Recognizing Emotions from EEG Data with Vision Transformers and Continuous Wavelet Transform

by

Agustin E. Lorenzo

(Under the direction of Neal Outland)

Abstract

A model that could accurately recognize emotions from physiological data would have potential for applications in human-computer interaction, psychotherapy, neuroscience, and medicine. Advancements in deep learning have steadily increased the feasibility of developing such a model, with the introduction of transformer architectures causing the most recent improvements in performance. However, vision transformers remain significantly less explored than their transformer + CNN counterparts. This paper demonstrates a method for emotion recognition from private EEG data through the use of 3D vision transformers and continuous wavelet transform features arranged in a video format. Averages for classification metrics were obtained using 5-fold cross validation, and results indicate that, when instances are clipped into non-overlapping segments, most models finetuned with public data before private data outperform their respective baselines, with potential for improved performance when accounting for overfitting.

Index words: Emotion recognition, Affective computing, Vision transformer, Electroencephalography, EEG, Wavelet transform, CWT, Computer vision, Machine learning, Classification, Computational neuroscience

Recognizing Emotions from EEG Data with Vision Transformers and Continuous Wavelet Transform

by

Agustin E. Lorenzo

B.A., The University of Georgia, 2024

A Thesis Submitted to the Graduate Faculty
of The University of Georgia in Partial Fulfillment
of the
Requirements for the Degree

Master of Science

Athens, Georgia

2025

Recognizing Emotions from EEG Data with Vision Transformers and Continuous Wavelet Transform

by

Agustin E. Lorenzo

Approved:

Major Professor: Neal Outland

Committee: Tianming Liu
Dean Sabatinelli

Electronic Version Approved:

Ron Walcott
Dean of the Graduate School
The University of Georgia
December 2025

Acknowledgments

I couldn't have completed this work without the support of the AI Institute. I am very grateful to have been awarded The Benjamin Lloyd Cloer Endowment in Artificial Intelligence, and the faculty here has helped me countless times throughout both of my degrees from UGA. I also want to thank my committee for helping me see this project through to the end, and my close friends and family for keeping me motivated through the process.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Emotions are complex mental states that can be assessed in several ways. The most straightforward and standard methods involve recording external signals like facial expression, speech, or eye blinking [2]. However, these methods are limited in their ability to accurately and consistently capture an individual's true emotional state, since external representations of emotions can easily be masked or misinterpreted [3]. More direct and objective methods of assessing emotion involve recording physiological signals through an electrocardiogram, electromyogram, or electroencephalogram (EEG). Of these, EEG is most commonly used due to its accessibility, low risk, and low cost [4].

The field focused on the development of technologies that can recognize, interpret, and respond to human emotions is known as affective computing. Currently, emotion recognition is regarded as the most important research topic within this field [5, 6]. The ability to accurately classify emotions on scales of valence and arousal would be useful in many cases. In the context of brain-computer interfaces, understanding the user's emotions could lead to better human-machine interactions [6]. There is also potential for uses in psychotherapy, criminal investigations [7], and medicine [8].

As the capability of machine learning models has steadily improved, the feasibility of developing an effective emotion recognition model has grown alongside it, which has brought more attention to the topic as of late. Recently, the introduction of transformer architectures to emotion recognition models has shown a promising improvement over previous machine learning methods. Recent approaches to emotion recognition that utilize transformers typically vary in terms of both model architecture and features extracted from EEG data. Of

these approaches, the vision transformer and continuous wavelet transform (CWT) are the least explored architecture and feature combination, respectively. Interestingly, this combination also shows promise, since CWT provides a visual input that lends itself to a vision transformer. This thesis demonstrates an approach to emotion recognition using vision transformers trained on CWT features extracted from EEG data. The proposed method allows for training vision transformers across multiple datasets with varying sets of EEG channels and shows the benefit of finetuning on public data before private data. The main contributions of this thesis are as follows:

- Introduce a novel approach in emotion recognition using 3D vision transformers and continuous wavelet transform.
- Create a method for ensuring similar, continuous orders from any set of EEG channels that enables model training across datasets.
- Demonstrate the impact of finetuning with public data before private data on model performance.
- Identify the effect of training on clipped and unclipped instances in emotion recognition.

The remainder of this paper is divided into four sections. Section 2 provides further background on relevant terms in affective computing, along with previous approaches to emotion recognition from EEG data. Section 3 specifies the specific methodology used for training the vision transformer; this includes details for the public and private datasets, feature extraction, hardware, method for determining similar EEG channel orders, and parameters used during model training. Section 4 provides the results of the 5-fold cross-validation procedures. Section 5 provides a discussion of results including implications, a conclusion for the paper, and suggestions for future work.

Chapter 2

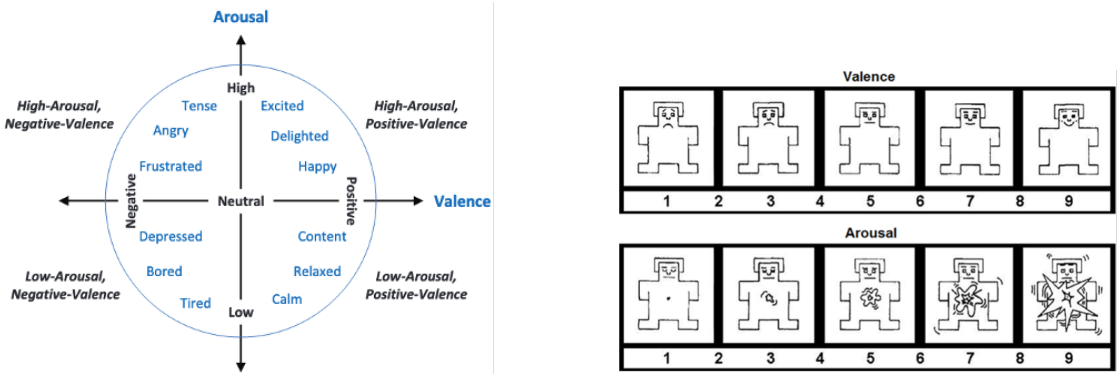
Literature Review

2.1 Background

The following sections contain definitions and overviews of relevant terms along with their relation to emotion recognition and affective computing.

2.1.1 Emotion

In the field of affective computing, the most widely accepted approach to measuring emotion is the circumplex model proposed by Russell, in which emotions are distributed along two axes, valence and arousal [9]. Valence refers to how positive or negative an emotion is, and arousal refers to the emotion's intensity. Together, these scales map most emotions onto a two-dimensional space. This is particularly useful for computation because it provides a quantitative measure that lends itself to direct comparisons between different emotions. To gage an individual's own judgment of their emotional state, they are typically given a Self-Assessment Manikins (SAM) scale-based questionnaire, where they can visually indicate their perceived degrees of valence and arousal [10]. After an experiment, the values indicated on the SAM questionnaire provide a reference for a stimulus' emotional effect on the participant. In emotion recognition, these values are eventually used as ground truth values during model training.



(a) Circumplex model [9] (b) SAM questionnaire [11]

Figure 2.1: Illustrations of valence and arousal

2.1.2 EEG

EEG is a form of neuroimaging that records the electrical activity of the brain through electrodes placed on the scalp [12]. This represents the level of activation for the corresponding brain region in response to the stimulus. In neuroscience, this is used to identify which brain regions are most active during a given task or stimulus. A main downside of EEG is the low spatial resolution caused by signals weakening as they travel through the skull, scalp, and hair of the participant. Additionally, because of the distance between deeper brain regions and the electrodes, EEG is not as effective for capturing brain activity under the cortex. However, it makes up for these weaknesses with the fact that its quick electrical measurements provide a high temporal resolution, especially when compared to other neuroimaging methods like magnetic resonance imaging. This makes EEG ideal for studying immediate brain responses to brief stimuli, which in turn makes it ideal for studying emotion.

Once the raw EEG data is collected, there are many features to select from in time, space, and frequency domains [13]. Spatial features encode each electode's position in relation to other electrodes. Features that pertain to the time domain can be represented by differential entropy (DE), which describes the time-series complexity of a signal [14, 15]. The frequency

domain divides the frequencies of the EEG signal into a set of ranges, or bands, where each band is associated with a state of mind [16, 4]. A common feature used to characterize the frequency domain is power spectral density (PSD), which describes the power per unit frequency interval [14, 17]. There are also time-frequency domain features, which can be extracted using various forms of wavelet transform [5]. Time-frequency features, like CWT, describe how the power of frequency components change over time, which can be plotted in a scalogram. An example of a CWT scalogram is shown in figure 2.2. Once these features are extracted, they can then be used as inputs to machine learning algorithms for emotion recognition.

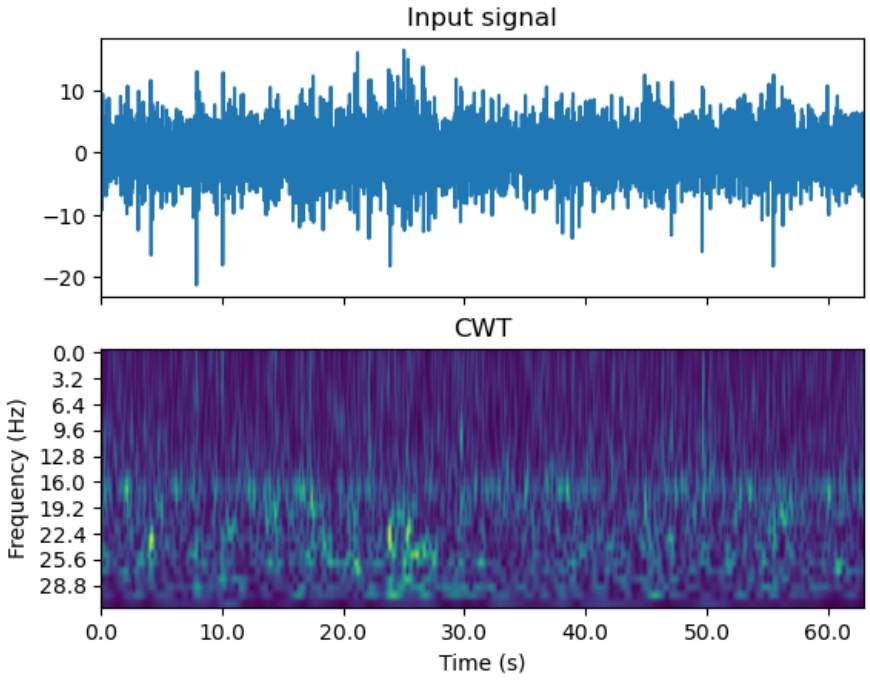


Figure 2.2: Example of CWT scalogram generated from an electrode's signal

2.1.3 Transformers

A transformer is a deep learning architecture proposed by Vanswani et al. [1] that was originally designed for machine translation. It has since been shown to be applicable to many contexts outside of translation, like large language models, computer vision, and speech

processing. The overall function of the transformer allows for an input sequence of symbols to be mapped to a sequence of continuous representations, which can then be used to predict an output sequence of symbols one at a time. In natural language processing (NLP), the input and output symbols are typically short sequences of characters, or tokens, and the continuous representation is a vector of real numbers that represents that token. This vector is also known as the token's embedding, and it is thought to capture the token's semantic meaning numerically.

