \end{aligned} \tag{5.5}$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ are functions for computing mean and standard deviation, respectively. N is the number of the repeated experiments and here we use $N = 100$. α and $\nu = N - 1$ are the significance level and the degree of freedom, respectively. For the purpose of this section, we adopt the significance level of 5%, then we can immediately read $t_{0.95,N-1}$ from t-distribution table (or t table). Now we can determine whether to reject H_0 by inspecting the computed CI_u. Specifically, $CI_u < 1$ implies that we can successfully reject H_0 and accept H_1, *i.e.* the AM $\boldsymbol{m}^{(k)}$ is predictable from voice. According to the experimental results, the probability that the aforementioned decision is correct is higher than 95%, *i.e.* statistically significant. In contrast, $CI_u \geq 1$ implies that we fail to reject H_0, for the current experimental results are not statistically significant enough. Note that failing to reject H_0 does not imply we accept H_0.

In a nutshell, we define a set of indicators $\boldsymbol{z} = \{z^{(1)}, z^{(2)}, ...z^{(K)}\}$. $\boldsymbol{z}^{(k)}$ is 1 if the k-th AM is predictable from voice. Otherwise $\boldsymbol{z}^{(k)}$ is 0, as given in Eqn. 5.6.

$$z^{(k)} = \begin{cases} 1, & \text{if } CI_u < 1. \\ 0, & \text{otherwise.} \end{cases} \tag{5.6}$$

Discussion. We emphasize that it is necessary to compute $\mathcal{E}_k^C$ and $\mathcal{E}_k$ on $\mathcal{D}_{v_2}$, rather than $\mathcal{D}_t$ or $\mathcal{D}_{v_1}$. This is because our estimators are trained on $\mathcal{D}_t$ and selected by the errors on $\mathcal{D}_{v_1}$, we can easily get significantly lower $\mathcal{E}_k$ than $\mathcal{E}_k^C$ on these datasets. In this case, type I error (false positive) happens, where an AM is actually not predictable from voice, but we reject H_0 and accept H_1. On the contrary, if we evaluate the estimator on $\mathcal{D}_{v_2}$, the type I error is less likely to happen. Type II error (false negative) happens when an AM is actually predictable from voice, but we fail to reject H_0. It is very likely to happen due to the small scale of the dataset, noises in data capturing, imperfect learning model, or even the improper hyperparameters, which indicates there remains considerable room for future exploration.

5.2.3 Reconstructing 3D Facial Shape

To reconstruct the 3D facial shape, we need to predict AMs of the voice recordings in $\mathcal{D}_e$ first. Each predicted AM is given by the ensemble of the trained estimators from the repeated experiments in 5.2.2.

Subsequently, we generate the 3D facial shapes based on the predicted AMs by an optimization-based method. To do so, we first project the 3D facial shapes into a low-dimensional linear space [76]. By adjusting the coefficients in low-dimensional space, we obtain different re-projected 3D facial shapes. The learning objective is to find a set of coefficients, such that the differences between the AMs of the re-projected 3D facial shape and the predicted AMs are minimized. Specifically, we construct a big matrix $B = [\boldsymbol{b}_1, \boldsymbol{b}_2, ...] \in \mathbb{R}^{3T \times |\mathcal{D}_t|}$ where each column $\boldsymbol{b}_i \in \mathbb{R}^{3T \times 1}$ is a long vector obtained by flattening a 3D facial shape $\boldsymbol{f}_i \in \mathbb{R}^{T \times 3}$. T is the number of vertices on 3D faces. Since $3T \gg |\mathcal{D}_t|$, we compute

the projection matrix $P \in \mathbb{R}^{3T\times d}(d \gg 3T)$ using eigenfaces [123] on B. Now any flattened 3D facial shape $\boldsymbol{b}$ can be approximated by re-projecting a low-dimensional vector $\boldsymbol{\beta} \in \mathbb{R}^{d\times 1}$ in the form of $P\boldsymbol{\beta}$.

We define the computation of AM as $Q_k(\boldsymbol{b}) : \boldsymbol{b} \mapsto \mathbb{R}$, which maps any flattened 3D facial shape $\boldsymbol{b}$ into the k-th AM of $\boldsymbol{b}$. Since $Q_k(\cdot)$ computes a distance, a proportion, or an angle of the 3D facial shape, it is a differentiable function. The optimizing objective is given below.

$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}} \ \lambda\|\boldsymbol{\beta}\|_2^2 + \sum_{k=1}^{K}(Q_k(P\boldsymbol{\beta}) - \hat{\boldsymbol{m}}^{(k)})^2 \cdot \boldsymbol{z}^{(k)} \tag{5.7}$$

where λ is the loss weight balancing two terms. The reconstructed 3D facial shape is given by $\hat{\boldsymbol{b}} = P\boldsymbol{\beta}^*$. The pipeline of SfV is summarized in Algorithm 2.

Algorithm 2 The experiment pipeline of SfV

randomly split the audiovisual dataset $\mathcal{D}$ into $\mathcal{D}_t \cup \mathcal{D}_{v_1} \cup \mathcal{D}_{v_2}$, and $\mathcal{D}_e$.
for $k \in [1, 2, ..., K]$ **do**
  repeat
    randomly split $\mathcal{D}_t \cup \mathcal{D}_{v_1} \cup \mathcal{D}_{v_2}$ into $\mathcal{D}_t$, $\mathcal{D}_{v_1}$, and $\mathcal{D}_{v_2}$
    train and select estimators (F_k and G_k) on $\mathcal{D}_t$ and $\mathcal{D}_{v_1}$, respectively (Eqn. 5.2)
    compute $\mathcal{E}_k^C$ and $\mathcal{E}_k$ on $\mathcal{D}_{v_2}$ (Eqn. 5.3 and Eqn. 5.4)
  until $\mu(\mathcal{E}_k/\mathcal{E}_k^C)$ is converged
  compute the upper bound of confidence interval CI_u (Eqn. 5.5)
  compute the indicator $\boldsymbol{z}$ (Eqn. 5.6).
end for
predict AMs of the voice recordings in $\mathcal{D}_e$. Each prediction is given by the ensemble of the trained estimators from the repeated experiments (Eqn. 5.3).
reconstruct 3D facial shapes with the predicted AMs (Eqn. 5.7)

5.3 Experiments

Dataset. We perform experiments on an audiovisual dataset $\mathcal{D}$ collected by researchers from Penn State University. The dataset consists of paired voice recordings and scanned 3D facial shapes from 1,026 people, with 364 males and 662 females. To prevent the estimation models from taking the gender shortcuts, we split the dataset $\mathcal{D}$ by gender, and experiments are individually performed on male and female subsets. For each subset, we adopt 7/1/1/1

Table 5.2: CNN architectures for F_k and G_k: details.

	voice feature	backbone				head
F_k	Spectrogram() Melscale() Log()	$(3,64)_{/2,1}$ $\begin{bmatrix}(3,64)_{/1,1}\\(3,64)_{/1,1}\end{bmatrix}$	$(3,96)_{/2,1}$ $\begin{bmatrix}(3,96)_{/1,1}\\(3,96)_{/1,1}\end{bmatrix}$	$(3,144)_{/2,1}$ $\begin{bmatrix}(3,144)_{/1,1}\\(3,144)_{/1,1}\end{bmatrix}$	$(3,216)_{/2,1}$ $\begin{bmatrix}(3,216)_{/1,1}\\(3,216)_{/1,1}\end{bmatrix}$	$(64,1)_{(1,0)}$
G_k						$(64,1)_{(1,0)}$ Exp()

