EEG2Face: EEG-driven Emotional 3D Face Reconstruction

by

Zhuldyz Kabidenova

Submitted to the Department of Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Data Science

at the

NAZARBAYEV UNIVERSITY

April 2025

Author .....
Department of Computer Science
30.04.2025

Certified by .....
Minho Lee
Assistant Professor
Thesis Supervisor

Certified by .....
Adnan Yazici
Associate Professor
Thesis Supervisor

Accepted by .....
Yelyzaveta Arkhangelsky
Dean, School of Engineering and Digital Sciences

EEG2Face: EEG-driven Emotional 3D Face Reconstruction

by

Zhuldyz Kabidenova

Submitted to the Department of Computer Science
on 30.04.2025, in partial fulfillment of the
requirements for the degree of
Master of Science in Data Science

Abstract

The increasing use of 3D facial avatars in digital communication highlights the critical challenge of accurately capturing and replicating genuine emotional expressions. While most existing methods rely on visual data to recreate facial dynamics, the potential to decode these dynamics directly from brain activity remains largely unexplored. In this work a novel machine-learning framework that reconstructs 3D facial expressions using EEG signals alone is proposed. By leveraging synchronized 3D pseudo-ground-truth extracted from the EAV dataset as supervision, our model decodes EEG signals into dynamic 3D face meshes, faithfully replicating the corresponding facial expressions. This approach bridges deep learning and neuroscience, presenting a first-of-its-kind system for neural signal-to-3D reconstruction. Our findings establish a robust baseline for EEG-driven facial expression synthesis, with broad implications for generative modeling, representation learning, and brain-computer interface technologies. The model and code are publicly available at *https://github.com/zizimars/EEG2Face*

Thesis Supervisor: Minho Lee
Title: Assistant Professor

Thesis Supervisor: Adnan Yazici
Title: Associate Professor

Acknowledgments

First and foremost, I would like to acknowledge myself for the dedication and perseverance invested in this thesis over the past two years. Through countless days and nights of work, I have grown tremendously as a researcher and scholar. The journey has been challenging but immensely rewarding, and I stand proud of what I have accomplished.

I express my deepest gratitude to my supervisor, Dr. Minho Lee, whose exceptional guidance was instrumental in developing the innovative concept of generating 3D faces from EEG signals. His profound expertise in Generative AI and neuroscience provided the foundation upon which this research was built. His mentorship has been invaluable throughout this academic journey.

My heartfelt thanks go to my family for their unwavering emotional support. Their encouragement gave me strength during the most challenging moments of this endeavor.

I am profoundly grateful to my colleague and boyfriend, Kirill, whose contribution to this work cannot be overstated. His technical assistance was crucial whenever I encountered obstacles, and his moral support was a constant source of strength. Our numerous discussions and brainstorming sessions helped clarify my vision whenever doubts arose. This thesis would not have been possible without his constant support.

I also wish to express my appreciation to my co-supervisor, Dr. Adnan Yazici, and the member of my thesis committee, Dr. Berdakh Abidullaev, for their thoughtful review, constructive feedback, and active engagement with my research. Their insights and recommendations significantly enhanced the quality and depth of this work.

Finally, I extend my gratitude to everyone who contributed directly or indirectly to the completion of this thesis. This achievement stands as a testament to the collective support I have been fortunate to receive.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Realistic facial animation represents a fundamental challenge in computer graphics and human-computer interaction, with applications spanning entertainment, communication, and assistive technologies. Animating 3D facial avatars from various input sources, such as vision, audio, or text, holds significant promise across multiple fields, including character animation in films and games, virtual telepresence for augmented and virtual reality (AR and VR), and the creation of digital personal assistants with human-like expressiveness [2, 44].

A major challenge in this domain is effectively balancing dynamic facial expressions with the precise movements required for accurate synchronization with input sources while ensuring natural emotional realism. This requires a robust disentanglement of key facial attributes, including pose, expression, identity, illumination, and lip articulation, which are critical for generating high-fidelity and expressive 3D avatars.

In the realm of 3D facial animation, there are two primary streams for avatar generation: face-based generation [28, 17, 14, 9] and speech-based generation [10, 35, 41]. Face-based generation techniques rely on visual cues extracted from facial expressions, capturing a range of fine-grained attributes essential for realism. These include facial geometry and structure, which define the unique shape and proportions of the face, pose and head orientation. Speech-

based generation focuses on audio cues, using spoken content to drive facial movements and expressions. One of its key strengths is its ability to ensure precise lip-speech synchronization, aligning mouth movements with phonetic cues to enhance intelligibility.

Meanwhile, accurate emotional expressions are essential for conveying subtle affective states, making interactions more natural and immersive [18, 9, 10, 35, 41]. However, 3D face models still struggle to capture and reproduce fine-grained emotions due to the limited availability of audio-visual datasets with emotion annotations. To address this, recent studies have attempted to disentangle content and emotion attributes, leading to a more structured feature space, improved interpolation controllability, and enhanced model stability.

Despite advances in facial animation technologies, current approaches face a fundamental limitation: they rely on external expressions that may not accurately reflect a person's internal emotional state. This disconnect creates an authenticity gap, particularly problematic for applications requiring genuine emotional representation or for individuals with limited facial mobility. Traditional methods cannot capture unexpressed or suppressed emotions, nor can they generate authentic expressions for individuals who have difficulty expressing emotions physically. This critical gap in existing approaches highlights the need for a novel method that can directly access and represent internal emotional states.

Electroencephalography (EEG) data offers a promising alternative input modality that can address these limitations. EEG represents a fundamentally different approach from audio-visual inputs, as its signal patterns directly capture neural activity associated with emotional states. One significant advantage of EEG data is its ability to directly measure brain activity related to emotions, providing a more authentic reflection of an individual's emotional state [24]. This approach has immense potential: by using EEG to drive 3D facial expressions, it is possible to create avatars where the displayed emotions are consistent with the actual feelings of the individual. This results in a "true-to-life" representation, where the emotion behind the expression matches the visual output.

The applications of such EEG-driven facial animation technology are diverse and far-reaching. In the medical field, this technology could serve as an assistive communication tool for individuals with facial paralysis, ALS, or other conditions that limit physical expressiveness, enabling more natural emotional communication through personalized avatars. For mental health applications, it could provide visualization tools for therapy sessions, helping patients better understand and express their emotions. In the entertainment industry, game developers could create characters that respond to players' actual emotional states, creating deeply personalized gaming experiences. Virtual reality environments could adapt to users' emotional responses in real-time, enhancing immersion and presence. Additionally, this technology could enable new forms of artistic expression, allowing creators to translate emotional states directly into visual representations without the limitations of physical expression.

When analyzing EEG data, traditional approaches typically focus on extracting neural signals while treating other components as unwanted noise to be filtered out. These conventional methods have shown limited success in connecting EEG signals to facial expressions, as they focus primarily on frequency-domain information that lacks direct correlation with visual facial representations [3, 22, 4].

This challenge is approached with a novel hypothesis: the components of EEG signals typically discarded as "noise" in standard Brain-Computer Interface (BCI) pipelines may actually contain valuable information about facial muscle activities [21]. Unlike traditional approaches, it is proposed that these overlooked signal components are more directly correlated with facial movements and expressions. Rather than filtering out these signals, it is suggested they can be leveraged to create a more direct bridge between brain activity and facial animation. The approach uses vision-guided learning to identify and extract emotion-related information from these previously ignored components of EEG data, enabling a more effective translation from brain signals to expressive facial animations.

The main contributions of this study are summarized as follows:

1. The first framework capable of direct 3D face reconstruction from EEG signals alone is developed, a significant advancement as no previous system has successfully synthesized facial geometry and appearance directly from brain activity.

2. A novel disentanglement mechanism that separates content and emotion attributes from EEG signals is introduced: the content EEG regressor captures externally-expressed emotional features via vision-guided contrastive learning, while the emotion EEG decoder encodes intrinsic emotional states into a latent representation. These are integrated into a unified 3D facial synthesis pipeline that reflects both internal and external emotional cues.

3. A benchmark EEG2Face dataset is presented by reorganizing the EAV dataset [27], specifically designed to facilitate research on EEG-driven 3D facial generation.

The 3D face generation was evaluated using both qualitative and quantitative methods. The model achieved performance comparable to vision-based approaches in representing content attributes, and controllability of facial expressions through manipulation of the EEG emotion latent code.

