Revisiting Synthetic Face Generation for Multimedia Applications

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

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

by

Aditya Agarwal
2021701011
aditya.ag@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500 032, INDIA
November 2023

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled **“Revisiting Synthetic Face Generation for Multimedia Applications”** by Aditya Agarwal, has been carried out under my supervision and is not submitted elsewhere for a degree.

Date	Adviser: Prof. C V Jawahar
Date	Adviser: Prof. Vinay P. Namboodiri

To my family

Acknowledgments

My Master's journey has been a wonderful adventure of my life. It was a difficult decision for me to take initially as that meant leaving a high-paying and comfortable job at Microsoft and delving back into the scrawny yet taxing life of a research student. As it turns out, it was one of the best decisions of my life, as I found two amazing advisors here at the CVIT Lab of IIIT Hyderabad - **Prof. C V Jawahar** and **Prof. Vinay Namboodiri**, who crafted, shaped, and even cushioned my mistakes at times. I have learnt from Prof. C V Jawahar to always "think about the bigger picture and to align my research endeavors with the bigger goal". This was a term that I did not resonate with initially, but is something that I have come to truly appreciate and understand better over time. Prof. Jawahar has been truly instrumental in setting the tone and direction of my research thesis, and in influencing my research interests. His advice on "working towards a problem that the community appreciates and to build a solution that addresses the needs of the community" has become the mantra of my research life. I have learnt a great deal from him on "driving multiple research threads" simultaneously and to convey my thoughts effectively in the weekly meetings to get the most out of my advisor's time. His wisdom, vision, and clarity have given me the opportunity to reflect on myself as a person and as a researcher. With Prof. Vinay, I always felt incredibly at ease at discussing research ideas, paper structuring, and pitching future ideas, etc. and felt that I could express even the stupidest of ideas without the problem of being judged. He has taught me to be a better researcher and a "critic of my own work". I am grateful for his trust in my intuition in pitching new research ideas and proposing solutions to the existing ones, and the advise to follow it. I have learnt from him on how to "convey my research ideas effectively", because at the end of the day "it's your job to convince the reviewers, and not his job to put effort into understanding yours". His last mile paper edits and restructuring are so intuitive and convincing, that I am still in awe of how quickly he comes up with alternate viewpoints of pitching research ideas and solutions. From him, I have learnt a great deal to enjoy and appreciate the process of doing research.

I did not have a great deal of friends here, but the ones that I made at CVIT were always cheerful, funny, and kept me going. Thanks to **Rudrabha** for being there for me as I navigated the complex manifold of research initially, often getting stuck in supoptimal minimas. **Zeeshan, Siddhant, Madhav**, and **Shubham** - you guys were the lifeline of the lab, and I will always remember our "non-alcoholic" parties and discussions. **Siddharth** and **Soumya** - thank you for all the badminton games and for keeping me motivated to stay fit initially. **Rupak, Ravi, Seshadri** - thank you for just existing in the

lab, I will miss our conversations. **Varun** - thanks for being the inquisitive junior, and for asking all the right questions.

I am also grateful for taking Mobile Robotics (MR) in my first semester. **Prof. Madhava Krishna** - I will cherish listening to you about your unique perspectives on life and the subconscious mind. Thank you for letting me be a part of your group, and for introducing me to the field of robotics.

Vishal, Shankara, and Gaurav are an amazing set of undergraduate students that I got to work with closely. Vishal and Shankara - I will remember our first robotics competition together, and how on the New Year's night, we were working on our submission, while there was loud music playing around and everyone else was partying. Gaurav - I will remember all the thought provoking research discussions, and how it challenged my own thinking and understanding of research ideas.

Bipasha has been a pillar of strength during my Master's journey. I would have never have considered the program full-time if not for her convincing. With her, my journey was very memorable and joyful. I enjoyed all our technical discussions and arguments, and our casual conversations about life. Challenging each other at every step, truly brought the best out of both of us.

A big thanks to my **dad, mom**, and my **sister** who I got to spend a great deal of time with during the Covid pandemic while working at CVIT. The comfort, warmth, and mental support that I found at my home was unparalleled. My dad is the most hard-working person in the world, and I look up to him everyday. My mom is the sweetest and the most caring person that I know. My sister would always be up for a board game or casual banter, despite her busy job. My naughty, selfish (in a cute way), and hyper-energetic **dog** reminds me of my own childhood in more ways than one. You are the glue of our family. I could not have asked for a better family.

Abstract

Videos have become an integral part of our daily digital consumption. With the widespread adoption of mobile devices, internet connectivity, and social media platforms, the number of online users and consumers has risen exponentially in recent years. This has led to an unprecedented surge in video content consumption and creation, ranging from short-form content on TikTok to educational material on Coursera and entertainment videos on YouTube. Consequently, there is an urgent need to study videos as a modality in Computer Vision, as it can enable a multitude of applications across various domains, including virtual reality, education, and entertainment. By understanding the intricacies of video content, we can unlock its potential and leverage its benefits to enhance user experiences and create innovative solutions.

Producing video content at scale can be challenging due to various practical issues. The recording process can take several hours of practice, and setting up the right studio and camera equipment can be time-consuming and expensive. Moreover, recording requires manual effort, and any mistakes made during the shoot can be difficult to rectify or modify, often requiring the entire video to be re-shot.

In this thesis, we aim to ask the question "Can synthetically generated videos take the place of real videos?" as automatic content creation can significantly scale digital media production and ease the process of content creation that can aid several applications. A form of human-centric representation that is becoming increasingly popular in the research community is the ability to generate talking-head videos automatically. Talking-head generation refers to the ability to generate realistic videos of a person speaking, where the generated video can be of a person that may not exist in reality or may exhibit significantly different characteristics than the original person. Recent deep learning approaches can synthesize synthetic talking-head videos at tremendous scale and quality, with diverse content and styles, that are visually indistinguishable from real videos. Therefore, it is imperative to study the process of generating talking-head videos as these videos can be used for a variety of applications, such as video conferencing, movie-making, broadcasting news, vlogging, and language learning among others. Consider a digital avatar reading news from a text transcript being broadcasted on news.

In this vein, this thesis aims to explore two prominent use cases of generating synthetic talking-heads automatically - the first one towards generating large-scale synthetic content to aid people in lipreading at scale. The second use case is for automating the task of actor-double face-swapping in the moviemaking industry. We study and elucidate the challenges and limitations of the existing approaches, propose solutions based on synthetic talking head generation, and show the superiority of our methods through extensive experimental evaluation and user studies.

In the first task, we address the challenges associated with learning to lipread. Lipreading is a primary mode of communication for people suffering from some form of hearing loss. Therefore, learning to lipread is an important aspect for hard-of-hearing people. However, learning to lipread is not an easy task and finding resources to improve one's lipreading skills can be challenging. Existing lipreading training websites that provide basic online resources to improve lipreading skills, are unfortunately, limited by real-world variations in the talking faces, cover only a limited vocabulary, and are available in a few select languages and accents. This leaves the vast majority of users without access to adequate lipreading training resources. To address this challenge, we propose an end-to-end pipeline to develop an online lipreading training platform using state-of-the-art talking head video generator networks, text-to-speech models, and computer vision techniques, to increase the amount of online content on the LRT platforms in an automated and cost-effective manner. We show that incorporating existing talking heading generator networks for the task of lipreading is not trivial, and requires careful adaptation. For instance, we develop an audio-video alignment module that aligns the speech utterance on the region with the mouth movements and adds silence around the aligned utterance. Such modifications are necessary to generate realistic-looking videos that don't cause distress to the lipreaders. We also design carefully thought out lipreading training exercises, conduct extensive user studies, and perform statistical analysis to show the effectiveness of the generated content in replacing the manually recorded lipreading training videos.

In the second problem, we address challenges in the entertainment industry. Body doubles play an indispensable role in the moviemaking industry. They take the place of actors in dangerous stunt scenes and in scenes where the same actor plays multiple characters. In all these scenes, the double's face is later replaced by the actor's face and expressions using CGI technology requiring hundreds of hours of manual multimedia edits on heavy graphical units costing millions of dollars and taking months to complete. As we show in this thesis, automated face-swapping approaches based on deep learning models are not suitable for the task of actor-double face-swapping, as they fail to preserve the actor's expressions. To address this, we introduce "video-to-video (V2V) face-swapping", a novel task of face-swapping that aims to (1) swap the identity and expressions of a source face video, and (2) retain the pose and background of the target face video. Our key technical contribution lies in i) devising a self-supervised training strategy, which uses a single video as the source and target, introduces pseudo motion errors on the source video, and the network fixes these pseudo errors to regenerate the source video; and ii) we build temporal autoencoding models inspired by VQVAE-2, that take two different motions as input, and produce a third coherent output motion.

In summary, this thesis unravels several tasks enabled by synthetic talking-head generation, and provides solutions for the lipreading community and the moviemaking industry. Our findings concretely point toward the notion of replacing real human talking-head videos with synthetically generated videos, thereby, scaling digital content creation to new heights, saving precious time and resources, and easing the life of humans.

***Keywords* –** lipreading, talking-head generation, face-swapping, video-to-video face-swapping, online lipreading training, digital content generation

Contents

List of Figures

Figure Page

List of Tables

Chapter 1

Introduction

The recent advancements in generative networks have transformed the way digital content is produced at scale. Previously, generating such content was a costly, time-consuming process, requiring specialized skills, extensive computational power, and manual effort. However, generative models can now be used to generate realistic images of objects, landscapes, and people for video games and movies, creating virtual worlds and avatars. In addition, text generation models are useful in technologies such as chatbots, language translation, and content generation on social media platforms. Music can be generated efficiently for production companies that need to create original music quickly. Generative networks can also augment existing datasets by generating new data points. These networks have paved the way for new forms of creative expression, allowing artists and designers to explore new styles and techniques, creating works that were previously impossible to produce. For example, generative art is becoming increasingly popular, with artists using GANs to create unique and abstract works that challenge traditional notions of art. The advancements in generative networks have opened up new opportunities for businesses, artists, and individuals to create and share content in exciting and innovative ways. As the technology continues to evolve, we can expect to see even more creative applications of generative networks in the future. Videos constitute the vast majority of the digital content that we consume everyday. With the boom in mobile penetration, internet connectivity, and increasing number of internet users and social media applications, video based content is being consumed at an enormous rate. However, creating high-quality video content manually is an expensive and time-consuming process. Therefore, it is important to study video generation techniques in Computer Vision. Automated video generation has the potential to revolutionize several industries such as film and entertainment, by generating dance videos, special effects, and realistic animations at scale. One such application of tremendous social impact that we explore in this thesis, is lipreading. We develop an end-to-end pipeline to produce synthetic lipreading content at scale, by utilizing existing talking-head generators and text-to-speech models. We show that adapting these models for the task of lipreading is non-trivial, and we develop computer-vision modules to address these. With the recent advancements in generative models, automatic video generation has become a reality [54, 59, 69, 48]. Moreover, the technology allows for generating videos conditioned on different input modalities such as speech or text, offering several applications such as generating cricket videos based on text commentary, or generating sign language videos automatically conditioned on text. By automating the process of video content creation, businesses and individuals can create high-quality video content efficiently and affordably, further democratizing access to high-quality content creation.

A large part of these videos that are gaining immense popularity and becoming seamlessly accessible are human-centric videos, known as “talking-head videos”. These include vlogs, lectures, news broadcasting, historical speeches and public talks, public conferences, interviews, and face-to-face conversational videos, such as video calls. One such application that we propose in this thesis is the task of video-to-video face-swapping, that requires video face-swapping by preserving the identity and expressions of the source (actor) and the background and pose of the target (double) video. We show that such a task is non-trivial as it involves merging two different motions - the actor’s face motion (such as eye, cheek, or lip movements) and the double’s head motion (such as pose and jaw motion). Our key contributions for solving this task are two fold - i) we propose a self-supervised training scheme, that uses a single video as source and target, and ii) temporal autoencoding modules that take in two different motions as input and produce a third coherent output motion.

We explore two practical applications arising out of synthetic talking face generation. Firstly, we investigate the challenges associated with traditional lipreading training platforms and question whether synthetically generated talking heads could replace real human talking-head videos. We develop a pipeline that utilizes a state-of-the-art talking head generator module to produce synthetic lipreading content at scale. We argue that manually created lipreading content on existing MOOC platforms are available in a few select languages and accents, have limited vocabulary and real-world variations, and are expensive and time-consuming to develop from scratch. We perform extensive user evaluations and statistical analysis, and concretely demonstrate the capability of an automated synthetic talking head generation pipeline in replacing real human talking-head videos. We explore another use case of synthetic talking heads in the entertainment industry. The ability to change the expressions or hair color of a person in post-production offers numerous possibilities in film-making and entertainment. One such application is actor-double face-swapping, which involves swapping the identity and expressions of the starring actor on the double’s face while retaining his pose and background information. Currently, movie producers use CGI technology requiring hundreds of hours of manual multimedia edits on heavy graphical units costing millions of dollars and taking months to complete. Thus, the production team is generally forced to avoid such scenes by changing the mechanics of the scene such that only the double’s body is captured to provide an illusion of the actor. We show that existing automated face-swapping methods swap only the identity without preserving the source expressions of the actor important for the scene’s context [41, 51, 39, 38, 33, 5]. In this vein, we introduce a new line of research, called video-to-video (V2V) face-swapping, a novel task of face-swapping that preserves the identity and expressions of the source (actor) face video and the background and pose of the target (double) video.

While deep learning has its own set of challenges associated with synthetically generating content such as deepfakes, impersonation, and forgery, these models have far more useful applications, which is the focus and inspiration of this thesis. However, we must be cautious about the impact of content generation and modification technologies on public discourse and people's rights, especially since deepfakes have been used maliciously as a source of misinformation, manipulation, harassment, and persuasion. Although a plethora of anti-forgery detection and deepfake detection techniques have been developed in recent years, our work focuses solely on the aspects of creative and assistive technologies enabled by these methods.

1.1 Synthetic Talking Head Generation

Talking-head generation is a type of video generation that involves creating a digital video of a person's head and upper body that appears to be speaking, while also accurately matching the audio content. Recent talking-head generation models [12, 18, 67, 42, 14] are trained on large amounts of video and audio data and generate realistic-looking videos that are visually indistinguishable from real videos. In this thesis, we explore two important use cases of generating synthetic talking heads automatically - the first one towards generating large-scale synthetic content to aid people in lipreading, and the second one for the use case of actor-double face-swapping in the movie making industry.

1.1.1 Conditioning Synthetic Talking Heads

The generation of synthetic talking heads conditioned on input modalities such as speech or text offers several interesting capabilities. For instance, generating talking-head videos with accurate lip-sync for any input speech segment, can be used for the task of automated movie-dubbing. In audio-driven talking face generation, the expressions, pose, and lip-sync in the target video are conditioned on the given input speech audio [43, 26, 34, 74, 56, 45, 20]. One of the recent works, Wav2Lip [43] accurately morphs the lip movements of the identities to be in sync with the corresponding speech for arbitrary identities in unconstrained settings. MakeItTalk [75] generates speaker-aware expressive talking-head videos from a single facial image with audio as the only input. [55] trains a joint system combining a talking face generation system with a text-to-speech system that can generate multilingual talking face video from only the text input. These works can synthesize natural multilingual speeches while maintaining the vocal identity of the speaker, as well as the lip movements synchronized to the given input speech. In this thesis, we propose a novel approach to automatically generate a large-scale database using synthetic talking heads for developing a lipreading training MOOCs platform. We use SOTA text-to-speech (TTS) [8] models and talking head generators [43] to generate training examples of driving face videos lip-synced to the driving speech segments automatically. We show that such an approach enables the scaling of lipreading training platforms to more identities, accents, languages, incorporate larger vocabulary, and accommodate variations in the generated content, making the process of lipreading training more rigorous.

1.1.2 Talking Head Manipulation, Editing, & Reenactment

Manipulating, editing, reenacting talking-head videos have several applications in the moviemaking and the entertainment industry. These approaches typically encode the input modality to a suitable latent space (that encodes rich semantics), followed by applying careful and meaningful edits in the underlying latent space, and realizing these edits in a coherent manner to learn the required transformation. Face manipulation animates the pose and expressions of a target image/video according to a given prior [64, 52, 50, 65, 41, 73, 56, 75]. [53] propose a one-shot facial geometry-aware emotional talking face generation method that can generalize to arbitrary faces. They provide speech content feature, along with an emotion input to generate emotion and speech-induced motion. Read Avatars [46] generate photo-realistic and lip synchronized video from audio and a reference video, with control over emotion. The same audio is used to generate videos in multiple emotions with control over the intensity of the emotion. Wav2Lip-Emotion [34] modifies facial expressions of emotion in videos of speakers. They tackle the important use case of modifying an actor's performance in post-production, by proposing a video-to-video translation approach, that is used to touch up emotion in the output video. A different direction of face reenactment animates the source face movements according to the driving video [57, 45, 58, 28, 50, 52]. The identity is not exchanged in these works. In FOMM [50], the target video sequence is animated based on the input source image's appearance information and the driving video's motion patterns. PCAVS [74] propose to generate pose-controllable talking faces. They modularize audio-visual representations by devising a low-dimensional pose latent code, that is complementarily learned. Their method generates accurately lip-synced videos whose poses are controllable by other videos. Face-Vid2Vid [64], is a SOTA face reenactment network that reenacts the pose and expressions of a target image using a source (driving) video. Another direction is face editing, which involves editing the expressions of a face video [23, 9, 10, 37]. Using this approach, one can directly edit the target video according to the source expressions. Image-based face editing works have gained considerable attention. However, realizing these edits on a sequence of frames without modeling the temporal dynamics often results in temporally incoherent videos. Recently, STIT [62] was proposed that can coherently edit a given video to different expressions by applying careful edits in the video's latent space. Their method modifies the expressions of a face video based on an input label. In this thesis, we propose a novel method for the task of actor-double face-swapping in the moviemaking industry called video-to-video face-swapping that relies on generating face-swapped synthetic talking heads.

1.1.3 Face Swapping and its applications

Manipulating digital images, especially the manipulation of human portrait images, has improved rapidly and has achieved photo-realistic results in most cases. A class of research on synthetic talking heads deals with an important and highly relevant aspect of face-swapping. Face swapping is an eye catching task that involves swapping the faces of two individuals in a photograph or a video, or replacing one person's face with another person's [41, 39, 51, 6, 30, 33, 38, 5, 7]. The use of face-swapping techniques has gained widespread popularity on the internet due to their highly realistic and intriguing results, with many amusing and entertaining spoof videos being shared on platforms like YouTube and TikTok. Additionally, there are now various commercial applications such as Snapchat and FaceApp that allow users to easily create fake images and videos using this technology.

1.1.3.1 Image Face Swapping

Swapping faces across images and videos has been well-studied over the years. These works aim to swap an identity obtained from a source video (or an image) with a target video of a different identity such that all the other target characteristics are preserved in the swapped output. Face swapping is defined as the task of replacing a source face's identity the a target face while retaining the content, pose, and background characteristics of the source face. In 2018, DeepFakes introduced a pipeline in replacing a source person's face with the target person's along with the same facial expression such as eye movement, facial muscle movement. DeepFaceLabs [41] introduced a technique to build enormous amount of high-quality face swapping videos for entertainment. They introduced an integrated open-source system with a clean-state design of the pipeline, achieving photorealistic face-swapping results without painful tuning. DFL has gained significant traction in the research and non-research community, with DFL-based YouTube videos garnering over 100 million hits on YouTube. Approaches like Deepfakes and Deepfacelabs require inference time optimization and finetuning and are not realtime. On the other hand, FSGAN [39] performs subject agnostic face-swapping on a pair of faces without requiring training on those faces. All these approaches however swap the entire identity of the source. Motion-coseg [51] specifically swaps the identity of single/multiple segments of a given source image (either hair or lips or mouth or nose) to a target video. However, all these approaches swap only the identity or a specific part of an image, which is not suitable for moviemaking industry where the expression of the source face need to be retained. While these works have achieved impressive results, they are not suitable for the task of face-swapping in the moviemaking industry, as these methods end up losing the source actor's expressions, which are of utmost importance. In this thesis, we introduce a novel task of video-to-video face-swapping where we swap temporally changing expressions along with the identity of the source taking less than a second to face swap in-the-wild videos on unseen identities.

