ADAPTIVE DEEP CANONICAL CORRELATION ANALYSIS–BASED MULTIMODAL SENTIMENT ANALYSIS

by

YUNHONG LIAO

A thesis submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN ELECTRICAL AND COMPUTER ENGINEERING

2025

Oakland University
Rochester, Michigan

Master Advisory Committee:

Jia Li, Ph.D., Chair
Wing-Yue Geoffrey Louie, Ph.D.
Hongwei Qu, Ph.D.

To my family

ACKNOWLEDGMENTS

My journey at Oakland University began in the summer of 2024, and this thesis reaches completion in the fall of 2025. I am thrilled to continue my PhD studies at Oakland University and look forward to creating many more wonderful memories in the years ahead.

I am deeply grateful to everyone who has helped me along the way. First and foremost, my sincere thanks go to my supervisor, Dr. Jia Li, who truly opened the door to research for me. Her patient guidance, steady encouragement, and wide-ranging expertise have been invaluable, instilling in me both enthusiasm and confidence for scientific exploration. I am also thankful to Dr. Lianxiang Yang and Dr. Gary Barber for their dedication to the exchange program, which afforded me a life-changing opportunity. Likewise, I extend my gratitude to my thesis committee members, Dr. Wing-Yue Geoffrey Louie and Dr. Hongwei Qu, for their support of my research. I would also like to thank Dr. Steven Louis for providing the LaTeX template that made this thesis possible. At last but not least, I would like to acknowledge the technical and financial support of the Automotive Research Center (ARC) in accordance with Cooperative Agreement W56HZV-19-2-0001 U.S. Army DEVCOM Ground Vehicle Systems Center (GVSC) Warren, MI. To my friends who have accompanied me on this journey—thank you for your companionship and support through both the challenging and the joyful moments.

Finally, my deepest appreciation goes to my family. Your unwavering support has allowed me to stand on your shoulders and see the world. Thank you for always being there without complaint, for your quiet support behind the scenes, and for respecting my decisions. I could not have accomplished this without you.

See you all when the maple leaves have turned red four more times.

DISTRIBUTION STATEMENT A. Approved for public release; distribution is unlimited.
OPSEC10166.

Yunhong Liao

ABSTRACT

ADAPTIVE DEEP CANONICAL CORRELATION ANALYSIS–BASED MULTIMODAL SENTIMENT ANALYSIS

by

Yunhong Liao

Adviser: Jia Li, Ph.D.

Emotion recognition plays a critical role in affective computing and human-computer interaction. While physiological signals such as electroencephalography (EEG) and eye-tracking offer valuable insights into emotional states, effectively fusing these heterogeneous modalities remains challenging due to differences in temporal scale, dimensionality, and signal characteristics. Traditional fusion methods employ fixed strategies that fail to adapt to dynamic changes in modality reliability and cross-subject variability, limiting their practical applicability.

This thesis presents two novel adaptive multimodal fusion frameworks for emotion recognition: Adaptive Deep Canonical Correlation Analysis (ADCCA) and Adaptive Cross-Modal Deep CCA Transformer (ACDCT). ADCCA extends traditional Deep CCA by incorporating an attention-based adaptive weighting mechanism that dynamically adjusts the contribution of EEG and eye-tracking modalities based on their instantaneous reliability. Building upon ADCCA, ACDCT introduces a Transformer-based architecture with cross-modal attention at the token level, enabling fine-grained temporal interactions between modalities. ACDCT further employs layer-wise DCCA regularization to maintain feature alignment across network depth, enhancing cross-subject generalization.

We evaluate both methods on SEED-IV (four-class) and SEED-V (five-class) datasets under subject-dependent (SD) and subject-independent (SI) protocols. ADCCA

achieves 87.5% (SD) and 70.7% (SI) accuracy on SEED-IV, and 88.4% (SD) and 55.2% (SI) on SEED-V, outperforming traditional fusion baselines. ACDCT demonstrates substantial improvements, achieving 92.6% (SD) and 77.6% (SI) on SEED-IV, and 90.4% (SD) and 75.3% (SI) on SEED-V. Most notably, ACDCT improves SI performance on SEED-V by 20.1 percentage points over ADCCA, transforming near-chance performance to highly competitive accuracy. Both methods exhibit 40-60% lower standard deviations compared to baselines, indicating substantially more stable performance.

The results demonstrate that adaptive fusion mechanisms and cross-modal temporal interactions are essential for effective multimodal emotion recognition. ACDCT's token-level bidirectional attention enables the model to discover fine-grained cross-modal dependencies that coarse fusion strategies miss, particularly benefiting cross-subject generalization where learning abstract, shared emotional patterns is critical. This work advances multimodal affective computing by providing interpretable, high-performance frameworks that address fundamental challenges in physiological signal fusion and cross-subject variability.

TABLE OF CONTENTS

TABLE OF CONTENTS—Continued

TABLE OF CONTENTS—Continued

TABLE OF CONTENTS—Continued

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

Introduction

1.1 Background and Significance

Emotions play a pivotal role in human decision-making and behavior [1, 2]. The field of *affective computing* was founded to enable computational systems to recognize and respond to human emotions, thereby making human–computer interactions more natural and effective [3]. Accurate emotion recognition is critical in diverse applications, including clinical therapy, intelligent tutoring systems, and safety-critical operations [4, 5].

Early work in affective computing focused on observable behavioral cues such as facial expressions, speech intonation, and text sentiment. However, unimodal approaches often prove fragile: emotional expression is multi-faceted, and single-modality systems fail under conditions such as occluded faces, noisy audio, or environmental variations. Furthermore, external behaviors can be consciously controlled or masked, whereas internal physiological signals are less susceptible to deliberate manipulation. These limitations have driven interest in multimodal emotion recognition approaches that integrate multiple information channels to capture a more complete representation of affective states.

Among physiological signals, electroencephalography (EEG) and eye-tracking have emerged as particularly valuable and complementary modalities for emotion recognition [6–8]. EEG measures brain electrical activity with millisecond temporal resolution, directly capturing neural dynamics associated with affective processes such as changes in frequency-band power during emotional arousal [9, 10]. While EEG provides rich temporal information, it suffers from susceptibility to artifacts (eye blinks, muscle movements) and relatively low spatial resolution [11, 12]. Eye-tracking, conversely,

records gaze patterns, saccades, blinks, and pupil dilation—behavioral indicators of arousal, cognitive load, and attention [13, 14]. Eye-tracking data are typically low-dimensional with high signal-to-noise ratios under controlled conditions, but their effectiveness depends on stimulus design and they provide limited direct information about emotional valence.

By combining these modalities, we can leverage their complementary strengths: EEG captures implicit neural responses, while eye metrics provide behavioral context. This fusion potentially achieves more reliable emotion recognition than either modality alone [15–18].

However, fusing EEG and eye-tracking data presents significant challenges. The modalities differ fundamentally in sampling rate (EEG typically hundreds of Hz versus eye-tracking tens of Hz), dimensionality (dozens of EEG channels versus a few eye features), semantics (neural activity versus behavioral metrics), and noise characteristics. Naive fusion strategies—such as feature concatenation or fixed weighting—rarely exploit the full complementarity of the signals and may amplify individual weaknesses. For instance, when one modality becomes temporarily uninformative (e.g., EEG contaminated by motion artifacts or eye-tracking loss during blinks), static fusion continues to weight it equally, degrading overall performance.

Prior studies have shown that fixed fusion rules fail to accommodate the dynamic nature of real-world multimodal data [19, 20]. This motivates adaptive fusion mechanisms that adjust each modality's contribution based on instantaneous reliability and relevance. Such quality-aware techniques are particularly critical for physiological signals, where data quality fluctuates due to electrode contact, participant fatigue, and environmental factors.

A second major challenge is cross-subject variability. Models trained and tested on the same individual (*subject-dependent* scenario) can achieve high accuracy by exploiting

person-specific signal patterns. However, performance often drops dramatically when applied to new subjects (*subject-independent* scenario) due to differences in EEG baselines, eye movement habits, and emotional response patterns [21, 22]. Bridging this gap is crucial for real-world deployment, requiring representations that capture emotion-relevant features while remaining invariant to individual-specific characteristics—a difficult task given the high-dimensional, idiosyncratic nature of physiological signals.

This thesis addresses these challenges through two strategic approaches. First, we develop adaptive, quality-aware fusion mechanisms that dynamically adjust the contribution of EEG versus eye-tracking based on signal reliability and context. Rather than fixed fusion rules, we employ learnable gating modules that shift the modality balance according to data quality—for instance, emphasizing eye-tracking when EEG is artifact-contaminated, and vice versa. Second, we introduce cross-modal temporal interactions that enable fine-grained information exchange across modalities over time. This allows signals from one modality at a given moment to influence the interpretation of the other modality at subsequent time steps, potentially uncovering temporal correlations and dependencies that static fusion would miss. We hypothesize that this structured cross-modal attention will particularly benefit subject-independent generalization by learning more abstract, shared patterns of emotional response.

To support these objectives, we employ Deep Canonical Correlation Analysis (DCCA) [23] as a foundational component. DCCA projects the two modalities into a common latent space where their representations are maximally correlated, serving as a feature alignment tool that reduces modality-specific disparities and facilitates subsequent fusion. We evaluate our methods on two public multimodal emotion datasets, SEED-IV and SEED-V [24, 25], under both subject-dependent and subject-independent protocols to rigorously assess generalization capability.

1.2 Research Questions

This thesis investigates the following research questions:

1. Adaptive Fusion: Can learned, quality-aware gating mechanisms outperform traditional fixed-weight fusion strategies for EEG and eye-tracking emotion recognition by dynamically adjusting modality contributions based on signal reliability and context?

2. Cross-Modal Temporal Interaction: Does incorporating fine-grained cross-modal attention in a shared latent space improve generalization in subject-independent settings by enabling temporal information exchange between modalities that captures complementary evidence beyond scalar fusion?

The first question addresses the need for fusion strategies that adapt to non-stationary data reliability. The second question examines whether deep temporal modeling of cross-modal interactions enhances robustness and generalization. Together, these questions guide the development of an emotion recognition system suitable for real-world deployment.

1.3 Research Objectives and Contributions

The primary objective of this research is to develop a multimodal emotion recognition framework that achieves high accuracy in subject-dependent scenarios, significantly improves generalization in subject-independent scenarios, and provides interpretable decision-making without relying on manually tuned fusion heuristics. We pursue this objective through two progressive methodological contributions:

(1) Adaptive Deep Canonical Correlation Analysis (ADCCA): We design an adaptive fusion framework employing modality-specific deep encoders that project EEG and eye-tracking features into a DCCA-aligned latent space. Learnable gating networks

produce quality-aware weights that modulate each modality's contribution to the final prediction, performing soft selection of the more informative modality at each moment. ADCCA directly addresses the adaptive fusion research question by quantifying the benefits of data-driven gating over fixed fusion and revealing conditions under which adaptivity yields the greatest gains.

(2) Adaptive Cross-Modal Deep CCA Transformer (ACDCT): We extend ADCCA by introducing Transformer-based cross-modal attention [26, 27] within the correlation-aligned space. ACDCT adds layers of intra- and inter-modal attention that enable continuous interaction between EEG and eye-tracking streams over temporal windows. The Transformer architecture captures long-range dependencies via self-attention and enables cross-modal information routing via cross-attention. By exchanging information at the token (time-step) level, ACDCT can discover fine-grained temporal alignments—such as associations between EEG activity spikes and corresponding eye movement patterns—that coarse scalar fusion would miss. We hypothesize that this structured interaction strengthens generalization, particularly in subject-independent scenarios where learning abstract, shared emotional patterns is essential.

We evaluate both models under identical experimental conditions using the same datasets (SEED-IV and SEED-V), preprocessing pipelines, and performance metrics to ensure fair comparison. Systematic ablation studies isolate the contribution of each component (e.g., adaptive gating, cross-modal attention, correlation alignment). We also conduct hyperparameter sensitivity analyses and robustness evaluations under simulated noise to verify generalizability of our findings.

The main contributions of this thesis are:

- ADCCA: A novel adaptive fusion method employing learnable gating mechanisms to dynamically weight EEG and eye-tracking modalities based on signal quality, improving fusion flexibility and reliability.

- ACDCT: An extended framework integrating cross-modal Transformer architecture for fine-grained temporal interactions within a DCCA-aligned latent space, achieving stronger generalization especially in cross-subject evaluations.

- Empirical validation: Extensive experiments on SEED-IV and SEED-V datasets demonstrating significant improvements in subject-independent scenarios while maintaining competitive subject-dependent performance, narrowing the SD-SI performance gap compared to baseline methods.

- Comprehensive analysis: Ablation studies quantifying component contributions, hyperparameter sensitivity analyses verifying robustness, and interpretability investigations examining gating weights and attention maps to reveal how models adapt to modality quality and route cross-modal information.

This work advances multimodal emotion recognition by introducing novel adaptive fusion and cross-modal attention techniques, rigorously evaluated on benchmark datasets. To our knowledge, this represents one of the first applications of Transformer-based cross-modal attention to physiological emotion data (EEG and eye-tracking) within a correlation alignment framework, demonstrating significant advantages for handling cross-subject variability.

1.4 Experimental Setup Overview

We evaluate our methods on two benchmark multimodal emotion datasets: SEED-IV and SEED-V [24, 25], containing synchronized EEG and eye-tracking

recordings. SEED-IV includes four emotion classes (happiness, sadness, fear, neutrality) across 15 subjects, while SEED-V extends to five classes (adding disgust) with 20 subjects.