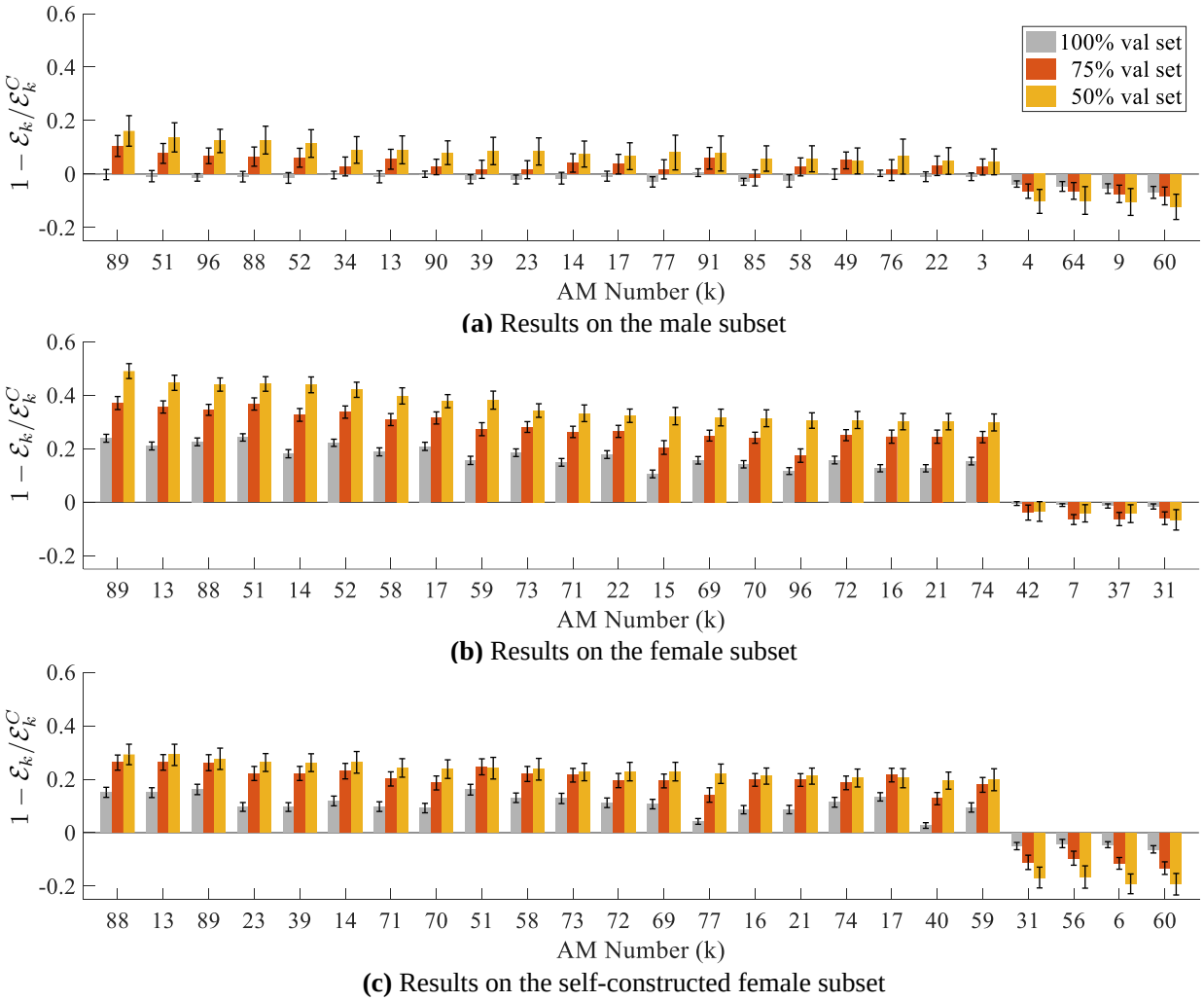
splitting for $\mathcal{D}_t/\mathcal{D}_{v_1}/\mathcal{D}_{v_2}/\mathcal{D}_e$. In training, the voice recordings are randomly trimmed to segments of 6 to 8 seconds, while we use the entire recordings in testing. The ground truth AMs are normalized to zero mean and unit variance. For voice features, we extract 64-dimensional log Mel-spectrograms using an analysis window of 25ms, with the hop of 10ms between frames. We perform mean and variance normalization of each Mel-frequency bin.

Training. The CNN architectures are given in Table 5.2. F_k and G_k share the backbone's learnable parameters but have individual parameters for their heads. Moreover, we adopt *exp* activation for the last layer of G_k to ensure the non-negative output. The numbers within the parentheses represent the size and number of filters, while the subscripts represent the stride and padding. So, for example, $(3,64)_{/2,1}$ denotes a 1D convolutional layer with 64 filters of size 3, where the stride and padding are 2 and 1, respectively. Modules in brackets are equipped with shortcut connections. For the variance head, we add an exponential activation to the last layer of G_k for non-negative positive output.

We follow the typical settings of stochastic gradient descent (SGD) for optimization. Minibatch size is 64. The momentum, learning rate, and weight decay values are 0.9, 0.1, and 0.0005, respectively. The training is completed at 5k iterations. To ensure statistical significance, we perform $N = 100$ repeated experiments to compute the CI_u.

5.3.1 The AM Prediction and Selection

For AM prediction, the estimation models are trained on $\mathcal{D}_t$ and selected based on their performance on $\mathcal{D}_{v_1}$ (hyperparameter tuning). For AM selection, the predictable AMs are selected based on the upper bound of the CI (CI_u) on $\mathcal{D}_{v_2}$. The performance can be evaluated



(a) Results on the male subset

(b) Results on the female subset

(c) Results on the self-constructed female subset

Figure 5.3: The normalized errors and CIs of 24 AMs on (a) male subset, (b) female subset, and (c) a smaller female subset.

by the mean error of each AM and its CI.

The Fig. 5.3 shows the results, including 20 AMs with highest $1 - CI_u$ and 4 AMs with lowest $1 - CI_u$. The gray bars are the results on the entire validation set $\mathcal{D}_{v_2}$, while the red and yellow ones are the results of 75% and 50% voice samples with lowest uncertainty $\hat{\boldsymbol{w}}$ in $\mathcal{D}_{v_2}$, respectively. The self-constructed female subset has the same size as the male subset. Higher $1 - CI_u$ indicates better results and the normalized error of 0 indicates the chance-level performance. As suggested by our hypothesis testing formulation, the AMs with $1 - CI_u > 0$ are considered predictable from voice. In this sense, we have discovered a number of predictable female AMs (see the gray bars and their CIs in Fig. 5.3(b)). By

