Learnable HMD Facial De-occlusion for VR Applications in a Person-Specific Setting

Thesis submitted in partial fulfillment
of the requirements for the degree of

*Master of Science in **Computer Science and Engineering** by Research*

by

Surabhi Gupta
2019701024
surabhi.gupta@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500 032, INDIA
November 2022

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled **"Learnable HMD Facial De-occlusion for VR Applications in a Person-Specific Setting"** by Surabhi Gupta, has been carried out under my supervision and is not submitted elsewhere for a degree.

Date	Adviser: Dr. Avinash Sharma
Date	Co-Adviser: Dr. Anoop Namboodiri

To my family and friends

Acknowledgments

I feel immense happiness in compiling my entire research journey into this thesis. From having no research clue to writing and publishing papers, the journey of research at IIITH has been one of the most memorable phases of my life. I would first like to express my heartfelt thanks to my research advisors, Dr. Avinash Sharma and Dr. Anoop Namboodiri, for their support and guidance in shaping my research career. I am grateful to Dr. Avinash for his constant motivation and direction in every phase of my research, which have helped me grow as an individual and an academic. His gesture of praising small milestones in my research work has always motivated me to work harder. He taught me how to handle pressure and rejections positively. Although my interactions with Dr. Anoop have been fewer, his dedication and passion for his work have forever inspired me. I thank him for including me in his projects, which proved a stepping stone to my research career.

I would also like to acknowledge and thank all my friends for making my college experience memorable and wonderful, even during the tough times of the pandemic. My sincere appreciation to my co-author, Ashwath Shetty, for his participation throughout the whole progress of this academic exploration with remarkable efforts and help. I would also like to mention Sarath Sivaprasad, Sindhu Hegde, Richa Mishra, Mounika Kalakanti, Ekta Gavas, Astitva Srivastava, Sai Sagar Jinka, Sai Amrit, Apoorva Srivastava, Shaily Mishra or their help and support and let me learn different and new information in their fields.

I want to express my special thanks of gratitude to Shubham Mante for being around and helping and supporting me throughout my master's journey. This journey would not have been possible without the love and blessings of my dearest family members. Their support and guidance have always been one of the most significant factors in all my achievements in life.

Last but not least, I would thank everyone for your support and cooperation during the data collection task done in the course of my research thesis. I would also like to appreciate all admins of Ada for providing me with computing resources, without which my research would not have been possible.

Abstract

Immersive technologies such as Virtual Reality (VR) and Augmented Reality (AR) are among the fasted growing and fascinated technology today. As the name suggests, these technologies promise to provide users with a much better experience using immersive head-mounted displays. Medicine, culture, education, and architecture are some areas that have already taken advantage of this technology. Popular video conferencing platforms such as Microsoft Teams, Zoom, and Google Meet is working to improve user experience by allowing users to use their digital avatars. Nevertheless, they lack immersiveness and realism. What if we extend these applications to virtual reality platforms so people can feel lively talking to each other? Integrating virtual reality platforms in collaborative spaces such as virtual telepresence systems have become quite popular after globalization since it enables multiple users to share the same virtual environment, thus mimicking real-life face-to-face interactions.

For a better immersive experience in virtual telepresence/communication systems, it is essential to recover the entire face, including the portion masked by the headsets (e.g., Head-Mounted Displays abbreviated as HMDs). Several methods have been proposed in the literature that deal with this problem in various forms, such as HMD removal and face inpainting. Despite some remarkable explorations, none of these methods promises to provide usable results as expected in virtual reality platforms. Addressing these challenges in the real-world deployment of AR/VR-based applications draws emerging attention. Considering the existing limitations and usability of previous solutions, we explore various research challenges and propose a practical approach to facial de-occlusion/HMD removal for virtual telepresence systems.

This thesis is well-documented to motivate and introduce the audience to various research challenges in facial de-occlusion, familiarizing them with existing solutions and their inapplicability in our problem domain, followed by the idea and formulation of our proposed solution to tackle the problem. With this view, the first chapter lays the outline of this thesis. In the second chapter, we propose a method for facial de-occlusion and discuss the importance of personalized facial de-occlusion methods in enhancing the sense of realism in virtual environments. The third chapter talks about the refinement of to previously proposed network in improving the reconstruction of the eye region. Last but not least, the final chapter briefly discusses the existing face datasets in face reconstruction and inpainting, followed by an overview of the dataset we collected for this work, from acquisition to making it usable for training deep learning models. In addition, we also attempt to extend image-based facial de-occlusion to video frames using off-the-shelf approaches, briefly explained in Appendix A.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Figure 1.1: Popular video calling applications for work, home and more. Source: News Article

Human collaboration is a vital component of any successful large corporation. Until recently, globalization has paved the way for remote communication. However, with the ongoing pandemic, virtual interaction and telepresence have almost become inevitable to meet the social needs of the people. It boosted the need for remote interaction and has positively impacted the work-life balance for the working professionals away from family and home. Traditionally, video conferencing is the most prominent solution for remote interactions, and a variety of platforms already exist for social meetings. With millions of people across the globe in different sectors, including multimedia, education, and healthcare, working remotely during Covid-19, there has been heavier reliance on existing video conferencing platforms such as Microsoft Teams[1], Zoom[2], and WhatsApp[3]. Nonetheless, these platforms

[1]https://www.microsoft.com/en-in/microsoft-teams/group-chat-software
[2]https://zoom.us/
[3]https://www.whatsapp.com/

are currently lacking regarding the visual representation of users, which hampers the lifelike experience. Users are represented by 2D images that lack immersion and do not look realistic. Figure 1.1 depicts a typical day-to-day scenario of a virtual meeting room that clearly shows the restrictive nature of these 2D-based meeting platforms. In general, multimedia applications that promise an immersive experience to users should deliver high-quality and realistic visual and audio content. However, current video conferencing platforms are restricted with 2D, thereby compromising the sense of realism and lacking immersiveness.

Figure 1.2: Applications of AR/VR in futuristic Metaverse. Sources: From Sofien Bouaziz's slides

This ongoing, unpredictable pandemic has emphasized that working from home will remain the norm for millions of employees. Studies [4, 3] suggest that since 2005, there has been a sudden hike in the number of people working remotely from home. Currently, most companies are flexible enough to provide lifetime Work From Home (WFH) to their employees, and many have preferred to work in a hybrid mode. However, in this mode of working remotely, the real-world experience is often compromised over virtual presence. Thus, it is essential to feel a sense of presence in digital meetings.

In cases where user experience is of prime importance, one of the most popular and widely preferred solutions is the ***Virtual Reality***. Virtual Reality (VR) has proven successful in enhancing the sense of social presence and interaction over 2D videoconferencing systems. It aids in bridging the digital and physical worlds by allowing you to take in information and content visually as you take in the world. VR lets you experience what it is like to go anywhere — from the front row of a concert to a business meeting room to distant planets in outer space. Moreover, with the Metaverse taking over the current technologies, the need to embrace virtual reality has become paramount to improving user experience. It has the potential and promises to become an extension of people's lives where they could opt to work, live and play continuously in real-time, and Augmented Reality (AR) and Virtual Reality (VR) are known to be the cornerstones of such projects. Figure 1.2 shows some future applications of AR/VR in Metaverse projects. Thus, it is essential to build solutions that can work flawlessly in hybrid modes, irrespective of the type of platform. So, how to get the best of both worlds? In the following section, we define our problem statement and propose a way to tackle this problem in a simple and cost-effective way.

1.2 Problem Defination

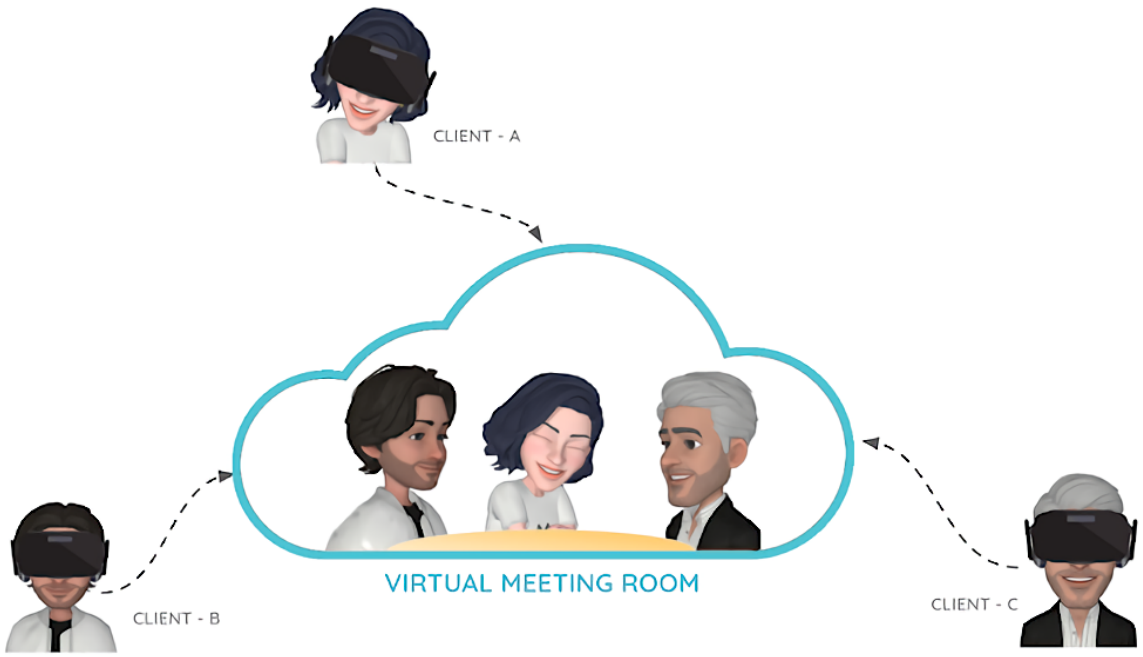

Figure 1.3: Illustration of our idea of setting up an environment with multiple people interacting with each other virtually.

With sufficient technological advancements, VR meetings may become the norm over 2D video-based meetings due to the benefits they offer in terms of users feeling a heightened social presence even post-pandemic. Furthermore, many people now prefer to work from home, especially in hybrid mode, since it gives them the privilege to spend time at home with their family besides working remotely during office hours. We attempt to build a similar solution by combining the power of virtual reality and deep learning to design an AR/VR telepresence system that can somewhat mimic real-life human interaction. We can have multiple clients connected to a virtual meeting room, interacting with each other using VR headsets. It will allow for better interactions, for example, face-to-face communication and immersive experience. Figure 1.5 illustrates the core idea of this work in setting up a realistic virtual environment that allows multiple users to interact simultaneously with a sense of being together in one place.

We address this problem using a popular computer vision technique called the image inpainting task. In earlier days, image inpainting was widely used to restore old corrupted photographs and watermark removal from images. We employ this technique in our problem statement to recover the missing region of the face occluded due to Head Mounted Displays (HMDs), where the input image corresponds to a 2D face image of a user wearing the VR headset, and the output image is a reconstructed full face, as shown in Figure 1.4.

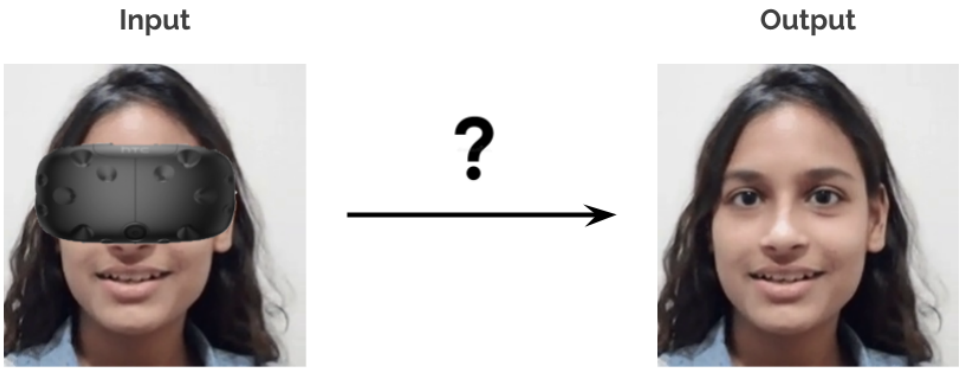

Figure 1.4: A facial de-occlusion task as an image inpainting problem.

1.3 Research Challenges

However, mimicking real-world face-to-face communication with all sorts of facial movements and expressions in the virtual world is not easy. It is a highly ill-posed problem that draws much attention to solving real-life problems. Several research challenges still exist in VR-based teleconferencing applications, some of which are yet to be addressed. Some of the common challenges include occlusion due to HMDs, low bandwidth network, low latency, expensive digital avatars, extreme headpose, poor illumination and complex expressions. Here, we discuss some of these in brief:

Figure 1.5: Major challenges in AR/VR based telepresence applications.

1.3.1 Occlusions by Head Mounted Displays (HMDs)

One of the biggest challenges with head-mounted displays/virtual reality headsets is the obstruction of certain facial features that restrict the full face's visibility. Additionally, reconstructing the full human face in the presence of occlusion is not an easy task. This is a highly ill-posed problem that is still in research for a long time. Human faces are an integral part of understanding and comprehending human behavior. Being subjective in nature, people have different facial structures and expressive styles. The region of the face occluded by virtual reality headsets contains the most detail about an individual, from tiny fine wrinkles that indicate that you like to laugh open mouth to a larger identifier that defines the shape of your eyes. In addition to this, facial features carry a wealth of social information necessary for effective communication. Thus, recovering missing facial regions with plausible and semantically consistent content with authenticity (preserving identity and expressive style) is an essential yet challenging task.

1.3.2 Low Bandwidth Network

Full motion video is considered the most complex and demanding component of video-based telepresence systems. However, one of the biggest implications of these systems is their high-bandwidth requirements for sharing data. Transmitting high-quality media such as images, graphics, and audio significantly requires higher bandwidth and good computing power for processing, especially videos. The need is even more crucial if a person interacts with others during a video call or meeting. It is desirable that the communication is smooth and continuous without any loss of quality and with low latency. A low bandwidth transmission medium with insufficient resources is unsuitable for dynamic graphics applications. This necessitates a simple yet cost-effective solution for hybrid teleconferencing systems that meet low bandwidth requirements.

1.3.3 Expensive Avatars

The advent of avatar-based animation has also proven valuable in enhancing user experience with realistic and highly detailed content. Big tech giants such as Meta and Microsoft have already started to

work on incorporating animatable avatars in their products. Microsoft is building a collaborative platform for virtual experiences called Mesh which they plan to integrate into Microsoft Teams by next year. It will allow users to ensure their presence in meetings without turning on their webcams using their own personalized and customizable avatars. However, these animation techniques need a large amount of multi-view calibrated data, expensive hardware for graphics and animation, and high computing resources for training the models. Besides, these models are restricted by licensing issues that incur extra efforts and costs. Hence, it might not be convenient for everyone to adopt it.

1.4 Research Landscape

Image inpainting is not a new field of research. It has been a prevalent area of research in the computer vision domain that has been used interchangeably with image restoration and retrieval. In olden times, image inpainting has been widely used as a technique to recover missing regions and restore corrupted images, primarily old photographs. The task of HMD removal from the face can be posed as reconstructing missing regions from face images. We shed some light on existing literature in similar areas to better understand our problem statement.

1.4.1 Generic Face Inpainting

Face inpainting (or face completion) is the task of generating plausible facial structures for missing pixels in a face image. It also refers to the task of recovering missing facial features in a given corrupted face image. This problem has been previously addressed in various forms such as caption/text/watermark removal, face mask removal, etc. [48] is a landmark guided face inpainting approach that has been trained and tested on the CelebA dataset with different types of masks. However, this method only works for frontal faces. Similarly, [34] is a generic edge-guided inpainting method that has also been validated on faces. Digitally removing face masks from an image is one example of image inpainting proposed in [22] that reconstructs the region occluded by the face mask with a natural and pleasing content.

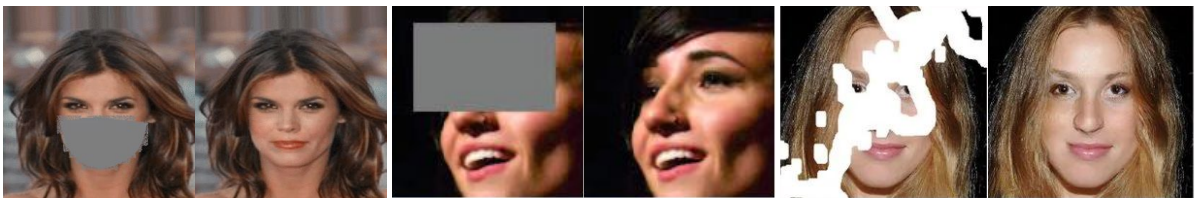

Figure 1.6: Different forms of generic face image inpainting.

