Incorporating Physiological Dynamics in Synthetic Face Video

M.Sc. Thesis

by

Vipendra Singh

DEPARTMENT OF MATHEMATICS

INDIAN INSTITUTE OF TECHNOLOGY INDORE

MAY 2025

Incorporating Physiological Dynamics in Synthetic Face Video

A THESIS

Submitted in partial fulfillment of the requirements for the award of the degree of

Master of Science

by

Vipendra Singh

(Roll No. 2303141018)

Under the guidance of

Dr. Puneet Gupta and Dr. Debopriya Mukherjee

DEPARTMENT OF MATHEMATICS

INDIAN INSTITUTE OF TECHNOLOGY INDORE

MAY 2025

INDIAN INSTITUTE OF TECHNOLOGY INDORE

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the thesis entitled **"Incorporating Physiological Dynamics in Synthetic Face Video"** in partial fulfillment of the requirements for the award of the degree of **MASTER OF SCIENCE** and submitted in the **DEPARTMENT OF MATHEMATICS**, **INDIAN INSTITUTE OF TECHNOLOGY INDORE**, is an authentic record of my own work carried out during the time period from July 2024 to May 2025 under the supervision of **Dr. Debopriya Mukherjee**, Assistant Professor, Department of Mathematics, IIT Indore and **Dr. Puneet Gupta**, Associate Professor, Department of Computer Science and Engineering, IIT Indore. The matter presented in this thesis by me has not been submitted for the award of any other degree of this or any other institute.

VSingh
29/05/2025

Signature of the student with date

(**Vipendra Singh**)

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

29/05/2025

Signature of Supervisor with date

(**Dr. Puneet Gupta**)

Debopriya Mukherjee 29/05/2025

Signature of Supervisor with date

(**Dr. Debopriya Mukherjee**)

Vipendra Singh has successfully given his M.Sc. Oral Examination held on 29th May, 2025.

1.

2. Debopriya Mukherjee

Signatures of Supervisors of M.Sc. Thesis

Date: 29/05/2025

V.K.

Signature of Convener, DPGC

Date: 29/05/2025

Dedicated to my
Mom *and* ***Dad***

" I never lose. I either win or learn. Success is not defined by the outcome, but by the lessons we take from our journey. Every challenge faced, every obstacle overcome, and every setback endured is an opportunity for growth. It is through our failures and the resilience to rise again that we truly discover the strength to succeed. Remember, each failure is a stepping stone toward eventual greatness."

– ***Nelson Mandela***

Acknowledgements

I am immensely thankful to my supervisors, Dr. Puneet Gupta and Dr. Debopriya Mukherjee, for their steadfast support and insightful guidance throughout my M.Sc. journey. Their encouragement was crucial; completing this degree would not have been possible without their mentorship. Their elegant approaches to problem-solving have inspired me to examine challenges from diverse perspectives. I conclude this program with a significantly enriched understanding of mathematics and computer science, owing largely to their remarkable guidance.

I would also like to express my sincere gratitude to the committee members, Prof. Swadesh Kumar Sahoo, Prof. V. Antony Vijesh, Dr. Sourav Mitra, and Dr. Charitha Cherugondi. Their valuable advice and support has been greatly appreciated. I am equally grateful to the HOD, Dr. Sanjeev Singh, DPGC convener, Dr. Vijay Kumar Sohani, and all our professors for their kind and insightful guidance. I sincerely thank Dr. Bibekananda Maji for his constant support and encouragement. His motivation and guidance have been truly inspiring throughout my journey.

My heartfelt appreciation goes to Mr. Anup Kumar Gupta and Miss. Tr-

ishna Saikia for encouraging me to strive for greater sophistication in AI and machine learning methodologies. I am also deeply grateful to every member of the Deep Intelligence Lab for their continuous support, insightful discussions, and encouragement throughout the course of my thesis work. Their willingness to share knowledge, offer assistance, and create an environment of collaborative learning has been invaluable to my academic and personal growth.

I owe a profound debt of gratitude to my family, whose unwavering support forms the foundation of my existence. Their love and encouragement have been a constant source of strength throughout my journey, and words cannot fully capture the depth of my appreciation. A special thanks to my little brother, Etendra, whose pure love, endless encouragement, and cheerful spirit have brought light and resilience to my path, even during the most challenging phases; and to my little sister, Yashma, whose innocent smile, unwavering affection, and joyful presence have been a constant source of comfort and strength throughout this journey. I am also profoundly grateful to my best friends, Ritik, Anuj and Vinay, who have always been my home, offering unconditional support and companionship at every step.

Finally, I want to thank the wonderful friends I have made here who will remain in my heart forever: Sagar, Shekhar, Debashish Boro, Rahul, Vishal, Praveen, Bhanu, Khusboo, Lwiwe and all my classmates. To my lovely juniors, thank you for always being there.

I would like to extend my special thanks to my seniors Chhavi didi, Madhurima didi, Nishi didi, Shubham Bhau, Rahul bhaiyaa, Yash bhaiyaa, Akash bhaiyya, Kanchan Bhaiyya and all other seniors. Your guidance and encouragement have been invaluable throughout my academic journey.

Special thanks to the administrative and support staff who have made my academic journey smoother with their assistance and dedication. Your behind-the-scenes efforts have not gone unnoticed, and I am truly grateful for your hard work. In addition, I want to acknowledge the library staff and the IT department for providing essential resources and technical support, which have been invaluable throughout my research and studies.

I am deeply grateful to the Department of Mathematics, IIT Indore, for providing me with a strong academic foundation and nurturing my interest in the subject. I would also like to thank the Department of Computer Science and Engineering for offering an intellectually stimulating environment through the Deep Intelligence Lab, where I had the opportunity to explore interdisciplinary ideas. Additionally, I am thankful to the Department of Science and Technology (DST), India, for establishing the Bhaskaracharya Lab, whose excellent facilities significantly supported the progress of my research.

My deepest gratitude goes to all those who have directly or indirectly contributed to my academic journey. Your support and encouragement have been a driving force in my achievements.

Abstract

Remote photoplethysmography enables non-contact Heart Rate(HR) assessment by studying skin color fluctuations captured in facial videos. However, the availability of real face video datasets with physiological ground truth is severely limited due to privacy concerns and data collection challenges. In response, this thesis presents a novel method for integrating physiological dynamics into synthetic face videos generated through advanced synthesis techniques, specifically designed for benchmarking rPPG algorithms in healthcare and telemedicine applications. By integrating a controllable rPPG signal and simulating physiological skin tone variations, we create synthetic videos that are able to reflect the target HR. This method ensures both visual and physiological realism, offering a valuable resource for the development and evaluation of rPPG methods. Our results demonstrate that the embedded physiological signals can be reliably extracted from the generated videos, validating the effectiveness of this synthetic dataset for advancing rPPG research.

Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

The development of remote photoplethysmography (rPPG) techniques has opened new avenues for contactless physiological monitoring by extracting vital signs from facial videos (Xiao et al., 2024). However, the effectiveness of rPPG models heavily depends on the availability of high-quality, diverse, and well-annotated facial video datasets. Privacy concerns, data protection regulations, and the logistical challenges involved in capturing large-scale facial videos with corresponding ground-truth physiological signals constrain the training and benchmarking of rPPG algorithms. This scarcity of suitable datasets limits progress in the field by hindering the generalizability and robustness of existing models. To overcome these challenges, researchers have increasingly explored synthetic data generation as a viable alternative to real-world datasets (Palazzo et al., 2018).

