UNIVERSITY OF CALIFORNIA

Los Angeles

A Deep Learning Approach to Facial 3D Reconstruction and Super-Resolution Rendering

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Computer Science

by

Tomoyori Iwao

2022

© Copyright by
Tomoyori Iwao
2022

ABSTRACT OF THE THESIS

A Deep Learning Approach to Facial 3D Reconstruction and Super-Resolution Rendering

by

Tomoyori Iwao
Master of Science in Computer Science
University of California, Los Angeles, 2022
Professor Demetri Terzopoulos, Chair

Reconstructing 3D facial shapes is of significant interest in Computer Vision and Computer Graphics. In recent years, deep learning methods have been proposed for cost-effectively acquiring accurate 3D facial shapes. To some extent, these methods can generate accurate 3D facial shapes from a single image. Most of them assume that the resolution of the input images is high enough to reconstruct facial shapes, but this is not always the case, such as in the 3D reconstruction of players in a dynamic sports scene. Due to the large distances between players, one normally cannot acquire a photo showing all the players in focus and some faces could appear blurry. To tackle this problem, we propose a new learning-based architecture that combines a Super-Resolution (SR) network with a 3D Facial Reconstruction (FR) network. First, we input a low-resolution facial photo to the SR network, which generates corresponding SR images. These are input to the FR network, which generates corresponding facial shapes. We generate multiple SR images and corresponding 3D facial shapes because the SR network is a non-deterministic algorithm whose output is not always optimal for 3D reconstruction and rendering. Hence, we choose the best pair considering the noise level of the SR images and the distribution of distances among those facial shapes in order to reconstruct a good-quality 3D facial shape despite the low resolution of the input image.

The thesis of Tomoyori Iwao is approved.

Achuta Kadambi

Chenfanfu Jiang

Demetri Terzopoulos, Committee Chair

University of California, Los Angeles

2022

To my family, friends, mentors, and Canon, Inc.
for all of their support of my studies at UCLA

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to Professor Demetri Terzopoulos. He suggested research directions when I was at a loss as to what to do with myself. He also advised me on how to organize the thesis and edited my draft. I also thank the other members of my thesis committee, Professors Chenfanfu Jiang and Achuta Kadambi, who provided feedback to improve this thesis.

I would like to express my thanks to my family for their encouragement and moral support. I am very grateful that my employer, Japan's Canon, Inc., gave me the opportunity to study at UCLA's Computer Science Department. I thank Professor Shigeo Morishima, who was my thesis advisor at Waseda University, for helping me prepare to study abroad. I am also thankful to Dr. Masaki Nakada, an alumnus of both the research labs of Professors Morishima and Terzopoulos, who helped prepare me for working at UCLA and for living in LA. Finally, I am delighted to acknowledge the UCLA faculty and student colleagues who taught me a great deal in my classes and campus life.

VITA

2012	B.S. (Engineering), Waseda University, Tokyo, Japan
2014	M.S. (Engineering), Waseda University, Tokyo, Japan
2014-present	Engineer, Canon,; Inc., Tokyo, Japan
2021-present	Graduate Researcher, UCLA Computer Graphics and Vision Laboratory, University of California, Los Angeles, USA

CHAPTER 1

Introduction

3D reconstruction is one of the most significant topics in the fields of Computer Vision and Computer Graphics. Reconstructed 3D shapes are widely used in virtual and augmented reality, 3D printing, digital archiving of cultural artifacts, electronic commerce, and many other applications. Searching for "3D reconstruction" on Google Scholar yields nearly 3 million publication results, including more than 400 thousand in the last 5 years.

However, reconstruction accuracy and the cost of capturing 3D shapes continue to challenge researchers. In particular, reconstructing 3D human models from multiple images should be a lucrative business, but it currently requires much time and money. In fact, to acquire photorealistic 3D human models, one must carefully arrange numerous advanced cameras in a studio and have them capture images simultaneously. Nevertheless, the quality of the reconstructed human shapes might not suffice for certain applications. The head and facial areas of the 3D model usually need special treatment because people are especially sensitive to reconstruction deficiencies such as partial collapse of the hair shape, inadequate protrusion of the nose, and missing details in the facial expression. Therefore, skilled 3D modelers usually must be hired to improve the reconstructed shapes, often at the cost of many hours of work. Although automated 3D reconstruction normally reduces the amount of effort required, the equipment and labor costs can still be high.

One of the most interesting uses of 3D human models is in the sports domain, to enable the audience to watch sports scenes free of viewpoint constraints; i.e., the viewpoint is not restricted to any of the broadcast cameras, and sports fans can freely change their views. Ideally, the viewer can observe the scene or any player by changing the viewpoint on their smartphones or tablet devices. For example, they can select overhead views to

see the formation of the sports team, or they can move their viewpoint close to any player. To enable this, dozens or hundreds of cameras are installed in a sports stadium. Next, the entire 3D scene is reconstructed from the images captured by the cameras. Finally, the unconstrained viewpoint video system renders the image from the desired viewpoint. We must use multiple cameras to reconstruct accurate 3D human shapes as described above and, therefore, to acquire accurate 3D models of entire sports scenes is nearly impossible now. A major challenge is that sports stadiums are large spaces and video cameras cannot capture all of the players in perfect focus. Moreover, players would occlude one another even from the points of view of hundreds or thousands of cameras.

Machine-learning-based 3D reconstruction methods have attracted much attention in recent years, and deep learning methods have been shown to reduce the number of images needed to reconstruct 3D shapes from hundreds down to ten or fewer. These neural network methods can also overcome challenges such as the collapse of reconstructed 3D object shape due to reflection anisotropies characteristic of hair and metal. Deep learning methods infer 3D shapes from 2D images by exploiting supervised learning from large datasets of input-output training pairs, in this context, single-view or multi-view input images and associated ground-truth output 3D shapes. Complete 3D shapes can be inferred even though some parts of objects are not well captured in the 2D input images due to occlusion or irregular light reflections, advantages that have been noted by researchers in many of the recent publications on accurate 3D reconstruction using deep learning methods. Nevertheless, the accuracy of reconstructed shapes has not yet reached a satisfactory level, especially when the resolution of the input image is not high, because the deep-learning-based 3D reconstruction network normally learns the relation between a high-resolution facial image and the associated facial shape. Even if the network is trained using low-resolution images, it would be difficult for deep learning methods to infer accurate 3D shapes from 2D images at different resolutions because training datasets usually lack coverage across a spectrum of resolution levels.

1.1 Thesis Contributions

To address some of the aforementioned challenges, in this thesis, we propose deep learning methods for reconstructing 3D human facial shapes from a single image by combining a Super-Resolution (SR) network with a 3D Facial Reconstruction (FR) network. Reducing the number of input images compared to classic multi-viewpoint 3D reconstruction, such as a multi-view-stereo method, we also concentrate on enhancing the accuracy of the shapes particularly when the resolution of the input image is not high.

Our algorithm is divided into three major parts: Super resolution, 3D facial reconstruction, and best pair selection. If there were no SR network before 3D reconstruction, a single facial photo could be input to the FR network. However, because the SR network is a non-deterministic algorithm whose output is not always optimal for 3D reconstruction and rendering, we should generate multiple SR images and corresponding 3D facial shapes. Examining only the SR output or FR output cannot determine the quality of the SR/FR pair because SR noise might influence the 3D reconstruction, and a failure to properly reconstruct 3D facial features implies an inadequate SR output. Hence, in order to reconstruct a high-quality 3D facial shape despite the low resolution of the input image, we choose the best pair considering both the noise level of the SR images and the distribution of distances among the reconstructed facial shapes.

In summary, the contributions of the thesis are as follows:

- We propose a new deep learning architecture that is composed of a Super-Resolution (SR) network and a 3D Facial Reconstruction (FR) network.
- We also propose a process that selects the best pair among multiple pairs of SR images and corresponding 3D facial shapes, which is important in combining an SR network with an FR network.
- We experimentally evaluate our method by applying it to real 3D facial data and demonstrate that despite low-resolution input images, the

method can reconstruct accurate 3D facial shapes and generate high-quality rendered images.

1.2 Thesis Overview

The remainder of this thesis is organized as follows:

Chapter 2 reviews related work on SR methods, 3D reconstruction methods, and evaluation metrics for images and shapes.

Chapter 3 develops our method. First, we overview our proposed architecture. Next, we describe the incorporated SR and FR networks. Finally, we explain how the best pair is selected from multiple input-output pairs generated by the networks.

In Chapter 4 we present our experiments and report the results generated by our method. We also numerically compare the results against those generated without using the SR network.

Finally, we present our conclusions and future research directions in Chapter 5.

CHAPTER 2

Related Work

Our architecture incorporates two networks, a Super Resolution (SR) network and a 3D Facial Reconstruction (FR) network. In this chapter, we first review SR methods. These methods, which have been researched since before 2000, have many applications such as enhancement of the resolution of microscopic and telescopic images. Next, we review 3D reconstruction methods. These methods have been proposed for more than 30 years and, with some modification or adjustment, some have been adapted to reconstruct 3D human faces. Finally, we review several metrics to evaluate images and shapes.

2.1 Super Resolution

Super Resolution (SR) algorithms generate high-resolution signals from low-resolution ones. Their application includes movies or TV shows, which were recorded at low resolution, being enhanced to higher resolution (Matsuo and Sakaida, 2017). SR originally started from the interpolation of an image. Nearest neighbor interpolation, bilinear interpolation, and bicubic interpolation are basic functions implemented in image processing libraries for many programming languages, and the performance of these algorithms is well researched (Mohamad et al., 2017). These methods can interpolate an image at low computational cost, but the quality of the interpolation is far from modern SR.

Today's SR methods are generally divided into two categories. One is a reconstruction-based method, and the other is a learning-based method. We will review both in turn.