1.4.2 HMD De-occlusion/ HMD Removal

Virtual reality headsets occlude a significant portion of the face that masks essential facial features such as eyes. HMD removal/de-occlusion refers to the task of removing the HMD digitally by reconstructing the missing region with plausible and relatable details. However, this is a highly ill-posed and complex problem. Several methods have been proposed in the past to recover these missing facial features with the help of a reference image. One such method has been introduced in [44, 35, 54] that utilizes an additional reference image and eye images (if available) to de-occlude HMD devices from face images. However, most of them are 3DMM [5] based methods.

Figure 1.7: Various approaches to HMD de-occlusion or HMD removal.

1.4.3 3D Model Vs Non-3D Model based Methods

Popular parametric face models such as 3D Morphable Face Models or 3DMMs were introduced for general face representation and image analysis. Due to their ability to model intrinsic properties of 3D faces, such as shape and skin texture, they have been widely used in face recognition, inpainting, computer graphics, and animations. These models have been used as a prior in HMD removal techniques such as [54, 35]. However, they are not good enough at capturing eyes, teeth, hairs, and skin details. They generate an overly smooth texture that lacks realism. [51] is a 3DMM [5] and GAN [15] based face de-occlusion method that can handle face images under challenging conditions. However, this method has not been tested on large occlusions created by HMD devices.

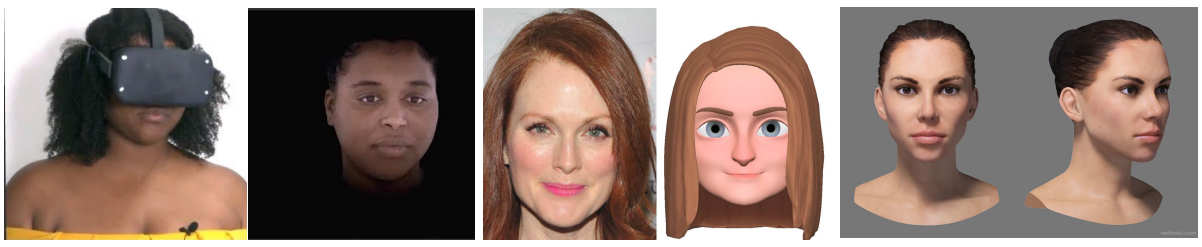

Figure 1.8: 3D model based face reconstruction methods.

1.4.4 Exisiting Face Datasets

Many publicly available face datasets exist in the literature and are widely used to train and test deep-learning models for various face-related tasks. However, depending on the task we need to perform, say face detection or landmark localization, or image synthesis, it is important that the dataset is well-prepared for use. It should include proper annotations with labels if required. Some popular and commonly used face datasets for training face image inpainting models include CelebA [30], Celeba-HQ [23], LFW [19] in images, and Vox-Celeb [33], FaceForensics [38] and Youtube faces [28] for videos. Figure 1.9 shows sample images from popular face image datasets.

1.4.4.1 CelebA

CelebFaces Attributes Dataset (CelebA) is a large-scale face dataset that covers significant pose variations and background clutter for more than 200K celebrity images, each with 40 attribute annotations. This dataset has varied variations with large quantities and rich annotations, including 202599 face images of 178x218 resolution, 10177 identities, 5 landmark locations, and 40 binary attribute annotations per image. The dataset is widely used as the training and test sets for several computer vision tasks such as face attribute recognition, face recognition, face detection, landmark localization, and face editing and synthesis.

1.4.4.2 Celeba-HQ

The CelebA-HQ dataset is another version of CelebA that consists of 30,000 face images at a higher resolution of 1024×1024. It has been introduced by Karras et al. in the paper, Progressive Growing of GANs for Improved Quality, Stability, and Variation. CelebA and CelebA-HQ are the two most widely used datasets for generic 2D face inpainting problems.

1.4.4.3 LFW

Labeled Faces in the Wild, popularly known as the LFW database, was created and maintained by researchers at the University of Massachusetts, Amherst. It is a database of face photographs prepared for studying the problem of unconstrained face recognition, consisting of 13233 images of 5749 people detected and centered by the Viola-Jones face detector and collected from the Internet. Some people pictured having two or more distinct photos in the dataset. However, the image resolution is limited to 250x250.

1.4.4.4 Vox-Celeb

VoxCeleb is a large-scale audio-visual dataset consisting of short clips of 3-5 min of human speech extracted from interview videos uploaded to YouTube. It contains speech for more than 7000 speakers

spanning a wide range of different ethnicities, accents, professions, and ages. Most of the clips are interview videos of various famous personalities.

1.4.4.5 FaceForensics

FaceForensics is a large-scale video dataset of 1004 faces containing more than 500k frames that can be used to study forgeries. All videos are downloaded from Youtube and are trimmed down to short continuous clips with mostly frontal faces detected by Viola-Jones face detector. The clips are manually screened to ensure high-quality videos with no occlusions. The videos are manipulated using state-of-the-art face editing methods. There is minimal variation in the head poses of actors.

1.4.4.6 Youtube faces

The YouTube faces is a face video dataset curated from videos uploaded from YouTube. It contains around 3425 video clips with 1595 unique people with at least two videos per person. This dataset is considered a standard dataset for face verification in videos.

Figure 1.9: Sample images from popular datasets. From top to bottom are samples from CelebA, Celeba-HQ and LFW dataset respectively.

1.5 Our Contributions

In this thesis, our key contributions focus on introducing person-specific learning-based paradigms for facial de-occlusion tasks. More specifically, we proposed a deep learning solution that uses an attention mechanism and eye-tracking priors. Finally, we also contributed with a dataset that we intend to release in the public domain for evaluating the proposed method in this field. Here we further expand on our detailed contributions.

1. Personlized facial de-occlusion network for virtual reality applications

    (a) We proposed an effective approach to facial de-occlusion/HMD removal using attention-based encoder-decoder architecture.

    (b) We also propose a mask-based loss function to enhance reconstruction in the masked region.

    (c) We also showed its application in virtual hybrid telepresence systems in low-bandwidth scenarios.

2. Enhancing facial de-occlusion using eye landmarks

    (a) To enhance reconstructions in the eye region and to capture subtle eye movements such as blinking, we used landmarks to guide the reconstruction.

    (b) We also propose a landmark heatmap prediction loss to regularize the inpainted reconstruction based on the predicted eye landmark heatmap.

    (c) We also utilize the eye images captured using modern HMD devices by providing the heatmap as an auxiliary input extracted from the eye landmarks.

3. Dataset collection and preparation

    (a) We proposed a simple yet cost-friendly approach to data collection for training our facial de-occlusion network.

    (b) We collected multiple short videos for various subjects in varying appearances, facial expressions, and head poses.

    (c) For every frame, we overlay a synthetically generated binary mask over the eye region generated using facial landmarks to create a corresponding occluded image.

Overall, we demonstrated the efficacy of our proposed facial de-occlusion framework using thorough experiments and analysis. We justified that our method yields superior results compared to other state-of-the-art face inpainting methods.

1.6 Thesis Organization

The rest of the thesis is organized as follows:

1. In Chapter 2, we investigate an ill-posed problem of facial de-occlusion of HMD/smart glasses in VR settings and proposed an image inpainting based solution.

2. Following this, in Chapter 3, we present an approach to enhance reconstruction in the eye region using the supervision of eye landmarks with modern eye-tracking HMD devices.

3. Finally, in Chapter 4, we introduce our proposed face dataset and discuss the setting up of recording device followed by the data acquisition process pin detail.

As a supplementatry, we also present an extension to our image-based facial de-occlusion network and introduces temporal smoothness to per-frames reconstructions, referred in Appendix A. We also explored exploiting temporal coherence (e.g., convLSTM framework) in order to achieve temporally consistent face de-occlusion.

Overall, we hope this thesis sheds light on various challenges prevalent when deep neural networks are deployed in practical settings. We believe that this thesis provides some ideas and baselines which can be taken up in the future to address these issues in a variety of settings with more success.

Chapter 2

Attention based Facial De-occlusion Network in VR Settings

The first chapter discussed the problem definition and various research challenges in VR-based applications. This chapter proposes a working solution for facial de-occlusion in VR settings. Traditionally, video conferencing is a widely adopted solution for remote communication, but a lack of immersiveness comes inherently due to the 2D nature of facial representation. The integration of Virtual Reality (VR) in a communication/telepresence system through Head Mounted Displays (HMDs) promises to provide users with a much better immersive experience. However, HMDs cause hindrance by blocking the facial appearance and expressions of the user. In this chapter, we propose a novel attention-enabled encoder-decoder architecture for HMD de-occlusion to overcome these issues. We also propose to train our person-specific model using short videos of the user, captured in varying appearances, and demonstrated generalization to unseen poses and appearances of the user. We report superior qualitative and quantitative results over state-of-the-art methods. We also present applications of this approach to hybrid video teleconferencing using existing animation and 3D face reconstruction pipelines [1]

2.1 Introduction

Globalization has led to an acute need for tele-interactions for effective communication that has been further boosted due to the current pandemic situation across the world. Traditionally, video conferencing is a widely adopted solution for telecommunication, but it lacks realism due to the 2D nature of facial representation. Virtual Reality (VR) based telepresence system provides a better immersive experience for remote conversation and collaboration. Nevertheless, HMDs significantly occlude the user's face, hindering facial appearance capture, including gaze and expressions. Therefore, HMD removal in images is vital for improving the user experience.

Traditionally, Analysis-by-Synthesis techniques for HMD de-occlusion proposed in the literature [42] animate parametric face models such as 3DMMs [5] using features extracted from an HMD occluded input image. However, such models often generate overly smooth geometrical details and compromise the realism of facial appearance. On the contrary, recent facial avatar-based models [31] achieve photorealistic

[1] Please refer to supplementary videos available on this link for better understanding and comparison.

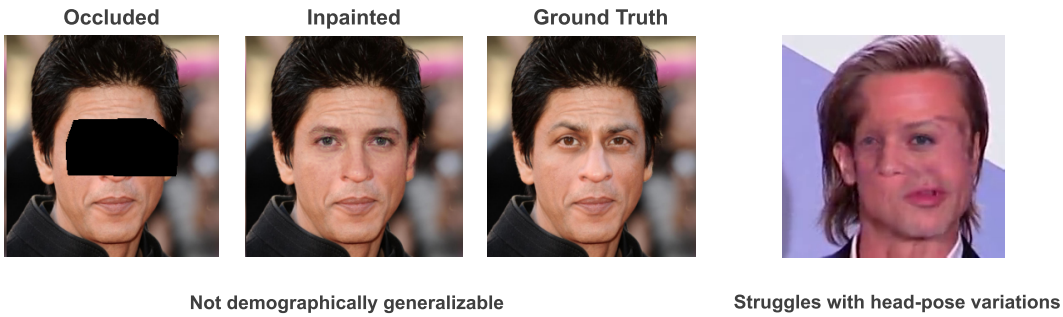

Figure 2.1: Failure cases of LaFIn [48]

results for HMD de-occlusion. However, avatar modeling methods require a large amount of calibrated multi-view data of a single user in different poses and expressions for avatar creation. [31] uses a setup consisting of 40 machine vision cameras capable of synchronously capturing 5120 $\times$ 3840 images at 30 frame per second (FPS). Thus, such avatar creation is non-trivial and challenging for a large user base to scale up. Additionally, it is a one-time process for each user. Therefore, such parametric model or avatar-based techniques have a significant limitation: they lack the user's actual appearance during the interaction (i.e., unable to model the everyday appearance of the user), hindering user experience.

HMD de-occlusion problem can also be posed as a face completion/inpainting problem. Existing face completion methods in literature [44, 48] attempt to learn a single inpainting network over a large training population, hoping for good generalization on unseen face images. However, these methods frequently suffer from issues like loss of identity and fail to generalize with even minor variations in head pose, as shown in Figure 2.1. Another set of methods in [35, 44] uses a reference image along with the occluded image to fill a masked region and preserve identity. However, as shown in Figure 2.1, their work fails to generalize with non-frontal head-poses. [35] requires additional information (depth and mask) and does not generate the entire face (with hairs, ears, and background), which hinders user experience.

The primary research challenge with the face completion/inpainting task comes from its ill-posed nature as a significant part of the face is occluded by HMD. Learning a single common face de-occlusion network with the capability to hallucinate diverse expressions in varying appearances and head poses across a large set of human faces is difficult to achieve. It is due to the broad space of facial geometry and appearance as well as the highly subjective way of articulating expressions/emotions across individuals [6]. Additionally, in the context of VR teleconferencing applications, the desired solution should be scalable, requiring minimal efforts and hardware at the user's end. An additional desired characteristic might be regarding integration ability in a hybrid VR teleconferencing setup where users with only video capability should also participate as in regular video conferencing.

To overcome these challenges, we propose to tackle this problem in a person-specific setting where we aim to train a dedicated model for each user to learn user-specific appearance, head-pose, gaze, and facial expression traits. The input to our method is a video frame with HMD occluded face, and our model

Figure 2.2: Our proposed approach can reconstruct high-quality unoccluded image from a given occluded face image.

generates a de-occluded plausible face by removing the occlusion. To achieve this, we introduce novel attention enabled encoder-decoder architecture and a novel training strategy to train our person-specific model using short videos (1-2 minutes) of the user. The video captures the varying appearance of the user with a variety of head poses and facial expressions without HMD occlusion. As a part of our training strategy, we first train the encoder-decoder module (sans attention) on large face image datasets to learn generic face appearance features. Subsequently, we finetune it on unoccluded user videos. Finally, we finetune our full model (encoder-decoder with attention) on the same training data with synthetic HMD masks. It allows our model to learn the person-specific facial geometry and expression traits and help generate occluded areas with varying appearances, poses, and expressions.

Our attention module allows the network to preserve the high-frequency appearance and background details (like hairs, wall texture, etc.) from the unoccluded part of the input HMD occluded image while generating the plausible facial appearance for the occluded part. Our novel mask-loss function helps the model to emphasize the occluded region. Figure 2.2 shows high-quality de-occlusion achieved by our method. It is important to note that learning a person-specific model is not equivalent to overfitting on a specific user as there is a significant change in user appearance, background, and lighting across sessions. Similar person-specific model learning has been successfully explored in this context [31] as well as another related context of egocentric frontal face recovery [13].

We conduct thorough empirical evaluation and report superior qualitative and quantitative results of our proposed method w.r.t. state-of-the-art methods. In addition to this, we also demonstrate the broader applicability of our proposed HMD de-occlusion method with two real-world use cases. First, we can use it to build hybrid VR systems by integrating it with video-driven face animation solutions such as [40, 45]. We also show how to integrate our method with a 3D face reconstruction pipeline to generate 3D face video for the VR teleconferencing system as a second use case.

To summarise, we make the following contributions in this chapter:

1. We present a deep learning framework for person-specific HMD de-occlusion. Our method does not rely on high-end hardware (e.g., HMD with gaze tracking) and calibrated data from the user (for avatar creation).

2. We introduced a novel attention module knitted with the encoder-decoder architecture that can use background and appearance details from the input image and allows the model to focus on inpainting the occluded region with plausible details.

3. We collected a small dataset of multiple users in different appearances, facial expressions, and head-poses that we intend to release to the academic community.

4. We present applications of our model where it can be integrated with neural animation models such as [40, 45] to animate an occluded video and can also be used to recover 3D face from occluded input.

2.2 Related Work

Our approach is an inpainting method that learns to fill in user-specific details. It is related to traditional inpainting methods and recent approaches that use a reference image to fill an occluded region faithfully. In the following section, we discuss the most relevant literature in detail.

2.2.1 Person Specific Models

Recent deep-learning advances in vision and graphics have led to the rise of personalized models that are animated/rendered using deep learning. The closest work to ours is [13]. They use a video-to-video GAN [15], which takes in egocentric frames of a person and generates the corresponding frontal view. These methods show the ability of person-specific models to capture high-frequency details. [31] learns auto-encoder network to predict view conditioned texture and mesh geometry from HMD occluded input. [14] trains a dynamic neural radiance field model on a short video (2-3 min) of the person with different expressions and poses. However, they require calibrated multi-view data from the user, adding additional hardware constraints. Animating these models is also expensive. Our approach requires a short uncalibrated video from the user for training and is considerably lightweight compared to avatar-based methods.