Simultaneously, the generation of authentic human facial videos has emerged as a critical research area in computer vision and deep learning, driven by applications in virtual communication, entertainment, digital avatars, and human-computer interaction. Synthetic face videos refer to computer-generated sequences that portray a human face exhibiting naturalistic expressions, movements, and speech patterns. These videos are typically created from static visual inputs, such as a single face image combined with auxiliary modalities such as audio or motion cues. The goal is to produce video sequences that are visually plausible, temporally coherent, and preserve the identity of the subject. By leveraging synthetic videos, it becomes possible to generate large-scale datasets while maintaining control over various factors such as identity, expression, and environmental conditions.

Synthetic face video generation leverages deep generative models to animate static inputs and simulate realistic facial dynamics. This process entails learning the intricate relationships between identity, expression, and temporal motion, enabling the creation of facial videos that closely mimic natural human behavior. Recent advances in deep learning have significantly improved the capacity of these systems to generalize across diverse identities and scenarios, enhancing their applicability in real-world contexts such as telepresence, personalized animation, and video content creation. Central to these advances is the power of deep learning architectures, which have revolutionized data-driven modeling across many domains.

Deep learning, a prominent branch of machine learning inspired by the human brain's structure and function, is foundational to this domain (LeCun, Bengio, and Hinton, 2015). Utilizing artificial neural networks with multiple processing layers, deep learning models extract hierarchical representations from complex, high-dimensional data including images, audio, and text. This capability supports end-to-end learning systems that automatically identify meaningful features and synthesize them for various tasks. In synthetic video generation, deep learning enables unified, data-driven modeling of facial structure, motion, and identity.

In parallel, integrating physiological parameters into facial video synthesis has attracted increasing attention as a means to enhance realism and unlock novel applications. Physiological parameters, measurable biological signals such as HR, respiration rate, or skin temperature, manifest subtly in facial regions through changes in skin tone or micro-movements and can be captured non-invasively via video (Davies and Morris, 1993). Estimating these parameters from facial videos, for example, by extracting HR and skin temperature from subtle skin color fluctuations or micro-expressions, is an emerging field within synthetic video generation. Integrating physiological cues into synthetic face videos improves their visual authenticity and opens new possibilities in health monitoring, affective computing, and biometric security. By incorporating these signals, synthetic videos become more dynamic and lifelike, bridging the virtual and physical worlds more effectively.

Physiological parameter estimation from facial videos has thus become a key focus area. It involves extracting vital signals such as HR or skin temperature directly from video data. Advances in deep learning, computer vision, and signal processing enable accurate detection of these subtle cues, facilitating non-invasive monitoring of an individual's physiological state (Gupta et al., 2023). This method holds great promise for personalized health diagnostics, remote patient monitoring, and enhancing interactive virtual environments with real-time physiological feedback. Despite these advances, the limited availability of large-scale annotated datasets continues to challenge further progress in this field.

To address the current limitations in dataset availability and realism, this thesis proposes a novel method for integrating realistic physiological dynamics, specifically rPPG signals, into synthetically generated facial videos. By incorporating controllable, ground-truth physiological signals into synthetic videos, we aim to create privacy-preserving, scalable, and physiologically efficient datasets. Such synthetic data can be leveraged to train and evaluate rPPG models, thereby advancing research in remote physiological monitoring without dependence on sensitive real-world recordings.

CHAPTER 2

Literature Review

Physiological signal estimation has traditionally relied on Photoplethysmography (PPG), a precise contact-based method that measures cardiovascular activity through direct skin contact. However, due to its intrusive nature and practical limitations, PPG is not always feasible for many applications. To address these challenges, Remote Photoplethysmography (rPPG) has been developed as a non-contact alternative, using video analysis of facial regions to estimate physiological signals remotely. This literature review starts by examining the core principles, limitations, and theoretical models of both PPG and rPPG. Next, it covers advanced generative modeling techniques, including Generative Adversarial Networks (GANs) and Diffusion Network, which are widely used for creating synthetic data. Finally, recent advancements in synthetic face video generation are discussed, with a focus on methods like SadTalker that enable realistic and controllable facial synthesis. Together, these topics provide the foundation for our method to integrating physiological dynamics into synthetic videos.

2.1 Photoplethysmography (PPG)

Photoplethysmography (PPG) is a high-precision light-based sensor technology used to measure blood flow volume, allowing the detection of fluctuations in HR (Sahin et al., 2021). It is a contact-based method that makes measurements on the surface of the skin. For example, smartwatches and pulse oximeters utilize this technique.

2.1.1 Working Principle of PPG

Photoplethysmography (PPG) operates by directing specific wavelengths of light typically green, red, or infrared onto the skin and measuring the amount of light that is either absorbed or reflected by the underlying tissues. The fluctuations in blood volume that occur during the cardiac cycle influence the intensity of the detected light. These variations in light absorption enable the estimation of physiological signals related to cardiac activity, such as HR (Sahin et al., 2021). As shown in Figure 2.1.

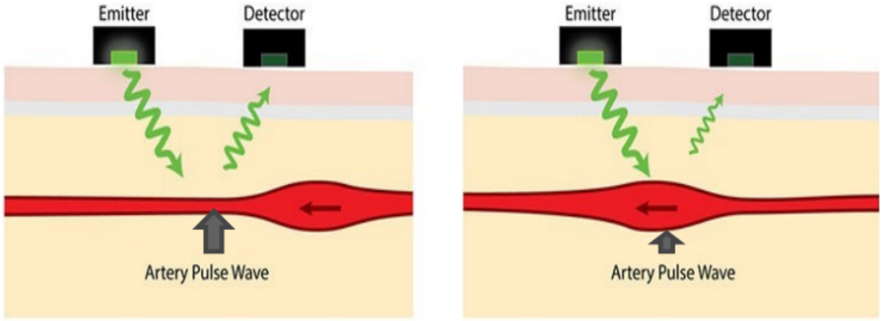

Figure 2.1: Working principle of PPG

2.1.2 Limitations of PPG

The contact-based techniques are unsuitable for monitoring:

- Patients with **sensitive skin** (Gupta, Bhowmick, and Pal, 2020).

- Patients with **damaged skin**, such as **burnt victims** (Gupta et al., 2023).
- **Neonates**, due to fragility and potential discomfort (Gupta, Bhowmick, and Pal, 2017).

While contact-based PPG methods have been widely adopted for HR monitoring, they are not ideal for long-term or continuous use. Extended wear of such devices may result in user discomfort and can even lead to skin irritation or infections, particularly in sensitive individuals or high-humidity environments. These limitations highlight the need for non-contact alternatives such as remote photoplethysmography (rPPG), which enables physiological monitoring without direct skin contact.

2.2 Remote Photoplethysmography (rPPG)

2.2.1 What is Remote Photoplethysmography (rPPG)

Remote Photoplethysmography (rPPG) is a contactless optical technique that estimates variations in blood volume by analyzing subtle color changes in the skin (Gupta et al., 2023). As shown in Figure 2.2, it only requires a **facial video** of a person to measure those subtle changes (Gupta et al., 2023).

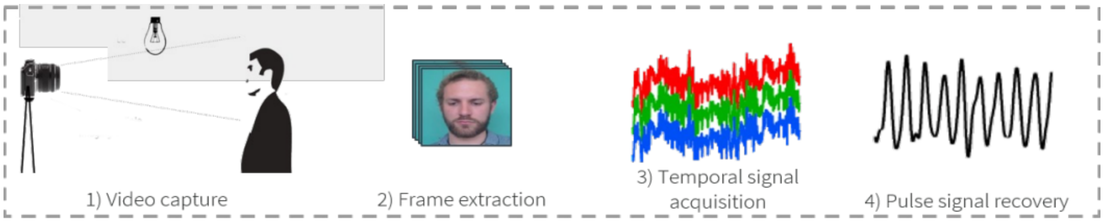

Figure 2.2: Working steps of rPPG

2.2.2 Eulerian Principle

In the Eulerian principle, variations in the intensity of reflected light rays are measured to extract rPPG information from face videos (Saikia et al., 2023) as shown in figure 2.3.