What sets the transformer apart from other models is its attention mechanism. This captures the relationships between all tokens in a sequence by computing a dot-product matrix with three vectors corresponding to each token: query Q, key K, and value V. The matrix of outputs is computed with the following formula:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

Where T is the current token's position in the sequence of inputs, and d_k is the dimensionality of the queries and keys. This matrix accounts for relationships between features that are relevant on the global level, as opposed to other deep learning models like convolutional neural networks (CNNs) that only account for the local level.

The first component of the transformer is the encoder module, which consists of N identical layers. Each layer is further divided into two sub-layers. The first of these sub-layers is the self-attention mechanism, and the second is a fully connected feed-forward network. The second component is the decoder module, which also consists of N identical layers. The decoder contains the same two sub-layers as the encoder, along with a third self-attention sub-layer in between that performs attention on the output of the encoder. The first self-attention sub-layer in the decoder is also masked to ensure that the current token doesn't attend to tokens that follow it, which prevents leakage from future tokens.

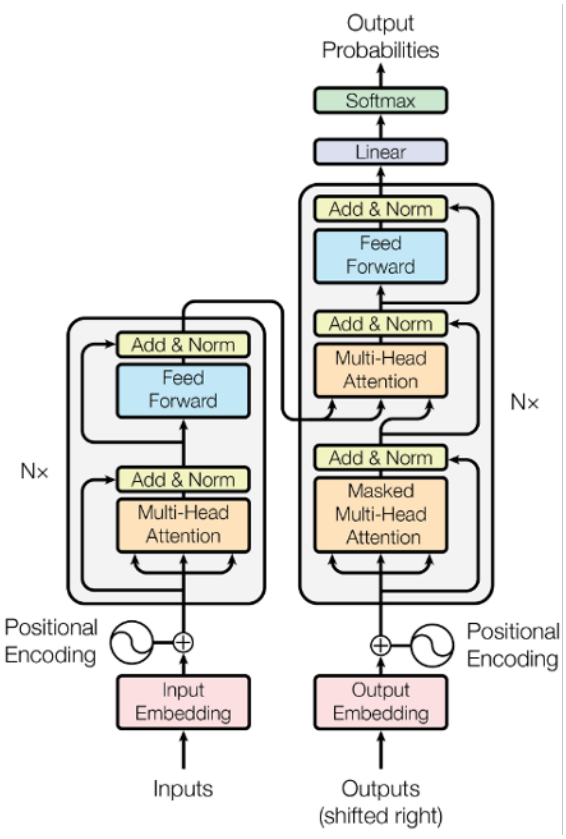
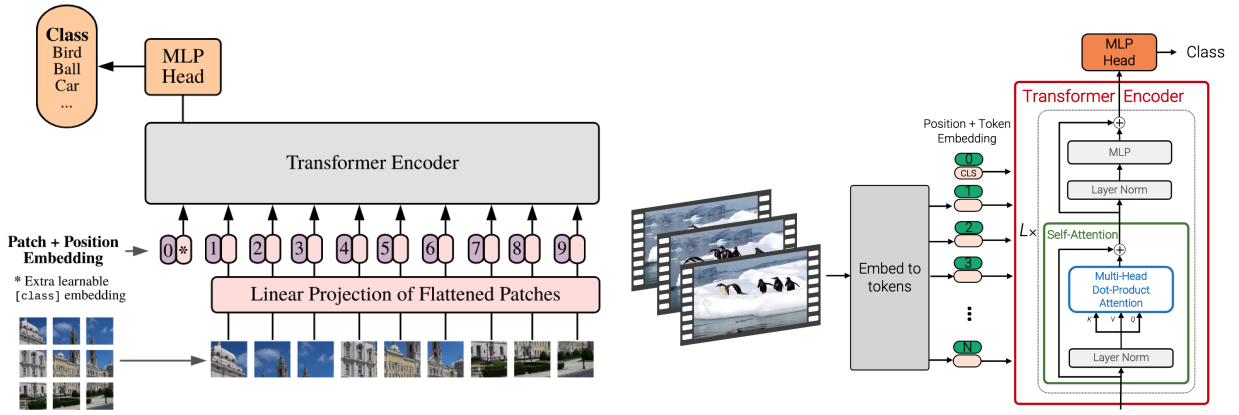


Figure 2.3: Transformer model architecture from Vanswani et al. [1]

While the original transformer architecture was ideal for NLP, it still needed to be adapted for use in other domains, like computer vision. The vision transformer extends the original transformer architecture to process visual inputs [18]. Before its introduction, standard transformers were typically used in conjunction with CNNs. Vision transformers were developed to eliminate reliance on convolutional architectures while reducing computational requirements. To achieve this, they divide images into patches, treating them similarly to tokens in NLP tasks. Just as standard vision transformers can be used for processing 2D images, 3D vision transformers, also known as video vision transformers, extend this further by processing spatio-temporal tokens extracted from video inputs [19].



(a) ViT adapted from [18] (b) ViViT adapted from [19]

Figure 2.4: Vision transformer model architectures.

2.2 Related Work

The simplest possible approach to emotion recognition involves specifying a threshold for an EEG feature so that if a given sample exceeds this threshold, it is classified as belonging to a particular emotional state [20]. However, because this threshold would vary for each subject, this method is not generalizable [10]. Inter-subject variability is considered to be the main obstacle in developing models based on EEG data [21]. For this reason, most approaches to emotion recognition utilize some form of machine learning. Machine learning methods in emotion recognition before the introduction of transformers can be divided into two groups: conventional machine learning and deep learning. Earlier attempts typically utilized conventional machine learning by feeding extracted EEG signal features into a classification algorithm [3]. Such algorithms include support vector machines (SVM) [22], random forests (RF) [23], and K-nearest neighbors (KNN) [23], [24].

Although some of these showed promising results, they were soon replaced by deep learning methods for a few reasons. First, extracting the relevant features required for these classification algorithms is difficult and requires expert knowledge [3]. Deep learning

approaches also became much more viable over time, and began to outperform these algorithms in other applications that involved learning complex patterns [7]. That is, deep learning had an advantage in emotion recognition because of its ability to nonlinearly transform EEG features into any vector space due to its universal approximation property [10]. Notable deep learning models in this context include recurrent neural networks (RNNs) [25], long short term memory (LSTM) [26], CNNs [27], and graph convolutional neural networks (GCNNs) [10].

However, these deep learning models come with their own set of limitations. Mainly, these models struggle with effectively capturing long-distance EEG features that may be relevant to the current emotion being evaluated. Neither CNNs nor RNNs consider non-local EEG features [28]. A majority of these models are also either based on CNNs or include a CNN as a main component, and both single-layer and multi-layer convolution varieties have their drawbacks; single-layer lacks the ability to account for long-distance features entirely, while multi-layer is much more computationally expensive [4]. Deep learning was an improvement over conventional machine learning, but the issue of accounting for long-distance features needed to be addressed before moving forward.

2.2.1 Transformers in Emotion Recognition

In recent years, transformers have become increasingly popular since their introduction in Vanswani et al. [1]. They have been shown to be significantly more effective than other machine learning models in tasks that involve processing sequential data, like computer vision and NLP, particularly due to their ability to account for long-range contextual information [8, 29]. This property of transformers served as the motivation behind their application to emotion recognition [21]. In fact, they have even been previously applied to recognizing emotions from textual data, and have shown promising results [30]. Currently, transformers are the most recent development in emotion recognition. Many newly proposed models include

transformer architectures as a main component of their model, and most report their models as achieving state-of-the-art results over non-transformer models.

Some of these models make use of transformers by combining them with other deep learning components in a larger pipeline. Guo et al. and Sun et al. used GCNNs alongside transformers and achieved high mean accuracies when classifying emotions across subjects from the DEAP dataset [21, 31]. Others have used CNNs, and most have achieved comparable mean accuracies [32, 33, 34, 35, 36, 37]. Although some transformer + CNN implementations achieved much lower mean accuracies, they still outperformed conventional machine learning algorithms like KNNs, SVMs, and CNNs alone [38]. Similarly, one approach used a combination of transformers and adversarial discriminative domain adaptation (ADDA) and achieved accuracies higher than ADDA alone [8]. The use of capsule networks also resulted in higher mean accuracies than previous deep learning approaches [39, 40, 41]

Many proposed models only use transformer architectures without any other additional machine learning components. Instead, these models typically vary by data preprocessing and the domain of the EEG feature input. Some make use of frequency features like DE and PSD [4, 14], while others make use of a spatial encoding [42]. However, most make use of an encoding that incorporates spatial and temporal features in some way [6, 4, 29, 10, 43, 44, 45]. Most of these approaches achieved relatively high mean accuracies. A list of proposed models and their performance on the DEAP dataset are given in table 2.2.

Of the proposed architectures, vision transformers are used less often than similar transformer + CNN counterparts, although they still achieve comparably high accuracy scores [46, 47]. CWT features are also some of the least often used features for model training, and there are even fewer approaches that utilize both vision transformers and CWT features together [48]. The limited exploration in both these aspects, along with the natural compatibility of CWT's visual output with vision transformers, serves as the motivation behind using this combination in the current work. Furthermore, prior work has been limited to 2D vision transformers exclusively. This paper is the first to introduce 3D vision transformers

in the context of emotion recognition. Comparison to other similar approaches are provided in table 2.1.