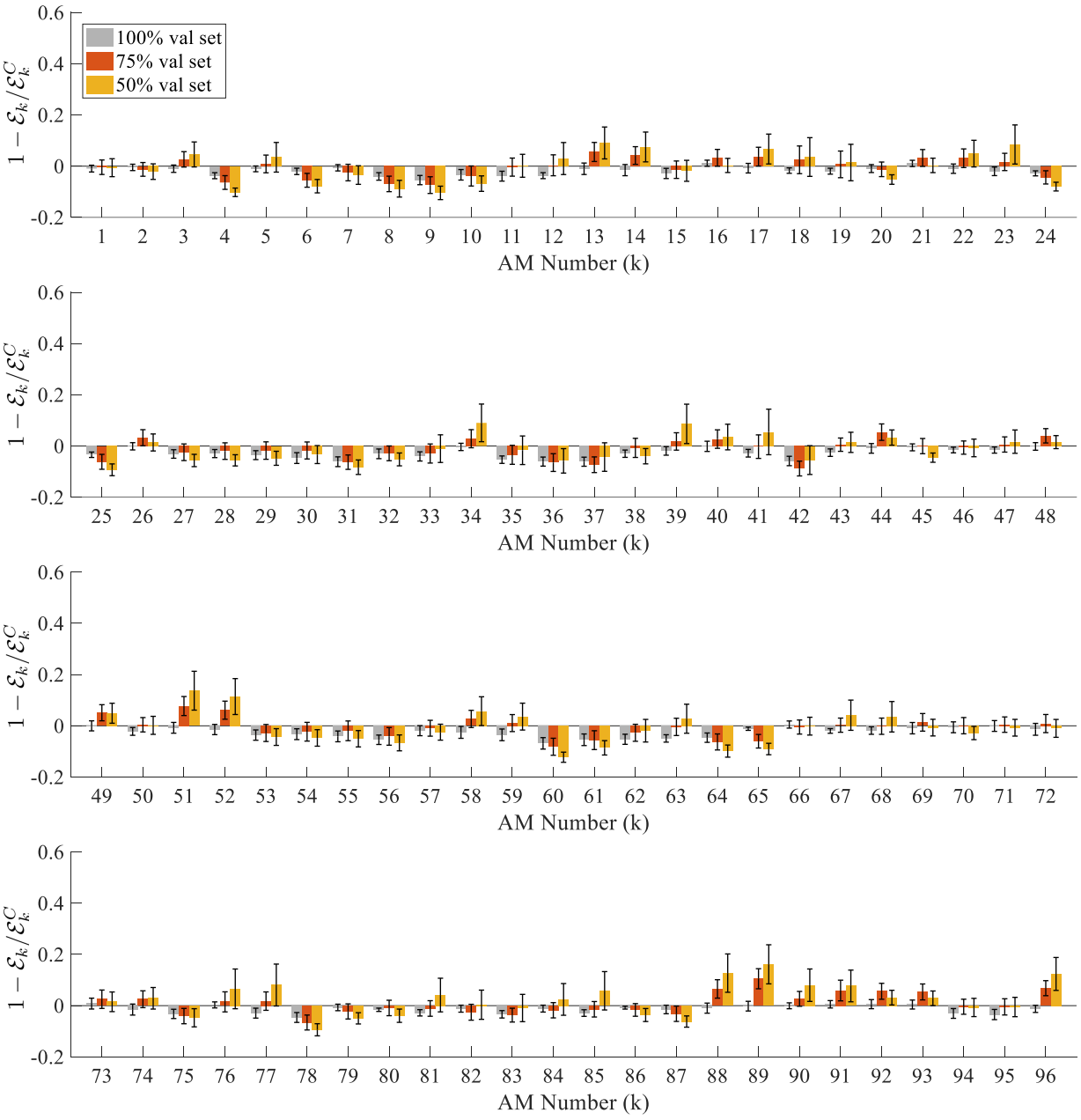


Figure 5.4: The normalized errors and CIs of 96 AMs on the male subset.

filtering out the voice samples with high uncertainties, we achieve even higher $1 - CI_u$ (see the red and yellow bars and their CIs). The improved performance indicates that more AMs are discovered as predictable from voice, including a few male AMs. The complete results of all AMs are given in Fig. 5.4 and 5.5. The results empirically demonstrate that the information of 3D facial shape is indeed encoded in the voices and can be discovered by the proposed SfV.

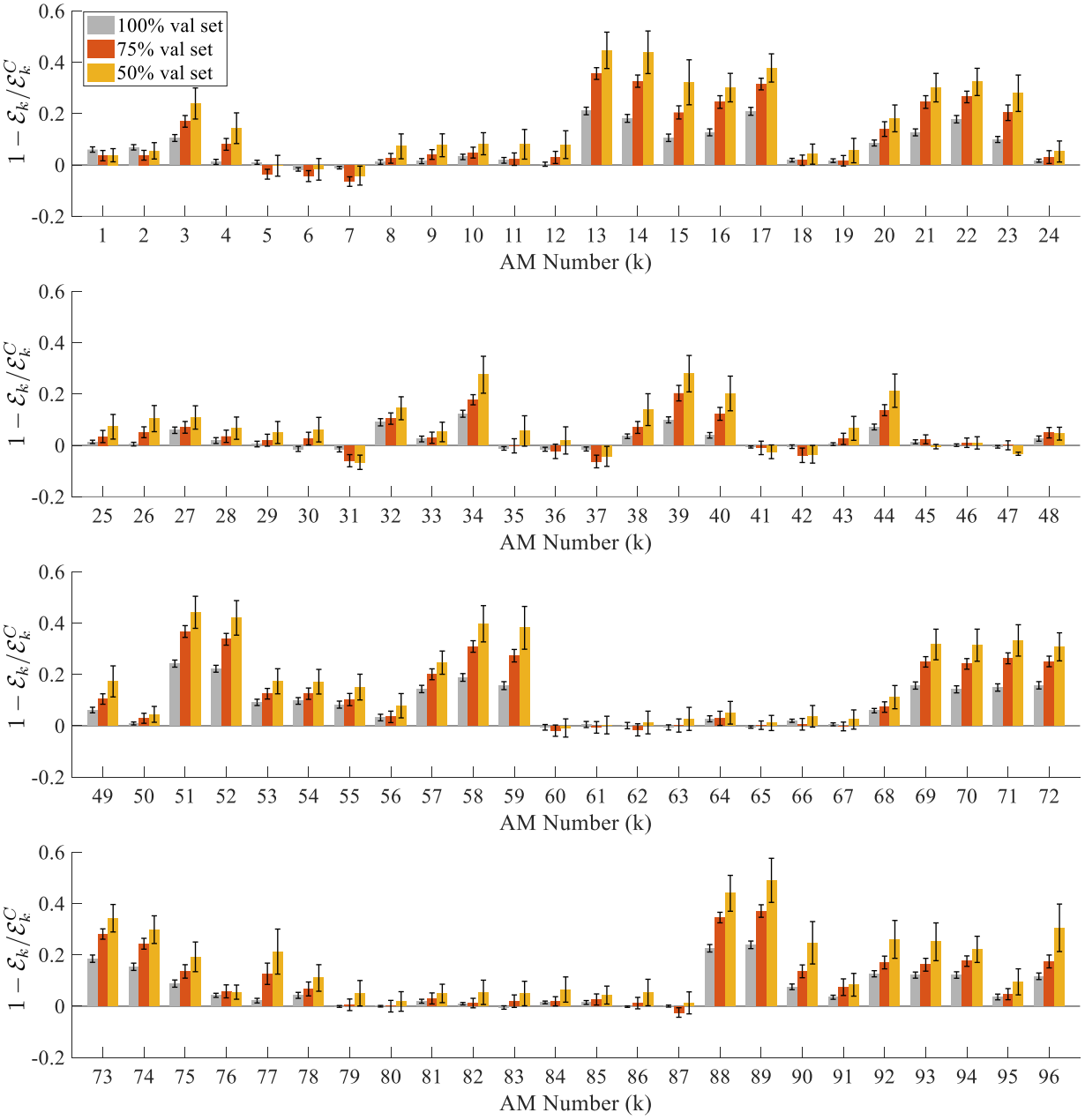


Figure 5.5: The normalized errors and CIs of 96 AMs on the female subset.