Reconstruction-based SR is installed mainly in a wide range of consumer products such as televisions. Such algorithms estimate an SR image from multiple low-resolution

images (Elad and Feuer, 1999; Zhang et al., 2010). The method hardly fails to produce an SR image, and it can run in real time if the required compute power and memory capacity are provided to achieve pixel alignment among multiple images.

For example, Farsiu et al. (2004) applied a low-pass filter to a high-resolution image, and sub-sampled it to generate multiple low-resolution images. They then reconstructed the original high-resolution image by combining those low-resolution images like a jigsaw puzzle. As such, reconstruction-based SR is based on the premise of the existence of the original high-resolution image. This limitation has been discussed for many years (Baker and Kanade, 2002).

Next, we discuss learning-based SR methods, in which machine-learning networks are often used. The degradation of the original images is simulated, and numerous original and degraded image pairs images are collected into a training dataset. The machine-learning network learns the degradation pattern. The trained network can then infer a high-resolution image from a low-resolution image. Such algorithms can generate a high-resolution from a single or a few low-resolution images, but if the input does not have features similar to the images in the training dataset, they do not work well.

Generally, cutting edge learning-based SR can be divided into two parts, based on Generative Adversarial Networks (GANs) or Diffusion Models. Applied to SR, GANs are referred to as SRGANs (Ledig et al., 2017). In the training process, the SRGAN generator outputs fake SR images to the discriminator, and the discriminator predicts whether or not the image is a fake high-resolution image. In this process, the generator becomes able to generate plausible SR images that can deceive the discriminator. Wang et al. (2018b) proposed the Enhanced SRGAN (ESRGAN), which removed batch normalization layers and replaced the Residual Block with a Residual-in-Residual Dense Block from the SRGAN. The latest version of SRGAN is Real-ESRGAN (Wang et al., 2021). These methods can generate high-quality SR images, but GANs have the potential problem that the training process is very difficult and unstable depending on the parameter tuning.

To overcome the problems of GAN, learning-based SR using Diffusion Models have

received substantial attention. The diffusion model was first proposed by Sohl-Dickstein et al. (2015). Recently, the model has become more stable and can generate promising results. At the beginning, the diffusion model gradually adds Gaussian noise to the training data and degrades the quality until the original training data becomes pure noise. Next, the neural network can be trained to learn the inverse of the noise-adding process. During inference, the trained network gradually removes image noise until a clean image is produced, which was interpreted as the process that generates potential SR data according to the gradient of the density of the data. One of the latest diffusion model methods is called SR3, proposed by Saharia et al. (2022), which can generate SR images that are comparable to the results generated by ESRGAN. SR3 is based on Denoising Diffusion Probabilistic Models (DDPM) and the U-Net architecture (Ronneberger et al., 2015). SR3 is easy to train and stable compared to GANs, but its inherent randomness can sometimes generate noisy and blurry results.

2.2 3D Reconstruction

3D reconstruction is one of the most important topics in Computer Vision and Computer Graphics. A great many methods have been proposed, but prior to around 2000, it was not often practical to use them because 3D reconstruction requires computational resources and time. However, rapid improvements in computational performance have made it possible to perform 3D reconstruction in laptop computers and smart phones.

One of the most popular uses of 3D reconstruction is in e-commerce (Lu et al., 2011; Vladimirov et al., 2021). For example, after registering their body shapes, customers buying clothes can check whether or not the clothing items will fit them well by examining their 3D models wearing virtual clothing, and customers can choose shoes by having their 3D foot models try on virtual shoes.

A traditional 3D reconstruction algorithm is the stereo-camera method (Hartley and Zisserman, 2003). Two cameras face the target object at a certain distance and acquire simultaneous photographs of the object. The reconstruction algorithm finds

corresponding points in the pair of images and, from the image disparities associated with the corresponding points, determines the distance from the well-calibrated cameras to those points on the surface of the object.

The stereo-camera algorithm can reconstruct 3D shapes, but two cameras cannot completely cover the surfaces of most target objects. They cannot observe the backs of objects and are subject to self-occlusions due to bumps on object surfaces or occlusions between objects when observing multiple objects of interest. To address this problem, the multi-view stereo method was proposed (Kang et al., 2001; Strecha et al., 2006). This algorithm uses more than two cameras, ideally dozens or hundreds of cameras, to capture unoccluded images of the target surfaces. Generally, the reconstructed shapes are accurate, but objects with anisotropic reflection areas make it difficult for the method to find corresponding points in the captured images because the appearance of the points changes depending on the viewing direction. Additionally, installing hundreds of cameras is expensive considering the equipment and labor costs. To reconstruct target objects accurately, one should calibrate the cameras well (Orteu et al., 1997) and control the depth of field to focus on the objects of interest, which is delicate work.

To address the above problems, deep-learning-based 3D reconstruction methods have recently been proposed. These methods can reconstruct a target 3D shape from either a single or a few images (Choy et al., 2016; Pontes et al., 2018; Wang et al., 2018a). First, the deep learning network can be trained to learn the relationship between images and the corresponding object shapes in the training set. The network can interpret the shape correlated with the outline and the shading of the object image. After the network learns the relationship, it can reconstruct the rough shape of an object, even if it has anisotropic reflection, because the network recognizes the object only by looking at the input image. Objects with some occlusions can be reconstructed well because the network can infer the occluded parts of the object. To train the network, constructing appropriate training datasets is also important. Databases including images and the target objects have been constructed for several years (Chang et al., 2015). Those databases are used by many researchers and more data will be added to them as a result of cooperation between

companies and academia.

The 3D reconstruction methods surveyed above are applicable to general objects. To enhance reconstruction quality, focusing on specific types of objects is a good idea. In this thesis, we focus on the reconstruction of 3D faces, which enables us to use some facial features such as the symmetry and the outlines of parts of the face.

Using a single or a several images, Choi et al. (2010) reconstructed 3D facial shapes by warping a generic 3D face model to match facial feature points in the images. A generic 3D face is often used based on the assumption that human faces are alike. Furthermore, if we use the generic face models based on gender and race, the reconstruction quality improves (Kaoru et al., 1995). Structure from Motion (SfM) is also a well-known algorithm to reconstruct faces. Several cameras or one moving camera are used to track feature points in the images captured by those cameras (Lee et al., 2011).

Deep learning methods are also used for 3D facial reconstruction. GANs and Convolutional Neural Networks (CNNs) are often used for 3D facial reconstruction (Gecer et al., 2019). However, most of the results cannot express facial details, such as wrinkles. Futhermore, training GANs is time consuming, and the robustness is low especially for in-the-wild facial images. Recently, Feng et al. (2021) proposed DECA (Detailed Expression Capture and Animation) to reconstruct facial shape with individual features. The detailed models are reconstructed based on FLAME (Li et al., 2017) geometry, a statistical 3D model that is composed of linear identity shape and expression spaces. This method can produce a displacement map that represents individual wrinkles and expressions after training. The latest version, MICA (MetrIC fAce) by Zielonka et al. (2022), has produced good results by inputting a sequence of face images. These methods can generate accurate 3D faces, but the quality of the output shape depends on the resolution of the input images. If the resolution of the input image is too low, this algorithm will not generate an accurate shape.

2.3 Evaluation Metrics

Many methods are used to evaluate the difference or the degradation of an image. One of the most popular is Mean Square Error (MSE), which can be calculated by averaging the squared difference of each pixel value between an original image and a processed image. MSE is often used to evaluate the quality of SR (Vandewalle, 2006) and it is also used as a loss function in an SR network (Dong et al., 2015).

Another metric for evaluating images is Signal to Noise Ratio (SNR), the ratio between an original image and the undesired noise in the processed image. This method is also used in the image processing area (Chen et al., 2012). Peak SNR (PSNR) measures the ratio between the maximum possible power of a signal and the noise deteriorating the signal. When the PSNR is high, the degradation of the image is small, and vice versa. PSNR is widely used as a metric for objective evaluation because it is generally highly correlated with subjective evaluations of image quality (Fardo et al., 2016; Anbarjafari and Demirel, 2010).

Sometimes, PSNR is inconsistent with subjective evaluation. Structural SIMilarity (SSIM) has been proposed for images for which PSNR is inappropriate. PSNR focuses on the perceptual sensitivity, whereas SSIM focuses on the degree of similarity of image structures. SSIM's evaluation formula is more complex than PSNR's and has some parameters. Therefore, PSNR and SSIM are used in a case-dependent manner (Hore and Ziou, 2010).

Finally, as a metric to evaluate the difference between two shapes, Hausdorff Distance (HD) is the most widely used (Unan et al., 2019). Mathematically, HD can represent the distance between two sets by calculating the maximum distance of a point in one set to the nearest point in the other set. In other words, any point in a set can reach one of the points in the other set by moving at most the HD.

CHAPTER 3

Methodology

In this chapter, we will introduce how we reconstruct accurate 3D facial shapes from a low-resolution image.

Figure 3.1 presents an overview of our architecture. First, a low-resolution image is input to the Super-Resolution (SR) network. Here, we show a low-resolution facial image of a sports player as input. The images of sport scenes might be captured as a sequence, but if one of the images is out of focus, the remaining images in the sequence will also be out of focus. We will not acquire any high-resolution information from the sequence. In this case, it is difficult to use a reconstruction-based SR method as described in Chapter 2.1. Therefore, we use a leaning-based SR method in our architecture.

