POLITECNICO DI TORINO

Master's Degree in Data Science and Engineering

Master's Degree Thesis

Cross-Domain Multimodal Emotion Recognition with Progressive Fusion and Conservative Domain Adaptation

Supervisor:

Prof. Giuseppe Rizzo

Candidate:

Xiyang Zu

October, 2025

Abstract

Emotion recognition technology plays a crucial role in human-computer interaction and affective computing applications, with potential applications spanning from mental health monitoring to educational technology. However, the majority of existing emotion recognition systems suffer from significant performance degradation when deployed across different environments and datasets, limiting their practical applicability. This highlights the critical need for robust cross-domain solutions that can maintain performance consistency across varied real-world conditions. With the advancement of multimodal learning and domain adaptation techniques, attention-based fusion models have shown promising results in emotion recognition tasks. However, cross-domain emotion recognition still faces substantial challenges due to domain distribution shifts, limited computational resources, and the need to balance source and target domain performance. Existing domain adaptation methods often exhibit training instability and catastrophic forgetting, where aggressive adversarial training sacrifices source domain performance for marginal target domain improvements, restricting the practical deployment of these systems.

To address these issues, this study proposes a progressive multimodal fusion architecture with ultra-conservative domain adaptation for cross-domain emotion recognition. We systematically evaluate three fusion strategies to identify the optimal multimodal integration approach, and implement a novel conservative alpha scheduling mechanism to ensure training stability. In addition, computational efficiency optimization is incorporated to enable practical deployment in resource-constrained scenarios. Specifically, we design a comprehensive fusion comparison framework with three different strategies—truly early fusion, progressive middle fusion with

multi-stage cross-attention, and weighted late fusion—to examine their effectiveness in cross-domain scenarios. Furthermore, we focus on transferring knowledge from controlled laboratory settings (RAVDESS) to naturalistic environments (CREMA-D), implementing an ultra-conservative domain adaptation strategy with alpha scheduling ranging from 0.0001 to 0.1, compared to traditional methods that scale to 1.0, along with strategic target domain subset selection to reduce computational requirements by approximately 5×.

The results demonstrate that progressive middle fusion combined with conservative domain adaptation significantly outperforms baseline approaches in cross-domain transfer tasks. In speaker-independent evaluation, our method maintains high source domain performance while substantially improving target domain accuracy compared to zero-shot transfer baselines. The conservative alpha scheduling strategy achieves superior training stability with notably lower loss variance compared to traditional adversarial methods, while the efficient data balancing approach reduces training time and computational overhead without compromising performance quality. These results demonstrate the strong generalization ability and computational efficiency of the proposed approach, highlighting its potential for practical cross-domain emotion recognition deployment in real-world applications where both performance consistency and resource constraints are critical considerations.

Acknowledgments

This thesis marks not merely the conclusion of my graduate studies, but the closing chapter of a transformative period that has fundamentally shaped who I am today. I entered this program during the pandemic—a time of global uncertainty—and embarked on a solitary journey to a foreign land, carrying little more than ambition and curiosity about what lay ahead.

For much of these past years, I found myself lost. I struggled to open my textbooks. My thesis lay abandoned for months. I questioned every choice I had made, that was perhaps my darkest period. Must we achieve specific goals to consider ourselves successful? I no longer think so. Everyone struggles with doubt. Even now, I wrestle with conflicting values and anxiety about career paths and the future. But the real question became: how do we maintain curiosity about the world and enthusiasm for life despite our confusion? That difficult period taught me an alternative way of living. I took on various part-time jobs to support myself. I committed to fitness and running. I hiked in the mountains and watched sunsets by the sea. Through continuous exploration, I rediscovered the joy and meaning in life.

I am deeply grateful to my supervisor for his patience and guidance, for allowing me the freedom to explore while providing the structure I needed to succeed. To my parents and my older sister, who supported my pursuit despite the distance it created between us. And to the most important person in my life, the girl I deeply cherish: your sincerity and beauty have profoundly influenced who I am today. Your companionship gave me solace in moments of loneliness, and the distance between us has only deepened our appreciation for each other. I also want to thank my friends here in Italy. Your friendship has been a precious treasure in my life, and I will forever cherish the warmth you brought to my experience in this foreign land.

These years in Italy—marked by struggle and discovery, by setbacks and growth—have revealed to me what kind of life I truly want to live. As I close this chapter, I carry with me not just a degree, but profound gratitude for this season of my youth that taught me resilience, authenticity, and the courage to define success on my own terms.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Motivation and Background Emotion recognition is an essential building block in intelligent systems, enabling machines to interpret and respond to human affective states. This capability underpins numerous applications, such as adaptive human-computer interaction, emotion-aware healthcare monitoring, and intelligent educational systems. However, while models perform well in controlled laboratory settings, their performance often deteriorates significantly when applied in real-world scenarios that differ in population characteristics, environmental conditions, and data acquisition methods.

Modern deployment environments—particularly in healthcare, education, and edge computing—impose stringent constraints on privacy, computational efficiency, and generalizability. For example, healthcare applications demand local processing on mobile devices to preserve patient privacy. Educational tools must adapt to diverse student populations with minimal delay. Meanwhile, edge devices such as smart speakers and wearable sensors must operate under limited memory and energy budgets while delivering real-time responses. These real-world requirements expose a persistent limitation in emotion recognition systems: domain generalization.

This domain gap—the mismatch between training data (e.g., RAVDESS) and deployment data (e.g., CREMA-D)—is a fundamental barrier to practical emotion recognition. Models trained on controlled datasets often struggle when exposed to variations in acoustic background, speaker demographics, and sensor quality. Such performance degradation renders many models unusable in real-world settings, particularly in scenarios where domain-specific data collection is infeasible or unethical.

Moreover, traditional domain adaptation methods rely heavily on large

target datasets and computationally intensive training procedures, which are impractical in constrained environments[1], [2]. As a result, many advanced systems remain confined to academic research, while deployed solutions rely on simplistic and brittle heuristics.

Challenges in Cross-Domain Multimodal Emotion Recognition The core difficulty in cross-domain emotion recognition lies in managing three interdependent challenges: domain shift, computational constraints, and multimodal degradation.

Domain shift occurs across multiple dimensions. In healthcare, for instance, models must generalize from clean lab-recorded speech to noisy, uncontrolled clinical recordings. In education, systems trained on adult datasets must interpret children's emotional cues, which differ in vocal tone and facial expression dynamics. Even hardware differences—such as camera resolution or microphone type—can lead to substantial performance drops if not adequately addressed.

Computational constraints are particularly salient in edge and mobile environments, where memory and processing budgets are extremely limited. Traditional adversarial domain adaptation methods—though powerful—are often computationally prohibitive due to their reliance on gradient reversal and iterative optimization. These methods are incompatible with deployment on personal devices such as tablets, smartphones, or classroom hardware.

Multimodal fusion under domain shift introduces further complexity. Audio and visual modalities often degrade asymmetrically across domains. For example, visual data may be affected by lighting changes while audio remains clear, or vice versa. Traditional fusion strategies assume stable modality reliability, which is not guaranteed in real-world deployments. Moreover, fusion strategies must remain computationally lightweight to meet latency requirements.

Finally, limited access to target domain data—due to privacy regulations, data scarcity, or diversity in user populations—prevents traditional supervised learning from being feasible. Thus, any viable solution must function under small-data conditions while avoiding overfitting and instability.

Research Objectives and Contributions This thesis proposes a resource-efficient framework for cross-domain multimodal emotion recognition, specifically tailored to operate under realistic deployment constraints. The contributions are as follows:

1. Conservative Domain Adaptation: A novel adversarial training schedule is proposed using ultra-conservative alpha scaling. This avoids the

optimization instability commonly observed when domain discriminators overpower emotion classifiers in small-data settings.

2. Progressive Multimodal Fusion Architecture: A middle-fusion strategy is introduced, leveraging attention-based cross-modal interactions to capture complementary cues while being resilient to domain-specific modality degradation.

3. Computational Efficiency Enhancements: An efficient training strategy is implemented by sampling small, representative subsets of the target domain (e.g., 25% of CREMA-D) guided by meta-learning principles. This reduces computational load without compromising adaptation quality.

4. Robustness Evaluation Framework: The model is tested under controlled degradation scenarios—including audio noise injection, visual blur, and mixed conditions—to systematically evaluate real-world reliability.

Technical Innovations and Practical Impact Our innovations provide a comprehensive solution to bridge the gap between laboratory research and deployment-ready systems:

1. The alpha scheduling algorithm mathematically controls adversarial loss contribution, ensuring domain alignment without compromising task performance.

2. The efficient fusion design ensures low computational overhead, allowing deployment on consumer-grade or embedded hardware.

3. The robustness framework provides a principled methodology for benchmarking models under realistic degradation, improving interpretability and reliability.

These innovations collectively enable real-time, privacy-preserving emotion recognition on devices such as mobile tablets, classroom systems, and wearable sensors. Furthermore, this work supports broader goals of democratizing AI technologies by enabling their application in low-resource settings across healthcare, education, and consumer domains.

Summary This study focuses on deep learning–based multimodal emotion recognition under resource and privacy constraints, with experiments centered on data augmentation, progressive cross-modal fusion, and cross-domain generalization. We demonstrate that an ultra-conservative adversarial schedule together with few-shot target sampling improves robustness and adaptability across recording conditions while reducing computational cost. In

addition, progressive middle fusion captures complementary affective cues from audio and video and remains stable when modality reliability shifts. Finally, we evaluate pre-trained backbone choices (EmoCatcher encoder; TimeSformer with parameter-efficient tuning) and compare domain-adaptive versus non-adaptive systems, as well as audio-only and video-only baselines, in multi-class classification using accuracy, macro-F1, domain gap, and domain-confusion scores.