1.1.3.2 Video Face Swapping

Unlike existing face-swapping approaches that swap a fixed identity component from one video to another video, "video-to-video (V2V) face-swapping" swaps expressions changing over time (a video) with another video with changing pose and background. Swapping faces across videos is non-trivial as it involves merging two different motions - the actor's face motion (such as eye, cheek, or lip movements) and the double's head motion (such as pose and jaw motion). This needs a network that can take two different motions as input and produce a third coherent motion. Fundamentally, a V2V face-swapping task, aims to (1) swap the identity and expressions of a source face video and (2) retain the pose and background of the target face video. In this thesis, we propose FaceOff, a V2V face-swapping system that operates by learning a robust blending operation to merge two face videos to a quantized latent space and then blends them in the reduced space, by being trained in a self-supervised manner.

1.2 Use Cases of Synthetic Talking Heads Covered in this Thesis

In this thesis, we explore two useful and important use cases of synthetic talking heads. The first is a fully automated approach towards building a large-scale lipreading training platform [2] that trains hard-of-hearing people to lipread better. The second is an automated video-to-video face-swapping approach for the moviemaking industry that tackles the challenging task of actor-double face-swapping [1], which is currently handled using expensive CGI techniques.

1.2.1 Lipreading training platform

Lipreading is a primary mode of communication for people with hearing loss. Many people with some form of hearing loss consider lipreading as their primary mode of day-to-day communication. The United States of America alone is home to 48 million people with some form of hearing loss, and about 500,000 Americans have a disabling hearing loss that noticeably disrupts communication. However, learning to lipread is not an easy task and finding resources to learn or improve one's lipreading skills can be challenging. People needing these skills undergo formal education in special schools and involve medically trained speech therapists. Today, online MOOCs platforms like Coursera and Udemy have become the most effective form of training for many types of skill development, and consequently, there exist several online lipreading training platforms. Inspired by the boom in online courses available for virtually every topic, we envision a MOOCs platform for LipReading Training (LRT) for the hearing disabled. Platforms like lipreading.org and lipreadingpractice provide basic online resources to improve lipreading skills. Unfortunately, these platforms cover only a limited vocabulary, the videos have minimal real-world variations in head-pose, camera angle, and distance to a speaker, making it difficult for a lipreader to adapt to the real world. Moreover, these resources are all available only in American or British-accented English, and it becomes challenging for people from other regions to adapt to their local accents and languages. Moreover, creating such resources is an extensive ordeal needing months of manual effort to record hired actors, requires expensive camera and studio environment, and professional editors. Because of the manual pipeline, such platforms are also limited in vocabulary, supported languages, accents, and speakers and have a high usage cost.

We approach this from a different angle and ask: "Can we replace real talking head videos used for training people suffering from hearing loss with synthetic versions of the same?" In this work, we propose an end-to-end automated pipeline to develop such a platform using state-of-the-art talking head video generator networks, text-to-speech models, and computer vision techniques, to generate training examples automatically. Our approach exponentially increases the amount of online content on the

LRT platforms in an automated and cost-effective manner. We also seamlessly increase the vocabulary and the number of speakers in the database. We also investigate the implications of our system for a range of deaf users and perform multiple experiments to show its effectiveness in replacing the manually recorded LRT videos. We complement these results through statistical analysis on extensive human evaluations carried out on carefully thought out lipreading exercises, and show that the users' performance on lipreading videos is not significantly different when switching from 'real' to 'generated' videos. Concretely our studies point toward the potential of synthetic talking head videos in replacing real human talking-head videos, which can be developed into a large-scale lipreading MOOCs platform that can potentially impact millions of people with hearing loss.

1.2.2 FaceOff: A novel task of video-to-video face-swapping

Doubles play an indispensable role in the movie-making industry. They take the place of the actors in dangerous stunt scenes. For instance, stunt doubles performed difficult and dangerous life-risking stunt scenes for actors Christian Bale and Arnold Schwarzenegger in the Batman and Terminator movies respectively. Similarly, they take the place of actors in movie scenes, where the same actor plays multiple characters, or for scenes that require multiple retakes. For example, Óscar Isaac played multiple characters with different personalities in the 'Moon Knight', Armie Hammer who played multiple roles of twin brothers in 'The Social Network', and Christian Bale and his body double who played the role of twin magician brothers in the movie 'The Prestige'. A different scenario is post-production scene modifications. If a dialogue is discovered in post-production that suits a scene better than the original, the entire scene is reset and re-shot. In the movie 'Justice League', Henry Cavill who plays Superman, had his moustache digitally erased using CGI as he was filming for a concurrent film 'Mission: Impossible - Fallout', which required him to grow a moustache. Yet another use case is using CGI to cast doubles to fill in for actors that are no longer available to shoot a particular movie scene. For instance, Paul Walker was replaced by his brothers, Caleb and Cody Walker, as his body doubles using CGI in 'Furious 7' due to Paul Walker's untimely demise in a car crash. It is for these reasons that producers and filmmakers are no strangers to using CGI and tricky digital maneuvers. In all of these scenes, the double's face is later replaced by the actor's face and expressions using CGI technology requiring hundreds of hours of manual multimedia edits on heavy graphical units costing millions of dollars and taking months to complete. For comparison, the Burly Brawl sequence in the classic movie 'Matrix Reloaded', which combined motion capture and CGI, took 27 days to film and cost a whooping $40 million to make. It is one of the most expensive action scenes in film. Thus, many a times, the production team is generally forced to avoid such scenes by changing the mechanics of the scene such that the double's body is captured to provide an illusion of the actor. This negatively impacts director's creativity; however, such adjustments are not always possible. An automated, inexpensive, and fast way can be to use face-swapping techniques that aim to swap an identity from a source face video (or an image) to a target face video. Fast and inexpensive computer vision based face-swapping techniques that aim to swap an identity between a source (actor) video and target (double) video can be considered. However, face-swapping swaps only the source identity while retaining the rest of the target video characteristics. In this case, the actor's expressions (source) are not captured in the output. This is because, existing face-swapping techniques use a discriminator-generator setup due to the absence of ground truth. The discriminator is responsible for monitoring the desired characteristics of the swapped output. However, using a discriminator leads to hallucinating components of the output that are different from the input - for instance, a modification in the identity of novel expressions. To tackle this, we introduce "video-to-video (V2V)" face-swapping as a novel task of face-swapping that aims to (1) swap the identity and expressions of a source face video and (2) retain the pose and background of the target face video. Unlike the face-swapping task that swaps a fixed identity component from one video to another, V2V face-swapping swaps the expressions changing over time (a video) with another video with changing pose and background (another video). FaceOff, a video-to-video face swapping system that can take two different motions as input and produce a third coherent output motion. Due to the absence of the ground truth, we devise a self-supervised training strategy for training our network, where we use a single video as the source and target. We then introduce pseudo motion errors on the source video, and train a network to 'fix' these pseudo errors to regenerate the source video.

1.3 Alternate applications (not covered in this Thesis)

In this section, we cover two alternate use cases of generating multimedia content automatically at scale. In the first work not covered in this thesis, we consider the task of learning a representation space of videos for automatically generating video data at scale [48]. In the second work, we rely on a data augmentation technique, for synthetically generating lipreading videos to augment the one-shot scenario of lipreading a patient suffering from ALS [47]. As discussed earlier, generating multimedia video content is a complex task that is accomplished by generating a set of temporally coherent images frame-by-frame [59, 61, 13, 54, 15, 69, 66]. Building a representation space for videos and conditioning it on different input modalities offers several exciting possibilities. For instance, consider generating sign-language videos conditioned on the textual input [27], or generating unconstrained sports videos conditioned on the text commentary. Existing state-of-the-art works treat video generation as the task of generating a sequence of temporally coherent frames. However, these methods come with a major limitation: they rely on an image space. This limits the applications of the learned space to image-based operations, such as animating images or editing on frames. Operations that require interpolating intermediate videos between two videos and generating future segment of a video, become difficult. To tackle this, we propose that videos can be represented as a single unit instead of being broken into a sequence of images. However, with existing video generator architectures, learning such a representation is difficult. In this work, we parameterize videos as a function of space and time using implicit neural representations (INRs). The dynamic dimension of videos (a few million pixels) is now reduced to a constant number of weights (a few thousand) required for the parametrization. A network can then be used to learn a prior over videos in this parameterized space [21].

In the second work, we consider the task of lipreading an ALS patient with very few examples. Lipreading is the task of visually recognizing speech from the mouth movements of a speaker. As discussed in Sec 1.2.1, lipreading is a mentally taxing exercise, requiring hundreds of hours of practice. In certain cases, patients suffering from neurologically degenerative diseases such as 'Amyotrophic Lateral Sclerosis' (ALS) [70, 36] often lose muscle control, consequently their ability to generate speech and communicate via lip movements. In such cases, talking to a person without voice may need you to to lipread them to understand the spoken words [22]. Applications and automated algorithms capable of lipreading a person can thus significantly improve the day-to-day communication of people dependent on lipreading. However, existing deep learning techniques [35, 17, 60] are inherently data-hungry, and require collecting large amounts of patient-specific data, as existing datasets on which these models are trained, don't capture the irregular and unreliable mouth movements exhibited by people suffering from medical disabilities. Mouthing words however is a tiring maneuver for people suffering from ALS, and thus a patient undergoes physical and mental stress during such data collection exercises. Manually labeling words mouthed by a person is time-consuming. It is thus crucial to build lipreading models that can work well on the minimum amount of manually labeled data. Inspired by several recent works that can generate realistic looking synthetic talking faces, we ask ourselves the question - "Can we use synthetic data to augment low data in the medical domain?". We use synthetically generated data along with very limited real world examples to train word-level lipreading models. We also adopt a domain adaptation technique to reduce the domain gap between the real and synthetically generated examples. We observe that the synthetic data helps the model learn the general underlying word-level characteristics of the classes, and the one-shot examples introduce the properties of personal speaking style in the model. Overall, we observe that augmenting existing one-shot data with synthetically generated examples, greatly improves the performance of our speaker-specific lipreading models.

1.4 Contributions

In this thesis, we revisit the role of synthetically generated talking head videos in several multimedia applications, and propose novel methods to address key challenges in existing works. Our core contributions are as follows -

- We propose two important use cases of synthetic talking heads - in lipreading training and in the movie-making industry.

- In the first work, we propose a novel approach of using synthetic talking heads to automatically generate a large-scale database for developing a lipreading training MOOCs platform.

- Our approach exponentially increases the amount of online content on an LRT platform in an automated and cost-effective manner, while seamlessly increasing the vocabulary and incorporating more number of speakers in the database. Additionally, our approach scales existing LRT platforms to incorporate multiple accents and languages.

- We show that naively adapting talking-head generator models is non-trivial, and design an audio-video alignment model that aligns the speech utterance on the region with the mouth movements and add silence around the uttered utterance. This is important in generating photorealistic videos that cause minimal distress to the lipreaders.

- We design carefully thought out lipreading training exercises to validate the design of our platform, and perform extensive human evaluations, concretely pointing toward the potential of our approach in developing a large-scale lipreading MOOCs platform.

- In the second work, we highlight several shortcomings of existing techniques in actor-double face-swapping, and introduce a novel task of video-to-video face-swapping.

- We devise a self-supervised training strategy that involves using a single video as the source and target. We introduce pseudo motion-errors on the source video, and train the network to 'fix' these pseudo errors to regenerate the source video.

- We show that the task of merging two different motions is non-trivial, and develop temporal autoencoding modules that take as input two different motions, and produce a third coherent output video.

1.5 Organization of Thesis

- In Chapter2, we introduce the task of using synthetic talking heads for training humans in lipreading. We analyze why lipreading is an important tool for individuals who are hard hearing and analyze critical drawbacks in existing lipreading training platforms. We propose a novel framework to automatically generate lipreading content at scale, and show the strengths through statistical anaylsis. Fundamentally, we ask the question "Can synthetically generated talking-head videos replace real human talking-head videos?"

- In Chapter3, we tackle the task of actor-double face-swapping in the moviemaking industry. We introduce a new line of research called video-to-video (V2V) face-swapping, that preserves the identity and expressions of the source, and pose and background of the target.

- In Chapter4, we present our concluding thoughts.

Chapter 2

MOOCs for Lipreading: Using Synthetic Talking Heads to Train Humans in Lipreading

In this chapter, we elaborate the task of using synthetic talking heads for training humans in lipreading. We first analyze why lipreading is an important tool for individuals who are hard of hearing, and how lipreading training platforms serve to train people who are hard of hearing to lipread better. We then analyze some critical drawbacks in existing lipreading training platforms, severely limiting their use and real-world applicability. We then explain our proposed approach "LRT MOOCs" for automatically generating lipreading content at scale, and show through statistical analysis the benefit of our proposed approach.

2.1 Introduction

Communication is a crucial ingredient that makes Humans the most intelligent species on the planet. While other animals also have different forms of communication, human language is more advanced by several orders of magnitude. But we are not inherently born with these skills! Then, how do we acquire them? Most of us learn linguistic skills through a formal education system consisting of schools, universities, and other organizations related to education. While this is still the most trusted & popular way of imparting education, the 21st century has seen an exponential rise in online forms of education like the Massive Open Online Courses (MOOCs). Online courses are generally designed to cover hundreds of topics in various domains, including language, and are often available free of cost. MOOCs have several advantages over the physical form of education. They are more accessible, cheap, and reachable to a broader audience. In today's world, it is natural to learn a whole new language from the comfort of your home by attending a high-quality MOOCs course.

Unfortunately, every person does not get the chance to learn linguistic skills like we usually do. Hearing loss is a common form of disability that can become a massive barrier to education! According

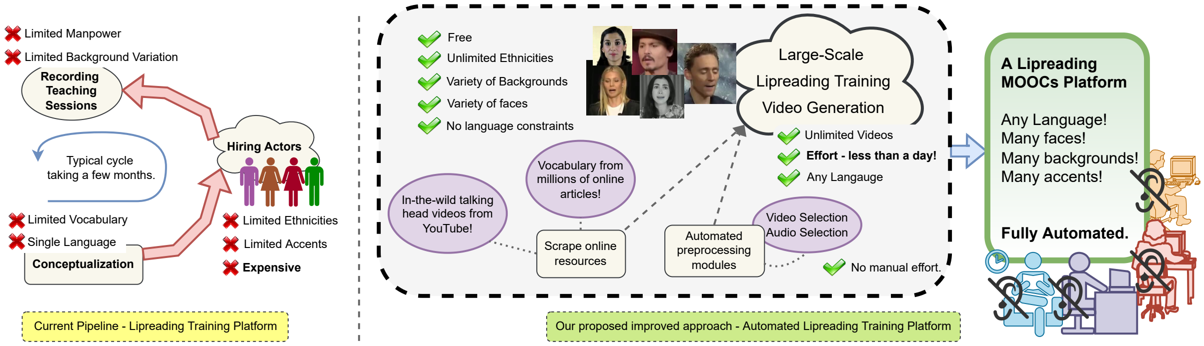
Limited Manpower
Limited Background Variation
Recording Teaching Sessions
Typical cycle taking a few months.
Hiring Actors
Limited Vocabulary
Single Language
Conceptualization
Limited Ethnicities
Limited Accents
Expensive
Current Pipeline - Lipreading Training Platform
Free
Unlimited Ethnicities
Variety of Backgrounds
Variety of faces
No language constraints
Large-Scale Lipreading Training Video Generation
Unlimited Videos
Effort - less than a day!
Any Langauge
Vocabulary from millions of online articles!
In-the-wild talking head videos from YouTube!
Scrape online resources
Automated preprocessing modules
Video Selection Audio Selection
No manual effort.
Our proposed improved approach - Automated Lipreading Training Platform
A Lipreading MOOCs Platform
Any Language!
Many faces!
Many backgrounds!
Many accents!
Fully Automated.

Figure 2.1: Lipreading is a primary mode of communication for people with hearing loss. The United States of America alone is home to 48 million people with some form of hearing loss. Despite these staggering stats, online lipreading training resources are scarce and available for only a handful of languages. However, hosting new lipreading training platforms is an extensive ordeal that can take months of manual effort. We propose a fully-automated approach to building large-scale lipreading training platforms. Our approach enables any language, any accent, and unlimited vocabulary on any identity! We envision a lipreading MOOCs platform to enable millions of people with hearing loss across the globe. In this work, we thoroughly analyze the viability of such an approach.

to organizations like WHO[1] and Washington Post[2], over 5% of the world's population (432 million adults and 34 million children) and at least 48 million Americans are deaf with some form of hearing loss. About $500,000$ Americans have a disabling hearing loss that noticeably disrupts communication.

2.1.1 Lipreading as a Mode of Communication

Lipreading is a primary mode of communication for people with hearing loss. The Scottish Sensory Censor (SSC)[3] quotes, "whatever the type or level of hearing loss, a child is going to need to lipread some of the time." However, learning to lipread is not an easy task! Lipreading can be thought of being analogous to "learning a new language" for people without hearing disabilities. People needing this skill undergo formal education in special schools and involve medically trained speech therapists. Other resources like daily interactions also help understand and decipher language solely from lip movements. However, these resources are highly constrained and inadequate for many patients suffering from hearing disabilities.

Inspired by the boom in online courses available for virtually every topic, we envision a MOOCs platform for **Lip**Reading **T**raining (LRT) for the hearing disabled.

[1] Deafness and Hearing Loss | WHO
[2] As wearing masks becomes the norm, lip readers are left out!
[3] Factors which help or hinder lipreading | SSC

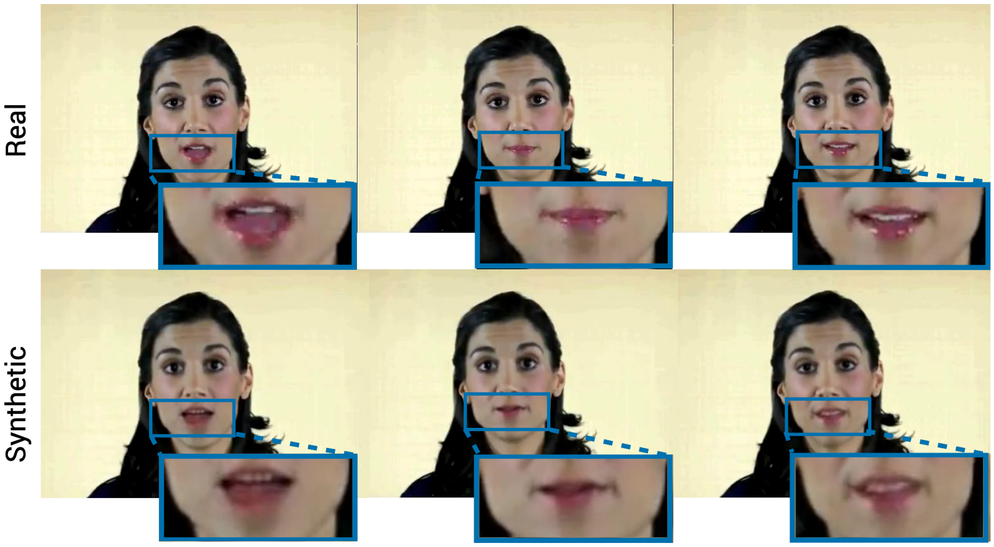

Figure 2.2: Talking-face video generated using our pipeline.