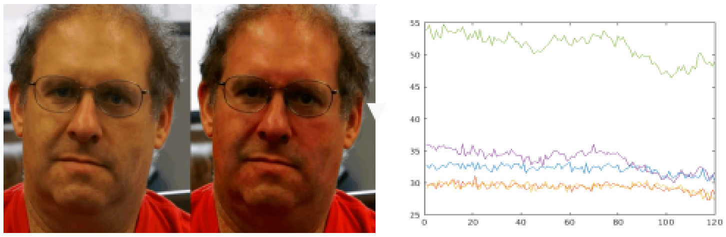

Figure 2.3: Eulerian technique which measures color variations occurring on the face and generates temporal signal by analyzing those variations

2.2.3 Lagrangian Principle

The Lagrangian technique extracts rPPG information from face videos by analyzing motion variations (Saikia et al., 2023) as shown in Figure 2.4

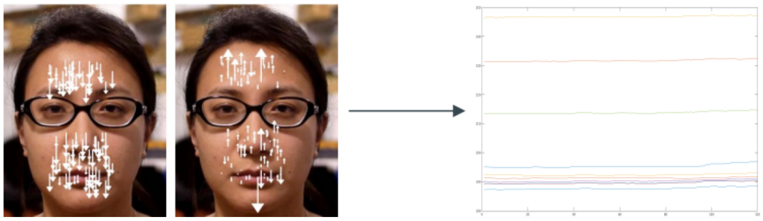

Figure 2.4: Lagrangian technique which measures motion variations occurring on the face and generates temporal signal by analyzing those variations

2.2.4 Applications of rPPG

Remote Photoplethysmography (rPPG) has emerged as a prominent technique due to its contactless nature and wide applicability across multiple domains:

- **Healthcare Monitoring:** Enables continuous, non-invasive HR and respiratory monitoring, particularly useful for telemedicine and remote patient care (Salim and Khidhir, 2024).

- **Fitness and Wellness:** Integrated into webcams and mobile cameras to estimate HR during fitness sessions without wearable devices (Salim and Khidhir, 2024).

- **Driver Fatigue Detection:** Used in automotive applications to monitor drivers' physiological states, potentially preventing accidents due to drowsiness (Tohma et al., 2021).

- **Security and Surveillance:** Supports lie detection and stress analysis by estimating physiological signals from facial videos (Tohma et al., 2021).

- **Neonatal and Burn Patient Care:** Offers a safe way to monitor vital signs in patients where contact-based sensors may cause harm or discomfort (Lee, Sivakumar, and Lim, 2024).

- **Emotion Recognition and Human-Computer Interaction:** Assists in determining user emotions or mental states by analyzing physiological signals (Lee, Sivakumar, and Lim, 2024).

2.3 Generative Adversarial Network

Generative Adversarial Networks (GANs) are a particular type of neural network that is employed for unsupervised and semisupervised generative tasks. It basically uses two branches of neural networks; one branch acts like a generator, and the other one acts like a discriminator. These two branches are trained simultaneously in a min-max game framework (Goodfellow et al., 2014).

2.3.1 Overview of GAN Structure

A GAN has two main components:

- The **generator** learns to generate data that closely resembles real data, and is indistinguishable from synthetic data (Goodfellow et al., 2014).
- The **discriminator** learns to differentiate between real data and synthetic data generated by the generator. It imposes penalty on the generator when it can identify the generator's output as synthetic (Goodfellow et al., 2014).

2.3.2 Working of GANs

The training process of a GAN involves the following steps:

- At the start, the generator creates synthetic data that the discriminator readily recognizes as counterfeit (Aggarwal, Mittal, and Battineni, 2021)
- The generated data is then passed directly to the discriminator for evaluation (Goodfellow et al., 2014).
- Using backpropagation, the discriminator's feedback guides the generator in adjusting its parameters to enhance data quality (Aggarwal, Mittal, and Battineni, 2021).
- Over time, the generator improves and begins producing more realistic data, eventually fooling the discriminator (Goodfellow et al., 2014).
- If training is successful, the discriminator becomes less effective at distinguishing real from fake, often misclassifying fake data as real (Aggarwal, Mittal, and Battineni, 2021).

The workflow of GAN is shown in Figure 2.5.

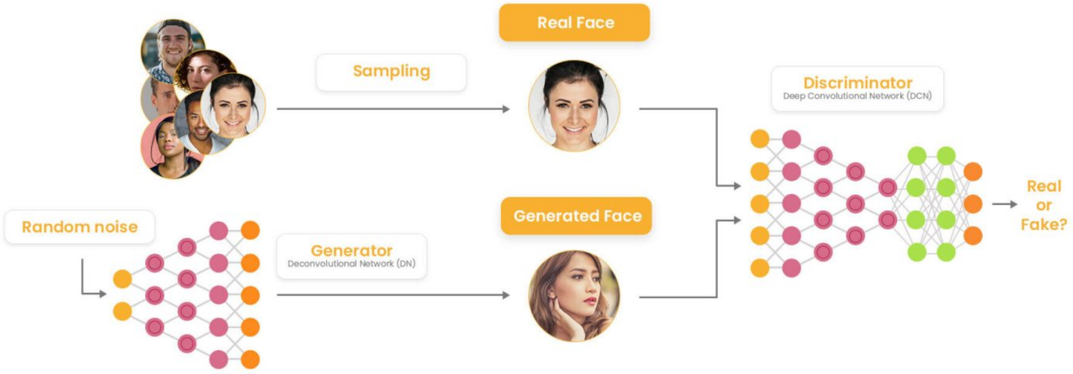

Figure 2.5: Architecture of Generative Adversarial Network (GAN)

2.3.3 GAN Objective Function

The objective of a Generative Adversarial Network (GAN) can be formulated as a minimax optimization problem involving two competing neural networks: the generator $\mathcal{G}$ and the discriminator $\mathcal{D}$. The discriminator aims to correctly identify whether a given sample originates from the true data distribution or has been generated by the generator. Whereas the generator seeks to produce data that is indistinguishable from real samples, thereby attempting to fool the discriminator.

$$\min_{\mathcal{G}} \max_{\mathcal{D}} V(\mathcal{D}, \mathcal{G}) = \mathbb{E}_{x \sim p_{\text{data}}(x)}[\ln \mathcal{D}(x)] + \mathbb{E}_{z \sim p_z(z)}[\ln(1 - \mathcal{D}(\mathcal{G}(z)))] \tag{2.1}$$

where:

- $x \sim p_{\text{data}}(x)$: Sample from real data distribution.
- $z \sim p_z(z)$: Random noise input.
- $\mathcal{G}(\cdot)$: Output of Generator.
- $\mathcal{D}(\cdot)$: Output of Discriminator.

2.3.4 Loss Functions

Discriminator Loss

$$\mathcal{L}_{\mathcal{D}} = -\mathbb{E}_{x\sim p_{\text{data}}(x)}[\ln \mathcal{D}(x)] - \mathbb{E}_{z\sim p_z(z)}[\ln(1-\mathcal{D}(\mathcal{G}(z)))] \tag{2.2}$$

Generator Loss

$$\mathcal{L}_{\mathcal{G}} = -\mathbb{E}_{z\sim p_z(z)}[\ln \mathcal{D}(\mathcal{G}(z))] \tag{2.3}$$

2.4 Diffusion Network

Diffusion is a process where we gradually add random noise to data in small steps. This is done using a sequence of steps, like in a Markov chain. Then, a network is trained to reverse this process, removing the noise step by step, until we get the desired data back from the noise (Ho, Jain, and Abbeel, 2020).