To intuitively locate the predictable AMs on the 3D face, we visualize them in Fig. 5.6. We clearly observe that most of the predictable AMs are around the nose and mouth, and many of them are shared between male and female subsets. This is consistent with the fact that the nose and mouth shapes affect pronunciation.

We also notice that the performance of the female dataset is much better than that of the male dataset. To investigate whether the improvements come from the larger data scale

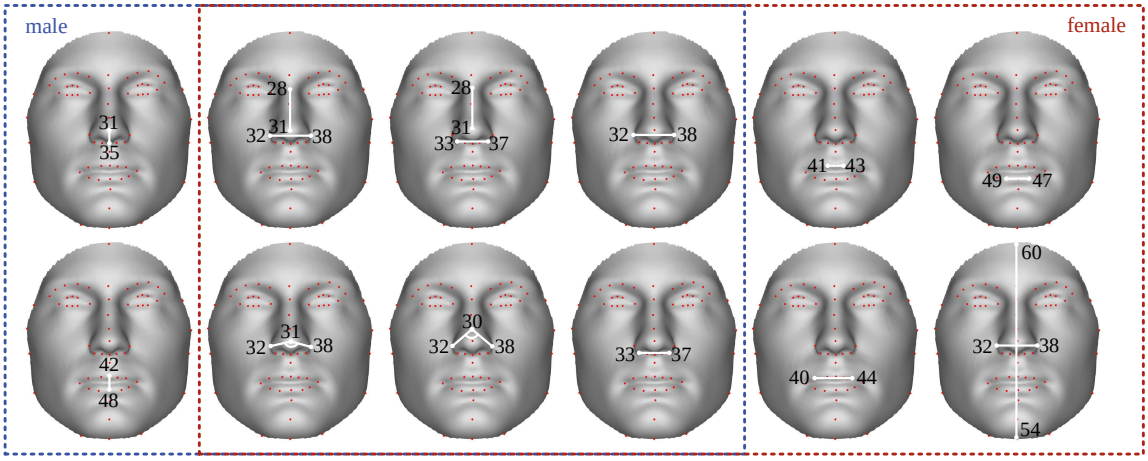


Figure 5.6: Visualization of the predictable AMs. Blue box: male, Red box: female.

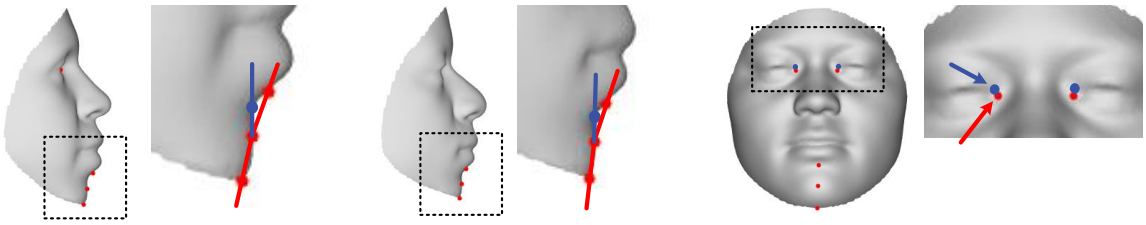


Figure 5.7: (a)(c)(e) are three samples with incorrect landmark labeling. (b)(d)(f) are the zoom-in views of (a)(c)(e), respectively. Red: the noisy label. Blue: the correct label.

(364 males *v.s.* 662 females), we perform another set of repeated experiments on a self-constructed female dataset, which has the same size as the male subset, *i.e. 362 females*. Surprisingly, the results on the new dataset are still better than those on the male subset, as shown in Fig. 5.3(c). This is possibly because the female subjects have higher nasalance scores on the nasal sentences [124] among other things, which provides useful information for predicting the AMs around the nose. Here we note that our experiments have revealed that measurements around the nose are highly correlated to voice. More investigations are left for future work.

On the other hand, some AMs have not been shown to be predictable from voice. This observation suggests that voices may only associate with a few specific regions of the 3D facial shape, like the nose and mouth. For the AMs with higher errors than chance-level, we do not claim they are not predictable from voice. Instead, we fail to demonstrate their

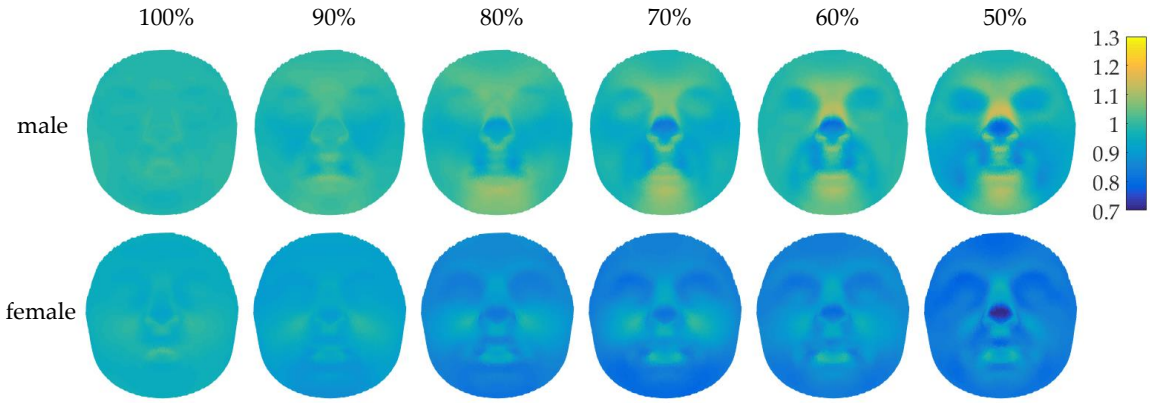


Figure 5.8: Error maps of the reconstructed 3D facial shapes for the male and female subsets. From left to right: the error maps corresponding to 100% (*i.e.* the entire test set) to 50% of the test set.

predictability based on our current empirical results. The possible reasons include imperfect modeling, limited data, data noise, etc. Here we illustrate the noises of landmark labeling in 3D facial shapes in Fig. 5.7. These noises may be detrimental to model learning and generalization.

5.3.2 The Reconstruction of 3D Facial Shape

In the last subsection, we have discovered a number of predictable AMs, from which we choose 10 AMs with the highest $1 - CI_u$ for the subsequent reconstructions on male and female subsets.

To evaluate the performance, we compute the per-vertex errors between the reconstructed 3D facial shape and their ground truths. We also filter out a portion of voice samples with the highest uncertainties and evaluate the errors in the remaining data. The filter out rate is from 0% to 50%, as shown from left to right in Fig. 5.8.