2.2.2 Image Inpainting Methods

Image inpainting describes the task of filling missing image regions with realistic content. Recent works [49, 50, 48] train a conditional GAN [21] on a face dataset as a solution to this problem. These

methods show impressive generalization to examples with frontal head pose and arbitrary occlusion. Another method, EdgeConnect [34] fills the missing region using edges as prior. However, these methods are biased to their training distribution as they do not generalize well to even slightly non-frontal head poses and suffer from a loss of identity. We train our method only on images of the same person with considerable variations in head poses and expressions that overcome the challenges of identity loss and generalization to various head poses.

2.2.3 HMD Removal

Exemplar guided image inpainting methods such as [35, 44] propose an image-based approach to HMD de-occlusion. They use a reference image to guide the inpainting procedure and learn a general model for the task. However, [44] fails to work well with cases of significant pose variations between the reference and occluded image. Also, [35] train and evaluate on synthetic data with additional depth information, which may not work or be available in a real-world teleconferencing scenario. We train and evaluate our model on real-world conversations and scenarios and show our model's ability to generalize to unseen appearances.

2.3 Methodology

2.3.1 Overview

The primary focus of our work is to learn a personalized model for face de-occlusion, particularly as an application in VR teleconferencing, where the face is partially occluded due to HMD. To tackle this, we formulate the face de-occlusion problem as an image inpainting task. Given an occluded face image as an input X_{occ}, our network aims to hallucinate the missing region with plausible and perceptually consistent facial details in order to reconstruct the generated unoccluded image, X_{rec} against the ground truth unoccluded image, X_{gt}.

Inspired by the autoencoder architecture proposed in [7], we use a novel attention enabled encoder-decoder framework with generative capabilities that learns to reconstruct high-fidelity unoccluded faces from HMD occluded input images. Additionally, we also propose a novel mask-based loss function and a novel training strategy to learn our model. Figure 2.3 shows the outline of our proposed architecture.

2.3.2 Proposed Architecture

2.3.2.1 Encoder-Decoder Module

Our encoder-decoder module comprises a stack of ResNet and inverted ResNet blocks. Each ResNet block consists of a set of convolutions with residual connections. For the inverted ResNet block, the first convolution in the ResNet block is replaced by a 4×4 deconv layer. We also provide the additional

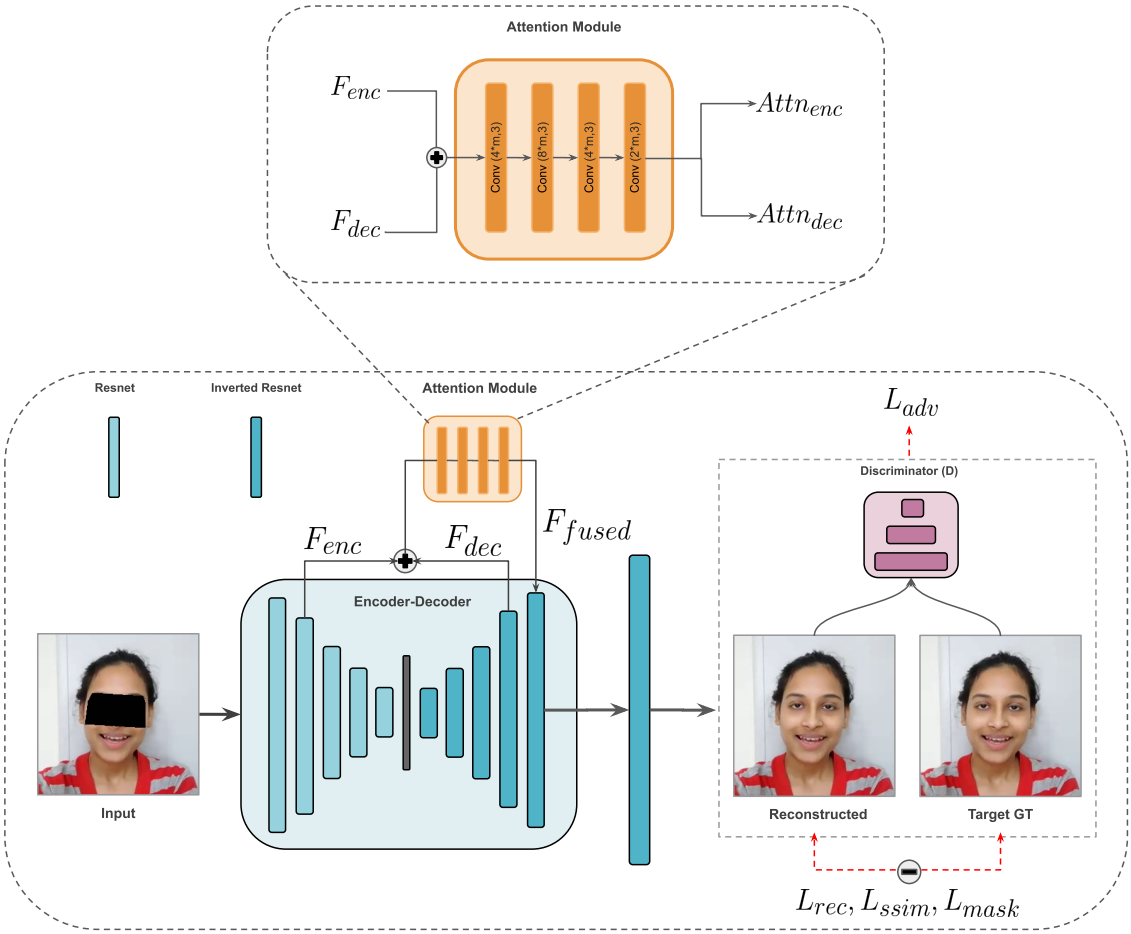

Figure 2.3: An overview of our proposed facial de-occlusion network.

generative capability to the network using an adversarial loss. The encoder learns a 256-dimensional feature representation of the input image. This bottleneck representation is subsequently fed to the decoder network to reconstruct the target image.

2.3.2.2 Attention Module

Inspired from existing literature on attention-based learning strategies as proposed in [39, 25], we append our encoder-decoder module with an attention module. We perform spatial attention by taking the encoder output from the second layer, F_{enc} of spatial dimension 64×64 and perform a channel-wise concatenation with the corresponding decoder layer output F_{dec} of the same spatial dimension. We feed it to our attention module, which subsequently generates attention maps of the same dimension as shown in Figure 2.3. We then decouple these attention maps and use them for a weighted fusion of respective feature maps (i.e., F_{enc} and F_{dec}). The fused feature maps are fed downstream to convolution layers to reconstruct the de-occluded face image.

Such attention-based spatial feature map fusion allows our network to preserve high-frequency appearance/background details (like hairs, wall texture, etc.) from the visible part of the input image while generating a plausible facial appearance for the occluded part. These attention maps can be learned using fully convolutional networks. As shown in Figure 2.3, the attention module consists of $Conv(4m, 3)$, $Conv(4m, 3)$, $Conv(8m, 3)$ and $Conv(2m, 3)$, where m denotes the base number of filters and $Conv(m, k)$ denotes a convolutional layer with output number of channels m and kernel size k. The final output with $2 * m$ channels is then split into two, $Attn_{enc}$ and $Attn_{dec}$, each of m channels and spatial dimension of 64×64. This acts as a attention mask for the inputs, F_{rec} and F_{dec} which is then fused again using a channel-wise summation according to Equation 2.1.

$$F_{fused} = F_{enc} \times Attn_{enc} + F_{dec} \times Attn_{dec} \tag{2.1}$$

2.3.2.3 Loss Function

We employ a combination of four different loss functions as our training objective. In order to penalize reconstruction errors, we use pixel-based $L1$ loss.

$$L_{rec} = \|X_{gt} - X_{rec}\|_1 \tag{2.2}$$

However, using only $L1$ reconstruction loss produces blurry outputs. To overcome this, we add a discriminator, D in the architecture to compute the adversarial loss. This adversarial loss term forces the encoder-decoder to reconstruct high-fidelity outputs by sharpening the blurred images. For D, we adopt the architecture of the DCGAN discriminator [37].

$$L_{adv} = log(D(X_{gt})) + log(1 - D(X_{rec})) \tag{2.3}$$

We also use SSIM based structural similarity loss, as defined in [7], that helps to improve the alignment of high-frequency image elements to stabilize the adversarial training.

$$L_{ssim} = SSIM(X_{rec}, X_{gt}) \tag{2.4}$$

To further improve the quality of reconstruction in the HMD occluded area of the generated image, we propose a novel *mask-based loss*. Here, we use the binary mask image as an additional supervision to the network along with input image while training. Minimizing this loss helps the model to emphasize more on quality of reconstruction in the masked region. This also helps to mitigate the blinking artifacts around the eye region for stable reconstructions.

We formulate the mask-based loss function as:

$$L_{mask} = \|I_{mask} \odot X_{gt} - I_{mask} \odot X_{rec}\|_1 \tag{2.5}$$

where, I_{mask} refers to single channel binary mask image where white pixels (1) correspond to occluded region and black pixels (0) correspond to the remaining unoccluded region and $\odot$ is element-wise multiplication. Thus, the final training objective loss function can be written as,

$$L_{final} = \lambda_{rec} * L_{rec} + \lambda_{adv} * L_{adv} + \lambda_{ssim} * L_{ssim} + \lambda_{mask} * L_{mask} \tag{2.6}$$

where, $\lambda_{rec}, \lambda_{adv}, \lambda_{ssim}$ and λ_{mask} are the corresponding weight parameters for each loss term.

2.3.3 Our Training Strategy

For learning a person-specific model, we adopt a two-step training process. We train the first and second steps on the person's unoccluded and occluded face images, respectively. In the first step, we freeze the attention module and only train the encoder-decoder to reconstruct the unoccluded images. We start with unsupervised training of only encoder-decoder on publicly available state-of-the-art face datasets such as VGGFace [8] and AffectNet [32] to leverage the inherent knowledge about the face structure. It enables the model to grasp knowledge about the basic facial structure and features such as eyes, nose, mouth, etc. Thus this step is common for all users. We then perform finetuning on users' images with a wide range of pose, expression, and appearance variations. This step helps the model learn the exact geometry of the user's face. We unfreeze our attention module in the second step and train the entire architecture on occluded images. It can be considered a self-supervised learning approach, where the input is an occluded image, and the target image is its corresponding unoccluded image. We, therefore, minimize the loss between the reconstructed face image and the unoccluded ground truth image. It enables the attention module to learn to retain the high-frequency details from the visible part of the occluded input image while performing a soft fusion with the generated image. Thus, our two-step training strategy yields superior de-occlusion with no explicit boundaries between occluded and visible regions of the reconstructed image.

2.4 Experiments and Results

2.4.1 Dataset

Our method uses monocular RGB video sequences. Hence, we captured various human subjects (around 20) in different appearances at a 1280 $\times$ 720 pixels resolution with a 30 FPS frame rate from a mobile phone camera. We collected 4-5 video sequences for each user, each of a length of around 1-2 min, i.e., approximately 8k-9k frames in total. Frames are cropped and scaled to 256 $\times$ 256. We use mutually exclusive sets of these video sequences from the same subject in different appearances to train and evaluate our user-specific model. The subjects were asked to engage in a day-to-day conversation and demonstrate variations in head poses.

It is important to note that this data is captured without any occlusion to the eye region to create ground-truth data for training and evaluation purposes. Thus, we add a synthetic mask that simulates an HMD for each video frame around the eye region. The placement of this binary mask is guided by the facial landmarks and placed over the eye region to occlude the face, which yields synthetic mask data

with ground-truth. During inference on real-world HMD occlusions, we first detect the HMD/smart-glass in the input video frames and replace it with a binary mask depicting the region that needs inpainting.

2.4.2 Implementation Details

For step one of training strategy (see Section A.3.3), we train the encoder-decoder architecture on unoccluded images with three loss functions (i.e., Equations 3.1, 3.2, 3.3) in a stage-wise manner, with each loss term being added in the training objective with every stage. We choose a batch size of 50 and an input resolution of 256×256. We train the network for 300, 100, and 300 epochs, respectively, for each loss function's incremental addition. For step two, we similarly train the encoder-decoder architecture with an additional attention module with mask loss (Equation 3.4) on occluded images for 300, 100, and 200 epochs, respectively, for each of the incremental addition of loss functions. We use the Adam optimizer [24] with a constant learning rate of 0.00002. We use $\lambda_{rec} = 1, \lambda_{adv} = 0.25, \lambda_{ssim} = 60$ and $\lambda_{mask} = 1$.

2.4.3 Evaluation Protocols

We choose SSIM (Structural Similarity Index Measure [46], PSNR (Peak Signal-to-Noise Ratio) [18] and LPIPS (Learned Perceptual Image Patch Similarity) [53] as our evaluation metrics for quantitative comparisons. For SSIM and PSNR, higher the value better the reconstruction quality, and for LPIPS, lower the value better the perceptual quality.

2.4.4 Quantitative & Qualitative Results

For qualitative evaluation, we evaluate our method with real occlusions, such as smart-glass, widely used in VR/AR applications. We overlay the area surrounding the smart glass with a synthetic mask generated (Section 2.4.1).

Figure 2.6 shows that our approach produces naturally-looking de-occluded faces that are semantically consistent with other frames in the sequence. In contrast, other state-of-the-art image inpainting methods like DeepFillv2 [50], LaFIn [48], EdgeConnect [34], when fine-tuned on images of the same user in different expressions, poses, and appearances, generates poor reconstruction results. As shown in the red box, these methods have a noticeable discrepancy between the left and right eyes. There are overlapping artifacts around the eye region indicated by a blue box. The yellow patch shows the inconsistency in skin color between the hallucinated and the rest of the face. This visual comparison strongly supports our idea of using a person-specific training approach rather than the generalized method since they do not guarantee to preserve identity and other high-frequency details such as appearance, pose, and expressions across frames. We also report de-occlusion results using our method in varying expressions and head poses in Figure 2.5 to verify the generalizability of the proposed model.

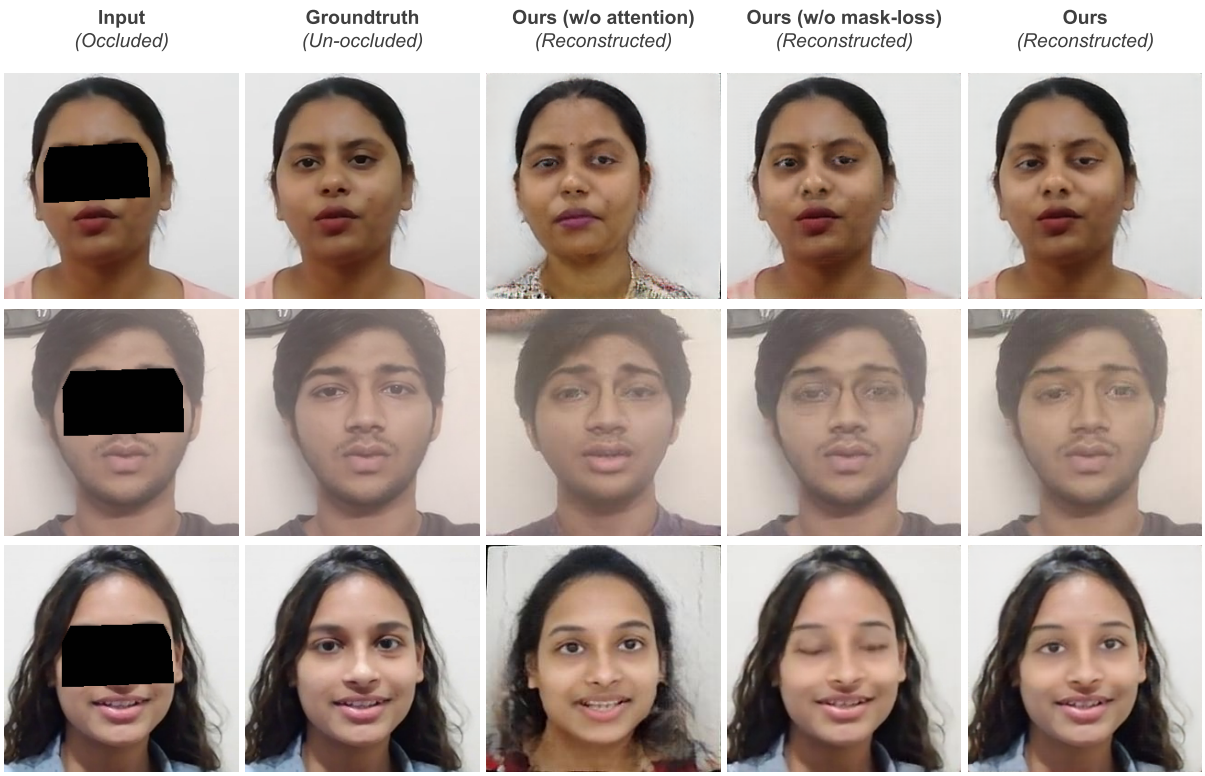

Figure 2.4: Visual results on unseen appearance demonstrating the effect of using attention and mask-loss. From left to right, third column shows results of our method without attention and mask-loss, fourth column shows results with only attention and fifth column shows results with both attention and mask-loss.