2.1.2 Limitations of Current Online Platforms for Lipreading Training

Platforms like lipreading.org[4] and lipreadingpractice[5] provide basic online resources to improve lipreading skills. These platforms allow users to learn limited levels of lipreading constrained by resources. Unfortunately, the amount of vocabulary systematically covered during the exercises is extremely narrow. The videos also have minimal real-world variations in head-pose, camera angle, and distance to a speaker, making it difficult for a lipreader to adapt to the real world. Finally, since these resources are all available only in American or British-accented English, it becomes challenging for people from other regions to adapt to their local accents and languages. All the above factors severely limit the quality of human training. Therefore, we believe it is quintessential to scale the current lipreading training platforms to incorporate extensive vocabulary and introduce variation in videos, languages, and accents. However, recording videos is a costly affair. It requires expensive camera equipment, studio environments, professional editors, and a substantial manual effort from the perspective of a speaker whose videos are being recorded.

2.1.3 Major difference: Replacing real talking head videos with synthetic versions

To resolve this issue, we approach this from a different angle and ask: "Can we replace real talking head videos used for training people suffering from hearing loss with synthetic versions of the same?" A synthetic talking head with accurate lip synchronization to a given text or speech signal can enable the scaling of LRT platforms to more identities, accents, languages, speed of speech, etc., making the training process more rigorous. We take advantage of the massive progress made by the computer vision

[4] lipreading.org

[5] lipreadingpractice.co.uk

community on synthetic talking head generation and employ a state-of-the-art (SOTA) algorithm [43], as mentioned below.

We propose a novel approach to automatically generate a large-scale database for developing an LRT MOOCs platform. We use SOTA text-to-speech (TTS) models [8] and talking head generators like Wav2Lip [43] to generate training examples automatically. Wav2Lip [43] requires driving face videos and driving speech segments (generated from the TTS in our case) to generate lip-synced talking head videos according to the driving speech. It preserves the head pose, background, identity, and distance of the person from the camera while modifying only the lip movements, as shown in Fig. 2.2.

2.1.4 Our Contributions

- Our approach can exponentially increase the amount of online content on the LRT platforms in an automated and cost-effective manner, and can seamlessly increase the vocabulary and the number of speakers in the database.

- We show that naively incorporating talking-head generator models for generating lipreading content is non-trivial, and design an audio-video alignment module that aligns the speech utterance on the region with the mouth movements and adds silences around the aligned utterance.

- We propose a large-scale database for developing an LRT MOOCs platform and conduct an extensive user study on carefully thought out lipreading training exercises.

- We show through statistical analysis that (1) the users' performance on lipreading videos is not significantly different when switching from 'real' to 'generated' videos, and (2) the benefit of lipreading platforms in one's native accent through an extensive user study.

2.2 Our Proposal: A Synthetic Talking Head Database

Our lipreading training database generation pipeline: (1) Scrapes a set of face videos automatically from the internet. This helps us cover a large number of identities, background variations, lip shapes, etc. (2) Post-processes the scraped videos to filter out invalid faces (such as drastic pose changes). (3) Automatically curates a vocabulary of many words and sentences from various online sources. (4) Generates synthetic speech utterances on the curated vocabulary. (5) Selects a driving face video and a speech utterance to generate synthetic talking head videos using a SOTA talking head generation model, Wav2Lip, in our case. Wav2Lip modifies the lip movements of the driving video according to the speech utterance. The rest of the video (background, pose, etc.) is retained. These synthetic videos (with or without speech) are used to train humans in lipreading. The overall pipeline is illustrated in Fig. 2.3.

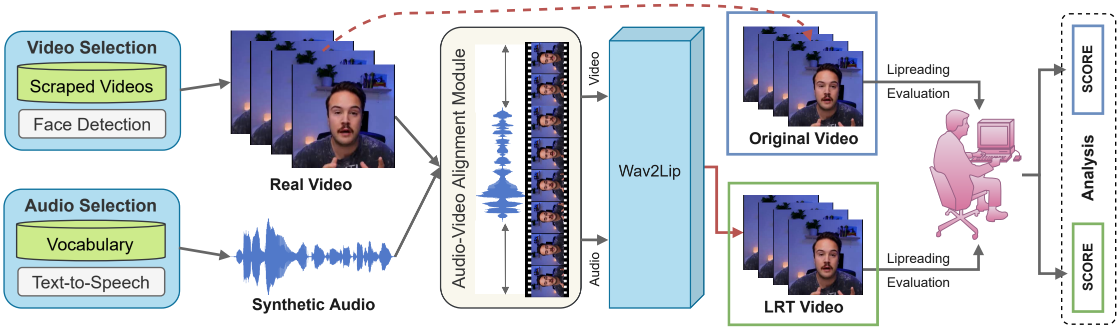

Figure 2.3: Proposed pipeline for generating large-scale lipreading training platform: **(a) Video Selection**: Videos are scraped from various online sources (such as YouTube) and invalid videos are filtered out. **(b) Audio Selection:** Synthetic speech utterances are generated using vocabulary curated from various online articles. **(c) Audio-Visual Alignment Module**: A video and a speech utterance is selected and aligned on each other such that the speech utterance overlaps with the region in the video with lip movements. **(d) Wav2Lip:** A state-of-the-art talking head generation model that modifies the lip movements of the video according to the speech utterance. **(e) User Evaluation:** A validation step to ensure that users perform comparably on real videos and synthetic videos generated using our approach.

2.2.1 Text-to-Speech System

We evaluate several TTS models: Fastspeech2 [8], Real time voice cloning [25], Glow-tts [29], and Tacotron2 [49] trained on LibriTTS [71] and LJSpeech [24]. We evaluate them at different speeds - 1×, 1.5×, 1.7×, 2×, pitch, and volume variations. We collect qualitative feedback from 30 participants without any hearing loss on the clarity of the generated speech and report the Mean Opinion Scores (MOS) in the supplementary. For our experiments conducted on American-accented English, we use Fastspeech2 with 1× speed configuration pretrained on LJSpeech. For Indianised English accent, we use an online TTS[7] with qualitatively similar performance to the speech generated by FastSpeech2. The TTS models used in our pipeline are configurable plug-and-play modules and can be replaced with any other TTS. This allows scalability and variations with little to no manual effort.

2.2.2 Synthetic Talking Head Videos

Since 2015, talking head generation models that modify the lip movements according to a given speech utterance have gained much traction in the computer vision community [32, 19, 68]. While some of these works generate accurate lip-sync, they are trained for specific speakers requiring large amounts of speaker-specific data. [3] can be remodeled for generating talking heads but require far more manual intervention limiting their use in our approach. Recent advances like LipGAN [26] and Wav2Lip [43] are perfect for our approach since they work for any identity without requiring speaker-specific data. Consequently, we adopt Wav2Lip in our pipeline. Wav2Lip takes a face video of any identity (driving face video) and audio (guiding speech) as inputs. The model then modifies the lip movements in the original video to match the guiding speech, as shown in Fig. 2.2. The rest of the video features, such

[7] http://ivr.indiantts.co.in/en/

as the background, identity, and face pose, are preserved. The algorithm also works for TTS-generated speech segments essential for our case.

2.2.3 Data Generation Pipeline

Data Collection Module: Random videos are first collected from various online sources such as YouTube. These random videos introduce real-world variations a lipreader encounters in real life, such as variations in the head-pose of the speaker, speaker's distance from the camera (lipreader), speaker's complexion, and lip structure. We post-process these videos with a face-detection model to detect valid videos. Valid videos are single-identity front-facing talking head videos with no drastic pose changes. Speech utterances are generated using TTS models on vocabulary curated automatically from online sources.

Audio-Video Alignment Module: In our next step, we randomly select a pair of driving speech and a face video. To generate lip-synced videos using Wav2Lip, we match the video and speech utterance length by aligning them and then padding the speech utterance with silence. Naively aligning the speech utterance on the driving video can lead to residual lip movements, as shown in Fig. 2.4, 'Misaligned Video' row. Wav2Lip does not modify the lip movements in the driving video in the silent region. As a result, the output contains residual lip movements (indicated in the red box) from the original video. This can confuse and cause distress to the user learning to lipread. Our audio-video alignment module aligns the speech utterance on the video region with lip movements, as shown in Fig. 2.4, 'Aligned Video' row. This way, Wav2Lip naturally modifies the original mouth movements to correct speech-synced mouth movements while keeping the regions with no mouth movements untouched.

We use lip-landmarks and the rate of change of the lip-landmarks between a predefined threshold of frames to detect mouth movements in the face videos. Once we have detected lip movements, we align the audio on the detected video region and add silences around the speech.

Data Generation: The aligned speech utterance and the face video are passed through Wav2Lip. Wav2Lip modifies the lip movements in the original video and preserves the original head movements, background, and camera variations, thus allowing us to create realistic-looking synthetic videos in the wild. Overall pipeline is illustrated in Fig. 2.3.

2.3 Evaluations: Human Lipreading Training

Lipreading is an involved process of recognizing speech from visual cues - the shape formed by the lips, teeth, and tongue. A lipreader may also rely on several other factors, such as the context of the conversation, familiarity with the speaker, vocabulary, and accent. Thus, taking inspiration from lipreading.org and readourlips.ca[8], we define three lipreading protocols for conducting a user study to evaluate the viability of our platform - (1) lipreading on isolated words (WL), (2) lipreading sentences

[8] https://www.readourlips.ca/

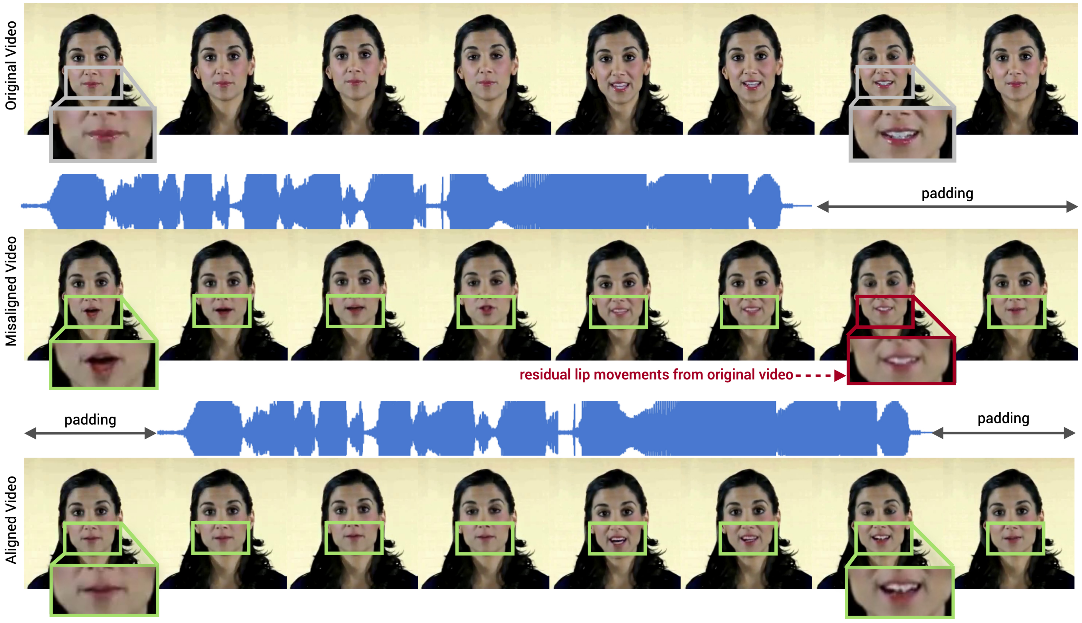

Figure 2.4: Audio-Video Alignment Module: Lip-sync models such as Wav2Lip modify the lip movements of an 'Original Video' (driving video) according to a given speech utterance. However, naively aligning the audio and video before passing through Wav2Lip can result in a 'Misaligned Video' with residual lip movements as indicated in red-boxes. We design an audio-video alignment module that detects the mouth movements in the original video. We then align the speech utterance on the region with the mouth movements and add silence around the aligned utterance. Wav2Lip then generates an 'Aligned Video' without any residual lip movements as indicated in green boxes.

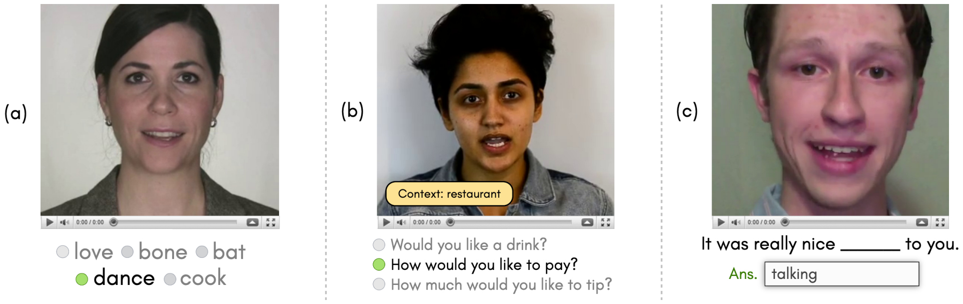
(a)
○ love ○ bone ○ bat
● dance ○ cook
(b)
Context: restaurant
○ Would you like a drink?
● How would you like to pay?
○ How much would you like to tip?
(c)
It was really nice ______ to you.
Ans. talking

Figure 2.5: Examples of different protocols used for our user study. (a) lipreading isolated words (WL): the speaker mouths a single word, and the user is expected to select one of the multiple choices presented. (b) lipreading sentences with context (SL): the speaker mouths an entire sentence. The user is presented with the context of the sentence and is expected to select one of the sentences in multiple choices, and (c) lipreading missing words in a sentence (MWIS): the speaker mouths an entire sentence. The user is presented with a sentence with blanks (masked words); the user needs to identify the masked word from the video and sentence context and answer in text format.

with context (SL), and (3) lipreading missing words in sentences (MWIS). These protocols rely on a lipreader's vocabulary and the role that semantic context plays in a person's ability to lipread.

2.3.1 Lipreading on isolated Words (WL)

The ability to disambiguate different words through visual lip movements helps shape auditory perception and speech production. In **w**ord-**l**evel (WL) lipreading, the user is presented with a video of an isolated word being spoken by a talking head, along with multiple choices and one correct answer. When a video is played on the screen, the user must respond by selecting a single response from the provided multiple choices. Visually similar words (homophenes) are placed as options in the multiple choices to increase the difficulty of the task. The difficulty can be further increased by testing for difficult words - difficulty associated with the word to lipread, e.g., uncommon words are harder to lipread. For the purpose of our study, we test the users only on the commonly known words. The multiple answer choices have been fixed to 5 options. An example of word-level lipreading is shown in Fig. 2.5 (a).

2.3.2 Lipreading Sentences with Context (SL)

In **s**entence-**l**evel (SL) lipreading, the users are presented with (1) videos of talking heads speaking entire sentences and (2) the context of the sentences. The context acts as an additional cue to the mouthing of sentences and is meant to simulate practical conversations in a given context. According to [4], the context of the sentences can improve a person's lipreading skills. Context narrows the vocabulary and helps in the disambiguation of different words. We evaluate our users in two contexts - A) Introduction - 'how are you?', 'what is your name?', and B) Lipreading in a restaurant - 'what would you like to order?'. Like WL lipreading, we provide the user with a fixed number of multiple choices and one correct answer. Apart from context, no other information is provided to the participants regarding the length or semantics of the sentence. Fig. 2.5 (b) shows an example of sentence-level lipreading with context.

2.3.3 Lipreading missing words in sentences (MWIS)

According to[9], an expert lipreader can discern only 40% of a given sentence or $4 - 5$ words in a 12 words long sentence. In this protocol, we try to emulate such an experience by **m**asking **w**ords **i**n the **s**entence (MWIS). The participants watch videos of sentences spoken by a talking head with a word in the sentence masked, as shown in Fig. 2.5 (c). Unlike SL mentioned in Sec. 2.3.2, the users are not provided with any additional sentence context. Lip movements are an ambiguous source of information due to the presence of homophenes. This exercise thus aims to use the context of the sentence to disambiguate between multiple possibilities and guess the correct answer. For instance, given the masked sentence "a cat sits on the $\{$masked$\}$," a lipreader can disambiguate between homophenes 'mat', 'bat', and 'pat' using the sentence context to select 'mat'. The user must enter the input in text format for the masked word as shown in Fig. 2.5 (c). Minor spelling mistakes are accepted.

2.4 User Study

In this section, we explain the collective background of our participants, the types of videos used for the study, and the design of our testing platform.

2.4.1 Participants

We perform our study on 50 participants with varying degrees of hearing loss with 32 male and 18 female participants. The average age of the participants in this study is 35 years, ranging from 29 years to 50 years. Participants in this study reside in the Indian states of Maharashtra and Rajasthan. 29 participants have a Master's degree, while the remaining 21 have a Bachelor's degree. All the participants in the study report having sensorineural hearing loss[10] and use hearing aids in their daily life along with lipreading and oral deaf speech as their primary mode of communication.

2.4.2 Dataset

We scrape real videos from lipreading.org and generate our synthetic videos on them. Lipreading.org videos allow us to (i) make a direct comparison between the real lipreading training videos and our synthetically generated videos and (ii) provides the correct answer to the video; this provides the correct ground truth label for the real videos later used for quantitative analysis.

[9] Speech Reading, Hearing Loss in Children | CDC

[10] What is Sensorineural Hearing Loss?

	Real	Synthetic	
Task	American	American	Indian
WL	80	800	800
SL	60	600	600
MWIS	70	700	700
Total	210	2100	2100

Table 2.1: No. of examples curated for each protocol in different English accents (American / Indian).

Primarily, we aim to compare a user's performance on the synthetic videos generated using our proposed pipeline against the real videos on lipreading.org. We use the three protocols explained in Sec. 2.3 for this purpose. Our synthetic videos are divided into: (1) non-native **A**merican-accented **E**nglish (AE) videos and (2) native **I**ndian-accented **E**nglish (IE) videos. Our users are of Indian origin.

Our synthetic dataset is created using 10 driving videos on 5 speakers. We scrape 80 labels from lipreading.org's single-word lipreading quiz for WL lipreading protocol. Using these, we generate $80 \times 10 = 800$ talking head videos - 10 variations per word. For SL lipreading, we scrape 60 questions from lipreading.org's sentence-level quiz across two contexts: introductions and lipreading in a restaurant. We generate $60 \times 10 = 600$ talking head videos - 10 variations for each sentence using these sentences. Lastly, we scrape 70 sentences from lipreading.org's missing words in sentences task and generate $70 \times 10 = 700$ talking head videos for the MWIS protocol. We generate these videos once using American-accented TTS and the second time using Indian-accented TTS. As shown in Table 2.5, we generate a total of 4200 synthetic videos and collect 210 real videos from lipreading.org across the protocols.

2.4.3 Test Design

Our primary goal is to validate that the synthetic talking head videos generated using our pipeline can replace real videos in terms of visual quality and ease of discernment.

Each participant participates in all 3 protocols. For each protocol, the user takes 3 quizzes corresponding to three datasets: (1) Real AE, (2) **Synth**etic **AE** (Synth AE), and (3) **Synth**etic **IE** (Synth IE). In total, a user attempts 9 quizzes. Quizzes are delivered through a web-based platform that we developed. Our users report taking the quizzes from a plethora of personal devices like PCs, laptops, Android and iPhone mobile devices and tablets. The number of days taken to complete a test is left at the user's discretion to prevent the user from feeling fatigued, as lipreading is an involved process and can be mentally taxing. The longest time taken by any user to complete our test is four days.