Anatomy of Diffusion Network

The process involves two phases: (i) Forward Diffusion Process and (ii) Reverse Diffusion Process.

2.4.1 Forward Diffusion Process

The network starts with the original data (e.g., an image) and progressively adds noise in small steps over time, turning it into random noise. Each step of this process is typically modeled using a simple distribution like a Gaussian, and the process is defined over several time steps, where the noise level increases gradually (Ho, Jain, and Abbeel, 2020).

We define a sequence of variables $\mathbf{x}_T, \mathbf{x}_{T-1}, \ldots, \mathbf{x}_1, \mathbf{x}_0$ where, $\mathbf{x}_T$ is pure noise, and $\mathbf{x}_0$ is the data sample. The forward diffusion process gradually adds noise to $\mathbf{x}_0$, resulting in $\mathbf{x}_T$.

Equation of Forward Diffusion Process

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\,\mathbf{x}_{t-1}, \beta_t I) \tag{2.4}$$

where:

- $q(\mathbf{x}_t|\mathbf{x}_{t-1})$ is the conditional probability of $\mathbf{x}_t$ given $\mathbf{x}_{t-1}$.
- $\mathcal{N}(\cdot)$ denotes a normal (Gaussian) distribution.
- β_t is the variance schedule at time t.
- I is the identity matrix.

Marginal Distribution

Rather than sequentially sampling $\mathbf{x}_t$ from $\mathbf{x}_{t-1}$ through all steps, we can solve for the direct relationship between $\mathbf{x}_t$ and $\mathbf{x}_0$. To do this, we marginalize over all intermediate states:

$$q(\mathbf{x}_t|\mathbf{x}_0) = \int q(\mathbf{x}_t|\mathbf{x}_{t-1})\, q(\mathbf{x}_{t-1}|\mathbf{x}_{t-2})\, \ldots\, q(\mathbf{x}_1|\mathbf{x}_0)\, d\mathbf{x}_{t-1}\, \ldots\, d\mathbf{x}_1 \tag{2.5}$$

Since the noise addition is Gaussian at each step, this leads to a closed-form solution:

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0, (1-\bar{\alpha}_t)I) \tag{2.6}$$

where

$$\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s) \tag{2.7}$$

Derivation of Cumulative Noise $\bar{\alpha}_t$

To better understand $\bar{\alpha}_t$, we rewrite the forward diffusion process (Equation 2.4) for a single step:

$$\mathbf{x}_t = \sqrt{1-\beta_t}\,\mathbf{x}_{t-1} + \sqrt{\beta_t}\,\epsilon, \tag{2.8}$$

where $\epsilon \sim \mathcal{N}(0, I)$ is Gaussian noise.

Applying this recursively for t steps from $\mathbf{x}_0$, and referring Equation 2.7, we get:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\,\epsilon_t, \tag{2.9}$$

where ϵ_t is the cumulative noise after t steps.

2.4.2 Reverse Diffusion Process

After the forward diffusion process, the model learns the reverse process. Particularly, the model is trained to reverse the diffusion by gradually denoising the noisy data and reconstructing the original data from the noise. This reverse process is also performed step by step and is typically modeled using a neural network. The functioning of the diffusion network is shown in Figure 2.6.

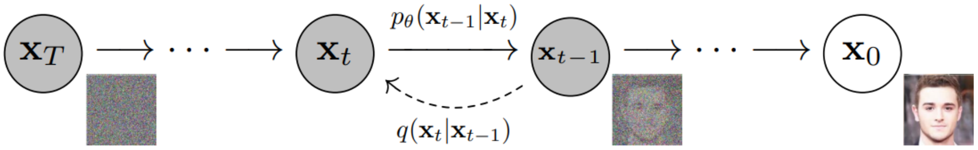

Figure 2.6: Diffusion Network

Reverse Diffusion Equation

The reverse process is modeled by a series of Gaussian distributions:

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)), \tag{2.10}$$

where:

- $\mathbf{x}_t$ is the noisy image at time step t.
- $\mu_\theta(\mathbf{x}_t, t)$ is the mean predicted by the model.

- $\Sigma_\theta(\mathbf{x}_t, t)$ is the predicted variance.

Mean of the Reverse Process

$$\mu_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(\mathbf{x}_t, t) \right), \tag{2.11}$$

where:

- α_t and β_t are time-dependent noise coefficients.
- $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$.
- $\epsilon_\theta(\mathbf{x}_t, t)$ is the noise predicted by the model.

Variance of the Reverse Process

The variance is typically either learned or fixed. In the variance-preserving setting:

$$\Sigma_\theta(\mathbf{x}_t, t) = \beta_t \tag{2.12}$$

2.4.3 Loss Function of a Diffusion Network

The goal is to train the model to learn the reverse process and recover the original data distribution. A simplified loss function based on the denoising process is often used.

Simplified Loss Function

$$L_{\text{simple}} = \mathbb{E}_{\mathbf{x}_0, \epsilon, t} \left[\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|^2 \right], \tag{2.13}$$

where:

- $\mathbf{x}_0$ is the original data sample.

- $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$, with $\epsilon \sim \mathcal{N}(0, I)$.
- $\epsilon_\theta(\mathbf{x}_t, t)$ is the predicted noise.
- ϵ is the true noise added to $\mathbf{x}_0$.

The model ϵ_θ is trained to predict the added noise ϵ at each time step, referring to Equation 2.9.

2.5 Synthetic Face Video Generation

2.5.1 SadTalker: Realistic and Stylized Audio-Driven Talking Face Generation

SadTalker presents a robust solution to the challenges of synthesizing realistic talking head videos from a single static image and audio input. It tackles common issues in earlier methods, such as unnatural head movement, expression distortion, and loss of identity, by introducing a 3D-aware and semantically disentangled framework (Zhang et al., 2023).

The method employs 3D Morphable Model (3DMM) motion coefficients as an intermediate representation, allowing the decoupling of facial expressions and head poses. This architecture is structured around three key modules:

- **ExpNet** is responsible for generating expression coefficients from audio, with a particular focus on lip motion. It uses lip-only expression targets for supervision and incorporates identity-specific priors. Additional perceptual losses, such as facial landmark and lip-reading losses, further enhance the accuracy of expression synthesis (Zhang et al., 2023).
- **PoseVAE** models head motion using a conditional variational autoencoder. Rather than predicting absolute head poses, it learns residual motion relative to an initial pose, resulting in smoother and more natural motion

sequences. Audio features and a style identity vector guide the generation of diverse head movements (Zhang et al., 2023).

- **3D-Aware Face Renderer** synthesizes photorealistic video frames by mapping the generated 3DMM motion coefficients into an unsupervised 3D keypoint domain. This allows the system to generate realistic animations without requiring a driving video as input (Zhang et al., 2023).

SadTalker demonstrates strong performance across multiple metrics, including lip synchronization accuracy, head pose diversity, and identity preservation. Its modular architecture supports the independent training of each component while allowing end-to-end inference during testing. Despite occasional artifacts in regions like the teeth and eyes, due to inherent limitations in 3DMM modeling, these issues can be alleviated through post-processing methods. In summary, SadTalker advances the state of the art in talking face generation from a single image by effectively combining realism, expressivity, and control within a unified framework.