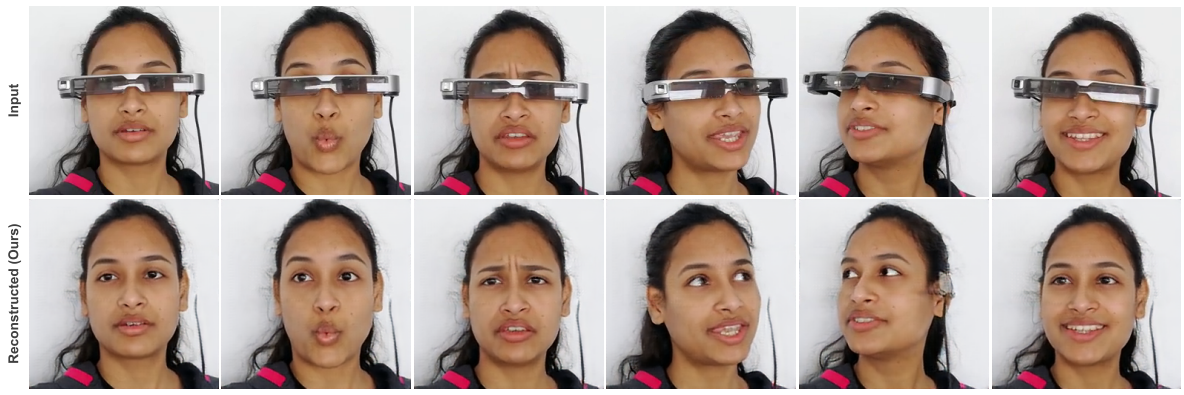

Figure 2.5: De-occlusion results using our method with large variations in head poses and expressions.

Table 2.1: Quantitative comparison with other methods on face reconstruction.

Method	SSIM↑	PSNR↑	LPIPS↓
LaFIn [48]	0.914	23.693	0.0601
EdgeConnect [34]	0.908	23.10	0.0689
DeepFillv2 [50]	0.845	19.693	0.117
Ours (w/o attention)	0.706	19.627	0.176
Ours	**0.938**	**30.59**	**0.029**

Table 2.1 reports the quantitative evaluation indicating the benefit of our approach for unseen appearances. It is important to note that we computed these quantitative results on data with a synthetic HMD mask to have a ground-truth to compare with. We can observe that our method achieves superior results in terms of all evaluation metrics. Though our method seems to perform only marginally better than LaFIn [48] and EgdeConnect [34] in terms of SSIM and LPIPS measures, there is a significant difference in terms of qualitative results. DeepFillv2 [50] fails poorly in hallucinating the missing region, thus reporting higher LPIPS and lower SSIM, PSNR compared to our method.

Table 2.2: Ablation study on different loss functions.

Method	SSIM↑	PSNR↑	LPIPS↓
Ours (L_{rec})	0.932	30.28	0.067
Ours (L_{rec}+L_{adv})	0.916	29.37	0.042
Ours (L_{rec}+L_{adv}+L_{ssim})	0.936	30.37	0.031
Ours (L_{rec}+L_{adv}+L_{ssim}+L_{mask})	**0.938**	**30.59**	**0.029**

2.4.5 Ablation Study

We also did an ablation study on the various loss functions used in our method. We observed that training only with L_{rec} generates a blurry image, whereas the addition of L_{adv} introduces more sharpness. Finally, L_{ssim} and L_{mask} make it more consistent with facial features in the original image. Table 2.2 reports quantitative evaluation indicating incremental importance of all loss functions in our formulation.

Figure 2.4 qualitatively shows the effect of adding attention and mask-loss into the network. We use unseen test examples with different appearances (not part of the training set). Without attention, we can observe that the model cannot capture the user's appearance and background details as it tries to

Figure 2.6: Qualitative comparison with SOTA inpainting methods on real-world occlusion (smart-glass). Zoom in for better details.

hallucinate it from the training examples, whereas introducing the attention into the network allows the model to use the high-frequency information from the input image. Furthermore, adding the mask loss introduces more consistency in the hallucination of the masked region. As can be observed in row 3 of Figure 2.4, introducing the mask allows the eyes to be open as it is with the ground truth face image.

Table 2.3: Ablation study on different dimensionality of z-vector.

#Dims.	SSIM↑	PSNR↑	LPIPS↓
99	0.918	29.025	0.042
256	**0.938**	**30.59**	**0.029**
512	0.935	29.12	0.031

Table 2.3 reports an ablation study on the dimensionality of z-vector. We achieved better results with $d = 256$.

2.5 Application to Hybrid Telepresence System

Recent works on face video animation, such as [40, 45], demonstrate that by just using sparse landmarks, a face image can be animated reasonably well from a reference image and show its application in low-bandwidth environments. We can easily integrate their method in our setup by first de-occluding the HMD followed by extracting reliable sparse landmarks for facial animation as shown in Figure 2.7. Thus, we can generate a consistent 2D video feed from the input-occluded video feed. Moreover, this animated face can also be used for per-frame 3D face reconstruction tasks [52] and fed to other VR teleconferencing users wearing a VR headset (as shown in Figure 2.8). Hence, our method allows VR and non-VR users to share a similar experience in a single hybrid teleconferencing application. For better understanding, please refer to video result 1.

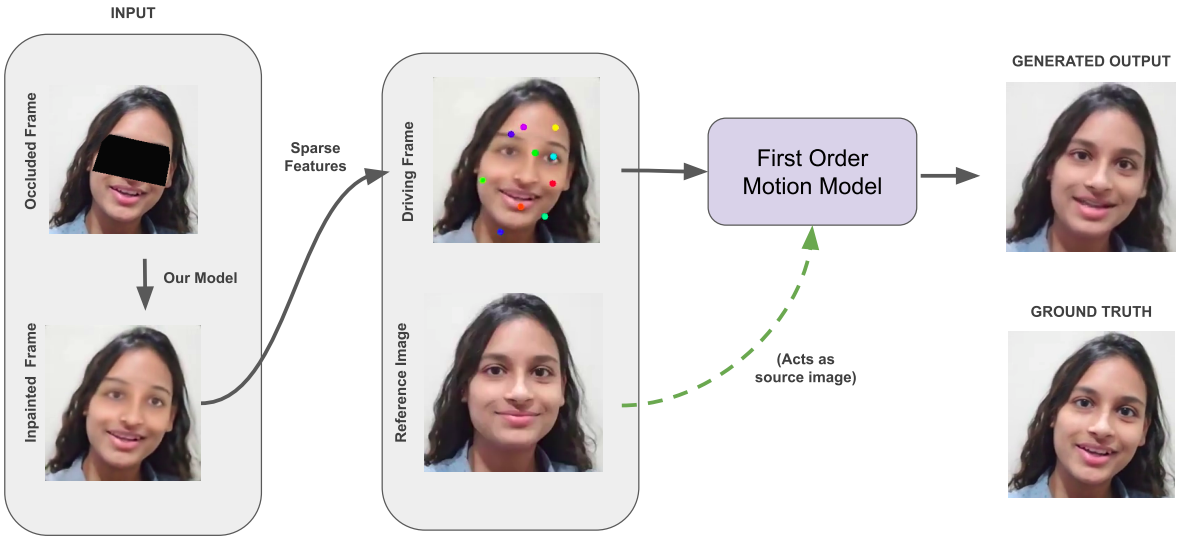

Figure 2.7: 2D de-occlusion using our method followed by the facial animation using FOMM [40].

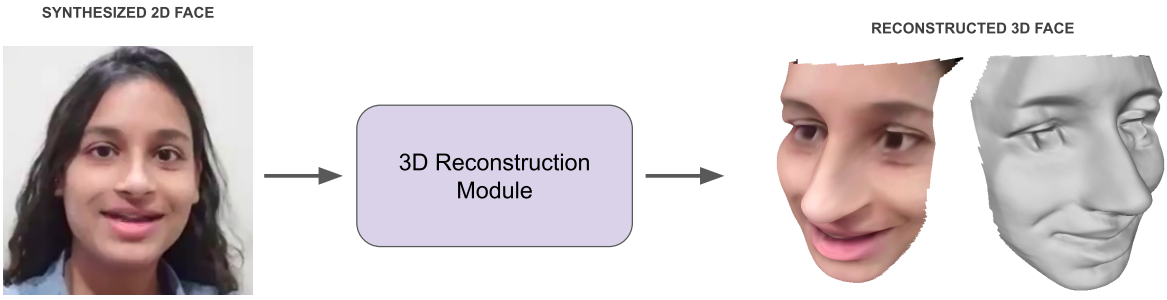

Figure 2.8: 3D reconstruction of de-occluded frame using DF2Net [52].

2.6 Discussion

As shown earlier, our proposed approach promises to give superior results, both qualitatively and quantitatively. We also notice that if HMD occlusion is more than 70 percent of the face, it becomes difficult for any existing methods to reconstruct plausible faces for varying expressions and head poses. Hence, these models can reconstruct canonical faces, which might not always be the case in telepresence systems.

In the scope of this chapter, our model does not explicitly handle eye movements since we are not providing any strong priors such as landmarks. Thus, it might not be able to capture accurate eyelid movement during blinks. However, we can easily incorporate eye tracking and gaze information for further refining our results. In the next chapter, we will focus on improving our proposed HMD de-occlusion by leveraging landmark information from modern HMD devices with eye-tracking capability and providing this as supervision to the model. This supervisory signal guides the network to produce stable reconstruction in the eye region consistent across frames.

2.7 Conclusions

As a conclusion to this chapter, we propose to learn a personalized model for face de-occlusion, particularly as an application in VR teleconferencing, where the face is partially occluded due to HMD. We formulate the face de-occlusion problem as an image inpainting task. Our proposed attention enabled encoder-decoder network takes an HMD occluded face as input and completes missing facial features, particularly the eye region. The experiments show that our method works reasonably well with the same person wearing different clothes, facial appearances, poses, and expressions. Experiments show that our proposed method reports superior qualitative and quantitative results over state-of-the-art methods.

Chapter 3

Enhanced Facial De-occlusion Network with Landmark Supervision

In the previous chapter, we presented an approach to facial de-occlusion using a deep-learning model. This chapter introduces a way to refine the previous method by leveraging modern HMD devices with eye tracking to handle complex eye motions. Face possesses a rich spatial structure that can provide valuable cues to guide various face-related tasks. The eyes are considered an important socio-visual cue for effective communication. They are an integral feature of facial expressions as they are an important aspect of interpersonal communication. However, virtual reality headsets occlude a significant portion of the face and restrict the visibility of certain facial features, particularly the eye region. Reproducing this region with realistic content and handling complex eye movements such as blinks is challenging. Previous facial inpainting methods are not capable enough to capture subtle eye movements. In view of this, we propose a working solution to refine the reconstructions, particularly around the eye region, by leveraging inherent eye structure. We introduce spatial supervision and a novel landmark predictor module to regularize per-frame reconstructions obtained from an existing image-based facial de-occlusion network. Experiments verify the usefulness of our approach in enhancing the quality of reconstructions to capture subtle eye movements. [1]

3.1 Introduction

Social telepresence and interaction are essential for human survival. Since globalization, there has been a considerable increase in the number of users interacting remotely, which has further witnessed a hike during the Covid-19 pandemic. With this unprecedented situation prevailing worldwide, there has been an inevitable need for remote interaction and communication. Working professionals have been pushed to either work entirely from home or adopt a hybrid working style to attend meetings. Thus, it is essential to feel interactive and lively during digital meetings. Traditional video conferencing platforms such as Microsoft Teams, WhatsApp, and Google Duo have gained immense popularity during the pandemic. However, they lack immersiveness and compromise realism that impacts the user's experience, which is undesirable. Even post-pandemic, these 2D applications will likely become a standard for social

[1] Please refer to supplementary videos available on this link for better understanding and comparison.

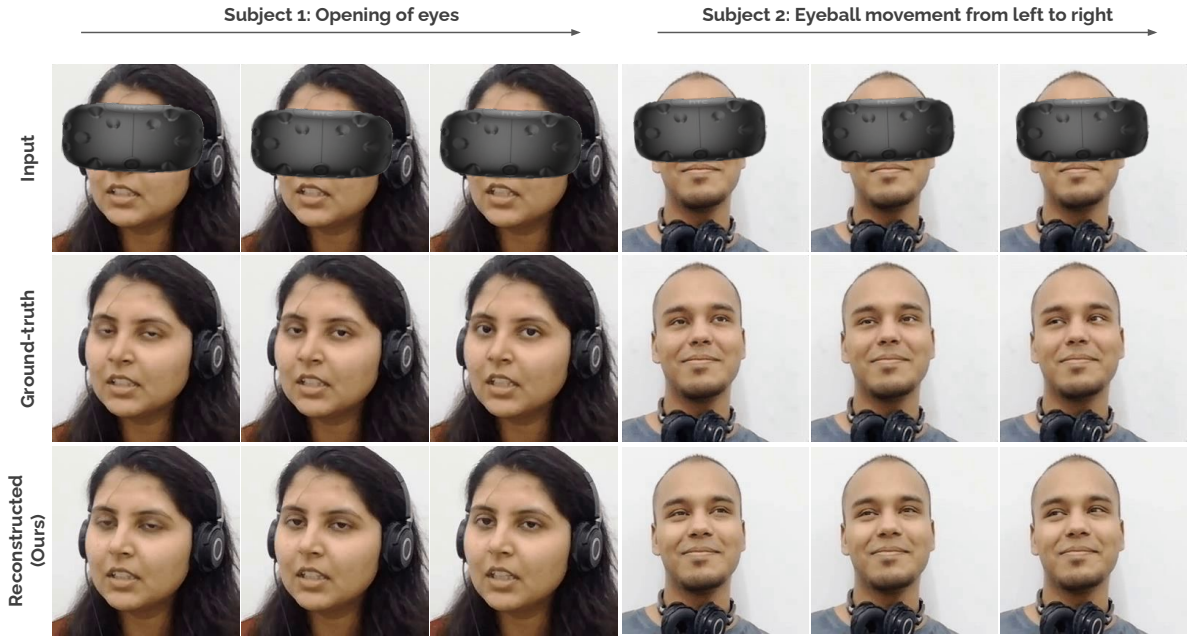

Figure 3.1: This figure shows the photo-realistic results generated by our proposed facial de-occlusion network, targeting complex eye motions.

interaction and meetings, thus necessitating the need to upgrade current technologies to meet the growing demands of people.

With the integration of virtual reality in a communication platform, the current technologies have witnessed a breakthrough in enhancing user experience with a sense of heightened social existence and interaction. Faces convey vital socio-visual cues that are important for effective communication. However, one of the major challenges with virtual reality, such as HMDs, is the occlusion it creates over the face when wearing these devices. These devices occlude almost 30-40 percent of the face, obscuring essential social cues, particularly the eye region, which hinders the user's experience. Several approaches have been proposed in the literature to tackle this problem, but none of them produces usable results that could be integrated into hybrid telepresence systems.

Existing face image inpainting approaches often suffer from incoherency in generating smooth reconstructions when applied to video frames. This incoherency is highly noticeable in the eye region, which is undesirable. It is generally visible as jittering in eyelids in successive video frames. Specific eye movements, such as blinking, are usually involuntary act in humans that is natural and unavoidable. Thus, it is important to retain this characteristic for effective communication. Synthesizing eyes, including iris and eyelid reconstruction with appropriate eye gaze, have been attempted before using 3D model-based approaches. Nonetheless, they require high-quality data and incur expensive training costs. [35] are 3D models based approaches to HMD de-occlusion. However, they work only with frontal face images and fail in cases of extreme head-poses.

Interestingly, one of the biggest advantages when dealing with digital face images in computer vision is its rich spatial structure. In digital images, this structure is generally represented in the form of 2D/3D coordinates, heatmaps, and edges and is provided as an auxiliary input to the network. Many works in the literature have exploited these spatial constraints for achieving better quality reconstruction in face inpainting and generation tasks. Recently, [47] proposed image-to-face video inpainting using spatio-temporal nested gan architecture. They used 3D residual blocks to capture inter-frame dependencies. The authors showed that conditioning facial inpainting on landmarks yielded stable reconstructions. Nonetheless, it is only validated with a specific type of circular mask that covers the eye, nose, and mouth. [48] is another face image inpainting method that is guided using facial landmarks. Personalized facial de-occlusion networks such as [17] have been proposed in the literature to generate plausible reconstructions. However, they are not controllable and thus cannot handle eye movements. Thus, we tackle this problem using an image generation/ image synthesis approach applied to faces using additional information that is easily accessible using modern HMD devices with eye-tracking capabilities.