Unsurprisingly, we achieve the lowest errors around the nose region for male and female subsets, consistent with the AM estimations. Moreover, the reconstruction errors decrease significantly by filtering out the voice samples with the highest uncertainties. This indicates that the learned uncertainty is effectively associated with the reconstruction quality and allows the system to decide whether to trust the model or not.

5.4 Discussion

We have proposed a shape from voice method, called SfV, for 3D facial shape reconstruction from voice. SfV bridges the gap between voices and faces by discovering the predictable AMs. Specifically, SfV consists of three key steps: (i) predicting AMs with uncertainty learning, (ii) selecting AMs with hypothesis testing, (iii) reconstructing 3D facial shapes with an optimization-based method. We have discovered a number of AMs that are predictable from voice. These AMs are mostly located around noses and mouths, which matches the voice production mechanism very well. We also achieved improved reconstruction on female noses based on the predictable AMs.

Chapter 6

Conclusions and Future Directions

6.1 Conclusions

In this thesis, we have focused on the problem of reconstructing human faces from voices. Specifically, we have presented how we approach this goal in a step-by-step fashion, summarized as follows.

- We have presented a disjoint mapping framework, called DIMNet, for matching voices to 2D face images. DIMNet can be viewed as a nonparametric approach for generating faces from voice. It can recognize the associations between voices and faces with high accuracy. Our experiments on several stratified evaluation sets have demonstrated that this approach achieves human-level performance.

- We have proposed a generative approach for reconstructing 2D face images from voice. The reconstructed faces are perceptually plausible and have identity associations with the true speakers. We have also illustrated the use of this approach in a real world application for public education.

- We have proposed a conditional estimation framework, called CEST, for 3D face reconstruction from video. CEST achieves more accurate reconstructions than state of

art techniques, and can be used for building audiovisual datasets with paired voices and 3D faces from online video data.

- We have presented a framework for reconstructing 3D facial shapes from voice based on anthropometric measurements. We have discovered many interesting associations between voices and faces. In particular, we observe strong identity associations between voices and the shapes of noses, which are different from the linguistic or affective associations. We hope our work can be insightful in understanding the associations between voices and faces and encourage more progress in this research topic.

6.2 Future Directions

The performance of face reconstruction from voice can be further improved from many perspectives. We summarize some of them below.

6.2.1 Data

More data. Reconstructing faces from voice is still at a very early stage, as illustrated in the Fig. 6.1. The largest dataset that we have used is still limited to only several hundred speakers. We believe that the performance of our proposed approaches can be significantly improved if we scale the dataset to several thousand speakers (or more). In the meantime, it is also important to balance the datasets by including sufficient samples from different demographic groups. This would be useful in the performance of stratified experiments and investigations of fairness issues.

Completed facial topology. The current topology of the 3D face does not include the tongue and neck. It would be interesting to include these as well, since they are highly important articulators. Discovering predictable measurements on them is expected to improve the quality and accuracy of the reconstruction. This will obviously reveal more audiovisual

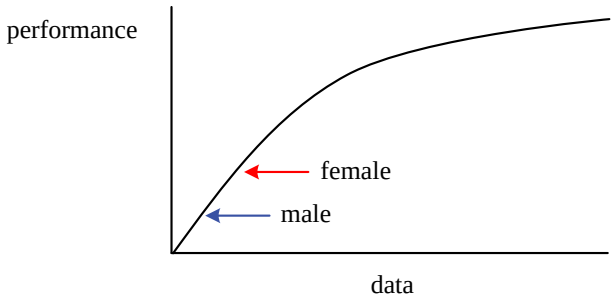


Figure 6.1: Illustration of the current stage of research.

associations between voices and faces. By including more and more articulartors in our framework, the facial topology can be accurately completed.

6.2.2 Method

Robust voice features. In this work, we adopt log Mel-spectrograms as the inputs to CNNs. These voice features yield the current best performance in our experimental settings. However, the Mel-spectrogram only contains the magnitudes of the frequency response, and the phase information is lost. So it is worth exploring the raw waveform as input, and using a voice encoder like SincNet [125] with it.

On the other hand, it is also promising to consider signal processing based acoustic features derived from the temporal domain (signal energy, zero-crossing rate, loudness, etc.), spectral domain (fundamental frequency, harmonics, spectral flux, etc.), composite domain (rhythm, melody, etc.), or even from the perceptual domain (voice qualities such as nasality, raspiness, breathiness, roughness, etc.).

Robust facial features. The anthropometric measurements that we used fall into distance, proportion, and angle categories. These categories are sensitive to the accuracy of 3D shape registration. If we can calibrate the registration or develop features (*e.g.* curve-based or surface-based features) that are less sensitive to noise in it, we should be able to discover more predictable facial features through empirical experiments.

It is also helpful to explore what facial features affect the reconstruction most and conversely, what reconstruction methods can fully make use of such facial features. These

directions can be explored even *without* any voice data. Observations in this context are expected to be very useful for enhancing the quality of reconstruction.

Self-supervised pretraining. In recent years, self-supervised pretraining has been an emerging topic and has demonstrated significant performance improvements on many speech and vision problems, especially those tasks with limited data. We can adopt a pretrained model in AM estimation to improve the model generalization and discover more predictable AMs.

6.2.3 Evaluation

Evaluation metrics. There are currently no objective measures for evaluating the accuracy of 2D face reconstruction. In this thesis, we have focused on the problem of evaluating 3D face reconstruction instead. The most commonly used metric for evaluating 3D facial shapes is either based on vertex-to-vertex or vertex-to-mesh comparisons. Such comparative metrics are sensitive to vertex permutation and are often not in agreement with human perception of shape, which is very subjective. For this reason, we can consider alternative representations for 3D facial shapes, *e.g.* implicit function [126], volumetric representation [127], etc. More importantly, it is necessary to develop robust quantitative metrics to evaluate the visual similarity of **3D facial shapes**.

Voice analysis. So far, what aspects of the voice signal are useful for face reconstruction is not fully explored. Future research will explore how the generated face changes with evidences in different durations of the voice signal, and with different articulatory units such as phonemes, syllables, etc. It is desirable to associate the reconstructed faces with specific voice features, since these associations can provide more insights for understanding the connections between voices and faces. The effect of reverberation, different kinds of noises and channel conditions on the quality of facial reconstruction can also be explored.

6.2.4 Reconstruction of Voices from Face

It is interesting to explore the reverse problem: that of generating voices from facial images, or face-based voice conversion [128]. We believe that more identity associations between voices and faces can be discovered by studying the problem from this perspective. The face-to-voice pipeline can also be incorporated into the existing voice-to-face framework, enabling cycle consistency for paired learning.