The rest of this thesis is organized as follows: Chapter 2 presents a literature review covering domain adaptation, multimodal fusion, and deployment challenges in affective computing. Chapter 3 describes our methodology, including the architecture design and training strategies. Chapter 4 details the experimental setup and results. Chapter 5 concludes the work and highlights its broader implications, then discusses limitations and future directions.

Chapter 2

State of the Art

2.1 Background and Applications of Emotion Recognition

Emotion recognition has emerged as a critical component in enabling machines to understand and respond to human affective states, bridging the gap between artificial intelligence systems and natural human communication. As comprehensively reviewed by Poria et al. [3], the ability to automatically detect and interpret emotions from multimodal signals—including speech, facial expressions, physiological signals, and text—has profound implications across diverse domains, transforming how humans interact with technology and each other.

Emotion Recognition in Human-Computer Interaction

In the realm of human-computer interaction (HCI), emotion-aware systems have revolutionized user experience by adapting their behavior based on detected emotional states [4]. Picard's seminal work on affective computing [5] laid the theoretical foundation for this field, proposing that computers should recognize, interpret, and simulate human emotions to achieve more natural and effective human-machine interaction. Building on this foundation, intelligent virtual assistants, conversational agents, and social robots now leverage emotion recognition to provide empathetic responses, thereby increasing user satisfaction and engagement [4].

Calvo et al.'s comprehensive handbook [4] documents how emotion recognition has been successfully deployed in customer service systems, where detecting customer dissatisfaction enables proactive intervention and improved service quality. The handbook also highlights applications in automotive safety, where driver emotion monitoring systems analyze facial expressions and vocal patterns in real-time to detect dangerous states such as fatigue, anger, or distraction, issuing alerts or activating autonomous driving features to reduce accident rates.

Applications in Healthcare and Mental Health

The healthcare sector has witnessed substantial benefits from emotion recognition technologies, particularly in mental health assessment and monitoring. As Cummins et al. [6] extensively review, traditional psychiatric evaluation relies heavily on subjective self-reports and clinical interviews, which can be biased or incomplete. Automated emotion recognition provides objective, continuous monitoring of patients' affective states, enabling early detection of conditions such as depression, anxiety, and post-traumatic stress disorder (PTSD).

Cummins et al. [6] document numerous studies demonstrating that automatic analysis of speech patterns can identify depressive symptoms with accuracy comparable to trained clinicians. The review highlights how vocal acoustic biomarkers—including prosody, speaking rate, and voice quality—serve as reliable indicators of mental health status. These findings have enabled teletherapy platforms to integrate emotion recognition for remote patient monitoring, particularly valuable during circumstances when in-person sessions are limited.

Beyond depression detection, Poria et al. [7] discuss applications in autism spectrum disorder (ASD) research, where emotion recognition technologies assist individuals who struggle with recognizing and expressing emotions. Interactive systems providing real-time feedback on emotional expressions serve as valuable training tools, helping patients develop social communication skills in controlled, supportive environments.

Educational Technology and Affective Learning

Educational technology has increasingly recognized the importance of students' emotional states in learning effectiveness. As detailed in Calvo et al.'s handbook [4], intelligent tutoring systems (ITS) equipped with emotion recognition capabilities can detect confusion, boredom, or frustration, adapting instructional strategies to maintain engagement and optimize

learning outcomes. The handbook cites empirical evidence showing that emotion-aware tutoring systems improve student performance by 15-20% compared to traditional systems by identifying moments of struggle and providing targeted support.

In massive open online courses (MOOCs) and remote learning environments, where direct teacher-student interaction is limited, automated emotion analysis helps instructors understand class-wide emotional trends [4]. The handbook discusses how analyzing aggregated emotional data reveals which course segments cause confusion or disengagement, informing iterative course improvements. Some platforms now employ real-time emotion monitoring to detect students requiring additional assistance, facilitating timely intervention [1, 2].

The integration of emotion recognition in educational contexts represents a shift toward affective learning paradigms, where pedagogical strategies are informed not only by cognitive metrics but also by learners' emotional states. This holistic approach acknowledges that emotions play a fundamental role in attention, memory formation, and motivation—key factors determining learning success.

Entertainment and Media Analysis

The entertainment industry leverages emotion recognition for content recommendation, audience analysis, and interactive experiences. Poria et al. [7] review how streaming platforms analyze viewers' emotional responses to recommend content matching their current mood or predicted preferences. During content production, filmmakers use emotion recognition to test audience reactions to different cuts, optimizing narrative pacing and emotional impact.

Video game development has particularly embraced emotion recognition to create adaptive, emotionally responsive gameplay. Calvo et al.'s handbook [4] describes games that detect player frustration and dynamically adjust difficulty levels, while horror games intensify suspense by monitoring fear responses. Virtual reality (VR) applications combine emotion recognition with immersive environments to create therapeutic experiences for phobia treatment or to enhance entertainment value through emotionally adaptive storytelling.

Market research represents another significant application domain. Poria et al. [7] note that companies employ emotion recognition to analyze consumer reactions to advertisements, product designs, and brand messaging. By capturing authentic emotional responses in naturalistic settings—rather than relying solely on post-exposure surveys—companies gain insights that inform

more effective marketing strategies. This approach addresses limitations of traditional market research methods, which often suffer from social desirability bias and poor recall.

The breadth of these applications underscores the transformative potential of emotion recognition technologies across human activity domains. However, as Poria et al. [7] emphasize, realizing this potential requires addressing fundamental challenges in multimodal signal processing, cross-context generalization, and ethical considerations regarding privacy and consent. The following sections examine technical approaches developed to address these challenges, focusing on audio-based methods (Section 2.2), video-based methods (Section 2.3), multimodal fusion strategies (Section 2.4), and domain adaptation techniques (Section 2.5) that enable emotion recognition systems to generalize across diverse datasets and real-world conditions.

2.2 Audio-based Emotion Recognition

Speech emotion recognition (SER) extracts affective states from acoustic and prosodic patterns in vocal signals. Emotions manifest through multiple acoustic dimensions—pitch variation, energy distribution, speaking rate, and spectral characteristics—making speech a remarkably rich channel for affective computing. Over the past two decades, the field has undergone a fundamental transformation: from carefully engineered acoustic features fed into conventional classifiers to deep learning architectures that discover hierarchical emotion representations directly from raw audio or spectrograms [8].

2.2.1 From Hand-crafted to Learned Representations

Traditional approaches to speech emotion recognition centered on extracting manually designed acoustic features. Researchers identified low-level descriptors (LLDs)—pitch (fundamental frequency F0), energy, formants, zero-crossing rate—and computed statistical functionals over these measures: means, standard deviations, extrema, and percentiles [5], [9]. The extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [10] exemplified this paradigm, offering 88 carefully curated features designed to capture emotion-relevant acoustic characteristics. These feature vectors were then classified using Support Vector Machines, Hidden Markov Models, or Gaussian Mixture Models.

This approach delivered moderate success on controlled laboratory datasets but revealed critical weaknesses. Manual feature engineering

demanded substantial domain expertise and proved labor-intensive. More fundamentally, hand-crafted features struggled to capture the complex temporal dynamics and subtle emotional nuances inherent in spontaneous speech. Performance degraded sharply when systems encountered diverse speakers, varied recording conditions, or cross-cultural expressions—precisely the scenarios encountered in real-world deployment [11].

The deep learning revolution fundamentally altered this landscape. Rather than manually specifying which acoustic patterns matter for emotion recognition, convolutional and recurrent neural networks learned hierarchical representations directly from data [12], [13]. A pivotal development was the adoption of mel-spectrograms as input representations—time-frequency visualizations that align with human auditory perception while capturing both spectral content and temporal evolution.

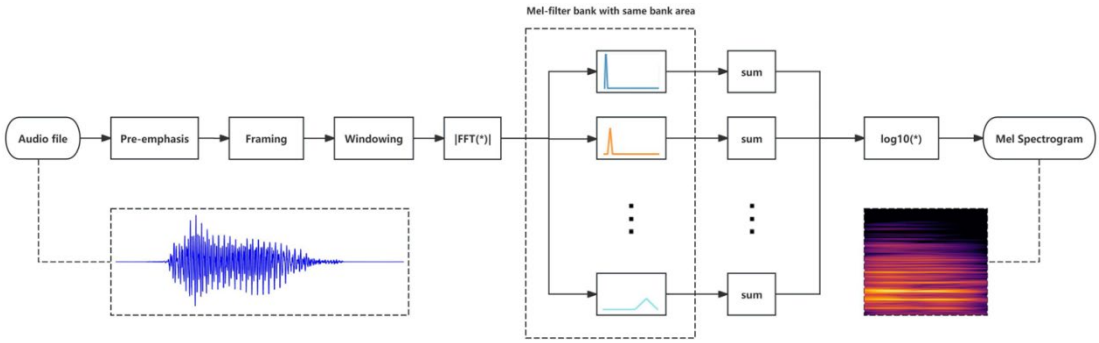

Figure 2.1: The technical architecture of converting audio files into Mel spectrograms from a raw audio file[14].

Mel-spectrogram Construction and Perceptual Relevance

The mel-spectrogram computation transforms raw audio waveforms into perceptually meaningful representations through several carefully designed stages. Given a raw audio signal $x(t)$ sampled at rate f_s (typically 22.05 kHz for speech), the process begins with Short-Time Fourier Transform (STFT) analysis. The signal is segmented into overlapping frames—commonly 2048 samples with 50% overlap (hop length of 1024 samples)—and each frame is multiplied by a window function (usually Hamming or Hann) to minimize spectral leakage. The discrete Fourier transform of each windowed frame produces a complex-valued spectrogram $S(f, t)$ representing frequency content evolving over time.