Chapter 2

Related works

This section presents a comprehensive review of existing face reconstruction methodologies, organized by input modalities including image/video, audio, and EEG. The analysis further examines available multimodal datasets utilized for facial reconstruction research, with particular emphasis on the distinctive capabilities of the EAV dataset [27] in facilitating novel EEG-driven face reconstruction paradigms.

2.0.1 Input modality in Face Reconstruction

Face Reconstruction from Images

The reconstruction of 3D facial geometry from 2D images has been a well-researched problem in computer vision, with its methods having evolved significantly over time. The primary focus has been on improving the accuracy of 3D facial representation while addressing challenges such as pose, shape, and expression variations. Early methods, such as 3DMM-CNN [42], introduced a deep learning framework to predict 3DMM parameters directly from images, combining the interpretability of model-based approaches with the efficiency of learning-based techniques. Building upon this, PRNet [15] predicts dense correspondence maps in UV space to regress dense 3D coordinates of facial surfaces. However,

these approaches often faced limitations in accurately capturing high-frequency facial details. Subsequent work, expressive parametric model FLAME [28] incorporates both shape and expression parameters making it suitable for facial performance capture and animation. RingNet [39] extends this concept by utilizing a neural network to predict FLAME parameters directly from images, achieving accurate and expressive reconstructions. Emotion-aware face reconstruction has gained traction for applications in affective computing and virtual avatar creation [14, 9, 47]. DECA [14] integrates detail maps into parametric models to enhance the fidelity of facial expressions. EMOCA [9] extends DECA by incorporating emotion recognition capabilities, enabling reconstructions that are both geometrically accurate and emotionally expressive. TokenFace [46] presents recent advancements, leveraging transformer architectures to achieve state-of-the-art results.

Audio-driven Talking Face Generation

Audio-driven 3D face reconstruction models represent a complementary approach, focusing on synchronizing facial animations with speech or vocal input. In this domain, significant contributions were made by [8], [37], [12], [45] and [10]. VOCA [8] is trained on a 4D face dataset and capable of animating previously unseen subjects across different languages without the need for retargeting. FaceFormer [12] employs an autoregressive transformer-based architecture to encode long-term audio context and the history of facial motions to predict sequences of animated 3D face meshes, thereby improving lip synchronization. MeshTalk [37] disentangles audio-correlated and uncorrelated facial movements using a categorical latent space, enabling realistic motion synthesis for the entire face. CodeTalker [45] frames animation as a code query task within a learned discrete motion codebook, reducing cross-modal mapping uncertainty and achieving vivid and realistic facial motions. Recognizing the importance of social expressions, LaughTalk [41] is a model designed to generate authentic laughter and smiles in 3D talking heads, enhance the social context and engagement of virtual avatars. Furthermore, EMOTE [10] generates 3D talking avatar models that preserve precise lip synchronization with speech input while simultaneously offering explicit mechanisms

for emotional expression modulation. These developments collectively contribute to more realistic and expressive speech-driven 3D facial animations, broadening the applicability of virtual avatars in various domains.

Face Reconstruction from EEG

The use of EEG signals for generating images has garnered significant attention in recent years, particularly in the context of reconstructing visual stimuli from neural activity. NeuroImagen [23] exemplifies this progress by introducing a framework that extracts multi-level semantics from EEG signals. These semantics include pixel-level information, such as saliency maps capturing color, position, and shape, as well as sample-level semantics aligned with image captions, which enhance the semantic accuracy of reconstructed images. Integrating these features into a latent diffusion model, NeuroImagen effectively reconstructs visual stimuli from noisy EEG data. Other studies have demonstrated the potential of EEG for tasks like facial identity discrimination. Nemrodov et al. [33] revealed that EEG data could robustly encode facial features, showing discrimination peaks at specific ERP components (e.g., N170 and N250) and highlighting the stability of neural-based face spaces. These findings underscore the suitability of EEG for capturing visual and structural attributes of faces. In addition to image reconstruction, EEG has been utilized for generating emotional expressions. For example, GAN-based approaches like [11] have been used to synthesize personalized emotional expressions conditioned on EEG signals, showcasing EEG's ability to encode emotional information.

However, despite the advancements in reconstructing visual stimuli and generating emotional expressions, there is no existing study that addresses the reconstruction of 3D facial geometry directly from EEG signals. The EEG2Face system represents a novel step in this direction, combining EEG data with the multimodal EAV dataset to bridge neural signals and dynamic 3D facial reconstruction. This work not only fills a significant gap in the field but also establishes a foundation for future research in EEG-driven 3D face generation.

2.0.2 Bench-mark Emotional EEG-Vision Dataset

The availability of high-quality datasets has played a critical role in advancing face reconstruction models across different modalities. In the case of audio-driven face reconstruction, datasets such as VOCASET [8], BIWI [13], MEAD [43], and others [35, 37], have become the cornerstone for research in this domain. These datasets typically include synchronized audio, video, and 3D facial scans of subjects speaking, allowing for the exploration of speech-driven facial animation and emotion-aware reconstructions. However, none of the above-mentioned datasets include EEG signals as part of their data modalities. Existing multimodal datasets that do incorporate EEG often suffer from limitations such as sparse electrode coverage [34], reliance on passive or static tasks [20, 19], or segmentation from unrelated data sources [29, 36]. This lack of robust, synchronized EEG and visual data has hindered the exploration of EEG-to-face reconstruction tasks.

The EAV dataset addresses these challenges as the first multimodal dataset specifically designed to collect synchronized EEG, video, and audio data in an interactive, conversational context. Unlike previous datasets, the EAV dataset have the following features: (1) the active conversational framework, enabling expressive data capture, (2) high-quality, competitive signal strength across all three modalities, and (3) balanced class annotations across five emotional categories, facilitated by a controlled, cue-based interaction design.

Chapter 3

Methodology

This section describes the datasets and models utilized for EEG and vision processing. The pretrained vision model, built on FLAME [28] and EMOCA [9] frameworks, generates 3D facial reconstructions and extracts emotional representations from image data. This model acts as a guiding teacher to supervise the training of our EEG-based model (Fig. 3-1). The EEG-based model consists of two key components: an EEG regressor ($\mathbf{E}_r$) that captures external facial representations, and an EEG emotion encoder ($\mathbf{E}_e$) that infers intrinsic emotional states. Additionally, the architecture incorporates emotion latent code ($\bar{\mathbf{z}}$) that enable precise manipulation of emotional expressions while maintaining consistent facial identity.

3.0.1 Data Configuration

This study utilizes the EAV dataset [27], a multimodal resource combining EEG and video recordings in cue-based conversational tasks. The dataset consists of listening-speaking iterations, each lasting 20 seconds, and provides a rich context for studying emotional expressions. The original EAV dataset was collected from 42 subjects [27]; however, this study focuses on a subset of 20 subjects. These individuals were selected based on balanced emotion recognition performance across EEG, audio, and visual modalities, and provided consent for the