Bibliography

[1] P. Belin, S. Fecteau, and C. Bedard, "Thinking the voice: neural correlates of voice perception," *Trends in cognitive sciences*, vol. 8, no. 3, pp. 129–135, 2004. ix, 3, 9, 29

[2] F. Wickelmaier, "An introduction to mds," *Sound Quality Research Unit, Aalborg University, Denmark*, vol. 46, no. 5, 2003. ix, 27, 28

[3] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," *Proc. Interspeech 2017*, pp. 2616–2620, 2017. x, 9, 17, 55, 61

[4] A. Tewari, M. Zollhöfer, P. Garrido, F. Bernard, H. Kim, P. Pérez, and C. 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*, 2018, pp. 2549–2559. x, 44, 45, 50, 63

[5] A. Tewari, M. Zollhofer, H. Kim, P. Garrido, F. Bernard, P. Perez, and C. Theobalt, "Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 1274–1283. x, 43, 44, 46, 48, 50, 55, 59, 64

[6] E. Richardson, M. Sela, R. Or-El, and R. Kimmel, "Learning detailed face reconstruction from a single image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1259–1268. x, 44, 46, 63, 64

[7] A. Tewari, F. Bernard, P. Garrido, G. Bharaj, M. Elgharib, H.-P. Seidel, P. Pérez, M. Zollhofer, and C. Theobalt, "Fml: Face model learning from videos," in *Proceed-*

ings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 10 812–10 822. x, 44, 45, 46, 47, 55, 60, 64, 65

[8] J. Booth, E. Antonakos, S. Ploumpis, G. Trigeorgis, Y. Panagakis, and S. Zafeiriou, "3d face morphable models" in-the-wild"," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 5464–5473. x, 63, 64

[9] M. Sela, E. Richardson, and R. Kimmel, "Unrestricted facial geometry reconstruction using image-to-image translation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1576–1585. x, 63, 64

[10] S. Sanyal, T. Bolkart, H. Feng, and M. 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*, 2019, pp. 7763–7772. x, 64

[11] J. Shang, T. Shen, S. Li, L. Zhou, M. Zhen, T. Fang, and L. Quan, "Self-supervised monocular 3d face reconstruction by occlusion-aware multi-view geometry consistency," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*. Springer, 2020, pp. 53–70. x, 47, 64

[12] P. Belin, "Similarities in face and voice cerebral processing," *Visual Cognition*, vol. 25, no. 4-6, pp. 658–665, 2017. 1

[13] W. A. Freiwald and D. Y. Tsao, "Functional compartmentalization and viewpoint generalization within the macaque face-processing system," *Science*, vol. 330, no. 6005, pp. 845–851, 2010. 1

[14] P. Belin and R. J. Zatorre, "Adaptation to speaker's voice in right anterior temporal lobe," *Neuroreport*, vol. 14, no. 16, pp. 2105–2109, 2003. 1

[15] H. L. Bear and R. Harvey, "Phoneme-to-viseme mappings: the good, the bad, and the ugly," *Speech Communication*, vol. 95, pp. 40–67, 2017. 1

[16] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, 1976. 1

[17] P. Belin, P. E. Bestelmeyer, M. Latinus, and R. Watson, "Understanding voice perception," *British Journal of Psychology*, vol. 102, no. 4, pp. 711–725, 2011. 1

[18] L. W. Mavica and E. Barenholtz, "Matching voice and face identity from static images." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 39, no. 2, p. 307, 2013. 1, 3

[19] M. Kamachi, H. Hill, K. Lander, and E. Vatikiotis-Bateson, "Putting the face to the voice': Matching identity across modality," *Current Biology*, vol. 13, no. 19, pp. 1709–1714, 2003. 1, 3, 9, 30

[20] A. W. Young, S. Frühholz, and S. R. Schweinberger, "Face and voice perception: Understanding commonalities and differences," *Trends in Cognitive Sciences*, vol. 24, no. 5, pp. 398–410, 2020. 1

[21] H. Zhou, Y. Liu, Z. Liu, P. Luo, and X. Wang, "Talking face generation by adversarially disentangled audio-visual representation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 9299–9306. 1

[22] E. Zakharov, A. Shysheya, E. Burkov, and V. Lempitsky, "Few-shot adversarial learning of realistic neural talking head models," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9459–9468. 1

[23] A. Richard, M. Zollhöfer, Y. Wen, F. De la Torre, and Y. Sheikh, "Meshtalk: 3d face animation from speech using cross-modality disentanglement," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1173–1182. 1

[24] L. Chen, R. K. Maddox, Z. Duan, and C. Xu, "Hierarchical cross-modal talking face generation with dynamic pixel-wise loss," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7832–7841. 1

[25] D. Das, S. Biswas, S. Sinha, and B. Bhowmick, "Speech-driven facial animation using cascaded gans for learning of motion and texture," in *European Conference on Computer Vision*. Springer, 2020, pp. 408–424. 1

[26] J. Thies, M. Elgharib, A. Tewari, C. Theobalt, and M. Nießner, "Neural voice puppetry: Audio-driven facial reenactment," in *European conference on computer vision*. Springer, 2020, pp. 716–731. 1

[27] P. Mermelstein, "Determination of the vocal-tract shape from measured formant frequencies," *The Journal of the Acoustical Society of America*, vol. 41, no. 5, pp. 1283–1294, 1967. 2, 29

[28] H. Teager and S. Teager, "Evidence for nonlinear sound production mechanisms in the vocal tract," in *Speech production and speech modelling*. Springer, 1990, pp. 241–261. 2, 29

[29] M. Gerek, A. Durmaz, U. Aydin, H. Birkent, Y. Hidir, and F. Tosun, "Relationship between nasal valve changes and nasalance of the voice," *Otolaryngology–Head and Neck Surgery*, vol. 147, no. 1, pp. 98–101, 2012. 2

[30] S. Marrakchi and H. I. Maibach, "Biophysical parameters of skin: map of human face, regional, and age-related differences," *Contact dermatitis*, vol. 57, no. 1, pp. 28–34, 2007. 2

[31] E. T. Stathopoulos, J. E. Huber, and J. E. Sussman, "Changes in acoustic characteristics of the voice across the life span: Measures from individuals 4–93 years of age," *Journal of Speech, Language, and Hearing Research*, 2011. 3

[32] R. Thornhill and A. P. Møller, "Developmental stability, disease and medicine," *Biological Reviews*, vol. 72, no. 4, pp. 497–548, 1997. 3

[33] H. Hollien and G. P. Moore, "Measurements of the vocal folds during changes in pitch," *Journal of Speech and Hearing Research*, vol. 3, no. 2, pp. 157–165, 1960. 3

[34] A. W. Ellis, "Neuro-cognitive processing of faces and voices," in *Handbook of research on face processing*. Elsevier, 1989, pp. 207–215. 3, 9, 29

[35] P. Belin, P. Bestelmeyer, M. Latinus, and R. Watson, "Understanding voice perception," *British Journal of Psychology*, vol. 108, pp. 711–725, 2011. 3, 9, 29

[36] H. A. McAllister, R. H. Dale, N. J. Bregman, A. McCabe, and C. R. Cotton, "When eyewitnesses are also earwitnesses: Effects on visual and voice identifications," *Basic and Applied Social Psychology*, vol. 14, no. 2, pp. 161–170, 1993. 3, 9, 30

[37] S. R. Schweinberger, D. Robertson, and J. M. Kaufmann, "Hearing facial identities," *Quarterly Journal of Experimental Psychology*, vol. 60, no. 10, pp. 1446–1456, 2007. 3, 9, 30

[38] S. R. Schweinberger, N. Kloth, and D. M. Robertson, "Hearing facial identities: Brain correlates of face–voice integration in person identification," *Cortex*, vol. 47, no. 9, pp. 1026–1037, 2011. 3, 9, 30

[39] A. Nagrani, S. Albanie, and A. Zisserman, "Seeing voices and hearing faces: Cross-modal biometric matching," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8427–8436. 4, 10, 11, 12, 17, 22, 23, 31, 35