The magnitude spectrogram $|S(f, t)|^2$ captures power distribution across frequencies but uses a linear frequency scale misaligned with human auditory

perception. The human cochlea processes sound using a quasi-logarithmic frequency scale—low-frequency differences (e.g., 100 Hz vs. 200 Hz) are perceptually more salient than equivalent high-frequency differences (e.g., 5000 Hz vs. 5100 Hz). The mel scale addresses this through the transformation:

$$m = 2595 \log_{10} \left(1 + \frac{f}{100}\right) \tag{2.1}$$

where m represents mel-frequency and f represents linear frequency in Hertz. This logarithmic mapping concentrates frequency resolution where human hearing is most sensitive while coarsening resolution at higher frequencies where emotional prosody carries less information [14], [15].

Converting the magnitude spectrogram to mel-scale involves applying a bank of triangular filters spaced according to mel-scale intervals. Typically, 128 mel-filters span the frequency range 0-8000 Hz, though emotion recognition often limits the upper bound to 6000 Hz since most prosodic information concentrates below this threshold. Each filter performs weighted integration of magnitude spectrogram bins falling within its frequency range, producing a mel-spectrogram $M(m,t)$ with dimensions $[n_{mels} \times n_{frames}]$. Finally, logarithmic compression—converting to decibel scale via $\log_{10}(M(m,t))$—yields the final representation that Convolutional Neural Networks (CNNs) process.

This representation offers substantial advantages for emotion recognition. The mel-frequency warping emphasizes prosodic features crucial for affective perception—fundamental frequency variations, formant structures, and harmonic relationships—while de-emphasizing high-frequency spectral details less relevant for emotion discrimination [16]. The time-frequency structure naturally suits convolutional processing: frequency patterns analogous to image features can be detected through learned filters, while temporal evolution unfolds along the time axis.

Early deep learning approaches applied CNNs to mel-spectrograms, treating them analogously to images. Convolutional layers with small kernels (e.g., 3×3 or 5×5) learned hierarchical acoustic patterns: low-level spectral edges and textures in initial layers, mid-level phonetic structures in intermediate layers, and high-level emotional signatures in deeper layers [5], [6]. Pooling operations—max-pooling or average-pooling applied spatially—introduced local translation invariance, allowing learned filters to recognize acoustic patterns regardless of slight temporal or frequency shifts[17], [18], [19].

2.2.2 Recurrent Architectures and Temporal Modeling

While CNNs effectively capture local spectro-temporal patterns, emotions unfold through extended sequences requiring models that maintain and update internal states based on temporal context. Recurrent Neural Networks, particularly Long Short-Term Memory (LSTM) networks [20] and Gated Recurrent Units (GRUs) [21], became instrumental for modeling speech's inherently sequential nature.

The standard Recurrent Neural Networks (RNNs) [22] maintains a hidden state h_t that updates at each time step according to:

$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t + b_h) \tag{2.2}$$

where x_t represents input at time t, W_{hh} and W_{xh} are learnable weight matrices, and b_h is a bias vector. This formulation enables the network to accumulate information across time, with h_t theoretically encoding all previous inputs $x_1, x_2, \ldots, x_t$. However, vanilla RNNs suffer from vanishing and exploding gradients during backpropagation through time, making them incapable of capturing long-range dependencies spanning more than 5-10 time steps—insufficient for speech emotion recognition where affective cues may be distributed across entire utterances [23].

LSTM Architecture and Gating Mechanisms

LSTMs address these limitations through a sophisticated gating mechanism that regulates information flow [24]. The LSTM cell maintains two states: a hidden state h_t (analogous to standard RNNs) and a cell state $c_t$that serves as a long-term memory. Three gates control cell state updates:

The *forget gate* determines which information from the previous cell state to discard:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{2.3}$$

where σ denotes the sigmoid function producing values in [0,1], and for the symbol $[\cdot,\cdot]$ represents concatenation. Values near 0 erase corresponding cell state components, while values near 1 preserve them.

The *input gate* decides which new information to add to the cell state. It operates in two stages: first, a sigmoid layer determines which values to update:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2.4}$$

Second, a tanh layer creates candidate values for addition:

$$\tilde{c}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{2.5}$$

The cell state then updates according to:

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{2.6}$$

where $\odot$ denotes element-wise multiplication. This formulation allows the network to selectively forget irrelevant past information while incorporating new task-relevant information—crucial for emotion recognition where only certain portions of an utterance carry affective signals.

Finally, the *output gate* determines which parts of the cell state to expose as the hidden state:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{2.7}$$

$$h_t = o_t \odot \tanh(c_t) \tag{2.8}$$

This gating architecture enables LSTMs to maintain information over hundreds of time steps, capturing prosodic patterns that unfold across entire utterances. For speech emotion recognition processing mel-spectrograms at a typical frame rate of ~43 Hz (with 512-sample hop length at 22.05 kHz sampling), even a 3-second utterance spans ~130 frames—well within LSTM modeling capacity.

GRUs: Simplified Temporal Modeling

GRUs offers a streamlined alternative to LSTMs with fewer parameters and comparable performance [25]. GRUs merge the forget and input gates into a single *update gate* and combine cell and hidden states, reducing computational overhead while maintaining long-range dependency modeling. The update gate controls information retention from the previous time step:

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t] + b_z) \tag{2.9}$$

A *reset gate* determines how much past information to discard when computing the candidate activation:

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t] + b_r) \tag{2.10}$$

$$\tilde{h}_t = \tanh(W_h \cdot [r_t \odot h_{t-1}, x_t] + b_h) \tag{2.11}$$

The final hidden state interpolates between previous and candidate activations:

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \tag{2.12}$$

This formulation provides similar expressive power to LSTMs with approximately 25% fewer parameters—a significant advantage when training on limited emotion datasets. In practice, bidirectional variants (Bi-LSTM, Bi-GRU) prove most effective for speech emotion recognition [26]. Bidirectional processing runs separate forward and backward RNNs, concatenating their hidden states:

$$h_t = \left[\vec{h}_t; \overleftarrow{h}_t\right] \tag{2.13}$$

where $\vec{h}_t$ encodes past context (frames 1 to t) and $\overleftarrow{h}_t$ encodes future context (frames t to T). This allows each position's representation to leverage information from the entire utterance—particularly valuable for emotion recognition where anticipatory prosody (voice quality changes preceding emotionally charged words) and carryover effects (emotional coloring persisting beyond triggering events) both contribute to affective perception.

2.2.3 Attention Mechanisms: Learning Temporal Saliency

Despite RNNs' effectiveness in capturing temporal dependencies, a fundamental challenge remains: not all portions of an utterance contribute equally to emotional expression. Emotionally charged words, prosodic peaks, and voice quality shifts often concentrate affective cues in brief intervals—perhaps 20-30% of utterance duration—while the remainder carries primarily linguistic content with minimal emotional information. Standard RNNs architectures process all time steps uniformly, potentially diluting emotionally salient signals.

Attention mechanisms address this through learned soft selection over temporal sequences [27], [28]. Rather than treating all RNNs hidden states equally, attention computes a weighted aggregation emphasizing emotionally relevant segments. The mechanism has become standard in modern speech emotion recognition systems, improving both accuracy and interpretability.

Bahdanau Attention: Additive Temporal Weighting

The Bahdanau attention mechanism, originally developed for neural machine translation [29], has been successfully adapted for speech emotion recognition.

Given a sequence of bidirectional GRU hidden states $h_1, h_2, \ldots, h_T$ where each $h_t \in \mathbb{R}^{2d}$ (concatenating forward and backward states from a GRU with d-dimensional hidden layers), attention computes a context vector c as a weighted sum:

$$c = \sum_{t=1}^{T} \alpha_t h_t \tag{2.14}$$

where attention weights α_t satisfy $\alpha_t \geq 0$ and $\sum_{t=1}^{T} \alpha_t = 1$. These weights reflect each time step's importance for the emotion classification decision. Computing appropriate weights requires learning an alignment model that scores how well position $t$matches the overall emotional content:

$$e_t = v_a^T \tanh(W_a h_t) \tag{2.15}$$

where $W_a \in \mathbb{R}^{d_a \times 2d}$ and $v_a \in \mathbb{R}^{d_a}$ are learnable parameters, and d_a is the attention dimension (commonly set equal to the hidden dimension d). The tanh nonlinearity allows the scoring function to learn complex, nonlinear relationships between hidden states and emotional salience. These raw scores are then normalized via softmax to produce valid probability distributions:

$$\alpha_t = \frac{\exp(e_t)}{\sum_{j=1}^{T} \exp(e_j)} \tag{2.16}$$

This formulation ensures attention weights sum to one and remain non-negative, enabling interpretation as a probability distribution over time steps. In practice, the attention mechanism learns to assign high weights ($\alpha_t > 0.1$) to 10-30% of frames, typically corresponding to prosodic peaks, emotionally charged words, or voice quality shifts indicative of affective state [30].

The resulting context vector $c$provides a fixed-dimensional utterance-level representation that emphasizes emotionally salient acoustic patterns while suppressing irrelevant variations from speaker identity or linguistic content. This representation then feeds into fully connected classification layers:

$$p(\text{ emotion } | \ c\ ) = \text{softmax}(W_c c + b_c) \tag{2.17}$$

where W_c and b_c are learnable parameters mapping the context vector to emotion class probabilities.