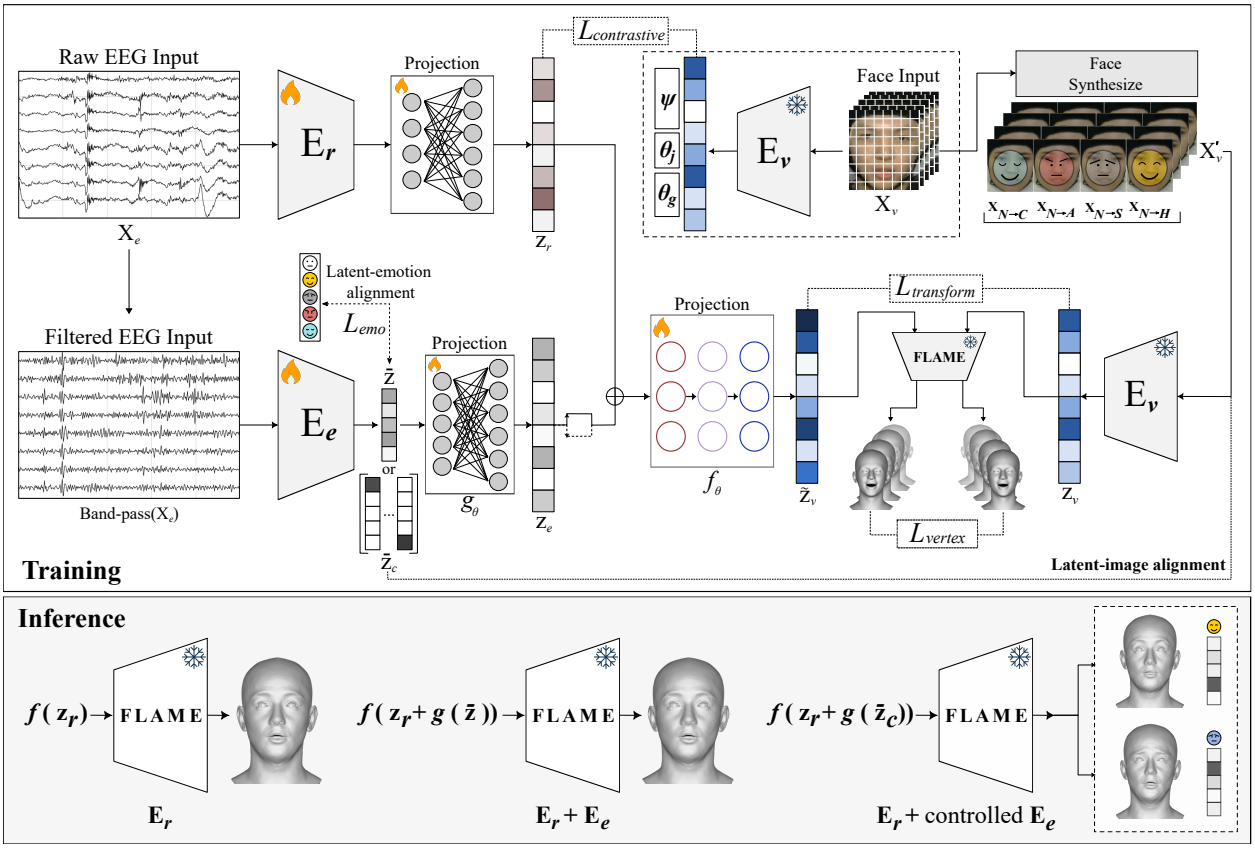

Figure 3-1: **Dual-pathway architecture for disentangled EEG representation learning.** The model consists of two parallel neural network streams that process 30-channel EEG signals. The left pathway (E_r) serves as the EEG regressor, applying spatial and temporal filtering followed by Conv2D layers. The right pathway (E_e) performs emotion recognition based on spectral filters. These two branches enable the extraction of disentangled content and emotion representations, each culminating in a separate fully connected layer.

use of their facial data in this research.

The original repository provided preprocessed EEG data segmented into 5-second intervals; however, this study utilizes the raw EEG data alongside the corresponding visual data with the original 20-second segments. From these segments, 100 speaking trials were extracted, each comprising 20-second emotional conversations categorized into five emotions: Neutral, Anger, Calmness, Happiness, and Sadness. The EEG signals were recorded using 30 electrodes and are represented as a tensor $\mathbf{X}_e \in \mathbb{R}^{N \times ch \times T}$, where N is the number of segments, $ch = 30$ is the number of EEG channels, and T denotes the number of temporal data points per segment, sampled at 100 Hz.

The videos were recorded at 30 FPS. Face regions were cropped following the DECA [14] procedure, with Mediapipe [32] used for facial landmark detection instead of FAN [5] due to its superior precision as demonstrated in [10] research. The processed face images were resized to 224×224 ($H \times W$). Each frame is represented as a tensor $\mathbf{X}_v \in \mathbb{R}^{N \times 3 \times H \times W}$, where N is the number of frames temporally aligned with EEG segments.

To ensure precise temporal alignment between EEG signals and facial video frames, the video modality was downsampled from 30 FPS to 10 FPS by sampling one frame every 100 ms. This matches the EEG sampling rate of 100 Hz, such that each face frame corresponds to 10 EEG time points.

Let X_e denote the EEG dataset and X_v denote the vision dataset. The datasets are defined as:

$$\mathbf{X}_e = \{x_1, x_2, \ldots, x_N\}, \quad x_i \in \mathbb{R}^{N \times ch \times t},$$

where $t = 10$ denotes the number of EEG time steps per image frame and $ch = 30$ are the number of channels,

$$\mathbf{X}_v = \{x_1, x_2, \ldots, x_N\}, \quad x_i \in \mathbb{R}^{N \times C \times H \times W}.$$

By construction, each x_i in X_v corresponds to the matching trial x_i in X_e, ensuring a consistent pairing of EEG and vision data for every index $i \in \{1, 2, \ldots, N\}$. In this study, the dataset was split into training and testing sets using a 6:2 ratio, following a subject-dependent scheme [27].

3.0.2 Preliminary in Vision-based reconstruction

To encode an input image into a latent representation, pretrained ResNet50-based encoder $\mathbf{E}_v$ [14] was employed to extract facial features. The extracted features are then passed through a fully connected layer to specifically regress the facial expression $\boldsymbol{\psi}_{\text{exp}} \in \mathbb{R}^{50}$, jaw pose $\boldsymbol{\theta}_{\text{jaw}} \in \mathbb{R}^3$ parameters of the FLAME model, while the shape $\boldsymbol{\beta} \in \mathbb{R}^{100}$ and global pose $\boldsymbol{\theta}_{\text{globalpose}} \in \mathbb{R}^3$ parameters are not used in this study.

This representation in this study is defined as:

$$\mathbf{z}_v = \mathbf{E}_v(\mathbf{X}_v), \quad \mathbf{z}_v = \{\boldsymbol{\psi}_{\text{exp}},\ \boldsymbol{\theta}_{\text{jawpose}}\} \in \mathbb{R}^{53}.$$

In this study, since global pose cannot be reliably predicted from EEG signals, it is fixed to zero, corresponding to a frontal view.

FLAME ($\mathrm{F}(\cdot)$) is a statistical 3D head model that employs standard vertex-based linear blend skinning (LBS) with corrective blendshapes. It generates a 3D mesh with vertices $\mathbf{v}$ by applying learned blendweights to deform a template mesh in the zero pose. This is defined as:

$$\mathbf{v} = \mathrm{F}(\mathbf{z}_v), \quad \mathbf{v} \in \mathbb{R}^{5023\times 3}.$$

The resulting 3D mesh is then rendered into a realistic facial image using a renderer $\mathrm{R}(\cdot)$, producing the output $\mathbf{I}_{3D}$.

3.0.3 EEG Modeling

The acquired raw EEG signal $\mathbf{X}_{\mathrm{e}} \in \mathbb{R}^{ch \times T}$ is modeled as the superposition of latent neural source activity $\mathbf{s}(t) \in \mathbb{R}^d$ transformed through a volume conduction operator $\mathcal{V} : \mathbb{R}^d \to \mathbb{R}^C$, and additive noise $\mathbf{n}(t) \in \mathbb{R}^C$ comprising sensor noise and physiological artifacts such as EOG and EMG. This relationship is formally expressed as:

$$\begin{aligned} \mathbf{X}_{\mathrm{e}}(t) &= \mathcal{V}(\mathbf{s}(t)) + \mathbf{n}(t) \\ \mathbf{n}(t) &= \mathrm{EOG}(t) + \mathrm{EMG}(t) + \mathrm{white}(t) \end{aligned}$$

where the operator $\mathcal{V}(\cdot)$ encapsulates the volume conduction from neural sources to scalp electrodes, and the noise term $\mathbf{n}(t)$ accounts for non-stationary components, including non-periodic physiological activities (e.g., EMG, EOG) and background white noise (e.g., heartbeat and sensor drift).

In typical HCI research, the goal is to extract discriminative patterns from the observed EEG signal $\mathrm{EEG}(t)$ and learn a function that maps these features to task-relevant labels. This process can be formally expressed as:

$$\hat{y} = f(\mathbf{X};\, \pi_f,\, \pi_s,\, \pi_t),$$

where $f(\cdot)$ is a task-specific decoder parameterized by spectral (π_f), spatial (π_s), and temporal (π_t) components. EMG and EOG studies may primarily focus on temporal dynamics (π_t), while EEG studies tend to focus more on frequency-domain representations (π_f).

In CNNs or transformer architectures, convolution operations in the initial layers play a crucial role: $1 \times t$ convolutions are typically used to learn spectral (temporal frequency) filters[25], while $ch \times t$ convolutions capture spatial patterns across EEG channels via channel-wise regression [3].