[40] ——, "Learnable pins: Cross-modal embeddings for person identity," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 71–88. 4, 10, 11, 12, 17, 26

[41] R. Wang, X. Liu, Y.-m. Cheung, K. Cheng, N. Wang, and W. Fan, "Learning discriminative joint embeddings for efficient face and voice association," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, pp. 1881–1884. 4

[42] C. Kim, H. V. Shin, T.-H. Oh, A. Kaspar, M. Elgharib, and W. Matusik, "On learning associations of faces and voices," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 276–292. 4, 10, 11, 12, 55

[43] T.-H. Oh, T. Dekel, C. Kim, I. Mosseri, W. T. Freeman, M. Rubinstein, and W. Matusik, "Speech2face: Learning the face behind a voice," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7539–7548. 4

[44] A. C. Duarte, F. Roldan, M. Tubau, J. Escur, S. Pascual, A. Salvador, E. Mohedano, K. McGuinness, J. Torres, and X. Giro-i Nieto, "Wav2pix: Speech-conditioned face generation using generative adversarial networks." in *ICASSP*, 2019, pp. 8633–8637. 4

[45] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680. 4, 30

[46] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014. 4, 30

[47] H.-S. Choi, C. Park, and K. Lee, "From inference to generation: End-to-end fully self-supervised generation of human face from speech," in *International Conference on Learning Representations*, 2019. 4

[48] H. Verma, P. Solanki, and M. James, "Acoustical and perceptual voice profiling of children with recurrent respiratory papillomatosis," *Journal of Voice*, vol. 30, no. 5, pp. 600–605, 2016. 5

[49] N. Schilling and A. Marsters, "Unmasking identity: speaker profiling for forensic linguistic purposes," *Annual Review of Applied Linguistics*, vol. 35, pp. 195–214, 2015. 5

[50] Q. Jin, A. R. Toth, T. Schultz, and A. W. Black, "Speaker de-identification via voice transformation," in *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*. IEEE, 2009, pp. 529–533. 8

[51] E. M. Newton, L. Sweeney, and B. Malin, "Preserving privacy by de-identifying face images," *IEEE transactions on Knowledge and Data Engineering*, vol. 17, no. 2, pp. 232–243, 2005. 8

[52] Y. Wen, M. A. Ismail, B. Raj, and R. Singh, "Optimal strategies for matching and retrieval problems by comparing covariates," *arXiv preprint arXiv:1807.04834*, 2018. 10, 21, 23, 24

[53] C. Lippert, R. Sabatini, M. C. Maher, E. Y. Kang, S. Lee, O. Arikan, A. Harley, A. Bernal, P. Garst, V. Lavrenko *et al.*, "Identification of individuals by trait prediction using whole-genome sequencing data," *Proceedings of the National Academy of Sciences*, vol. 114, no. 38, pp. 10 166–10 171, 2017. 13

[54] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proceedings of the British Machine Vision Conference 2015*. BMVA Press, 2015, pp. 41.1–41.12. 17, 35

[55] J. S. Chung and A. Zisserman, "Out of time: automated lip sync in the wild," in *Asian Conference on Computer Vision*. Springer, 2016, pp. 251–263. 17

[56] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016. 18

[57] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015. 18, 36, 56

[58] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105. 18, 36

[59] C. D. Manning, P. Raghavan, and H. Schütze, "Evaluation in information retrieval," in *Introduction to Information Retrieval*. Cambridge University Press, 2008, ch. 8, pp. 159–160. 20

[60] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008. 27

[61] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *ECCV*, 2016. 28

[62] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *CVPR*, 2017. 28

[63] M. Bahari, M. McLaren, D. van Leeuwen *et al.*, "Age estimation from telephone speech using i-vectors," *Proceedings of Interspeech 2012*, 2012. 29

[64] M. Kotti and C. Kotropoulos, "Gender classification in two emotional speech databases," in *2008 19th International Conference on Pattern Recognition*. IEEE, 2008, pp. 1–4. 29

[65] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and simile classifiers for face verification," in *2009 IEEE 12th International Conference on Computer Vision*. IEEE, 2009, pp. 365–372. 29

[66] R. Singh, "Reconstruction of the human persona in 3d, and its reverse," in *Profiling Humans from their Voice*. springer nature Press, 2020, ch. 10. 30

[67] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," *arXiv preprint arXiv:1605.05396*, 2016. 30

[68] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5907–5915. 30

[69] X. Xia, R. Togneri, F. Sohel, and D. Huang, "Auxiliary classifier generative adversarial network with soft labels in imbalanced acoustic event detection," *IEEE Transactions on Multimedia*, 2018. 30

[70] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks)," in *International Conference on Computer Vision*, 2017. 35, 54, 55

[71] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015. 36

[72] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014. 36, 57

[73] Y. Wen, M. Al Ismail, W. Liu, B. Raj, and R. Singh, "Disjoint mapping network for cross-modal matching of voices and faces," in *International Conference on Learning Representations*, 2018. 41

[74] R. Singh, *Profiling humans from their voice*. Springer, 2019. 41

[75] R. Singh, B. Raj, and D. Gencaga, "Forensic anthropometry from voice: an articulatory-phonetic approach," in *2016 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE, 2016, pp. 1375–1380. 41, 69

[76] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 1999, pp. 187–194. 42, 47, 52, 68, 70, 76

[77] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou, "Facewarehouse: A 3d facial expression database for visual computing," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 3, pp. 413–425, 2013. 42

[78] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, and P. Liu, "A high-resolution spontaneous 3d dynamic facial expression database," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, 2013, pp. 1–6. 42

[79] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3d facial expression database for facial behavior research," in *7th international conference on automatic face and gesture recognition (FGR06)*. IEEE, 2006, pp. 211–216. 42

[80] A. Lattas, S. Moschoglou, B. Gecer, S. Ploumpis, V. Triantafyllou, A. Ghosh, and S. Zafeiriou, "Avatarme: Realistically renderable 3d facial reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 760–769. 42, 50

[81] S. Sengupta, A. Kanazawa, C. D. Castillo, and D. W. Jacobs, "Sfsnet: Learning shape, reflectance and illuminance of facesin the wild'," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6296–6305. 42, 44, 46

[82] L. Tran and X. Liu, "Nonlinear 3d face morphable model," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7346–7355. 43, 44, 47, 48, 50

[83] ——, "On learning 3d face morphable model from in-the-wild images," *IEEE transactions on pattern analysis and machine intelligence*, 2019. 44, 45, 46, 47, 50, 52, 60

[84] S. Wu, C. Rupprecht, and A. Vedaldi, "Unsupervised learning of probably symmetric deformable 3d objects from images in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1–10. 44, 45, 54

[85] I. Kemelmacher-Shlizerman and S. M. Seitz, "Face reconstruction in the wild," in *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 1746–1753. 44, 45, 47

[86] P. Garrido, M. Zollhöfer, D. Casas, L. Valgaerts, K. Varanasi, P. Pérez, and C. Theobalt, "Reconstruction of personalized 3d face rigs from monocular video," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 3, pp. 1–15, 2016. 44, 47