Beyond improving classification accuracy, attention weights offer valuable interpretability. Visualizing α_t across time reveals which temporal segments drive emotion predictions—for instance, attention might concentrate on

prosodic peaks for anger detection (high pitch and energy), voice breaks and pauses for sadness recognition (reduced pitch range and energy), or rapid pitch modulation for happiness detection (varied prosody and elevated baseline pitch). This interpretability proves valuable for debugging model behavior, validating that learned patterns align with human intuition about emotional expression, and building trust in deployed systems [31].

2.2.4 Self-supervised Pre-training and Transfer Learning

A more recent development leverages massive-scale self-supervised pre-training on unlabeled audio data to learn robust general-purpose speech representations [32]. Models like Wav2Vec 2.0 [33] and HuBERT [34] train on hundreds of thousands of hours of diverse speech, vastly exceeding the scale of any labeled emotion dataset—even the largest emotion corpora contain only thousands of labeled utterances.

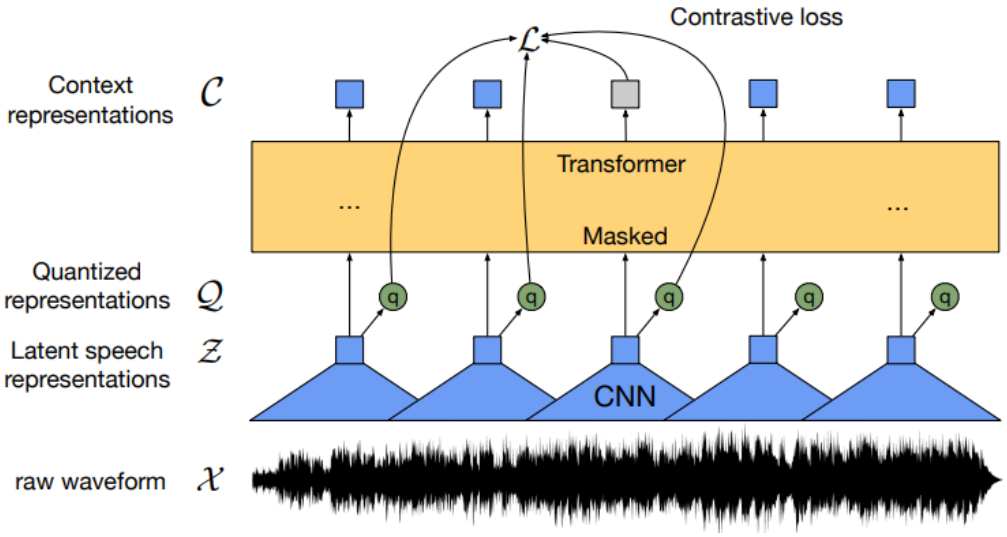

Figure 2.2: Wav2Vec 2.0

Wav2Vec 2.0: Contrastive Predictive Learning

Wav2Vec 2.0 learns speech representations through a masked prediction objective resembling BERT's masked language modeling [33], [35]. The architecture comprises two components: a convolutional feature encoder that processes raw audio waveforms, and a transformer encoder that contextualizes these features through self-attention.

The feature encoder applies seven convolutional layers with strides that

progressively downsample the raw waveform by 160x, that is approximately mapping 20ms of audio to a single time step. This multi-scale processing extracts acoustic features at a frame rate suitable for phonetic modeling (~50 Hz). These features are then randomly masked: approximately 6.5% of time steps are replaced with a learned mask embedding, and the model must predict masked representations from surrounding context.

The prediction task employs contrastive learning: for each masked position, the model must identify the true latent representation among a set of K distractors (typically $K = 100$) sampled from other time steps in the same utterance. Specifically, given context representations c_t and a set of candidates quantized features $\{q_t, q_{t,1}, \dots, q_{t,K}\}$ where q_t is the true target and $q_{t,i}$ are distractors, the model maximizes:

$$\mathcal{L}_{\text{contrastive}} = -\log \frac{\exp\left(\frac{\text{sim}(c_t, q_t)}{\tau}\right)}{\sum_{i=0}^{K} \exp\left(\frac{\text{sim}(c_t, q_{t,i})}{\tau}\right)} \tag{2.18}$$

where $\text{sim}(\cdot,\cdot)$ computes cosine similarity and τ is a temperature parameter. This objective encourages representations where true targets lie closer to context vectors than distractors in embedding space.

Crucially, the targets q_t are themselves learned through vector quantization—the continuous feature encoder outputs are discretized into a finite vocabulary of ~300 tokens learned via k-means clustering or Gumbel-softmax relaxation. This quantization forces the model to discover discrete acoustic units resembling phonemes, enabling phonetic-level learning without phonetic labels [36].

Pre-trained Wav2Vec 2.0 models, trained on 960 hours of LibriSpeech data [37], capture rich acoustic and phonetic structures that transfer to downstream tasks. For emotion recognition, practitioners typically freeze the convolutional encoder while fine-tuning the transformer encoder and adding task-specific classification heads. The pre-trained representations encode prosodic contours, voice quality characteristics, and phonetic structures—all relevant for emotion perception—though they optimize for linguistic content rather than paralinguistic cues.

HuBERT: Clustering-based Masked Prediction

HuBERT (Hidden-Unit BERT) [34] offers an alternative self-supervised approach, simplifying the training objective while achieving comparable or superior results. Rather than jointly learning representations and discrete targets via contrastive learning, HuBERT separates these stages. The model

first clusters frame-level features from a previous iteration (or from MFCC features in the first iteration) to generate pseudo-labels. A masked prediction loss then trains a BERT-like transformer to predict cluster assignments for masked frames:

$$\mathcal{L}_{\text{HuBERT}} = -\sum_{t \in \mathcal{M}} \log p(\, c_t \mid c_{\setminus t} \,) \tag{2.19}$$

where $\mathcal{M}$ denotes masked time steps, c_t is the cluster assignment, and $c_{\setminus t}$ represents context from unmasked frames. This simpler objective avoids contrastive sampling while still encouraging the model to discover discrete acoustic units. Iterative refinement—alternating between clustering learned representations and training with updated cluster labels—progressively improves representation quality. After several iterations, HuBERT representations encode phonetic, prosodic, and speaker characteristics that transfer effectively to emotion recognition.

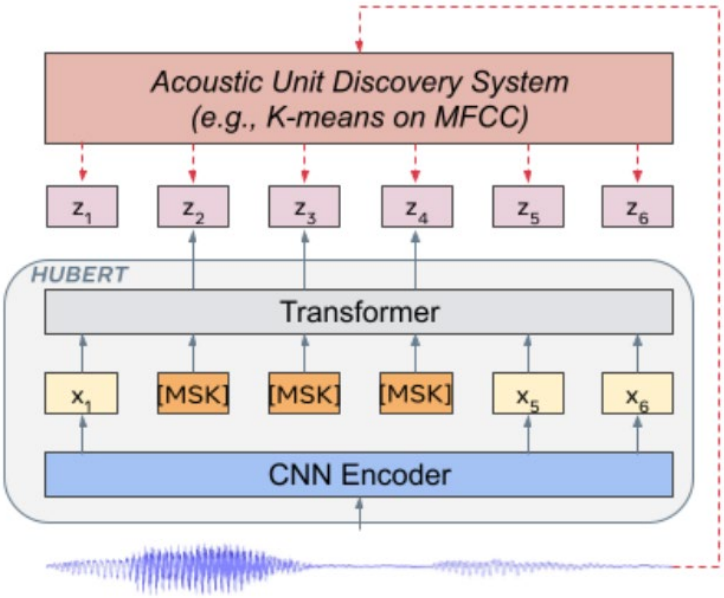

Figure 2.3: HuBERT

Challenges in Transferring ASR Models to SER

Despite their promise, directly applying automatic speech recognition (ASR) pre-trained models to speech emotion recognition presents challenges [9], [13]. ASR systems optimize for recognizing linguistic content—the "what" of speech—while actively suppressing paralinguistic cues like prosody, voice quality, and speaking style that convey emotional information—the "how" of speech. This fundamental objective mismatch means that representations ideal

for transcription may be suboptimal for emotion recognition.

Effective transfer learning strategies must balance preserving useful acoustic representations while adapting to affective characteristics. Layer-wise learning rate decay has proven beneficial: lower transformer layers capturing general acoustic features update conservatively (learning rates 10-100x smaller than standard values), while higher layers capturing task-specific patterns learn more aggressively [38]. Some practitioners freeze the feature encoder entirely and fine-tune only the contextualizing transformer, while others employ gradual unfreezing—initially training only the classification head, then progressively unfreezing deeper layers as training proceeds.

Despite these challenges, transfer learning from self-supervised models shows particular promise in low-resource scenarios where labeled emotional speech is scarce. Fine-tuning on as few as 1-2 hours of labeled emotion data can achieve performance approaching systems trained on 10× more data from scratch [39]. This efficiency proves crucial for cross-lingual emotion recognition and adapting to specialized domains (e.g., clinical speech analysis) where large-scale labeled datasets are impractical to collect.

Persistent Challenges and Future Directions

Despite these advances, speech emotion recognition confronts persistent challenges[19], [40], [41]. Emotions manifest through subtle acoustic variations easily overshadowed by speaker-specific characteristics, linguistic content, and recording conditions. A speaker's habitual voice quality, speaking rate, and pitch range introduce substantial variability unrelated to emotion—what constitutes elevated pitch for one speaker may represent baseline pitch for another. Disentangling emotion-relevant acoustic variations from speaker identity remains an open challenge.