The user is presented with 20 questions/videos for each quiz. A word/sentence is first randomly sampled from the database for each question. One of the 10 variations of the sampled word/sentence present in the database is then randomly chosen. The audio is removed from the videos before displaying to the users. We ensure that words/sentences are not repeated across the quizzes in a single protocol to prevent bias by familiarization. We also ensure that the difficulty of lipreading across all the datasets

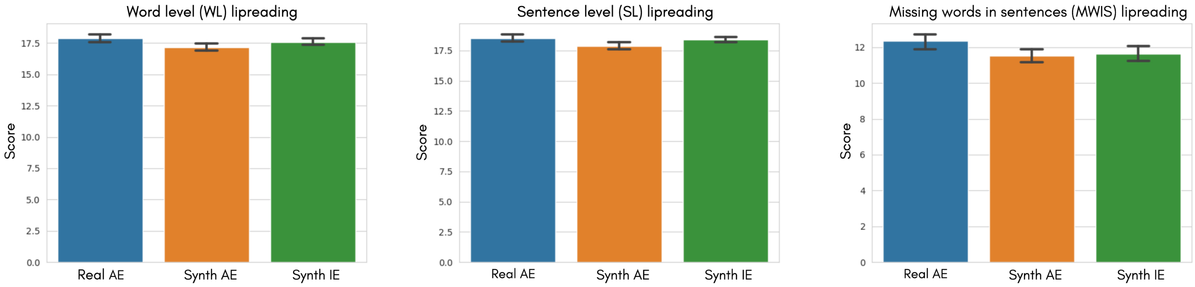

Figure 2.6: Mean user performance on the three lipreading protocols. Error bars are the standard errors of the mean.

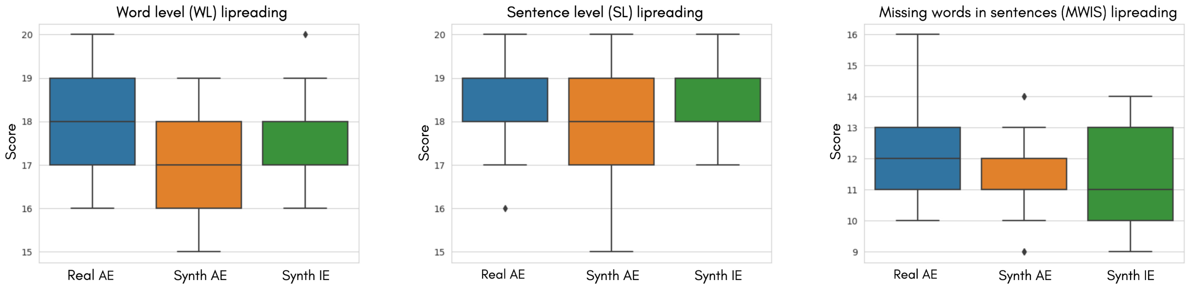

Figure 2.7: Box plots depicting the distribution of scores on the three lipreading protocols. Horizontal lines within the rectangles represent median scores. The top and bottom of the rectangles correspond to the first and third quartiles; the horizontal lines at the ends of the vertical "whiskers" represent the minimum and maximum scores, and the diamonds represent scores outside this range.

and protocols is kept consistent. The user is rewarded 1 point for each correct attempt, and the score is computed out of 20. We expect the user to finish a single test in one sitting. For a fair comparison, we do not inform the user if they are being tested on real or synthetic data.

2.5 Results and Discussion

In this section, we conduct statistical analysis to verify **(T1)** If the lipreading performance of the users remains comparable across the real and synthetic videos generated using our pipeline. Through this, we will validate the viability of our proposed pipeline as an alternative to the existing online lipreading training platforms. **(T2)** If the users are more comfortable lipreading in their native accent/language than in a foreign accent/language. This would validate the need for bootstrapping lipreading training platforms in multiple languages/accents across the globe.

Fig. 2.6 plots the standard errors of the mean. Fig. 2.7 presents the boxplot across the three lipreading protocols.

Synthetic videos as a replacement for real videos: To validate **(T1)**, the difference in the user scores across the real and synthetic videos should be statistically insignificant. Since our conclusion depends on the evidence for a null hypothesis (no difference between the categories), just the absence of evidence is not enough to support the hypothesis. Therefore, we perform a Bayesian Equivalence Analysis using the Bayesian Estimation Supersedes the t-test (BEST) [31] to quantify the evidence in favor of our model. BEST estimates the difference in means between two distributions/groups and yields a probability distribution over the difference. Using this method, we compute (1) the mean credible value as the best guess of the actual difference between the two distributions and (2) the 95% Highest Density Interval (HDI) as the range where the actual difference is with 95% credibility. For the difference in the two distributions to be statistically significant, the difference in their mean scores should lie outside the 95% HDI.

We report the BEST statistics on Real AE and Synth AE studies for all three lipreading protocols in Table 2.2. We also report the t-statistic and p-value using the standard two-tailed t-test. From Table. 2.2, it is clear that the BEST statistic lies within the acceptable 95% HDI for all three protocols indicating that the difference in the scores between the two groups is statistically insignificant. This suggests that our pipeline is a viable alternative to the existing manually curated talking-head videos.

Native vs Non-native accented lipreading: To validate **(T2)**, the difference in the user scores between native and non-native accented English should be statistically significant. Since our participant pool is from India, we compare the user scores on Synth IE and Synth AE. We perform a two-sample Z test to validate the statistical significance since our sample size is large (> 30). To this end, we propose Null Hypothesis **H0**: the difference in the mean scores between Synth IE and Synth AE is statistically insignificant, and consequently, the Alternate Hypothesis **H1**: the difference in the mean scores between the Synth IE and Synth AE is statistically significant. We compute the z statistics and report the p-value for the 90% confidence interval (significance value α=0.1) in Table 2.3 for the three protocols. We observe that the Z test statistic lies outside the 90% critical value accepted range for two tasks, WL and SL, indicating that the difference in their mean values is statistically significant in favor of IE, and we reject **H0** in favor of **H1** for these protocols. For MWIS protocol, the p-value is > 0.1, and the z statistic falls within the acceptable 90% confidence interval, indicating that the difference in their mean scores is not statistically significant. Thus, we fail to reject **H0** in this case. The overall results support our claim that lipreading on native accents makes much difference in the performance of a lipreader, and they are more comfortable in lipreading native accents. Moreover, it reinforces the importance of our platform.

Developing a lipreading training database for each new accent using real videos is a non-trivial, exhausting, and time-consuming task. Our platform could thus be quickly adopted to add any new language/accent as long as a TTS model for that language/accent is available.

Discussion: We note that the lipreaders score relatively higher for the SL protocol. The context of the sentence narrows the vocabulary space and helps disambiguate homophenes. MWIS is the most challenging protocol as it involves the user's needing to retrieve the correct word from their memory

	95% HDI	**Mean**	**MGD**	**t-value**	**p-value**
WL	(-0.254, 1.63)	0.701	0.706	1.676	0.103
SL	(-0.226, 1.62)	0.671	0.647	1.540	0.133
MWIS	(-0.366, 1.98)	0.793,	0.824	1.517	0.139

Table 2.2: We perform BEST statistical analysis and compute the 95% HDI range of the difference in means of the real and synthetic distributions. Mean is the distribution of means. We also report the p-values and t-values from a standard t-test for comparison.

	p-value	**accepted range**	**z statistic**
WL	0.0786	(-1.645 : 1.645)	1.758
SL	0.0171	(-1.645 : 1.645)	2.384
MWIS	0.705	(-1.645 : 1.645)	0.378

Table 2.3: Two-sample z-test on synthetic Indian-accented English (IE) and American-accented English videos (AE). The significance level α is kept at 0.1. The null-hypothesis is rejected if the z statistic falls outside the 90% critical value accepted range. Consequently, the p-value is also less than the significance value α in that case.

instead of classifying from the given choices. It also involves mapping the masked word from sentences to its corresponding mouthing in the videos. Thus, the users score relatively low on MWIS.

As a conclusion of the user study, we present evidence that synthetic videos can potentially replace real videos. We show that the drop in user performance across Real AE and Synth AE is statistically insignificant across all the protocols. We also show that users are more comfortable lipreading in a native accent through paired z-test, highlighting the dire need to bootstrap lipreading platforms in multiple languages/accents at scale.

2.5.1 Statistical Analysis

We present additional graphs of the user study conducted on the three lipreading protocols - (1) lipreading isolated words (WL), (2) lipreading sentences with context (SL), and (3) lipreading missing words in sentence (MWIS). Fig. 2.8 represents the mean user performance on the three protocols against standard deviation, and Fig. 2.9 represents the mean user performance against the 95% confidence interval (Eqn. 2.2) of the mean. Standard error indicated in the main paper is computed using Eqn. 2.1. The blue bars indicate scores on real videos with American-accented English, the orange bars indicate synthetic American-accented English, and the green bars indicate synthetic Indian-accented English. The synthetic data is generated using our pipeline.

$$\text{standard error} = \frac{\sigma}{\sqrt{n}} \tag{2.1}$$

$$\text{ci} = \hat{x} \pm z\left(\frac{s}{\sqrt{n}}\right) \tag{2.2}$$

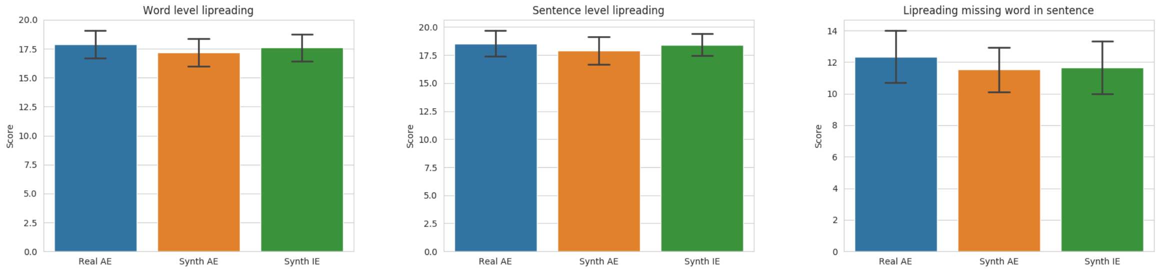

Figure 2.8: Mean user performance of the three lipreading protocols. The error bars are the standard deviation of the distribution.

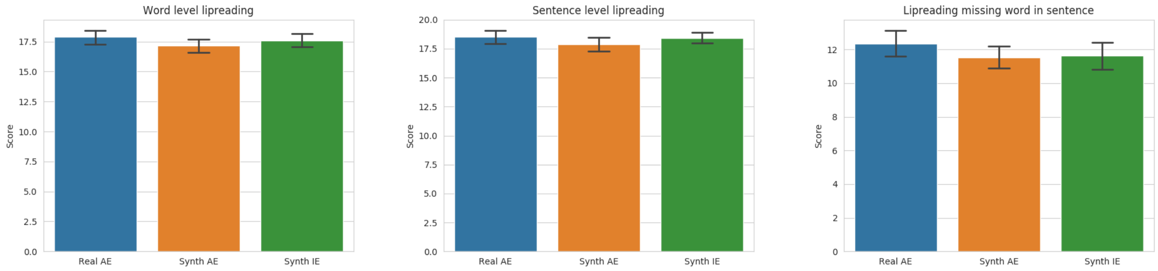

Figure 2.9: Mean user performance on the three lipreading protocols. The error bars are the 95% confidence interval of the mean.

As mentioned in the main paper, we perform the Bayesian Estimation Supersedes the t-test (BEST) [31] for comparing the lipreading scores of the users across the real and synthetically generated videos. BEST estimates the difference in means between the two groups and yields a probability distribution over the difference. From the distributions, the mean credible value as the best guess of the actual difference and the 95% Highest Density Interval (HDI) as the range where the actual difference is with 95% credibility are computed. We validate if the ideal difference between the two groups lies in the 95% HDI. If it does, the difference between the two groups is not statistically significant; otherwise, the difference is statistically significant. In Fig. 2.10, we show the graph of the distribution of the difference of means between the real and synthetically generated American-accented English (AE) videos for the three lipreading protocols. Please note that the ideal mean difference for all three lipreading tasks lies in the 95% HDI, indicating that the difference in the lipreading scores across the synthetic and real datasets are statistically insignificant. The graphs denote the distribution of the difference of means for the three protocols - WL, SL, and MWIS.

To validate if lipreading native-accented videos affect lipreading performance, we conduct a statistical analysis of the user's performance on the synthetically generated Indian-accented English (IE) and

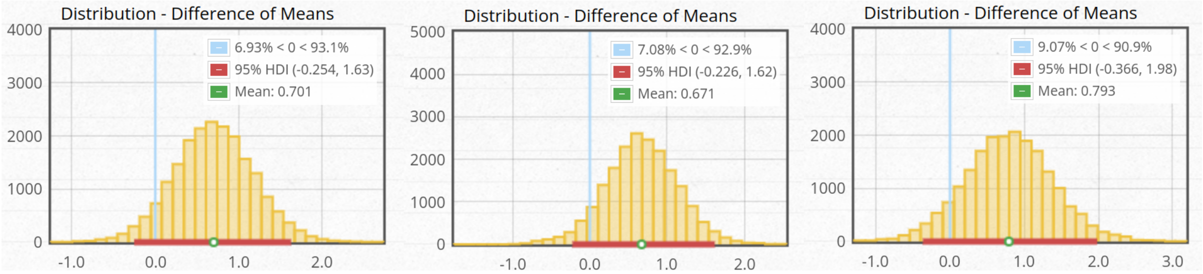

Figure 2.10: Distribution of difference of means performed using the Bayesian Analysis on real and synthetically generated American-accented English (AE). The graphs for the protocols are displayed in the following order: (1) lipreading words (WL), (2) lipreading sentences (SL), (3) lipreading missing words in sentence (MWIS). The 95% HDI interval is represented using the horizontal red line.

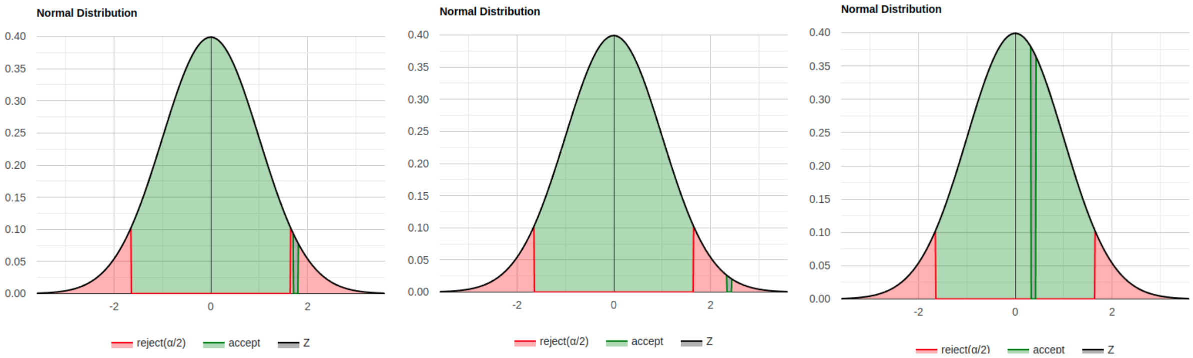

Figure 2.11: z statistic for the 90% confidence interval computed using the two-sample z-test for synthetically generated Indian-accented English (IE) and American-accented English (AE) videos. The critical z-value corresponding to the 90% confidence interval is ±1.96.

American-accented English (AE). Since the participants of our user study are from India, our expectation is that their lipreading scores on IE should be better than their scores on the test with AE, even though the users are comfortable with both accents. We conduct the two-sample z-test as our sample size is large (>30) for comparing the scores of the users across the synthetically generated IE and AE and plot the graph of the z-statistic for the 90% confidence interval. The 90% confidence interval is the acceptable region from -1.96 to $+1.96$ and is represented by the green region, and the region lying outside in red is the rejection region. The graph of z-statistic for the three lipreading protocols is shown in Fig. 2.11. The z-statistic is given by the formula:

$$z = \frac{(\bar{x_1} - \bar{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{2.3}$$

where $\bar{x_1}$ and $\bar{x_2}$ denote the sample average of Indian and American accents. s_1 and s_2 denote the standard deviation of the two groups, and n_1 and n_2 represent the sample size of the two groups. The population's average is represented using μ_1 and μ_2.

For the difference in the means of the two groups to be statistically insignificant, the expected difference ($\mu_1 - \mu_2$) between the population's average should be 0. Consequently, our null hypothesis is **H0**: the difference in the mean scores between Synth IE and Synth AE is not statistically significant, i.e., $\mu_1 - \mu_2 = 0$. Our alternate hypothesis is **H1**: the difference in the mean scores between Synth IE and Synth AE is statistically significant. If the z-statistic lies outside the acceptable range, H0 is rejected in favor of H1, indicating that the difference in the mean scores of the two groups is statistically significant. From the graphs 2.11, we observe that the z-statistic lies outside the acceptable range for two tasks - (1) lipreading words (WL) and (2) lipreading sentences (SL). Consequently, the p-value is also lower than the significance value (α = 0.1).

2.5.2 Quantitative and Qualitative Results

Even though all the individual modules in our pipeline can be replaced with other equivalent modules, we provide quantitative and qualitative metrics of the modules used in our pipeline and other similar modules. We evaluate the choice of the TTS model by conducting a user study comprising 30 participants. We compare three recent SOTA TTS works - (1) FastSpeech2 [8], (2) Tacotron2 [49], and (3) Glow-TTS [29] by performing Mean Opinion Score (MOS) [11] evaluation to evaluate the perceptual quality of the models. The MOS evaluation scores for the 95% confidence interval are provided in Table 2.4.

We provide quantitative scores for comparing the real and synthetically generated videos using the lipsync model. For quantitative comparison, we report the LSE-D [43] scores for comparing the lipsync performance of real videos against videos generated synthetically using Wav2Lip [43].

Method	MOS
Tacotron2	3.85 ± 0.08
Glow-TTS	3.96 ± 0.06
FastSpeech2	3.98 ± 0.04
GT	4.53 ± 0.07

Table 2.4: Mean Opinion Scores (MOS) evaluations of different TTS models with 95% confidence interval.

Method	LSE-D
Wav2Lip	6.902
GT	6.718

Table 2.5: LSE-D metric computed for the synthetically generated videos using Wav2Lip model against real videos.

2.6 Conclusion and Future Work

Lipreading is a widely adopted mode of communication for people with hearing loss. However, online resources for lipreading training are scarce and limited in many factors, such as vocabulary, speakers, and languages. Moreover, launching a new platform in a new language is costly, requiring months of manual effort to record training videos on hired actors. In this work, we analyze the viability of using synthetically generated videos to replace real videos for lipreading training. We propose an end-to-end automated and cost-effective pipeline for generating lipreading videos and carefully design a set of protocols to evaluate the generated videos. We perform statistical analysis to validate that the difference in user performance on real and synthetic lipreading videos is statistically insignificant. We also show the advantage of lipreading in native accents, thus highlighting the dire need for lipreading training in many languages and accents. In this vein, we envision a MOOCs platform for training humans in lipreading to potentially impact millions of people with hearing loss across the globe.

One major learning from the user study was that the expressions of the talking heads played a pivotal role in deciphering the spoken content for the hard-of-hearing people. Consequently, it would be extremely beneficial to generate synthetic talking-heads with control over the expressions. A desirable quality would be to detect and transfer the expressions automatically from the intonation of the speech segment or from the sentiment of the spoken text. Towards this direction, we envision and explore preserving the expressions of synthetic talking-heads for an important albeit different scenario in the movie-making industry in the next chapter.

Chapter 3

FaceOff: A Video-to-Video Face Swapping System

In this chapter, we explore an alternate usecase of synthetic talking heads in the moviemaking industry. Moviemakers are no strangers to the important role of body doubles in the moviemaking industry. They take the place of actors in dangerous stunt scenes or scenes where the same actor plays multiple characters. The double's face is later replaced with the actor's face and expressions manually using expensive CGI technology, costing millions of dollars and taking months to complete. An automated, inexpensive, fast way can be to use face-swapping techniques that aim to swap an identity from a source face video (or an image) to a target face video. However, such methods cannot preserve the source expressions of the actor important for the scene's context. We tackle this challenge, by introducing a new line of research called video-to-video (V2V) face-swapping, that can preserve (1) the identity and expressions of the source (actor) face video and (2) the background and pose of the target (double) video. We highlight the key drawbacks of existing face swapping approaches, and how they are undersuited for the task of V2V face-swapping. We also curate and benchmark V2VFaceSwap, a V2V face-swapping test dataset made of unconstrained YouTube videos on unseen identities, backgrounds, and lighting conditions. We also compare our approach against existing approaches on several benchmark metrics and inference times.