This work aims to generate high-quality facial reconstructions in and around the eye region with consistent eye motions in the presence of occluders such as HMDs. Our primary focus is to handle instability during eye movements that are noticeable mainly around the eye region. For this, we leverage the spatial property of faces, i.e., facial landmarks, to guide the model to synthesize the eye region with minimal artifacts that look realistic and plausible. Figure 3.1 presents high-quality and photo-realistic results produced by our method showing the efficacy of our approach of using spatial supervision to control complex eye motions such as eye blinks and rolling of eyeballs.

To summarize, we make the following contributions:

1. We propose a potential solution to refine the reconstructions in the eye region.

2. To achieve this, we leverage the spatial constraints such as landmarks to improve upon consistency in the eye region by feeding eye landmarks heatmaps as an auxiliary input to the network along with occluded face image.

3. To further improve the fidelity of the reconstruction, we use an additional loss function to regularize the training based on the landmarks.

3.2 Related Work

To see what is not present in the image is one of the most exciting yet challenging tasks in computer vision. We often refer to it as image restoration/image inpainting in the digital domain. It has applications in medical image processing, watermark removal, restoring old photographs, and object removal. Inpainting has been an active research topic for many years, and several works have been proposed in the literature. Recently, this area has seen tremendous interest in image synthesis/image completion in AR/VR. This section will discuss the most relevant existing works in detail.

3.2.1 Facial De-occlusion and HMD Removal Methods

De-occluding face images in the presence of large occluders such as HMDs is highly an ill-posed problem. Several works such as [35], [44] have been proposed in the literature to address this issue. However, none promises to provide usable results in practice as these have only been validated for frontal face images with rectangular masks. Since they use an additional reference image of the person, they fail in cases of different pose variations between occluded and reference images. Recently, [17] presented an approach to tackle the problem of facial de-occlusion by training a person-specific model in VR settings. It generates plausible and natural-looking reconstructions but might fail to maintain smooth eye movements across consecutive frames. To address this issue, we can use extra information provided by modern HMD devices equipped with eye-tracking to generate consistent eye motions.

3.2.2 Structure-guided Image Inpainting

Figuring out missing regions without any prior information is a difficult task. Many previous image inpainting methods, such as [34], [48] use edges, landmarks, and other structural information as an auxiliary input to guide the reconstructions. This extra supervision has proven effective in helping the model fill the missing region with appropriate content. Nonetheless, these are image-based approaches and might not guarantee to generate consistent results across frames. Thus, we cannot directly use these methods to generate smooth reconstructions, particularly in the eye region.

[12], a closed-to-open eye in-painting eye inpainting framework based on GAN architecture has been proposed in the literature. It uses a reference image of the person to inpaint the eye region such that the closed eye of the person in the input image is reconstructed to be open.

3.3 Proposed Method

This chapter primarily focuses on improving the synthesis of the eye region in the per-frame reconstructions from our facial de-occlusion network, which was explicitly not handled in the previous chapter.

3.3.1 The Architecture

We built upon the architecture as proposed in [17] and use an attention-enabled encoder-decoder architecture followed by a novel Landmark Heatmap Predictor (LHP) that acts as a regularizer to enhance the reconstruction in and around the eye region. We train this network in an end-to-end fashion in two stages using a dedicated loss function. We consider [17] as our baseline method. Figure 3.2 illustrates an overview of the proposed pipeline.

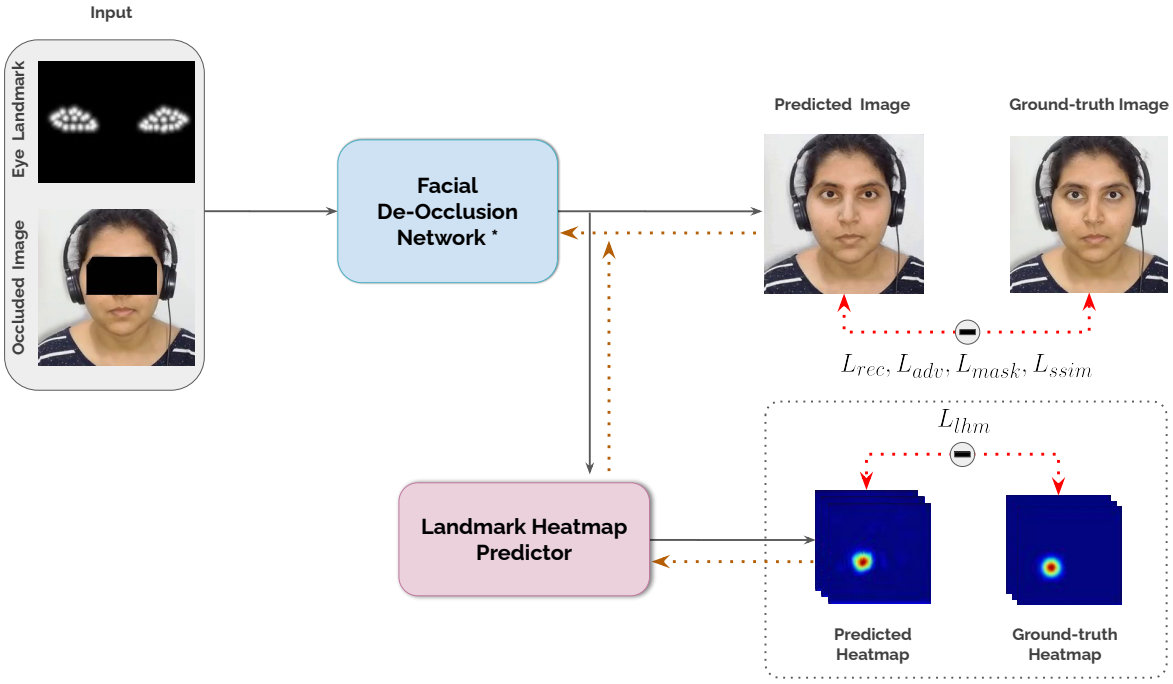

Figure 3.2: Illustration of our proposed architecture.

3.3.1.1 Attention-based encoder-decoder

We utilize an encoder-decoder architecture with an attention module to inpaint the missing regions of the face, particularly the eye region. The primary function of the attention module is to focus on reconstructing this region with high-frequency details such as hair, facial accessories, and appearances. It also helps the model to generalize to unseen and novel appearances, hairstyles, etc. It comprises ResNet and Inverted ResNet layers, with a bottleneck layer of 99 dimensions.

3.3.1.2 Landmark heatmap predictor module

We employ another encoder-decoder network to refine the reconstruction around the eye region. The primary aim of this network is to predict the eye landmark heatmap of the reconstructed image, based on which we can regularize the final reconstructed image using a loss function. This landmark heatmap predictor network is composed of ResNet and Inverted ResNet layers. The input to this network is the reconstructed image, X_{recon} produced from a facial de-occlusion network. The output is a 42-channeled landmark heatmap where each channel corresponds to one of 42 eye landmarks.

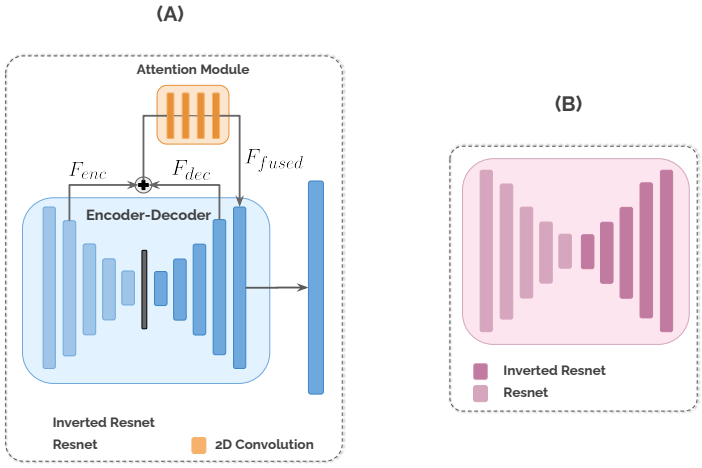

Figure 3.3: Two main module of our network. (A) Facial de-occlusion network as proposed in Chapter 2 and (B) Landmark Heatmap Predictor (LHM) module

.

3.3.2 Spatial Supervision using Landmarks

Per-frame predictions from traditional image-based facial de-occlusion network such as [17] suffer from temporal discontinuity and flickering, especially in eyelids. Therefore, to stabilize the eye movements, we leverage eye landmarks as an auxiliary input to guide smooth reconstructions in the eyelids that are much more realistic and consistent. This supervision helps the model preserve the structure of the eyelids. For better and enriched representations, we prefer 2D heatmaps over 2D coordinates. Each landmark is represented by a separate heatmap, interpreted as a grayscale image. We convolve all heatmaps to a single-channel grayscale image which is then concatenated with the occluded RGB input image in the channel dimension. This is further fed to the facial de-occlusion network to generate plausible and stable reconstructions.

3.3.3 Loss Functions

The primary goal of this pipeline is to generate plausible facial inpainted reconstructions consistent with other frames in sequence while preserving the landmark structure of the eyes. To serve this purpose, we use the following loss functions:

The first loss function ensures that the generated reconstruction is in close proximity to the ground-truth unoccluded image. Thus, we formulate pixel-based $L1$ loss to penalize reconstruction errors.

$$L_{rec} = \|X_{gt} - X_{rec}\|_1 \tag{3.1}$$

However, using only reconstruction loss generates blurry reconstructions. Thus we use a discriminator, D, in the pipeline to compute the adversarial loss that forces the encoder-decoder to reconstruct high-

fidelity outputs by sharpening the blurred images.

$$L_{adv} = log(D(X_{gt})) + log(1 - D(X_{rec})) \tag{3.2}$$

To further stabilize the adversarial training, we use SSIM based structural similarity loss, as defined in [7], that helps to improve the alignment of high-frequency image elements.

$$L_{ssim} = SSIM(X_{rec}, X_{gt}) \tag{3.3}$$

In order to emphasize the quality of reconstruction in the masked region, i.e., invalid pixels, we use a mask-based loss function. Here, we use the binary mask image as additional supervision to the network and input image while training. This helps mitigate the blinking artifacts around the eye region for stable reconstructions.

$$L_{mask} = \|I_{mask} \odot X_{gt} - I_{mask} \odot X_{rec}\|_1 \tag{3.4}$$

where, I_{mask} refers to single channel binary mask image where white pixels (1) correspond to occluded region and black pixels (0) correspond to the remaining unoccluded region and $\odot$ is element-wise multiplication.

Apart from providing a landmark heatmap along with occluded input, we also regularize the reconstructions based on landmarks using a loss function. To prevent irregularities in the eye region and preserve eyelid shape, we utilize landmark heatmap prediction loss that regularizes the inpainted reconstructions based on predicted eye landmark heatmaps, $LHP(X_{rec})$ and ground-truth eye landmark heatmaps, H. Here, for each landmark $l_i \in R^2$, H_i consists of a 2D normal distribution centered at li and a standard deviation of σ.

$$L_{lhm} = \|H - LHP(X_{rec})\|_2 \tag{3.5}$$

Thus, the final training objective loss function can be written as,

$$L_{final} = \lambda_{rec} * L_{rec} + \lambda_{adv} * L_{adv} + \lambda_{ssim} * L_{ssim} + \lambda_{mask} * L_{mask} + \lambda_{lmh} * L_{lmh} \tag{3.6}$$

where, $\lambda_{rec}, \lambda_{adv}, \lambda_{ssim}, \lambda_{mask}$ and λ_{lmh} are the corresponding weight parameters for each loss term.

3.4 Experiments and Results

3.4.1 Dataset and Training Settings

3.4.1.1 Dataset preparation

We train the network on different face video sequences for multiple identities. We train a person-specific model for every identity on 4-5 sequences captured in various appearances, including apparel,

hairstyle, facial accessories, and different head poses. Videos are recorded at a resolution of 1280 x 720 at 30 fps and then cropped to 256 x 256 for training. Note that there is no overlap between the training and test set. To test the ability of our model to generalize to novel appearances, we validate it with completely unseen videos that are not seen during the training process. For the provision of spatial supervision, we use Mediapipe [16] to detect and localize 42 landmarks around the eye, including iris landmarks. As discussed in Section 3.3.2, we create a heatmap for every landmark coordinate. Since we extract pseudo landmarks directly from unoccluded ground truth that is inherently spatially aligned with the occluded face, we directly append landmark heatmaps with the occluded input image without any further processing.

3.4.1.2 Inference with real occlusion

The eye information might not be directly accessible during inference when wearing regular virtual reality headsets. Fortunately, modern devices allow eye tracking using IR cameras mounted inside headsets. We can extract this information from eye images captured using these cameras. Unfortunately, the images captured by these cameras are not aligned with the face image. Hence, we need to calibrate both the eye and face camera as proposed in [42], [54] to align the eye images with the face image coordinate system. However, due to the unavailability of these headsets, we opt for pseudo landmarks extracted from ground-truth images to provide supervision to the model. As discussed, we extract these landmarks using the Mediapipe [16] face landmark detector. It is to be noted that these landmarks do not adhere well to an anatomically defined point across every video frame and thus have local noise in them generated due to the inaccuracy of the facial landmark detector.

3.4.1.3 Training strategy

We follow a similar two-stage training strategy proposed in [17]. In the first stage, we only train the encoder-decoder network without an attention module on unoccluded images of the person using the first three losses aforementioned in Section 3.3.3, each added incrementally after 400, 50, and 250 epochs, respectively. In the second stage, we fine-tune the same encoder-decoder with the attention module and the landmark prediction module on occluded images of the same person using two additional loss functions. We use the same three losses as the first stage. Apart from this, we also use a landmark heatmap prediction loss to regularize the reconstructions generated from the facial de-occlusion network and a mask-based loss to minimize reconstruction errors in the masked region.

3.4.2 Results

In this section, we present the results of our method and discuss its superiority over existing approaches. We first compare the visual quality of the reconstruction generated by our method with popular state-of-the-art inpainting methods, followed by a quantitative analysis using standard evaluation metrics. To

further validate the efficacy of our approach, we also report an ablation study conducted in the scope of this work.

3.4.2.1 Qualitative comparison:

For visual comparisons, we evaluate our method against various image inpainting methods across 20 subjects. Results highlighted in Figure 3.4, 3.5 and 3.6 show that the reconstructions generated using our method are visually pleasing and consistent across frames compared to other inpainting methods. Reconstructions generated using our approach, as shown in row (C) of Figure 3.4 show the significance of landmark supervision and regularization loss in capturing eye movements such as blinks.

However, predictions using other approaches are often incoherent across frames. As visible in row (F), DeepFillv2 [50] fails poorly to generate plausible reconstruction in the eye region. LaFIn [48] and Edge-Connect [34] generate superior reconstructions compared to DeepFillv2, however, it cannot handle eye movements. Besides, there is a noticeable discrepancy in the left and right eyes that looks unnatural. Baseline [17] produces naturally-looking reconstructions but cannot handle eye blinks. For better comparison, refer to the supplementary video 1. In Figure 3.7, we also show the reconstruction error ($l2$ error) between the results generated by different image inpainting methods and the ground truth for better justification.

3.4.2.2 Quantitative Comparison

To quantify the quality of reconstructions, we use standard image quality metrics such as SSIM [46], PSNR [18], and LPIPS [53]. Table 3.1 shows the quantitative comparison of our proposed method with other state-of-the-art methods. As reported, our method **(in bold)** performs better in all evaluation metrics than other face inpainting methods. For SSIM and PSNR, a higher value indicates better reconstruction quality and vice-versa. Similarly, for LPIPS, a lower value indicates better perceptual quality and vice-versa.

Method	SSIM↑	PSNR↑	LPIPS↓
LaFIn [48]	0.914	23.693	0.0601
EdgeConnect [34]	0.908	23.10	0.0689
DeepFillv2 [50]	0.845	19.693	0.117
Ours	**0.949**	**31.417**	**0.0235**

Table 3.1: Quantitative comparision of our method with and without landmarks with other image inpainting methods.

Figure 3.4: Qualitative comparison with SOTA image inpainting methods. From row (A-G) are Occluded (input), Original (ground-truth), Ours, Baseline [17], Edge-connect [34], DeepFillv2 [50] and LaFIn [48] respectively. From left to right are consecutive frames of unseen testing video.