Cultural differences in emotional expression further complicate generalization. Display rules—socially learned norms governing appropriate emotional expression—vary dramatically across cultures. What constitutes an angry tone in Western contexts may sound neutral in East Asian cultures emphasizing emotional restraint, while expressions considered neutral in individualist societies may signal mild displeasure in collectivist contexts. Most existing models train primarily on Western speakers, raising questions about cross-cultural applicability.

Additionally, most existing datasets comprise acted or elicited emotions recorded in controlled settings, limiting ecological validity. Professional actors produce exaggerated, stereotypical expressions designed for clarity rather than reflecting the subtle, context-dependent emotional displays characteristic of spontaneous affective behavior. Whether models trained on acted speech will

recognize naturalistic emotions in real-world conversational contexts remains uncertain. The field continues evolving toward systems that can operate across diverse speakers, languages, and recording conditions while respecting cultural variation in emotional expression—challenges that motivate the domain adaptation techniques discussed in Section 2.5.

2.3 Video-based Emotion Recognition

Video-based emotion recognition leverages visual cues—primarily facial expressions, but also head pose, gaze direction, and body language—to infer affective states from image sequences. Human faces constitute the most expressive channel for emotional communication: the Facial Action Coding System (FACS) identifies over 40 distinct action units that combine to produce thousands of possible expressions. The challenge lies in automatically detecting these subtle muscular movements and their temporal dynamics from video data captured under varying lighting conditions, camera angles, and occlusions [42], [43].

The field has evolved from frame-based analysis using traditional computer vision techniques to sophisticated deep learning models that jointly process spatial appearance and temporal dynamics. Modern approaches must address fundamental challenges including inter-subject variability in expression intensity, cultural differences in emotional display, and the distinction between spontaneous and posed expressions. Real-world applications demand robustness to partial occlusions (hands covering face, glasses), non-frontal poses, and low-resolution imagery captured in uncontrolled environments—scenarios where traditional geometric feature extraction fails catastrophically.

2.3.1 From Frame-based to Spatiotemporal Modeling

Early video-based emotion recognition systems treated facial expression analysis as a frame-level classification problem. These approaches extracted geometric features—facial landmark positions and their configurations—combined with appearance-based features like Local Binary Patterns (LBP) or Histogram of Oriented Gradients (HOG). The extracted features fed into classifiers such as Support Vector Machines or Hidden Markov Models to recognize discrete emotion categories [43], [44]. While computationally efficient, this paradigm suffered from critical limitations: accurate facial landmark detection required as a preprocessing step proved fragile under head pose variations, and frame-by-frame processing discarded the temporal evolution of expressions.

The deep learning revolution transformed facial expression analysis by enabling end-to-end learning directly from raw pixel intensities. Convolutional neural networks, initially applied to static images, learned hierarchical representations where lower layers detect edges and textures while higher layers recognize facial parts and expression-specific patterns. Transfer learning from large-scale face recognition datasets (VGGFace, FaceNet trained on millions of identities) provided robust feature extractors that could be fine-tuned for emotion recognition with limited labeled data—a practical necessity given the scarcity of large-scale annotated emotional video datasets [45].

However, as Kollias and Zafeiriou emphasize [46], treating video as collections of independent frames discards crucial temporal information. Emotions unfold dynamically over time, exhibiting distinct onset, apex, and offset phases. Genuine surprise, for instance, shows rapid onset (less than 0.5 seconds from neutral to peak expression) followed by quick decay, while posed surprise often exhibits slower, more deliberate timing. Capturing these temporal dynamics became essential for distinguishing authentic emotions from deliberate expressions and achieving robust recognition in naturalistic settings where single-frame ambiguity is common.

Spatiotemporal Feature Learning Architectures

Modeling spatiotemporal patterns requires architectures that jointly process spatial appearance within frames and temporal evolution across frames. Early deep learning approaches employed two-stream architectures: separate convolutional networks processed RGB frames (appearance stream) and optical flow fields (motion stream), with late fusion combining their predictions. While conceptually simple, this approach required expensive optical flow computation and failed to capture joint spatiotemporal patterns—spatial and temporal information processed independently cannot learn correlations between appearance changes and motion patterns that characterize emotional expressions.

Three-dimensional Convolutional Neural Networks (3D CNNs) offered a more integrated solution by extending 2D convolutions along the temporal dimension [47]. A 3D convolutional filter with dimensions $k_h \times k_w \times k_t$ (height, width, time) applies across spatial and temporal dimensions simultaneously:

$$y(x, y, t) = \sum_{i,j,\tau} w(i, j, \tau) \cdot x(x + i, y + j, t + \tau) + b \tag{2.20}$$

where w represents learnable filter weights and b is a bias term. This joint spatiotemporal convolution directly learns features capturing both appearance

patterns (facial configurations at individual frames) and motion patterns (how these configurations change over time)—for instance, a 3D filter might learn to detect the rapid eyebrow raise characteristic of surprise or the progressive mouth corner movement indicating a genuine smile.

Notable 3D CNN architectures include C3D (Convolutional 3D) [48], [49], which demonstrated that 3×3×3 filters applied over 16-frame clips achieve strong performance on action recognition benchmarks, and I3D (Inflated 3D), which initializes 3D filters by "inflating" weights from 2D CNNs pre-trained on ImageNet [50]—replicating 2D filters across the temporal dimension and normalizing by the temporal extent. This inflation strategy enables effective transfer learning from image recognition to video understanding, leveraging the vast labeled image data to initialize spatiotemporal models.

However, 3D CNNs introduce substantial computational overhead. A standard 2D convolutional layer with C_{in} input channels, C_{out} output channels, and $k \times k$ spatial filters require $C_{in} \times C_{out} \times k \times k$ parameters. The corresponding 3D layer with temporal extent k_t requires $C_{in} \times C_{out} \times k \times k \times k_t$ parameters—a multiplicative increase proportional to temporal kernel size. For typical configurations ($k_t = 3$ to 5), this translates to 3-5× more parameters per layer, dramatically increasing memory requirements and training data needs. Deep 3D CNNs easily exceed GPU memory constraints and overfit on limited emotion datasets [51], [52].

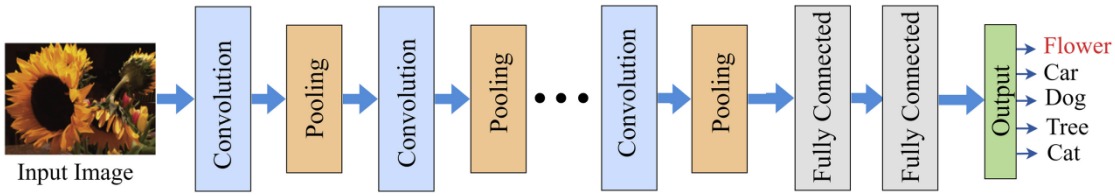

Figure 2.4: Traditional structure of a CNNs model

Factorized spatiotemporal convolutions address this efficiency challenge by decomposing 3D convolutions into separate spatial (2D) and temporal (1D) operations. The (2+1)D convolution architecture exemplifies this strategy [47]: a 3D convolution with $k \times k \times k_t$ filter is decomposed into a 2D spatial convolution ($k \times k \times 1$) followed by a 1D temporal convolution ($1 \times 1 \times k_t$). Mathematically:

$$\text{3D conv: } y = \sigma(W_{3D} * x + b) \tag{2.21}$$

is approximated by:

$$\text{(2+1)D: } z = \sigma(W_{2D} * x + b_1), y = \sigma(W_{1D} * z + b_2) \tag{2.22}$$

where σ denotes nonlinearity and $*$ represents convolution. This factorization reduces parameter count significantly: instead of $C_{in} \times M \times k \times k \times k_t$ parameters for a 3D convolution with M intermediate channels, the factorized version requires $C_{in} \times M \times k \times k + M \times C_{out} \times 1 \times k_t$ parameters. For typical values ($M = C_{out}$, $k = 3$, $k_t = 3$), this achieves approximately 30-40% parameter reduction while maintaining comparable expressive power. The factorized design enables training deeper networks on limited computational resources while still capturing essential spatiotemporal patterns.

2.3.2 Vision Transformers and Self-Attention for Video Understanding

The transformer architecture, originally proposed for natural language processing, has recently revolutionized computer vision. The Vision Transformer (ViT) demonstrated that pure attention-based models could match or exceed CNNs performance on image classification when trained on sufficient data [53]. ViT divides images into fixed-size patches (typically 16×16 pixels), linearly embeds each patch, and processes the sequence of patch embeddings using standard transformer encoder layers with multi-head self-attention mechanisms.

Self-Attention Mechanism: Learning Spatial Relationships

The core innovation enabling transformers is the self-attention mechanism, which computes interactions between all pairs of elements in a sequence [54], [55]. For an input sequence of patch embeddings $\{x_1, x_2, ..., x_N\}$ where $x_i \in \mathbb{R}^d$, self-attention transforms each element by aggregating information from all other elements through learned similarity weights.

The mechanism employs three learnable linear projections to generate queries (Q), keys (K), and values (V):

$$Q = XW_Q, K = XW_K, V = XW_V \tag{2.23}$$

where $X \in \mathbb{R}^{N \times d}$ stacks input embeddings and $W_Q, W_K, W_V \in \mathbb{R}^{d \times d_k}$ are learnable weight matrices. The query for position $i$is matched against keys from all positions to compute attention weights:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.24}$$

The softmax operation over QK^T produces a matrix of attention weights α_{ij} indicating how much position $i$should attend to position j. The scaling factor $\sqrt{d_k}$ prevents dot products from growing too large in magnitude, which would push softmax into regions with vanishingly small gradients. Each output y_i is then computed as a weighted sum of all value vectors:

$$y_i = \sum_{j=1}^{N} \alpha_{ij} v_j \tag{2.25}$$

This formulation enables each patch to selectively attend to relevant patches throughout the image, learning spatial relationships without convolutional inductive biases. For facial expression recognition, self-attention learns to correlate mouth regions with eye regions—for instance, detecting that a smiling mouth should co-occur with activated eye muscles (Duchenne smile) versus a mouth-only smile indicating posed expression [53].