CHAPTER 3

Proposed Method

The primary objective of this work is to integrate physiological dynamics into synthetic face videos, enhancing their realism by integrating subtle yet meaningful signals such as HR through natural skin color variations. Our method begins with generating a synthetic face video using SadTalker, which animates a static image based on an audio input to produce realistic facial movements. From this video, we extract facial frames precisely using facial landmark detection tools such as MediaPipe or OpenFace (Challa, Krishna, Chakravarthi, et al., 2023), ensuring a focused region for physiological modulation. We then generate an rPPG signal modeled as a sinusoidal waveform corresponding to a target HR, which acts as the physiological reference. Using the **CIELAB** color space (Weatherall and Coombs, 1992), we perform frame-by-frame modulation on the extracted facial frames, adjusting both the luminance (L^*) and chrominance (a^*) channels in accordance with the generated rPPG signal. This modulation simulates the subtle skin color changes caused by blood volume variations under the skin. Finally, we extract the rPPG signal from the modulated video frames to verify that the integrated physiological information accurately reflects the intended HR. This method ensures that the synthetic face videos exhibit natural expressions and movements while integrating authentic physiological dynamics through subtle skin color fluctuations, consistent and measurable changes but typically imperceptible to the naked eye.

3.1 Synthetic Face Video Generation

To begin the process, we generate the synthetic face video using the **SadTalker** framework. SadTalker is used to generate facial animations driven by the corresponding audio input. This framework allows for realistic facial expression synthesis and head pose estimation, providing us with the initial synthetic face video.

However, the generated face video might exhibit unrealistic physiological attributes, such as abnormal HR variations. Thus, additional steps are needed to integrate real physiological dynamics, such as HR, into the synthetic video.

3.2 Face Frame Extraction

After generating the synthetic face video, we use *Mediapipe* or *OpenFace* to precisely extract the face frames from the video. These tools are used to ensure that only the facial region is considered, excluding background and non-facial areas. *Mediapipe's Face Mesh* is particularly effective in detecting facial landmarks, enabling us to extract accurate face frames with minimal interference from the background.

3.2.1 Mediapipe/OpenFace Extraction Process

1. **Face Detection**: The face is detected in each frame.
2. **Landmark Detection**: Mediapipe or OpenFace is used to extract facial landmarks, ensuring the precision of the face frame extraction.
3. **Face Frame Isolation**: Using the detected landmarks, the face region is isolated, leaving only the face area for further processing.

3.3 rPPG Signal Generation and Modification

We generate a synthetic *rPPG signal* corresponding to a desired target HR, designed to simulate periodic skin reflectance changes associated with blood volume pulse. This signal is then used to modulate pixel intensities in the CIELAB color space, particularly the L^* and a^* channels, which are more sensitive to physiological variations. The b^* channel, being less relevant to rPPG dynamics, is left unmodified.

3.3.1 Synthetic Temporal Signal Generation

To simulate the cardiac pulse, we generate a simple sine wave at a frequency corresponding to the target HR, h_t of frequency f (in Hz). Let f denote this frequency and t the time vector sampled at the video frame rate. The synthetic signal $s_{\text{syn}}(f, t)$ is defined as:

$$s_{\text{syn}}(f, t) = \sin(2\pi f t) \tag{3.1}$$

Here, $f = \frac{h_t}{60}$, where HR in beats per minute (BPM), and $t = \frac{n}{\text{fps}}$, with n as the frame index and fps the frame rate. This generates a temporally smooth and periodic signal that mimics the pulsatile nature of real rPPG signals.

3.3.2 Signal-Based Pixel Modulation in Lab Space

Once the rPPG signal is generated, we modulate the L^* (lightness) and a^* (green–red opponent) channels of the CIELAB representation of each facial frame. These channels are known to exhibit stronger sensitivity to hemoglobin-related absorption changes than the b^* (blue–yellow) channel, which remains unaltered.

For each frame, the synthetic signal value at time t is scaled and added to the L^* and a^* channel values within the face region, defined by a binary mask.

3.3.3 Skin Color Modulation

1. **HR Signal**: A synthetic rPPG signal is generated based on the desired HR, using a simple sine function as defined in Equation 3.1.

2. **Skin Color Adjustment in Lab Space**: Instead of directly modifying the RGB channels, we convert each face frame to the CIELAB color space, which better aligns with human perceptual sensitivity. The skin region is modulated by altering the L^* (lightness) and a^* (green–red) channels as follows:

$$L_t^* = L_t^* + \alpha \cdot \mathrm{rPPG}(t) \tag{3.2}$$

$$a_t^* = a_t^* + \beta \cdot \mathrm{rPPG}(t) \tag{3.3}$$

   Where $\mathrm{rPPG}(t)$ is defined as the synthetic signal $s_{\mathrm{syn}}(f,t)$, L_t^* and a_t^* represent the channel intensities at time t, and α and β are the modulation strengths. The b^* channel is left unchanged due to its weaker relevance to blood volume pulse variations.

3.4 Overlaying Modified Face Frames onto Real Video

After modulating the skin tone using the synthetic rPPG signal, the modified face frames are composited back onto the original video frames to achieve a photorealistic appearance with embedded physiological dynamics.

3.4.1 Overlay Process

1. **Face Frame Modification**: The face-only frames are converted to CIELAB color space, modulated according to the synthetic rPPG signal (in L^* and a^* channels), and converted back to RGB for rendering.

2. **Frame Overlay**: The modified face regions are seamlessly overlaid onto the original video frames using a binary face mask, ensuring that the temporal modulation affects only the facial skin and not the surrounding regions (such as, background, hair, or clothing).

3.5 HR Validation of Modified Synthetic Video

Finally, to validate the effectiveness of the physiological dynamics incorporation, we extract the rPPG signals from the final modified synthetic video frames, followed by HR estimation from those rPPG signals.

3.5.1 The rPPG Signal Extraction from Synthetic Face Video

Frame Extraction from the Video

The modified synthetic video consists of a sequence of frames captured at a fixed frame rate (FPS). We extract frames from the input video:

$$I_t = V(t), \quad t = 1, 2, 3, \ldots, n \tag{3.4}$$

where I_t is the frame at time t, and n is the total frames of input video V. Frame extraction from video is shown in Figure 3.1.

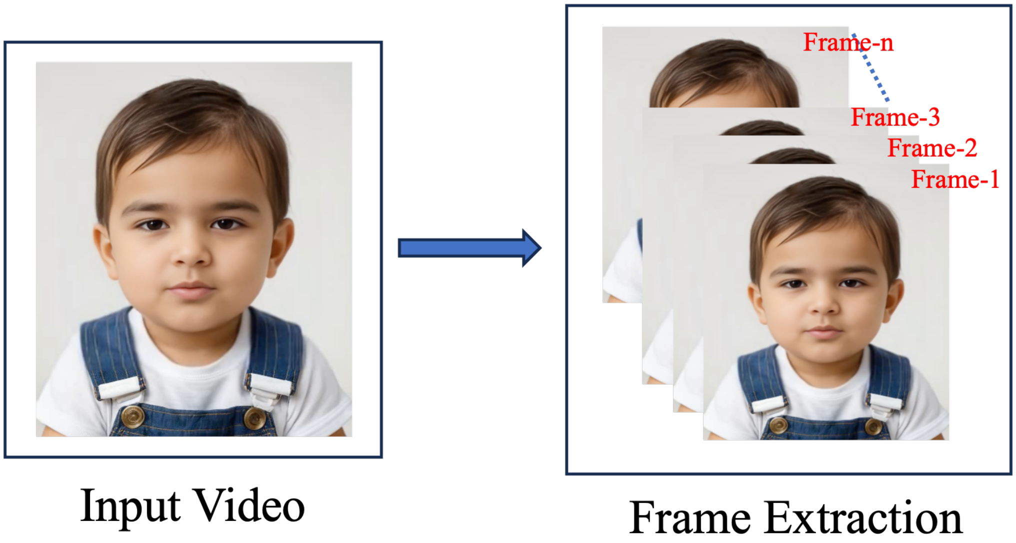

Figure 3.1: Frame Extraction

Face Localization and Landmark Detection

We detect the face in each frame using a face detection algorithm (e.g., Mediapipe or OpenCV). The same algorithm is used to find landmark points of the detected face, represented as:

$$l_t = (x, y) \tag{3.5}$$

where (x, y) are the coordinates of the landmark point l_t of the detected face. Face localization is shown in Figure 3.2.