Considering the stability of the training and the quality of the outputs, we selected SR3 proposed by Saharia et al. (2022). Unlike GAN-based methods, which can generate high-quality SR outputs, but known to suffer from an unstable learning process, SR3 is

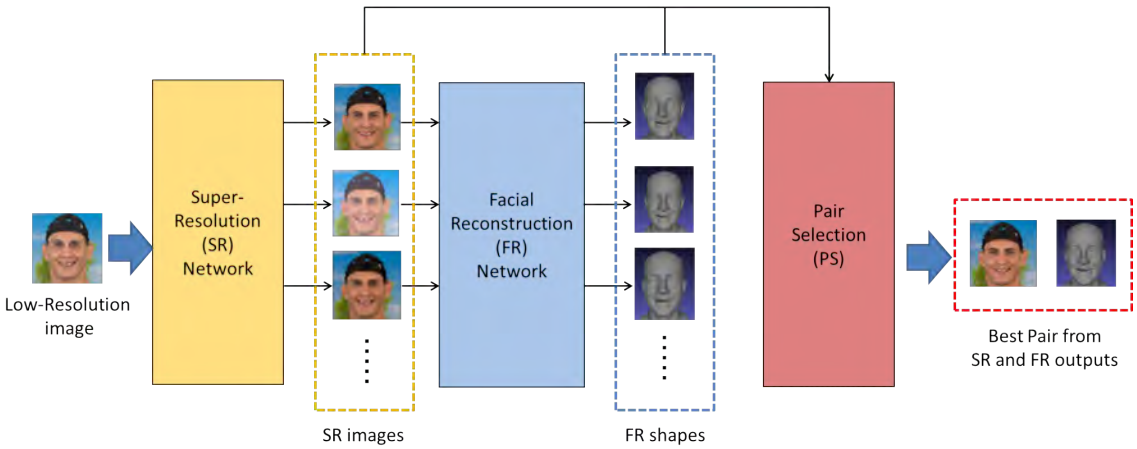


Figure 3.1: Overview of Our Architecture

stable and features the Diffusion Model along with a U-Net architecture. However, the randomness inherent to the Diffusion Model changes the output quality of SR3 each time it is invoked on the same input. To tackle this problem, we select the best among multiple SR3 outputs by assessing both the noise level of the SR outputs and the quality of the facial shapes from a Facial Reconstruction (FR) network. That is to say, we generate multiple SR images using SR3, and have the FR network output the corresponding facial shape for each image, and we then select the best pair.

We reviewed FR methods in Section 2.2. Considering the fact that we input a single still facial image per one facial shape, we selected DECA (Detailed Expression Capture and Animation) proposed by Feng et al. (2021). However, DECA needs a clear image to reconstruct a detailed 3D facial shape. Therefore, we input multiple SR images to DECA and acquire multiple facial shapes. Some of the SR images could have strong noise or some poor facial features. We therefore choose the best pair of an SR image and the corresponding FR output based on the quality of the SR, which could have some noise, and the FR, which could fail to adequately reconstruct parts of the face.

Finally, in the process of Pair Selection in Figure 3.1, we choose the best pair from the SR and FR outputs. As discussed in Section 2.3, many metrics have been proposed to evaluate the quality of images and shapes. As an evaluation indicator for SR images we use PSNR because generally PSNR can evaluate the degradation of images well and it is highly correlated with subjective evaluation. As an indicator to evaluate the FR outputs, we use Hausdorff Distance (HD), which can represent the distance between two sets. We calculate PSNRs between an SR output and the other SR outputs one by one, and average the PSNRs. For all of the SR outputs, we calculate averaged PSNRs. We also calcualte HD for all of the FR outputs. Ultimately, we estimate scores for all the PSNR and HD pairs, and select the pair with the best score. In this way, our architecture can generate a high-quality SR image and the corresponding facial shape.

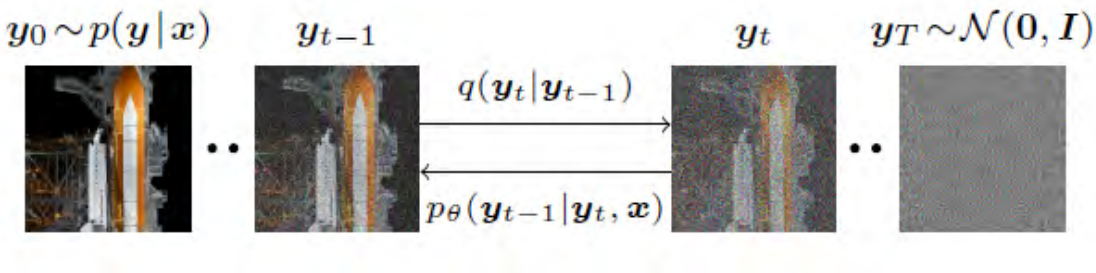


Figure 3.2: Overview of Diffusion Process (from the paper by Saharia et al. (2022))

3.1 Super Resolution

In this section, we introduce the key components of SR3: the Diffusion Model (DM) and the U-Net.

3.1.1 Diffusion Model

The DM assumes that the original signal should become pure Gaussian noise by gradually adding noise to it. Modeling the inverse process of becoming pure noise is an avenue to generating an SR image from a noisy image.

Figure 3.2 shows an overview of diffusion process. The DM starts from pure noise $\boldsymbol{y}_T$ and in T noise removing steps can generate the image $\boldsymbol{y}_0$. To remove the noise, the DM should know how the noise was added to the original image. The noise adding process is called forward diffusion, and the noise removing process is called reverse diffusion.

The forward diffusion process adds Gaussian noise to a state $\boldsymbol{y}_{t-1}$ and transitions the state to $\boldsymbol{y}_t$. If the intensity of the noise is β_t, this process can be described as

$$q(\boldsymbol{y}_t|\boldsymbol{y}_{t-1}) = \mathcal{N}(\boldsymbol{y}_t; \sqrt{1-\beta_t}\boldsymbol{y}_{t-1}, \beta_t\mathbf{I}), \tag{3.1}$$

and

$$q(\boldsymbol{y}_{1:T}|\boldsymbol{y}_0) = \prod_{t=1}^{T} q(\boldsymbol{y}_t|\boldsymbol{y}_{t-1}). \tag{3.2}$$

These equations mean that an arbitrary state $\boldsymbol{y}_t$ at step t can be expressed using $\boldsymbol{y}_{t-1}$. Thus, we iterate this transition t's time and can express $\boldsymbol{y}_t$ by the initial state $\boldsymbol{y}_0$. Letting

$\alpha_t = 1 - \beta_t$, $\overline{\alpha}_t = \Pi_{i=1}^{t} \alpha_i$, and $\boldsymbol{z} \sim \mathcal{N}(0, \mathbf{I})$, we obtain

$$\begin{aligned} \boldsymbol{y}_t &= \sqrt{\alpha_t}\boldsymbol{y}_{t-1} + \sqrt{1-\alpha_t}\boldsymbol{z}_{t-1} \\ &= \sqrt{\alpha_t\alpha_{t-1}}\boldsymbol{y}_{t-2} + \sqrt{1-\alpha_t\alpha_{t-1}}\overline{\boldsymbol{z}}_{t-2} \\ &= \sqrt{\overline{\alpha}_t}\boldsymbol{y}_0 + \sqrt{1-\overline{\alpha}_t}\boldsymbol{z} \end{aligned} \tag{3.3}$$

$$q(\boldsymbol{y}_t|\boldsymbol{y}_0) = \mathcal{N}(\boldsymbol{y}_t; \sqrt{\overline{\alpha}_t}\boldsymbol{y}_0, (1-\overline{\alpha}_t)\mathbf{I}). \tag{3.4}$$

The reverse diffusion process can be defined as $q(\boldsymbol{y}_{t-1}|\boldsymbol{y}_t)$. When β_t is sufficiently small, $q(\boldsymbol{y}_{t-1}|\boldsymbol{y}_t)$ follows a Gaussian distribution. This can be approximated by a neural network p_θ parameterized by θ.

$$p_\theta(\boldsymbol{x}_{0:T}) = p(\boldsymbol{y}_T)\prod_{t=1}^{T} p_\theta(\boldsymbol{y}_{t-1}|\boldsymbol{y}_t), \tag{3.5}$$

$$p_\theta(\boldsymbol{y}_{t-1}|\boldsymbol{y}_t) = \mathcal{N}(\boldsymbol{y}_{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{y}_t, t)\sum_\theta(\boldsymbol{y}_t, t)). \tag{3.6}$$

To handle $q(\boldsymbol{y}_{t-1}|\boldsymbol{y}_t)$ easily, we condition it by $\boldsymbol{y}_0$. Assuming that $q(\boldsymbol{y}_{t-1}|\boldsymbol{y}_t, \boldsymbol{y}_0)$ also follows a Gaussian distribution, the average $\boldsymbol{\mu}$ and variance σ^2 are expressed as follows:

$$q(\boldsymbol{y}_{t-1}|\boldsymbol{y}_0, \boldsymbol{y}_t) = \mathcal{N}(\boldsymbol{y}_{t-1}|\boldsymbol{\mu}, \sigma^2\mathbf{I}), \tag{3.7}$$

$$\boldsymbol{\mu} = \frac{\sqrt{\overline{\alpha}_{t-1}}\beta_t}{1-\overline{\alpha}_t}\boldsymbol{y}_0 + \frac{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})}{1-\overline{\alpha}_t}\boldsymbol{y}_t, \tag{3.8}$$

$$\sigma^2 = \frac{(1-\overline{\alpha}_{t-1})\beta_t}{1-\overline{\alpha}_t}. \tag{3.9}$$

Considering the above equations, given the noise vector $\boldsymbol{\epsilon}$ and a source image $\boldsymbol{x}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$, $f \in \{1, 2\}$, and $\overline{\alpha} \sim f(\overline{\alpha})$, we can set the objective function of our neural network as follows:

$$\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})}\mathbb{E}_{(\boldsymbol{\epsilon},\overline{\alpha})}\left\|p_\theta(\boldsymbol{x}, \sqrt{\overline{\alpha}}\boldsymbol{y}_0 + \sqrt{1-\overline{\alpha}}\boldsymbol{\epsilon}, \overline{\alpha}) - \boldsymbol{\epsilon}\right\|_f^f \tag{3.10}$$