3.1 Introduction

Having doubles[1] for the starring actors in movies is an indispensable component of movie-making. A double may take the actor's place during stunt scenes involving difficult and dangerous life-risking acts. They may even stand-in for the actor during regular fill scenes or multiple retakes. For instance, 'The Social Network' extensively used body doubles as a stand-in for actor Armie Hammer who played multiple roles of twin brothers[2][3][4].

[1] https://en.wikipedia.org/wiki/Double_(filmmaking)

[2] Captain America - Skinny Steve Rogers Behind the Scenes

[3] How CGI made Cody and Caleb as PAUL WALKER — VFX

[4] Armie Hammer Didn't Play Both Winklevoss Twins Social Network

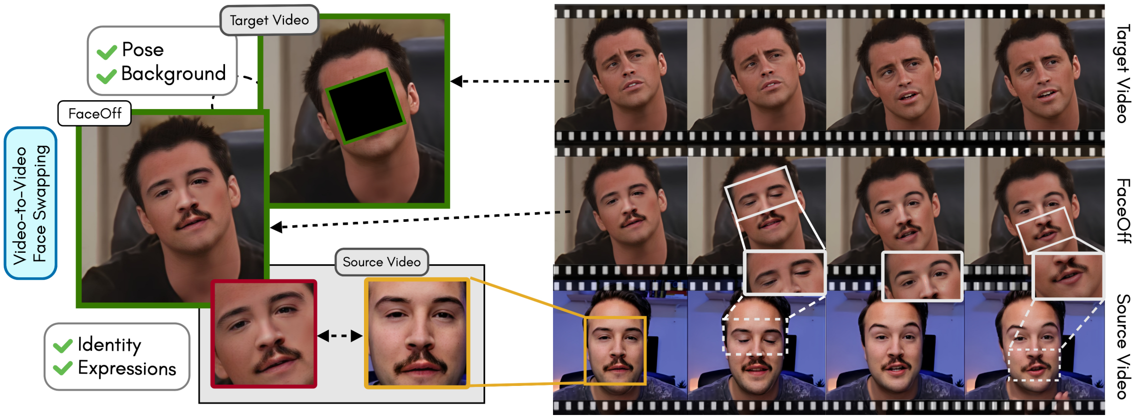

Figure 3.1: We introduce video-to-video (V2V) face-swapping, a novel task of face-swapping that aims to swap the **identity** and **expressions** from a source face video to a target face video. This differs from the face-swapping task that aims to swap only an identity. There are many downstream applications of V2V face-swapping, such as automating the process of an actor replacing their double in movie scenes, which today is handled manually using expensive CGI technology. In this example, Nolan, an actor (source video), is recording his dialogues and expressions at the convenience of his home. Joey Tribiani (target video) is acting as his double in a scene of the famous sitcom FRIENDS. FaceOff face-swaps Nolan into the scene. Please note the zoomed-in source (yellow box) and face-swapped (red box) output. In this output, although the source face pose and skin complexion have changed and blended with the background, identity and expressions are preserved.

3.1.1 How the movie-making industry handles such tasks?

In such scenes, the double's face is later replaced by the actor's face and expressions using CGI technology requiring hundreds of hours of manual multimedia edits on heavy graphical units costing millions of dollars and taking months to complete. Thus, the production team is generally forced to avoid such scenes by changing the mechanics of the scene such that only the double's body is captured to provide an illusion of the actor. This may act as a constraint on the director's creativity. However, such adjustments are not always possible.

A different scenario is post-production scene modifications. If a dialogue is discovered in post-production that suits a scene better than the original, the entire scene is reset and re-shot. We propose that the actor could instead record in a studio and get their face superimposed on the previous recording. In fact, like other industries, the movie industry is also headed in the direction where actors can work from home. In today's era, CGI technologies can produce incredible human structures, scenes, and realistic graphics. However, it is known that they struggle to create realistic-looking skin[5]. As shown in Fig. 3.1, an actor could lend their identity and expressions from the comfort of their home or studio while leaving the heavy-duty to graphics or a double. CGI technologies needed for such tasks are manually operated, expensive, and time-consuming.

[5] Why It's SO HARD To Do CGI Skin!

3.1.2 Why are existing approaches inept for this task?

To automate such tasks, fast and inexpensive computer vision based face-swapping [41, 51, 39, 38, 33, 5] techniques that aim to swap an identity between a source (actor) video and target (double) video can be considered. However, such techniques cannot be directly used. Face-swapping swaps only the source identity while retaining the rest of the target video characteristics. In this case, the actor's expressions (source) are not captured in the output.

3.1.3 What is the video-to-video (V2V) face-swapping task?

”Video-to-video (V2V) face-swapping” is a novel task of face-swapping that aims to **(1)** swap the identity and expressions of a source face video and **(2)** retain the pose and background of the target face video. The target pose is essential as it depends on the scene's context. E.g., a stunt man performs at an outdoor location dealing with machines or talking to a fellow double; the actor acts in front of a green screen at a studio. Here, the double's pose is context-aware, and the actor only improvises. Unlike the face-swapping task that swaps a fixed identity component from one video to another video, V2V face-swapping swaps expressions changing over time (a video) with another video with changing pose and background (another video), making our task video-to-video.

3.1.4 Overview of this work

Swapping faces across videos is non-trivial as it involves merging two different motions - the actor's face motion (such as eye, cheek, or lip movements) and the double's head motion (such as pose and jaw motion). This needs a network that can take two different motions as input and produce a third coherent motion. We propose **FaceOff**, a video-to-video face swapping system that reduces the face videos to a quantized latent space and blends them in the reduced space. A fundamental challenge in training such a network is the absence of ground truth. Face-swapping approaches [51, 39, 41] use a discriminator-generator setup for training the networks. The discriminator is responsible for monitoring the desired characteristic of the swapped output. However, using a discriminator leads to hallucinating components of the output different from the input - for instance, modified identity or novel expressions. Thus, we devise a self-supervised training strategy for training our network: We use a single video as the source and target. We then introduce pseudo motion errors on the source video. Finally, we train a network to 'fix' these pseudo errors to regenerate the source video. FaceOff can face-swap unseen cross-identities directly at inference without any finetuning. Moreover, unlike most face-swapping methods that need inference time optimization ranging from 5 minutes to 24 hours on high-end GPUs, FaceOff face-swaps videos in just one forward pass, taking less than a second. A key feature of FaceOff is that it preserves at least one of the input expressions (source in our case), whereas, as we show later, existing methods fail to preserve either of the expressions (source or target expressions). Lastly, we curate and benchmark V2VFaceSwap, a V2V face-swapping test dataset made of instances from unconstrained YouTube videos on unseen identities, backgrounds, and lighting conditions.

To summarize, the major contributions of our work are:

- We introduce V2V face-swapping, a novel task of face-swapping that aims to swap source face identity and expressions while retaining the target background and pose.
- We propose FaceOff: a V2V face-swapping system trained in a self-supervised manner. This is done by using the same video as source and target, introducing pseudo motion-errors in the source video, and ”fixing” them to regenerate the original video using a reconstruction loss objective. This loss objective helps preserve the identity and expressions of the source actor.
- We show that the task of merging two different motions is non-trivial, and develop temporal autoencoding models (adapted from VQVAE-2), that take as input two different motions, and produce a third coherent output motion.
- Our approach works on unseen identities directly at the time of inference without any additional finetuning or inference time optimization, and takes less than a second to infer.
- We release the V2VFaceSwap test dataset and establish a benchmark for the V2V face-swapping task.

3.2 Delving into the Existing Face Swapping, Face Manipulation, Face Reenactment, and Face Editing models

Table 3.1 provides a comparison between the existing tasks and FaceOff. FaceOff aims to solve a unique challenge of V2V face-swapping that has not been tackled before.

3.2.1 Face Swapping

Swapping faces across images and videos have been well-studied [41, 39, 51, 6, 30, 33, 38, 5, 7] over the years. These works aim to swap an identity obtained from a source video (or an image) with a target video of a different identity such that all the other target characteristics are preserved in the swapped output. DeepFakes[6], DeepFaceLabs [41], and FSGAN [39] swap the entire identity of the source; Motion-coseg [51] specifically swaps the identity of single/multiple segments of a given source image (either hair or lips or nose, etc.) to a target video. Unlike these approaches that swap only the identity or a specific part of an image, we swap temporally changing expressions along with the identity of the source. Moreover, FSGAN takes 5 minutes of inference time optimization, DeepFaceLabs and DeepFakes take up to 24 hours of inference time optimization on high-end GPUs. FaceOff takes less than a second to face swap in-the-wild videos of unseen identities.

[6] https://github.com/deepfakes/faceswap

	Source		Target	
Method	**Identity**	**Expression**	**Pose**	**Background**
Face Swapping	✓	×	✓	✓
Face Reenactment	×	✓	×	✓
Face Editing	×	×	✓	✓
FaceOff (Ours)	✓	✓	✓	✓

Table 3.1: Comparison of FaceOff with existing tasks. ✓and × indicate if the characteristic is preserved and lost respectively. FaceOff solves a unique task of preserving source identity and expressions that has not been tackled before.

3.2.2 Face Manipulation

Face manipulation animates the pose and expressions of a target image/video according to a given prior [64, 52, 50, 65, 41, 73, 56, 75]. In audio-driven talking face generation [43, 26, 34, 74, 56, 45, 20], the expressions, pose, and lip-sync in the target video are conditioned on a given input speech audio. Unlike such works, we do not assume an audio prior for our approach.

3.2.3 Face Reenactment

A different direction of **face reenactment** animates the source face movements according to the driving video [57, 45, 58, 28, 50, 52]. The identity is not exchanged in these works. This can tackle a special case of our task – when the target and source have the same identity. Here, a target image can be re-enacted according to the source video expressions. As we show in Section 3.6.3, FaceOff captures the micro-expression of the driving video, unlike the existing approaches. This is because we rely on a blending mechanism - allowing a perfect transfer of the driving expressions.

3.2.4 Face Editing

Another direction that handles this special case is **face editing**, which involves editing the expressions of a face video. Using this approach, one can directly edit the target video according to the source expressions. Image-based face editing works such as [23, 9, 10, 37] have gained considerable attention. However, realizing these edits on a sequence of frames without modeling the temporal dynamics often results in temporally incoherent videos. Recently, STIT [62] was proposed that can coherently edit a given video to different expressions by applying careful edits in the video's latent space. Despite the success, these techniques allow limited control over expression types and variations. Moreover, obtaining a correct target expression that matches the source expressions is a manual hit and trial. FaceOff can add micro-expressions undefined in the label space simply by blending the emotion from a different video of the same identity with the desired expressions.

3.3 Learning to Swap Faces in Videos

We first describe the mechanism behind our proposed approach. We then highlight two key aspects of our approach - i) Self-supervised training, and ii) Merging videos as quantized latents.

3.3.1 FaceOff: Face Swapping in Videos

We aim to swap source face video with a target face video such that (1) the identity and the expression of the source video are preserved and (2) the pose and background of the target video are retained. To do this, we learn to blend the foreground of the source face video with the background and pose of the target face video (as shown in Fig. 3.3) such that the blended output is coherent and meaningful. This is non-trivial as it involves merging two separate motions. Please note that we only aim to blend the two motions; thus, the desired input characteristics – identity, expressions, pose, and background – are naturally retained from the inputs without additional supervision. The main challenge is to align the foreground and background videos so that the output forms a coherent identity and has a single coherent pose. All the other characteristics are reconstructed from the inputs. Our core idea is to use a temporal autoencoding model that merges these motions using a quantized latent space. Overall, our approach relies on (1) Encoding the two input motions to a quantized latent space and learning a robust blending operation in the reduced space. (2) A temporally and spatially coherent decoding. (3) In the absence of ground truth, a self-supervised training scheme.

3.3.2 Merging Videos using Quantized Latents

We pose face-swapping in videos as a blending problem: given two videos as input, blend the videos into a coherent and meaningful output. We rely on an encoder to encode the input videos to a meaningful latent space. Our overall network is a special autoencoder that can then learn to blend the reduced videos in the latent space robustly and generate a blended output. We select our encoder model carefully, focusing on "blending" rather than learning an overall data distribution. Encoder networks with a continuous latent space reduce the dimension of a given input, often down to a single vector that can be considered a part of an underlying distribution. This latent vector is highly stochastic; a very different latent is generated for each new input, introducing high variations that a decoder needs to handle. Recently, "vector quantization" was proposed in [40, 16, 44]. Quantization reduces the variation in latents by fixing the number of possible latent codes. However, retaining the input properties using a single quantized latent vector is impossible. Thus the inputs are reduced to a higher dimensional quantized space (such as 64×64) such that properties of the input needed for a full reconstruction are preserved. We adopt such an encoder in our proposed autoencoder for encoding our videos. As shown in Fig. 3.2, our encoder is a modified VQVAE2 [44] encoder that encodes videos instead of images. We introduce temporal modules made of non-linear 3D convolution operations to do so.

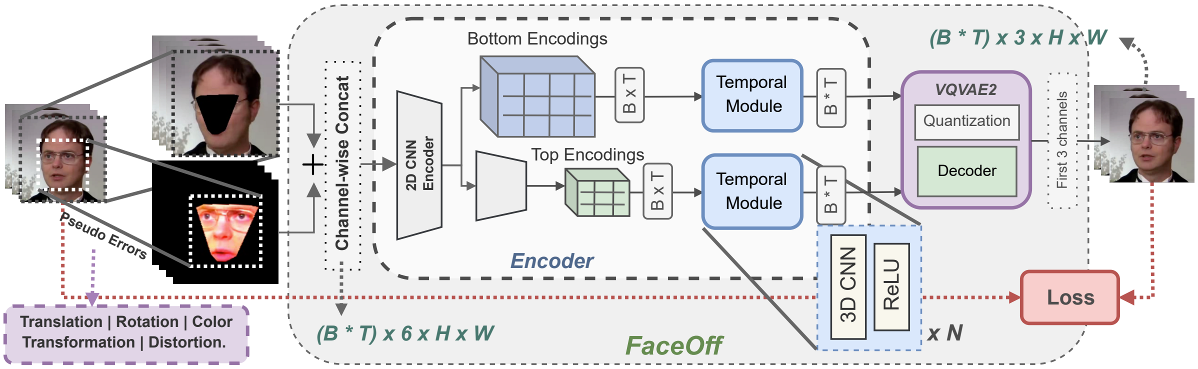

Figure 3.2: FaceOff is a temporal autoencoder operating in a hierarchical quantized latent space. We use a self-supervised training scheme to train FaceOff using a distance loss on the exact output-ground truth pairs. In the scheme, we first extract the face, f, and background, b, from a single video, s. We then apply "pseudo errors" made of random rotation, translation, scaling, colors, and non-linear distortions to modify f. Next, modified f (acting as a source) and b (acting as a target) are concatenated at each corresponding frame channel-wise to form a single video input. This video input is then reduced and blended, generating a coherent and meaningful output. This output is expected to match the source video, s.

The input to our encoder is a single video made by concatenating the source foreground and target background frames channel-wise, as shown in Fig. 3.3. Like VQVAE2, our encoder first encodes the concatenated video input framewise into 32×32 and 64×64 dimensional top and bottom hierarchies, respectively. Before the quantization step at each of these hierarchies, our temporal modules are added that process the reduced video frames. This step allows the network to backpropagate with temporal connections between the frames. The further processing is then again done framewise using a standard VQVAE2 decoder. In practice, we observed that this temporal module plays an important role in generating temporally coherent outputs, as we show through ablations in Sec. 3.7. Our special autoencoder differs from standard autoencoders in the loss computation step. Instead of reconstructing the inputs, a six-channel video input – the first three channels belonging to the source foreground and the last three channels belonging to the target pose and background – FaceOff aims to generate a three-channel blended video output. Therefore, the loss computation is between a ground truth three-channel video and the three-channel video output.

3.3.3 Self-supervised Training Approach

Existing face-swapping approaches employ generators and discriminators to train their networks. These discriminators are classifiers that indicate a relationship between the generator's outputs and underlying data distribution, such as an identity or an expression distribution. In such a setup, the generators are encouraged to hallucinate some aspects of the outputs to match the discriminator's data distribution causing it to output novel identities or expressions. We show this phenomenon in Fig. 3.4. A

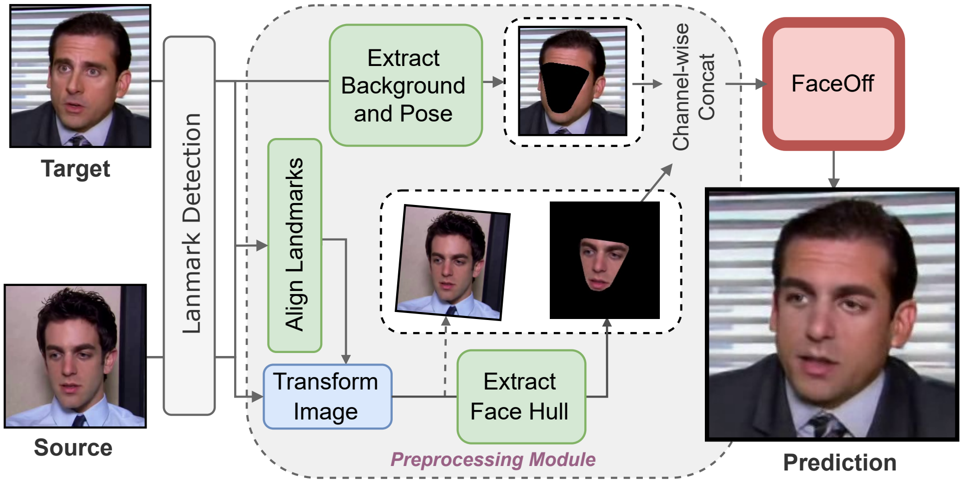

Figure 3.3: Inference pipeline: FaceOff can be directly inferred on any unseen identity without any finetuning. At inference, the source video is first aligned frame-by-frame using the target face landmarks. FaceOff then takes (1) foreground of the aligned source video and (2) background and pose of the target video as input and generates the output.

hard distance loss (e.g., Euclidean distance) indicating the exact output-ground truth relationship instead of a stochastic discriminator loss can be used to overcome this issue. In V2V face-swapping, retaining the exact source expressions is essential. Thus, we train our network using a distance loss by devising a self-supervised training scheme that forces the network to reconstruct a denoised version of a given input video.

To understand the training scheme, we first look at the challenges we encounter when trying to blend two motions naively. First, there is a global and local pose difference between the faces in the source and target videos. We fix the global pose difference by aligning (rotating, translating, and scaling) the source poses according to the target poses using face landmarks, as shown in Fig. 3.3. However, the local pose difference is not overcome this way, and we observe temporal incoherence across the frames. Next, we observe a difference in the foreground and background color (illumination, hue, saturation, and contrast). Thus, we train our network to solve these known issues by reproducing these errors during training. As illustrated in Fig. 3.2, we train our model in the following manner: (1) Take a video, say s. (2) From s, extract the face region, say f; and the background region, say b. (3) Introduce pseudo errors (rotation, color, scale, etc.) on f. (4) Construct the input v by concatenating f and b channel-wise at every corresponding frame. (5) Train the network to construct s from v. Although we train the network using the same identity in the self-supervised scheme, it can face-swap unseen identities directly at inference without any finetuning.