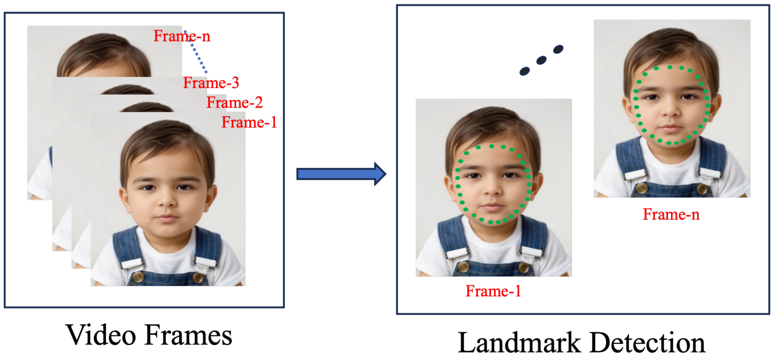

Figure 3.2: Landmark Detection

Selecting the Region of Interest (ROI) Using (x, y) Coordinates

The forehead is an ideal region for the extraction of rPPG. We extract the ROI R_t within the detected face with the help of detected landmarks:

$$R_t = (x_{\text{ROI}}, y_{\text{ROI}}, w_{\text{ROI}}, h_{\text{ROI}}) \tag{3.6}$$

where $x_{\text{ROI}}, y_{\text{ROI}}$ are the top-left coordinates of the ROI inside the face, and $w_{\text{ROI}}, h_{\text{ROI}}$ define its size as shown in Figure 3.3.

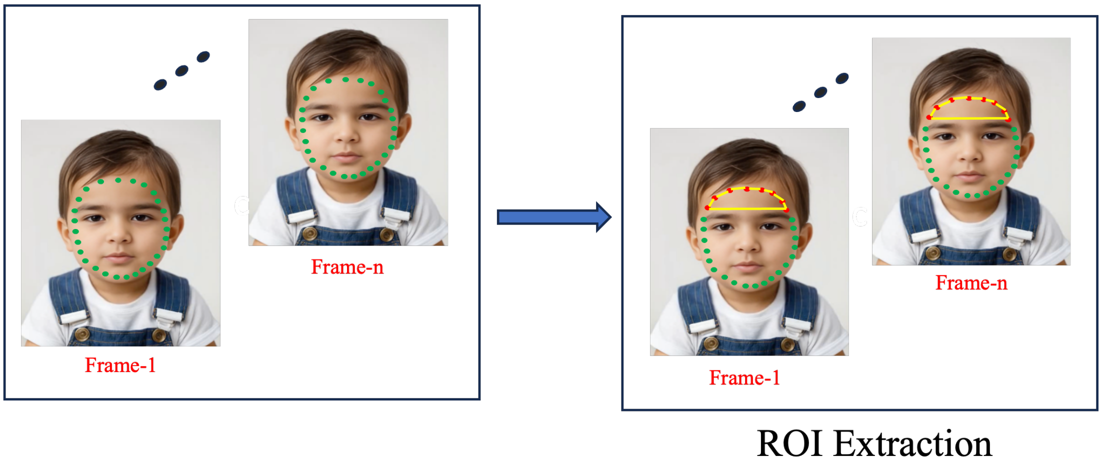

Figure 3.3: ROI extraction

Dividing ROI into Blocks for Better Signal Extraction

To reduce the noise effect, we divide the ROI into small, equal-sized blocks and extract rPPG signal for each block:

$$R_t = \bigcup_{i=1}^{m} \mathbf{b}_i \tag{3.7}$$

where b_i refers to the i^{th} block, and m is the total number of blocks present in R_t as shown in Figure 3.4.

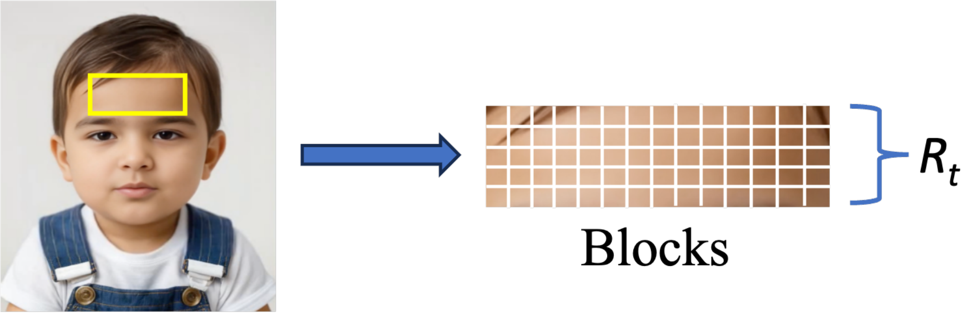

Figure 3.4: ROI division into blocks

The rPPG Signal Extraction from Each Block

We use the Eulerian principle to extract the rPPG signal from each block. The temporal signal corresponding to the i^{th} block is formulated as:

$$\mathbf{b}_i = \left[p_i^1, p_i^2, p_i^3, \ldots, p_i^n\right] \tag{3.8}$$

$$p_i^f = \frac{\sum_{(u,v)\in B_i^f} I_g(u,v)}{\eta} \tag{3.9}$$

Where:

- $\mathbf{b}_i$ → rPPG signal extracted from the i^{th} block using Eulerian principle.
- n → Total number of video frames considered.
- p_i^f → Average green channel intensity of the i^{th} block in frame f.

- B_i^f → Set containing all pixel locations within the i^{th} block of frame f.
- η → Number of pixels present in the set B_i^f.
- $I_g(u, v)$ → Green channel intensity value at the pixel coordinate (u, v).

3.5.2 Pulse Signal Extraction Using BSS

After extracting rPPG signals from each block of the selected ROI from equation (3.8). We apply Blind Source Separation (BSS) to extract the pulse signal by decomposing the mixed signals into independent components:

$$\mathbf{S} = \text{BSS}([\mathbf{b}_1, \mathbf{b}_2, ..., \mathbf{b}_m]) \tag{3.10}$$

Among the separated components, the principal component that corresponds to the physiological HR range (0.7–4.0 Hz) is selected as the pulse signal:

$$\mathbf{s}_{\text{ps}} \in \mathbf{S} \tag{3.11}$$

Pulse signal extraction using BSS is shown in Figure 3.5

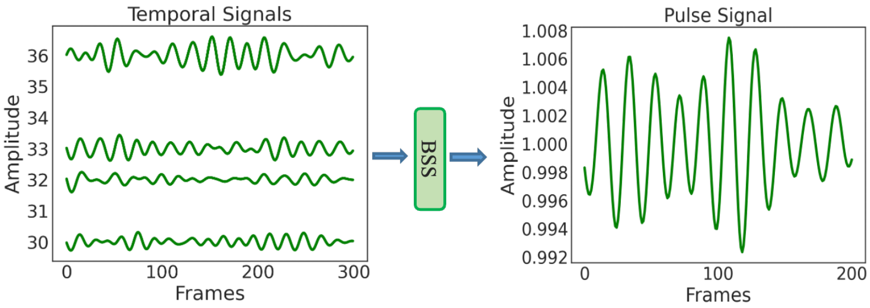

Figure 3.5: Pulse signal extracted using BSS.