Each dataset is evaluated under two protocols. In the subject-dependent (SD) protocol, training and testing use data from the same individuals (separate sessions/trials), simulating personalized systems. In the subject-independent (SI) protocol, models train on a subset of subjects and test on entirely unseen subjects, providing a stringent test of generalization. SD results reflect capacity to capture individual-specific patterns, while SI results indicate ability to learn universal emotion-related features that transfer across people.

Primary performance metrics include classification accuracy and macro-averaged F1 score (Macro-F1), which computes per-class F1 scores and averages them to ensure balanced evaluation across classes. We report 95% confidence intervals via statistical bootstrapping at the subject level and perform paired significance tests (t-tests or Wilcoxon signed-rank tests) when comparing models. These statistical measures establish whether observed improvements are meaningful rather than chance fluctuations.

All methods undergo identical preprocessing: EEG signals are band-pass filtered with artifact removal, then segmented into temporal windows; eye-tracking data (gaze coordinates, pupil diameter) are interpolated and temporally aligned to EEG; both modalities are normalized and projected via DCCA into a shared feature space. This standardization isolates the impact of fusion strategies and model architectures.

Robustness is assessed by introducing controlled artificial noise during testing—Gaussian noise or channel dropout to simulate EEG artifacts, and occlusion periods to mimic eye-tracking loss—to evaluate sensitivity to real-world data quality degradation. Hyperparameter sensitivity analyses vary key parameters (latent space dimensionality, temporal window length, Transformer depth/heads) to verify that

performance gains hold across reasonable configuration ranges rather than at isolated optima. Detailed experimental procedures appear in Chapters 3 and 6.

1.5 Thesis Organization

The remainder of this thesis is organized as follows:

Chapter 2 reviews relevant background and literature, including affective computing fundamentals, fusion paradigms (early, late, and hybrid strategies), Canonical Correlation Analysis and DCCA [23, 28], and Transformer-based cross-modal attention [26, 27]. This positions our approach within existing work and identifies gaps our contributions address.

Chapter 3 describes the SEED-IV and SEED-V datasets [24, 25], including data collection procedures, modalities, preprocessing, and synchronization. It defines the subject-dependent and subject-independent evaluation protocols and outlines data splits and baseline comparisons.

Chapter 4 presents the Adaptive DCCA (ADCCA) methodology. We detail the problem formulation, model architecture (EEG and eye-tracking feature extractors, DCCA alignment, adaptive gating), design motivations, and training procedures (loss functions, optimization techniques). This chapter provides a complete understanding of ADCCA as the first-stage solution.

Chapter 5 introduces the Adaptive Cross-Modal Deep CCA Transformer (ACDCT), extending ADCCA with Transformer-based cross-modal attention. We describe the architecture incorporating stacked Transformer blocks that enable both self-attention (within each modality) and cross-attention (between modalities), highlighting new capabilities for capturing temporal dependencies and cross-modal interactions. The chapter discusses training considerations for managing model complexity and maintaining correlation alignment across layers.

Chapter 6 presents experimental results and in-depth analysis. We report performance on SEED-IV and SEED-V under SD and SI protocols, comparing ADCCA and ACDCT against baselines using accuracy, macro-F1, and confidence intervals. Ablation studies isolate component contributions (adaptive gating, cross-attention, DCCA alignment). Robustness tests with artificial noise assess degradation under signal corruption. Visualizations of gating weights and attention patterns demonstrate interpretability, revealing how models adapt to modality quality and route cross-modal information. The findings confirm that ACDCT achieves superior SI generalization while maintaining SD performance, affirmatively answering our research questions.

Chapter 7 concludes with a summary of contributions and findings, reflecting on implications for affective computing. We acknowledge limitations (computational overhead, data requirements) and outline future directions including personalization techniques for new users, online adaptation methods, and efficiency improvements (model compression, channel selection) to enhance practical deployment.

CHAPTER TWO

Background and Literature Review

2.1 EEG and Eye Tracking in Emotion Recognition

Electroencephalography (EEG) measures brain electrical activity and has been extensively used in affective computing due to its direct reflection of neural processes underlying emotions [6, 7, 29]. Changes in EEG patterns (e.g., power in certain frequency bands or event-related potentials) correlate significantly with emotional states, making EEG a powerful modality for emotion recognition [8, 30]. However, EEG signals are often noisy and vary between individuals and recording sessions, making reliance on EEG alone challenging.

Eye movement signals, captured via eye tracking, offer a complementary behavioral perspective on emotion. Features such as gaze direction, blink rate, and pupillary response provide insight into a person's attention and arousal levels, which are modulated by emotion. Notably, pupil dilation is a well-known indicator of heightened arousal or cognitive load [13, 14], and certain gaze patterns or blink behaviors have been associated with emotional stimuli. These ocular cues are relatively easy to acquire and are non-intrusive, though they may be influenced by conscious control or environmental factors (e.g. a subject may avert gaze or suppress expressions). In contrast, EEG signals—being an "objective" physiological measure—are not under conscious control and can reveal hidden affective changes even when external expressions are limited [6].

By combining EEG with eye tracking, we can leverage both internal neural responses and external behavioral indicators of emotion. The modalities capture different aspects of affective response: EEG reflects central nervous system activity, while eye movements reflect peripheral responses related to attention and autonomic arousal.

Multimodal emotion recognition studies have demonstrated that fusing EEG and eye-tracking data yields higher accuracy than either modality alone [15, 16]. The SEED-IV dataset, designed with simultaneous EEG and eye-tracking recordings, reflects the growing interest in this bimodal approach [24]. The inclusion of both modalities is motivated by the expectation that eye movement features can enhance or disambiguate EEG-based predictions when recognizing complex emotional responses.

2.2 Data Preprocessing and Synchronization

Before multimodal data can be fused for analysis, each signal type must undergo careful preprocessing and synchronization. EEG signals typically require artifact removal and filtering to improve signal-to-noise ratio. Common steps include band-pass filtering (e.g., 0.1–75 Hz) to remove DC drift and high-frequency noise, notch filtering to eliminate powerline interference, and Independent Component Analysis (ICA) to remove ocular and muscle artifacts [31]. Feature extraction from EEG may involve computing time-domain statistics or frequency-domain features (power spectral densities, differential entropy in specific bands) known to reflect emotional states [6]. Modern deep learning approaches may instead operate on minimally processed EEG to learn features automatically.

Eye-tracking data (e.g. pupil diameter, gaze coordinates, blink events) also require preprocessing. This can include smoothing or interpolation to handle blink-induced signal loss, and extraction of higher-level features such as average gaze fixation duration or saccade frequency. Pupillary measurements might be z-score normalized per subject to account for individual differences in pupil range. In some cases, blinks and saccades themselves are treated as discrete events that can be counted in an interval or aligned with EEG events.

Crucially, when combining EEG and eye modalities, temporal synchronization is required so that data from both sources refer to the same moments or events. EEG is often

recorded at a high sampling rate (e.g. 128–1000 Hz), whereas eye trackers may sample at a different rate (commonly 60–250 Hz). To synchronize, one approach is to resample one modality (or both) to a common timeline, or to use timestamped event markers recorded simultaneously in both data streams. For instance, in the SEED-IV dataset the EEG signals are sampled at 200 Hz while eye movement data are recorded at a different rate; alignment is achieved by recording a common clock and trial onset markers, ensuring that each EEG sample window corresponds to the correct segment of eye-tracking data [32]. In practice, researchers often segment both EEG and eye data into time epochs corresponding to emotional stimuli or trials (e.g. 2–5 second segments) and then extract synchronized feature vectors from each modality for those epochs. This preprocessing and alignment step is fundamental to guarantee that subsequent multimodal analysis truly reflects concurrent brain and eye responses to the same emotional experience.

2.3 Fusion Strategies and Adaptivity

Given well-preprocessed EEG and eye features, a key question is how to effectively fuse the two modalities to perform emotion recognition. In multimodal machine learning, fusion paradigms are generally categorized by the stage at which modalities are combined [19, 33]. In an early fusion (feature-level fusion) approach, EEG and eye features are concatenated or otherwise merged into a joint feature vector, which is then fed into a machine learning model (e.g. a classifier or neural network). Early fusion allows the model to learn direct interactions between EEG and eye features, potentially capturing complex cross-modal patterns. However, it assumes that the two modalities are already well-aligned and comparable; differences in feature scale or timing must be carefully handled. Early fusion can also lead to a very high-dimensional input, increasing the risk of overfitting when training data is limited.

In contrast, late fusion (decision-level fusion) keeps the modalities separate through most of the processing. Each modality is analyzed by its own model (for example, one classifier for EEG and another for eye tracking), and the outputs (decisions or confidence scores) are combined at the end. This could be a simple averaging or voting, or a learned weighted combination of the modality-specific decisions. Late fusion is more robust to misalignment or noise in one modality (since each stream is processed independently) and allows using modality-specific modeling techniques. However, it may fail to exploit informative interactions between modalities that occur at a fine-grained level (because the modalities meet only at the final decision).

Hybrid or mid-level fusion strategies seek a compromise, combining information at multiple stages [34, 35]. For example, one can perform some early fusion at an intermediate layer of a deep neural network (after each modality has been partially processed), or iterate between separate and combined layers. This allows the model to learn both modality-specific patterns and cross-modal interactions. Recent deep learning architectures often adopt such hybrid fusion by design – e.g., feeding each modality through several dedicated layers, then merging for a joint representation, and possibly even splitting again into separate streams later (to preserve modality-specific nuances) [20, 36, 37].

An important consideration in fusion is adaptivity – the capability of the model to dynamically adjust the contribution of each modality. Not all modalities are equally informative at all times: for instance, eye movement might be particularly telling during a jump scare in a horror video (large pupil dilation), whereas EEG might be more sensitive during a subtle mood change not accompanied by obvious eye changes. A rigid fusion (especially simple early fusion) might not account for these context-dependent variations in modality relevance. Adaptive fusion approaches introduce mechanisms to weight or select modalities based on their estimated reliability or importance. One simple form is to

learn weights for each modality's feature set. More advanced techniques estimate uncertainty for each modality's prediction and weight them accordingly [38]. In multi-task learning, Kendall et al. [38] used learned uncertainty to balance loss contributions, and a similar idea can be applied in multimodal fusion by giving higher weight to the modality with lower uncertainty or noise in a given scenario.

Another adaptation scenario is handling missing or degraded modality data. In real-world settings, one modality might occasionally be unavailable (e.g. eye tracker failure or EEG signal artifacts). Robust fusion systems should degrade gracefully by relying on the remaining modality. Some recent works explicitly address this by training models that can reconstruct or translate one modality from another, or by employing generative models to impute missing signals [39, 40]. While our focus is on when both EEG and eye inputs are present, these strategies underscore the importance of flexible fusion mechanisms.

In the context of EEG and eye movement, a number of neural network architectures have been proposed to exploit their complementarity [41, 42]. For instance, a Memory Fusion Network was introduced to learn synchronous and asynchronous interactions between multiple modalities over time, using gated memory cells to adaptively emphasize each modality's contributions at each time step [43]. Such networks can capture complex temporal dynamics where one modality might lead or lag the other in reflecting an emotional response [44, 45]. Overall, the choice of fusion strategy can significantly affect performance: feature-level fusion can harness rich interactions but needs careful normalization and alignment, whereas decision-level fusion is simpler but potentially suboptimal in capturing nuanced correlations. Many state-of-the-art solutions, therefore, integrate adaptivity – through learned weights, gating mechanisms, or attention modules – to get the best of both worlds, improving robustness and accuracy in multimodal emotion recognition [20].

2.4 Aligned Feature Space and Correlation Analysis

One major challenge in multimodal learning is that different modalities reside in very different feature spaces; direct fusion (especially early fusion) might be suboptimal if the model cannot easily reconcile EEG features with eye movement features. A powerful approach to addressing this is to project both modalities into a shared aligned feature space where their representations are comparable and maximally correlated. Classic statistical techniques for two-view data, like Canonical Correlation Analysis (CCA), aim to find linear projections of two sets of variables (here, EEG features and eye features) onto a common latent space such that the correlation between the projected variables is maximized [28]. CCA effectively identifies the common signal amid two modalities, filtering out components that are private to one modality. In our context, this could correspond to isolating the latent emotional factors that drive coherent changes in both brain signals and eye movements.

Linear CCA is limited by its linear nature and by the requirement of paired data for calculating covariance across modalities. To overcome these limitations, researchers have developed nonlinear and data-driven extensions. Deep Canonical Correlation Analysis (DCCA) uses deep neural networks (e.g. multilayer perceptrons or convolutional networks) to learn nonlinear transformations of each modality such that the resulting high-level features are maximally correlated [23]. The networks are trained jointly with a correlation-based objective (instead of or in addition to the usual task objective). By doing so, DCCA can capture complex nonlinear relationships between EEG and eye features that linear CCA would miss. For example, an EEG pattern corresponding to heightened anxiety might correlate with a nonlinear combination of eye features (like an increase in blink rate together with a specific gaze dispersion pattern); DCCA could learn to extract a representation of "anxiety level" from each modality that aligns these effects.

An evolution of this idea is the use of deep canonically correlated autoencoders (DCCAE), which combine the correlation objective with reconstruction objectives for each modality's autoencoder [46]. This helps ensure that the learned representations not only correlate across modalities but also retain useful information from the original signals. Such representations can be more robust and generalizable. In a broader sense, these methods fall under multi-view representation learning, which seeks common embeddings for data from different views (modalities) [47]. Surveys of multi-view learning highlight that correlation-based alignment (CCA and its variants) is a fundamental technique for fusing information from heterogeneous sources while reducing cross-modal discrepancies [47].