Two transformer-based modules are introduced in the next sections: a facial regressor E_r

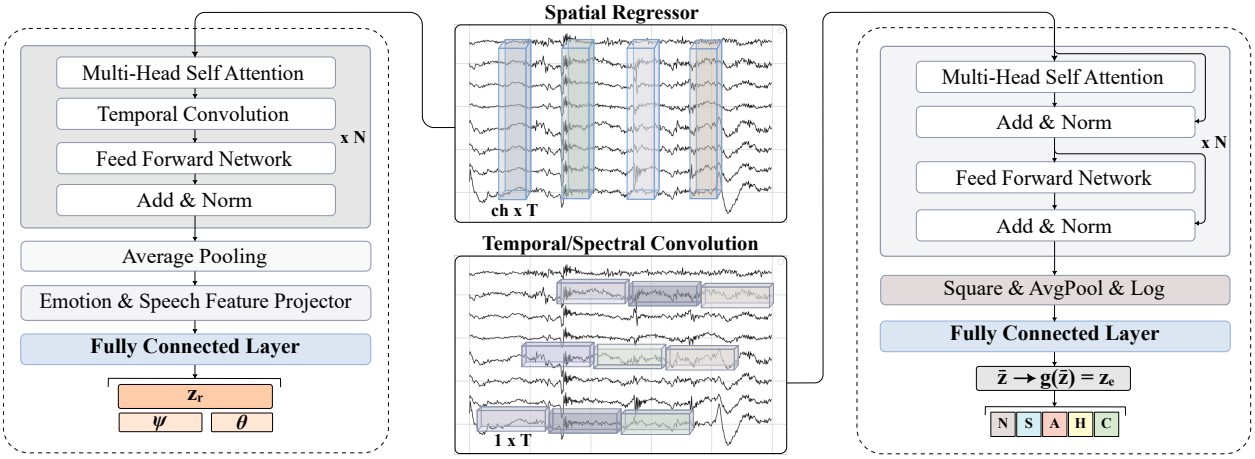

Figure 3-2: **Dual-pathway architecture for disentangled EEG representation learning.** Two parallel neural network streams processing 30-channel EEG signals begin with TCNs, with the left pathway using 16 filters followed by Conv2D layers, while the right pathway employs TCN with 40 filters followed by Patch Embedding and Multi-Head attention mechanisms. These architectures enable extraction of distinct representational features, culminating in separate fully-connected layers that encode content and emotional information.

for modeling outward facial expressions, and an emotion encoder E_e for capturing intrinsic emotional states from EEG.

3.0.4 EEG Content Regressor

A compact transformer-based regression model (see Figure 3-2 (left)), denoted as E_r, is proposed to generate temporally aligned regression outputs that correspond to the vision-based content representation $\mathrm{E}_v(\mathbf{X}_v)$. Given an input EEG signal $\mathbf{X}_e \in \mathbb{R}^{N\times ch\times T}$, a 2D convolution with kernel size $ch \times t_s$ is first applied to extract localized spatial-temporal features:

$$\mathbf{Z}_0 = \mathrm{Conv2D}(\mathbf{X}_e) \in \mathbb{R}^{N\times d\times 1\times T},$$

where the filter size is $ch \times t_s$, d denotes the number of filters (i.e., the output feature dimension), and *same padding* is applied along the temporal axis to preserve the length T. The patch embedding step is omitted for simplicity, and the output is directly used as the input sequence to the transformer. The resulting sequence is processed by L layers of

transformer blocks:

$$\mathbf{Z}_\ell = \text{TransformerLayer}_\ell(\mathbf{Z}_{\ell-1}), \quad \ell = 1, \ldots, L,$$

producing the final encoded sequence $\mathbf{Z}_L \in \mathbb{R}^{N \times T \times d}$.

To obtain a fixed-length regression feature, attention-based pooling is applied over the temporal axis:

$$\mathbf{z}_r = \sum_{t=1}^{T} \alpha_t \mathbf{Z}_L[:, t, :], \quad \alpha_t = \frac{\exp(\mathbf{q}^\top \mathbf{Z}_L[:, t, :])}{\sum_{k=1}^{T'} \exp(\mathbf{q}^\top \mathbf{Z}_L[:, k, :])},$$

where $\mathbf{q} \in \mathbb{R}^d$ is a learnable query vector shared across batches.

This process yields the final output of the EEG facial regressor:

$$\mathbf{z}_r = \text{E}_r(\mathbf{X}_e) \in \mathbb{R}^{N \times d}.$$

To map the EEG embedding $\mathbf{z}_r$ into the vision domain the transformation function $f_\theta(\cdot)$ is defined as follows:

$$\tilde{\mathbf{z}}_v = f_\theta(\mathbf{z}_r), \quad \tilde{\mathbf{z}}_v \in \mathbb{R}^D.$$

3.0.5 EEG Emotion Encoder

Inspired by EEGNet [25], ShallowNet [40], and typical BFCSP approaches [3], the proposed transformer-based EEG emotion decoder (Fig. 3-2 (right)) integrates their core architectural components.

Given an input EEG signal $\mathbf{X}_e \in \mathbb{R}^{N \times ch \times T}$, a 1D depthwise convolution with f kernels of size $1 \times t_s$ is first applied:

$$\mathbf{X}_f = \text{Conv1D}(\mathbf{X}_e) \in \mathbb{R}^{N \times f \times ch \times T}, \quad \mathbf{X}_f = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_f\},$$

where f denotes the number of convolutional kernels and t_s represents the kernel size along the temporal axis. The *same padding* is employed to preserve the original temporal length T.

For a sequence of EEG feature slices $\mathbf{x}_t \in \mathbb{R}^{ch\times 1}$ over T time steps, a learnable spatial projection $\mathbf{W}_f \in \mathbb{R}^{ch\times d}$ is applied. The projected outputs are concatenated to form a sequence of token embeddings:

$$\mathbf{Z}_0 = [\mathbf{x}_1^\top \mathbf{W}_f;\, \mathbf{x}_2^\top \mathbf{W}_f;\, \ldots;\, \mathbf{x}_T^\top \mathbf{W}_f] \in \mathbb{R}^{N\times T\times d},$$

where d indicates the embedding dimension, and each projected token corresponds to one time step.

The resulting sequence $\mathbf{Z}_0$ is then processed by a transformer encoder composed of L stacked layers of self-attention and feed-forward modules:

$$\mathbf{Z}_\ell = \mathrm{TransformerLayer}_\ell(\mathbf{Z}_{\ell-1}), \quad \ell = 1, \ldots, L,$$

yielding the final encoded representation $\mathbf{Z}_L \in \mathbb{R}^{N\times T'\times d}$.

The attention output is then pooled and transformed using a logarithmic nonlinearity [40]:

$$\bar{\mathbf{z}}_e = \mathrm{Softmax}\left(\mathrm{FC}\left(\log\left(\mathrm{AvgPool}(\mathbf{Z}_L)\right)^2\right)\right),$$

where $\bar{\mathbf{z}}_e$ denotes the latent code aligned with the five emotion classes.

3.0.6 Emotion Disentanglement

To stabilize and enhance the controllability of the latent code $\bar{\mathbf{z}}_e$, the neutral face $\mathbf{x}_{v,\mathrm{neu}}$ is first retargeted to create five emotional facial images using the retargeting model [16]. These retargeted emotional faces preserve the identity and content of the original neutral face while changing only the upper facial expressions to represent different emotions. All retargeted faces correspond to the same EEG trial, ensuring cross-modal consistency, as

shown in Figure 3-1.

Let $\mathbf{x}_v$ represent the vision data, which consists of the neutral face and its retargeted emotional variants. Specifically, the retargeted emotional facial images are denoted as:

$$\mathbf{x}_s = [\mathbf{x}_{N\to A} \,|\, \mathbf{x}_{N\to C} \,|\, \mathbf{x}_{N\to S} \,|\, \mathbf{x}_{N\to H}]$$

where each $\mathbf{x}_{N\to E}$ represents the neutral face transformed into a specific emotional state E (e.g., angry (A), calm (C), sad (S), and happy (H)). Importantly, the retargeted images are paired with the EEG trial that was originally matched with the neutral image.

Each retargeted emotional image is associated with a one-hot encoded vector in the latent space $\mathbf{z}_{t,s}$ that corresponds to its specific emotion (e.g., N A, C, S, or H). In this manner, the EEG decoder can generate controllable emotional expressions by manipulating the components of $\mathbf{z}_{t,s}$, enabling precise control over the intensity and type of emotion.

3.0.7 Loss function

A structured training pipeline is proposed, consisting of three stages: (1) independent training of the EEG facial regressor $\mathbf{E}_r$ and the EEG emotion decoder $\mathbf{E}_e$, (2) training of the regression heads $f_\theta(\cdot)$ and $g_\theta(\cdot)$ to project EEG-derived features into the vision-based latent space for retarget and alignment.