Author	**Features**	**Architecture**	**Data**
Guo et al. [46]	spatiotemporal	ViT + CNN	DEAP, SEED
Awan et al. [47]	DWT	ViT + CNN	DEAP, AMIGOS
Arjun et al. [48]	CWT, temporal	ViT	DEAP
Wang et al. [49]	spatiotemporal	ViT	Private
Current Work	CWT	3D ViT	DEAP, Private

Table 2.1: Comparison between current approach and similar approaches with vision transformers.

Model	Features	Architecture	Valence (%)	Arousal (%)
SAG-CET [21]	spatiotemporal	TF + GCNN	98.89	98.92
Guo et al. [46]	spatiotemporal	ViT + CNN	92.44	92.85
Sun et al. [31]	DE	TF + GCNN	95.91	94.61
STS-TF [6]	spatiotemporal	TF	84.75	82.16
Awan et al. [47]	DWT	ViT + CNN	97.4	97.4
ADDA-TF [8]	spatiotemporal	TF + ADDA	61	64
AMDET [28]	DE	TF	97.48	96.85
Bi-ANN [4]	spatiotemporal	TF	96.96	96.64
BiCCT [32]	spatiotemporal	TF + CNN	94.41	95.15
Asif et al. [29]	spatiotemporal	TF	90.07	84.52
CARL-DSAN [42]	spatial	TF	67.18	67.6
TNAS [10]	spatiotemporal	TF	98.66	98.68
ST-TCNN [36]	spatiotemporal	TF + CNN	96.34	96.95
ERTNet [37]	spatiotemporal	TF + CNN	59.6	63.9
Arjun et al. [48]	CWT	ViT	97	95.75
Arjun et al. [48]	temporal	ViT	99.4	99.1
MEEG-TF [35]	DWT, DE, ES, PSD, temporal	TF + CNN	96	96.8
MES-CTNet [39]	DE, SE, PSD	TF + Capsule	98.31	98.28
TSFFN [33]	spatiotemporal	TF + CNN	98.27	98.53
MSDTTs [34]	spatiotemporal	TF + CNN	93.39	94.36
Capsule-TF [40]	DE	TF + Capsule	96.75	96.88
SECT [14]	DE, PSD	TF	66.83	65.31
EeT [50]	spatiotemporal, frequency	TF	92.86	93.34
TC-Net [41]	CWT	TF + Capsule	98.76	98.81
TcT [43]	spatiotemporal	TF	96.76	97.02
MACTN [38]	temporal	TF + CNN	66.1	67.8
TR&CA [44]	spatiotemporal	TF	95.18	95.58
TSERT [45]	spatiotemporal	TF	67.59	68.87

Table 2.2: Average cross-subject accuracies for valence and arousal on the DEAP dataset along with each model's architecture and features used including spatial, temporal, frequency, DE, PSD, discrete wavelet transform (DWT), CWT, energy spectrum (ES), and sample entropy (SE)

Chapter 3

Methodology

3.1 Datasets

3.1.1 A Database for Emotion Analysis using Physiological Signals (DEAP)

DEAP [51] is a dataset commonly used for training emotion recognition models. This dataset includes recordings of participants' physiological responses to stimuli through EEG, peripheral physiological signals like temperature and respiration, and video of the participant's face. The stimuli used for this dataset are minute-long highlights of music videos that were rated by the participants on scales of arousal, valence, dominance, liking, and familiarity. Valence and arousal were rated on continuous scales of 1-9, and the EEG signal was recorded through 32 channels at 512 Hz. Recordings were gathered from 32 participants responding to 40 videos.

In this study, a preprocessed version of DEAP was used to train the model. The creators of the dataset preprocessed the data by downsampling to 128 Hz, removing EOG artifacts, filtering out frequencies below 4.0 Hz and above 45 Hz, averaging to the common reference, and removing the 3 second pre-trial baseline. Emotional classes were determined by using valence thresholds of 3 and 6 to divide samples into unpleasant, neutral, and pleasant categories. 60 second samples were also split into 2 second clips to match the private dataset's sample length, resulting in 30 clips per sample and 1,200 clips per participant. In total, 38,400 instances were extracted from the DEAP dataset.

3.1.2 Private Dataset

The private dataset used contains 512 Hz, 128-channel EEG recordings of participants responding to images and videos depicting emotional stimuli [52]. Participants consist of 45 University of Georgia students with a mean age of 19.05 (SD = 1.03). Of the 45 participants, there were 32 females and 13 males, with 2 Black, 7 Asian, 1 Hispanic, 2 Indian, 1 Middle-Eastern, 2 Multiracial/Multiethnic, 29 White, and 1 participant who chose not to identify their age or ethnicity. Participants were shown 90 images and 90 videos that were hand-selected to illicit unpleasant, neutral, or pleasant responses. These content categories were also used as emotional classes for each sample.

The videos are 10-second clips taken from video sharing websites on the internet. These clips were picked with the intent of minimizing the effect of artificial elements that may break the viewer's belief that the clips depict true events. Videos were also selected with roughly similar levels of volume and motion to avoid a confound between emotion and action [53]. Image stimuli were obtained by taking the most representative frames from each of the videos. Because images were presented for only 2 seconds, samples from video stimuli were split into non-overlapping 2 second segments to ensure that all instances in the dataset have the same duration.

After EEG data were collected, participants used a computer-based version of the SAM questionnaire to rate each video in terms of valence and arousal on scales of 1–9 in units of 0.1. Before training, the data was preprocessed by removing the baseline before video onset, applying low-pass and high-pass filters, re-referencing to the average reference, and down sampling to 128 Hz. Artifacts were also removed using the SCADS pipline on a per-trial basis, leaving an average of 76.3 image trials and 72.9 video trials retained per participant. In total, 20,236 instances were extracted from both image and video trials.

3.2 Feature Extraction

CWT values were calculated using The fast Continuous Wavelet Transform (`fCWT`) Python library [54]. For each trial per participant, CWT values for frequencies between 4 – 45 Hz were calculated for all electrodes in the dataset, and every 4 values were averaged. The number of frequencies also matched the number of channels, to ensure a square frame when reshaped later, resulting in a ((n channels * n frequencies), (128 Hz * s seconds)/4) shaped 2D image displaying each CWT value for its corresponding channel and frequency at a point in time.

However, when CWT features are arranged in this way, changes along the y-axis do not represent any meaningful relationship in the real data. For example, in this arrangement, CWT values for one channel's highest frequencies are placed next to values for another channel's lowest frequencies; although there may be some underlying relationship between them, their placement is still arbitrary. For this reason, the data is reshaped into an (n frequencies, n channels, (128 Hz * s seconds)/4) 3D array, similar to the approach proposed by Li et al. [25]. Here, each n x n frame displays CWT values for each frequency along the y-axis and each channel along the x-axis at one point in time. These frames are then stacked to create 'video' inputs that are given to the model. An example frame is provided in figure 3.1, and the pipeline for creating a video from CWT scalograms is shown in figure 3.2

3.2.1 Determining Channel Order with Hilbert Curves

Transformers are highly dependent on the sequence of the input data. Vision transformers and video vision transformers learn from spatial relationships between patches in the same way that transformers in NLP learn from the order of input tokens. In the current project, it is important to ensure that EEG channels are ordered in a way that maintains spatial continuity between electrodes, so that spatial relationships between pixels correspond to spatial relationships in the real world. Furthermore, since the model is finetuned using multiple

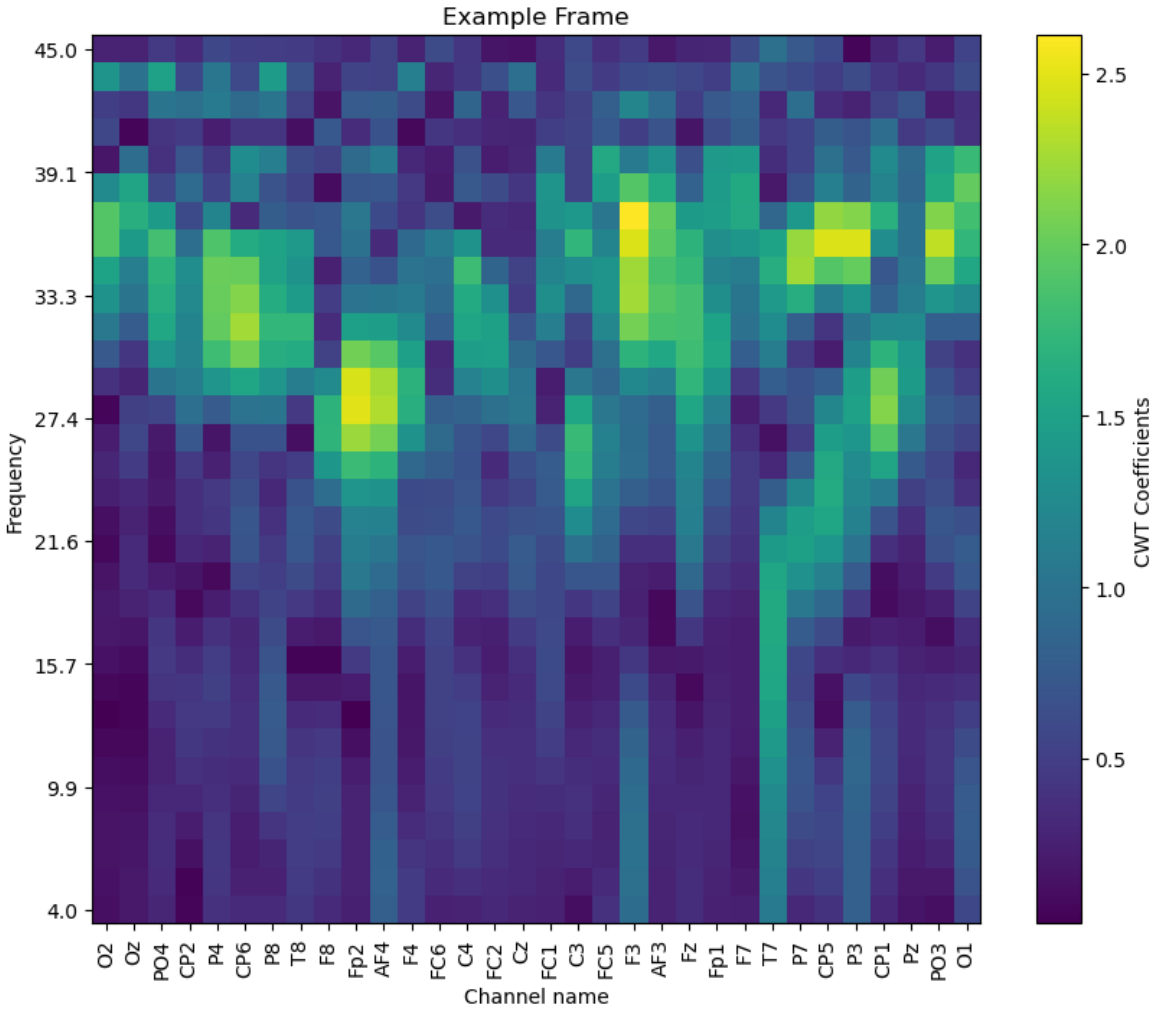


Figure 3.1: An example frame showing CWT values for each channel at one point in time.

datasets with varying numbers of channels, it is also important to ensure that relative positions in the frame always correspond to similar brain regions, regardless of the resolution. Otherwise, patterns learned from one dataset would become meaningless as the model was trained and evaluated on a different dataset with a different arrangement. This creates the need for a general method for obtaining similar orders from any number of EEG channels. Doing so would ensure that relative positions in those orders would always correspond to similar brain regions.