In practice, applying DCCA to EEG and eye tracking data would involve training two deep subnetworks (one per modality) on synchronized data, optimizing to maximize the correlation between the network outputs. If successful, the resulting shared latent variables can be interpreted as high-level features describing the participant's emotional state, abstracting away modality-specific noise. These aligned features can then be fed to a classifier or further multimodal model for emotion recognition. Prior work in other domains validates this strategy: for instance, DCCA has been used to align audio and video features in speech recognition, or text and visual features in sentiment analysis, yielding improved performance when modalities are not perfectly aligned in time [48]. By analogy, for EEG and eye tracking, an aligned space could help the model consider "EEG and eye movement saying the same thing" when both modalities reflect the same emotional stimulus, thus strengthening the signal, or could allow the model to identify and focus on the common emotional signature even if one modality has irrelevant variations.

It's worth noting that correlation is one way to measure alignment, but there are other criteria as well. Some works maximize mutual information between modalities, and others use adversarial learning to encourage modality-invariant representations. The

essential idea is to ensure the model captures what is shared between EEG and eye signals (presumably, the underlying emotion) while discarding modality-specific artifacts. This aligned feature space forms a foundation for higher-level fusion and is particularly useful when modalities have different statistical properties or when we expect some degree of asynchrony or variation in how they express the emotional cues.

2.5 Cross-Modal Temporal Interaction with Attention

Emotional responses are temporal by nature: they evolve over time and often involve dynamic patterns in both EEG and eye tracking signals. A critical aspect of multimodal emotion analysis is modeling the temporal dynamics and interactions across modalities. Recently, attention mechanisms have become a cornerstone for modeling sequential data and cross-modal relationships. Attention was first popularized in the context of sequence-to-sequence models in NLP and was generalized in the Transformer architecture [26]. In essence, an attention mechanism enables a model to focus on the most relevant parts of a sequence (or another sequence) when making a prediction, by computing attention weights that highlight important features while suppressing less relevant ones.

For multimodal sequences like EEG and eye tracking, attention can operate in several ways:

- Intra-modal (self-)attention: Each modality sequence can benefit from attention over its own time steps to capture long-range dependencies. For example, a self-attention over EEG could learn that patterns at the beginning of a stimulus (say, a sudden spike in beta rhythm) combined with patterns later (sustained theta rhythm) are jointly indicative of a certain emotion. Similarly, self-attention on eye movement might relate an initial wide gaze sweep to later focused stare as part of an emotional response.

- Cross-modal attention: One modality can use attention to seek relevant information in the other modality's sequence. For instance, when processing the EEG at a particular time, the model can attend to the eye tracking sequence to find which eye features at which timestamps are most strongly associated with the current EEG context. This is especially valuable if the two signals are not strictly aligned or synchronous – attention can learn the temporal offset or mapping between EEG and eye events.

By incorporating such mechanisms, models can learn cross-modal temporal interactions. A prominent example is the Multimodal Transformer proposed by Tsai et al. [27] for unaligned sequential modalities: it uses cross-attention to allow each modality to dynamically attend to time steps of the other modalities, effectively learning the latent alignment. This approach was initially applied to language, vision, and acoustic signals, but the concept extends naturally to any multimodal sequence. The Transformer's multi-head attention structure can capture different types of relationships in parallel, e.g. one attention head might focus on aligning peaks of arousal in EEG with moments of pupil dilation, while another head might correlate eye gaze aversion with certain EEG frontal asymmetry patterns of emotion.

Attention-based multimodal models have shown state-of-the-art performance in various affective computing tasks. For example, Yu et al. [49] introduced a hierarchical transformer for multimodal sentiment analysis that first extracts features within each modality and then uses cross-modal attention at a higher level, demonstrating the benefit of targeted attention in fusion. Similarly, in the vision-language domain, co-attention networks (where image regions and text words attend to each other) like ViLBERT and LXMERT achieved large improvements by letting each modality query the other for relevant context [50, 51]. In our EEG-eye context, an attention-based model could, for instance, learn that a certain EEG oscillation pattern is only significant for emotion if it

coincides with a prolonged gaze fixation (detected in the eye data)—a rule that simple early fusion might not easily learn. The attention mechanism would give high weight to the coincident EEG-eye events and lower weight to EEG signals at times when eye data indicates the subject was distracted.

Another advantage of attention is interpretability. By examining the attention weights, we can gain insight into what the model found important. For example, attention weights might reveal that the model consistently looks at pupil dilation shortly after an emotional stimulus onset to corroborate the EEG-indicated arousal, aligning with known physiological responses. This can help in explaining the model's decisions and build trust in multimodal BCI applications.

In summary, attention mechanisms empower models to handle the complexity of temporal alignment and interaction between EEG and eye tracking signals. They provide a flexible framework where the model itself learns when and how to integrate information across time and modalities, rather than assuming a fixed fusion at each time step. This leads to more effective modeling of emotions that unfold over time and are expressed through multiple channels simultaneously. As a result, attention-based cross-modal temporal models have become a leading approach in recent literature for emotion recognition and broader multimodal sequence modeling [27].

2.6 Generalization Across Subjects

A persistent challenge in EEG-based emotion recognition (and physiological computing in general) is the variability of signals across different individuals. Each person's EEG has unique characteristics due to differences in brain anatomy, electrode impedance, baseline neural activity, and idiosyncratic reactions to emotional stimuli. Eye movement patterns can also vary with individual traits or strategies (for example, some people tend to avert gaze when upset, others might stare). Consequently, models trained

on data from a specific group of subjects often experience a drop in performance when applied to new subjects not seen during training. This issue of cross-subject generalization is well-documented: classification accuracies that are high in a within-subject setting (where training and testing data come from the same person) can significantly decrease in a cross-subject setting [6, 7].

Improving cross-subject generalization is critical for developing practical emotion recognition systems. One straightforward approach is to gather large and diverse training datasets encompassing many subjects, so that the model can learn to ignore subject-specific features in favor of more universal patterns. This is partly the motivation behind newer datasets like SEED-V, which includes 20 subjects across multiple sessions [25]. However, collecting and labeling physiological data from many users is expensive and time-consuming, especially for emotions which may require carefully elicited responses.

Thus, a line of research focuses on techniques for domain adaptation or subject-invariant feature learning. The goal is to make the model's internal representation of the data more universal, stripping away the subject-dependent noise. For instance, one can apply normalization strategies to EEG features (such as z-score normalization per subject) to reduce inter-subject differences in scale. More sophisticated methods use machine learning approaches: one example is the work of Li et al. [21], who proposed a Self-Organized Graph Neural Network that models EEG electrodes as graph nodes and includes a mechanism to account for individual differences, achieving better cross-subject emotion recognition. The graph structure can capture spatial relationships in EEG sensors while an adaptive module adjusts to each subject's topology or distribution, thereby improving generalization.

Another example is the use of domain adaptation algorithms. Wu et al. [52] introduced a multi-source domain adaptation framework using Graph Convolutional

Networks (GCNs) to align feature distributions from multiple source subjects to a target subject. By leveraging data from several source domains (subjects) and learning mappings to the target domain, their model can mitigate the shift in EEG feature distributions between subjects. Techniques such as adversarial training (where a domain discriminator is used to encourage the model to produce indistinguishable features for different subjects) have also been explored in EEG-based emotion recognition and have shown promising results in making features more subject-invariant.

It is also worth noting that the inclusion of eye tracking data might help generalization to some extent. Some eye movement features (like reflexive pupil response to brightness or certain universal gaze patterns to stimuli) could be more consistent across people than raw EEG signals. The multimodal approach could thus buffer the variability—if one modality varies widely, the other might anchor the prediction. Nevertheless, if both modalities are subject-specific, the model might actually overfit to personal idiosyncrasies present in both EEG and eye data.

In practice, rigorous evaluation protocols are used to assess cross-subject generalization. A common approach is leave-one-subject-out (LOSO) cross-validation: the model is trained on all but one subject and tested on the held-out subject, repeating this for each subject. This provides a measure of how well the model might perform on an entirely new person. Some studies use a variant where a few subjects are held out for testing and possibly a few for validation (mimicking a train/validation/test split by subject). The results consistently indicate that cross-subject emotion recognition is much more challenging than subject-specific training, with a notable performance gap. This has driven the community to prioritize generalization in recent years.

2.7 Evaluation Methodologies and Reporting Results

Proper evaluation and reporting are essential to validate any emotion recognition approach. In multimodal EEG and eye tracking research, several factors must be clearly reported:

Performance Metrics: Most works report classification accuracy for emotion recognition (especially when emotions are treated as discrete classes). However, considering class imbalance or the ordinal nature of certain emotion dimensions, additional metrics are often used: precision, recall, F1-score for each class (or macro-averaged), and sometimes Cohen's kappa or area under the ROC curve if appropriate. For regression tasks (e.g. predicting a valence level), mean squared error or correlation with ground truth might be reported. It is advisable to report more than one metric since a single metric like accuracy can be misleading in imbalanced scenarios (e.g. if one emotion class dominates). In the multimodal context, one might also report the performance of single-modality models (EEG-only, eye-only) alongside the fused model to quantify the fusion benefit.

Cross-Validation Protocol: As discussed, reporting whether the evaluation is within-subject or cross-subject (or cross-session) is crucial. Early EEG emotion studies often used within-subject *k*-fold cross-validation (where data from each subject is split into training/testing folds). While this can show the model's capacity to learn personal emotion cues, it does not demonstrate generalization. Therefore, recent studies have increasingly adopted cross-subject testing protocols. For transparency, researchers should clearly state how data is split. If LOSO cross-validation is used, the average performance across all test subjects (and its standard deviation) is typically reported. If a separate hold-out set is used (like a subset of subjects as a fixed test set), results on that set are given. Reproducibility and fair comparison demand that these protocols are standardized or at least well documented.

Statistical Significance: Given the relatively small sample sizes in many EEG datasets, random chance or particular subject outliers could influence results. It is good practice to run models multiple times with different random initializations or data shuffles and report an average and variance. Statistical tests (e.g. paired t-tests or Wilcoxon signed-rank tests) can be used to verify if improvements of a proposed method over baselines are significant. Not all papers do this, but in a rigorous thesis, noting whether an accuracy gain is likely due to chance strengthens the credibility of claims.

Reproducibility: With complex deep learning models, it is important to report implementation details (network architecture, hyperparameters, training epochs, etc.) and, when possible, to share code. In the literature, there is a movement toward common benchmarking. For example, LibEER [53] provides a comprehensive library and benchmark for EEG-based emotion recognition, offering standard data preprocessing pipelines and evaluation tools for several datasets. Such efforts encourage consistent evaluation and make it easier to compare new models against established baselines under the same conditions.

When reporting multimodal fusion results, one should also consider ablation studies and case analyses. Ablation experiments might involve disabling the fusion (testing each modality alone), or removing an adaptive component (to show its effect on performance). This helps identify which parts of the system contribute most. Case studies could examine particular instances where fusion helped (e.g. a case where EEG-alone was wrong but EEG+eye was correct, illustrating a scenario where the eye data resolved ambiguity) and cases where fusion failed (perhaps indicating when modalities conflict or when one modality dominates incorrectly).

In summary, a solid evaluation of an EEG eye tracking emotion recognition model should demonstrate not only high accuracy but also robust performance across subjects and conditions. It should convincingly show that the fusion of modalities yields a tangible

benefit. Clear reporting of protocols and metrics allows the research community to trust the results and build upon them. As the field advances, we expect to see more standardized benchmarks and public challenges that enforce these best practices, leading to more generalizable and reproducible findings in multimodal emotion recognition [54].

CHAPTER THREE

DATASETS: SEED, SEED-IV, AND SEED-V

3.1 SEED

The original SEED dataset established the recording and labeling protocol followed by later releases. It contains 62-channel EEG from 15 subjects (7 male, 8 female; mean age 23.27 ± 2.37 years), each completing three sessions separated by about a week; each session includes 15 trials (film clips) designed to elicit positive, neutral, or negative emotions [21, 55]. Stimuli are Chinese movie excerpts of approximately 4 minutes each. Within a session, the procedure is: a 5 s on-screen hint, the clip itself, a 45 s self-assessment, and a 15 s rest interval; clips targeting the same emotion are not shown consecutively [31]. Recordings use an ESI NeuroScan 62-channel cap (10–20/10–10 layout) with synchronized triggers; the public release provides raw and preprocessed files, with downsampling to 200 Hz and a 0–75 Hz band-pass applied in the supplied preprocessed version [31]. Labels are assigned at the trial level (1 negative, 0 neutral, +1 positive).

SEED includes an explicitly multimodal branch: SEED Multimodal contains EEG and eye movement data for 12 of the 15 subjects (the remaining 3 have EEG-only), acquired with SMI eye-tracking glasses and time-aligned to EEG via shared triggers [55]. The dataset's sessioned design and balanced class scheduling make it suitable for both subject-dependent (SD) and subject-independent (SI) evaluation; typical SD splits hold out trials or sessions within a subject, while SI evaluations reserve entire subjects for testing to avoid data leakage across partitions [21].

3.2 SEED-IV

SEED-IV extends SEED to four discrete categories: happy, sad, fear, and neutral. It contains data from 15 subjects (three sessions per subject), and each session has 24 trials arranged as 6 per category with counterbalanced order [24, 32]. EEG is recorded with the same 62-channel NeuroScan system and synchronized stimulus triggers; the public materials and widely used derivatives provide either the raw recordings (often originally at 1 kHz) or downsampled 200 Hz preprocessed versions, depending on the file set used [32, 53]. Eye-movement streams (gaze coordinates and pupil size) are captured with a head-mounted tracker and aligned to EEG via the shared trigger pulses [24].