To train the EEG facial regressor $\mathbf{E}_r$, a one-sided contrastive loss is employed to align the EEG representation with the corresponding vision latent code:

$$\mathcal{L}_{\text{Info}} = \arg\min_{\theta} \; -\frac{1}{N} \sum_i \log \frac{\exp\left(\text{sim}(\mathbf{E}_r(\mathbf{x}_e^i;\theta),\ \mathbf{E}_v(\mathbf{x}_v^i))\right)}{\sum_j \exp\left(\text{sim}(\mathbf{E}_r(\mathbf{x}_e^i;\theta),\ \mathbf{E}_v(\mathbf{x}_v^j))\right)}$$

Here, $\text{sim}(a, b)$ denotes the cosine similarity between vectors a and b.

The EEG emotion decoder $\mathbf{E}_e$ is optimized using a standard cross-entropy loss:

$$\mathcal{L}_{\text{emo}} = \text{CE}(\text{Softmax}(\mathbf{E}_e(\mathbf{X}_e;\, \theta)),\ \mathbf{Y}),$$

where $\mathbf{Y}$ denotes the discrete ground-truth emotion labels.

Once the EEG encoders are trained, their parameters are frozen. The regression functions $f_\theta(\cdot)$ and $g_\theta(\cdot)$ are then optimized to transform the content and emotion features into the target latent space of the vision model:

$$\mathcal{L}_{\text{mse}} = \|\mathbf{E}_v(\mathbf{X}'_v) - f_\theta\left(\mathbf{E}_r(\mathbf{X}_e) + g_\theta(\mathbf{E}_e(\mathbf{X}_e))\right)\|_2$$

Geometric consistency in the FLAME output space is measured by defining a vertex-level loss between the predicted and reference representations. Specifically, the FLAME mesh generated from the EEG-based latent prediction is compared with the one produced from the vision-based reference:

$$\mathcal{L}_{\text{vertex}} = \left\|\text{F}\left(f_\theta(\mathbf{z}_r + g_\theta(\bar{\mathbf{z}}))\right) - \text{F}(\mathbf{z}_v)\right\|_2,$$

where $\mathbf{z}_r = \mathbf{E}_r(\mathbf{X}_e)$, $\bar{\mathbf{z}} = \mathbf{E}_e(\mathbf{X}_e)$, and $\mathbf{z}_v = \mathbf{E}_v(\mathbf{X}_v)$.

The total loss is a weighted combination of these terms:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{mse}} + \lambda_2 \mathcal{L}_{\text{vertex}} \tag{3.1}$$

During training, the emotion latent vector $\bar{\mathbf{z}}$ is randomly dropped in 50% of the samples. In such cases, only the content vector $\mathbf{z}_r$ is used for supervision, enabling the model to operate under different inference modes.

3.0.8 Implementation

For the EEG Content Regressor, each convolutional layer uses a kernel size of 1×30, a stride of 1×5, and 16 filters. In EEG Emotion Decoder, the spectral convolution uses a kernel size of 1×30 with a stride of 1×5, while the spatial convolution has a kernel of 30×1 with 16 filters per layer. The transformer module consists of $L = 2$ layers with $h = 8$ attention heads, a hidden dimension of $d = 128$, and a feed-forward network (FFN) expansion factor of 4. The model is trained for 400 epochs with a batch size of 64 using the Adam optimizer with a learning rate of 0.001. The EEG Emotion Decoder is then frozen, and the entire loss is trained with $\lambda_1 = 0.7$, $\lambda_2 = 0.3$, a batch size of 64, and 300 epochs.

3.0.9 Inference

During inference, our model provides three distinct types of facial reconstructions based on the available latent components: (1) *Content-only*, where the 3D face is generated as FLAME($f(\mathbf{z}_r)$) using only the EEG regressor output; (2) *Emotion-enhanced*, where both content and internal emotion are combined as FLAME($f(\mathbf{z}_r + g(\bar{\mathbf{z}}))$); and (3) *Emotion modulation*, where a manually controlled emotion vector is applied, resulting in FLAME($f(\mathbf{z}_r + g(\bar{\mathbf{z}}_c))$). In addition, more nuanced or blended emotional expressions can be achieved by interpolating the latent vectors, such as FLAME($f(\mathbf{z}_r + g(\text{Softmax}(\bar{\mathbf{z}} + \bar{\mathbf{z}}_c)))$).

Chapter 4

Results

Vision- or speech-based 3D face reconstruction requires multimodal audio-video data, accompanied by frame-wise mesh vertex ground truth labels [8, 13, 43, 35, 37]. Quantitative evaluation encompasses various key metrics, including reconstruction error [14], vertex displacement [41], and the realism of facial synthesis [18], particularly in regions such as the lips and eyes [12], as well as emotion representation [30, 10, 6]. As the first study on EEG2Face reconstruction, comprehensive quantitative metrics and qualitative analysis are proposed.

4.0.1 Qualitative evaluation

To evaluate the qualitative performance of our EEG2Face model, the reconstructed 3D facial expressions are compared with the ground truth video frames from the EAV dataset, as well as with its pseudo-ground truth 3D supervision. Figure 4-1 presents ten key facial expressions selected from the emotion classes: neutral, angry, calm, happy, and sad. Since EEG signals correlate weakly with head tilts and spatial positioning, the global posture was fixed for both our model's reconstructions and the pseudo-ground truth representations. The results demonstrate that our model effectively captures a wide range of expressions, including closed and open-mouthed smiles, sadness reflected in raised outer eyebrows and lowered eyelids, and

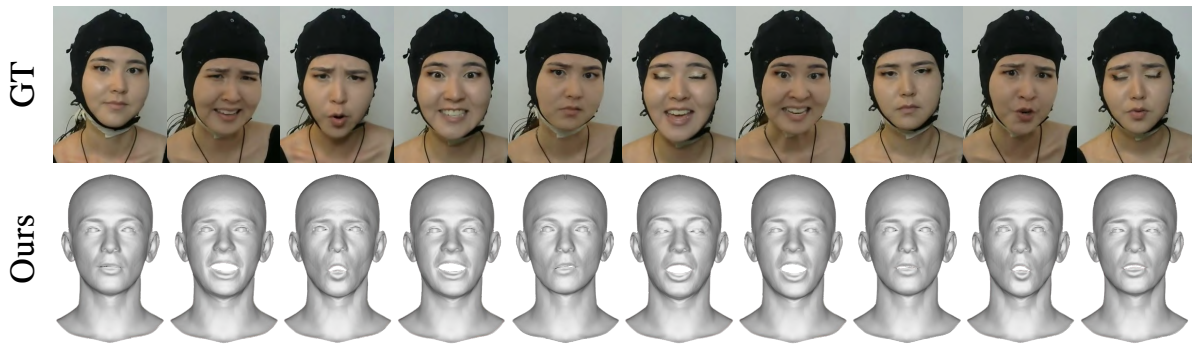

Figure 4-1: Comparison of facial expression reconstruction from EEG signals. The figure presents a sequence of facial expressions across three rows: original video frames (GT), our EEG2Face model's reconstructed expressions (Coarse), and detailed version of the second row (Detail).

a surprised expression characterized by strongly raised eyebrows. Additionally, the model accurately identifies mouth opening and closure as well as forward-stretched lips, highlighting its ability to translate EEG signals into meaningful facial deformations.

The qualitative evaluation demonstrates our model's ability to control emotional expression while preserving facial identity. To investigate the controllability of the emotion latent space, interpolation is performed between the class-specific dimensions of anger and happiness in the predicted emotion vector y_{emo}. The resulting 3D facial reconstructions exhibit a smooth and semantically coherent transition, demonstrating the model's ability to generate continuous emotional expressions through disentangled and controllable latent representations.