Multi-head Attention: Capturing Multiple Relationships

Single attention heads learn a particular type of relationship (e.g., spatial proximity, semantic similarity). Multi-head attention extends this by learning multiple parallel attention functions, each potentially capturing different relationship types:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W_O \tag{2.26}$$

where $\text{head}_i = \text{Attention}(QW_Q^i, KW_K^i, VW_V^i)$, and each head uses separate projection matrices W_Q^i, W_K^i, W_V^i with reduced dimensionality $d_k = d/h$ (where h is the number of heads), maintaining constant computational cost. Different heads learn complementary attention patterns—some may focus on local texture (wrinkles, skin tension), others on geometric configurations (relative positions of facial landmarks), and yet others on symmetric patterns (bilateral activation of facial muscles). The concatenated outputs are projected through W_O to produce the final representation.

For image classification, ViT typically uses 8-12 attention heads per layer and stacks 12-24 transformer encoder layers, creating deep architectures that progressively refine representations. Each encoder layer consists of multi-head self-attention followed by a position-wise feedforward network, with layer

normalization and residual connections ensuring stable gradient flow during training.

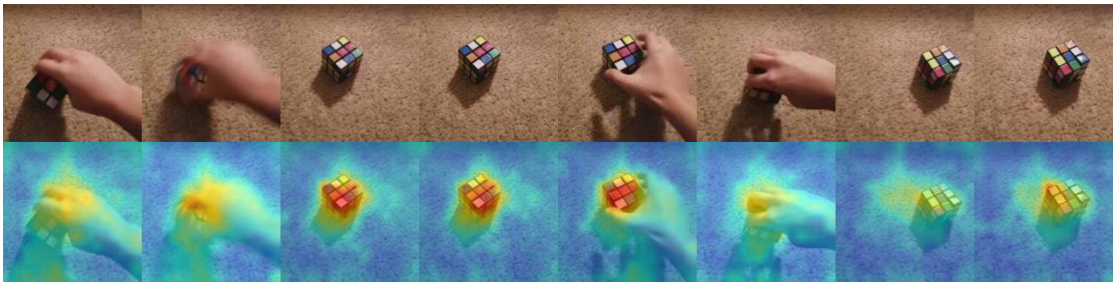

Figure 2.5: A TimeSformer Instance

Note: This is the example used in the TimeSformer paper to demonstrate that the model can learn to attend to the relevant regions in the video in order to perform complex spatiotemporal reasoning. we can see that the model focuses on the configuration of the hand when visible and the object-only when not visible.

2.3.3 TimeSformer: Divided Attention for Video

Extending transformers to video understanding presents unique challenges due to the significantly longer sequences resulting from spatiotemporal patch extraction. An image with $H \times W$ pixels divided into $P \times P$ patches yields $N = (H/P) \times (W/P)$ patches. For video with Tframes, naive spatiotemporal patching produces $T \times N$ patches—for typical configurations (224×224 images, 16×16 patches, 8 frames), this yields 8×196 = 1,568 patches. Self-attention's $O(L^2)$ computational complexity (where L is sequence length) makes joint spatiotemporal attention prohibitively expensive: computing attention over 1,568 patches requires ~2.5 million pairwise comparisons, compared to 38,000 for a single frame.

The TimeSformer architecture [56] addresses this challenge through a divided attention mechanism that decomposes spatiotemporal attention into separate spatial and temporal attention operations. This factorization dramatically reduces computational complexity while maintaining the capacity to model spatiotemporal dependencies.

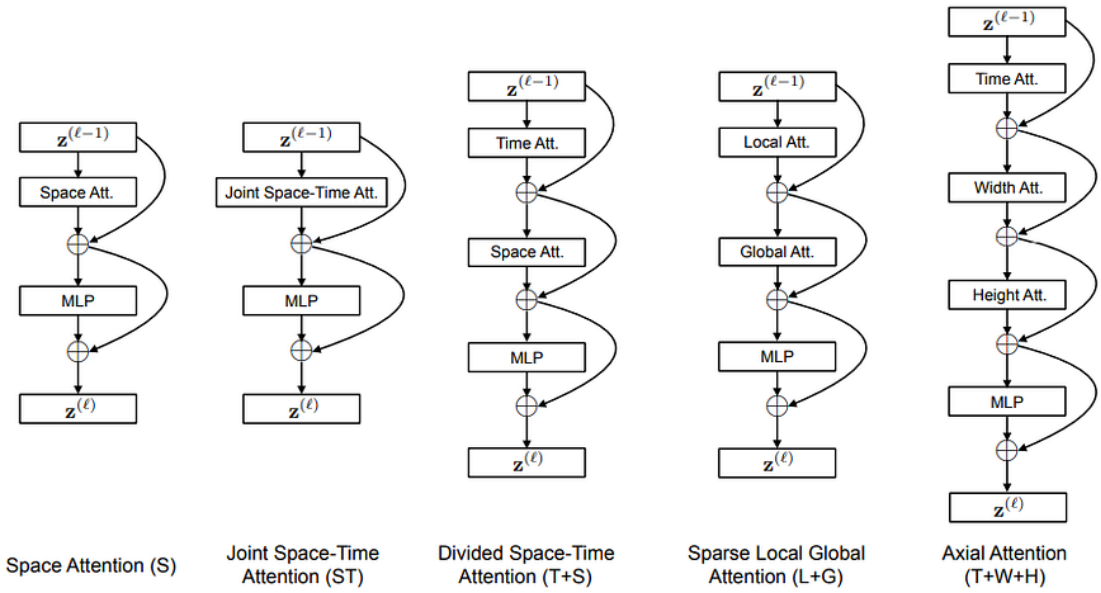

Figure 2.6: Video self-attention blocks of TimeSformer

Divided Attention: Spatial and Temporal Factorization

TimeSformer processes video through five sequential attention blocks, as showing in Figure 2.6, each implementing divided attention:

1. **Temporal Attention:** For each spatial position (p, q), compute attention across all frames at that fixed spatial location. Given patch embeddings $z^{(t,p,q)}$ for frame t and spatial position (p, q), temporal attention updates:

$$z_{temp}^{(t,p,q)} = z^{t,p,q} + \text{Attention}\left(Q^{(t,p,q)}, \left\{K^{(t',p,q)}\right\}_{t'=1}^{T}, \left\{V^{(t',p,q)}\right\}_{t'=1}^{T}\right) \tag{2.27}$$

This operation captures motion and temporal evolution at each spatial location—for instance, how a specific facial region (corner of mouth, eyebrow position) changes across frames.

2. **Spatial Attention:** For each frame t, compute attention across all spatial positions within that frame. This models spatial relationships at fixed temporal locations:

$$z_{\text{spat}}^{(t,p,q)} = z_{\text{temp}}^{(t,p,q)} + \text{Attention}\left(Q^{(t,p,q)}, \{K^{(t,p',q')}\}_{p',q'}, \{V^{(t,p',q')}\}_{p',q'}\right) \tag{2.28}$$

Spatial attention learns which facial regions should be correlated for

emotion recognition—eyes with mouth (for genuine smiles), eyebrows with eyes (for surprise), forehead with mouth (for anger).

The divided attention factorization reduces computational complexity from $O((T \times N)^2)$ for joint spatiotemporal attention to $O(T \times N \times T) + O(T \times N \times N) = O(T \times N \times (T + N))$. For $T = 8$ and $N = 196$, joint attention requires ~6.1M operations while divided attention requires only ~320K operations—approximately 19× reduction. This efficiency gain enables processing longer video sequences and training deeper models within memory constraints [56].

Positional Encoding and Learned Embeddings

Unlike CNNs with built-in translation equivariance through weight sharing, transformers process patch embeddings as sets without inherent spatial or temporal ordering. Positional encodings inject spatial and temporal location information. TimeSformer uses learned positional embeddings: each patch position (t, p, q) receives an additive embedding:

$$z_{\text{input}}^{(t,p,q)} = \text{PatchEmbed}(x^{(t,p,q)}) + E_{\text{pos}}^{(t)} + E_{\text{pos}}^{(p,q)} \tag{2.29}$$

where $E_{\text{pos}}^{(t)}$ encodes temporal position and $E_{\text{pos}}^{(p,q)}$ encodes spatial position. These embeddings, learned during training, enable the model to leverage spatiotemporal structure—for instance, learning that central facial patches carry more emotional information than peripheral background patches, or that expression onset (early frames) should attend differently than apex (middle frames) or offset (late frames) [56].

Pre-training and Transfer Learning

TimeSformer benefits significantly from pre-training on large-scale action recognition datasets like Kinetics-400 (containing 400 action classes across 240,000 training videos). Pre-training provides several advantages for emotion recognition: First, the model learns general spatiotemporal patterns—how objects and faces move, how appearance changes over time—that transfer across tasks. Second, pre-trained attention mechanisms learn to focus on relevant spatial regions (faces, hands, bodies) and temporal segments (action onsets, peaks). Third, initialization from pre-trained weights enables training on smaller emotion datasets (thousands rather than millions of samples) while avoiding overfitting.