Trial structure follows SEED (baseline–stimulus–rest) with dataset-specific clip durations (shorter than SEED in many releases) and fixed per-session scheduling to maintain class balance [52]. The dataset is designed for SD and SI protocols. In SD, cross-session or cross-trial splits within each participant are common; in SI, leave-one-subject-out (LOSO) or equivalent rotations are standard. Because inter-subject variability is pronounced, SEED-IV is widely used to benchmark adaptive fusion between EEG and eye tracking and to study generalization beyond fixed fusion rules [52]. For reproducibility across methods, common preprocessing includes band-pass filtering, re-referencing, ocular/muscular artifact handling (regression or ICA) for EEG, and resampling plus interpolation for brief dropouts in eye tracking, followed by segmentation into fixed-length windows with consistent stride before feature learning [24, 32].

3.3 SEED-V

SEED-V expands the taxonomy to five emotions—disgust, fear, sad, neutral, happy—and increases subject count. The official description reports 20 subjects (10 male, 10 female), recruited from Shanghai Jiao Tong University, each assessed for normal vision/hearing and stable mental state; sessions are separated in time and accompanied by

personality screening (EPQ) [25]. In widely used processed distributions, each subject completes three sessions with 15 trials per session (3 per category), yielding a consistent, balanced design for five-way classification [56]. EEG again uses the 62-channel NeuroScan setup; eye movement (gaze and pupil) is recorded and synchronized via hardware triggers. Public releases commonly provide precomputed features (e.g., differential entropy) alongside raw or downsampled signals to facilitate standardized baselines [56].

The five-class label space stresses both the discriminative capacity of encoders and the ability of fusion mechanisms to resolve borderline cases. As with SEED-IV, both SD and SI protocols are standard; SI typically follows LOSO or similar rotations, with strict separation of all windows from held-out subjects to avoid leakage. Reporting both accuracy and macro-F1 is recommended to guard against mild class imbalance introduced by unusable trials or subject-level dropouts [54]. Precise synchronization and per-trial timestamps (baseline start, stimulus onset/offset, rest) support frame-accurate alignment for models that rely on cross-modal temporal interaction [25].

CHAPTER FOUR

Adaptive DCCA-Based Multimodal Emotion Recognition Method

4.1 Method Overview

4.1.1 Problem Definition and Challenges

In multimodal physiological signal emotion recognition, the core challenge lies in effectively integrating heterogeneous information from different physiological sources. EEG signals provide rich neural activity information, containing high-dimensional temporal data from multiple channels that directly reflect the brain's emotion processing. Eye movement signals, conversely, have lower dimensions but carry behavioral information closely related to emotional states, such as gaze patterns and pupil responses. These modalities differ significantly in temporal scale, data dimensionality, and noise characteristics, presenting substantial challenges for multimodal fusion.

Traditional fusion methods often employ fixed integration strategies unable to adapt to changes in modality contributions under varying circumstances. For example, when subjects are highly focused, eye movement signals may contain more useful information, whereas during high emotional arousal, EEG signal changes may be more significant. Additionally, dynamic fluctuations in signal quality demand adaptive fusion capabilities. We propose an adaptive fusion solution to address these challenges.

4.1.2 Method Framework Design

The design of the Adaptive DCCA framework follows principles of modularity and end-to-end learning. The entire framework consists of three main modules: the feature extraction module learns high-level representations from raw signals, the adaptive fusion module achieves dynamic multimodal integration, and the emotion classification module produces the final emotion prediction. This modular design not only provides a

clear system structure but also facilitates targeted optimization and improvement of specific components.

Our design philosophy emphasizes preserving the uniqueness of each modality while exploring inter-modal correlations. Each modality employs a specially designed feature extraction network that considers its signal characteristics. During fusion, rather than simply concatenating or averaging features, we uncover latent associations between modalities through Deep Canonical Correlation Analysis (DCCA) and achieve adaptive weighting via an attention-based mechanism. This design enables the model to utilize complementary information from each modality while flexibly adjusting fusion strategies based on context. An overview of the proposed framework is illustrated in Figure 4.1.

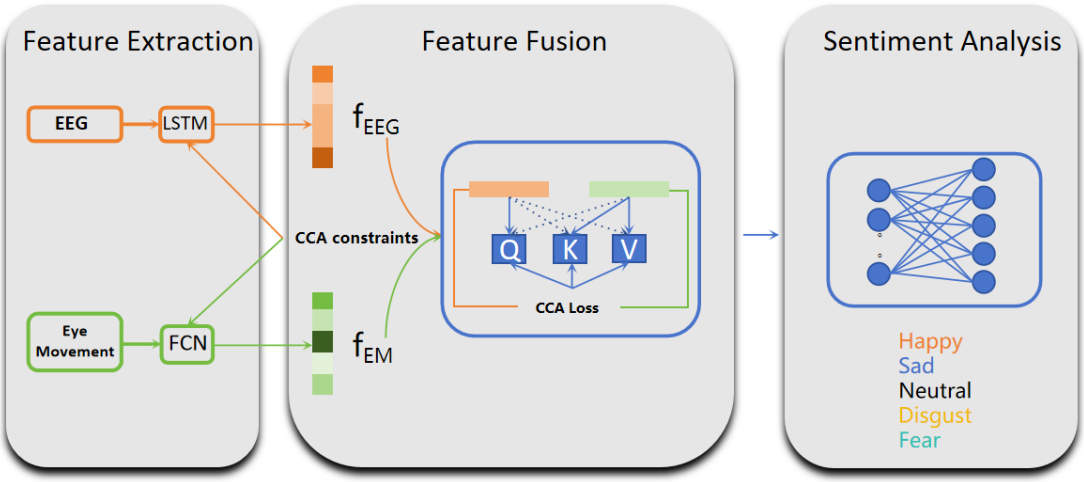

Figure 4.1: Overview of the proposed Adaptive DCCA Framework. EEG and Eye movement features are extracted using LSTM and FCN respectively. The adaptive weight mechanism dynamically adjusts each modality's contribution during fusion, enhancing sentiment classification performance by prioritizing informative features.

4.2 Feature Extraction Network Design

4.2.1 EEG Feature Extraction Network

Feature extraction from EEG signals is a crucial step in emotion recognition [29, 41]. Considering the multi-channel characteristics and complex spatiotemporal dynamics of EEG signals, we designed a feature extraction architecture based on Long Short-Term Memory (LSTM) networks [57]. The choice of LSTM is motivated by its strength in processing time-series data, particularly its ability to effectively capture long-range temporal dependencies in the signal [42].

In the network design, we first introduce a spatial attention mechanism to handle multi-channel EEG data. Different regions of the human brain play distinct roles in emotion processing. For example, the frontal lobe is closely associated with emotion regulation, while the temporal lobe participates in emotional memory processing. The spatial attention mechanism can automatically learn the importance of different EEG channels, allowing the network to focus more on brain regions relevant to the current emotional state. This data-driven channel weighting method avoids the subjectivity of manual feature selection.

The spatially attention-weighted EEG signals are then fed into a bidirectional LSTM network. The bidirectional structure enables the network to utilize both past and future context information simultaneously, which is particularly important for understanding the dynamic progression of emotional responses. LSTM's gating mechanism effectively mitigates the gradient vanishing problem of traditional RNNs when processing long sequences, enabling the network to capture persistent features of emotional states.

The design of the network's output layer has also been carefully considered. We not only make use of the LSTM's final hidden state but also introduce temporal pooling

mechanisms to aggregate information over the entire sequence. This design captures both instantaneous emotional responses and sustained emotional states, providing more comprehensive EEG feature representations.

4.2.2 Eye Movement Feature Extraction Network

Although eye movement signals have relatively lower dimensionality than EEG, they contain rich cognitive and emotional information. We designed a feature extractor based on a Fully Connected Network (FCN) that is specifically optimized for the characteristics of eye movement data. Unlike EEG, each dimension of eye movement data has a clear physical meaning – for instance, fixation coordinates reflect the spatial distribution of visual attention, pupil diameter changes relate to emotional arousal, and blink frequency can indicate cognitive load or fatigue.

The first layer of the eye movement network is a feature-level attention mechanism for adaptively selecting the most relevant eye movement features. This design accounts for the fact that different emotional states may manifest predominantly through different eye movement indicators. For example, positive emotions may lead to more active visual exploration patterns, manifested as increased saccade frequency, while negative emotions like sadness may result in more prolonged fixations and reduced overall eye movement activity. The feature attention mechanism enables the network to adjust its focus on different eye movement indicators based on the current emotional context.

To capture the temporal dynamics of eye movement signals, we integrate temporal convolutional layers into the network. Although the sampling rate of eye tracking is typically lower than that of EEG, the temporal patterns of eye movements still carry important emotional information. For example, pupil responses to emotional stimuli exhibit characteristic temporal profiles, including an initial rapid dilation followed by a

slower recovery. Temporal convolutional layers can automatically learn these temporal patterns and extract emotion-related dynamic features.

Deeper layers of the eye movement network use multiple fully connected layers for feature transformation and abstraction. We apply batch normalization and dropout techniques between layers [58, 59], which accelerate the training process and improve model generalization. The final eye movement feature representation thus integrates information from multiple aspects, including spatial distribution, temporal dynamics, and feature importance.

4.3 Adaptive Deep Canonical Correlation Analysis

4.3.1 Principles and Advantages of Deep CCA

Deep Canonical Correlation Analysis (DCCA) is an extension of traditional Canonical Correlation Analysis (CCA) within a deep learning framework. Traditional CCA aims to find linear combinations of two sets of variables that maximize the correlation between these combinations. However, when dealing with complex physiological signals, linear assumptions are often oversimplified and cannot capture the nonlinear relationships between modalities. DCCA greatly enhances CCA's expressive power by introducing deep neural networks to learn nonlinear transformations of the data.

In our framework, DCCA's role is to learn a shared representation space for EEG and eye movement features. Let f_{EEG} and f_{EM} denote the extracted features from EEG and eye movement signals, respectively. After processing through their respective deep networks, these features are projected into a shared latent space to obtain representations O_{EEG} and O_{EM}. The DCCA objective is to maximize the canonical correlation between these representations by minimizing the correlation loss:

$$\mathscr{L}_{\text{CCA}} = -\text{corr}(O_{\text{EEG}}, O_{\text{EM}}) \tag{4.1}$$

where $\text{corr}(\cdot, \cdot)$ computes the canonical correlation between the two modality representations.

This alignment is not a simple feature matching, but rather a semantic-level fusion of information. For example, enhanced oscillatory activity in EEG that reflects emotional arousal may correspond to pupil dilation in eye movements. DCCA can automatically discover and strengthen such cross-modal associations.

Another major advantage of DCCA lies in its robustness to noise. By maximizing the correlation between the two modalities, DCCA tends to extract common, stable information present in both signals while suppressing modality-specific noise. This property is particularly important for physiological signal processing, as both EEG and eye movement signals are susceptible to various sources of interference. Additionally, the DCCA objective serves as a form of regularization, preventing the model from overfitting to patterns found in only one modality.

4.3.2 Design Motivation for Adaptive Weighting Mechanism

Traditional DCCA-based fusion assigns equal importance to both modalities, overlooking several practical issues. First, the information content and reliability of each modality can vary under different conditions. For example, in visual stimulation experiments, eye movement signals may provide more direct indicators of emotional responses, whereas in resting states EEG may carry more pertinent information. Second, individual differences can affect modality contributions: some individuals express emotion more through physiological arousal changes (captured well by EEG), while others exhibit changes primarily in attention patterns (reflected in eye movements).

Based on these observations, we propose an adaptive weighting mechanism. The core idea is to let the model learn which modality to rely on more in specific situations. This weighting is learned in a data-driven manner through end-to-end training, avoiding

the limitations of manually set fusion rules. An attention network is employed as the means to implement this mechanism, dynamically calculating modality weights based on the quality and relevance of the input features.

Adaptive weighting not only improves fusion flexibility but also enhances the model's interpretability. By analyzing the learned attention weight distribution, we can understand the relative importance of each modality under different emotional states. This interpretability is significant for understanding the physiological underpinnings of emotions and helps establish user trust in practical applications.

4.3.3 Technical Implementation of the Fusion Process

The implementation of the adaptive fusion process involves the collaboration of multiple technical components. First, EEG and eye movement feature vectors obtained from their respective extraction networks need to be aligned in dimensionality. Since the original feature dimensions of the two modalities may differ, we project each into a common latent space of the same dimension using learned projection layers. This projection is not a simple linear transform but a nonlinear mapping implemented via multi-layer perceptrons (MLPs) to ensure rich representational capacity.

The calculation of attention weights is a key step in the fusion process. The attention network takes the features from both modalities as input and produces a quality score for each modality through a series of nonlinear transformations. These scores reflect the information content and reliability of each modality's features. To ensure the weights are comparable and properly scaled, we apply a softmax function to normalize the scores so that the weights for both modalities sum to 1.

The final fused feature representation is computed as a weighted combination of the DCCA-aligned modality representations:

$$O_{\text{fusion}} = \alpha \cdot \text{Attention}_{\text{EEG}} \cdot O_{\text{EEG}} + \beta \cdot \text{Attention}_{\text{EM}} \cdot O_{\text{EM}} \tag{4.2}$$

where α and β are learnable parameters that balance the contributions of EEG and eye movement features, and $\text{Attention}_{\text{EEG}}$ and $\text{Attention}_{\text{EM}}$ are the computed attention weights for each modality. This formulation allows the model to dynamically adjust the importance of each modality based on their relative informativeness in the current context.