The Figure 4-2 shows a visualization of facial expression transitions. The top three illustrate transition sequences from neutral (N) to target emotions: happiness (H), sadness (S), and anger (A). At mid-level intensity (central images), distinct emotional characteristics begin to emerge. For happiness, a subtle upward curl of the mouth corners and a slight elevation of the cheeks are observed. In the sadness transition, brows start to furrow slightly inward with the inner corners dropping, while the mouth corners turn subtly downward. The anger transition shows initial tension around the brow region with a slight narrowing of the eye aperture and horizontal compression of the lips. At maximum intensity (rightmost images),

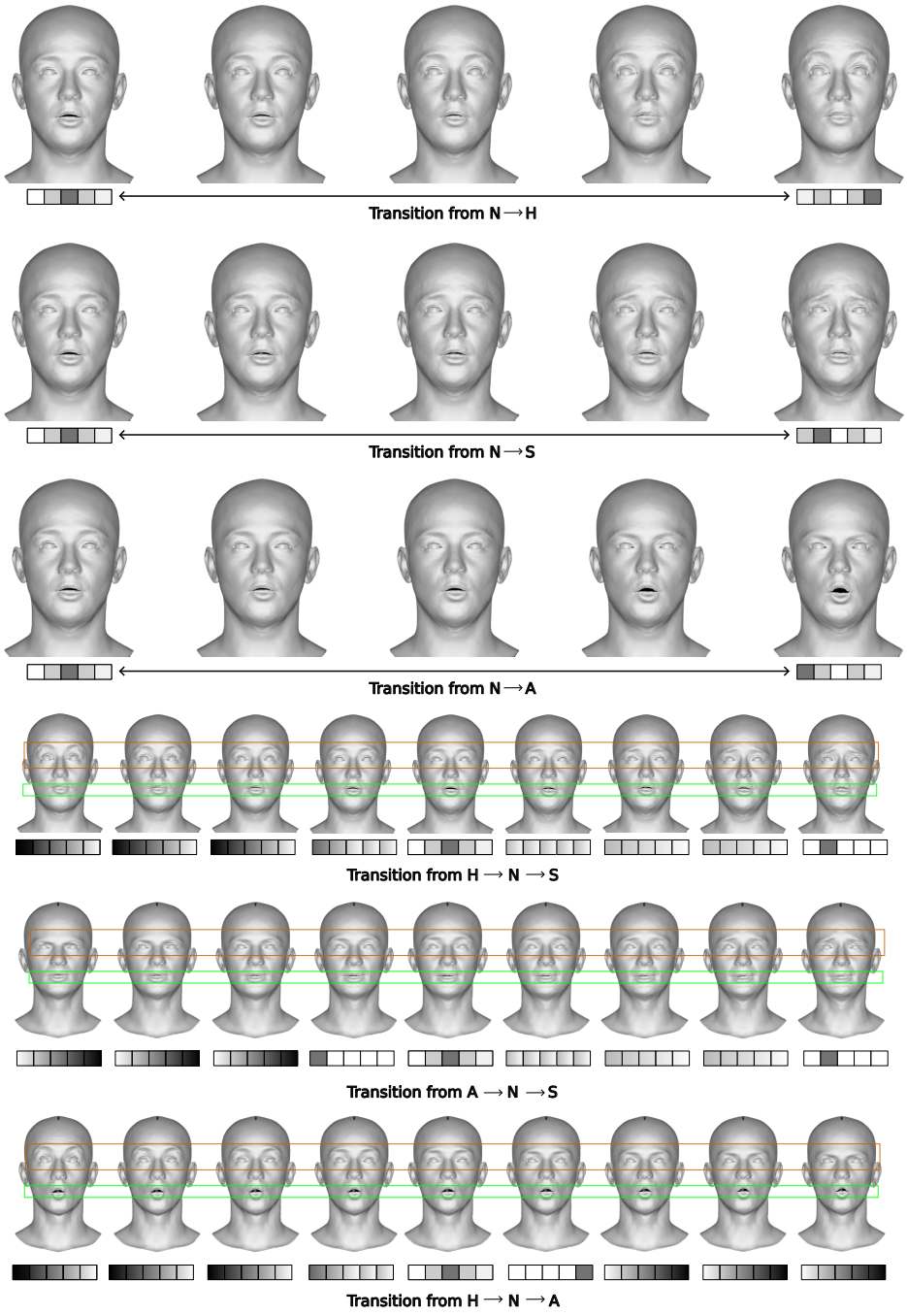

Figure 4-2: Emotion control. Each row illustrates the generation of three emotional expressions – Happy (H), Sad (S), and Angry (A) – from the same EEG trial. The facial reconstructions are generated using frozen EEG encoders $\mathbf{E}_r$ and $\mathbf{E}_e$, while the emotional variation is controlled by adjusting the latent code $\bar{\mathbf{z}}_e$.

the emotion expressions reach their peak. The happiness expression displays pronounced elevation of the cheeks, widened mouth corners with visible teeth, and subtle crow's feet around the eyes. In the sadness expression, brows are significantly lowered at their inner corners, nasolabial folds deepen, and the mouth adopts a distinctive downturned shape with a slight depression in the center of the lower lip. The anger expression exhibits strongly furrowed brows that angle downward toward the nose bridge, narrowed eyes, flared nostrils, and compressed lips with tension visible in the chin area.

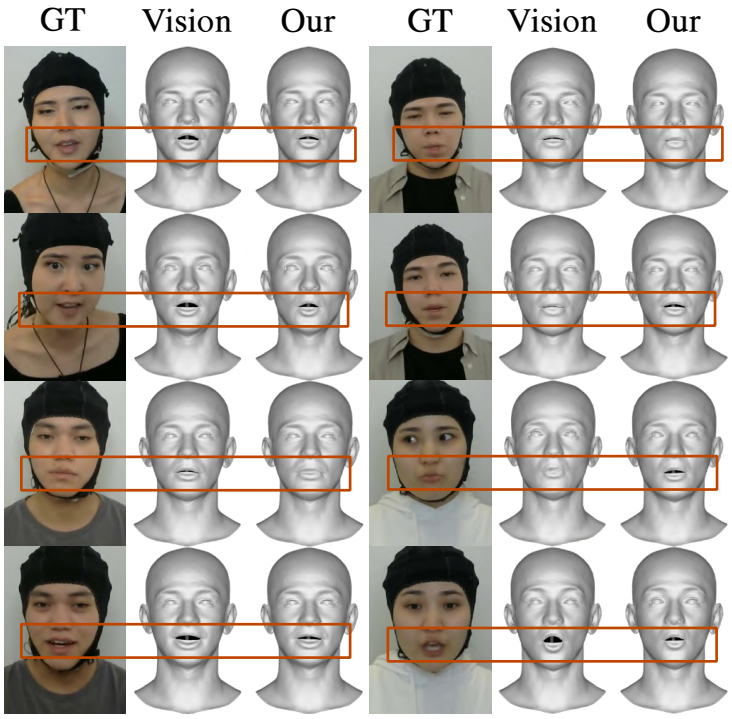

Figure 4-3: Comparison of 3D reconstructions with detailed displacements. **Top:** Vision input (ground truth), **Middle:** EEG2Face results, **Bottom:** Vision2Face (EMOCA) results. While the model consistently captures upper-face features such as eyebrow and eye region expressions, inaccuracies are observed in mouth shapes. Specifically, the model sometimes predicts mouths that are not sufficiently open or not fully closed, highlighting a limitation in decoding nuanced mouth dynamics from EEG signals.

In the bottom three rows of Figure 4-2 the model's performance in maintaining content stability during emotional transitions is examined, illustrating three critical emotional transitions: Happy-Neutral-Sad, Angry-Neutral-Sad, and Happy-Neutral-Angry. As highlighted

by the red boundary around the upper facial regions, our model successfully modulates emotional intensity across these transitions while preserving content consistency, indicated by the green boundary surrounding the mouth area. This visual evidence supports our hypothesis that emotional expression and content representation can be disentangled, with upper facial features primarily conveying affective states while maintaining the underlying EEG signal content integrity. The graduated transitions between emotional states further demonstrate the model's ability to generate realistic intermediary expressions rather than abrupt shifts, suggesting robust comprehension of the emotional continuum. These results validate the effectiveness of our synchronization and synthesis approaches in producing naturalistic facial animations driven by EEG signals.

Figure 4-3 demonstrates a comparative analysis of our EEG2Face system against ground truth (GT) expressions and vision-based predictions (Vision) across four subjects. The red boundary boxes highlight the mouth region, where our evaluation reveals certain limitations in reconstruction accuracy. While our model demonstrates reasonable performance in many cases, the selected examples illustrate challenging scenarios where mouth shape reproduction is imperfect. The reconstructions occasionally struggle with precise mouth aperture-producing shapes that are either not sufficiently open or not fully closed compared to ground truth. It's important to note that these examples represent more challenging cases rather than typical performance. This difficulty in perfectly capturing mouth dynamics likely stems from the inherent complexity of mapping EEG signals to the intricate muscular control required for precise mouth articulation. Despite these challenging cases, our system maintains overall facial expression coherence and demonstrates that EEG signals can indeed drive facial animation, though with occasional limitations in capturing the full range of mouth movements.