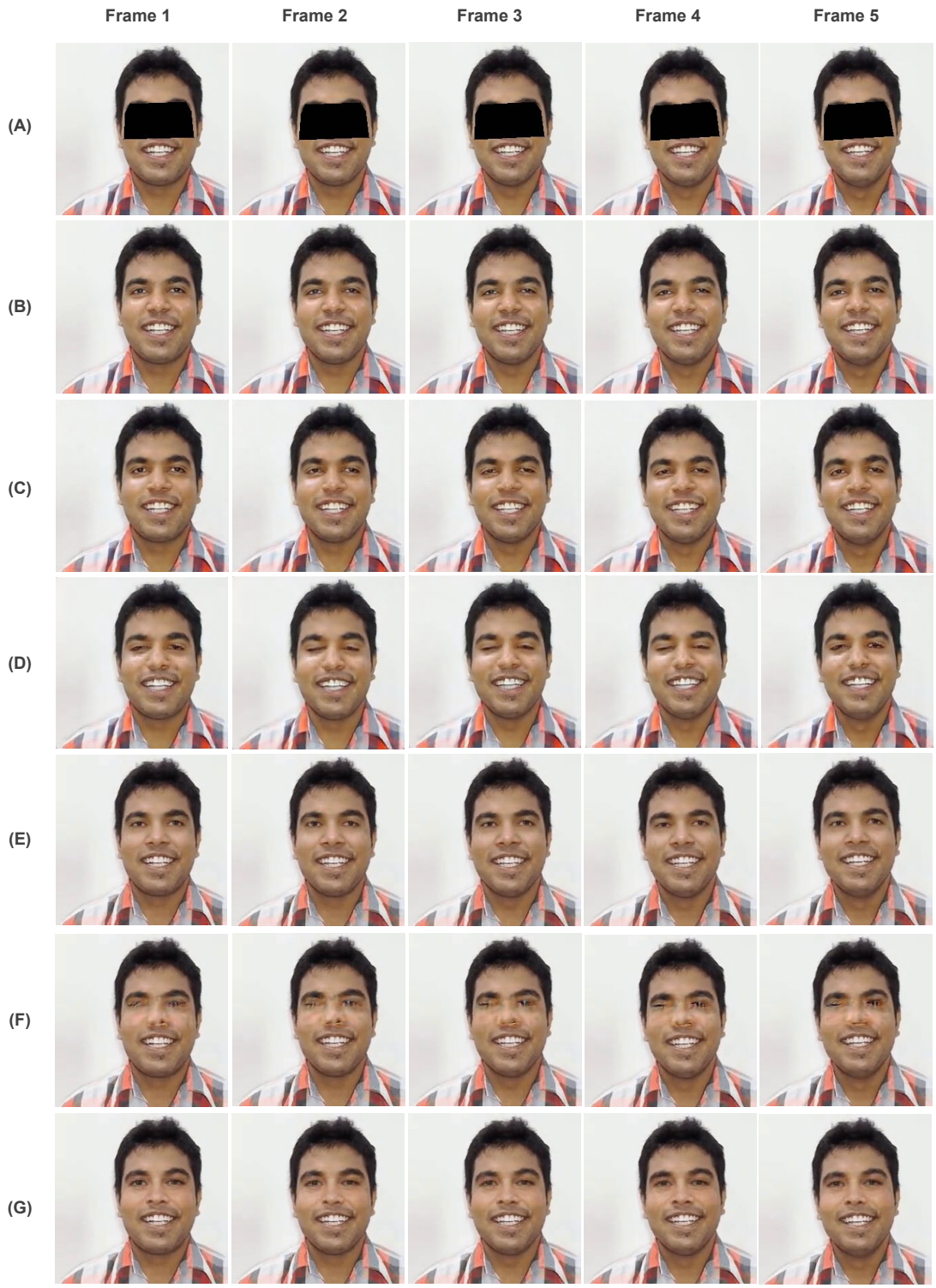

Figure 3.5: Qualitative comparison with SOTA image inpainting methods. From row (A-G) are Occluded (input), Original (ground-truth), Ours, Baseline [17], Edge-connect [34], DeepFillv2 [50] and LaFIn [48] respectively. From left to right are consecutive frames of unseen testing video.

Figure 3.6: Qualitative comparison with SOTA image inpainting methods. From row (A-G) are Occluded (input), Original (ground-truth), Ours, Baseline [17], Edge-connect [34], DeepFillv2 [50] and LaFIn [48] respectively. From left to right are consecutive frames of unseen testing video.

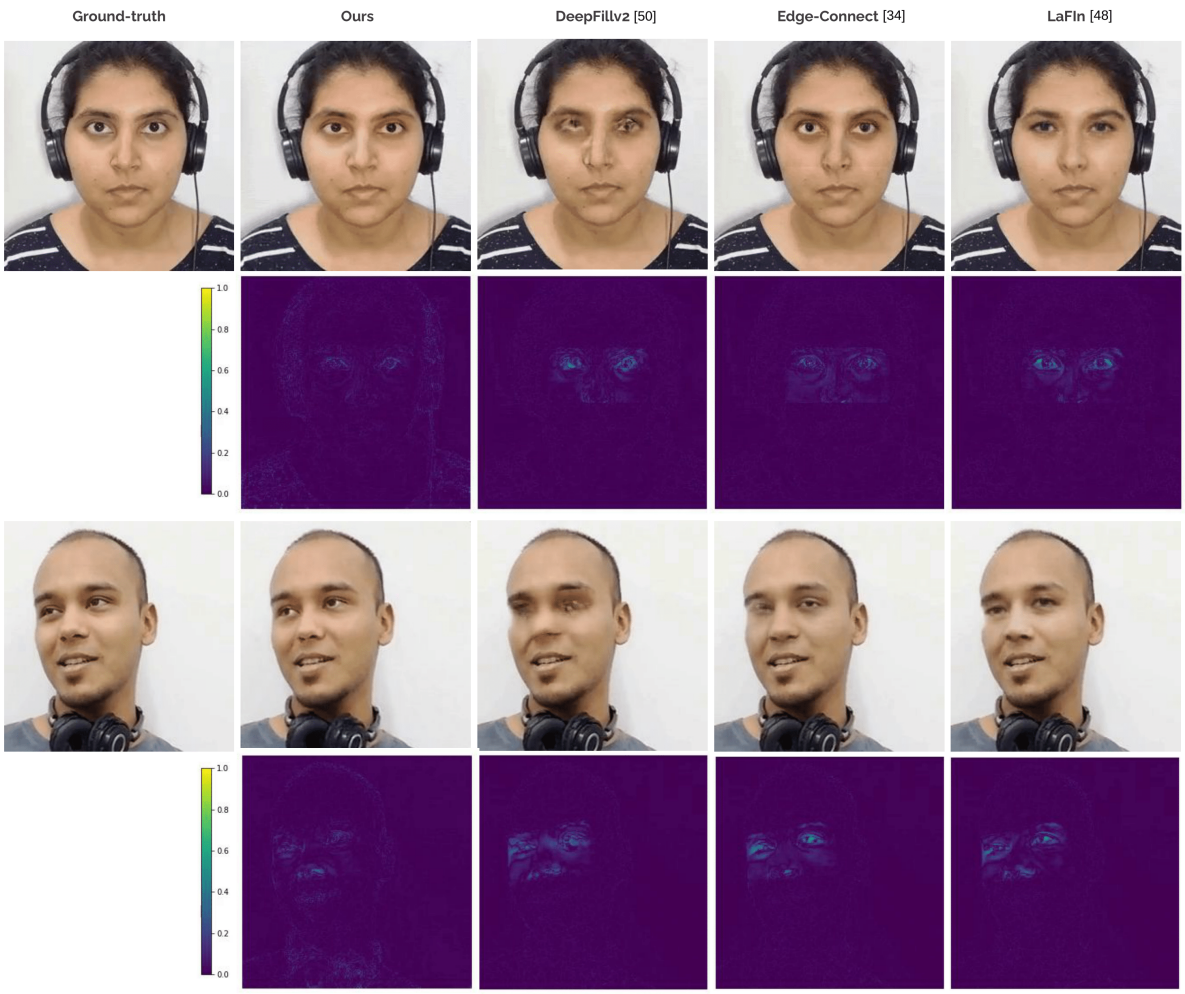

Figure 3.7: Qualitative result that showing the reconstruction error ($l2$ error) between the results generated by different image inpainting methods and the ground truth.

3.5 Ablation Studies

We perform several ablation studies to understand the various aspects of our model. We first analyze the effect of providing spatial supervision to the model in enhancing reconstruction quality. As depicted in Figure 3.8 and 3.9, our model with landmarks produces aesthetically pleasing eyes and preserves eyelid shape in contrast to the one without landmarks supervision. It is due to the guidance provided by the landmarks that helps the model enforce consistency in eye movements, including the opening and closing of eyes. However, this does not ensure eye movements are temporally coherent. Secondly, we show the effect of using a regularizing loss function based on landmarks heatmap to penalize the errors caused by the model. Table 3.2 reports the positive impact of using eye landmarks in guiding the reconstruction in the eye region.

Method	SSIM↑	PSNR↑	LPIPS↓
Baseline[17]	0.918	29.025	0.042
Ours (with LHM)	0.926	29.272	0.0418
Ours (with LHM + L_{lmh})	**0.949**	**31.417**	**0.0235**

Table 3.2: Ablation study showing the significance of using landmark supervision on the reconstruction quality.

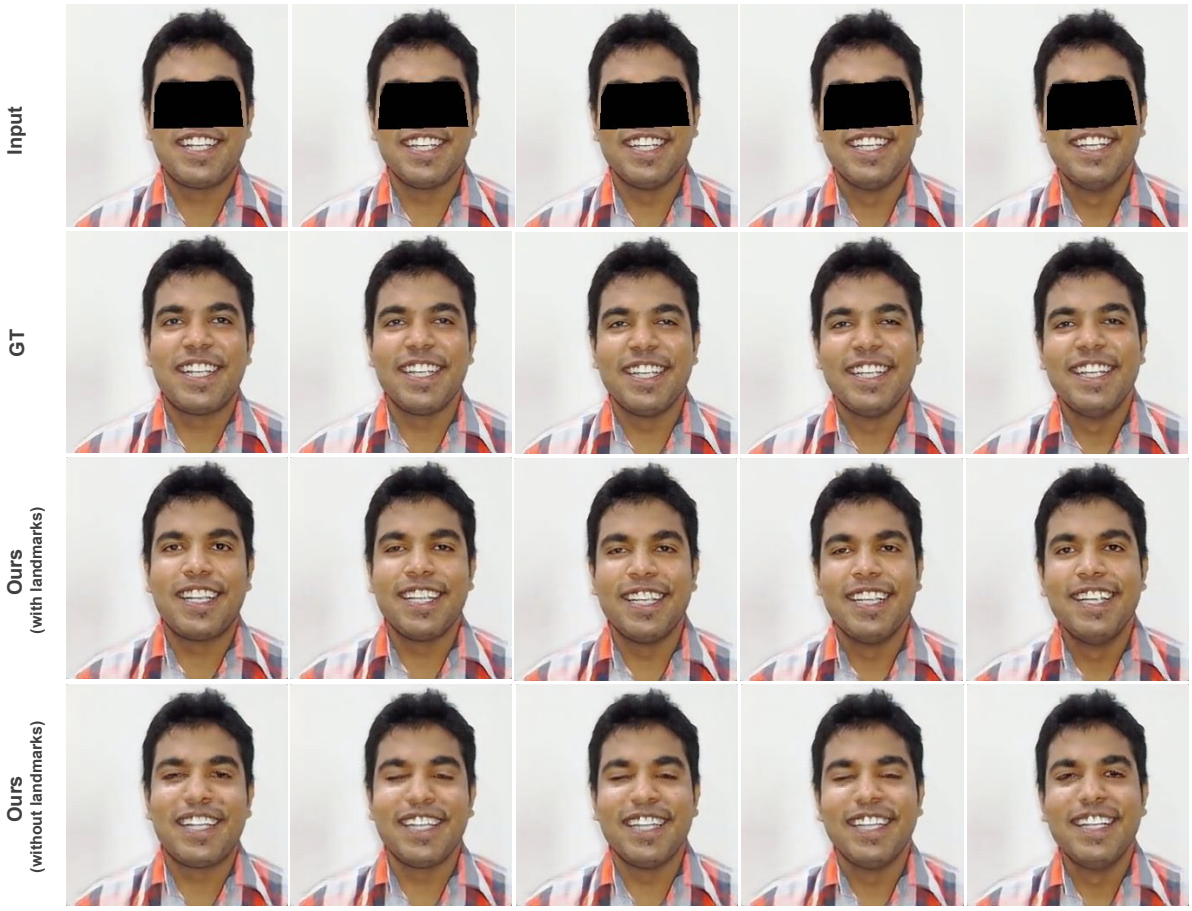

Figure 3.8: Testing results showing the effect of using landmarks as auxillary input to the network. From row (1-4) are occluded (input), original (ground-truth), results with and without landmarks respectively. From left to right is temporal continuously images of original 30fps videos.

Figure 3.9: Testing results showing the effect of using landmarks as auxillary input to the network. From row (1-4) are occluded (input), original (ground-truth), results with and without landmarks respectively. From left to right is temporal continuously images of original 30fps videos.

3.6 Conclusion

This chapter presents an enhancement in existing facial de-occlusion networks by explicitly focusing on improving eye synthesis. We show that providing landmark information during the inpainting process can yield superior quality and photorealistic reconstructions, including the eye region. We discuss how this information can be retrieved: 1) by extracting pseudo landmarks from ground-truth images and 2) using modern HMD devices capable of tracking eye movements. To further enhance the generated output, we propose a landmark-based loss function that act as a regularizing term to improve reconstruction quality and helps capture subtle eye movements such as eye blinks. We conducted qualitative and quantitative analysis and reported superior results with other SOTA inpainting methods to justify the usefulness of our approach.

Chapter 4

Our Dataset

This chapter describes the dataset we used for our previous experiments and constitutes a crucial part of this thesis work. In deep learning research, the performance of a model is largely dependent upon the quality and type of dataset we use for training and evaluation purposes. Preparing a suitable dataset is the crucial step in training all deep-learning models. This chapter discusses existing face datasets widely used for training these models related to tasks such as face recognition and other face-oriented problems. Following this, we also talk about the need to propose another face dataset for our problem statement, a simple yet cost-effective method to obtain a face dataset using readily available low-cost devices and apply off-the-shelf approaches to manipulate and make it fit for training deep learning models. This chapter briefly discusses the data acquisition and preparation of the face dataset we used for our problem statement. Dataset is available on this *website.* It is to be noted that out of 50 subjects for whom we collected the dataset, we used the data of only 20 subjects for evaluation of our proposed approaches and reporting results. The data for additional 30 subjects were collected at a later stage to release it for research and academic purposes.

4.1 Introduction and Motivation

Deep-learning models are notoriously data-hungry. The significant advances in computer vision are mainly due to an abundance of publicly available state-of-the-art image and video datasets. The Internet is the most common source for providing all forms of data, including billions of images and videos. Ever since the birth of the digital era and the availability of web-scale data exchanges, there has been a tremendous increase in the amount of data generated every day, which is incessantly growing. As of February 2020, more than 500 hours of video were uploaded to YouTube every minute. Most popular computer vision datasets, such as ImageNet, VoxCeleb, and CelebA, have been hand-crafted using the data uploaded and created on the Internet, mainly on social media platforms such as Meta, Youtube, Instagram, and many more. Depending upon the type of task, these are then pre-processed and annotated with additional information such as labels.

However, depending on the complexity and the domain of the computer vision task we are trying to solve, the available datasets might not be suitable. For example, in tasks like facial inpainting and reenactment in VR settings, the data needed to train the model requires face videos of the person wearing headsets. Several works have been explored in this direction; however, none have open-sourced the dataset. This motivates us to build such a dataset and make it available to be used for academic and research purposes. With the next iteration of the Internet, the metaverse, an increase in AR/VR-based research is expected. The availability of such datasets can prove immensely beneficial to the computer vision community, especially for AR/VR-based applications.

4.2 Our Dataset

Understanding and recovering missing or corrupted regions in images are essential for many applications, e.g., image restoration, inpainting, and de-occlusion. In the case of facial de-occlusion in VR settings, public datasets consisting of videos of the person wearing virtual reality headsets/glasses with associated ground truth face images are needed. However, it might not be feasible in practice since these headsets physically occlude the region around the eyes. One option to deal with this is to model the headsets synthetically. Most common 2D-based methods such as [48, 50] overlay either use headset images or binary mask images to indicate the region of occlusions. 3D-based methods like [35] built a data synthesization pipeline to create a synthetic dataset of RGB-D images of faces based on 3D face models such as BFM and 3DMM. To our knowledge, there is no publicly available dataset of VR-based face videos that consists of different identities in different appearances, poses, and facial expressions.