Rather than relying on simple weighted addition, the attention mechanism uses a soft alignment strategy: each latent representation O_{EEG} and O_{EM} is first projected into an attention space using a small multi-layer perceptron, which outputs modality-specific attention scores. These scores are then normalized via a softmax function before multiplying each modality's representation. This approach allows the network to focus on salient features within each modality and suppress less informative aspects, especially when facing noisy or incomplete inputs.

To further enhance the expressive power of the fused features, we introduce residual connections and feature enhancement layers in the fusion module. Residual connections pass the original modality features directly to the fusion output, helping to avoid information loss in deeper layers of the network. Meanwhile, additional transformation layers are applied to the fused output to extract higher-level emotional representations. These additions refine the fused features and contribute to improved classification performance.

4.4 Emotion Classification and Model Optimization

4.4.1 Classification Network Design

Once the multimodal features are fused, a classification network maps these features to the final emotion categories. The design of the classification network must balance model complexity with generalization ability. We employ a Multi-Layer Perceptron (MLP) with an appropriate number of hidden layers and neurons. The depth and width of this network are carefully tuned to ensure it has sufficient capacity to learn

complex emotional patterns while avoiding overfitting, given the limited size of training data.

Between the hidden layers, we apply various regularization techniques to improve generalization. Dropout is used to randomly drop certain neural connections during training, forcing the network to learn more robust and distributed feature representations. Batch normalization is also applied to accelerate training convergence and provide additional regularization. The combination of these techniques significantly improves the model's performance on validation and test sets.

The output layer of the classifier uses a softmax activation function to convert the network's outputs into a probability distribution over emotion classes. The predicted sentiment label $\hat{Y}$ is computed as:

$$\hat{Y} = C(O_{\text{fusion}}) \tag{4.3}$$

where $C(\cdot)$ represents the classification network that maps the fused features to the emotion categories. This probabilistic output is used for making the final classification decision and also provides a measure of confidence for each prediction. In practical applications, the confidence information can help identify cases where the model is uncertain, which could trigger additional processing steps or human intervention as needed.

4.4.2 Loss Function Design and Optimization Strategy

Training the model involves minimizing a joint loss function tailored to our multimodal setup. The overall loss contains three main components: the classification loss, the DCCA correlation loss, and regularization terms [60]. The classification loss is a standard cross-entropy loss that directly optimizes the emotion recognition accuracy. The DCCA loss term ensures that the learned features from both modalities maintain a high

correlation in the shared latent space. The regularization terms include weight decay and an attention smoothing term, which together help prevent overfitting and encourage a reasonable distribution of attention weights.

Directly optimizing the DCCA objective can be challenging due to operations like covariance matrix computation and inversion, which might lead to numerical instability. To address this, we adopt an alternating optimization strategy during training. In each training iteration, we first fix the DCCA projection matrices and update the network parameters (feature extractors, attention network, classifier) via backpropagation using the current estimate of the correlation loss. Then we fix the network parameters and update the DCCA projection matrices by solving the corresponding generalized eigenvalue problem to maximize correlation. This alternating procedure ensures training stability while effectively optimizing both the classification and correlation objectives.

Another important aspect of our optimization strategy is learning rate scheduling. We use a cosine annealing schedule to gradually reduce the learning rate as training progresses. This strategy employs a relatively higher learning rate in the early epochs to quickly explore the parameter space, and then diminishes the learning rate in later epochs for fine-grained tuning. In practice, this scheduling has been shown to help the model converge to better local optima.

4.4.3 Training Techniques and Implementation Details

Successfully training a deep multimodal model requires careful attention to various implementation details. Data preprocessing is one of the most critical steps. For EEG signals, we apply a band-pass filter in the range of 1–50 Hz to remove high-frequency noise and baseline drift. This preprocessing step ensures that the EEG signals are cleaned of artifacts while preserving the relevant frequency components associated with emotional processing. Eye movement data undergoes a normalization procedure that scales raw gaze

coordinates and pupil size metrics into a consistent range. This ensures that outliers or abrupt changes in fixation do not disproportionately affect the downstream fusion process.

By harmonizing the input representations of these two modalities, the proposed framework can better learn consistent embeddings during the DCCA alignment stage. We also use Independent Component Analysis (ICA) to eliminate common artifacts such as electrooculographic (eye-blink) and electromyographic (muscle) noise from EEG signals. For eye movement signals, preprocessing includes outlier detection and smoothing to clean the gaze and pupil size data, ensuring stable feature extraction.

We also employ data augmentation techniques to improve the model's generalization. For EEG, augmentation methods such as randomly shifting the time windows and adding Gaussian noise are used to simulate variability in signal timing and noise conditions. For eye movement data, we apply random scaling of signal magnitude and temporal perturbation (e.g., slight time warping) as effective augmentation strategies. These techniques simulate real-world variations in the data, making the model more robust to such variations.

Moreover, batch processing strategies are adjusted due to the DCCA component. Since computing DCCA involves estimating covariance matrices on each mini-batch, the batch size cannot be too small or the correlation estimates become unstable. We determined an appropriate batch size through preliminary experiments to ensure reliable covariance estimation without sacrificing training efficiency with overly large batches. When training on multiple GPUs, we use synchronized batch normalization to maintain consistent statistics across different devices.

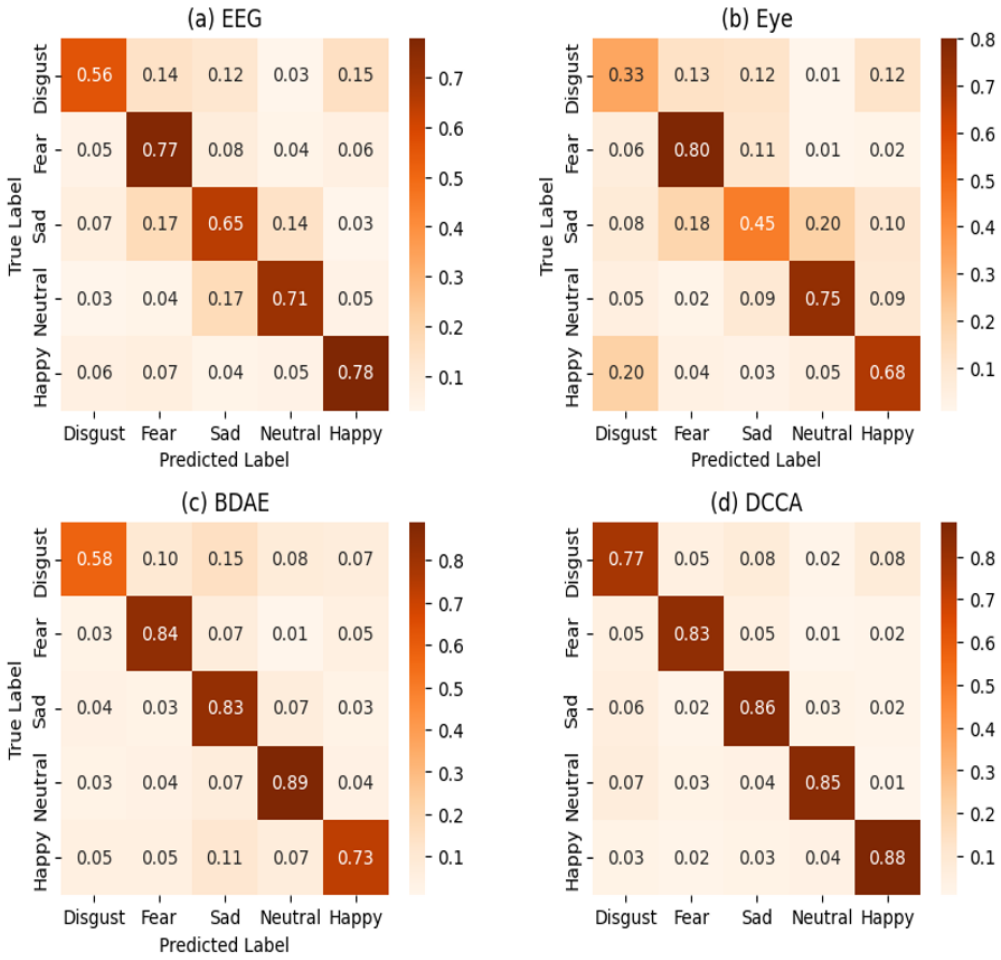

Figure 4.2: Confusion matrices comparison for multimodal emotion recognition on the SEED-V dataset (subject-dependent setting). Adaptive DCCA effectively reduces confusion between classes by adaptively weighting EEG and eye movement signals.

CHAPTER FIVE

Adaptive Cross-Modal DCCA Transformer (ACDCT)

5.1 Overview

This chapter introduces the Adaptive Cross-Modal Deep CCA Transformer (ACDCT), an advanced framework that extends ADCCA by incorporating fine-grained temporal interactions between EEG and eye-tracking modalities. While ADCCA performs scalar fusion in a correlation-aligned space, ACDCT employs stacked cross-modal Transformer blocks to enable bidirectional information exchange at the token level. The framework retains quality-aware gating mechanisms to dynamically emphasize the more reliable modality while facilitating complementary evidence exchange over time. Layer-wise DCCA regularization maintains modality alignment across network depth, reducing heterogeneity and stabilizing attention mechanisms.

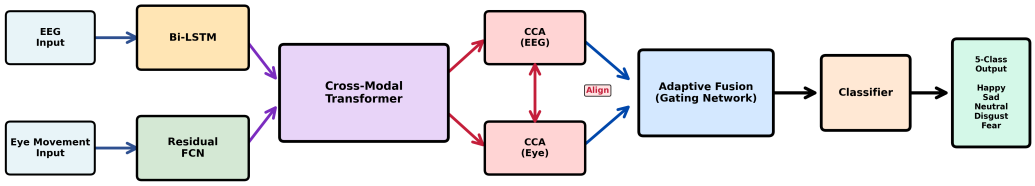

Figure 5.1: Overview of the ACDCT Framework.

5.2 Method Description

5.2.1 Problem Formulation

Let $X \in \mathbb{R}^{T \times C_e}$ denote an EEG window with T time steps and C_e channels, and $Y \in \mathbb{R}^{T' \times C_g}$ represent the synchronized eye-tracking window with T' time steps and C_g features (gaze coordinates and pupil measurements). The goal is to learn a function that maps these multimodal inputs to emotion class predictions while adaptively fusing information from both modalities.

5.2.2 ACDCT Algorithm

The complete ACDCT framework consists of three main stages: initial encoding and alignment (Algorithm 1), stacked cross-modal Transformer layers (Algorithm 2), and quality-aware global fusion (Algorithm 3). This modular design highlights the key innovations while maintaining clarity.

Algorithm 1 ACDCT Overall Framework

Input: EEG sequence $X \in \mathbb{R}^{T \times d_e}$, Eye features $Y \in \mathbb{R}^{d_g}$
Initialization: Initialize model parameters Θ
Encode EEG: $\mathbf{h}_x = \text{BiLSTMWithAttention}(X) \in \mathbb{R}^d$
Encode Eye: $\mathbf{h}_y = \text{ResidualFCN}(Y) \in \mathbb{R}^d$
Cross-modal interaction: $(\tilde{\mathbf{h}}_x, \tilde{\mathbf{h}}_y) = \text{CrossModalTransformer}(\mathbf{h}_x, \mathbf{h}_y)$
Project to CCA space: $\mathbf{z}_x = \tanh(\text{LN}(W_x \tilde{\mathbf{h}}_x)) \in \mathbb{R}^{d'}$
Project to CCA space: $\mathbf{z}_y = \tanh(\text{LN}(W_y \tilde{\mathbf{h}}_y)) \in \mathbb{R}^{d'}$
Compute CCA loss: $\mathscr{L}_{\text{cca}} = -\sum_{i=1}^{r} \sqrt{\lambda_i}$
Adaptive fusion: $\mathbf{z} = \text{AdaptiveFusion}(\mathbf{z}_x, \mathbf{z}_y)$
Main classification: $\hat{y}_{\text{main}} = \text{Classifier}_{\text{main}}(\mathbf{z})$
Auxiliary classification: $\hat{y}_x = \text{Classifier}_x(\mathbf{z}_x)$, $\hat{y}_y = \text{Classifier}_y(\mathbf{z}_y)$
Output: $\hat{y}_{\text{main}}$ =0