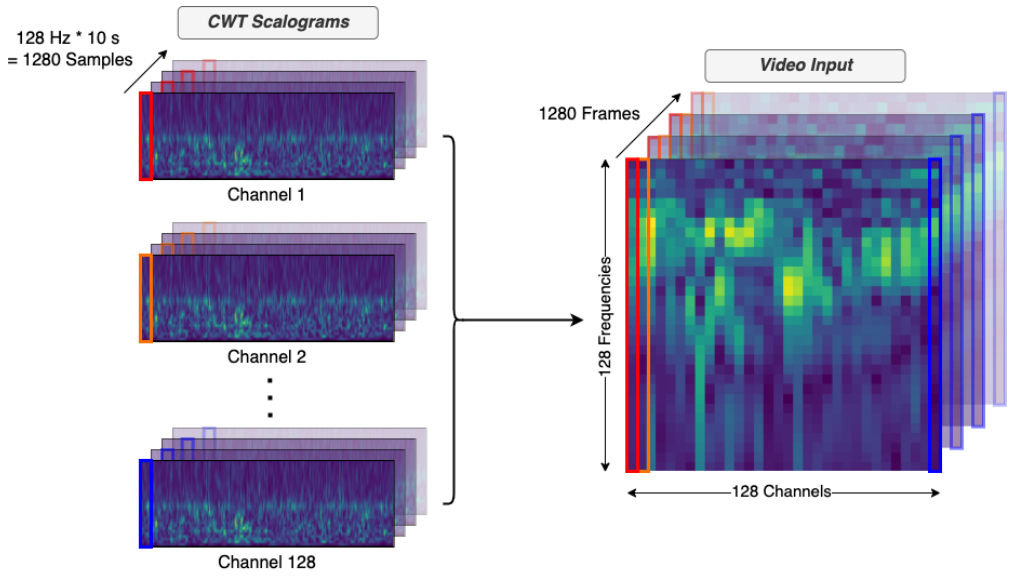
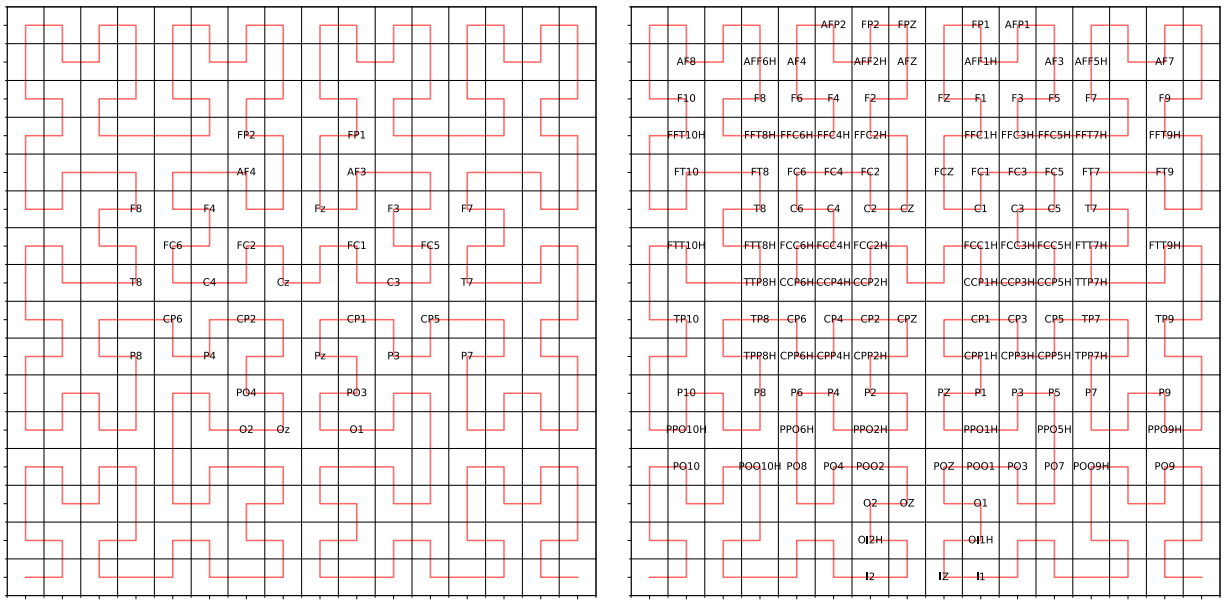


Figure 3.2: Process for creating video input from 128 channel scalograms

Essentially, this method must provide a way to convert 2D locations of electrodes placed on a scalp to a 1D list of channels. In mathematics, this problem is solved with Hilbert curves, which are curves that can be used to fill a square space. In other words, a Hilbert curve can map all points in 2D space onto a 1D line while maintaining continuity. In this case, a Hilbert curve can be used to traverse a 2D matrix containing either 32 or 128 EEG channels. As the curve traverses the matrix, it visits each brain region in a similar order, and adds the names of electrodes to a list as it encounters them along its path. During feature extraction, these lists are used to reorder the channels in both datasets before CWT values are calculated. Figure 3.3 provides the Hilbert curves used for both datasets.



(a) 32 channels from DEAP dataset.

(b) 128 channels from private dataset.

Figure 3.3: Hilbert curves used to determine continuous and similar orders of EEG channels between datasets. The curve begins in the bottom left corner and traverses through every point until it reaches the bottom right corner.

3.3 Model and Training

The model used is Google's base Video Vision Transformer (ViViT) model with 89,236,992 parameters pretrained on the Kinetics-400 dataset [19]. The model's classifier head was adjusted to account for 3 classes, and the transformer layers were initially frozen before training. The model was finetuned in two phases, first with instances from the DEAP dataset and then with instances from the private dataset. All instances were transformed to match the expected input of the ViViT by resizing to (224, 224) without antialiasing, repeating each frame twice to match 3-channel RGB format, and normalizing by the base model's image processor mean and standard deviation. The process for matching the model's expected number of frames is discussed in section 3.3.1.

There were three separate training cases in which the model was trained with 1, 2, or none of the final transformer layers unfrozen. Baseline models were also trained by finetuning the ViViT on the private dataset directly without finetuning on DEAP first. 5-fold cross validation was used to obtain averages for classification metrics including accuracy, precision, recall, ROC AUC, and F1 score. An overall confusion matrix for each model was obtained by summing the model's responses to test splits from each fold. Each model was trained for 20 epochs with a cross-entropy loss criterion that incorporated class weights, `AdamW` optimizer, and learning rate of 3e-5. As for hardware, the model was trained on a Linux computer provided by the University of Georgia's Institute for Artificial Intelligence containing an Intel Core i9 CPU with 3.00 GHz, an Nvidia RTX A5000, and 251 GiB of RAM.

3.3.1 Clipped and Unclipped Private Data

When determining which regions of the brain correspond to an emotional stimulus, it is important to evaluate the participant's response to the stimulus as a whole rather than isolated portions. When the private data was originally collected, the participant's emotional response relied on watching the entire narrative presented in the 10s video stimuli [52]. However, previous work in affective computing has also shown that clipping participant responses into non-overlapping clips can improve model performance [48]. To identify how clipping private data may affect model performance, the entire process is repeated with instances that represent entire 10s unclipped samples. Both clipped and unclipped data are processed by changing the stride used to match the model's expected input of 32 frames.

For both clipped and unclipped samples, every 4 CWT values were averaged to let 32 frames represent a second of data. This also further matches the original frame arrangement from Li et al., which similarly averaged CWT values to reduce the number of frames per second of data [25]. Then, from a 2-second, 64-frame instance corresponding to either an image or clipped video stimulus, every other frame is selected by using a stride of 2 to match the original approach for training the base ViViT [19]. For 10-second, 320-frame instances

representing unclipped video stimuli, a stride of 10 is used. When using unclipped private data, 60s DEAP samples are also split into 10s clips, instead of 2s clips, and processed with a stride of 10 to match the private data. Doing so results in the expected 32 frames for all instances.

Chapter 4

Results

Results are divided into two sections. Section 4.1 contains results from finetuning with unclipped 10s private data samples processed with a stride of 10, and section 4.2 contains results from finetuning with clipped 2s private data samples processed with a stride of 2. Both sections contain confusion matrices, average classification metrics, and training/evaluation loss and accuracy curves for each model. The loss and accuracy curves were generated by averaging the curves recorded during each fold's training process. Curves are grouped together by the number of transformer layers unfrozen when finetuned on private data. Models are labeled using the following key:

Model	**Description**
P0	Baseline model, private data only, 0 layers unfrozen
P1	Baseline model, private data only, 1 layer unfrozen
P2	Baseline model, private data only, 2 layers unfrozen
D0P0	Model trained on both DEAP and private data, with 0 layers unfrozen during training on both datasets
D1P1	Model trained on both DEAP and private data, with 1 layer unfrozen during training on both datasets
D2P2	Model trained on both DEAP and private data, with 1 layer unfrozen during training on both datasets
D1P0	Model trained on both DEAP and private data, with 1 layer unfrozen during training with DEAP, 0 unfrozen for private data
D2P0	Model trained on both DEAP and private data, with 2 layers unfrozen during training with DEAP, 0 unfrozen for private data
D2P1	Model trained on both DEAP and private data, with 2 layers unfrozen during training with DEAP, 1 unfrozen for private data

Table 4.1: Key containing model names and training process descriptions.