[87] K. Genova, F. Cole, A. Maschinot, A. Sarna, D. Vlasic, and W. T. Freeman, "Unsupervised training for 3d morphable model regression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8377–8386. 44, 52

[88] I. Kemelmacher-Shlizerman and R. Basri, "3d face reconstruction from a single image using a single reference face shape," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 2, pp. 394–405, 2010. 45

[89] F. Shi, H.-T. Wu, X. Tong, and J. Chai, "Automatic acquisition of high-fidelity facial performances using monocular videos," *ACM Transactions on Graphics (TOG)*, vol. 33, no. 6, pp. 1–13, 2014. 45, 47

[90] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3d solution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 146–155. 46, 50, 55, 62

[91] A. D. Bagdanov, A. Del Bimbo, and I. Masi, "The florence 2d/3d hybrid face dataset," in *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding*, 2011, pp. 79–80. 46

[92] H. Kim, M. Zollhöfer, A. Tewari, J. Thies, C. Richardt, and C. Theobalt, "Inversefacenet: Deep monocular inverse face rendering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4625–4634. 46

[93] Y. Zhou, J. Deng, I. Kotsia, and S. Zafeiriou, "Dense 3d face decoding over 2500fps: Joint texture & shape convolutional mesh decoders," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1097–1106. 47, 48, 50

[94] R. Garg, A. Roussos, and L. Agapito, "Dense variational reconstruction of non-rigid surfaces from monocular video," in *Proceedings of the IEEE Conference on computer vision and pattern recognition*, 2013, pp. 1272–1279. 47

[95] R. Szeliski, *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010. 49, 51, 52

[96] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. 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*. Ieee, 2009, pp. 296–301. 50

[97] B. Gecer, S. Ploumpis, I. Kotsia, and S. Zafeiriou, "Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1155–1164. 50

[98] H. Wei, S. Liang, and Y. Wei, "3d dense face alignment via graph convolution networks," *arXiv preprint arXiv:1904.05562*, 2019. 50

[99] O. Bottema, "On the area of a triangle in barycentric coordinates," *Crux Mathematicorum*, vol. 8, no. 8, pp. 228–231, 1982. 52

[100] R. Basri and D. W. Jacobs, "Lambertian reflectance and linear subspaces," *IEEE transactions on pattern analysis and machine intelligence*, vol. 25, no. 2, pp. 218–233, 2003. 52

[101] R. Ramamoorthi and P. Hanrahan, "A signal-processing framework for inverse rendering," in *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 2001, pp. 117–128. 52

[102] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "Maskgan: Towards diverse and interactive facial image manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5549–5558. 53

[103] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou, "Joint 3d face reconstruction and dense alignment with position map regression network," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 534–551. 55, 62

[104] A. S. Jackson, A. Bulat, V. Argyriou, and G. Tzimiropoulos, "Large pose 3d face reconstruction from a single image via direct volumetric cnn regression," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1031–1039. 55, 62

[105] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241. 56

[106] P. J. Besl and N. D. McKay, "Method for registration of 3-d shapes," in *Sensor fusion IV: control paradigms and data structures*, vol. 1611. International Society for Optics and Photonics, 1992, pp. 586–606. 62

[107] Y. Liu, A. Jourabloo, W. Ren, and X. Liu, "Dense face alignment," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 1619–1628. 62

[108] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2387–2395. 63

[109] P. Garrido, M. Zollhöfer, C. Wu, D. Bradley, P. Pérez, T. Beeler, and C. Theobalt, "Corrective 3d reconstruction of lips from monocular video," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 6, pp. 1–11, 2016. 65

[110] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *ICCV*, 2015. 65

[111] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017. 67

[112] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *Advances in neural information processing systems*, vol. 29, 2016. 67

[113] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut learning in deep neural networks," *Nature Machine Intelligence*, vol. 2, no. 11, pp. 665–673, 2020. 68

[114] T. Vampola, J. Horáček, V. Radolf, J. G. Švec, and A.-M. Laukkanen, "Influence of nasal cavities on voice quality: Computer simulations and experiments," *The Journal of the Acoustical Society of America*, vol. 148, no. 5, pp. 3218–3231, 2020. 69

[115] M. Wyganowska-Świątkowska, I. Kowalkowska, K. Mehr, and M. Dąbrowski, "An anthropometric analysis of the head and face in vocal students," *Folia Phoniatrica et Logopaedica*, vol. 65, no. 3, pp. 136–142, 2013. 69

[116] M. Wyganowska-Świątkowska, I. Kowalkowska, G. Flicińska-Pamfil, M. Dąbrowski, P. Kopczyński, and B. Wiskirska-Woźnica, "Vocal training in an anthropometrical aspect," *Logopedics Phoniatrics Vocology*, vol. 42, no. 4, pp. 178–186, 2017. 69

[117] D. Ghafourzadeh, C. Rahgoshay, S. Fallahdoust, A. Beauchamp, A. Aubame, T. Popa, and E. Paquette, "Part-based 3d face morphable model with anthropometric local control," in *Proceedings of Graphics Interface 2020*, ser. GI 2020. Canadian Human-Computer Communications Society, 2020, pp. 7 – 16. 70, 71

[118] Z. Zhuang, D. Landsittel, S. Benson, R. Roberge, and R. Shaffer, "Facial anthropometric differences among gender, ethnicity, and age groups," *Annals of occupational hygiene*, vol. 54, no. 4, pp. 391–402, 2010. 70, 71

[119] Z. Shan, R. T.-C. Hsung, C. Zhang, J. Ji, W. S. Choi, W. Wang, Y. Yang, M. Gu, and B. S. Khambay, "Anthropometric accuracy of three-dimensional average faces compared to conventional facial measurements," *Scientific Reports*, vol. 11, no. 1, pp. 1–12, 2021. 70, 71

[120] L. G. Farkas, O. G. Eiben, S. Sivkov, B. Tompson, M. J. Katic, and C. R. Forrest, "Anthropometric measurements of the facial framework in adulthood: age-related changes in eight age categories in 600 healthy white north americans of european ancestry from 16 to 90 years of age," *Journal of Craniofacial Surgery*, vol. 15, no. 2, pp. 288–298, 2004. 70, 71

[121] N. Ramanathan and R. Chellappa, "Modeling age progression in young faces," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 1. IEEE, 2006, pp. 387–394. 70, 71

[122] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" *Advances in neural information processing systems*, vol. 30, 2017. 70, 73

[123] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of cognitive neuroscience*, vol. 3, no. 1, pp. 71–86, 1991. 77

[124] E. J. Seaver, R. M. Dalston, H. A. Leeper, and L. E. Adams, "A study of nasometric values for normal nasal resonance," *Journal of Speech, Language, and Hearing Research*, vol. 34, no. 4, pp. 715–721, 1991. 82

[125] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 1021–1028. 87

[126] J. Chibane, T. Alldieck, and G. Pons-Moll, "Implicit functions in feature space for 3d shape reconstruction and completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6970–6981. 88

[127] S. Isler, R. Sabzevari, J. Delmerico, and D. Scaramuzza, "An information gain formulation for active volumetric 3d reconstruction," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 3477–3484. 88

[128] H.-H. Lu, S.-E. Weng, Y.-F. Yen, H.-H. Shuai, and W.-H. Cheng, "Face-based voice conversion: Learning the voice behind a face," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 496–505. 89