To identify the primary frequency component of the extracted pulse signal, we apply the Fast Fourier Transform (FFT) on $(\mathbf{s}_{\text{ps}})$:

$$S_{\text{ps}}(f') = \mathcal{F}(\mathbf{s}_{\text{ps}}) \tag{3.12}$$

where $\mathcal{F}$ represents the Fourier Transform operation. The resulting frequency spectrum $S_{\text{pulse-signal}}(f')$ provides information on periodic variations in the intensity of skin color due to blood flow. After extracting the frequency spectrum $S_{\text{pulse-signal}}(f')$, the HR is computed from the extracted spectrum based on the principle that the dominant frequency component reflects cardiovascular activity (Gupta, Bhowmick, and Pal, 2017). The frequency with the highest magnitude, f' , corresponds to the HR frequency, as shown in the Figure 3.6 and the final HR value, h'_t, is given by:

$$h'_t = round\,(f' \times 60) \tag{3.13}$$

where round$(\cdot)$ denotes the rounding operation.

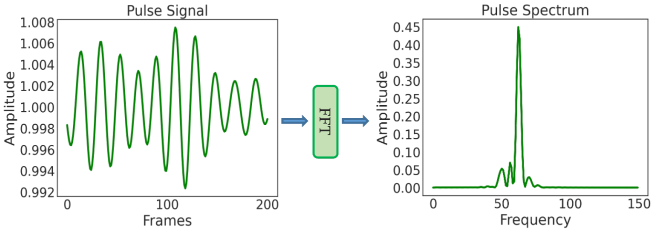

Figure 3.6: Obtaining pulse signal's spectrum.

Finally, we validate whether the extracted heart rate (HR), h'_t, matches the original target HR, h_t, which was incorporated into the face frames, by using equation (3.1). A very low difference between the extracted heart rate (HR), h'_t, and the original target HR, h_t, confirms that the *physiological dynamics* (i.e., HR modulation) have been effectively embedded into the synthetic face video. This method ensures that the generated synthetic face videos exhibit realistic *physiological dynamics*, such as skin color variations driven by HR, thereby providing a natural simulation of human facial appearance.

.

CHAPTER 4

Results and Discussion

4.1 Evaluation Protocol

4.1.1 Introduction to the Evaluation Protocol

To evaluate the effectiveness of incorporating physiological dynamics, we generate synthetic face videos using a static image and an audio signal, followed by post-generation modulation of the facial region using a predefined rPPG waveform. The physiological signal is embedded by altering pixel intensities within selected facial areas primarily the cheeks and forehead according to the temporal dynamics of the target waveform.

Once the HR incorporated synthetic videos are generated, we apply signal processing techniques to assess whether the embedded rPPG signals can be precisely retrieved. Specifically, Independent Component Analysis (ICA) is applied to the forehead region to extract the physiological signal. The extracted signal is then compared with the original reference waveform to evaluate fidelity.

This evaluation method enables precise, ground-truth-based assessment of both the integrating and retrieval processes, confirming the ability of the synthetic pipeline to produce realistic and analyzable physiological dynamics.

4.1.2 Qualitative Results

To assess the visual quality of our synthetic face frames after incorporating physiological dynamics, we conducted a frame-by-frame analysis of the rPPG modulated outputs. The goal was to ensure that rPPG-based modulation did not introduce visible artifacts or disrupt the natural appearance of facial textures, expressions, or lighting conditions.

To embed the rPPG signal, each face frame was first converted from the RGB color space to the perceptually uniform CIELAB color space. Subtle intensity modulation was then applied to the L^* (lightness) and a^* (green–red) channels based on the synthetic rPPG signal. After modulation, the frames were converted back to RGB for rendering. Due to the localized and low-amplitude nature of the modifications, the visual differences between original and modulated frames remained imperceptible to the naked eye. Photorealism was successfully preserved across a variety of poses and facial expressions.

As illustrated in Figure 4.1, the rPPG-based modulation process preserves the visual quality of the face images. The modulation was carefully applied only to selected facial regions specifically the cheeks and forehead, using facial landmark-based masking. Sensitive areas such as the eyes, lips, and background were deliberately excluded to maintain natural appearance and avoid unwanted distortions.

Although the current analysis focuses on individual frames, preliminary observations reveal a high degree of visual consistency across temporally adjacent frames. This suggests the potential for smooth temporal coherence once the modulated frames are reassembled into a video. A dynamic evaluation through

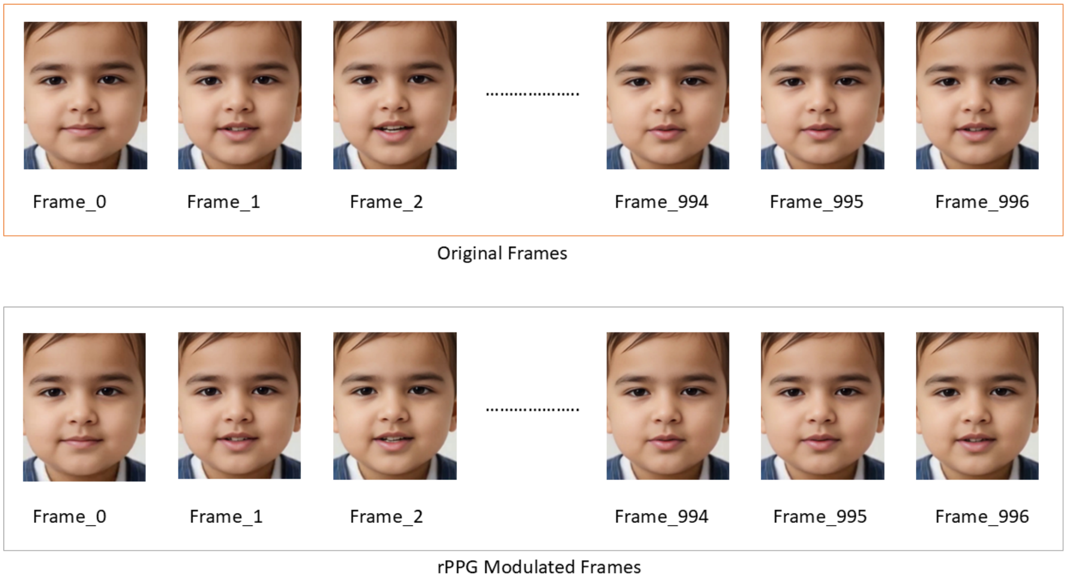

Figure 4.1: Sample comparison between original (top row) and rPPG-modulated (bottom row) face frames.

playback will be conducted to further validate motion continuity and perceptual realism. The modulation preserves facial texture, lighting, and overall realism while integrating the physiological signal. Overall, these results confirm that the proposed modulation pipeline effectively embeds physiological signals while maintaining the visual integrity of the face. This balance is essential for downstream applications such as rPPG signal extraction and physiological monitoring from synthetic videos.

4.1.3 Quantitative Evaluation

Synthetic Video Evaluation

In this section, we evaluate the effectiveness of our synthetic video generation pipeline by comparing the HR signals incorporated into the facial frames with those extracted using the rPPG extraction method. This comparison quantifies the fidelity of the embedded physiological dynamics and the accuracy of signal recovery.

Serial No.	**Incorporated HR (BPM)**	**Frequency (Hz)**	**Extracted HR (BPM)**	**Absolute Error HR (BPM)**
Subject 1	54.00	0.90	54.16	0.16
Subject 2	60.00	1.00	60.18	0.18
Subject 3	72.00	1.20	72.22	0.22
Subject 4	90.00	1.50	90.27	0.27
Subject 5	120.00	2.00	120.36	0.36
Subject 6	150.00	2.50	150.45	0.45
Subject 7	168.00	2.80	168.51	0.51
Subject 8	180.00	3.00	180.54	0.54
Subject 9	210.00	3.50	210.63	0.63
Subject 10	240.00	4.00	239.22	0.78

Table 4.1: Comparison of Incorporated and Extracted HRs with Absolute Error

The table above presents the quantitative evaluation of our method, in which a predefined HR signal was integrated into synthetic face videos and subsequently extracted using our signal processing pipeline. The precision of this extraction is assessed using the Absolute Error (AE), which measures the deviation between the incorporated and extracted HRs. The consistently low AE values across all test cases highlight the effectiveness of our method in accurately recovering the embedded physiological signals. These results validate the reliability of our rPPG extraction pipeline in controlled synthetic settings.