Algorithm 2 Cross-Modal Transformer

Input: EEG features $\mathbf{h}_x \in \mathbb{R}^d$, Eye features $\mathbf{h}_y \in \mathbb{R}^d$, Number of layers L
Add modality embeddings: $\mathbf{h}_x^{(0)} = \mathbf{h}_x + \mathbf{e}_x$, $\mathbf{h}_y^{(0)} = \mathbf{h}_y + \mathbf{e}_y$
for $\ell = 1$ to L **do**
Self-attention for EEG: $\hat{\mathbf{h}}_x^{(\ell)} = \mathbf{h}_x^{(\ell-1)} + \text{MHA}(\text{LN}(\mathbf{h}_x^{(\ell-1)}))$
Self-attention for Eye: $\hat{\mathbf{h}}_y^{(\ell)} = \mathbf{h}_y^{(\ell-1)} + \text{MHA}(\text{LN}(\mathbf{h}_y^{(\ell-1)}))$
Cross-attention EEG to Eye: $\mathbf{c}_x^{(\ell)} = \text{MHA}(\text{LN}(\hat{\mathbf{h}}_x^{(\ell)}), \text{LN}(\hat{\mathbf{h}}_y^{(\ell)}))$
Cross-attention Eye to EEG: $\mathbf{c}_y^{(\ell)} = \text{MHA}(\text{LN}(\hat{\mathbf{h}}_y^{(\ell)}), \text{LN}(\hat{\mathbf{h}}_x^{(\ell)}))$
Combine: $\bar{\mathbf{h}}_x^{(\ell)} = \hat{\mathbf{h}}_x^{(\ell)} + \mathbf{c}_x^{(\ell)}$, $\bar{\mathbf{h}}_y^{(\ell)} = \hat{\mathbf{h}}_y^{(\ell)} + \mathbf{c}_y^{(\ell)}$
Feed-forward for EEG: $\mathbf{h}_x^{(\ell)} = \bar{\mathbf{h}}_x^{(\ell)} + \text{FFN}(\text{LN}(\bar{\mathbf{h}}_x^{(\ell)}))$
Feed-forward for Eye: $\mathbf{h}_y^{(\ell)} = \bar{\mathbf{h}}_y^{(\ell)} + \text{FFN}(\text{LN}(\bar{\mathbf{h}}_y^{(\ell)}))$
end for
Output: $\tilde{\mathbf{h}}_x = \mathbf{h}_x^{(L)}$, $\tilde{\mathbf{h}}_y = \mathbf{h}_y^{(L)}$ =0

Algorithm 3 Adaptive Fusion

Input: Aligned features $\mathbf{z}_x, \mathbf{z}_y \in \mathbb{R}^{d'}$
Compute adaptive gates: $[\alpha, \beta] = \text{GateNetwork}([\mathbf{z}_x; \mathbf{z}_y])$
Weighted fusion: $\mathbf{z}_{\text{fused}} = \alpha \cdot \mathbf{z}_x + \beta \cdot \mathbf{z}_y$
Transform: $\mathbf{z} = \text{Transform}(\mathbf{z}_{\text{fused}})$
Output: $\mathbf{z}$ =0

5.2.3 Modality-Specific Encoders

To ensure fair comparison with ADCCA and maintain comparable model capacity, we reuse the compact encoders from Chapter 4. The EEG encoder employs channel-attention mechanisms followed by bidirectional LSTM or temporal convolutional

networks to produce token embeddings $\mathbf{H}_x \in \mathbb{R}^{T \times d}$. The eye-tracking encoder uses feature-level attention to adaptively select relevant indicators (gaze coordinates, pupil diameter, blink frequency), followed by lightweight temporal processing to output $\mathbf{H}_y \in \mathbb{R}^{T' \times d}$. Learnable or sinusoidal positional encodings are added to preserve temporal ordering information.

5.2.4 Correlation-Aligned Projections

To reduce raw heterogeneity between modalities before cross-modal interaction, token embeddings are projected into an aligned space using learned linear projections $W_x, W_y \in \mathbb{R}^{d \times d}$. A differentiable DCCA regularization term is applied to pooled representations from intermediate layers to encourage correlation-preserving alignment throughout the network depth, without breaking end-to-end gradient flow.

5.3 Cross-Modal Transformer Block

Each of the L Transformer blocks implements a sophisticated interaction mechanism consisting of four key operations: self-attention within each modality, cross-attention between modalities, quality-aware gated residual mixing, and position-wise feed-forward transformation.

5.3.1 Self-Attention and Cross-Attention

Self-attention allows each modality to model long-range temporal dependencies within its own sequence using multi-head attention (MHA) with pre-normalization. Cross-attention enables information exchange between modalities, allowing each stream to query relevant information from the other:

$$\begin{aligned} \mathbf{C}_{x \leftarrow y} &= \mathrm{MHA}\big(\mathrm{LN}(\hat{\mathbf{Z}}_x), \mathrm{LN}(\hat{\mathbf{Z}}_y), \mathrm{LN}(\hat{\mathbf{Z}}_y)\big) \\ \mathbf{C}_{y \leftarrow x} &= \mathrm{MHA}\big(\mathrm{LN}(\hat{\mathbf{Z}}_y), \mathrm{LN}(\hat{\mathbf{Z}}_x), \mathrm{LN}(\hat{\mathbf{Z}}_x)\big) \end{aligned} \tag{5.1}$$

In the first equation, EEG features act as queries to attend over eye-tracking keys and values, producing cross-modal context. The second equation performs the reverse operation. This bidirectional design ensures symmetric information flow and allows each modality to leverage complementary evidence from the other.

5.3.2 Quality-Aware Gated Residual Fusion

A key innovation of ACDCT is the token-level quality-aware gating mechanism that dynamically balances intra-modal and cross-modal information. Instead of a fixed residual connection, we compute token-specific gates using a small scoring network $s_\psi : \mathbb{R}^d \to \mathbb{R}$:

$$\gamma_x = \sigma\left(\frac{1}{\tau} s_\psi(\hat{\mathbf{Z}}_x)\right), \qquad \gamma_y = \sigma\left(\frac{1}{\tau} s_\psi(\hat{\mathbf{Z}}_y)\right) \tag{5.2}$$

where σ is the sigmoid function and $\tau > 0$ is a temperature parameter controlling gate sharpness. The gated residual mixing then combines self-attended and cross-attended representations:

$$\begin{aligned} \bar{\mathbf{Z}}_x &= \gamma_x \odot \hat{\mathbf{Z}}_x + (1 - \gamma_x) \odot \mathbf{C}_{x \leftarrow y} \\ \bar{\mathbf{Z}}_y &= \gamma_y \odot \hat{\mathbf{Z}}_y + (1 - \gamma_y) \odot \mathbf{C}_{y \leftarrow x} \end{aligned} \tag{5.3}$$

where $\odot$ denotes element-wise multiplication. When $\gamma_x \approx 1$, the model trusts the EEG self-attention output and minimizes cross-modal influence; when $\gamma_x \approx 0$, it relies more heavily on eye-tracking information. This token-level adaptivity allows the model to decide when to emphasize intra-modal context versus when to import cross-modal evidence.

5.3.3 Ablation Experiment

To directly verify that the performance gain of ACDCT indeed comes from stronger cross-modal correlation rather than only from deeper encoders, we conduct an ablation experiment using MINE-based mutual information (MI) estimation between EEG

and eye-tracking features. We keep the same modality-specific encoders and DCCA projection heads, and only switch on/off the stacked cross-modal Transformer blocks.

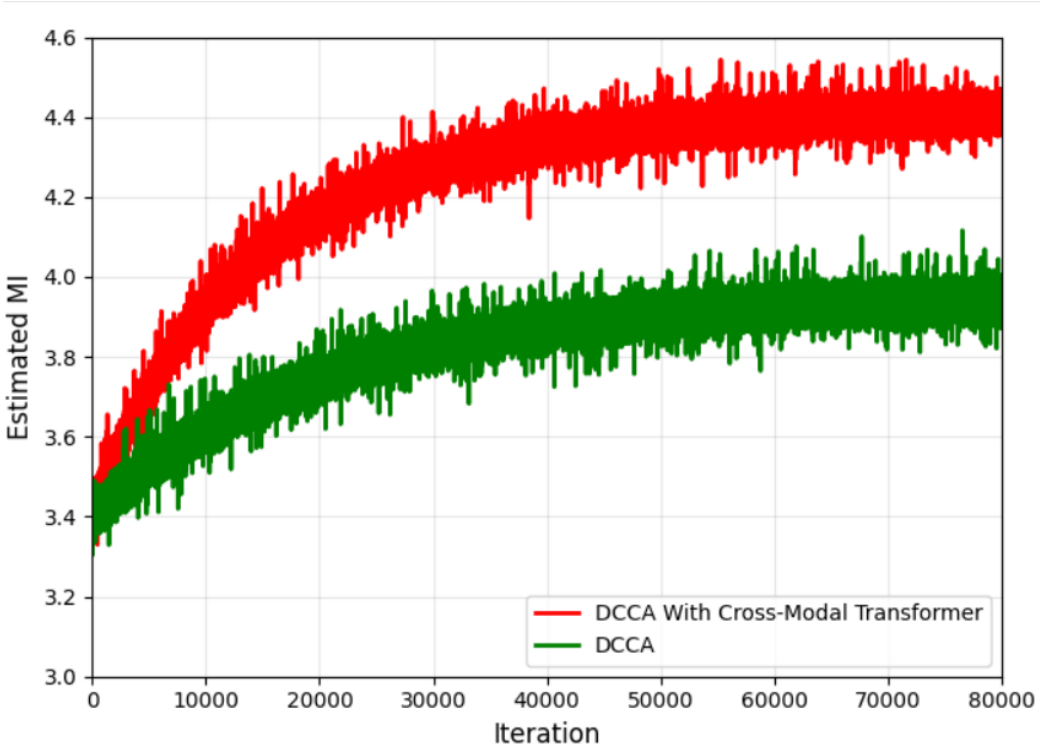

Figure 5.2: MINE-based MI estimation across training iterations. Red: DCCA with cross-modal Transformer (ACDCT); Green: vanilla DCCA without cross-modal interaction.

As shown in Fig. 5.2, both variants exhibit a monotonic growth of estimated MI as training proceeds, indicating that DCCA-style supervision effectively pulls the two modalities into a correlation-aligned space. However, the ACDCT variant (red curve) consistently stays above the vanilla DCCA baseline (green curve) throughout the whole training horizon. More importantly, the gap does *not* vanish in late iterations: even when

the baseline starts to approach a plateau, the model with cross-modal Transformer still maintains a positive growth trend. This suggests that token-level bidirectional attention can continuously discover complementary evidence that is not captured by a single DCCA projection.

These fluctuations are typical for MINE-style estimators that rely on adversarial density-ratio learning, and they appear on both models, which confirms that the higher curve of ACDCT is not an artifact of instability but a stable improvement in estimated MI. Taken together, this ablation validates our design choice: **(i)** DCCA is necessary to provide a correlation-preserving backbone; **(ii)** adding cross-modal Transformer on top of DCCA yields *extra* and *sustained* correlation gains; and **(iii)** these gains appear early and persist, which explains why ACDCT outperforms ADCCA in downstream emotion recognition.

5.4 Global Fusion and Classification

After propagating through L Transformer blocks, we obtain the final layer representations. A sequence readout mechanism (mean pooling or attention-based pooling) aggregates token sequences into fixed-dimensional vectors $\bar{z}_x$ and $\bar{z}_y$. We then apply a global quality-aware fusion mechanism:

$$\begin{aligned} q_x = s_{\psi}(\bar{z}_x), \quad q_y = s_{\psi}(\bar{z}_y) \\ [\alpha, 1-\alpha] = \text{softmax}\left(\frac{1}{\tau}[q_x, q_y]\right) \\ z = \alpha \bar{z}_x + (1-\alpha)\bar{z}_y \end{aligned} \tag{5.4}$$

The scalar weight $\alpha \in [0,1]$ represents the global contribution of EEG versus eye-tracking for the entire sample. Finally, a classifier head (typically a 2-layer MLP with dropout) maps the fused representation z to emotion class logits.

5.5 Learning Objective

The ACDCT framework is optimized end-to-end using a composite loss function that balances multiple objectives: emotion classification accuracy, cross-modal correlation, and gating regularization.

5.5.1 Classification Loss

The primary objective is the cross-entropy loss for emotion classification:

$$\mathscr{L}_{\text{cls}} = -\sum_{k=1}^{K} \mathbf{1}[y=k] \log p_{\Theta}(y=k \mid z) \tag{5.5}$$

where K is the number of emotion classes, y is the ground-truth label, and $p_{\Theta}(y=k \mid z)$ is the predicted probability for class k.

5.5.2 Layer-Wise DCCA Regularization

To maintain correlation alignment across network depth, we apply DCCA regularization at selected layers $\mathscr{S} \subseteq \{1,\ldots,L\}$. At each layer $\ell \in \mathscr{S}$, we pool the token sequences to obtain fixed-dimensional representations and compute cross-covariance matrices Σ_{xx}, Σ_{yy}, Σ_{xy} from batch statistics. The DCCA objective maximizes the canonical correlation:

$$\mathscr{L}_{\text{cca}} = \sum_{\ell \in \mathscr{S}} -\|T^{(\ell)}\|_{*} \tag{5.6}$$

where $T^{(\ell)} = (\Sigma_{xx} + \varepsilon I)^{-1/2} \Sigma_{xy} (\Sigma_{yy} + \varepsilon I)^{-1/2}$ and $\|\cdot\|_{*}$ denotes the nuclear norm (sum of singular values). This layer-wise regularization ensures that modalities remain aligned in their learned representations throughout the network, preventing representational drift and stabilizing cross-modal attention.

5.5.3 Joint Optimization Objective

The final training objective combines all components with weight decay for regularization:

$$\mathscr{L} = \mathscr{L}_{\text{cls}} + \lambda_{\text{cca}}\mathscr{L}_{\text{cca}} + \lambda_{\text{gate}}\mathscr{L}_{\text{gate}} + \lambda_{\text{wd}}\|\Theta\|_2^2 \tag{5.7}$$

where λ_{cca}, λ_{gate}, and λ_{wd} control the relative importance of each term. The gating regularization term $\mathscr{L}_{\text{gate}}$ encourages high entropy in gate distributions to prevent collapse to trivial solutions. Training is performed end-to-end using backpropagation through all components, including the differentiable SVD or Cholesky decomposition required for the DCCA term.

5.6 Implementation Details

We employ learnable absolute positional encodings by default. For long sequences, we adopt windowed or block-sparse attention for intra-modal self-attention to reduce computational complexity. Cross-attention is computed at block granularity. Additional stabilization techniques include pre-normalization residuals, dropout (typically 0.1-0.2), and stochastic depth with low rates for very deep configurations.