To generate SR images, we used a neural network trained on the Flickr-Faces-HQ

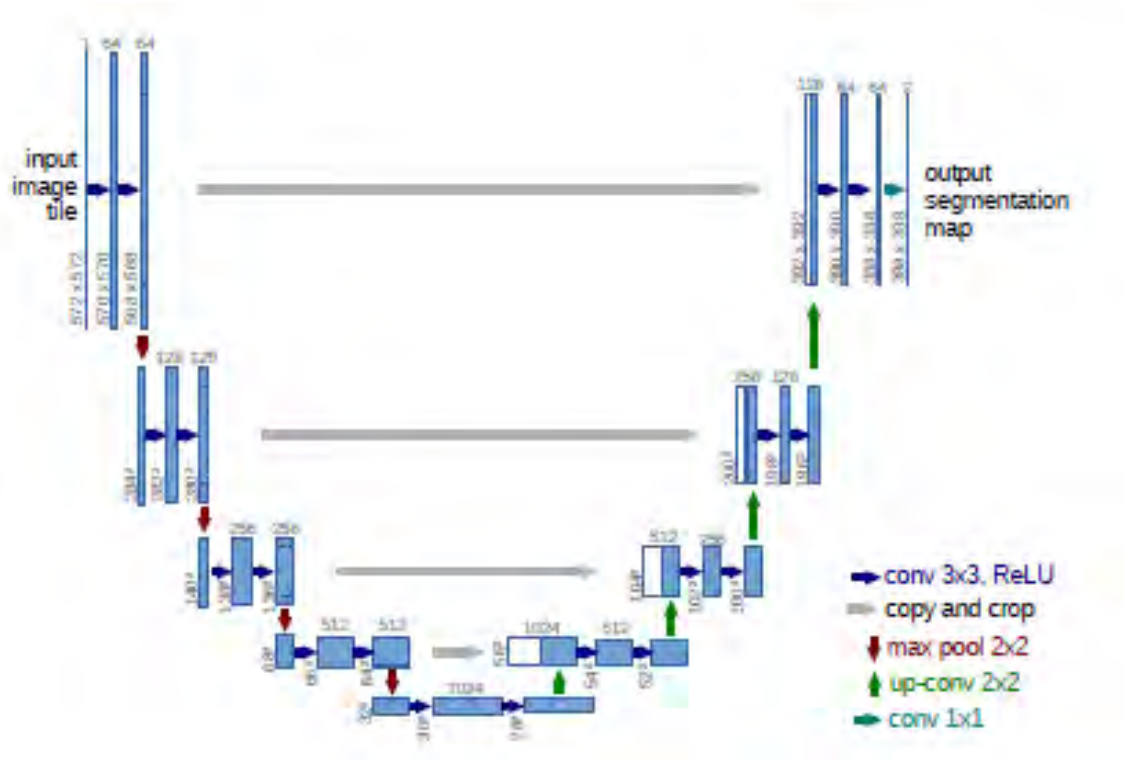


Figure 3.3: The architecture of U-Net (from the paper by Ronneberger et al. (2015))

(FFHQ) dataset (Karras et al., 2019).

3.1.2 U-Net

The U-net (Ronneberger et al., 2015) was first proposed as a model for semantic segmentation in the biomedical imaging field. We show the overview of the original U-Net in Figure 3.3, whose encoder-decoder architecture takes the shape of the letter "U."

The encoder convolves and down-samples the input image several times, extracting feature maps of the image at multiple scales. A model such as ResNet, which is often used in the image processing area, can be used. The decoder deconvolves the feature maps and up-samples them. Simply up-sampling makes it difficult to accurately detect the positions of objects. Therefore, U-Net combines the feature maps in the encoder with the feature maps in the decoder, as represented by the gray arrows in Figure 3.3. In this way, the information of the feature map in the encoder can be conveyed to the decoder, which makes detecting object positions easy when up-sampling. The output of the U-net

is a segmentation probability map of the same size as the input image.

After U-Net was introduced, many variants were proposed. SR3's U-Net is similar to the structure used for the Denoising Diffusion Probabilistic Model (DDPM) proposed by Ho et al. (2020). SR3 made some changes to DDPM's U-Net. For example, Saharia et al. (2022) replaced the DDPM residual blocks with BigGAN's residual blocks, and changed the skip connections. Following the work by Saharia et al. (2022), we use

$$f(\overline{\alpha}) = \sum_{t=1}^{T} \frac{1}{T} U(\overline{\alpha}_{t-1}, \overline{\alpha}_t) \tag{3.11}$$

for a distribution $\overline{\alpha}$. We also set $T = 2000$ in a time step $t \sim \{0..T\}$, along with $\overline{\alpha} \sim U(\overline{\alpha}_{t-1}, \overline{\alpha}_t)$.

3.2 Facial Reconstruction

DECA (Feng et al., 2021), which we use in our architecture to reconstruct 3D facial shapes from SR images, is characterized as a technique that divides facial parameters into general expression ones and individual detailed ones. By changing the general expression parameters, DECA changes the expression of the individual face while keeping person-specific features like wrinkles. Here, we do not use the animation component of DECA, but we use the reconstruction from a single facial image component.

Figure 3.4 shows an overview of DECA. The left box illustrates the training part of DECA and the right box illustrates the application part.

A facial image is input to two encoders. Encoder E_c infers general expression parameters, such as camera parameters, albedo, lighting, a shape, a pose, and a facial expression. By extracting these parameters, DECA generates a rough shape, and minimizes the loss between the rendered rough shape I_r and the input image. As for loss functions, landmark re-projection loss, eye closure loss, photometric loss, identity loss, and shape consistency loss are used. Encoder E_d extracts individual detail parameters. Both of the parameters extracted by E_c and E_d are integrated and the detailed shape is generated. Subsequently,

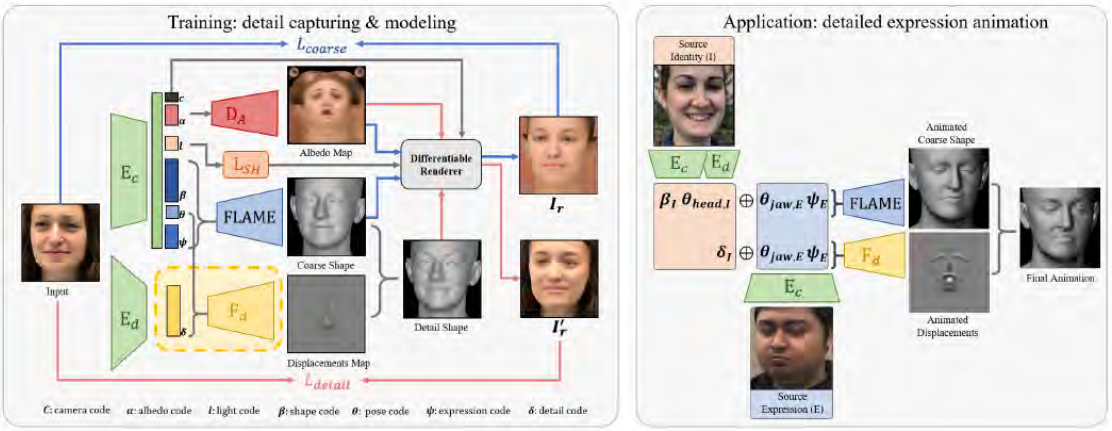


Figure 3.4: Overview of DECA (from the paper by Feng et al. (2021))

DECA minimizes the loss between the rendered detail shape I'_r and the input image. The training of DECA was completed using three publicly available datasets: VGGFace2 (Cao et al., 2018), BUPT-Balancedface (Wang et al., 2019), and VoxCeleb2 (Chung et al., 2018).

As shown in Figure 3.4, DECA can generate a detailed facial shape from a single input image. The network can generate a facial animation by changing the general expression parameters extracted by E_c, but we do not use this facility.

3.3 Pair Selection

After we generate multiple SR images from our SR network and FR shapes from our FR network, we choose the best pair from them. First, we evaluate the SR images based on PSNR, as follows:

$$\text{PSNR} = 10 \log \frac{\text{MAX}_I^2}{e^2}, \tag{3.12}$$

$$\text{with } e^2 = \frac{1}{wh} \sum_{i=0}^{w-1} \sum_{j=0}^{h-1} |X(i,j) - Y(i,j)|^2, \tag{3.13}$$

where the original image is X, the processed image is Y, the height of the image is h, the width of the image is w, and the max pixel value is MAX_I.

We calculate PSNRs between an SR output and the other n SR outputs one by one and average the PSNRs as follows:

$$\mathrm{PSNR}_t = \frac{1}{n-1} \sum_{k=1,k\neq t}^{n} 10 \log \frac{\mathrm{MAX}_I^2}{e_k^2}, \tag{3.14}$$

$$\text{with } e_k^2 = \frac{1}{wh} \sum_{i=0}^{w-1} \sum_{j=0}^{h-1} \left| X_t(i,j) - Y_k(i,j) \right|^2, \tag{3.15}$$

where t indexes the target SR output.

Next, we calculate the averaged Hausdorff Distance (HD) for each FR shape. The HD for two different two shapes X and Y is formulated as follows:

$$d(X,Y) = \max\{\sup_{x\in X} \inf_{y\in Y} d(x,y), \sup_{y\in Y} \inf_{x\in X} d(x,y)\}, \tag{3.16}$$

where x and y denote points of shapes X and Y, respectively. As stated in Section 2.3, the HD represents the distance between two shapes by calculating the maximum distance of a set to the nearest point in the other set. Here, we use a part of the HD calculation to compute the Root Mean Squared (RMS) of the distance from points on one mesh to the ones on the target mesh. We refer to this RMS as RMSHD, and write it as follows:

$$\mathrm{HD}_t(X,Y) = \frac{1}{n-1} \sum_{k=1,k\neq t}^{n} \mathrm{RMSHD}(X_t, Y_k), \tag{3.17}$$

$$\text{with } \mathrm{RMSHD}(X,Y_k) = \sqrt{\frac{1}{n(X)} \sum_{x\in X} (\inf_{y\in Y_k} d(x,y))^2}, \tag{3.18}$$

where n is the total number of the FR outputs and the index of the target FR output is t.