4.1 Unclipped Private Data

Confusion matrices and classification metric averages for models finetuned with unclipped private data indicate a slight increase in performance over a random guesser. Although some accuracy curves tend to fluctuate around chance-level accuracy, some models, like P0, P1, P2, and D2P1 show a general upward trend. Models with unfrozen layers, however, show clear signs of overfitting. Unfreezing one layer before finetuning on private data resulted in steady increases in evaluation loss, and unfreezing two layers resulted in even steeper increases, despite greater decreases in training loss. Yet, there still seems to be potential for these models as the highest metric averages also come from models with more unfrozen layers, like P2 and D2P1. This suggests that taking further measures to prevent overfitting could yield improved results that better represent the capabilities and benefits of finetuning more transformer layers over the classifier head alone.

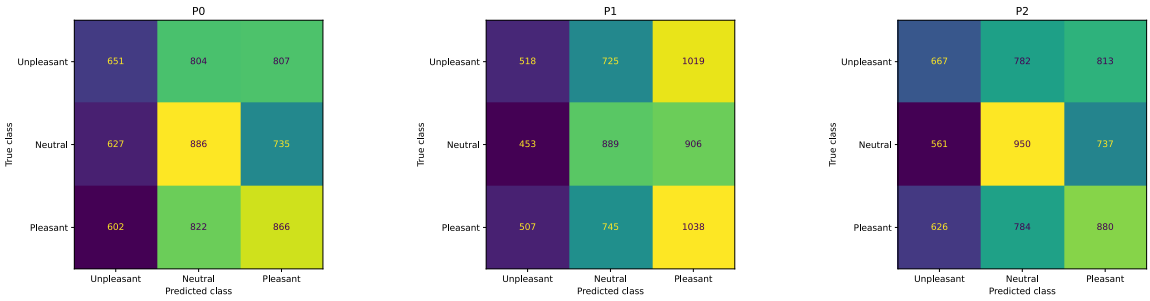


Figure 4.1: Confusion matrices for baseline models finetuned with unclipped private data alone.

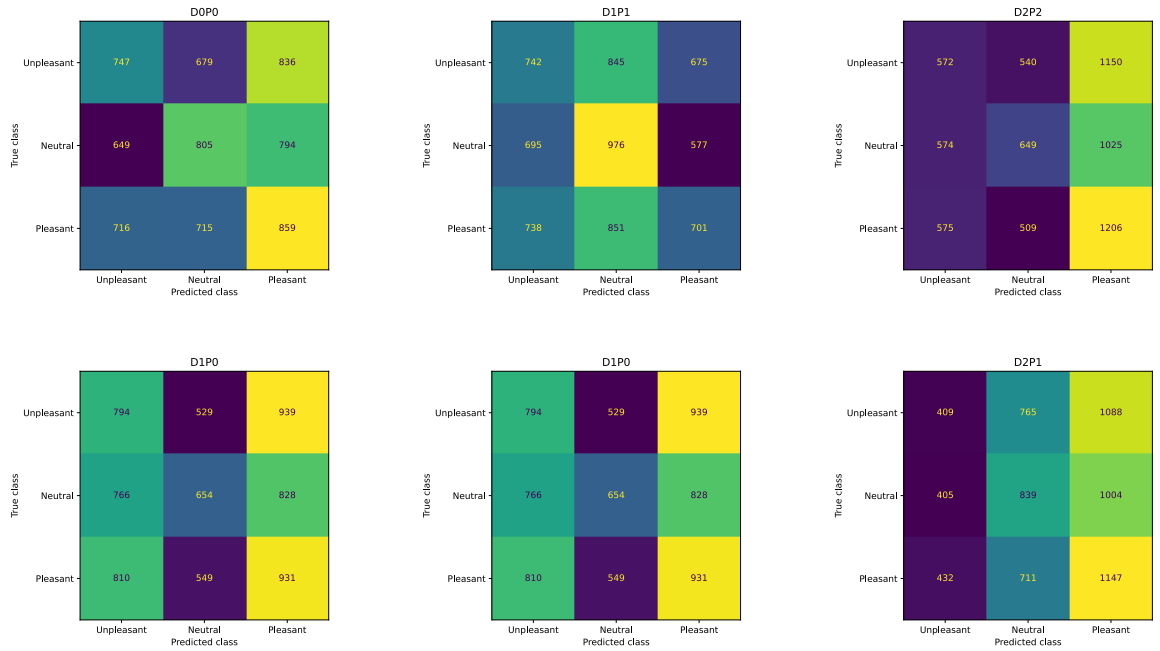


Figure 4.2: Confusion matrices for models finetuned with both DEAP and unclipped private data.

Model	**Accuracy**	**F1**	**Recall**	**Precision**	**ROC AUC**
P0	0.3534	0.3412	0.3534	0.3535	0.5008
P1	0.3596	0.3212	0.3592	0.3542	0.5128
P2	**0.3672**	**0.3573**	**0.3673**	**0.3674**	**0.5282**
D0P0	0.3546	0.3453	0.3545	0.3537	0.5086
D1P1	0.3557	0.3342	0.3560	0.3537	0.5086
D2P2	0.3569	0.3246	0.3560	0.3581	0.5181
D1P0	0.3499	0.3461	0.3495	0.3526	0.5096
D2P0	0.3444	0.3379	0.3439	0.3445	0.5028
D2P1	0.3522	0.3249	0.3517	0.3517	0.5150

Table 4.2: Average classification metrics for models finetuned with unclipped private data.

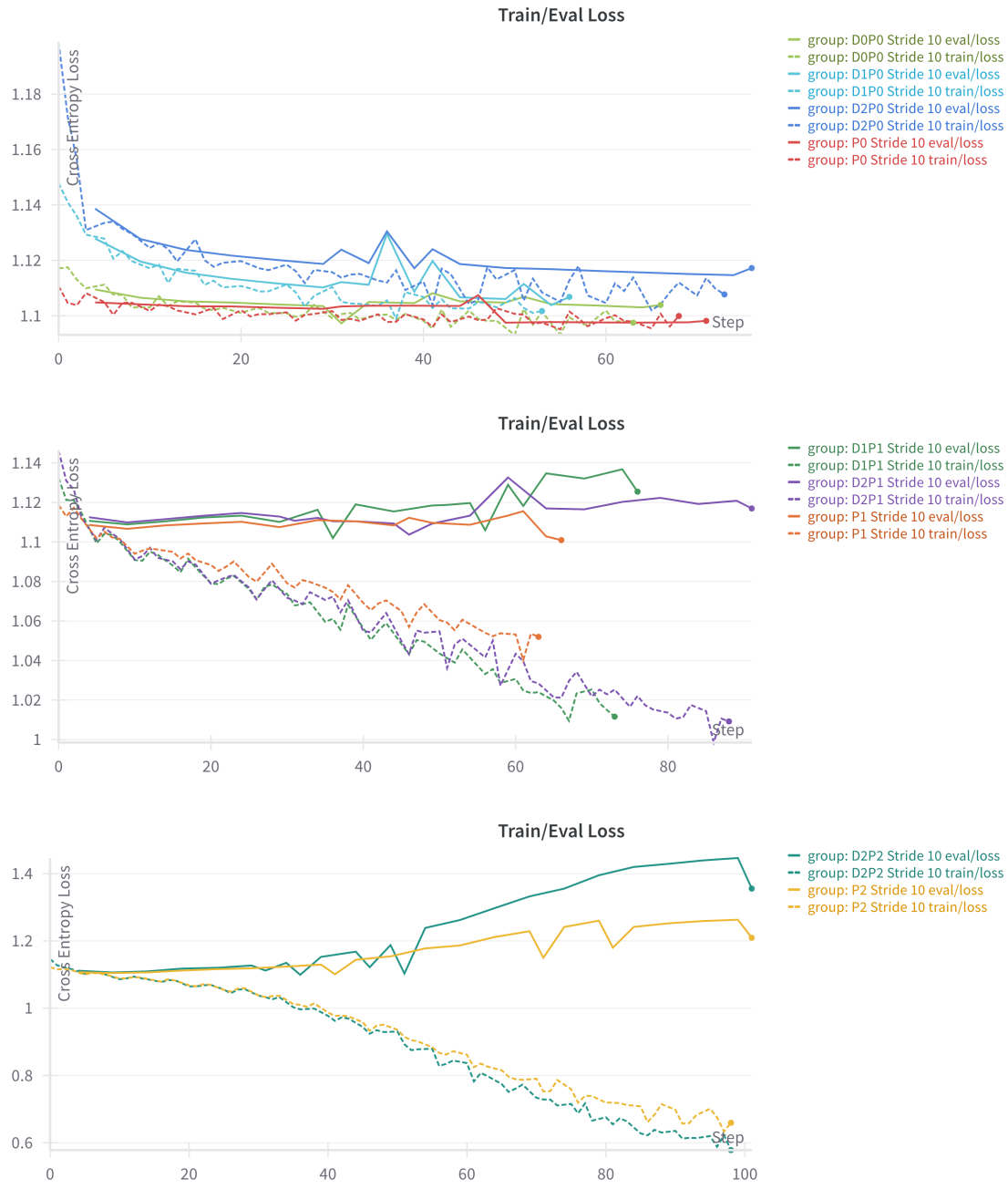


Figure 4.3: Training and evaluation loss curves for models finetuned with unclipped private data.