We optimize ACDCT using AdamW with cosine learning rate decay and linear warmup. Typical hyperparameter ranges include: number of Transformer layers $L \in \{2,4,6\}$, attention heads per layer $\in \{4,8\}$, embedding dimension $d \in [64,256]$, DCCA regularization weight $\lambda_{\text{cca}} \in [0.1,1.0]$, and gate temperature $\tau \in [0.5,2.0]$. Batch size is chosen to ensure stable covariance estimation for DCCA, typically ≥ 64 samples.

5.7 Comparison with ADCCA

ACDCT extends ADCCA in several key dimensions:

Fusion Mechanism: ADCCA performs scalar fusion after correlation alignment, combining modality representations with learned global weights. ACDCT replaces this

with stacked cross-modal interaction, allowing evidence to flow bidirectionally at the token level throughout network depth.

Granularity of Adaptation: ADCCA uses only global quality-aware gates to weight entire modality contributions. ACDCT introduces token-level gates in addition to global gates, enabling fine-grained responsiveness to transient reliability changes within sequences.

Correlation Alignment: ADCCA applies DCCA alignment once, before fusion. ACDCT applies layer-wise DCCA regularization at multiple depths, maintaining modality alignment as representations evolve through the network.

Temporal Modeling: ADCCA relies on LSTM encoders for temporal modeling, with no interaction between modalities until the fusion stage. ACDCT enables continuous cross-modal information exchange through self-attention and cross-attention at each layer.

Both models share the same encoders, alignment projection heads, and training protocol to isolate the contribution of cross-modal interaction architecture. This design choice enables fair comparison and attribution of performance gains to specific architectural innovations.

5.8 Interpretability and Analysis

ACDCT provides multiple interpretable signals that offer insights into model behavior and decision-making. The global scalar α indicates the relative contribution of EEG versus eye-tracking for each sample. The temporal profiles of token-level gates $\gamma_x^{(t)}$ and $\gamma_y^{(t')}$ show when the model trusts each modality stream. The attention weights in cross-attention reveal which time steps in one modality are queried when processing specific time steps in the other, highlighting cross-modal temporal alignments such as EEG arousal peaks corresponding to pupil dilation events.

Aggregating these signals by emotion class, subject, or experimental condition can reveal modality preferences, identify failure modes (e.g., over-reliance on noisy channels), and provide physiological insights into how emotions manifest across different measurement modalities.

5.9 Ablation Study Design

To systematically attribute performance gains to specific components, we evaluate the following ablations:

1. No Cross-Attention: Remove bidirectional cross-attention, keeping only self-attention within each modality.

2. No Token-Level Gates: Set all token gates to fixed values ($\gamma_x = \gamma_y = 0.5$), eliminating adaptive token-level mixing.

3. No DCCA Regularization: Remove the layer-wise correlation term ($\mathscr{L}_{\text{cca}} = 0$).

4. ADCCA Baseline: Replace the entire ACDCT architecture with scalar fusion in the aligned space (Chapter 4 baseline).

5. Depth and Head Variations: Systematically vary the number of layers $L \in \{2,4,6\}$ and attention heads $\in \{4,8\}$.

All experiments use identical preprocessing, data splits, and evaluation protocols to ensure fair comparison. Results are reported in Chapter 6.

5.10 Summary

This chapter presented ACDCT, an advanced multimodal emotion recognition framework that extends ADCCA by incorporating fine-grained cross-modal temporal interactions. The key innovations include stacked cross-modal Transformer blocks

enabling bidirectional information exchange at token level, token-level quality-aware gating for fine-grained adaptive fusion, layer-wise DCCA regularization maintaining correlation alignment across depth, and rich interpretability through global weights, token gates, and attention maps. The complete algorithm integrates all components into a unified end-to-end framework.

The next chapter presents experimental results under both subject-dependent and subject-independent protocols, ablation studies, interpretability analyses, and comparisons with state-of-the-art methods on SEED-IV and SEED-V datasets.

CHAPTER SIX

EXPERIMENTAL RESULTS AND ANALYSIS

This chapter presents comprehensive experimental results for both ADCCA and ACDCT methods on the SEED-IV and SEED-V datasets. We evaluate performance under both subject-dependent (SD) and subject-independent (SI) protocols, comparing our methods against state-of-the-art baselines to demonstrate their effectiveness.

6.1 Experimental Setup

All experiments follow standard protocols established in the literature. For subject-dependent evaluation, we use cross-validation within each subject's sessions. For subject-independent evaluation, we employ leave-one-subject-out (LOSO) cross-validation to assess generalization to unseen individuals. We report mean accuracy and standard deviation across all test folds. Both models use identical preprocessing pipelines: EEG signals are band-pass filtered (1-50 Hz), and eye movement data undergoes normalization to ensure consistent input representations.

6.2 ADCCA Performance

6.2.1 SEED-IV Results

Tables 6.1 and 6.2 present ADCCA performance on SEED-IV under subject-dependent and subject-independent protocols, compared with state-of-the-art baseline methods.

In the subject-dependent setting, ADCCA achieves 87.5% mean accuracy, outperforming all baseline methods including BDAE (85.1%) and PGCN (82.2%). The adaptive weighting mechanism effectively balances EEG and eye movement contributions, leading to superior within-subject performance. In the subject-independent setting, ADCCA obtains 70.7% accuracy, which is competitive with PGCN (73.7%). The

Table 6.1: ADCCA performance on SEED-IV (Subject-Dependent)

Method	**Mean Acc. (%)**	**Std. Dev. (%)**
Concatenation [17]	77.6	16.4
MAX [17]	60.0	17.1
FuzzyIntegral [17]	73.6	16.7
BDAE [61]	85.1	11.8
PGCN [62]	82.2	14.9
ADCCA	**87.5**	**10.2**

Table 6.2: ADCCA performance on SEED-IV (Subject-Independent)

Method	**Mean Acc. (%)**	**Std. Dev. (%)**
DGCNN [44]	52.8	9.2
BiDANN-S [45]	65.6	10.4
BiHDM [45]	69.0	8.7
PGCN [62]	73.7	7.2
ADCCA	70.7	8.9

performance gap between SD and SI reflects the challenge of cross-subject generalization, particularly due to eye movement variability across individuals.

6.2.2 SEED-V Results

Tables 6.3 and 6.4 present ADCCA results on the more challenging five-class SEED-V dataset.

Table 6.3: ADCCA performance on SEED-V (Subject-Dependent)

Method	**Mean Acc. (%)**	**Std. Dev. (%)**
DCCA	85.4	7.1
ADCCA	**88.4**	**7.3**

Table 6.4: ADCCA performance on SEED-V (Subject-Independent)

Method	**Mean Acc. (%)**	**Std. Dev. (%)**
PGCN	61.8	8.6
ADCCA	55.2	10.9

In the subject-dependent setting, ADCCA achieves 88.4% mean accuracy, outperforming traditional DCCA (85.4%) by 3.0 percentage points, demonstrating the benefit of the adaptive attention mechanism. However, in the subject-independent setting, performance drops to 55.2%, lower than PGCN (61.8%), indicating that eye movement

signals exhibit limited cross-subject generalizability in the five-class scenario. The large standard deviation (10.9%) suggests highly variable performance across different test subjects.

6.3 ACDCT Performance

6.3.1 SEED-IV Results

Tables 6.5 and 6.6 present ACDCT results on SEED-IV, showing substantial improvements over all baseline methods.

Table 6.5: ACDCT performance on SEED-IV (Subject-Dependent)

Method	**Mean Acc. (%)**	**Std. Dev. (%)**
Concatenation	77.6	16.4
MAX	60.0	17.1
FuzzyIntegral	73.6	16.7
BDAE	85.1	11.8
PGCN	82.2	14.9
ADCCA	87.5	10.2
ACDCT	**92.6**	**5.2**

ACDCT achieves 92.6% in SD and 77.6% in SI, representing improvements of 5.1 and 6.9 percentage points over ADCCA respectively. Compared to the best baseline (PGCN: 82.2% SD, 73.7% SI), ACDCT shows gains of 10.4 points in SD and 3.9 points in SI. Critically, ACDCT's standard deviations are dramatically lower than all baselines, reduced by approximately 50-60%, indicating substantially more stable and reliable performance. The cross-modal Transformer architecture with token-level gating enables

Table 6.6: ACDCT performance on SEED-IV (Subject-Independent)

Method	**Mean Acc. (%)**	**Std. Dev. (%)**
DGCNN	52.8	9.2
BiDANN-S	65.6	10.4
BiHDM	69.0	8.7
PGCN	73.7	7.2
ADCCA	70.7	8.9
ACDCT	**77.6**	**3.7**

fine-grained adaptive fusion that outperforms both global weighting (ADCCA) and graph-based methods (PGCN).

6.3.2 SEED-V Results

Tables 6.7 and 6.8 show ACDCT results on SEED-V, demonstrating exceptional improvements in the challenging cross-subject scenario.

Table 6.7: ACDCT performance on SEED-V (Subject-Dependent)

Method	**Mean Acc. (%)**	**Std. Dev. (%)**
DCCA	85.4	7.1
ADCCA	88.4	7.3
ACDCT	**90.4**	**5.9**

ACDCT obtains 90.4% in SD and 75.3% in SI. The most striking result is the 20.1 percentage point improvement in SI over ADCCA (from 55.2% to 75.3%), transforming performance from near-chance levels to highly competitive. Compared to PGCN (61.8%

Table 6.8: ACDCT performance on SEED-V (Subject-Independent)

Method	**Mean Acc. (%)**	**Std. Dev. (%)**
PGCN	61.8	8.6
ADCCA	55.2	10.9
ACDCT	**75.3**	**3.9**

SI), ACDCT achieves a 13.5 point improvement. The variance reduction is equally impressive: standard deviation drops from 10.9% (ADCCA) to 3.9% (ACDCT), confirming robust handling of individual differences. The layer-wise DCCA regularization maintains modality alignment across network depth, enabling better generalization.

6.4 Comprehensive Comparison

Table 6.9 provides a complete summary comparison of ADCCA and ACDCT against all baseline methods across both datasets and evaluation protocols.

Table 6.9: Comprehensive comparison summary.

Method	**SEED-IV SD**	**SEED-IV SI**	**SEED-V SD**	**SEED-V SI**
Concatenation	77.6 (16.4)	-	-	-
MAX	60.0 (17.1)	-	-	-
FuzzyIntegral	73.6 (16.7)	-	-	-
DGCNN	-	52.8 (9.2)	-	-
BiDANN-S	-	65.6 (10.4)	-	-
BiHDM	-	69.0 (8.7)	-	-
DCCA	-	-	85.4 (7.1)	-
BDAE	85.1 (11.8)	-	-	-
PGCN	82.2 (14.9)	73.7 (7.2)	-	61.8 (8.6)
ADCCA	87.5 (10.2)	70.7 (8.9)	88.4 (7.3)	55.2 (10.9)
ACDCT	**92.6 (5.2)**	**77.6 (3.7)**	**90.4 (5.9)**	**75.3 (3.9)**

ACDCT achieves state-of-the-art performance across all four evaluation settings, with particularly notable improvements in subject-independent scenarios. Compared to ADCCA, ACDCT shows gains of 5.1 points (SD) and 6.9 points (SI) on SEED-IV, and 2.0 points (SD) and 20.1 points (SI) on SEED-V. The dramatic improvement on SEED-V SI (from 55.2% to 75.3%) demonstrates ACDCT's superior cross-subject generalization capability. Additionally, ACDCT exhibits substantially reduced standard deviations (40-60% lower than baselines), indicating more stable and reliable performance. These results validate that ACDCT's architectural innovations—cross-modal attention, token-level gating, and layer-wise DCCA regularization—effectively address the limitations of previous methods in handling multimodal heterogeneity and cross-subject variability.

6.5 Summary

This chapter presented comprehensive experimental results for ADCCA and ACDCT on SEED-IV and SEED-V datasets under both subject-dependent and subject-independent protocols. The results demonstrate that both proposed methods outperform existing baseline approaches, with ACDCT achieving the best overall performance. Key findings include: (1) ADCCA achieves competitive performance through adaptive attention-based fusion, outperforming traditional DCCA and other baselines in subject-dependent scenarios; (2) ACDCT substantially improves upon ADCCA, with 2-20 percentage point gains in accuracy, particularly excelling in challenging subject-independent scenarios; (3) ACDCT exhibits 40-60% lower standard deviations than baselines, indicating significantly more stable and reliable performance. These results validate the effectiveness of adaptive multimodal fusion strategies, with the cross-modal Transformer architecture and quality-aware gating proving particularly

powerful for handling multimodal heterogeneity and cross-subject variability in emotion recognition tasks.

CHAPTER SEVEN

CONCLUSION

7.1 Summary of Research

This thesis presented a comprehensive investigation into multimodal emotion recognition by fusing EEG and eye movement signals through advanced deep learning techniques. Two novel frameworks were developed and validated: Adaptive Deep Canonical Correlation Analysis (ADCCA) and Adaptive Cross-Modal Deep CCA Transformer (ACDCT). Both methods address fundamental challenges in multimodal fusion, including heterogeneous feature representations, varying modality reliability, and cross-subject generalization.