Finally, we calculate the score for each pair as follows:

$$S_t = \alpha \frac{\mathrm{PSNR}_t}{\max_{k\in n}\{\mathrm{PSNR}_k\}} + \beta \frac{\min_{k\in n}\{\mathrm{HD}_k\}}{\mathrm{HD}_t}, \tag{3.19}$$

where the weight for the SR outputs is α and that for the FR outputs is β. We used

$\alpha = 0.5$, $\beta = 0.5$. We then select the best pair as P_b where

$$b = \arg\max_{k \in n} \{S_k\}. \tag{3.20}$$

CHAPTER 4

Results

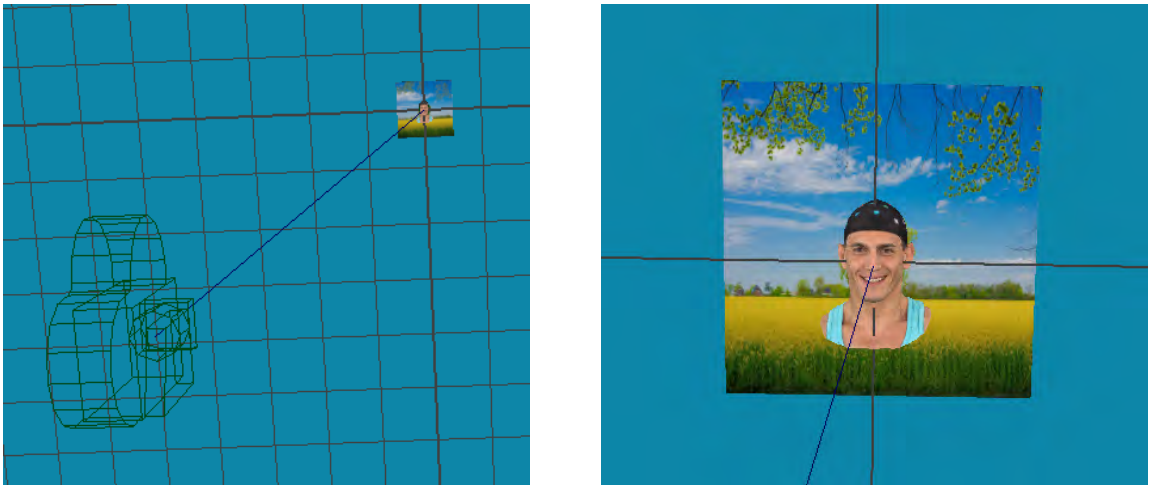
In this chapter, we apply our method to low-resolution images and confirm that it can super-resolve those images and generate 3D facial shapes. First, we conduct an experiment that demonstrates the validity of our method. In this experiment, we use 3D facial data that were captured by a 3D face scanner. These scanned data were edited by a modeling artist to enhance the quality of the shapes. We set these data as target shapes for our method and calculate the difference between them and the shape generated by our method. Second, we conduct an experiment to show that our method can be applied to various sports scenes, specifically low resolution photos of baseball, basketball, rugby, and soccer action.

4.1 Objective Evaluation

4.1.1 Reconstructing the Front of a Male Face

To evaluate our method objectively, initially we must prepare target data to be reconstructed. Our method super-resolves a low-resolution and generates a 3D facial shape from it. Therefore, for the target data, we need a pair of a low-resolution image and the corresponding facial shape. We also require knowledge of the viewpoint and view direction of the camera that acquired the low resolution image in order to calculate the difference between the original facial shape and a generated one. Generally, this kind of dataset is difficult to obtain.

There are some databases that include hundreds of pairs of facial images and corresponding 3D facial shapes, but those images are at a high resolution. Even if we made



(a) Camera, Face, and Background (b) Close-Up View of Face and Background

Figure 4.1: Virtual 3D Scene for Facial Image Rendering

the resolution lower by some filter, it would be somewhat artificial.

To overcome the problem, we use facial data that was captured by a 3D scanner. We downloaded 3D facial shapes from *Turbosquid*,[1] in which a great number of 3D shapes can be purchased.

The first target shape is that of a smiling man called "Mike." We first set the Mike shape in a 3D space and set up a virtual camera that renders an image. Figure 4.1 shows the virtual scene. Using Maya,[2] a popular modeling/animation/rendering software package widely used in the gaming and movie industries, we determined the position and the view direction of the virtual camera to render frontal images of the Mike face in 64×64 pixel resolution as is appropriate for our application (Figure 4.1a). Figure 4.1b shows a close-up view of the target 3D facial shape in front of a background.

A trimmed 64×64 facial image is shown in Figure 4.2a. Next, we simulate the blur that is not uncommon in sports images by applying to this image a 3-pixel sized Gaussian Filter,

$$G(x, y) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right), \tag{4.1}$$

[1] See https://www.turbosquid.com/

[2] See https://www.autodesk.com/products/maya/

(a) Trimmed 64×64 Image

(b) Gaussian-Filtered Image

Figure 4.2: Processing the Mike Image

Table 4.1: PSNRs of SR outputs

ImageNum	1	2	3	4	5	6	7	8	9	10
PSNR [dB]	24.1	23.1	21.6	17.1	18.8	22.8	24.3	22.1	24.0	24.2

with $\sigma = 0.6$. The blurred image is shown in Figure 4.2b.

The blurred image is input to our SR network. As discussed in Chapter 3, we generate multiple images from our SR network. In this experiment we set the number of image outputs to $n_{\mathrm{SR}} = 10$. We show the results from our Super-Resolution (SR) network in Figure 4.3. Some outputs have heavy noise that will deteriorate the quality of the rendered image and make it difficult for our 3D Facial Reconstruction (FR) network to detect the feature points in the face. Some outputs have indistinct facial parts that will lead to failure in the corresponding 3D shape generation. To detect heavy noise in SR outputs, we use PSNR as a metric, as discussed in Chapter 3. We calculate the PSNR between pairs of output images, and we continue this process for all of the outputs. Subsequently, we average the PSNRs for each output.

We show the results of averaged PSNRs in Table 4.1. The averaged PSNR is worse in Image 4 and Image 5. They appear very noisy, which indicates that PSNR can measure the noise level of the SR outputs.

Figure 4.3: Outputs From the SR network

Table 4.2: Hausdorff Distances of FR outputs

ShapeNum	1	2	3	4	5	6	7	8	9	10
RMSHD [mm]	1.09	1.20	1.35	1.87	1.32	1.00	1.42	1.10	1.26	1.26

After calculating the PSNRs, we input the SR outputs to our FR network. Figure 4.4 shows reconstructed facial shapes. Even though those shapes look similar, the distance between two shapes shows that the outputs from our FR network differ depending on the quality of the SR outputs. To calculate the differences between FR outputs, we used Hausdorff Distance as a metric, per equations (3.17) and (3.18). We calculated the Root Mean Squared of Hausdorff Distance (RMSHD) between one output shape and the other output shapes, one by one, and we continued this process for all of the outputs. Subsequently, we averaged the RMSHDs for each output. We show the averaged RMSHD results in Table 4.2.

Finally, we selected the best pair from SR and FR outputs by using the averaged

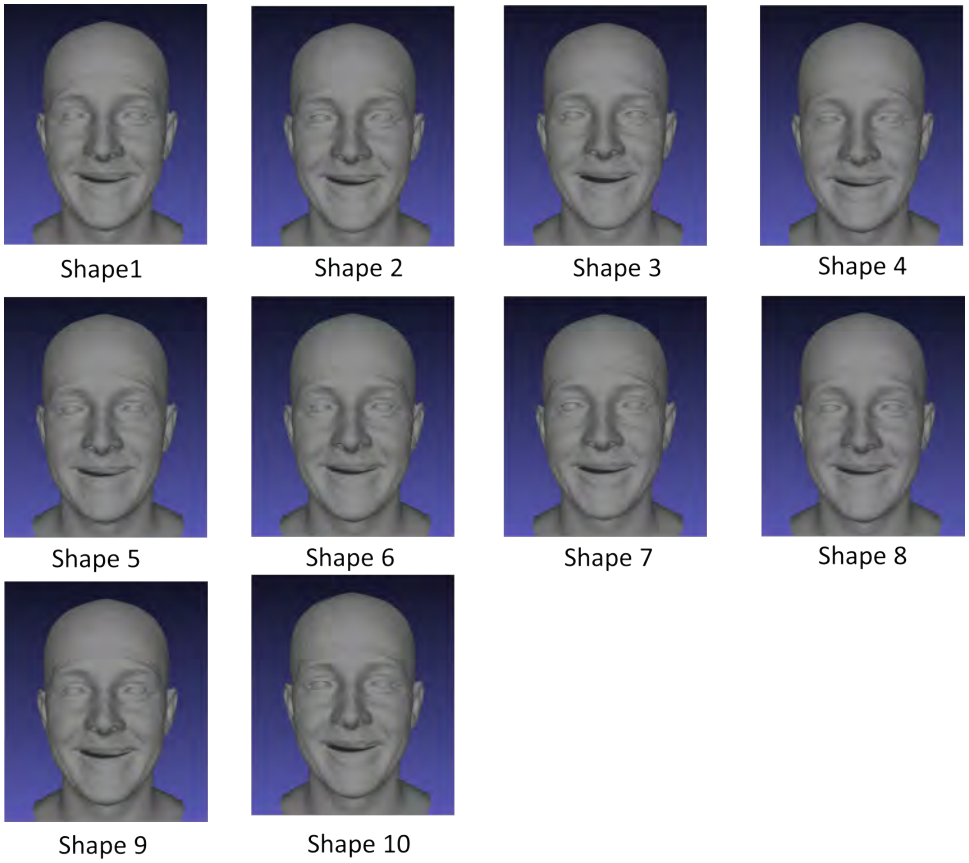


Figure 4.4: Outputs From the FR network

Table 4.3: Final Scores

ShapeNum	1	2	3	4	5	6	7	8	9	10
Score	0.96	0.90	0.82	0.62	0.77	0.97	0.95	0.92	0.89	0.90

PSNRs and Hausdorff Distances. According to (3.19), the final scores were calculated as shown in Table 4.3. Considering the results in the table, Image 6 and Shape 6 are the best pair, and Image 4 and Shape 4 are the worst pair.