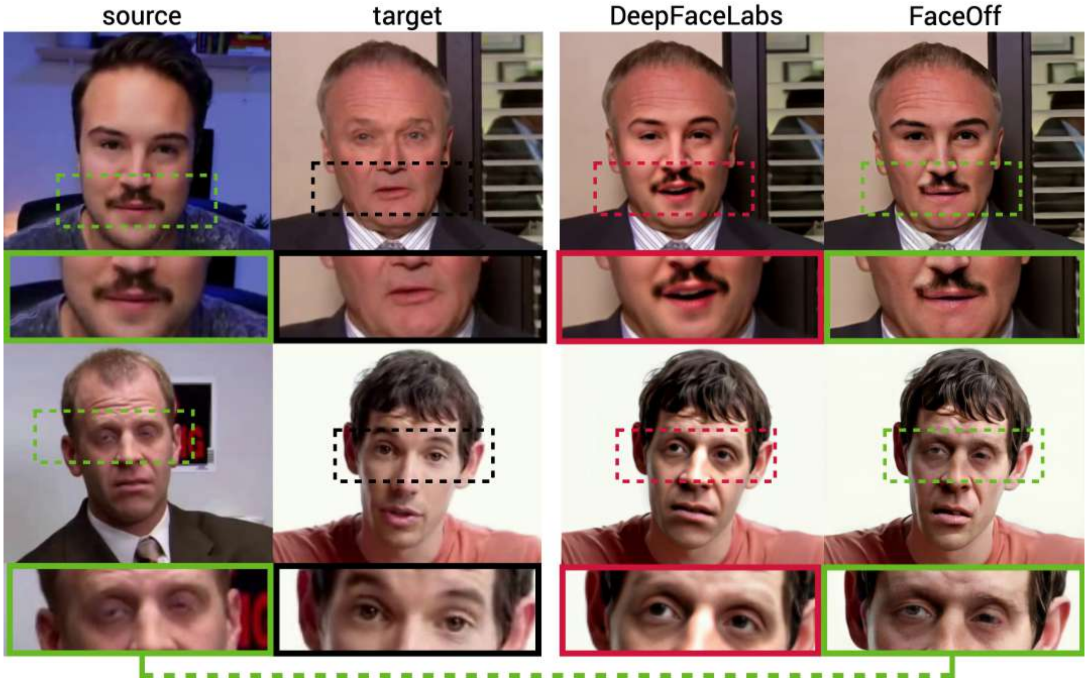

Figure 3.4: Existing face-swapping methods [41, 51, 39] use a generator-discriminator training strategy. This results in outputs with novel expressions as explained in Sec. 3.3.3. We show this phenomenon on DeepFaceLabs [41]. The expressions in the output (red boxes) do not match either of the inputs, source, or target. E.g., direction of eye gaze (second row) or overall laugh expression (first row). FaceOff successfully preserves the source expressions (green boxes).

3.3.4 Reproducing Inference Errors at Training

Given two talking-head videos, source and target, denoted by S and T, respectively, our aim is to generate an output that preserves (1) the identity and the emotions from S and (2) the pose and background from T. We assume the number of frames, denoted by N, in S and T are equal. Given two frames, $s_i \in S$ and $t_i \in T$ such that $i = 1...N$, we denote $f_{s_i} \in F_s$ and $b_{t_i} \in B_t$ as the foreground and background of s_i and t_i, respectively. Given F_s and B_t as input, the network fixes the following issues:

First, the network encounters a local pose difference between f_{s_i} and b_{t_i}. This pose difference can be fixed using an affine transformation function: $\delta(f_{s_i}, b_{t_i}) = m(rf_{s_i} + d) + m(rb_{t_i} + d)$ where m, r, and d denote scaling, rotation, and translation. Face being a non-rigid body; the affine transformation only results in the two faces with a perfect match in the pose but a mismatch in shape. One can imagine trying to fit a square in a circle. One would need a non-linear function to first transform the square to a shape similar to the circle so that they fit. We denote this non-linear transformation as a learnable function $\omega(f_{s_i}, b_{t_i})$. Being non-linear, a network can perform such transformations on the input frames as long as both faces fit. These transformations can be constrained using a distance loss to encourage spatially-consistent transformations that generate a meaningful frame. However, these spatially-consistent transformations may be temporally-incoherent across the video. This would result

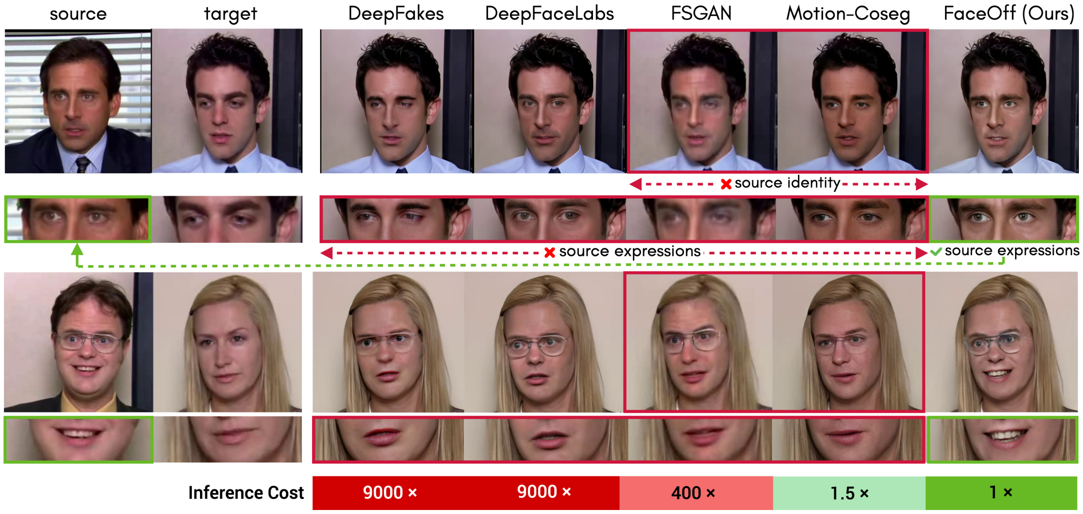

Figure 3.5: "Inference Cost" denotes the time taken for a single face-swap. FSGAN, with 400× FaceOff's inference cost, fails to swap the identities fully. DeepFakes and DeepFaceLabs swap the identities successfully but are 9000× less efficient than FaceOff. FaceOff perfectly swaps source identity and expressions. None of the other methods can swap source expressions.

in a video with a face that wobbles, as shown in Sec. 3.7. Thus, we constrain the transformations as $\omega(f_{s_i}, b_{t_i}, f_{s_k}, b_{t_k})$ where $k = 1..N$ such that $k \neq i$. Here, the transformation on the current frame is constrained by the transformations on all the other frames in the video. This is enabled by the temporal module, as explained in Sec. 3.3.2. Lastly, the network encounters a difference in color between f_{s_i} and b_{t_i} that is fixed as $c(f_{s_i}, b_{t_i})$.

As shown in Fig. 3.2, at the time of training $S = T$. For each frame $s_i \in S$, we first extract the foreground, $f_{s_i} \in F_s$ (acting as the source), and the background, $b_{t_i} \in B_t$ (acting as the target) from s_i. Next, we apply random rotation, translation, scaling, color, and distortion (Barrel, Mustache) errors on f_{s_i}. The training setting is then formulated as:

$$\Phi : \Omega(\delta, \omega, c) \tag{3.1}$$

$$J = \frac{1}{N}\sum_{i=1}^{N}[s_i - \Phi(f_{s_i}, b_{t_i}, f_{s_k}, b_{t_k})] + P(F_s, B_t) \tag{3.2}$$

where Ω is a learnable function, J is the overall cost of the network to be minimized, and P is a perceptual metric (LPIPS [72] in our case), and $k = 1 \dots N$ such that $k \neq i$.

3.4 Network Design

We adopt the architecture of VQVAE2 [44]. VQVAE2 encodes the input into multiple hierarchies: top and bottom. We adopt the same architecture but modify it in two fundamental ways. (1) VQVAE2 is

	Quantitative Evaluation					Human Evaluation		
Method	**SPIDis ↓**	**LMD ↓**	**TL-ID ↑**	**TG-ID ↑**	**FVD ↓**	**Identity ↑**	**Exps. ↑**	**Ntrl. ↑**
Motion-coseg [51]	0.48	0.59	0.872	0.893	293.652	6.82	5.81	7.44
FSGAN [39]	0.49	0.57	0.914	**0.923**	**242.691**	7.84	6.83	**8.31**
FaceOff (Ours)	**0.38**	**0.41**	**0.925**	0.915	255.980	**9.64**	**9.86**	8.18

Table 3.2: Quantitative metrics on V2VFaceSwap dataset. DeepFakes and DeepFaceLabs take up to 24 hours for best inference on a single face-swap [41]; thus, we do not compare with them. The metrics used for comparisons are explained in Sec. 3.6. For fair comparisons, FSGAN scores are reported without any inference time optimization. Although FSGAN has a slightly better FVD and Naturalness (Ntrl.) score, it fails to swap the identity fully, as can be clearly seen from SPIDis, LMD, and Identity metric. Moreover, the difference in the FVD of FSGAN and FaceOff is not statistically significant perceptually [63].

an autoencoding network and thus computes the distance between the input and the output of dimension $H \times W \times C$ – the height, the width of the image, and the number of input channels, respectively. In our case, input is a channel-wise concatenation of the source foreground, f_{s_i}, and target background, b_{t_i}, giving a dimension of $H \times W \times 6$, and thus, the output generated by our network is of the same dimension $H \times W \times 6$. During training, instead of the input, we compute the loss against the ground truth video, s_i, of dimension $H \times W \times 3$. Thus, we only consider the first three channels of $H \times W \times 6$ output at the network's output. Similarly, we only consider the first three channels as our output at the inference. (2) VQVAE2 operates at a frame level and thus cannot model temporal properties. Thus, we add temporal modules in the network just before the quantization block. At each hierarchy, the encoder produces a latent of dimension $(B * T) \times C \times H \times W$. Here, we expand the batch dimension to convert the flattened input into videos. These video latents of dimension $B \times T \times C \times H \times W$ are then passed through the temporal block made of 3D convolution and ReLU layers (see Fig. 3.2). Post this step, we again convert the batch dimension to $(B * T)$. The losses are then applied frame-by-frame. The temporal layers learn to identify the properties across the video and produce a blended encoding even with a frame-by-frame loss. At this point, the encoder outputs are quantized, and we adopt the decoder architecture of VQVAE2 for decoding the latent.

3.5 Experimental Setup

3.5.1 Hardware Setup

All of our models are trained and inferred on NVIDIA GTX 3080 Ti using 4 GPUs and 1 GPU respectively.

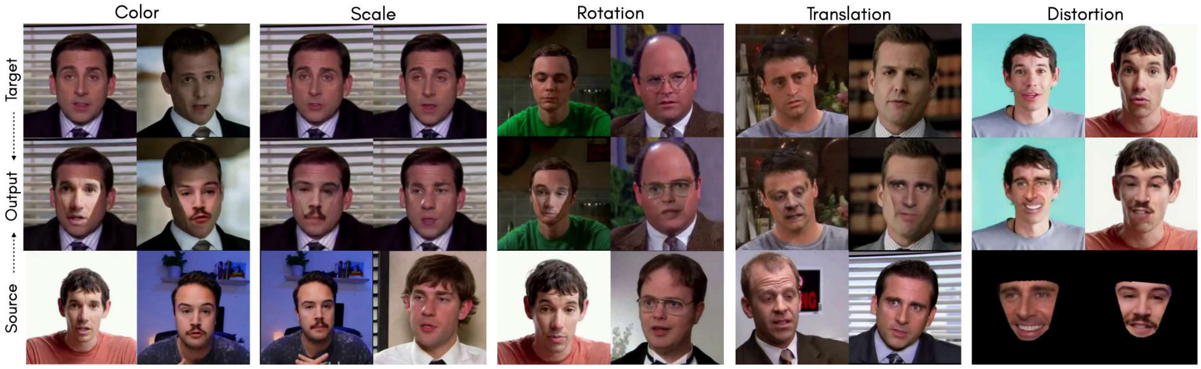

Figure 3.6: Ablation Experiments. In each of the experiment, we remove the type of error mentioned at the time of self-supervised training. Here, we present the results of the trained models at the inference on cross-identity.

	Name	Nationality	YouTube Channel
1.	Anfisa Nava	Russia	ANFISAofficial
2.	Sejal Kumar	India	sejalkumar7theclothingedit
3.	Johnny Harris	USA	johnnyharris
4.	BestDressed	USA	bestdressed
5.	Jack Edwards	UK	thejackexperience

Table 3.3: Speakers in the training dataset collected from publicly available YouTube VLOG videos.

3.5.2 Dataset

To create the training dataset, we curate publicly available unconstrained YouTube VLOG videos. It includes five different YouTubers; the specifications of the same are provided in Table 3.3. The data amounts to a total of 15 hours of video divided equally among all speakers. All the speakers speak in English, although they have different accents based on their nationality. The details of the videos, along with the timestamp, will be released publicly to promote future research.

The test set is also curated from unconstrained YouTube videos. The videos have a different identity, background, and light setting from the training set. Furthermore, they are selected from a widely varying timeline ranging from the 1990s to the late 2021s! This ensures we cover different video capture technologies, compression techniques, etc. Specifically, the videos are collected from Sitcom snippets, interviews, and movies. Some examples are The Office (Sitcom), Alex Honnold's Interviews, Think Media's tutorials, and FRIENDS (Sitcom).

3.5.3 Human Evaluation

We conduct human evaluations as part of our qualitative evaluations, primarily to assess the quality of video-to-video face swapping achieved by our network. We randomly select ten videos from our curated dataset, and the results from all the comparisons and our network are displayed in a random order to the user. A pool of 50 participants is asked to assign a score between 1-10, indicating the perceptual quality of the generated videos. Our participant pool comprised people aged between 25-45 years of age. At the time of rating, every user was asked to rate a video on a scale of $1 - 10$, 1 and 10 being the worst and the best, respectively. Each user was shown a source video, a target video, and the final swapped video. The swapped video could be randomly from FSGAN, Motion-coseg, or FaceOff. Each user saw 10 instances of each category during rating. They had to answer the following three questions: (1) How natural does this video (swapped) look? (2) How similar is the expression in the swapped video to the source expression? and (3) How similar is the identity in the swapped video to the source identity? No additional directions were given to the users for rating. Along with the rating, they were also asked to submit their subjective opinion on the naturalness aspect of the swapped video. The mean opinion scores of all the users are reported. We also try to summarize their opinion in this section.

As was observed in Table 3.2, we outperformed the existing approaches in preserving the source identity in both quantitative and qualitative evaluation. However, FSGAN was voted slightly better qualitatively for the naturalness factor. Hereon, we will discuss the naturalness factor of the observed videos. Out of the three, the highest variations in the user rating were observed to be in Motion-coseg. FaceOff had the least variation in rating, and almost all the videos appear natural. Although FSGAN was rated highest in terms of naturalness, the users commented that the output had unnatural color. Despite the drawback, the users agreed that the overall expression and the swapped person looked natural. It is to be noted that despite FSGAN being voted as producing more natural outputs than FaceOff, the task of identity swapping was unanimously voted to be superior in FaceOff. Although FSGAN preserved the source identity and looked more natural, the users agreed that the output had a little match with either of the expressions - source or target. This meant that the model took leeway in creating expressions as long as the output looked natural.

3.6 Experiments and Results

In this section, we try to answer the following questions: (1) How well can we preserve the source identity compared to the alternate approaches? (2) How well do we preserve the expressions of the input videos? (3) How efficient is FaceOff when compared to other techniques?

We compare FaceOff against different tasks: “face-swapping”, “face reenactment”, and “face editing”. Please note that none of these methods can fully solve the task of V2V face-swapping that we aim to solve. Specifically, V2V face-swapping aims to (1) swap source identity and expressions and (2) retain the target pose and background. In the next section, we introduce the quantitative and qualitative metrics used for comparison.

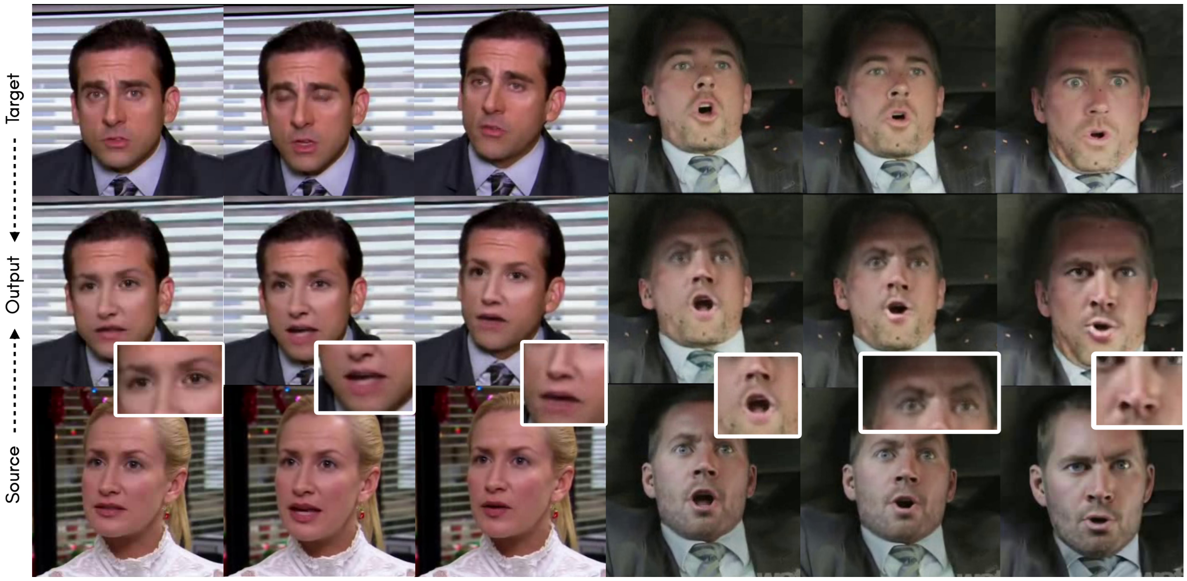

Figure 3.7: Qualitative results of FaceOff. Note that there is a significant difference in the source and target expressions in all the cases. FaceOff swaps the source expressions (mouth, eyes, etc.) and identity; and retains the target pose and background.

3.6.1 Metrics and Dataset

Quantitative Metrics: **(1) S**ource-**P**rediction **I**dentity **Dis**tance **(SPIDis)**: computes the difference in identity between face images. It is computed as the Euclidean distance between the face embeddings generated using dlib's face detection module. **(2) F**réchet **V**ideo **D**istance **(FVD)**, as proposed in [63], computes the temporal coherence in the generated video output. **(3) L**and**m**ark **D**istance **(LMD)**: evaluates the overall face structure and expressions of the source and swapped output. To compute LMD, the source and the swapped face landmarks are normalized: faces are first centered and then rotated about the x-axis so that the centroid and angle between the eye coordinates align a mean image. Next, the faces are scaled to the mean image. Euclidean distance between the normalized swapped and source video landmarks gives the LMD. We compute LMD between the source and the output face expressions (excluding the landmarks of the face permiter). **(4) T**emporally **L**ocally **(TL-ID)** and **T**emporally **G**lobally **(TG-ID) Id**entity Preservation: proposed in [62]. They evaluate a video's identity consistency at a local and global level. For both metrics, a score of 1 would indicate that the method successfully maintains the identity consistency of the original video.

Qualitative Metrics: A mean absolute opinion score on a scale of $1-10$ is reported for **(1) Identity**: How similar is the swapped-output identity with the source identity? **(2)** Expressions **(Exps.)**: How similar is the swapped-output expression with the source expression?, and **(3)** Naturalness **(Ntrl.)**: Is the generated output natural?