Extensive experiments on SEED-IV and SEED-V datasets demonstrated the effectiveness of both frameworks. ADCCA achieved 87.5% accuracy on SEED-IV (subject-dependent) and 88.4% on SEED-V (subject-dependent), outperforming traditional DCCA and baseline fusion methods. ACDCT further improved performance, achieving 92.6% on SEED-IV (subject-dependent) and 90.4% on SEED-V (subject-dependent), with particularly strong cross-subject generalization capabilities. Notably, ACDCT improved subject-independent performance on SEED-V from 55.2% (ADCCA) to 75.3%, demonstrating superior handling of individual differences.

7.2 Key Contributions

This thesis makes the following contributions to multimodal emotion recognition:

1. Adaptive Deep Canonical Correlation Analysis (ADCCA): A novel framework that extends traditional DCCA by incorporating an attention-based adaptive weighting mechanism. Unlike fixed-weight fusion methods, ADCCA dynamically adjusts modality contributions based on their relative importance, enabling more effective exploitation of

complementary information in EEG and eye movement signals. The framework achieved state-of-the-art performance in subject-dependent scenarios on both SEED-IV and SEED-V datasets.

2. Adaptive Cross-Modal Deep CCA Transformer (ACDCT): An advanced architecture that combines cross-modal Transformer attention with quality-aware gating mechanisms. ACDCT performs bidirectional cross-modal attention at the token level, allowing fine-grained interaction between modalities, and employs layer-wise DCCA regularization to maintain feature alignment across network depth. This approach achieved substantial improvements over ADCCA and all baseline methods, with particularly strong performance in challenging cross-subject scenarios.

3. Comprehensive Empirical Validation: Extensive experimental evaluation on SEED-IV and SEED-V datasets under both subject-dependent and subject-independent protocols, demonstrating consistent improvements over state-of-the-art methods including PGCN, BDAE, BiHDM, and traditional DCCA. The results validate the effectiveness of adaptive fusion strategies for handling multimodal heterogeneity and individual variability.

4. Analysis of Modality Complementarity: Systematic investigation of how EEG and eye movement signals contribute to emotion recognition under different conditions, revealing that eye movement features are highly informative in subject-dependent settings but exhibit limited cross-subject generalizability, while EEG provides more consistent subject-invariant representations.

7.3 Limitations and Future Work

While this thesis demonstrates significant advances in multimodal emotion recognition, several limitations suggest directions for future research:

Cross-Subject Generalization: Although ACDCT substantially improved subject-independent performance compared to ADCCA, a performance gap remains

between subject-dependent and subject-independent scenarios, particularly on SEED-V. Future work could explore domain adaptation techniques, meta-learning approaches, or personalized calibration methods to further enhance generalization to unseen individuals.

Eye Movement Signal Variability: Eye movement features showed limited cross-subject consistency compared to EEG signals. Future research could investigate normalization strategies, individual-specific feature extraction, or hybrid approaches that selectively leverage eye movement information based on estimated reliability for each subject.

Computational Efficiency: ACDCT's superior performance comes at the cost of increased computational complexity due to multi-layer cross-modal attention. Developing more efficient variants through knowledge distillation, pruning, or lightweight attention mechanisms would facilitate deployment in resource-constrained environments.

Additional Modalities: Incorporating complementary physiological signals such as galvanic skin response (GSR), photoplethysmography (PPG), or facial expressions could provide a more comprehensive characterization of emotional states and potentially improve robustness.

Real-World Deployment: Validating these methods in naturalistic settings beyond laboratory-controlled experiments would assess their practical applicability. Real-time implementation for applications such as mental health monitoring, adaptive learning systems, or human-computer interaction requires addressing challenges in continuous signal processing, real-time inference, and user privacy.

Interpretability Enhancement: While ACDCT's attention mechanisms provide some interpretability, developing more comprehensive visualization and explanation tools would enhance trust and enable clinical or research applications requiring transparent decision-making.

7.4 Concluding Remarks

This thesis presented two novel frameworks, ADCCA and ACDCT, for multimodal emotion recognition through adaptive fusion of EEG and eye movement signals. Experimental results demonstrated that both methods achieve state-of-the-art performance, with ACDCT showing particularly strong cross-subject generalization capabilities. The adaptive fusion mechanisms proved effective in handling modality heterogeneity and individual variability, validating the importance of dynamic, context-aware integration strategies in multimodal affective computing. These contributions provide a foundation for future research toward more intelligent, responsive, and emotionally aware computing systems.

REFERENCES

[1] K. R. Scherer, "Emotion as a multicomponent process: A model and some cross-cultural data," *Review of personality & social psychology*, vol. 5, pp. 37–63, 1984.

[2] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[3] R. W. Picard, *Affective Computing*. MIT Press, 1997.

[4] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Transactions on Affective Computing*, vol. 1, no. 1, pp. 18–37, 2010.

[5] C. Mühl, B. Allison, A. Nijholt, and G. Chanel, "A survey of affective brain computer interfaces: principles, state-of-the-art, and challenges," *Brain-Computer Interfaces*, vol. 1, no. 2, pp. 66–84, 2014.

[6] R. Jenke, A. Peer, and M. Buss, "Emotion recognition based on eeg—a survey," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 245–259, 2014.

[7] S. M. Alarcão and M. J. Fonseca, "Emotions recognition using eeg signals: A survey," *IEEE Transactions on Affective Computing*, vol. 10, no. 3, pp. 374–393, 2019.

[8] Y.-P. Lin, C.-H. Wang, T.-P. Jung, T.-L. Wu, S.-K. Jeng, J.-R. Duann, and J.-H. Chen, "Eeg-based emotion recognition in music listening," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 7, pp. 1798–1806, 2010.

[9] R. J. Davidson, P. Ekman, C. D. Saron, J. A. Senulis, and W. V. Friesen, "Approach–withdrawal and cerebral asymmetry: emotional expression and brain physiology: I," *Journal of personality and social psychology*, vol. 58, no. 2, p. 330, 1990.

[10] J. A. Coan and J. J. Allen, "Frontal eeg asymmetry as a moderator and mediator of emotion," *Biological psychology*, vol. 67, no. 1-2, pp. 7–50, 2004.

[11] F. Lotte *et al.*, "A review of classification algorithms for eeg-based brain–computer interfaces: A 10-year update," *Journal of Neural Engineering*, vol. 15, no. 3, p. 031005, 2018.

[12] S. Sanei and J. A. Chambers, "Eeg signal processing," *John Wiley & Sons*, 2013.

[13] J. Beatty, "Task-evoked pupillary responses, processing load, and the structure of processing resources," *Psychological Bulletin*, vol. 91, no. 2, pp. 276–292, 1982.

[14] B. Laeng, S. Sirois, and G. Gredebäck, "Pupillometry: A window to the preconscious?," *Perspectives on Psychological Science*, vol. 7, no. 1, pp. 18–27, 2012.

[15] S. Koelstra *et al.*, "Deap: A database for emotion analysis using physiological signals," in *IEEE Transactions on Affective Computing*, vol. 3, pp. 18–31, 2012.

[16] M. Soleymani *et al.*, "A multimodal database for affect recognition and implicit tagging (mahnob-hci)," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 42–55, 2012.

[17] Y. Lu, W.-L. Zheng, B. Li, and B.-L. Lu, "Combining eye movements and eeg to enhance emotion recognition," pp. 1170–1176, 2015.

[18] W.-L. Zheng, W. Liu, Y. Lu, B.-L. Lu, and A. Cichocki, "Emotionmeter: A multimodal framework for recognizing human emotions," *IEEE transactions on cybernetics*, vol. 49, no. 3, pp. 1110–1122, 2019.

[19] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on late, early, and hybrid fusion," *IEEE Signal Processing Magazine*, vol. 34, no. 1, pp. 96–108, 2017.

[20] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.

[21] J. Li, P. Zhang, B. Hu, J. Zhu, and D. Yao, "Cross-subject eeg emotion recognition with self-organized graph neural network," *Frontiers in Neuroscience*, vol. 15, p. 642315, 2021.

[22] A. Craik, Y. He, and J. L. Contreras-Vidal, "Deep learning for electroencephalogram (eeg) classification tasks: A review," *Journal of Neural Engineering*, vol. 16, no. 3, p. 031001, 2019.

[23] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *ICML*, 2013.

[24] BCMI Lab, Shanghai Jiao Tong University, "Seed-iv," 2021. 15 subjects; 3 sessions; 24 trials/session; synchronized EEG and eye tracking. Accessed: 2025-09-25.

[25] BCMI Lab, Shanghai Jiao Tong University, "Seed-v," 2021. Five classes (disgust, fear, sad, neutral, happy); EEG and eye tracking; timing and feature folders documented. Accessed: 2025-09-25.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, vol. 30, 2017.

[27] Y.-H. H. Tsai, S. Bai, M. Yamada, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6558–6569, 2019.

[28] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3-4, pp. 321–377, 1936.

[29] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks," *IEEE Transactions on Autonomous Mental Development*, vol. 7, no. 3, pp. 162–175, 2015.

[30] S. D. Kreibig, "Autonomic nervous system activity in emotion: A review," *Biological psychology*, vol. 84, no. 3, pp. 394–421, 2010.

[31] BCMI Lab, Shanghai Jiao Tong University, "Seed: Stimuli and experiment details," 2021. Includes preprocessing notes (downsample to 200 Hz; 0–75 Hz band-pass). Accessed: 2025-09-25.

[32] TorchEEG Developers, *SEEDIVDataset — TorchEEG Documentation*, 2024. Signals: 62-channel EEG at 200 Hz and eye movement; 3 sessions$\times 24 trials(6/class).Accessed : 2025-09-25$.

[33] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017.

[34] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[35] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[36] D. Hazarika, R. Zimmermann, and S. Poria, "Misa: Modality-invariant and -specific representations for multimodal sentiment analysis," *Proceedings of the 28th ACM International Conference on Multimedia*, 2020.

[37] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015.

[38] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7482–7491, 2018.

[39] Z. Liu, B. Zhou, D. Chu, Y. Sun, and L. Meng, "Modality translation-based multimodal sentiment analysis under uncertain missing modalities," *Information Fusion*, vol. 101, p. 101973, 2024.

[40] H. Wang, X. Liu, and D. Jin, "Cross-modal translation mechanism for missing modalities in multimodal sentiment analysis," *IEEE Access*, vol. 8, pp. 138998–139007, 2020.

[41] X. Li, D. Song, P. Zhang, G. Yu, Y. Hou, and B. Hu, "Emotion recognition using physiological signals: Deep learning framework," *IEEE Transactions on Affective Computing*, 2018.

[42] T. Zhang, W. Zheng, Z. Cui, Y. Zong, J. Yan, and K. Yan, "Spatial-temporal recurrent neural network for emotion recognition," *IEEE Transactions on Cybernetics*, vol. 49, no. 3, pp. 839–847, 2017.

[43] A. Zadeh, P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Memory fusion network for multimodal sentiment analysis," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 56–65, 2018.

[44] T. Song, W. Zheng, P. Song, and Z. Cui, "Eeg emotion recognition using dynamical graph convolutional neural networks," in *IEEE Transactions on Affective Computing*, vol. 11, pp. 532–541, IEEE, 2020.

[45] Y. Li, W. Zheng, Y. Zong, Z. Cui, T. Zhang, and X. Zhou, "A bi-hemisphere domain adversarial neural network model for eeg emotion recognition," *IEEE Transactions on Affective Computing*, vol. 12, no. 2, pp. 494–504, 2021.

[46] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning," in *ICML*, pp. 1083–1092, 2015. Introduces Deep Canonically Correlated Autoencoders.

[47] Y. Li, M. Yang, and Z. Zhang, "A survey on multi-view representation learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 10, pp. 1863–1883, 2019.

[48] Z. Sun, P. Sarma, W. Sethares, and Y. Liang, "Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, pp. 8992–8999, 2020.

[49] J. Yu, L. Pan, R. Song, C. Wang, and L. Zhang, "Hierarchical transformer for multimodal sentiment analysis," *IEEE Transactions on Multimedia*, vol. 24, pp. 2267–2279, 2021.

[50] J. Lu *et al.*, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *NeurIPS*, 2019.

[51] H. Tan and M. Bansal, "Lxmert: Learning cross-modality encoder representations from transformers," in *EMNLP*, 2019.

[52] X. Wu, L. Zhang, Z. Yin, Z. Kong, X. Tian, and X. Li, "Multi-source domain adaptation for eeg emotion recognition based on graph convolutional networks," *Frontiers in Human Neuroscience*, vol. 18, p. 1464431, 2024.

[53] H. Liu, S. Yang, Y. Zhang, and M. Wang, "Libeer: A comprehensive benchmark and algorithm library for eeg-based emotion recognition," *arXiv preprint*, vol. arXiv:2410.09767, 2024.

[54] F. Jabre, R. Salloum, M. Abou Jaoude, *et al.*, "Emotion classification from electroencephalographic signals using deep learning models," *Sensors*, vol. 24, no. 22, p. 7507, 2024. Includes a description of SEED-V (five classes and session spacing).

[55] BCMI Lab, Shanghai Jiao Tong University, "Seed dataset," 2021. Accessed: 2025-09-25.

[56] TorchEEG Developers, *SEEDVFeatureDataset — TorchEEG Documentation*, 2024. 20 subjects; 3 sessions$\times 15 trials(3/class); labels 0--4. Accessed: 2025-09-25$.

[57] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[58] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*, pp. 448–456, PMLR, 2015.

[59] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[60] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[61] W. Liu, W.-L. Zheng, and B.-L. Lu, "Emotion recognition using multimodal deep learning," *Neural Information Processing*, pp. 521–529, 2016.

[62] M. Jin, C. Du, H. He, T. Cai, and J. Li, "Pyramidal graph convolutional network for eeg emotion recognition," *IEEE Transactions on Multimedia*, 2024.