Using heat maps, we show in Figure 4.5 the RMSHD between our best output and the original target shape, between our worst output and the original target shape, and between our output without SR and the original target shape. This clearly shows that our method can reconstruct the front of the face accurately from a low resolution image compared to reconstructing the face without SR. This heat map also shows that even if we used SR before FR, the output from FR might be worse than a result generated

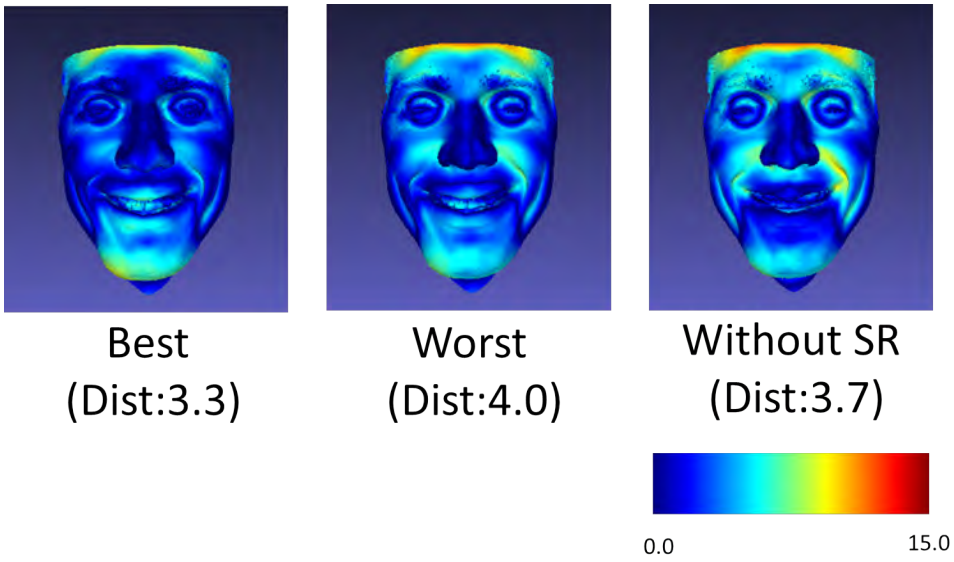


Figure 4.5: Comparison of Shape Differences

Figure 4.6: Rendered Frontal Images of a Male Face

without SR. Therefore, it appears beneficial that we generate multiple pairs of SR images and corresponding 3D facial shapes and select the best pair among them. Notably, our method can reduce the error values around the cheek and eyes dramatically.

Finally, in Figure 4.6 we show renderings from the above three FR shapes and SR images. The image rendered from the best image-shape pair is more vivid than the one generated without SR and not noisy compared to the worst one. The best one can express the details of the face, and the direction of the face matches that of the original face in our 3D space.

Table 4.4: Averaged PSNRs of Male Face Profile

ImageNum	1	2	3	4	5	6	7	8	9	10
PSNR [dB]	29.1	29.2	27.2	21.6	24.0	28.0	27.6	28.7	27.4	28.6

Table 4.5: Averaged RMSHDs of Male Face Profile

ShapeNum	1	2	3	4	5	6	7	8	9	10
RMSHD [mm]	0.529	0.614	1.08	0.644	0.646	0.537	0.756	0.508	0.651	0.907

Table 4.6: Final Scores of Male Face Profile

ShapeNum	1	2	3	4	5	6	7	8	9	10
Score	0.988	0.830	0.795	0.756	0.813	0.957	0.865	0.990	0.892	0.854

4.1.2 Other Cases

We also applied our method to other cases. First, using the 3D Mike model we rendered an image from a different camera viewpoint. Setting the target face as the center of rotation, we rotated the camera by 30 degrees horizontally. We then applied our method to the rendered profile image and obtained the SR and FR outputs. We selected the best pair from them and compared the RMSHD.

We show an original blurred image and SR outputs in Figure 4.7. We also calculated averaged PSNRs (Table 4.4). The corresponding FR outputs are shown in Figure 4.8 and averaged RMSHDs are shown in Table 4.5. Then, we calculated the final scores (Table 4.6). Based on the final scores, we selected Image 8 and Shape 8 as the best pair. Using heat maps in Figure 4.9, we show the RMSHD between our best output and the original target shape, between our worst output and the original target shape, and between our output without SR and the original target shape. Finally, Figure 4.10 shows the rendered images.

As we discussed in Chapter 4.1.1, worse averaged PSNRs indicate heavy noise in an SR image. For example, the averaged PSNR is the worst for Image 4 in Table 4.4. As expected, Image 4 is very noisy, which indicates that PSNR can measure the noise level

Figure 4.7: Original Image and SR Outputs of Male Face Profile

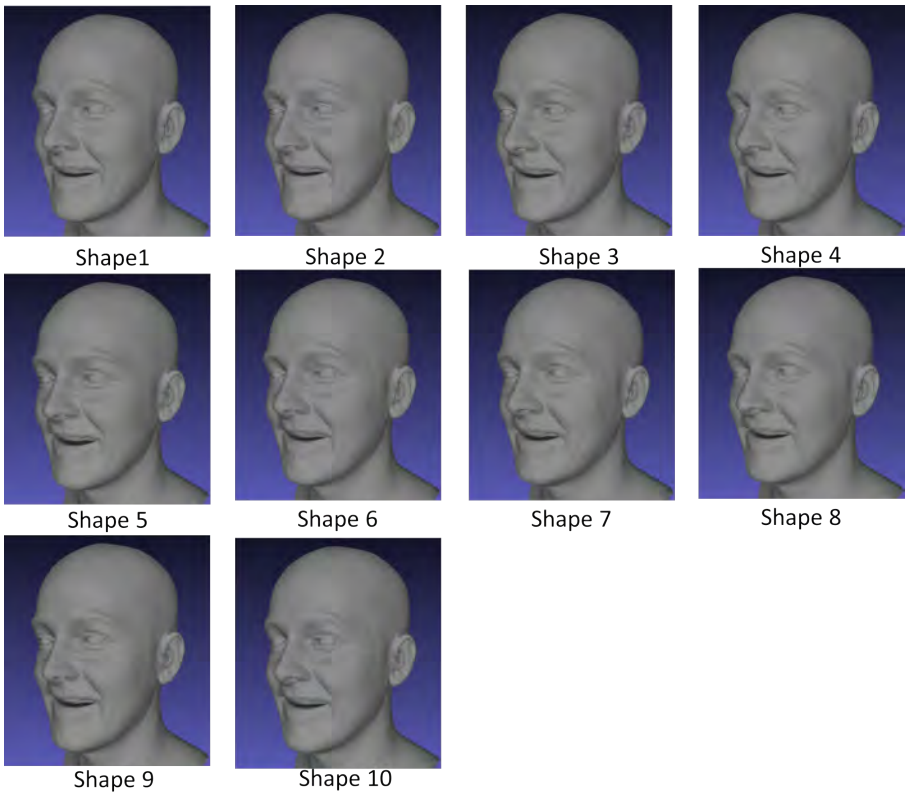


Figure 4.8: FR Outputs of Male Face Profile

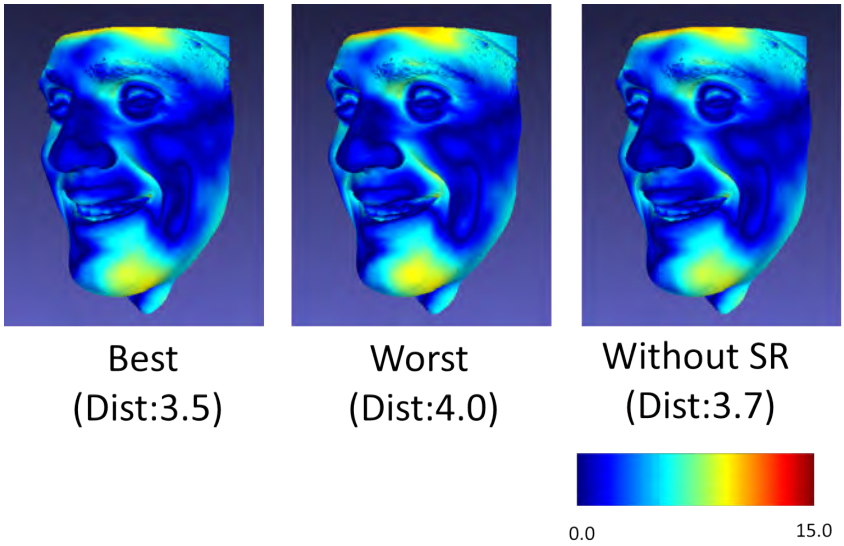


Figure 4.9: Comparison of Shape Differences of Male Face Profile

Figure 4.10: Rendered Profile Images of a Male Face

of the SR outputs in this experiment.

The FR output shapes also look similar in this experiment, but the RMSHD between two shapes shows the difference of the outputs from our FR network. For instance, in Table 4.5, the RMSHDs of Shape 3 and Shape 10 are larger than the others, which means that these two shapes differ most as facial shapes.