For emotion recognition applications, practitioners typically fine-tune pre-

trained TimeSformer models on emotion-specific datasets [57], [58]. The pre-trained weights provide strong initialization, while task-specific fine-tuning adapts representations to affective characteristics—learning to attend to subtle facial muscle movements (micro-expressions) and emotion-specific temporal patterns (the sustained nature of sadness versus the brief peak of surprise). Fine-tuning often employs lower learning rates for early transformer layers (capturing general visual features) and higher rates for later layers (capturing emotion-specific patterns), a strategy analogous to the layer-wise learning rate decay discussed for speech models in Section 2.2.4.

Interpretability through Attention Visualization

Beyond performance gains, TimeSformer's attention mechanisms provide interpretability. Visualizing attention weights reveals which spatial regions and temporal segments drive emotion predictions. For a surprise prediction, temporal attention weights might peak at frames showing the expression onset—the moment eyebrows raise and eyes widen—while spatial attention concentrates on eye and eyebrow regions. For sadness, attention may distribute more uniformly across frames, reflecting the sustained nature of sad expressions, while spatially focusing on mouth regions (downturned corners) and eyes (reduced aperture, lack of crinkling).

This interpretability proves valuable for validating that models learn meaningful patterns aligned with psychological theories of emotional expression rather than spurious correlations in training data (e.g., background objects, lighting conditions). Attention visualizations also facilitate error analysis—examining cases where predictions fail can reveal systematic biases, such as over-reliance on particular facial regions or temporal segments that fail to generalize across individuals or contexts [59].

2.3.4 Face Detection and Preprocessing Pipeline

Robust face detection and preprocessing constitute critical prerequisites for video-based emotion recognition, as facial alignment and normalization significantly impact downstream model performance. Modern systems employ deep learning-based face detectors that jointly localize faces and facial landmarks in a single forward pass.

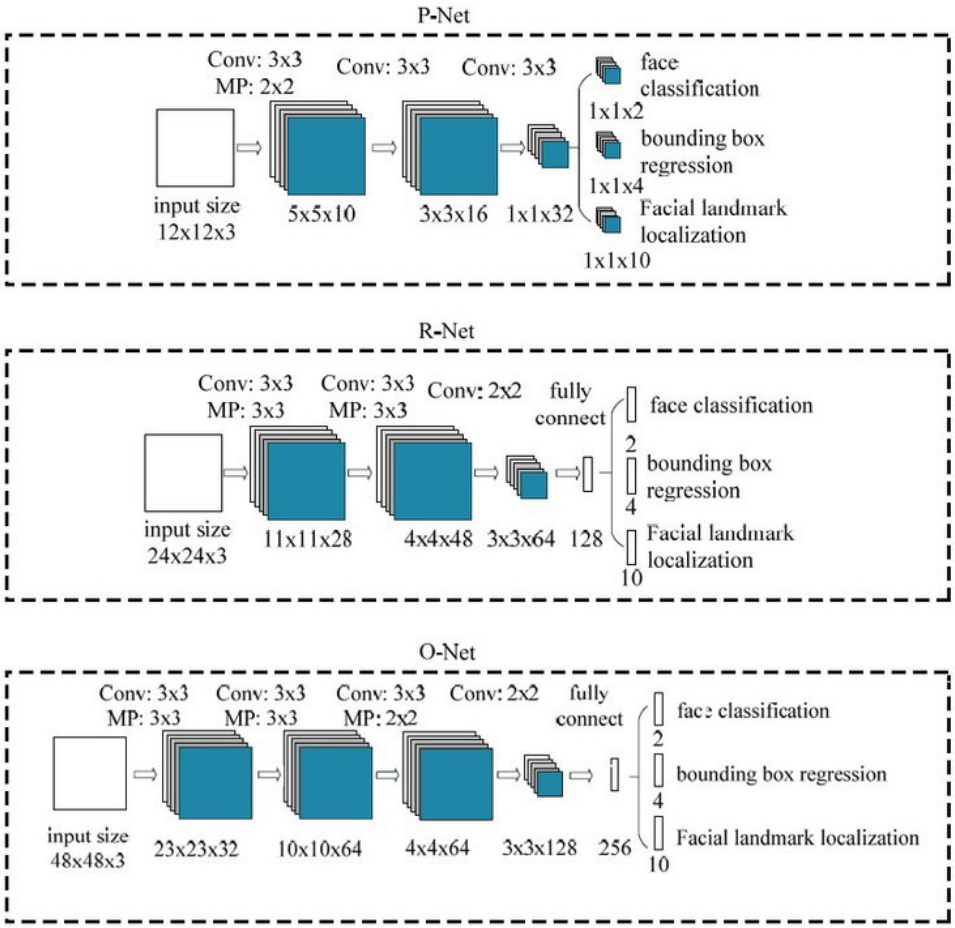

Figure 2.7: MTCNN architecture

Note: The architecture consists of three networks (P-Net, R-Net, and O-Net) that progressively refine face detection and alignment.

Multi-task Cascaded Convolutional Networks (MTCNN) [60] exemplifies contemporary face detection approaches. MTCNN employs a three-stage cascade: a Proposal Network (P-Net) rapidly scans images at multiple scales to generate candidate face regions, a Refinement Network (R-Net) filters candidates and refines bounding boxes through regression, and an Output Network (O-Net) produces final detections along with five facial landmarks (eyes, nose, mouth corners). This coarse-to-fine strategy achieves high accuracy even for small, blurred, or partially occluded faces while maintaining computational efficiency through early rejection of obvious non-face regions.

The preprocessed face sequences serve as input to emotion recognition models. For TimeSformer specifically, faces are cropped with appropriate margins (typically 20-30% padding beyond detected bounding boxes to include contextual regions), resized to 224×224 pixels, and normalized to [0,1] range.

Frame sampling strategies also play important roles: uniform sampling at fixed intervals (e.g., selecting 8 frames from a 3-second clip) provides temporal coverage, while adaptive sampling based on optical flow or expression intensity can emphasize emotionally salient moments. For the domain adaptation experiments discussed later, maintaining consistent preprocessing across RAVDESS and CREMA-D datasets proves crucial—differences in face detection, cropping, or resolution can introduce artificial domain shifts that confound the evaluation of adaptation effectiveness.

The quality of face detection directly propagates through the entire pipeline. Detection failures (missed faces or false positives) prevent emotion recognition entirely, while imprecise localization (slightly off-center crops, inconsistent scaling) degrades recognition accuracy by 5-10% [61]. For video sequences, temporal consistency in face tracking—smoothing bounding boxes across frames rather than independently detecting per frame—reduces jittering and provides stable input to temporal models. These preprocessing considerations, while often overlooked in methodological descriptions, critically impact real-world system performance.

2.4 Multimodal Fusion Strategies for Emotion Recognition

Multimodal fusion aims to integrate complementary information from heterogeneous modalities to achieve more robust and accurate emotion recognition than unimodal approaches. As Rahman et al. [62] comprehensively document in their recent survey, audio and video signals capture distinct aspects of emotional expression—speech conveys prosodic and vocal characteristics while facial expressions reveal visual affective cues—making their combination particularly effective for emotion analysis. The central challenge lies in determining how and when to combine these modalities to maximize their complementary strengths while mitigating individual weaknesses.

Fusion strategies can be broadly categorized into three paradigms based on the processing stage at which integration occurs: early fusion (feature-level), late fusion (decision-level), and middle fusion (intermediate-level). Each approach presents distinct trade-offs between modeling capacity, computational efficiency, and robustness to modality-specific noise or missing data. Recent advances have leveraged attention mechanisms and transformer architectures to enable more sophisticated fusion strategies that dynamically weight modality contributions based on context [63].

2.4.1 Early Fusion: Feature-level Integration

Early fusion, also termed feature-level fusion, concatenates representations from different modalities at an early processing stage before feeding the combined feature vector into a unified classifier. In the context of audio-visual emotion recognition, this typically involves extracting acoustic features (e.g., mel-spectrograms processed through CNNs or RNNs) and visual features (e.g., facial appearance encoded by vision transformers), concatenating these embeddings, and training a joint classifier on the combined representation.

The primary advantage of early fusion lies in its computational efficiency and simplicity. By processing modalities jointly from an early stage, the model can potentially learn cross-modal correlations and interactions that might be missed by separate processing pipelines. A basic implementation concatenates audio and video embeddings and passes them through fully connected layers for classification. This approach enables end-to-end training with direct backpropagation of gradients through both modality encoders.

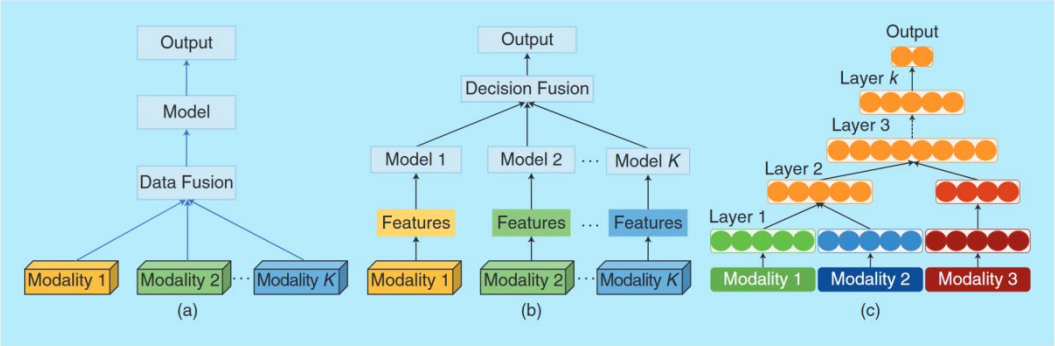

Figure 2.8: An illustration of various fusion models for multimodal learning. (a) Early or data-level fusion, (b) late or decision-level fusion, and (c) intermediate fusion [64]

However, Rahman et al. note several limitations [62]. Early fusion assumes tight temporal synchronization between modalities, which may not hold in real-world scenarios where audio and video signals experience different processing delays or quality degradations. Additionally, simple concatenation treats all features equally, failing to account for varying reliability or relevance of different modalities across contexts. When one modality contains predominantly noise or irrelevant information, it can negatively impact the combined representation.