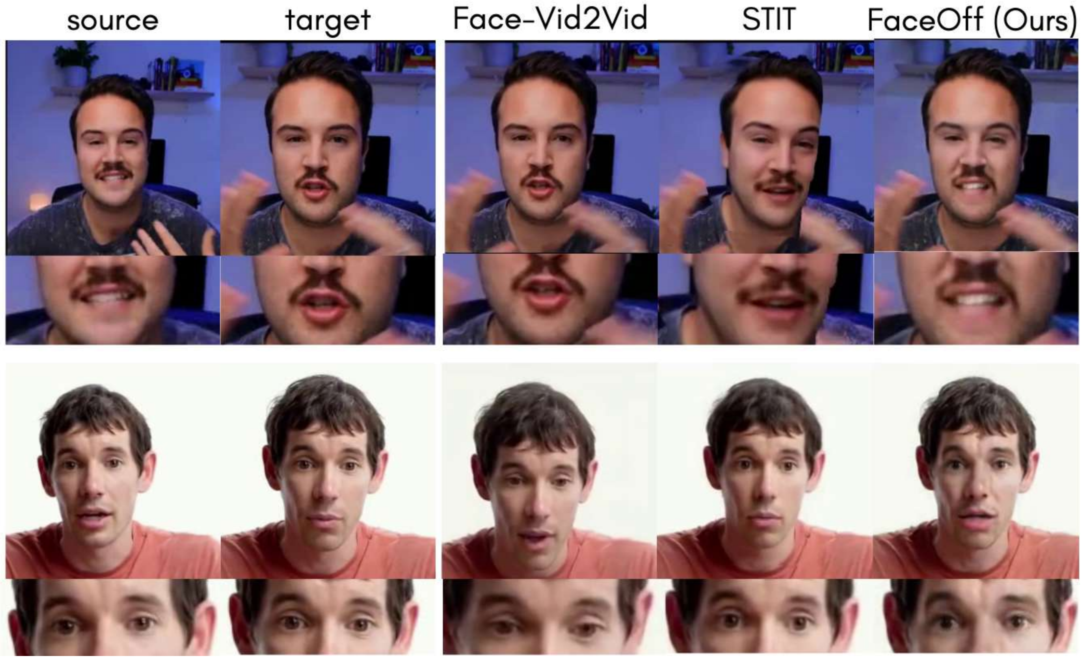

Figure 3.8: Qualitative demonstration of Face Manipulation. As can be seen, none of the methods, except FaceOff, preserve the source expressions or pose information perfectly.

Experimental Dataset: We benchmark the V2VFaceSwap dataset made of unconstrained YouTube videos with many unseen identities, backgrounds, and lighting conditions. The supplementary paper reports further details about the dataset and evaluation setup.

3.6.2 Face-Swapping Results

Fig. 3.5 and Table 3.2 present a qualitative and quantitative comparison, respectively, between the existing methods and FaceOff. Fig. 3.7 demonstrates FaceOff's face-swapping results on videos. As shown in Fig. 3.5, FaceOff successfully swaps the identity and expressions of the source face video. Existing methods cannot swap the source expressions, which shows that FaceOff solves a unique challenge of V2V face-swapping. An interesting finding of our experiments is that the existing methods do not preserve any input expressions – source or target – at the output and generate novel expressions, e.g., novel gaze direction or mouth movements. This phenomenon is also demonstrated in Fig. 3.4. FSGAN and Motion-Coseg fail to swap the identity entirely. This is further corroborated through quantitative metrics in Table 3.2. FaceOff shows an improvement of $\sim 22\%$ and $\sim 28\%$ on SPIDis and LMD over FSGAN, indicating FaceOff's superiority.

FSGAN achieves a slightly better FVD and is voted more natural in human evaluation. This is expected as FSGAN does not change the target identity much and retains the original target video making it more natural to observe. FaceOff swaps identity near-perfectly. Moreover, existing methods only have a single target motion to follow. FaceOff tackles an additional challenge of motion-to-motion swapping that needs source-target pose alignment at every frame. This requires FaceOff to generate a novel motion such that the identity, expressions, and pose in the motion look natural and match the inputs. Despite this challenge, the difference between FSGAN and FaceOff's FVD is not perceptually significant [63]. DeepFaceLabs and DeepFakes swap identity well but are $9000\times$ more computationally expensive than FaceOff, making FaceOff much more scalable and applicable in the real world.

3.6.3 Target Face Manipulation Results

Given that the source and target have the same identity, the problem reduces to the following - transfer expressions from a source video to a target video. This is fundamentally the setting of "face reenactment." One could also modify the expression of the target by identifying and quantifying the source expressions and using a "face-editing" network to edit the target expressions. Fig. 3.8 presents a qualitative comparison between FaceOff, "face reenactment" (Face-Vid2Vid) and "face-editing" (STIT).

Face Reenactment: We compare against Face-Vid2Vid [64], a SOTA face reenactment network that reenacts the pose and expression of a target image using a source (driving) video. As shown in Fig. 3.8, FaceOff preserves the source's micro-expression, such as exact mouth opening and eye-frown. FaceOff relies on a deterministic distance loss, so it can retain the precise input expressions in the output. Moreover, FaceOff retains the temporal target pose and background, whereas Face-Vid2Vid modifies a static frame.

Face Editing: Using a powerful neural network, one can introduce the desired expressions in a video by performing edits. We compare our method against STIT [62]. STIT modifies the expressions of a face video based on an input label. We observe the source expression and manually try out various intensities of the "smile" emotion ranging from negative to positive direction. As seen in Fig. 3.8, although STIT can change the overall expression, it needs a manual hit-and-trial to pinpoint the exact expression. It also lacks personalized expression (amount of mouth opening, subtle brow changes). Also, each and every expression cannot be defined using a single label, and introducing variations in emotion along the temporal dimension is hard. With our proposed method, one can incorporate any emotion in the video (as long as we have access to a source video).

3.7 Ablation Study

We investigate the contribution of different modules and errors in achieving FaceOff. Fig. 3.9 demonstrates the performance of FaceOff without the proposed temporal module. As shown, although at a frame level, the output is spatially-coherent, as we look across the frames, we can notice the temporal incoherence. The face seems to 'wobble' across the frames - squishing up and down. In fact, without the temporal module, the network does not understand an overall face structure and generates unnatural frames (marked in red). Jumping from one red box to another, we can see that the face structure has completely changed. This suggests that constraining the network by the neighboring frames using the

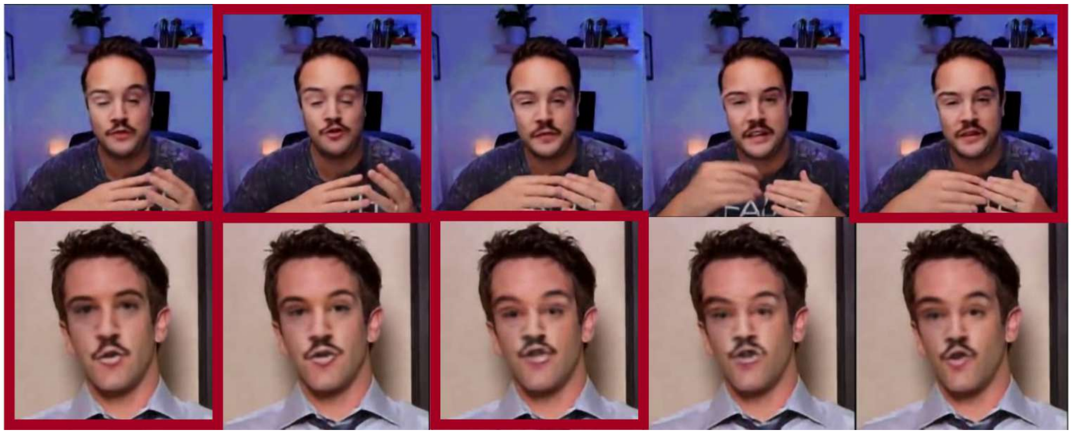

Figure 3.9: FaceOff without Temporal Module. As we jump from one frame to another (red boxes), we can observe a ”wobble effect”: significant change in the facial structure (elongated and then squeezed). This occurs as the model does not have an understanding of the neighboring frames while generating the current frame.

Component	SPIDis ↓	LMD ↓	FVD ↓
FaceOff	**0.38**	**0.41**	**255.980**
w/o Temporal.	0.71	0.49	350.60
w/o Rotation	0.65	0.44	292.76
w/o Color	0.74	0.42	303.35
w/o Translation	0.58	0.47	271.82
w/o Distortion	0.55	0.45	285.54

Table 3.4: We remove different components and errors and evaluate their contributions in achieving FaceOff.

temporal module enables the network to learn a global shape fitting problem, consequently generating a temporally coherent output.

Table 3.4 presents the quantitative contribution of the temporal module and each of the errors used for self-supervised training. The metrics indicate that each of them contributes significantly to achieving FaceOff.

As mentioned in Section 3.3, we introduce five types of pseudo errors: rotation, translation, scaling, distortion, and color, at the time of training to emulate the different errors we face during inference. In this section, we perform an ablation to show the effects (at the time of inference) of removing each error during training. In each subsection, we try to remove the errors one at a time. i.e., as we remove rotation, the remaining four errors are still present while training. To showcase the clear distinction between the foreground and the background, we turn off the color error for all the ablations.

As clearly depicted in Fig. 3.6, each error causes a degradation in the output. The leftmost column in the figure shows the effect of not introducing the color normalization error. This leads to sub-optimal blending between the source and target face with significant artifacts. Similarly, the scale and rotation pseudo errors are also extremely important, as shown in the same figure, Fig. 3.6. Removing the scaling error causes the blended face to be on a different scale. On the other hand, the rotation error forces the faces to be aligned, making it easier for the algorithm to blend. Finally, without the translation error, the source face does not fit the target face giving rise to an unstructured output. A conjunction of these different errors leads to a setting where the model can blend the given videos spatially and temporally. Affine transformation is a combination of scaling, rotation, and translation. Therefore, removing one of these errors does not confuse the model of the underlying task of alignment. The model still performs the task well and can fit the irregular face shape into the background. However, distortion error (as shown in the figure's last column) is very important. Without the distortion error (which is, in fact, the non-linear transformation), the model struggles to warp the face in a way that best fits the background. This causes the foreground to go out of the background and generate unnatural outputs.

3.8 Conclusion

We introduce "video-to-video (V2V) face-swapping", a novel task of face-swapping. Unlike face-swapping, which aims to swap an identity from a source face video (or an image) to a target face video, V2V face-swapping aims to swap the source expressions along with the identity. To tackle this, we propose FaceOff, a self-supervised temporal autoencoding network that takes two face videos as input and produces a single coherent blended output. As shown in the experimental section, FaceOff swaps the source identity much better than the existing approaches while also being $400\times$ computationally efficient. It also swaps the exact source identity that none of the methods can do. V2V face-swapping has many applications; a significant application can be automating the task of replacing the double's face with the actor's identity and expressions in movies. We believe our work adds a whole new dimension to movie editing that can potentially save months of tedious manual effort and millions of dollars.

3.8.1 Our results and Potential Applications

Our approach has several potential applications, especially in multimedia, entertainment, and education.

We demonstrate two such applications in this paper. The first is depicted in Fig. 3.7, which shows a real-use case of Paul Walker. In post-production, the VFX team replaced the face of Cody and Caleb Walker, who acted as Paul's double[1]. The team underwent extensive graphical post-processing to superimpose Paul's face from previous recordings of Cody and Caleb. In Fig. 3.1, we demonstrate another result of FaceOff. Here, we simulate a scenario of body doubles. Nolan, the actor in the source video, is 'working from home' recording his dialogues and expressions at the convenience of his home. Joey Tribiani, the double in the target video, acts in the famous sitcom FRIENDS. FaceOff swaps Nolan into the scene in one forward pass! We show such an application in the supplementary video and encourage our readers to view the result of double-actor V2V face-swapping. FaceOff can potentially save millions

[1] Redevelopment of Walker's character

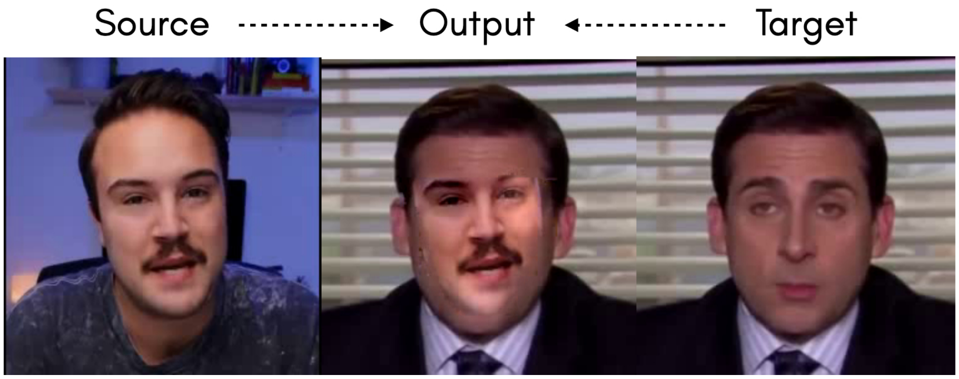

Figure 3.10: Sample output of blending using the classical technique of Poisson blending.

of dollars and reduce months of post-production edits to merely a few minutes of touch-ups on top of the FaceOff output!

Another application of our work is post-production movie editing. Today, multiple scenes are anticipated in advance to avoid retakes during post-production. Our work will encourage the movie-production team to become more flexible with doubles and post-production movie edits.

FaceOff also has huge potential in the advertisement sector and could be a potential futuristic technique for making advertising videos. Today, the VFX and CGI take abundant resources for V2V face swapping, whereas, with our work, one could replace themselves in a sitcom in less than a second. This could also become a potential teaching technique. For example, creating light-hearted advisory videos about vital life lessons for students. Our work can also be applied in animation[2] to swap an existing face/background in multiple scenes.

3.9 Additional Results

3.9.1 Poisson Blending vs Neural Blending

In this section, we observe that the blending approach fails to produce convincing results by simply applying a heuristic blending technique like Poisson blending on the heuristically aligned frames. The neural blending approach learns a non-linear transformation and blending strategy on the given input that cannot be emulated with a heuristic blending approach like Poisson blending. Poisson blending performs blending well when the source and the target faces are well aligned. It fails to generalize to cases where there is a difference between the source and the target faces, and learning an affine transformation no longer suffices. Faces are rigid bodies, and a rigid-body transformation does not suffice for cases with considerable head differences between the source and the target frames.

Moreover, Poisson blending requires precise alignments and masks to paste the source face onto the target face. A sample output of Poisson blending is shown in Fig. 3.10. The blending was performed

[2] List of recycled animation in Disney movies

after the heuristic alignment step, as shown in Fig. 3.3. As can be seen, even though the images were blended, the output looks unnatural and distorted.

3.9.2 Accuracy vs Inference Trade-Off

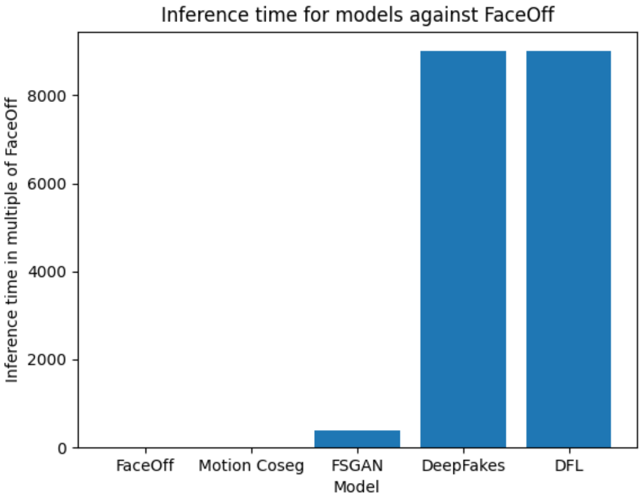

Figure 3.11: Comparison on the time needed for performing video-to-video face swapping. Faceoff is considered to need $1\times$ inference time optimization, every other model is plotted relative to FaceOff's inference time. Motion Co-seg: $1.5\times$, FSGAN: $400\times$, DeepFakes: $9000\times$, and DeepFaceLabs: $9000\times$.

In graph 3.11, we demonstrate the huge disparity between the inference times of our approach against SOTA approaches DeepFakes, DeepFaceLabs (denoted by DFL), Motion Coseg, and FSGAN. Our approach and motioncoseg are one-shot approaches, and do not require further finetuning. Fsgan provides two modes of inference, a faster inference and an inference that requires finetuning the output. We used the second approach to further improve Fsgan's results and achieved the finetuning in 5 minutes for qualitative results. Quantitative scores were computed without any optimization. Deepfakes and Deepfacelabs require considerable amount of time to achieve reasonable face-swapping and they work on a pair of videos with heavy compute. Even though our approach is one shot, we outperform existing approaches in the SPI metric, as shown above. We achieve the best SPI of 0.38 over all the baseline approaches.

3.10 Limitations and Future Work

Our work fundamentally lacks two areas: (1) Pose difference in the 'Z' direction (normal to the image) between the source and target. The network struggles to generate coherent outputs. As can be seen in Fig. 3.12, the lips and the overall production seem unnatural. Going beyond 2D images and exploring the space of 3D modeling could be an exciting way to approach this issue. Extracting 3D information from 2D images has witnessed tremendous research recently, and one could model the face as a 3D model. Aligning the face model in the 3D space allows incorporating more degrees of freedom, and consequently better output generation. (2) Difference in face ornaments. As can be seen in Fig. 3.12, artifacts such as part of the hair and spectacles are visible in the output. As we avoid adding a discriminator, the model does not learn to 'remove' any input part to make the output more realistic. For future work, one could experiment with soft discriminators such that there are minimum hallucinations.

Lastly, we extract the source face using the eye and mouth region landmarks. However, a part of one's identity also includes the head region. We do this to preserve the pose of the target. In the specific use-case we tackle, a double is selected such that the head of the double is similar, if not the same, to the actor (see Fig. 3.1). Thus, extracting only the face region is sufficient for preserving the identity in our case. However, to preserve the entire identity, one would have to move from face-swapping to head-replacement [41], which would also be an interesting direction of exploration. Here, one would need to be able to transfer the head pose of the target to the source head while preserving the other necessary characteristics.

3.11 Ethical Issues

Unlike other generative works in similar settings, we do not re-enact a given identity according to a driving video. Our work focuses on swapping relevant parts of the source video onto the target video so that the expression and lip movements of the source video are preserved. At the same time, the head motion and background remain the same as the target video. This ensures that the generated identity and the spoken content in the generated video match the source speaker (extensively evaluated in Table 3.2). Thus, body doubles and doppelgangers of celebrities cannot be directly used to re-enact a target celebrity video since the final generated identity will be copied from the source. However, since our work deals with modifying critical facial features of the target identity, we decide to take further steps to ensure fair use. We will only release the code after signing legal agreements with the users to maintain records. We will also use a visible watermark on the generated video to ensure they remain identifiable and fake.