Regarding the heat maps in Figure 4.9, the best pair, Image 8 and Shape 8, which were selected based on the final score, has a smaller RMSHD compared to the worst one and the one generated without SR. Figure 4.10 also shows that the image and the shape that were selected by our method can be rendered with a noiseless texture and facial details. As described above, even though we change the viewpoint and direction of the

Table 4.7: Averaged PSNRs of Female Face Front

ImageNum	1	2	3	4	5	6	7	8	9	10
PSNR [dB]	26.6	26.4	16.6	25.4	25.5	26.9	25.0	24.7	14.3	27.0

Table 4.8: Averaged RMSHDs of Female Face Front

ShapeNum	1	2	3	4	5	6	7	8	9	10
RMSHD [mm]	2.1	2.2	1.7	1.5	1.6	1.6	1.5	1.7	1.0	1.5

Table 4.9: Final Scores of Female Face Front

ShapeNum	1	2	3	4	5	6	7	8	9	10
Score	0.86	0.84	0.75	0.97	0.95	0.96	0.96	0.91	0.34	0.99

Table 4.10: Averaged PSNRs of Female Face Profile

ImageNum	1	2	3	4	5	6	7	8	9	10
PSNR [dB]	25.7	22.7	27.6	26.0	24.6	19.2	27.0	26.7	25.7	27.5

Table 4.11: Averaged RMSHDs of Female Face Profile

ShapeNum	1	2	3	4	5	6	7	8	9	10
RMSHD [mm]	1.0	0.9	1.2	0.9	1.0	2.0	1.0	0.9	1.0	1.0

Table 4.12: Final Scores of Female Face Profile

ShapeNum	1	2	3	4	5	6	7	8	9	10
Score	0.92	0.90	0.87	0.97	0.86	0.53	0.89	0.95	0.81	0.82

virtual camera, our method can generate a more accurate facial shape and a high-quality rendered image than can be generated without SR outputs.

We also applied our method to frontal and left side views of a woman's face. For the frontal view, we tabulate the averaged PSNRs in Table 4.7, the averaged RMSHDs in Table 4.8, and the final scores in Table 4.9. For the side view, we tabulate the averaged PSNRs in Table 4.10, the averaged RMSHDs in Table 4.11, and the final scores in Table 4.12. Figure 4.11 shows the heat maps for the frontal view and Figure 4.12 shows them for the left side view. The heat maps indicate that our method can generate accurate

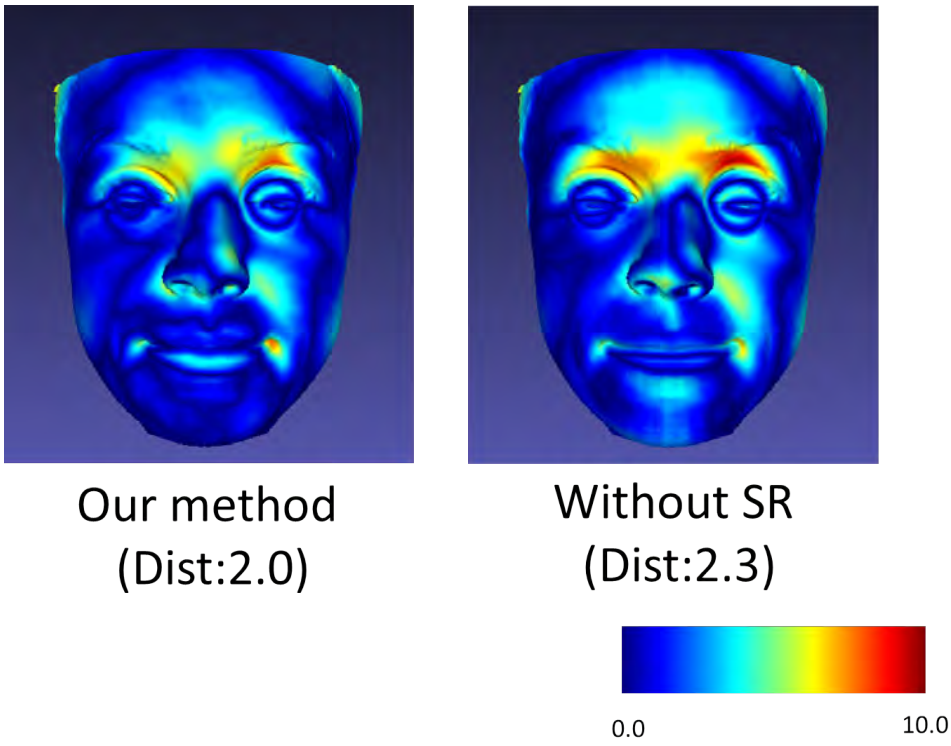


Figure 4.11: Comparison of Shape Differences of Female Face Front

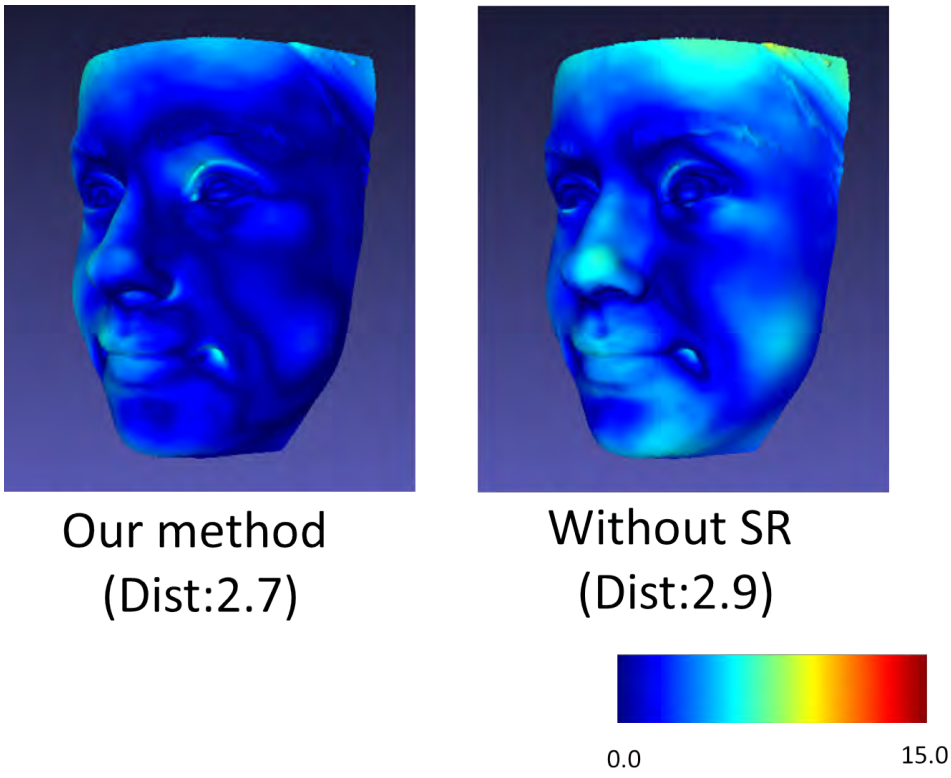


Figure 4.12: Comparison of Shape Differences of Female Face Profile

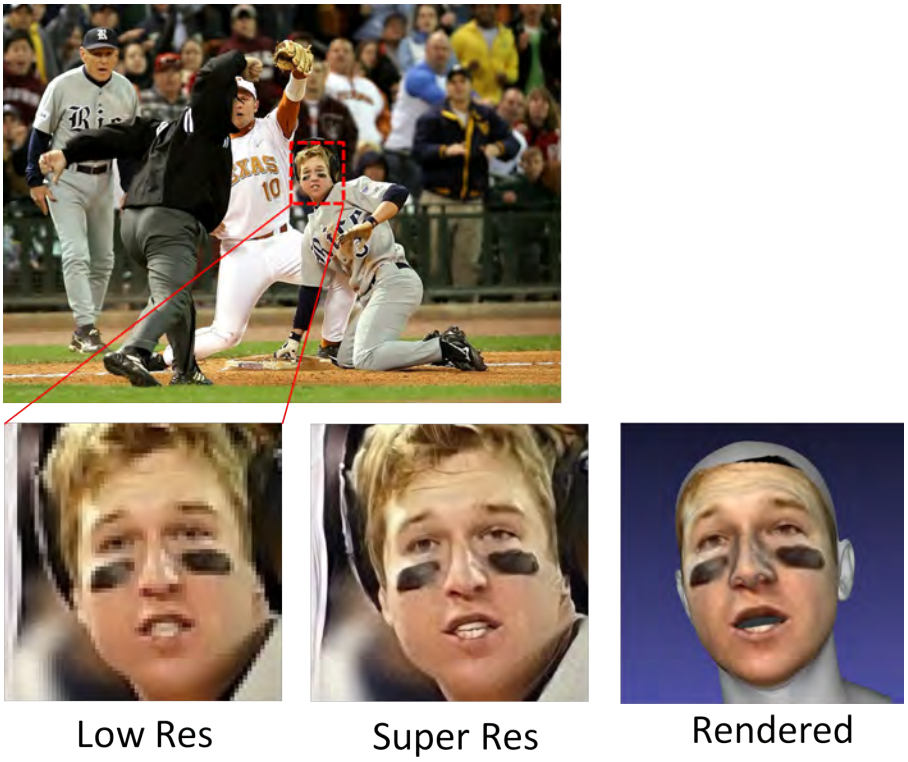


Figure 4.13: Baseball Scene

facial shapes from the best pair of SR and FR outputs for both views of the woman's face.

Considering the numerical analyses above, we conclude that our method can generate more accurate facial shapes from low resolution images than can be generated without the benefit of SR.

4.2 Application to Sports Scenes

In this section, we apply our method to various kinds of sports scenes.