4.2.1 Data Recording Setup

We focus on setting up a simple yet cost-effective approach for collecting our face dataset to train the facial de-occlusion network. To record the face sequences of the participants, we used a mobile phone camera that captures videos at 1280 x 720 px at a frame rate of 30 fps. We ensured uniform lighting and background during the entire session. Figure4.2 shows the data recording setup installed for the data collection task.

We recorded the data for around 50 subjects of Indian origin, preferably 20-40 years, including males and females. Multiple sessions were recorded in different appearances and facial accessories, including jewellery, simple make-up, headphones, and hair accessories, to maintain variability in the dataset. Every subject was asked to converse in a room with another person over an extended period. The intention was to produce natural behavior, despite the unnatural circumstances. This included day-to-day facial and eye movements with expressions such as happy, sad, surprise, anger, and neutral. We also post-process the video sequences. We crop the video frames using the smallest crop containing all the bounding boxes. The process is repeated until the end of the sequence. All sequences are resized to 256×256 preserving

Figure 4.1: Sample frames collected for various identities for training purposes.

the aspect ratio. We then split each video into frames for training. Some samples collected for different identities in varying appearances, poses, and expressions are shown in Figure 4.1 and 4.5.

4.2.2 Creation of Synthetic Mask

We synthetically generate binary occlusion masks to account for the non-availability of ground-truth data for training purposes. Every pixel representing occlusion is represented by 255 or 1, and the rest are denoted by 0. Most of the previous methods generally use rectangle/square masks or irregular-shaped masks [29] to indicate the region in the image that is corrupted or missing. However, it might not be appropriate to use these masks in order to imitate the shape and structure of virtual reality headsets since the occlusion created by virtual reality headsets is not always rectangular or square. Additionally, the headsets also occlude some part of the background when there is a change in head poses, especially non-frontal. Thus for better comparison, we generated a polygon mask guided by facial landmarks to represent the headset occlusion accurately.

State-of-the-art landmark detectors such as Dlib [2] fail to detect faces (thus landmarks) in case of extreme head poses. Thus, we use off-the-shelf Google Mediapipe [16] dense face-mesh to extract relevant facial landmarks around the eye region, which are robust and accurate enough for our task. It is to be noted that we only use 6 out of 468 vertices (or landmarks) (3 each on the left and right side) around

Figure 4.2: Data recording setup for the collection of our face dataset.

the eye region. We ensure that the headsets occlude the region surrounding the area covered by these landmarks. During training time, we overlay the generated polygon mask over a ground-truth unoccluded image to produce a corresponding corrupted input image. This is done for every frame of the training video sequences. Figure 4.4 shows the sample data that is fed to the network during training time.

4.2.3 Data with real occlusion

To test with real occlusion, we collected data with Epson Moverio BT-200 Smart Glasses. Since our network has been trained to work with binary masks, we must pre-process the above data to make it suitable for testing. Hence as discussed in Section 4.2.2, we overlay a binary mask over the eye region to cover the smart glass entirely to generate occluded input for the network.

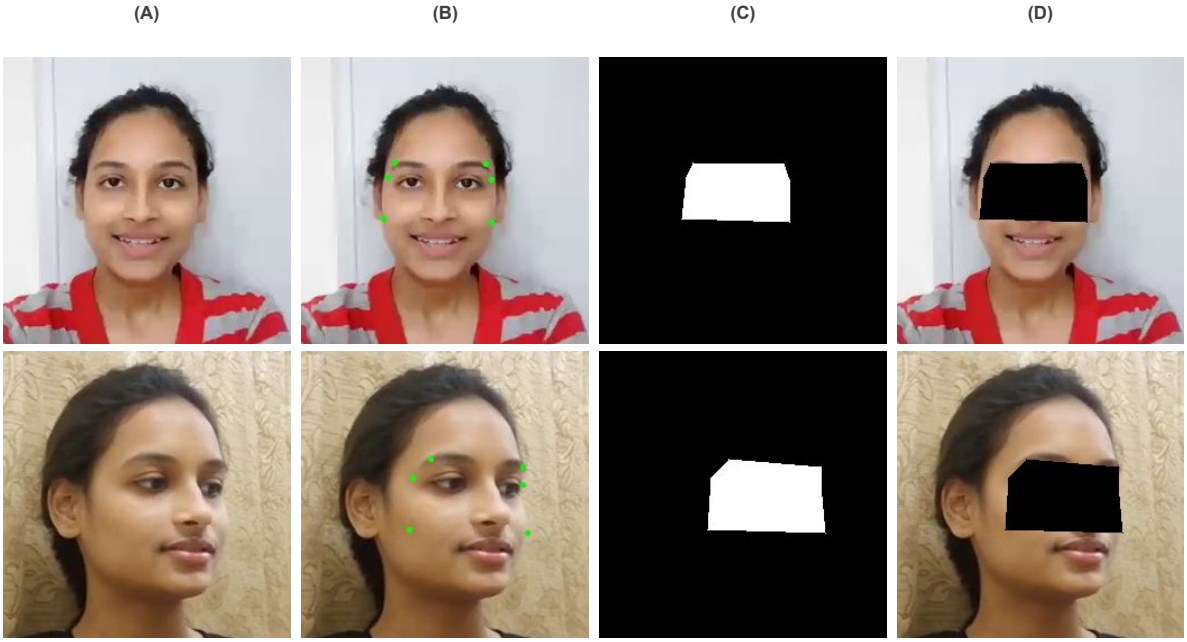

Figure 4.3: Dataset generated using above mention method. From left to right, (A) Ground-truth unoccluded image, (B) Eye region landmarks extracted from [16] to create synthetic occlusion, (C) Binary mask image generated using extracted landmarks, and (D) Corresponding occluded image given as an input during training and testing.

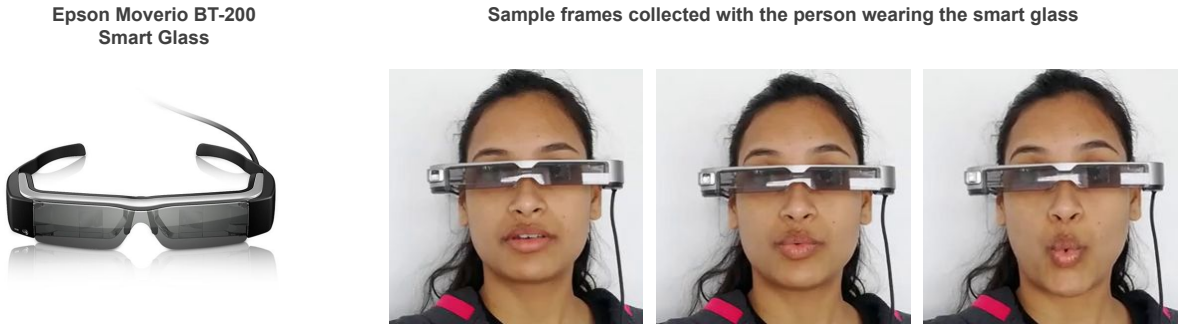

Figure 4.4: Sample data with real occlusion (smart-glass).

Figure 4.5: Sample frames collected for various identities for training purposes.

Chapter 5

Future Work and Conclusion

Virtual communication has become inevitable as the world becomes more dependent on internat and global platforms. With the Covid-19 pandemic bringing unprecedented transitions in social and work life, 2D video conferencing applications have realized an impressive surge in their users. Since post-pandemic, concepts like 'hybrid workforce' and 'flexible work' have been the preferred mode of interaction. The flexibility and convenience of virtual communication provide working professionals to attend meetings remotely from home, thus allowing them to spend more time with their families. Thus, it is essential to upgrade current technologies and build scalable applications so that users can experience similar lifelike in-person interactions virtually, depending on their mode of interaction.

This thesis introduced an efficient and usable approach to facial de-occlusion for virtual telepresence systems that promises to provide a better immersive experience to users. Facial de-occlusion or HMD removal is a highly ill-posed problem that has recently attracted considerable attention due to its applications in Metaverse projects. We followed a two-stage person-specific training strategy for preserving high-frequency user-specific details. With thorough qualitative and quantitative analysis, we justified the efficacy of our approach in generating high-quality realistic reconstructions, even with novel and unseen appearances. In addition, we also collected our dataset for multiple users with varied variations in appearances, poses, and facial expressions, which we plan to release for academic/research purposes. From an application perspective, we showed how our approach is easily integrable with facial animation models that are effective for low-bandwidth situations. We believe this work can be valuable to the research community and benefit other AR/VR-based applications.

Face reconstruction is an exciting area of research; indeed, there is a scope for improving our proposed approach by extending it to the 3D domain, which will be of immense interest to people working in 3D reconstruction, multimedia, computer graphics, and animation. There can be a possible improvement in the temporal domain and its extension to video-based facial de-occlusion. As additional work, in Appendix A, we discussed temporal coherency and proposed possible approaches to generate temporally consistent predictions. We look forward to working and exploring this field one step ahead by solving existing challenges and making life easier and better for the people around us.

Appendix A

Temporally Consistent Face Image Inpainting

In the scope of this thesis, we dealt with video frames independently. However, as a future direction of this work, we also explored temporally consistent face image inpainting, where we tried to extend the previously proposed image-based face de-occlusion to video frames. Achieving high-fidelity inpainted reconstructions with minimal flickering artifacts is of prime importance in video conferencing/facial animation applications. Inpainting results should possess high visual quality with respect to reconstruction performance, realism, and temporal consistency, i.e., they should faithfully recreate missing contents in a way that resembles the real world and exhibits minimal flickering artifacts. Directly using existing image inpainting solutions for per-frame video reconstruction perform poorly. The performance even degrades more when dealing with face image inpainting in VR settings since a large portion of the face is occluded by virtual reality devices. Additionally, it is challenging to model the complex structure of the face and maintain coherency in facial movements simultaneously. To this end, we propose various approaches to generate temporally consistent reconstructions.

A.1 Introduction

Videos establish a more personal connection with the audience than other media types, such as images, audio, and text. They are interactive and quickly help grab attention and keep the user engaged. In multimedia, video-based applications such as video conferencing/telepresence systems, it is of utmost importance that the resultant video has high reconstruction quality with minimum temporal artifacts. A naive approach to extend single-image methods to video is to apply them on a frame-by-frame basis. However, directly applying image processing algorithms to video level often leads to temporal discontinuities and noticeable flickering in the resulting video. Since the model does not explicitly capture inter-frame dependencies, the generated output might suffer from temporal and flickering artifacts which is not acceptable in real-time applications.

In the past, a similar direction of work has been explored, and this problem was initially addressed by Blind Video Temporal Consistency [26] using a gradient-domain technique. Since then, many methods such as [27, 9, 36] have been proposed in the literature. [36] is one such video inpainting method

that uses a similar temporal module consisting of ConvGRU to post-process the inpainted frames to improve coherency in the resultant video. However, these methods are not validated to improve temporal inconsistency and flickering artifacts with inpainted faces. There have also been an active research in video inpainting. Traditional face video inpainting methods, such as CombCN [43] is built by combining 2D and 3D CNNs. Since, the method has only been tested on low-resolution videos of aligned faces and vehicles with fixed square holes, it is difficult to use it for real-world occlusions such as virtual reality headsets. [11] is another deep video inpainting method that uses a learnable gated temporal shift module with 2D convolutions in a U-net kind of architecture to recover face videos with free-form masks. Though the method deals with large and moving occlusion, still the inpainted video lacks realism and suffers from inconsistency all over the face. Similarly, [10] also deal with free-form video inpainting using 3D convolutions and temporal patchGAN loss to enhance temporal consistency in videos. Another set of works on inpainting has been done in the form of object removal from images and videos.

Recently, [47] proposed an image to face video inpainting by using spatio-temporal nested gan architecture. They used 3D residual blocks to capture inter-frame dependencies. The authors showed that conditioning facial inpainting on landmarks yielded stable reconstructions. Nonetheless, it is only validated with a specific type of circular mask that covers the eye, nose, and mouth. [48] is another face image inpainting method that is guided using facial landmarks.

Keeping in mind the drawbacks of extending image-based approaches to video frames, we propose an approach to temporally consistent face image inpainting. This appendix focuses on extending the image-based facial de-occlusion network to another dimension by making temporally coherent predictions using popular approaches used in the literature. This work can prove valuable for AR/VR-based hybrid telepresence systems and can benefit Metaverse applications.

The following points can summarise this work:

1. This is supplementary to the previous work, where we try to address the concept of temporality in per-frame predictions.
2. Directly using the previous proposed image-based facial de-occlusion network on videos results in flickering and temporal artifacts in the eye region, which is undesirable.
3. To tackle this, we propose two approaches; the first one utilizes optical flow, and the second is a convLSTM-based recurrent approach inspired by [26].

A.2 Related Work

A.2.1 Video Temporal Consistency

Temporal consistency in videos is essential for multimedia applications where user experience is a priority. Since videos are just a collection of consecutive images or frames, a simple way to extend the image-based method to videos is to apply them frame-by-frame. However, following this approach to

image processing methods, such as colorization and relighting, does not produce consistent results when applied to video frames. To address this issue, [26] proposed a practical approach to ensure temporal consistency in per-frame outputs of these image-based methods. Since then, many researchers have adopted this approach to solve similar problems related to image relighting [9] and video inpainting [36].

A.2.2 Optical flow for Temporality

Interestingly, videos are just an extension of images in the temporal dimension. Where image only has spatial information, videos, on the other hand, comprise both spatial and temporal information in them. Understanding video structure is vital, especially in multimedia, computer graphics, and animation. Optical flow is one of the most powerful and commonly used mechanisms for smoothing temporally inconsistent input frames to generate temporally consistent output frames in videos. Popular flow estimation methods widely used in literature include FlowNet2 [20] and RAFT [41]. In general, optical flow is calculated between two frames, but interestingly, there has been some research in predicting optical flow using a single video frame [1].

A.2.3 Recurrent Networks

Recurrent Neural Networks (RNN) are used when the data is in the form of a sequence, as in texts, videos, etc. These recurrent networks recognize data's sequential characteristics and use patterns to predict the following likely scenario. It has the capability to memorize and store previous results that can be used to predict the subsequent output. LSTM are the building blocks of RNNs. Convolutional LSTM (ConvLSTM) is a type of RNN with convolutional structures in both the input-to-state and state-to-state transitions used for spatio-temporal prediction. This ConvLSTM can be embedded as a layer to capture the spatial-temporal correlation of natural videos.

A.3 Proposed Method

Inspired by existing literature such as [26], we try two well-known approaches to target temporal consistency in per-frame reconstructions.

A.3.1 Approach 1: Using Optical Flow

We incorporate optical flow in our facial de-occlusion network for generating temporally smooth reconstructions. However, due to occlusion by HMD, the information in the masked region is unavailable for calculating the flow. Thus, we utilize the previous temporally consistent output frame, O_{T-1} to predict the next frame, $Pred_T$ using the single-image optical flow prediction network. This predicted next frame could be considered as the reference frame that is provided along with the current occluded input frame, I_T to predict the temporally consistent output frame, O_T. This output is then fed back to the

optical flow prediction network to process next frames in sequence. The proposed optical-flow-based facial de-occlusion network pipeline is demonstrated in Figure. A.1.

We use the same attention-based facial de-occlusion network proposed in chapter 2 for generating temporally consistent output frames. For the single-image optical flow prediction network, we use an encoder-decoder network that takes in input, a 3-channel RGB image, and generates a 1-channel flow output. We use the generated flow to warp the previous temporally consistent output frame, O_{T-1} to produce the next predicted frame, $Pred_T$.

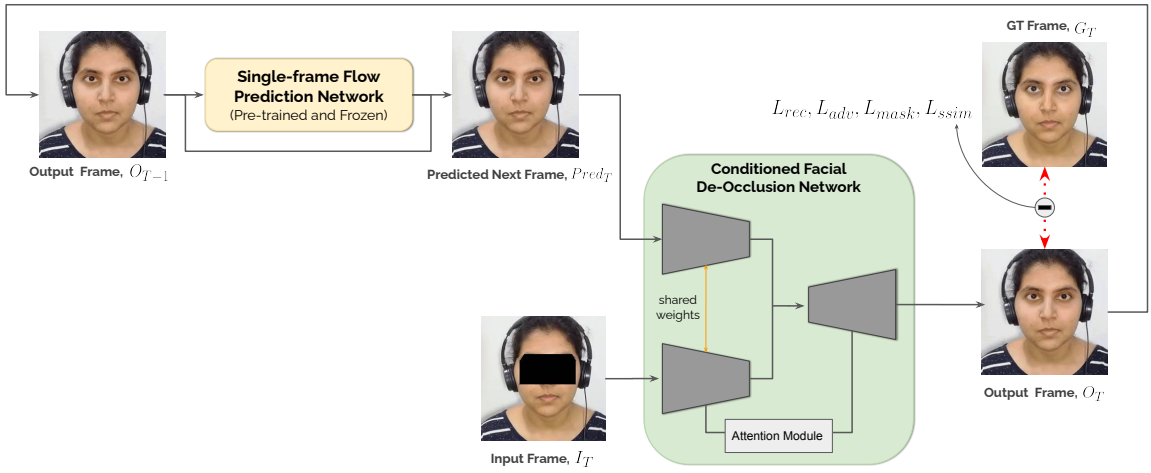

Figure A.1: Illustration of our proposed optical-flow based architecture.