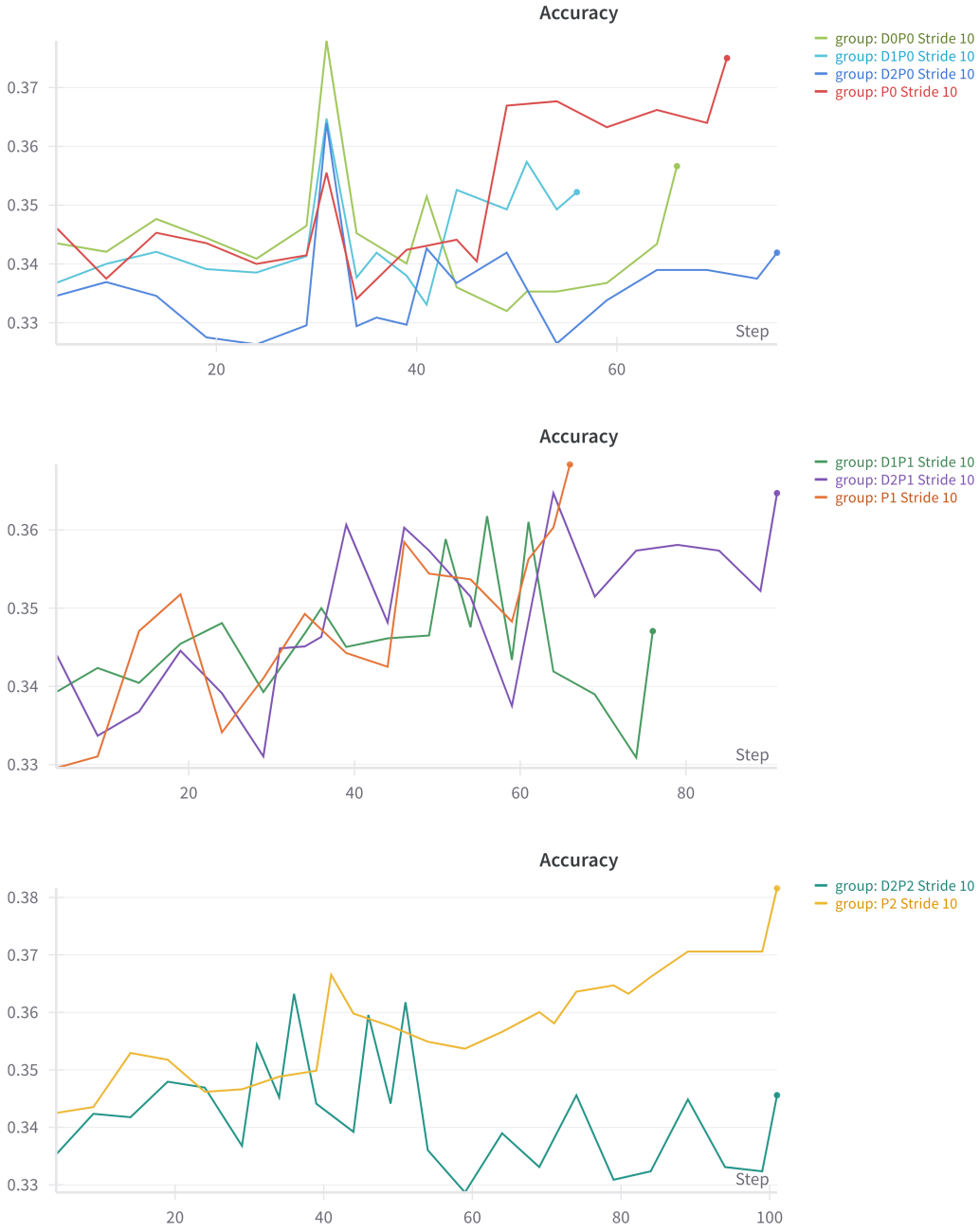


Figure 4.4: Accuracy curves for models finetuned with unclipped private data.

4.2 Clipped Private Data

Finetuning on 2s clipped private data resulted in a further improvement in performance over unclipped data. Most accuracy curves show steadier increases above chance-level accuracy with more consistent upward trends. Furthermore, almost all models finetuned on both datasets outperform their baseline counterparts, unlike models trained on unclipped data. That is, with clipped private data, D0P0, D1P0, and D2P0 all outperform P0, and D2P1 and D1P1 both outperform P1. However, overfitting is still an issue. Overfitting follows a similar pattern as before, with more unfrozen layers resulting in higher evaluation loss curves. D2P2, the only model to underperform its corresponding baseline, P2, begins with higher accuracies than the baseline, but it quickly falters as the model overfits the data with a substantial increase in evaluation loss. As with the unclipped data, taking measures to prevent overfitting in the future may further improve performance in a way that better represents the capabilities of these models in emotion recognition. Nonetheless, because most cases show that finetuning with both public and private data results in either comparable or improved performance over baselines finetuned with private data alone, the proposed method effectively allows for model training across multiple datasets with varying sets of EEG channels.

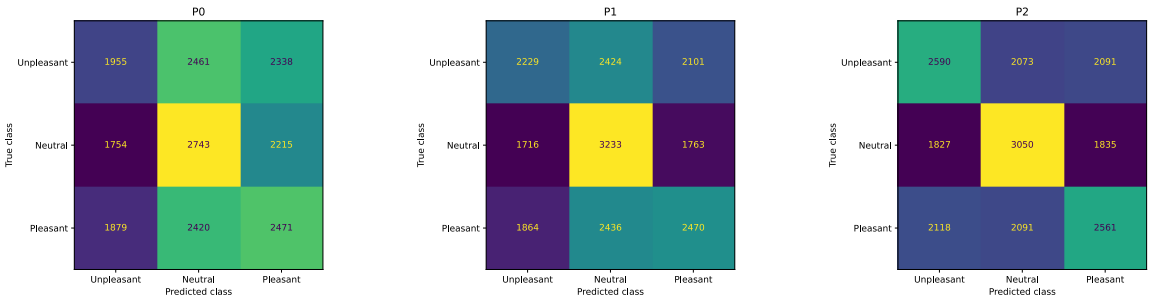


Figure 4.5: Confusion matrices for baseline models finetuned with clipped private data alone.

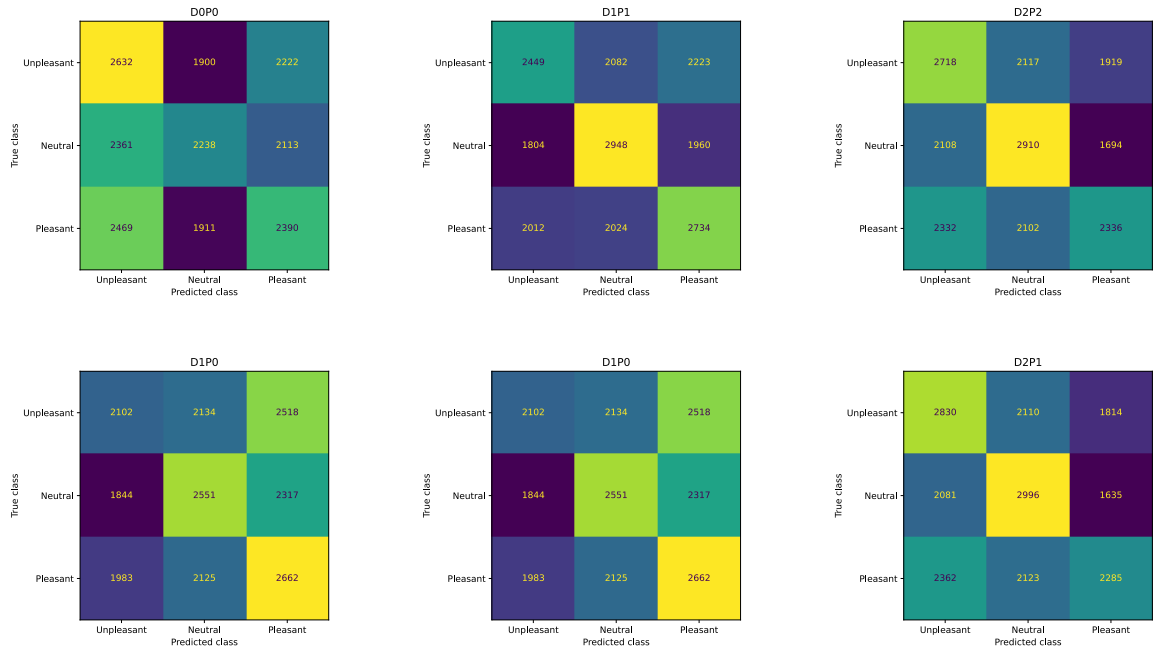


Figure 4.6: Confusion matrices for models finetuned with both DEAP and clipped private data.

Model	**Accuracy**	**F1**	**Recall**	**Precision**	**ROC AUC**
P0	0.3543	0.3478	0.3544	0.3536	0.5224
P1	0.3920	0.3872	0.3922	0.3924	0.5668
P2	**0.4053**	**0.4044**	**0.4054**	**0.4049**	**0.5774**
D0P0	0.3588	0.3537	0.3587	0.3596	0.5232
D1P1	0.4018	0.3987	0.4019	0.4022	0.5761
D2P2	0.3936	0.3860	0.3937	0.3968	0.5677
D1P0	0.3615	0.3569	0.3615	0.3610	0.5260
D2P0	0.3543	0.3540	0.3543	0.3543	0.5176
D2P1	0.4008	0.3980	0.4010	0.4019	0.5765

Table 4.3: Average classification metrics for models finetuned with clipped private data.