In Figure 4.13, we show the result when applying our method to a baseball scene. In this context, we can see that our method is effective and can select a good pair from SR outputs and FR outputs. The original low-resolution image includes a baseball helmet, but the SR network works. Looking at the rendered image, even though the direction of the face is a bit different from the original one, the sharpness and the detail of the face are well expressed.

Figure 4.14: Basketball Scene

In Figure 4.14, we show the result when applying our architecture to a basketball scene. Looking at the rendered image, the direction of the face is nearly the same as the original one. The low resolution picture includes a lot of noise, but the SR network works to some extent.

In Figure 4.15, we show the result when applying our architecture to a rugby scene. A low-resolution face image is well super-resolved by our SR network. Looking at the rendered image, the direction of the face is nearly the same as the original and the facial parts are well reconstructed by our FR network.

Finally, in Figure 4.16, we show the result when applying our architecture to a soccer scene. An original low-resolution facial image is super-resolved well by our SR network. As for the rendered image, generally, the facial parts are well reconstructed, but there is a lack of expression around the eyes.

We conclude that our method can be applied to various kinds of sports scenes. We

Figure 4.15: Rugby Scene

Figure 4.16: Soccer Scene

can generate detailed facial shapes and a good-quality rendered images based on the combination of our SR and FR networks and the selection of the best pairs from the outputs of these networks.

CHAPTER 5

Conclusions and Future Work

5.1 Conclusions

In this thesis, we addressed the problem of reconstructing textured 3D models of human faces from within sports action images. Anticipated future application of this technology is in unconstrained-viewpoint observation of sports action. To this end, we proposed a deep-learning-based architecture and method that combines a Super-Resolution (SR) network and a 3D Facial Reconstruction (FR) network. We have shown that our method can generate accurate 3D facial shapes and good-quality rendering images even though the input image resolution is not high enough to be usable by previous 3D reconstruction networks.

Our method also tackles the problem that the output SR images are not stable because of the non-determinism of SR networks. The output SR image is often noisy and suffers some indistinct facial features, which undermines 3D facial shape reconstruction. If the output SR image is noisy, when it is used as a facial texture by the rendering algorithm, the rendered image will also be noisy. Therefore, we generate multiple pairs of SR images and corresponding facial shapes, and select the best pair based on analysis of the noise-level difference among SR images and the distance among the generated shapes.

We also showed the validity of our method by numerically comparing the results that it generates against those generated without SR. If we do not use an SR method before 3D facial reconstruction, the quality of the facial shapes declines.

5.2 Future Work

We weighted the noise level and shape distance equally, and that sufficed to select good pairs, but image resolution and noise can influence weight selection, and modification of the weights could improve our results. For instance, if the noise in an input image is too high, it might be difficult to use a noise level as a dependable indicator. In this case, we should lower the weight for the noise level and raise the weight for the distance.

Another direction of future research would be to improve our SR network or to remove improper inputs for our SR network. For example, as shown in Chapter 4, some input images are not well super-resolved when the facial parts in them are undetectable. To tackle this problem, we should change our SR network or construct different databases that include facial images at different resolutions with corresponding facial shapes. Considering a real-world application, we might have to decide what are proper inputs and remove improper ones before our network processes them.

Finally, we would like to enable our network to learn end-to-end. We now use different datasets when training our SR network and our FR network. In the future, we should use a single training database of images at low resolution with corresponding facial shapes. We currently select a good pair by using the information of generated facial shapes. Accordingly, we believe that training SR parameters based on shape information could improve the quality of SR images as well as the shapes based on those SR images.

Anbarjafari, G. and Demirel, H. (2010). Image super resolution based on interpolation of wavelet domain high frequency subbands and the spatial domain input image. *ETRI Journal*, 32(3):390–394. 10

Baker, S. and Kanade, T. (2002). Limits on super-resolution and how to break them. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(9):1167–1183. 6

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., and Zisserman, A. (2018). Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic face & Gesture Recognition (FG 2018)*, pages 67–74. IEEE. 17

Chang, A. X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al. (2015). Shapenet: An information-rich 3D model repository. *arXiv preprint arXiv:1512.03012*. 8

Chen, Y., Ji, Y., Zhou, J., Chen, X., and Shen, W. (2012). Computation of signal-to-noise ratio of airborne hyperspectral imaging spectrometer. In *2012 International Conference on Systems and Informatics (ICSAI2012)*, pages 1046–1049. IEEE. 10

Choi, J., Medioni, G., Lin, Y., Silva, L., Regina, O., Pamplona, M., and Faltemier, T. C. (2010). 3d face reconstruction using a single or multiple views. In *2010 20th International Conference on Pattern Recognition*, pages 3959–3962. 9

Choy, C. B., Xu, D., Gwak, J., Chen, K., and Savarese, S. (2016). 3D-R2N2: A unified approach for single and multi-view 3d object reconstruction. In *European Conference on Computer Vision*, pages 628–644. Springer. 8

Chung, J. S., Nagrani, A., and Zisserman, A. (2018). Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*. 17

Dong, C., Loy, C. C., He, K., and Tang, X. (2015). Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307. 10

Elad, M. and Feuer, A. (1999). Super-resolution reconstruction of image sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(9):817–834. 6

Fardo, F. A., Conforto, V. H., de Oliveira, F. C., and Rodrigues, P. S. (2016). A formal evaluation of psnr as quality measurement parameter for image segmentation algorithms. *arXiv preprint arXiv:1605.07116*. 10

Farsiu, S., Robinson, M. D., Elad, M., and Milanfar, P. (2004). Fast and robust multiframe super resolution. *IEEE Transactions on Image Processing*, 13(10):1327–1344. 6

Feng, Y., Feng, H., Black, M. J., and Bolkart, T. (2021). Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4):1–13. 9, 12, 16, 17

Gecer, B., Ploumpis, S., Kotsia, I., and Zafeiriou, S. (2019). GANFIT: Generative adversarial network fitting for high fidelity 3D face reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1155–1164. 9

Hartley, R. and Zisserman, A. (2003). *Multiple View Geometry in Computer Vision*. Cambridge University Press. 7

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851. 16

Hore, A. and Ziou, D. (2010). Image quality metrics: PSNR vs. SSIM. In *20th International Conference on Pattern Recognition*, pages 2366–2369. IEEE. 10

Kang, S. B., Szeliski, R., and Chai, J. (2001). Handling occlusions in dense multi-view stereo. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001)*, volume 1, pages I–I. IEEE. 8

Kaoru, I., lchikawa Ryoichi, Masabumi, N., and Jeorge, K. (1995). Sex differences in the shapes of several parts of the young japanese face. *Applied Human Science : Journal of Physiological Anthropology*, 14(4):191–194. 9

Karras, T., Laine, S., and Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410. 15

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al. (2017). Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4681–4690. 6

Lee, S. J., Park, K. R., and Kim, J. (2011). A SfM-based 3D face reconstruction method robust to self-occlusion by using a shape conversion matrix. *Pattern Recognition*, 44(7):1470–1486. 9

Li, T., Bolkart, T., Black, M. J., Li, H., and Romero, J. (2017). Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17. 9

Lu, K., Wang, Z., and Li, X. (2011). Online 3D presentation system for goods. In *IET International Communication Conference on Wireless Mobile and Computing (CCWMC 2011)*, pages 500–502. 7

Matsuo, Y. and Sakaida, S. (2017). Super-resolution for 2K/8K television using wavelet-based image registration. In *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 378–382. IEEE. 5

Mohamad, B., Yaakob, S., A. Raof, R. A., Nazren, A., and Nasrudin, M. W. (2017). An analysis of performance for commonly used interpolation method. *Advanced Science Letters*, 23(6):5147–5150. 5

Orteu, J.-J., Garric, V., and Devy, M. (1997). Camera calibration for 3D reconstruction: Application to the measure of 3D deformations on sheet metal parts. In *Conference on Vision Systems – Algorithms, Methods, Components, and Applications of Lasers, Optics and Vision in Manufacturing*, page 12 p., Munich, Germany. 8

Pontes, J. K., Kong, C., Sridharan, S., Lucey, S., Eriksson, A., and Fookes, C. (2018). Image2mesh: A learning framework for single image 3D reconstruction. In *Asian Conference on Computer Vision*, pages 365–381. Springer. 8

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer. 7, 15

Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D. J., and Norouzi, M. (2022). Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 7, 11, 13, 16

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR. 7

Strecha, C., Fransens, R., and Van Gool, L. (2006). Combined depth and outlier estimation in multi-view stereo. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 2394–2401. IEEE. 8

Unan, M., An, J., Seimenis, I., Shah, D. J., and Tsekos, N. V. (2019). 3D reconstruction of tubular structure using radially deployed projections. In *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 322–327. 10

Vandewalle, P. (2006). *Super-resolution from unregistered aliased images*. PhD thesis, EPFL, Lausanne, Switzerland. 10

Vladimirov, I., Nikolova, D., and Terneva, Z. (2021). Overview of methods for 3D reconstruction of human models with applications in fashion e-commerce. In *2021 56th International Scientific Conference on Information, Communication and Energy Systems and Technologies (ICEST)*, pages 19–22. 7

Wang, M., Deng, W., Hu, J., Tao, X., and Huang, Y. (2019). Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 692–702. 17

Wang, N., Zhang, Y., Li, Z., Fu, Y., Liu, W., and Jiang, Y.-G. (2018a). Pixel2Mesh: Generating 3D mesh models from single RGB images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 52–67. 8

Wang, X., Xie, L., Dong, C., and Shan, Y. (2021). Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1905–1914. 6

Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., and Change Loy, C. (2018b). ESRGAN: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV) workshops*, pages 0–0. 6

Zhang, L., Zhang, H., Shen, H., and Li, P. (2010). A super-resolution reconstruction algorithm for surveillance images. *Signal Processing*, 90(3):848–859. 6

Zielonka, W., Bolkart, T., and Thies, J. (2022). Towards metrical reconstruction of human faces. *European Conference on Computer Vision*. 9