Figure 4-4 illustrates photorealistic 3D face generation using albedo maps $\mathbf{a}$, where the albedo is either extracted from the participant (first row) or taken from a random template (second row). Given an EEG signal and the model's latent representations – facial regressor $\mathbf{z}_r$,

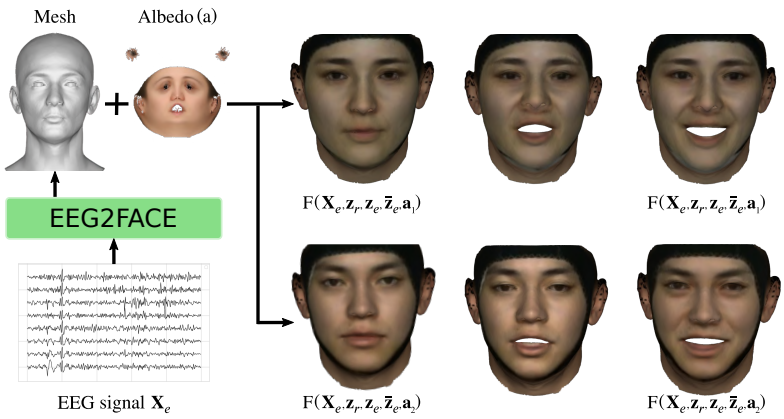

Figure 4-4: Our model enables photorealistic neural avatar generation from EEG signals. For the same EEG signal $\mathbf{X}$, different emotion codes $\bar{\mathbf{z}}_i$ and albedo maps $\mathbf{a}_i$ allow our generator F to synthesize diverse facial expressions and appearances.

emotion component $\mathbf{z}_e$, and emotion latent code $\bar{\mathbf{z}}_e$ – a realistic face is synthesized using the generation function $F(\mathbf{z}_r, \mathbf{z}_e, \bar{\mathbf{z}}_e, \mathbf{a})$.

4.0.2 Quantitative evaluation

For quantitative analysis the Lip Vertex Error (LVE) [37, 12] and the Emotional Vertex Error (EVE) [35]. LVE is defined as the ℓ_2 error among all lip vertices, while EVE measures the ℓ_2 error in emotion-related regions, including the forehead, cheeks, and eye vertices. To assess the accuracy of lip movements, the lip synchronization metric utilized in the MeshTalk [37] study is adopted. Specifically, the maximum L2 error across all lip vertices is computed for each frame, serving as a quantitative measure of synchronization accuracy. This approach allows for an objective evaluation of how closely the predicted lip movements align with the corresponding real data.

Our study evaluates five distinct neural network architectures for EEG-based emotional face reconstruction: ShallowConvNet [40], EEGNet [25], TCT [31], EEGTransformer [1], and our proposed $\mathbf{E}_r$ Transformer-based model. The original backbone structures of these models are maintained, while suitably modifying them for regression and classification tasks.

Table 4.1: Quantitative comparison of EEG encoder architectures across three facial geometry metrics: Lip Vertex Error (LVE), Mouth Corner Error (MCE), and Eye Vertex Error (EVE). Lower values indicate better performance.

Method	**LVE (mm)**	**MCE (mm)**	**EVE (mm)**
[25]	1.0975	2.4331	0.5385
[40]	1.0439	1.2331	0.4554
[31]	0.8658	2.7552	0.5055
[1]	0.7856	0.8329	0.2699
Ours	0.6647	0.4864	0.1541

The quantitative evaluation results demonstrate the superior performance of our proposed Transformer-based architecture for EEG-based emotional face reconstruction compared to other EEG encoding approaches. As shown in Table 4.1, $\mathbf{E}_r$ achieves the lowest error rates in two critical facial feature metrics: LVE at 0.6647×10^{-4} mm, MCE at 0.4864×10^{-4} mm, and EVE at 0.1541×10^{-4} mm. Our model shows substantial improvements over established architectures, with nearly twice the accuracy of ShallowConvNet in LVE measurement and almost three times better performance than TCT in MCE. The improvement is particularly pronounced in capturing eye region dynamics, where our approach reduces the error by almost half compared to EEGTransformer, the next best performing model. These consistent performance advantages across all metrics validate that our transformer-based architecture $\mathbf{E}_r$ design effectively captures the complex spatio-temporal relationships between EEG signals and facial geometry, especially in regions requiring fine-grained control.

4.1 Discussion

4.1.1 Model Architectures

In this study, an EEG-based model was developed using a transformer backbone, with the architecture deliberately kept as simple as possible. Specifically, positional encodings and class tokens were not incorporated. This decision was motivated in part by the goal of maintaining model compactness; however, it was primarily influenced by the observation that negligible performance gains were achieved when these components were included. It is hypothesized that, in a subject-dependent setting, the available data per subject is insufficient to benefit from class-token-based summarization. Furthermore, for relatively well-defined tasks such as regression or emotion classification, it was found that compact models perform sufficiently well without additional architectural complexity.

Previous EEG transformer studies have also reported that deeper layers can sometimes degrade performance. This observation is consistent with the findings in this work, where the attention and feedforward modules in the early transformer layers were found to act as effective spectral, temporal, and spatial filters. As a result, a shallow structure with $L = 2$ transformer layers was selected.

4.1.2 Limitations

A key challenge in 3D facial reconstruction from non-visual modalities – such as audio or EEG – is the acquisition of appropriate ground truth (GT) 3D meshes. While high-quality 3D meshes can be obtained using dedicated scanning equipment or specific vision datasets [38, 7], such resources are often limited, expensive, and difficult to scale.

To address this, many recent studies adopt an alternative strategy: leveraging vision-based 3D reconstruction models to obtain *pseudo-ground-truth*. This approach is consistent with prior works in audio-driven face synthesis, such as VOCA [8], MeshTalk [37], and Face-

Former [12], where fitted 3D Morphable Model (3DMM) parameters extracted from RGB images are treated as ground truth for supervision.

In this study, the EMOCA model [9] is used to obtain pseudo-ground-truth supervision. While our EEG-based model demonstrates competitive performance in reconstructing 3D facial expressions, its accuracy is inevitably affected by the quality of the vision model providing supervision. As illustrated in Figure 4-3, some mismatches can be observed in the reconstructed expressions. These discrepancies may arise from either limitations in the vision-based supervision or the EEG model's capacity to effectively regress toward the vision-derived latent space.

Unlike the vision model, which directly maps to the generated 3D facial meshes [8, 37, 9], EEG signals do not have a direct correspondence to 3D facial geometry, making the learning of such a mapping inherently more challenging. Furthermore, the quality of EEG signals can degrade over time due to external factors such as increased electrode impedance or signal drift, which may lead to a decline in model performance [26].

However, in terms of emotional representation, our approach outperforms the vision-only baseline by leveraging the complementary nature of EEG signals. While the EEG facial regressor ($\mathbf{E}_r$) is guided by vision supervision, the emotion decoder ($\mathbf{E}_e$) is trained solely from the EEG oscillations. This enables the model to capture intrinsic emotional states that may not be externally expressed, resulting in a richer and more comprehensive representation of emotion.

Meanwhile, facial morphology tends to exhibit consistent and shared structural patterns across individuals, enabling the development of generalized facial extractors (e.g., $\mathbf{E}_v$). In contrast, EEG signals are highly subject-specific and often require user-dependent calibration [26, 27, 4]. Although our study adopts a subject-dependent design, extending the proposed system to a subject-independent setting remains an important direction, especially as recent BCI research emphasizes the development of generalized models capable of operating

across users without the need for individual calibration [22].

Chapter 5

Conclusion

In this study, a novel framework for EEG-driven 3D face generation has been presented, in which expressive facial geometry is reconstructed directly from brain signals. A transformer-based EEG encoder were leveraged to disentangle facial structure and emotional states, thereby enabling controllable face synthesis through interpretable latent representations.

This work is the first to demonstrate that EEG signals alone can drive photorealistic facial synthesis with semantically meaningful emotional control. This outcome highlights the strong representational capacity of the disentangled EEG features.

This approach opens up new possibilities across multiple application domains, particularly in neural avatar generation. In BCI research, an alternative communication pathway is provided for users with impaired facial expressions, such as individuals with ALS or facial paralysis. In psychiatry, a means is offered to monitor emotional states in nonverbal populations, where traditional behavioral cues may be absent.

Bibliography

[1] Resha Dwika Hefni Al-Fahsi. EEG Motor Imagery Classification Using CNN, Transformer, and MLP.