The Mean Absolute Error (MAE) between the incorporated and extracted HRs across all synthetic videos was found to be remarkably low at **0.41 BPM**, demonstrating the high precision and reliability of our rPPG extraction pipeline. Additionally, the **Mean Squared Error (MSE)** was computed to be **0.207 BPM^2**, further confirming the minimal deviation between the intended and recovered HR values. Most notably, the **Pearson Correlation Coefficient (PCC)** between the incorporated and extracted HRs was calculated to be **0.99998**, indicating an almost perfect linear agreement.

These metrics collectively highlight the strong fidelity, robustness, and realism of the synthetic physiological dynamics embedded within the generated face videos, validating the effectiveness of our method in accurately simulating and recovering vital signs from visual data.

4.1.4 Result Analysis on Synthetic Data

To evaluate the effectiveness of the proposed rPPG signal extraction pipeline, experiments were conducted exclusively on synthetically generated face videos embedded with controlled physiological dynamics. In these videos, a clean sinusoidal rPPG signal was intentionally incorporated during the modulation process to serve as a known ground truth for validation.

The results demonstrate a high degree of fidelity between the incorporated and extracted HR signals. The Mean Absolute Error (MAE) between the incorporated and extracted HRs was found to be exceptionally low, indicating minimal deviation. Additionally, the Mean Squared Error (MSE) remained negligible, and the Pearson Correlation Coefficient (PCC) approached unity, reflecting a near-perfect linear correlation between the signals.

These results validate two essential aspects: (i) the efficiency and robustness of the rPPG extraction method under ideal, controlled conditions, and (ii) the effectiveness of the synthetic video generation pipeline in integrating realistic, retrievable physiological dynamics without compromising visual quality.

Such high accuracy in a controlled synthetic environment provides a reliable testbed for benchmarking rPPG algorithms and facilitates further development and validation before deployment in real-world scenarios.

CHAPTER 5

Conclusion and Future Work

5.1 Conclusion and Future Work

5.1.1 Conclusion

In this thesis, we presented a novel method for incorporating physiological dynamics specifically, remote photoplethysmography (rPPG) signals into synthetically generated face videos. Starting from a static face image and an audio signal, we generated realistic talking face videos and subsequently modulated specific facial regions (forehead and cheeks) using a predefined rPPG waveform. Careful landmark based masking ensured that sensitive regions such as the eyes, lips, and background remained unaffected, preserving visual realism.

Quantitative analysis demonstrated the high fidelity of the embedded signals. Our rPPG extraction pipeline was able to retrieve the incorporated HR signals with a **Mean Absolute Error (MAE)** of **0.41 BPM**, **Mean Squared Error (MSE)** of **0.207 BPM^2**, and **Pearson Correlation Coefficient (PCC)** of **0.99998**, indicating almost perfect agreement between the embedded and recovered signals. These results confirm the effectiveness of our modulation strategy and the robustness of our extraction pipeline under controlled conditions.

This work serves as a step forward in bridging the domains of facial video synthesis and physiological signal modeling, enabling the creation of synthetic datasets with controllable and ground-truth-aligned physiological dynamics.

5.1.2 Future Work

Although the proposed method demonstrates strong potential, several enhancements can be pursued to increase its utility and realism:

- **Physiological Signal Diversity:** Future extensions could explore integrating more complex and realistic physiological patterns, such as non-sinusoidal waveforms, signal variability, and natural noise better approximating real cardiovascular behavior and improving generalization to real-world conditions.

- **Dynamic Environmental Conditions:** The current setup assumes a controlled environment. Incorporating dynamic lighting, background variation, and head motion into the synthetic videos could provide a more rigorous testbed for evaluating rPPG extraction algorithms under realistic challenges.

- **Extension to Multi-modal Physiological Embedding:** Beyond HR, this method can be extended to include other physiological signals such as respiratory rate or facial thermal patterns, enabling richer synthetic datasets for multi-modal health monitoring and signal fusion research.

- **Integration with Learning-based rPPG Models:** The synthetic videos generated through our pipeline can be used to train or pre-train deep learning models for rPPG signal estimation, offering a data-rich, noise-controlled environment for improving model robustness in low-data real-world settings.

Bibliography

Aggarwal, Alankrita, Mamta Mittal, and Gopi Battineni (2021). "Generative adversarial network: An overview of theory and applications". In: *International Journal of Information Management Data Insights* 1.1, p. 100004.

Challa, Nagendra Panini, ES Phalguna Krishna, S Sreenivasa Chakravarthi, et al. (2023). "Facial Landmarks Detection System with OpenCV Mediapipe and Python using Optical Flow (Active) Approach". In: *2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*. IEEE, pp. 92–96.

Davies, Brian and Tim Morris (1993). "Physiological parameters in laboratory animals and humans". In: *Pharmaceutical research* 10.7, pp. 1093–1095.

Goodfellow, Ian et al. (2014). "Generative adversarial nets". In: *Advances in Neural Information Processing Systems*. Vol. 27.

Gupta, Anup Kumar et al. (2023). "RADIANT: Better rPPG estimation using signal embeddings and transformer". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4976–4986.

Gupta, Puneet, Brojeshwar Bhowmick, and Arpan Pal (2017). “Serial fusion of Eulerian and Lagrangian approaches for accurate heart-rate estimation using face videos”. In: *2017 39th annual international conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, pp. 2834–2837.

— (2020). “MOMBAT: Heart rate monitoring from face video using pulse modeling and Bayesian tracking”. In: *Computers in Biology and Medicine* 121, p. 103813.

Ho, Jonathan, Ajay Jain, and Pieter Abbeel (2020). “Denoising diffusion probabilistic models”. In: *Advances in Neural Information Processing Systems.* Vol. 33, pp. 6840–6851.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). “Deep learning”. In: *nature* 521.7553, pp. 436–444.

Lee, Ru Jing, Saaveethya Sivakumar, and King Hann Lim (2024). “Review on remote heart rate measurements using photoplethysmography”. In: *Multimedia Tools and Applications* 83.15, pp. 44699–44728.

Palazzo, Simone et al. (2018). “Generating synthetic video sequences by explicitly modeling object motion”. In: *Proceedings of the European Conference on Computer Vision Workshops*, pp. 1–7.

Sahin, Sarker Md et al. (2021). “Non-contact heart rate monitoring from face video utilizing color intensity”. In: *Journal of Multimedia Information System* 8.1, pp. 1–10.

Saikia, Trishna et al. (2023). “HREADAI: Heart rate estimation from face mask videos by consolidating Eulerian and Lagrangian approaches”. In: *IEEE Transactions on Instrumentation and Measurement* 73, pp. 1–11.

Salim, Ali S and Abdul Sattar M Khidhir (2024). “A Comprehensive Review of rPPG Methods for Heart Rate Estimation”. In: *Open Access Library Journal* 11.11, pp. 1–18.

Tohma, Akito et al. (2021). “Evaluation of remote photoplethysmography measurement conditions toward telemedicine applications”. In: *Sensors* 21.24, p. 8357.

Weatherall, Ian L and Bernard D Coombs (1992). “Skin color measurements in terms of CIELAB color space values”. In: *Journal of Investigative Dermatology* 99.4, pp. 468–473.

Xiao, Hanguang et al. (2024). “Remote photoplethysmography for heart rate measurement: A review”. In: *Biomedical Signal Processing and Control* 88, p. 105608.

Zhang, Wenxuan et al. (2023). “SadTalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8652–8661.