A.3.2 Approach 2: Extending Blind Video Temporal Consistency

Apart from an optical-flow-based approach, we also try a convLSTM-based encoder-decoder architecture to capture temporal information by leveraging the information from multiple consecutive frames. This framework was previously proposed in the work [26], and since then, it has been widely used to solve problems related to inconsistency in video frames. However, this approach has not been validated for the problems related to inpainting. The proposed ConvLSTM-based pipeline integrated with the facial de-occlusion network is demonstrated in Figure. A.2.

Unlike [26], we do not have ground-truth frames during inference. However, we utilize them during training and assume they are unavailable during test time. Thus, our method is independent of the actual ground-truth frame. We first generate the temporally inconsistent de-occluded frame using our facial de-occlusion network [17] for the occluded input, I_T. In this context, the generated image is referred to as a processed frame, P_T. Along with the previously generated temporally consistent output, O_{T-1}, we fed the processed frame, P_T to Consistency Network that outputs the residual or the error. We then add

this residual image with the processed frame, P_T to generate temporally consistent output for the current frame, O_T. This output is then reused to process the successive frames in sequence.

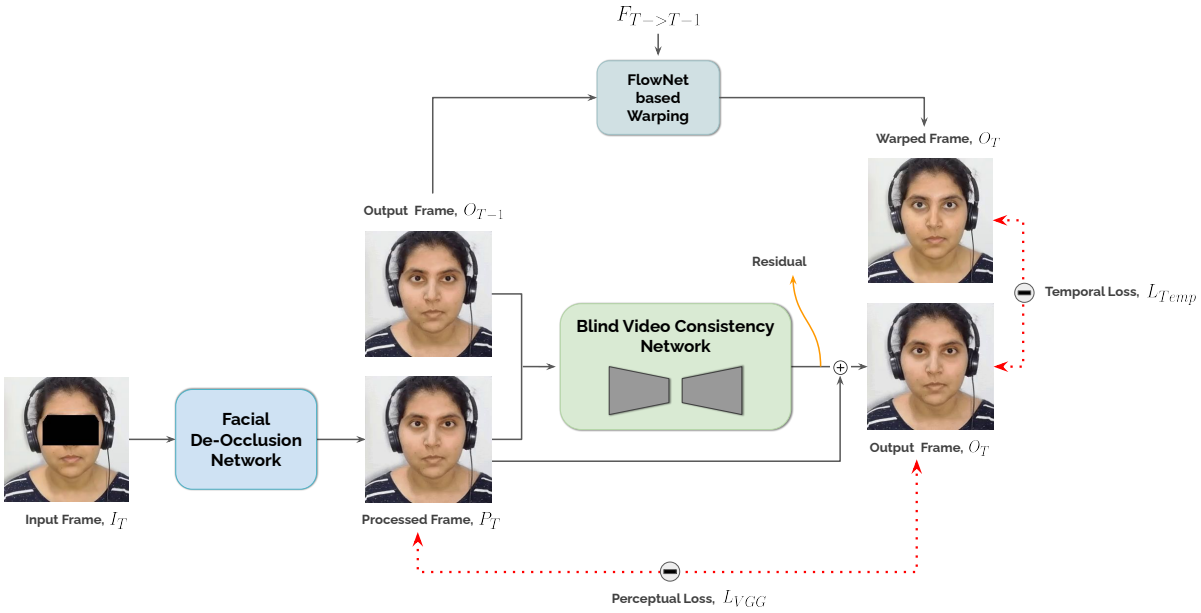

Figure A.2: Illustration of our proposed architecture built upon Blind Video Consistency [26].

A.3.3 Our Training Strategy

For training the optical-flow-based network, we used a single-frame flow prediction network to predict the flow. We train this network until convergence on unoccluded images of the user for 3 frames in sequence and backpropagate the loss while processing the third frame. We then freeze this network and train the entire network in two stages on unoccluded and occluded images of the user using the loss function proposed in Section 2.3.2.3 of Chapter 2. The dataset used to train this network consist of 14 different sequences with 300 frames each.

For training the ConvLSTM-based network, we used off-the-shelf, FlowNet2 to calculate the flow between two ground-truth consecutive frames and use this flow to warp predicted frames during the training process. Please note that we freeze pre-trained FlowNet2 and the pre-trained image-based facial de-deocclusion network and only train the ConvLSTM-based consistency module. This module is trained from scratch on occluded images with a sequence size of 11 frames and batch size of 4 on around 4k frames. We use two loss functions, perceptual loss, L_{VGG} and temporal loss L_{temp}. The dimensionality of the z-vector of the facial de-occlusion network is fixed to 99 and 256 for optical-flow-based and ConvLSTM-based approaches, respectively. We use the dataset proposed in Chapter 2 of this thesis for training and inference purposes.

A.4 Results and Analysis

In this section, we present results generated by our proposed pipelines. Figure A.3 and A.5 reports the results generated by the optical-flow and blind video consistency based approaches, respectively. It is to be noted that although the optical-flow-based method produces better reconstruction than the image-based facial de-occlusion network, it fails to work in cases of non-frontal poses as depicted in Figure A.4. Even, there is slight variation in facial structure and identity. ConvLSTM based approach produces comparable results with image-based approach but does not contribute anything to ensure temporal smoothness during eye blinks. Both the methods do not seem to improve temporal consistency as expected. However, qualitatively, the results are comparable to the outputs yielded by the previous image-based approach and do not seem to deteriorate much.

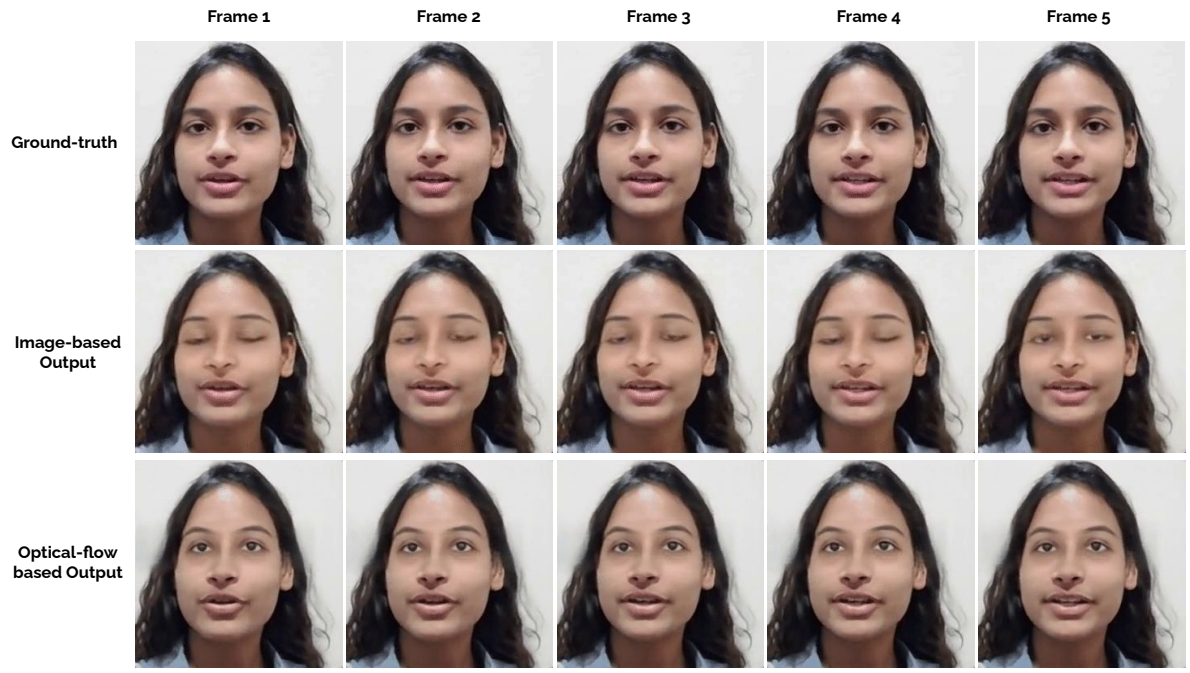

Figure A.3: Qualitative results produced from our proposed optical-flow based approach to generated temporally consistent results. Here, image based output refers to the output generated by facial de-occlusion network introduced in Chapter 3.

A.5 Conclusion

Here, we attempted to solve a complex problem of handling inconsistency in per-frame reconstruction when our image-based de-occlusion network was directly applied to videos. We presented two popular

approaches that have been used in the past to deal with problems related to temporal consistency in videos. However, none of the two approaches seemed to work well in handling inconsistencies from the image-based facial de-occlusion network. The possible reason behind this could be attributed to the missing eye region that contributes to an important region of faces in articulating complex facial expressions and movements.

Figure A.4: Failure cases of proposed optical-flow based method in non-frontal head poses.

Figure A.5: Qualitative results produced from our proposed by ConvLSTM based approach to generated temporally consistent results. Here, image based output refers to the output generated by facial de-occlusion network introduced in Chapter 2.

Related Publications

- **Attention based Occlusion Removal for Hybrid Telepresence Systems**, Surabhi Gupta, Ashwath Shetty, Avinash Sharma. **In CRV:** Conference on Robots and Vision 2022, Toronto, Ontario.
- **Facial De-occlusion Network for Virtual Telepresence Systems**, Surabhi Gupta, Ashwath Shetty, Avinash Sharma. **In CVPRW:** CVPR Workshop on Computer Vision for Augmented and Virtual Reality, New Orleans, LA, 2022.
- **Supervision by Landmarks: An Enhanced Facial De-occlusion Network for VR-based Applications**, Surabhi Gupta, Sai Sagar Jinka, Avinash Sharma, Anoop Namboodiri. **In ECCVW:** 1st International Workshop and Challenge on People Analysis: From Face, Body and Fashion to 3D Virtual Avatars (In conjunction with ECCV 2022).

Bibliography

[1] http://cs231n.stanford.edu/reports/2016/pdfs/223_Report.pdf.

[2] http://dlib.net/.

[3] https://www.brookings.edu/blog/up-front/2020/04/06/telecommuting-will-likely-continue-long-after-the-pandemic/.

[4] https://www.techrepublic.com/article/why-remote-work-has-grown-by-159-since-2005/.

[5] 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*, SIGGRAPH '99, 1999.

[6] K. W. Bowyer, K. Chang, and P. Flynn. A survey of approaches and challenges in 3d and multi-modal 3d+ 2d face recognition. *Computer vision and image understanding*, 2006.

[7] B. Browatzki and C. Wallraven. 3fabrec: Fast few-shot face alignment by reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[8] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. 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, 2018.

[9] S. Chandran, Y. Hold-Geoffroy, K. Sunkavalli, Z. Shu, and S. Jayasuriya. Temporally consistent relighting for portrait videos. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, 2022.

[10] Y.-L. Chang, Z. Y. Liu, K.-Y. Lee, and W. Hsu. Free-form video inpainting with 3d gated convolution and temporal patchgan, 2019.

[11] Y.-L. Chang, Z. Y. Liu, K.-Y. Lee, and W. Hsu. Learnable gated temporal shift module for deep video inpainting, 2019.

[12] B. Dolhansky and C. Canton-Ferrer. Eye in-painting with exemplar generative adversarial networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7902–7911, 2018.

[13] M. Elgharib, M. Mendiratta, J. Thies, M. Niessner, H.-P. Seidel, A. Tewari, V. Golyanik, and C. Theobalt. Egocentric videoconferencing. *ACM Transactions on Graphics (TOG)*, 2020.

[14] G. Gafni, J. Thies, M. Zollhöfer, and M. Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[15] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks, 2014.

[16] I. Grishchenko, A. Ablavatski, Y. Kartynnik, K. Raveendran, and M. Grundmann. Attention mesh: High-fidelity face mesh prediction in real-time, 2020.

[17] S. Gupta, A. Shetty, and A. Sharma. Attention based occlusion removal for hybrid telepresence systems. In *19th Conference on Robots and Vision (CRV)*, 2022.

[18] A. Horé and D. Ziou. Image quality metrics: Psnr vs. ssim. *ICPR*, 2010.

[19] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, University of Massachusetts, Amherst, 2007.

[20] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017.

[21] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[22] Y. Jiang, F. Yang, Z. Bian, C. Lu, and S. Xia. Mask removal: Face inpainting via attributes. *Multimedia Tools and Applications*, pages 1–13, 2022.

[23] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation.

[24] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2014.

[25] A. Kubade, D. Patel, A. Sharma, and K. Rajan. Afn: Attentional feedback network based 3d terrain super-resolution. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[26] W.-S. Lai, J.-B. Huang, O. Wang, E. Shechtman, E. Yumer, and M.-H. Yang. Learning blind video temporal consistency, 2018.

[27] C. Lei, Y. Xing, and Q. Chen. Blind video temporal consistency via deep video prior. In *Advances in Neural Information Processing Systems*, 2020.

[28] T. H. Lior Wolf and I. Maoz. Face recognition in unconstrained videos with matched background similarity. Technical report, 2011.

[29] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro. Image inpainting for irregular holes using partial convolutions.

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

[31] S. Lombardi, J. Saragih, T. Simon, and Y. Sheikh. Deep appearance models for face rendering. *ACM Transactions on Graphics*, 2018.

[32] A. Mollahosseini, B. Hasani, and M. H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017.

[33] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman. Voxceleb: Large-scale speaker verification in the wild. *Computer Science and Language*, 2019.

[34] K. Nazeri, E. Ng, T. Joseph, F. Z. Qureshi, and M. Ebrahimi. Edgeconnect: Generative image inpainting with adversarial edge learning, 2019.

[35] N. Numan, F. ter Haar, and P. Cesar. Generative rgb-d face completion for head-mounted display removal. In *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 2021.

[36] S. W. Oh, S. Lee, J.-Y. Lee, and S. J. Kim. Onion-peel networks for deep video completion, 2019.

[37] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks, 2016.

[38] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. Faceforensics: A large-scale video dataset for forgery detection in human faces. *arXiv*, 2018.

[39] A. Show. Tell: Neural image caption generation with visual attention kelvin xu. *Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, Yoshua Bengio arXiv (2015-02-10) https://arxiv. org/abs/1502.03044 v3*.

[40] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe. First order motion model for image animation. *Advances in Neural Information Processing Systems*, 2019.

[41] Z. Teed and J. Deng. RAFT: recurrent all-pairs field transforms for optical flow. 2020.

[42] J. Thies, M. Zollöfer, M. Stamminger, C. Theobalt, and M. Nießner. FaceVR: Real-Time Facial Reenactment and Eye Gaze Control in Virtual Reality. 2016.

[43] C. Wang, H. Huang, X. Han, and J. Wang. Video inpainting by jointly learning temporal structure and spatial details. In *AAAI*, 2019.

[44] M. Wang, X. Wen, and S.-M. Hu. Faithful face image completion for hmd occlusion removal. In *2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. IEEE, 2019.

[45] T.-C. Wang, A. Mallya, and M.-Y. Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[46] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 2004.

[47] Y. Wu, V. Singh, and A. Kapoor. From image to video face inpainting: Spatial-temporal nested gan (stn-gan) for usability recovery. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020.

[48] Y. Yang, X. Guo, J. Ma, L. Ma, and H. Ling. Lafin: Generative landmark guided face inpainting, 2019.

[49] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. Free-form image inpainting with gated convolution. 2018.

[50] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. Generative image inpainting with contextual attention. 2018.

[51] X. Yuan and I. K. Park. Face de-occlusion using 3d morphable model and generative adversarial network. *CoRR*, ICCV, 2019.

[52] X. Zeng, X. Peng, and Y. Qiao. Df2net: A dense-fine-finer network for detailed 3d face reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

[53] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[54] Y. Zhao, Q. Xu, W. Chen, C. Du, J. Xing, X. Huang, and R. Yang. Mask-off: Synthesizing face images in the presence of head-mounted displays. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, 2019.