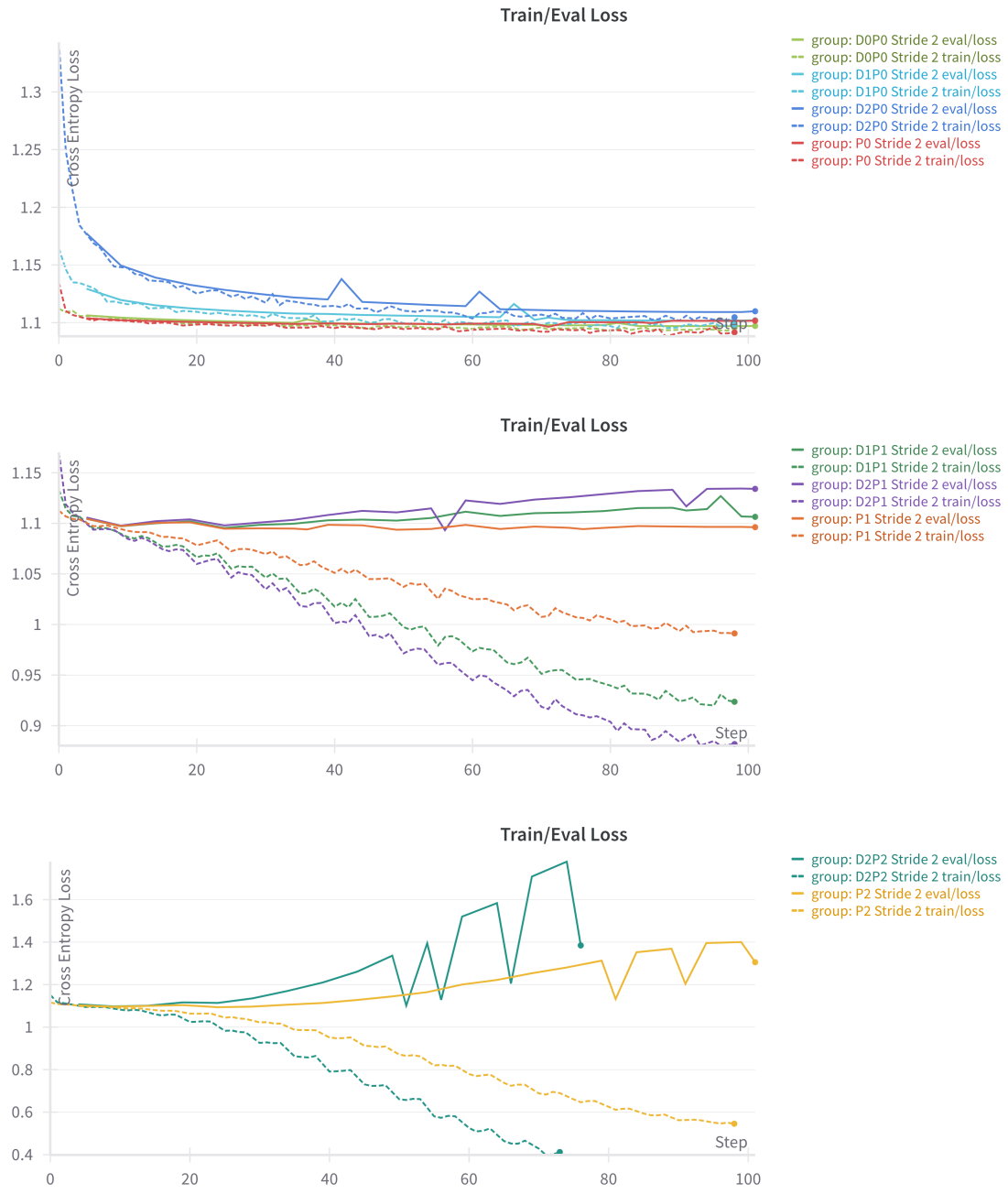


Figure 4.7: Training and evaluation loss curves for models finetuned with clipped private data.

Figure 4.8: Accuracy curves for models finetuned with clipped private data.

Chapter 5

Discussion

Classification metrics from all cross-validation experiments show that these models are far from being able to consistently recognize emotions from real-world data. When compared to valence accuracies from table 2.2, the current models fall significantly behind previous vision transformer based models trained on both public [48] and private datasets [49]. Although some of the current highest-performing models, like P2 and D2P1, show the most potential for future improvement with adjustments in training procedures, they are still only slightly better than random guessers with ROC AUC scores of 0.5774 and 0.5765.

The main factor holding these models back seems to be the issue of overfitting, likely caused by insufficient data. 38,400 public instances and 20,236 private instances was seemingly enough for finetuning the classifier head alone, as loss curves for P0 models show an expected decay during training, but the demand for more data grew with each unfrozen transformer layer. P1 and P2 models were able to obtain higher average classification metrics, but not without increasingly higher validation loss curves as well. Unfreezing more transformer layers has greater potential for improved performance, but it must be supported with overfitting countermeasures in order for models to reach this potential.

Where these findings show promise is in the fact that models benefited from finetuning on multiple datasets with varying sets of EEG channels. Had the proposed method of reordering channels with Hilbert curves failed, models would have essentially restarted the learning process when presented with new instances from the private dataset; if the second phase of finetuning with private data featured a new arrangement of frame regions representing different areas of the brain, models would not be able to build upon the patterns learned

in the first phase of finetuning with public data. In this case, because starting from a pretrained model trained with a generalized dataset, like ViViT with Kinetics-400, would be an advantageous starting point compared to a model pretrained on data arranged in the wrong order, we would expect models to perform worse than their baselines.

Since finetuning with unclipped instances resulted in metrics comparable to baselines, and finetuning with clipped instances resulted in improved metrics, the proposed method allows for models to build upon patterns learned across multiple datasets. This approach of arranging CWT features in a video format and ordering EEG channels with Hilbert curves should then be seen as a starting point for future research. Now that such an approach has been established, future work can iterate on and improve this process to realize the full potential of 3D vision transformers in emotion recognition.

5.1 Implications

The findings of this work hold promise for improving 3D vision transformers in emotion recognition, which in turn may benefit certain efforts in human-computer interaction, medicine, and neuroscience. The main benefit of the current method is in its capability to train models on features extracted across multiple datasets. This removes the restriction of training with less data without sacrificing spatial continuity between datasets, which will further encourage models to learn more generalizable patterns of emotion rather than relying on the specific characteristics of a single dataset. This is especially valuable for transformer-based models, which are heavily dependent on these spatial relationships.

The advantages of using 3D vision transformers in emotion recognition are similar to the advantages of using vision transformers over previous CNN-transformer architectures. Although vision transformers require more training data, they do not fall victim to some of the biases inherent to CNNs, like translation equivariance and locality [18]. Vision transformers have also been shown to exceed previous state of the art CNN models in image classification tasks while keeping computational costs relatively low [18]. Vision transformers

may not be as effective at smaller scales, but this could be accounted for by training across many public datasets with the proposed method.

Consequently, these aspects also help in the overall advancement of emotion recognition, which could impact multiple different fields. For human-computer dialogue systems, recognizing the emotional state behind the user's spoken words could lead to more appropriate and humanized responses that improve the conversation experience [55]. In medicine, previous models have utilized speech responses to emotion-eliciting videos to detect mood disorders like unipolar depression and bipolar disorder [56]. Using similar methods with brain responses, like those captured in EEG data, circumvents the issue of reliability that comes with non-physiological data [3]. Finally, in neuroscience, a consistent and accurate emotion recognition model could be used to help label the true emotions elicited by unique or ambiguous stimuli. The current work supports these ends by exploring a new model architecture and feature extraction process in emotion recognition.

5.2 Conclusion

This paper demonstrated a new approach to emotion recognition from EEG data that utilized 3D vision transformers and CWT features extracted from both public and private datasets. The proposed method arranges CWT features in a video format and ensures continuity between videos extracted from datasets with varying sets of EEG channels. When trained on unclipped private data, models finetuned on both datasets achieved comparable results to baseline models finetuned on public data alone, which indicates that this method enables training across multiple datasets without negatively impacting performance. Results also show that clipping instances can improve model performance when compared to models trained on unclipped instances. When trained on clipped instances specifically, most models outperform their respective baselines, which further highlights the benefit of finetuning with public data before private data.

5.3 Future Work

Future research should first address the issue of overfitting to further evaluate the potential capabilities of 3D vision transformers in emotion recognition. This can be done by increasing dataset size with data augmentation, using larger pretrained video vision transformers with more parameters, or adjusting training hyperparameters to promote better generalization. From there, other topics of interest might include comparing multiple pretrained models, or finetuning with multiple public datasets.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. ukasz Kaiser, and I. Polosukhin, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017.

[2] J. Li, W. Pan, H. Huang, J. Pan, and F. Wang, "STGATE: Spatial-temporal graph attention network with a transformer encoder for EEG-based emotion recognition," *Frontiers in Human Neuroscience*, Apr. 2023.

[3] W. Lu, T.-P. Tan, and H. Ma, "Bi-Branch Vision Transformer Network for EEG Emotion Recognition," *IEEE Access*, vol. 11, pp. 36233–36243, 2023.

[4] X. Zhong, Y. Gu, Y. Luo, X. Zeng, and G. Liu, "Bi-hemisphere asymmetric attention network: Recognizing emotion from EEG signals based on the transformer," *Applied Intelligence*, vol. 53, pp. 15278–15294, June 2023.

[5] L. Gong, M. Li, T. Zhang, and W. Chen, "EEG emotion recognition using attention-based convolutional transformer neural network," *Biomedical Signal Processing and Control*, vol. 84, p. 104835, July 2023.

[6] W. Zheng and B. Pan, "A spatiotemporal symmetrical transformer structure for EEG emotion recognition," *Biomedical Signal Processing and Control*, vol. 87, p. 105487, Jan. 2024.

[7] J.-Y. Guo, Q. Cai, J.-P. An, P.-Y. Chen, C. Ma, J.-H. Wan, and Z.-K. Gao, "A Transformer based neural network for emotion recognition and visualizations of crucial EEG channels," *Physica A: Statistical Mechanics and its Applications*, vol. 603, p. 127700, Oct. 2022.

[8] S. Sartipi and M. Cetin, "Adversarial Discriminative Domain Adaptation and Transformers for EEG-based Cross-Subject Emotion Recognition," in *2023 11th International IEEE/EMBS Conference on Neural Engineering (NER)*, pp. 1–4, Apr. 2023.

[9] J. A. Russell, "Affective space is bipolar," *Journal of Personality and Social Psychology*, vol. 37, no. 3, pp. 345–356, 1979.

[10] X. Li, Y. Zhang, P. Tiwari, D. Song, B. Hu, M. Yang, Z. Zhao, N. Kumar, and P. Marttinen, "EEG Based Emotion Recognition: A Tutorial and Review," *ACM Computing Surveys*, vol. 55, pp. 1–57, Apr. 2023.

[11] S. Langeslag, "Effects of organization and disorganization on pleasantness, calmness, and the frontal negativity in the event-related potential," *PLOS ONE*, vol. 13, p. e0202726, Aug. 2018.

[12] D. G. Graetzer, "Electroencephalography (EEG).," *Magill's Medical Guide (Online Edition)*, Sept. 2023.