[2] Oleg Alexander, Mike Rogers, William Lambeth, Matt Chiang, and Paul Debevec. The digital emily project: photoreal facial modeling and animation. In *Acm siggraph 2009 courses*, pages 1–15. 2009.

[3] Kai Keng Ang, Zheng Yang Chin, Haihong Zhang, and Cuntai Guan. Filter Bank Common Spatial Pattern (FBCSP) in Brain–Computer Interface. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 2390–2397, 2008.

[4] Ji-Seon Bang, Min-Ho Lee, Siamac Fazli, Cuntai Guan, and Seong-Whan Lee. Spatio-spectral Feature Representation for Motor Imagery Classification Using Convolutional Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, 33(7):3038–3049, 2021.

[5] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE international conference on computer vision*, pages 1021–1030, 2017.

[6] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. IEMOCAP: Interactive Emotional Dyadic Motion Capture Database. *Language resources and evaluation*, 42:335–359, 2008.

[7] Darren Cosker, Eva Krumhuber, and Adrian Hilton. A facs valid 3d dynamic action unit database with applications to 3d dynamic morphable facial modeling. In *2011 international conference on computer vision*, pages 2296–2303. IEEE, 2011.

[8] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael Black. Capture, learning, and synthesis of 3D speaking styles. *Computer Vision and Pattern Recognition (CVPR)*, pages 10101–10111, 2019.

[9] Radek Daněček, Michael J Black, and Timo Bolkart. Emoca: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20311–20322, 2022.

[10] Radek Daněček, Kiran Chhatre, Shashank Tripathi, Yandong Wen, Michael Black, and Timo Bolkart. Emotional speech-driven animation with content-emotion disentanglement. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–13, 2023.

[11] Masoumeh Esmaeili and Kourosh Kiani. Generating personalized facial emotions using emotional eeg signals and conditional generative adversarial networks. *Multimedia Tools and Applications*, 83(12):36013–36038, 2024.

[12] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Faceformer: Speech-driven 3d facial animation with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18770–18780, 2022.

[13] Gabriele Fanelli, Juergen Gall, Harald Romsdorfer, Thibaut Weise, and Luc Van Gool. A 3-d audio-visual corpus of affective communication. *IEEE Transactions on Multimedia*, 12(6):591–598, 2010.

[14] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4):1–13, 2021.

[15] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. 2018.

[16] Jianzhu Guo, Dingyun Zhang, Xiaoqiang Liu, Zhizhou Zhong, Yuan Zhang, Pengfei Wan, and Di Zhang. Liveportrait: Efficient portrait animation with stitching and retargeting control. *arXiv preprint arXiv:2407.03168*, 2024.

[17] Aaron S Jackson, Adrian Bulat, Vasileios Argyriou, and Georgios Tzimiropoulos. Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. In *Proceedings of the IEEE international conference on computer vision*, pages 1031–1039, 2017.

[18] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu, Chen Change Loy, Xun Cao, and Feng Xu. Audio-driven emotional video portraits. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14080–14089, 2021.

[19] Stamos Katsigiannis and Naeem Ramzan. Dreamer: A database for emotion recognition through eeg and ecg signals from wireless low-cost off-the-shelf devices. *IEEE journal of biomedical and health informatics*, 22(1):98–107, 2017.

[20] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing*, 3(1):18–31, 2011.

[21] Mangesh Ramaji Kose, Mitul Kumar Ahirwal, and Anil Kumar. A new approach for emotions recognition through eog and emg signals. *Signal, Image and Video Processing*, 15(8):1863–1871, 2021.

[22] O-Yeon Kwon, Min-Ho Lee, Cuntai Guan, and Seong-Whan Lee. Subject-independent Brain–Computer Interfaces Based on Deep Convolutional Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, 31(10):3839–3852, 2019.

[23] Yu-Ting Lan, Kan Ren, Yansen Wang, Wei-Long Zheng, Dongsheng Li, Bao-Liang Lu, and Lili Qiu. Seeing through the brain: image reconstruction of visual perception from human brain signals. *arXiv preprint arXiv:2308.02510*, 2023.

[24] Peter J Lang, Margaret M Bradley, and Bruce N Cuthbert. Emotion, attention, and the startle reflex. *Psychological review*, 97(3):377, 1990.

[25] Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. EEGNet: a Compact Convolutional Neural Network for EEG-based Brain–Computer Interfaces. *Journal of neural engineering*, 15(5):056013, 2018.

[26] Min-Ho Lee, O-Yeon Kwon, Yong-Jeong Kim, Hong-Kyung Kim, Young-Eun Lee, John Williamson, Siamac Fazli, and Seong-Whan Lee. Eeg dataset and openbmi toolbox for three bci paradigms: An investigation into bci illiteracy. *GigaScience*, 8(5):giz002, 2019.

[27] Min-Ho Lee, Adai Shomanov, Balgyn Begim, Zhuldyz Kabidenova, Aruna Nyssanbay, Adnan Yazici, and Seong-Whan Lee. EAV: EEG-Audio-Video Dataset for Emotion Recognition in Conversational Contexts. *Scientific Data*, 11(1):1026, 2024.

[28] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017.

[29] Ya Li, Jianhua Tao, Linlin Chao, Wei Bao, and Yazhu Liu. Cheavd: a chinese natural emotional audio–visual database. *Journal of Ambient Intelligence and Humanized Computing*, 8:913–924, 2017.

[30] Yihong Lin, Liang Peng, Jianqiao Hu, Xiandong Li, Wenxiong Kang, Songju Lei, Xianjia Wu, and Huang Xu. Emoface: Emotion-content disentangled speech-driven 3d talking face with mesh attention. *arXiv preprint arXiv:2408.11518*, 2024.

[31] Yanling Liu, Yueying Zhou, and Daoqiang Zhang. Tct: Temporal and channel transformer for eeg-based emotion recognition. In *2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 366–371, 2022.

[32] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019.

[33] Dan Nemrodov, Matthias Niemeier, Ashutosh Patel, and Adrian Nestor. The neural dynamics of facial identity processing: insights from eeg-based pattern analysis and image reconstruction. *Eneuro*, 5(1), 2018.

[34] Cheul Young Park, Narae Cha, Soowon Kang, Auk Kim, Ahsan Habib Khandoker, Leontios Hadjileontiadis, Alice Oh, Yong Jeong, and Uichin Lee. K-emocon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations. *Scientific Data*, 7(1):293, 2020.

[35] Ziqiao Peng, Haoyu Wu, Zhenbo Song, Hao Xu, Xiangyu Zhu, Jun He, Hongyan Liu, and Zhaoxin Fan. Emotalk: Speech-driven emotional disentanglement for 3d face animation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20687–20697, 2023.

[36] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*, 2018.

[37] Alexander Richard, Michael Zollhöfer, Yandong Wen, Fernando De la Torre, and Yaser Sheikh. Meshtalk: 3d face animation from speech using cross-modality disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1173–1182, 2021.

[38] Kathleen M Robinette, Hans Daanen, and Eric Paquet. The caesar project: a 3-d surface anthropometry survey. In *Second international conference on 3-D digital imaging and modeling (cat. No. PR00062)*, pages 380–386. IEEE, 1999.

[39] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J. Black. Learning to regress 3d face shape and expression from an image without 3d supervision. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7755–7764, 2019.

[40] Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep Learning with Convolutional Neural Networks for EEG Decoding and Visualization. *Human brain mapping*, 38(11):5391–5420, 2017.

[41] Kim Sung-Bin, Lee Hyun, Da Hye Hong, Suekyeong Nam, Janghoon Ju, and Tae-Hyun Oh. Laughtalk: Expressive 3d talking head generation with laughter. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6404–6413, 2024.

[42] A. Tran, Tal Hassner, Iacopo Masi, and Gérard G. Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1493–1502, 2016.

[43] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *ECCV*, 2020.

[44] Isabell Wohlgenannt, Alexander Simons, and Stefan Stieglitz. Virtual reality. *Business & Information Systems Engineering*, 62:455–461, 2020.

[45] Jinbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun, Jue Wang, and Tien-Tsin Wong. Codetalker: Speech-driven 3d facial animation with discrete motion prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12780–12790, 2023.

[46] Tianke Zhang, Xuangeng Chu, Yunfei Liu, Lijian Lin, Zhendong Yang, Zhengzhuo Xu, Chengkun Cao, Fei Yu, Changyin Zhou, Chun Yuan, and Yu Li. Accurate 3d face reconstruction with facial component tokens. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9033–9042, October 2023.

[47] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *European Conference on Computer Vision*, 2022.