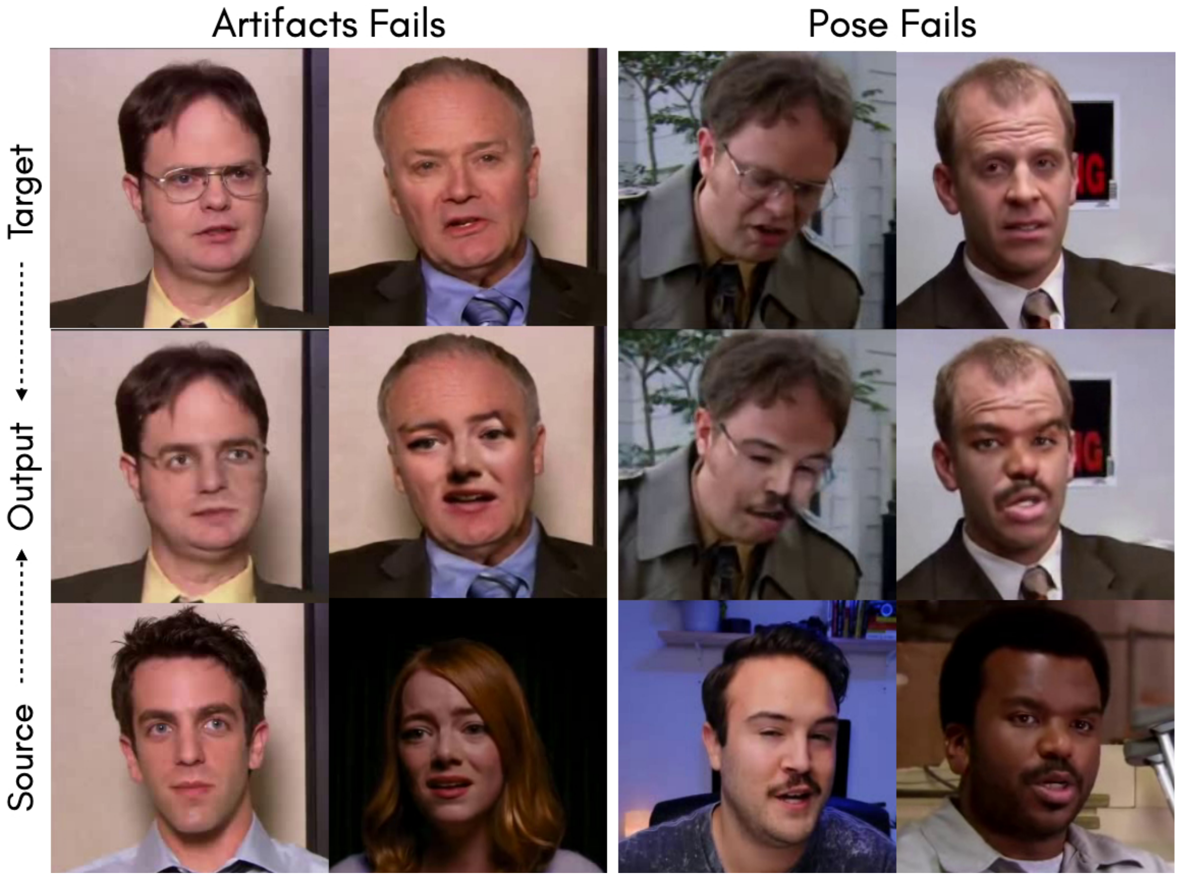

Figure 3.12: Limitations of our approach. Artifacts such as hair strands and spectacles are visible. In case of extreme pose change, the network struggles to produce a coherent output.

Chapter 4

Conclusions

In this thesis, we systematically investigated the use of synthetic talking-head videos in generating digital content. Specifically, we studied the challenges and limitations associated with existing approaches that rely on generating content manuallly. Fundamentally, our thesis explores the question - "Can synthetic talking heads replace real human talking-head videos?". We study this from the aspect of generating synthetic talking head videos as a replacement for real talking heads in online lipreading training and the use of automated face-swapping techniques for the task of actor-double face-swapping in the moviemaking industry as an alternative to manual and time consuming CGI techniques. We proposed novel solutions based on synthetic talking head generation, which have the potential to significantly improve these applications and make them more accessible and cost-effective.

In Chapter 2, we investigate the challenges associated with lipreading on existing online lipreading training platforms. Lipreading is a primary mode of communication for many people suffering from some form of hearing impairment. Consequently, they rely on speech therapists or online lipreading training platforms to learn to lipread. However, as we showed, existing platforms are severely limited - they do not incorporate real-world variations in the speaker's pose or style of speaking, have limited vocabulary, and are available in a few select languages and accents. Moreover, bootstrapping a lipreading platform from scratch is an expensive and time-consuming ordeal, requiring setting up expensive studio and camera equipment, recording on hired actors, and taking months of manual effort to complete. Such a costly endeavor effectively leaves the majority of the users reliant on online lipreading training. We proposed an end-to-end pipeline to develop such a platform using state-of-the-art talking head video generator networks, TTS models, and computer vision techniques, to generate lipreading examples automatically. Through our experimental evaluations, user studies, and statistical analysis, we present conclusive and concrete evidence that synthetic talking head videos can serve as a replacement for real human talking head videos. Our approach incorporating several interconnected deep learning modules is modular, and can bootstrap a new lipreading training platform in any language and accent in a few hours without any manual effort or intervention. Our user studies and field evaluations indicate that the hard-of-community deeply appreciates our work and provides a solution to a socially relevant problem that can positively impact millions of people over the world.

In Chapter 3, we investigated the challenges associated with using body doubles in the moviemaking industry. Doubles play an indispensable in the moviemaking industry as they take the place of actors in dangerous stunt scenes or in scenes where the same actor plays multiple characters. In these scenes, the double's face is later replaced by the actor's face and expressions using CGI technology that require hundreds of hours of manula multimedia edits on heavy graphical units costing millions of dollars and taking months to complete. Automated face-swapping techniques using deep learning models could be potentially applied to perform such a task automatically. However, as we show in this thesis, existing face-swapping techniques fail to preserve the expressions of the actor and only swap the identity. To tackle this, we proposed video-to-video face-swapping, a novel task of face-swapping, that aims to swap the identity and expressions from the source face video (or image) while preserving the pose and background of the target face video. Our approach relies on a self-supervised temporal autoencoding network, that takes two face videos as input and produces a single coherent blended output. Our metrics show that our technique swaps the source identity much better than existing approaches while also being computationally efficient. Our approach adds a whole new dimension to movie editing that can potentially save months of tedious manual effort and millions of dollars.

Future Work: Overall, the research asks important questions around the use of talking head videos and if they could replace real human talking head videos. The research demonstrates the vast potential of synthetic talking head generation in several avenues - such as the entertainment industry, education, and in virtual reality. Generating data automatically in a cost-effective and scalable fashion that is visually indistinguishable from real talking-head videos could significantly increase content creation that could aid several applications. For instance, in the lipreading training platform, building talking-head generator models that incorporate prosody can make lipreading training more effective, as facial expressions are an important cue for lipreading. Similarly, in FaceOff, we observed noticeable artefacts when there is a significant difference in the face pose of the source and target actors. These artefacts are expected, as the task of video face-swapping was limited to the space of images (2D). By inferring 3D information from 2D images, such as through analysis by synthesis approaches, video face-swapping can be significantly improved by merging the motion of the videos in the 3D space, allowing more control over the blended output due to higher degrees of freedom, and generating the 2D videos through image-based rendering.

Related Publications

- **Towards MOOCs for Lipreading: Using Synthetic Talking Heads to Train Humans in Lipreading at Scale**, Aditya Agarwal*, Bipasha Sen*, Rudrabha Mukhopadhyay, Vinay Namboodiri, C V Jawahar. *In IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) 2023.*
- **FaceOff: A Video-to-Video Face Swapping System**, Aditya Agarwal*, Bipasha Sen*, Rudrabha Mukhopadhyay, Vinay Namboodiri, C V Jawahar. *In IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) 2023.*

Other publications:

- **INR-V: A Continuous Representation Space for Video-based Generative Tasks**, Bipasha Sen*, Aditya Agarwal*, Vinay Namboodiri, C V Jawahar. *In Transactions on Machine Learning Research (TMLR) 2023*.
- **Personalized One-Shot Lipreading for an ALS Patient**, Bipasha Sen*, Aditya Agarwal*, Rudrabha Mukhopadhyay, Vinay Namboodiri, C V Jawahar. *In The British Machine Vision Conference (BMVC) 2021*.
- **Fréchet Semantic Distance: A new metric to evaluate the underlying semantics of generated images**, Sai Niranjan, Rudrabha Mukhopadhyay, Bipasha Sen, Aditya Agarwal , Vinay Namboodiri, C V Jawahar. *Under review*.

(*equal contribution)

Bibliography

[1] A. Agarwal, B. Sen, R. Mukhopadhyay, V. P. Namboodiri, and C. V. Jawahar. Faceoff: A video-to-video face swapping system. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3495–3504, January 2023.

[2] A. Agarwal, B. Sen, R. Mukhopadhyay, V. P. Namboodiri, and C. V. Jawahar. Towards moocs for lipreading: Using synthetic talking heads to train humans in lipreading at scale. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2217–2226, January 2023.

[3] D. Aneja, D. McDuff, and S. Shah. A high-fidelity open embodied avatar with lip syncing and expression capabilities. In *2019 International Conference on Multimodal Interaction*, ICMI '19, page 69–73, New York, NY, USA, 2019. Association for Computing Machinery.

[4] S. B, G. S, and T.-M. N. Effects of context type on lipreading and listening performance and implications for sentence processing. In *J Speech Lang Hear Res*. JJournal of Speech, Language, and Hearing Research (JSLHR), 2015.

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

[6] R. Chen, X. Chen, B. Ni, and Y. Ge. SimSwap. In *Proceedings of the 28th ACM International Conference on Multimedia*. ACM, oct 2020.

[7] Y.-T. Cheng, V. Tzeng, Y. Liang, C.-C. Wang, B.-Y. Chen, Y.-Y. Chuang, and M. Ouhyoung. 3d-model-based face replacement in video. SIGGRAPH 2009, 01 2009.

[8] C.-M. Chien, J.-H. Lin, C. yu Huang, P. chun Hsu, and H. yi Lee. Investigating on incorporating pretrained and learnable speaker representations for multi-speaker multi-style text-to-speech, 2021. ICASSP 2021.

[9] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. 2017. CVPR 2018.

[10] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha. Stargan v2: Diverse image synthesis for multiple domains, 2019. CVPR 2020.

[11] M. Chu and H. Peng. An objective measure for estimating mos of synthesized speech. pages 2087–2090, 01 2001.

[12] J. S. Chung, A. Jamaludin, and A. Zisserman. You said that? In *British Machine Vision Conference*, 2017.

[13] A. Clark, J. Donahue, and K. Simonyan. Efficient video generation on complex datasets. *ArXiv*, abs/1907.06571, 2019.

[14] F. A. de Barros Reis, P. Dornhofer Paro Costa, and J. M. de Martino. Deeply emotional talking head: A generative adversarial network approach to expressive speech synthesis with emotion control. In *ACM SIGGRAPH 2020 Posters*, SIGGRAPH '20. Association for Computing Machinery, 2020.

[15] Z. Ding, X.-Y. Liu, M. Yin, and L. Kong. Tgan: Deep tensor generative adversarial nets for large image generation. *arXiv preprint arXiv:1901.09953*, 2019.

[16] P. Esser, R. Rombach, and B. Ommer. Taming transformers for high-resolution image synthesis, 2020. CVPR 2021.

[17] D. Feng, S. Yang, S. Shan, and X. Chen. Learn an effective lip reading model without pains. *arXiv preprint arXiv:2011.07557*, 2020.

[18] O. Fried, A. Tewari, M. Zollhöfer, A. Finkelstein, E. Shechtman, D. B. Goldman, K. Genova, Z. Jin, C. Theobalt, and M. Agrawala. Text-based editing of talking-head video. *ACM Trans. Graph.*, 38, 2019.

[19] O. Fried, A. Tewari, M. Zollhöfer, A. Finkelstein, E. Shechtman, D. B. Goldman, K. Genova, Z. Jin, C. Theobalt, and M. Agrawala. Text-based editing of talking-head video. *ACM Trans. Graph.*, 38, 2019.

[20] P. Garrido, L. Valgaerts, H. Sarmadi, I. Steiner, K. Varanasi, P. Pérez, and C. Theobalt. Vdub: Modifying face video of actors for plausible visual alignment to a dubbed audio track. *Comput. Graph. Forum*, 34(2):193–204, may 2015.

[21] D. Ha, A. Dai, and Q. V. Le. Hypernetworks. *arXiv preprint arXiv:1609.09106*, 2016.

[22] C. Hopwood. Concentration fatigue, Feb 2021.

[23] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks, 2016. CVPR 2017.

[24] K. Ito and L. Johnson. The lj speech dataset. `https://keithito.com/LJ-Speech-Dataset/`, 2017.

[25] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. L. Moreno, and Y. Wu. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 4485–4495, Red Hook, NY, USA, 2018. Curran Associates Inc.

[26] P. K R, R. Mukhopadhyay, J. Philip, A. Jha, V. Namboodiri, and C. V. Jawahar. Towards automatic face-to-face translation. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, page 1428–1436, New York, NY, USA, 2019. Association for Computing Machinery.

[27] P. Kapoor, R. Mukhopadhyay, S. B. Hegde, V. Namboodiri, and C. V. Jawahar. Towards automatic speech to sign language generation, 2021. INTERSPEECH 2021.

[28] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Nießner, P. Pérez, C. Richardt, M. Zollöfer, and C. Theobalt. Deep video portraits. *ACM Transactions on Graphics (TOG)*, 37(4):163, 2018.

[29] J. Kim, S. Kim, J. Kong, and S. Yoon. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 8067–8077. Curran Associates, Inc., 2020.

[30] I. Korshunova, W. Shi, J. Dambre, and L. Theis. Fast face-swap using convolutional neural networks, 2016. ICCV 2017.

[31] J. Kruschke. Bayesian estimation supersedes the t test. *Journal of experimental psychology. General*, 142, 07 2012.

[32] R. Kumar, J. M. R. Sotelo, K. Kumar, A. D. Brébisson, and Y. Bengio. Obamanet: Photo-realistic lip-sync from text. *ArXiv*, abs/1801.01442, 2018.

[33] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping, 2019. CVPR 2020.

[34] I. Magnusson, A. Sankaranarayanan, and A. Lippman. Invertible frowns: Video-to-video facial emotion translation, 2021. ACM Multimedia 2021.

[35] B. Martinez, P. Ma, S. Petridis, and M. Pantic. Lipreading using temporal convolutional networks, 2020. ICASSP 2022.

[36] P. Masrori and P. V. Damme. Amyotrophic lateral sclerosis: a clinical review. *European Journal of Neurology*, 27(10):1918–1929, July 2020.

[37] M. Mirza and S. Osindero. Conditional generative adversarial nets, 2014. arXiv.

[38] J. Naruniec, L. Helminger, C. Schroers, and R. Weber. High-resolution neural face swapping for visual effects. *Computer Graphics Forum*, 39:173–184, 07 2020.

[39] Y. Nirkin, Y. Keller, and T. Hassner. Fsgan: Subject agnostic face swapping and reenactment, 2019. ICCV 2019.

[40] A. v. d. Oord, O. Vinyals, and K. Kavukcuoglu. Neural discrete representation learning, 2017. NIPS 2017.

[41] I. Perov, D. Gao, N. Chervoniy, K. Liu, S. Marangonda, C. Umé, M. Dpfks, C. S. Facenheim, L. RP, J. Jiang, S. Zhang, P. Wu, B. Zhou, and W. Zhang. Deepfacelab: Integrated, flexible and extensible face-swapping framework, 2020.

[42] K. R. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 484–492, 2020.

[43] K. R. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 484–492, 2020.

[44] A. Razavi, A. v. d. Oord, and O. Vinyals. Generating diverse high-fidelity images with vq-vae-2, 2019. NIPS 2019.

[45] Y. Ren, G. Li, Y. Chen, T. H. Li, and S. Liu. Pirenderer: Controllable portrait image generation via semantic neural rendering, 2021. ICCV 2021.

[46] J. Saunders and V. Namboodiri. Read avatars: Realistic emotion-controllable audio driven avatars, 2023.

[47] B. Sen, A. Agarwal, R. Mukhopadhyay, V. Namboodiri, and C. V. Jawahar. Personalized one-shot lipreading for an als patient. *2021 British Machine Vision Conference (BMVC)*, 2021.

[48] B. Sen, A. Agarwal, V. P. Namboodiri, and C. Jawahar. INR-v: A continuous representation space for video-based generative tasks. *Transactions on Machine Learning Research*, 2022.

[49] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783, 2018.

[50] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe. First order motion model for image animation. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[51] A. Siarohin, S. Roy, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe. Motion-supervised co-part segmentation. 2020. ICPR 2020.

[52] A. Siarohin, O. J. Woodford, J. Ren, M. Chai, and S. Tulyakov. Motion representations for articulated animation. 2021. CVPR 2021.

[53] S. Sinha, S. Biswas, R. Yadav, and B. Bhowmick. Emotion-controllable generalized talking face generation, 2022. IJCAI2022.

[54] I. Skorokhodov, S. Tulyakov, and M. Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2, 2021. CVPR 2022.

[55] H.-K. Song, S. H. Woo, J. Lee, S. Yang, H. Cho, Y. Lee, D. Choi, and K. wook Kim. Talking face generation with multilingual tts, 2022. CVPR 2022.

[56] J. Thies, M. Elgharib, A. Tewari, C. Theobalt, and M. Nießner. Neural voice puppetry: Audio-driven facial reenactment. 2019. ECCV 2022.

[57] J. Thies, M. Zollhöfer, and M. Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Trans. Graph.*, 38(4), jul 2019.

[58] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: Real-time face capture and reenactment of rgb videos. 2020.

[59] Y. Tian, J. Ren, M. Chai, K. Olszewski, X. Peng, D. N. Metaxas, and S. Tulyakov. A good image generator is what you need for high-resolution video synthesis. In *International Conference on Learning Representations*, 2021.

[60] A. Torfi, S. M. Iranmanesh, N. Nasrabadi, and J. Dawson. 3d convolutional neural networks for cross audio-visual matching recognition. *IEEE Access*, 5:22081–22091, 2017.

[61] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz. Mocogan: Decomposing motion and content for video generation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1526–1535, 2017.

[62] R. Tzaban, R. Mokady, R. Gal, A. H. Bermano, and D. Cohen-Or. Stitch it in time: Gan-based facial editing of real videos, 2022. SIGGRAPH 2022.

[63] T. Unterthiner, S. van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly. Towards accurate generative models of video: A new metric challenges, 2018. ICLR Workshop 2019.

[64] T.-C. Wang, A. Mallya, and M.-Y. Liu. One-shot free-view neural talking-head synthesis for video conferencing, 2020.

[65] W. Wu, Y. Zhang, C. Li, C. Qian, and C. C. Loy. Reenactgan: Learning to reenact faces via boundary transfer, 2018. ECCV 2018.

[66] W. Yan, Y. Zhang, P. Abbeel, and A. Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021.

[67] X. Yao, O. Fried, K. Fatahalian, and M. Agrawala. Iterative text-based editing of talking-heads using neural retargeting. *ACM Trans. Graph.*, 40:20:1–20:14, 2021.

[68] X.-W. Yao, O. Fried, K. Fatahalian, and M. Agrawala. Iterative text-based editing of talking-heads using neural retargeting. *ArXiv*, abs/2011.10688, 2020.

[69] S. Yu, J. Tack, S. Mo, H. Kim, J. Kim, J.-W. Ha, and J. Shin. Generating videos with dynamics-aware implicit generative adversarial networks. In *International Conference on Learning Representations*, 2022.

[70] S. Zarei, K. Carr, L. Reiley, K. Diaz, O. Guerra, P. Altamirano, W. Pagani, D. Lodin, G. Orozco, and A. Chinea. A comprehensive review of amyotrophic lateral sclerosis. *Surgical Neurology International*, 6(1):171, 2015.

[71] H. Zen, R. Clark, R. J. Weiss, V. Dang, Y. Jia, Y. Wu, Y. Zhang, and Z. Chen. Libritts: A corpus derived from librispeech for text-to-speech. In *Interspeech*, 2019.

[72] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric, 2018. CVPR 2018.

[73] Z. Zhang, L. Li, Y. Ding, and C. Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3660–3669, 2021.

[74] H. Zhou, Y. Sun, W. Wu, C. C. Loy, X. Wang, and Z. Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[75] Y. Zhou, X. Han, E. Shechtman, J. Echevarria, E. Kalogerakis, and D. Li. MakeltTalk. *ACM Transactions on Graphics*, 39(6):1–15, dec 2020.