[13] S. M. Alarcão and M. J. Fonseca, "Emotions Recognition Using EEG Signals: A Survey," *IEEE Transactions on Affective Computing*, vol. 10, pp. 374–393, July 2019.

[14] Z. Bai, F. Hou, K. Sun, Q. Wu, M. Zhu, Z. Mao, Y. Song, and Q. Gao, "SECT: A Method of Shifted EEG Channel Transformer for Emotion Recognition," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, pp. 4758–4767, Oct. 2023.

[15] R.-N. Duan, J.-Y. Zhu, and B.-L. Lu, "Differential entropy feature for EEG-based emotion classification," in *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*, pp. 81–84, Nov. 2013.

[16] E. Niedermeyer and F. L. Da Silva, "Electroencephalography, basic principles, clinical applications, and related fields," Jan. 1982.

[17] O. Dressler, G. Schneider, G. Stockmanns, and E. F. Kochs, "Awareness and the EEG power spectrum: Analysis of frequencies," *BJA: British Journal of Anaesthesia*, vol. 93, pp. 806–809, Dec. 2004.

[18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale."

[19] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "ViViT: A Video Vision Transformer," 2021.

[20] Y. Liu, O. Sourina, and M. K. Nguyen, "Real-Time EEG-Based Human Emotion Recognition and Visualization," in *2010 International Conference on Cyberworlds*, pp. 262–269, Oct. 2010.

[21] Y. Guo, B. Zhang, X. Fan, X. Shen, and X. Peng, "A Comprehensive Interaction in Multiscale Multichannel EEG Signals for Emotion Recognition," *Mathematics*, vol. 12, p. 1180, Apr. 2024.

[22] J. Atkinson and D. Campos, "Improving BCI-based emotion recognition by combining EEG feature selection and kernel classifiers," *Expert Systems with Applications*, vol. 47, pp. 35–41, Apr. 2016.

[23] J. Liu, H. Meng, A. Nandi, and M. Li, "Emotion detection from EEG recordings," in *2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, pp. 1722–1727, Aug. 2016.

[24] M. Li, H. Xu, X. Liu, and S. Lu, "Emotion recognition from multichannel EEG signals using K-nearest neighbor classification," *Technology and Health Care: Official Journal of the European Society for Engineering and Medicine*, vol. 26, no. S1, pp. 509–519, 2018.

[25] X. Li, D. Song, P. Zhang, G. Yu, Y. Hou, and B. Hu, "Emotion recognition from multi-channel EEG data through Convolutional Recurrent Neural Network," in *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 352–359, Dec. 2016.

[26] S. Alhagry, A. A. Fahmy, and R. A. El-Khoribi, "Emotion Recognition based on EEG using LSTM Recurrent Neural Network," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 8, no. 10, 2017/11/31.

[27] S. Hwang, K. Hong, G. Son, and H. Byun, "Learning CNN features from DE features for EEG-based emotion recognition," *Pattern Analysis and Applications*, vol. 23, pp. 1323–1335, Aug. 2020.

[28] Y. Xu, Y. Du, L. Li, H. Lai, J. Zou, T. Zhou, L. Xiao, L. Liu, and P. Ma, "AMDET: Attention based Multiple Dimensions EEG Transformer for Emotion Recognition," *IEEE Transactions on Affective Computing*, pp. 1–11, 2023.

[29] M. Asif, A. Gupta, A. Aditya, S. Mishra, and U. S. Tiwary, "Brain Multi-Region Information Fusion using Attentional Transformer for EEG Based Affective Computing," in *2023 IEEE 20th India Council International Conference (INDICON)*, pp. 771–775, Dec. 2023.

[30] A. Khare, S. Parthasarathy, and S. Sundaram, "Self-Supervised Learning with Cross-Modal Transformers for Emotion Recognition," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, pp. 381–388, Jan. 2021.

[31] M. Sun, W. Cui, S. Yu, H. Han, B. Hu, and Y. Li, "A Dual-Branch Dynamic Graph Convolution Based Adaptive TransFormer Feature Fusion Network for EEG Emotion Recognition," *IEEE Transactions on Affective Computing*, vol. 13, pp. 2218–2228, Oct. 2022.

[32] G. Cao, L. Yang, C. Tang, Q. Zhang, and H. He, "BiCCT: A Compact Convolutional Transformer for EEG Emotion Recognition," in *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 4792–4799, Dec. 2023.

[33] J. Sun, X. Wang, K. Zhao, S. Hao, and T. Wang, "Multi-Channel EEG Emotion Recognition Based on Parallel Transformer and 3D-Convolutional Neural Network," *Mathematics*, vol. 10, no. 17, p. 3131, 2022.

[34] C. Cheng, Y. Zhang, L. Liu, W. Liu, and L. Feng, "Multi-Domain Encoding of Spatiotemporal Dynamics in EEG for Emotion Recognition," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, pp. 1342–1353, Mar. 2023.

[35] H. Sun, L. Yang, Q. Wang, D. Liu, and P. Ni, "MEEG-Transformer: Transformer Network based on Multi-domain EEG for Emotion Recognition," in *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, (Istanbul, Turkiye), pp. 3332–3339, IEEE, Dec. 2023.

[36] X. Yao, T. Li, P. Ding, F. Wang, L. Zhao, A. Gong, W. Nan, and Y. Fu, "Emotion Classification Based on Transformer and CNN for EEG Spatial–Temporal Feature Learning," *Brain Sciences*, vol. 14, p. 268, Mar. 2024.

[37] R. Liu, Y. Chao, X. Ma, X. Sha, L. Sun, S. Li, and S. Chang, "ERTNet: An interpretable transformer-based framework for EEG emotion recognition," *Frontiers in Neuroscience*, vol. 18, p. 1320645, Jan. 2024.

[38] X. Si, D. Huang, Y. Sun, and D. Ming, "Temporal Aware Mixed Attention-based Convolution and Transformer Network (MACTN) for EEG Emotion Recognition," May 2023.

[39] Y. Du, H. Ding, M. Wu, F. Chen, and Z. Cai, "MES-CTNet: A Novel Capsule Transformer Network Base on a Multi-Domain Feature Map for Electroencephalogram-Based Emotion Recognition," *Brain Sciences*, vol. 14, no. 4, p. 344, 2024.

[40] S. Ke, C. Ma, W. Li, J. Lv, and L. Zou, "Multi-Region and Multi-Band Electroencephalogram Emotion Recognition Based on Self-Attention and Capsule Network," *Applied Sciences*, vol. 14, no. 2, p. 702, 2024.

[41] Y. Wei, Y. Liu, C. Li, J. Cheng, R. Song, and X. Chen, "TC-Net: A Transformer Capsule Network for EEG-based emotion recognition," *Computers in Biology and Medicine*, vol. 152, p. 106463, Jan. 2023.

[42] Z. Wang, Y. Wang, X. Wan, and Y. Tang, "Cerebral asymmetry representation learning-based deep subdomain adaptation network for electroencephalogram-based emotion recognition," *Physiological Measurement*, vol. 45, p. 035004, Mar. 2024.

[43] Y. Liu, Y. Zhou, and D. Zhang, "TcT: Temporal and channel Transformer for EEG-based Emotion Recognition," in *2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 366–371, July 2022.

[44] G. Peng, K. Zhao, H. Zhang, D. Xu, and X. Kong, "Temporal relative transformer encoding cooperating with channel attention for EEG emotion analysis," *Computers in Biology and Medicine*, vol. 154, p. 106537, Mar. 2023.

[45] Z. Wang, Y. Wang, C. Hu, Z. Yin, and Y. Song, "Temporal-spatial Representation Learning Transformer for EEG-based Emotion Recognition," Nov. 2022.

[46] Z. Guo, J. Wang, B. Zhang, Y. Ku, and F. Ma, "A dual transfer learning method based on 3D-CNN and vision transformer for emotion recognition," vol. 55, no. 3, p. 200.

[47] A. W. Awan, I. Taj, S. Khalid, M. U. Syed, A. S. Imran, M. Usman Akram, and U. A. Muhammad, "Advancing Emotional Health Assessments: A Hybrid Deep Learning Approach Using Physiological Signals for Robust Emotion Recognition," vol. 12, pp. 141890–141904.

[48] A. Arjun, A. S. Rajpoot, and M. Raveendranatha Panicker, "Introducing Attention Mechanism for EEG Signals: Emotion Recognition with Vision Transformers," in *2021*

43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), (Mexico), pp. 5723–5726, IEEE, Nov. 2021.

[49] D. Wang, J. Lian, H. Cheng, and Y. Zhou, "Music-evoked emotions classification using vision transformer in EEG signals," vol. 15, p. 1275142.

[50] J. Liu, H. Wu, L. Zhang, and Y. Zhao, "Spatial-temporal Transformers for EEG Emotion Recognition," Sept. 2022.

[51] S. Koelstra, C. Muhl, M. Soleymani, Jong-Seok Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A Database for Emotion Analysis Using Physiological Signals," *IEEE Transactions on Affective Computing*, vol. 3, pp. 18–31, Jan. 2012.

[52] A. H. Farkas, M. C. Gehr, H. Jia, and D. Sabatinelli, "Measuring Realistic Emotional Perception With EEG: A Comparison of Multimodal Videos and Naturalistic Scenes," vol. 62, no. 1, p. e14765.

[53] D. Sabatinelli, A. H. Farkas, and M. C. Gehr, "Moving toward reality: Electrocortical reactivity to naturalistic multimodal emotional videos," vol. 61, no. 6, p. e14526.

[54] L. Arts and E. van den Broek, "The fast continuous wavelet transformation (fCWT) for real-time, high-quality, noise-resistant time–frequency analysis.," *Nat Comput Sci*, vol. 2, no. 1, pp. 47–58, 2022.

[55] N. Hajarolasvadi and H. Demirel, "3D CNN-Based Speech Emotion Recognition Using K-Means Clustering and Spectrograms," *Entropy*, vol. 21, p. 479, May 2019.

[56] K.-Y. Huang, C.-H. Wu, and M.-H. Su, "Attention-based convolutional neural network and long short-term memory for short-term detection of mood disorders based on elicited speech responses," *Pattern Recognition*, vol. 88, pp. 668–678, Apr. 2019.