Advanced early fusion strategies address these limitations through learned integration mechanisms. Rather than simple concatenation followed by linear

transformation, more sophisticated architectures employ gated fusion units or attention-weighted combination of modality features before classification, providing some flexibility while maintaining the early fusion paradigm.

2.4.2 Late Fusion: Decision-level Combination

Late fusion, or decision-level fusion, takes the opposite approach by training separate classifiers for each modality and combining their predictions to produce a final decision. In audio-visual emotion recognition, this involves training independent audio and video emotion classifiers, obtaining probability distributions over emotion classes from each modality, and aggregating these predictions through averaging, voting, or learned combination weights. The key advantage of late fusion lies in its modularity and robustness. Since modalities are processed independently, the approach naturally handles asynchronous inputs and modality-specific characteristics without requiring aligned feature spaces. If one modality is corrupted or missing, the system can fall back to predictions from available modalities without catastrophic failure. This robustness makes late fusion particularly attractive for real-world applications where sensor reliability varies.

Traditional late fusion employs simple averaging of class probabilities or majority voting among modality-specific predictions. While straightforward, these methods treat all modalities as equally reliable regardless of input characteristics. Recent approaches have introduced learned fusion weights that dynamically adjust modality contributions based on their confidence or relevance. Wagner et al. [65] demonstrate that training a small meta-classifier to combine modality-specific predictions based on their entropy or consistency can significantly improve performance over fixed weighting schemes.

A more sophisticated variant employs neural networks to learn optimal combination strategies. By concatenating probability distributions or logits from individual modality classifiers and training a fusion network to predict final emotions, the system learns context-dependent weighting that adapts to varying modality reliabilities. This approach maintains late fusion's modularity while enabling more nuanced integration than simple averaging.

Despite these advantages, late fusion's separate processing of modalities prevents learning of cross-modal interactions during feature extraction [62]. Subtle correlations between audio and visual cues—such as synchronization between lip movements and speech prosody—cannot be captured when modalities are encoded independently. This limitation has motivated the development of middle fusion strategies.

2.4.3 Middle Fusion: Attention-based Integration

Recent work has explored multi-stage fusion architectures that enable iterative refinement of multimodal representations. Rather than performing fusion in a single step, these approaches employ multiple layers of cross-modal interaction, allowing increasingly abstract representations to be refined through successive information exchange. Each fusion layer computes bidirectional attention between modalities, applies residual connections to preserve modality-specific information, and employs layer normalization for training stability. This progressive refinement enables the model to capture both low-level cross-modal correspondences (e.g., lip-speech synchronization) and high-level semantic alignments (e.g., crying sounds accompanying sad facial expressions). Mittal et al. [66] demonstrate that employing multiplicative interactions across multiple stages significantly improves multimodal emotion recognition performance compared to single-stage fusion. Their work highlights the importance of hierarchical integration where early layers capture basic alignments while deeper layers learn complex semantic relationships. Similarly, progressive attention architectures can be designed with multiple cross-attention layers, where each layer refines representations from the previous stage. By stacking attention mechanisms with residual connections, the architecture enables complex multi-hop reasoning across modalities—for instance, attending to facial muscle movements that correlate with vocal strain, which in turn indicates emotional intensity. The progressive fusion strategy offers several advantages over single-stage approaches. The iterative refinement process allows the model to gradually integrate information, starting with obvious alignments and progressing to subtle cross-modal patterns. Layer normalization between attention stages ensures stable gradient flow during training, enabling deeper fusion networks. As documented in recent multimodal benchmarks [64], hierarchical fusion architectures consistently outperform both early and late fusion baselines across diverse emotion recognition tasks, with particularly strong improvements on subtle or complex emotional states that require nuanced cross-modal reasoning.

2.4.4 Comparative Analysis and Design Considerations

The choice among fusion strategies involves fundamental trade-offs between modeling capacity, computational efficiency, and practical robustness. Early fusion excels in scenarios with clean, well-synchronized multimodal data where computational resources are limited, as it requires training only a single joint classifier. However, its rigid integration makes it vulnerable to modality-specific corruptions and synchronization errors.

Late fusion provides maximum modularity and fault tolerance, making it suitable for applications where modality availability varies or where separately pre-trained unimodal models must be integrated. The inability to learn cross-modal interactions during feature extraction, however, limits its capacity to capture subtle multimodal patterns that require joint reasoning.

Middle fusion strategies, particularly attention-based approaches, offer the best of both paradigms by enabling cross-modal learning while maintaining some modality-specific processing. The computational overhead of attention mechanisms is offset by superior performance on complex emotion recognition tasks where multimodal cues interact in sophisticated ways. Multi-stage fusion architectures further enhance this capability by enabling hierarchical cross-modal reasoning [63].

Empirical comparisons consistently demonstrate that attention-based middle fusion achieves the highest accuracy on challenging emotion recognition benchmarks, particularly for subtle or complex emotional states [19]. The interpretability provided by attention weights also offers valuable insights into which cross-modal patterns drive predictions, facilitating model debugging and trust in deployment scenarios. These advantages make progressive attention-based fusion the preferred approach for state-of-the-art multimodal emotion recognition systems.

2.5 Domain Adaptation in Affective Computing

A critical challenge in deploying emotion recognition systems in real-world applications is the domain shift problem—models trained on one dataset often exhibit severe performance degradation when applied to data from different sources, recording conditions, or populations. In affective computing, this challenge is particularly acute due to substantial variability across emotional expression datasets in terms of actor demographics, recording environments, annotation protocols, and the fundamental distinction between acted versus spontaneous emotional displays [3]. A model achieving 85% accuracy on laboratory-collected acted emotions may drop to 50-60% accuracy on naturalistic emotional expressions captured in the wild, rendering the system impractical for deployment [67], [68].

Domain adaptation techniques aim to bridge this gap by learning representations that transfer effectively across domains, enabling models trained on labeled source data to generalize to unlabeled or sparsely labeled target data [67]. For multimodal emotion recognition, this challenge is compounded by the need to align not only individual modality features but also cross-modal interactions across domains. As Baltrušaitis et al. [69] comprehensively document, multimodal learning introduces unique challenges

for domain transfer beyond those encountered in unimodal settings. The fundamental question becomes: how can we leverage abundant labeled data from controlled laboratory datasets (e.g., RAVDESS [70]) to build models that perform reliably on diverse real-world datasets (e.g., CREMA-D [71]) where emotional expressions vary significantly in authenticity, intensity, and cultural context.

2.5.1 The Cross-corpus Emotion Recognition Challenge

Cross-corpus generalization represents one of the most persistent challenges in speech and video emotion recognition. Even when datasets share the same emotion label space, systematic differences in data collection methodology introduce distribution shifts that severely impact model performance. As documented in recent surveys of speech emotion recognition [68], models achieving over 80% accuracy on within-corpus test sets frequently fall below 45-50% accuracy when evaluated on held-out corpora, even when both datasets ostensibly capture the same emotion categories.

The sources of domain shift in emotion recognition are multifaceted. At the acoustic level, recording equipment quality, background noise characteristics, and room acoustics introduce systematic biases—RAVDESS recordings use professional studio equipment with minimal background noise, while CREMA-D employs consumer-grade cameras with variable acoustic conditions. At the expression level, acted emotions (RAVDESS) exhibit exaggerated prosodic patterns and stereotypical facial configurations designed for clarity, whereas more naturalistic expressions (CREMA-D) display subtle, context-dependent emotional cues that may deviate from prototypical patterns.

Beyond these technical differences, fundamental mismatches in emotion taxonomies complicate cross-corpus transfer. RAVDESS employs an 8-category emotion model (neutral, calm, happy, sad, angry, fearful, disgust, surprised) reflecting Ekman's basic emotions framework, while CREMA-D uses a 6-category model (neutral, happy, sad, angry, fear, disgust) that omits "calm" and "surprised" [70], [71]. This label space mismatch necessitates careful mapping strategies when transferring models between datasets—simply collapsing RAVDESS's 8 classes into CREMA-D's 6 classes requires decisions about whether to merge "calm" with "neutral" or discard calm samples entirely, and whether to map "surprised" to "happy" based on valence similarity or treat it as an outlier class.

Demographic and cultural factors introduce additional complications. Actor diversity varies significantly across datasets: RAVDESS features North American actors with balanced gender representation, while CREMA-D

includes more diverse ethnic backgrounds and age ranges. Cultural display rules governing emotional expression—such as the suppression of negative emotions in collectivist cultures or exaggeration of surprise in Western contexts—mean that even identical emotion labels may correspond to different behavioral patterns across populations. A model learning to recognize "anger" from RAVDESS's theatrical expressions may fail to identify more subdued angry expressions common in CREMA-D's diverse actor pool.

2.5.2 Domain Adversarial Neural Networks

Domain adversarial training has emerged as a powerful approach for learning domain-invariant representations in emotion recognition. The foundational work by Ganin et al. [72] introduced the Domain-Adversarial Neural Network (DANN) architecture, which explicitly encourages feature extractors to produce representations that are discriminative for the primary task (emotion classification) while being indistinguishable across source and target domains. This is achieved through an adversarial training objective where a domain classifier attempts to identify which domain a sample originates from, while the feature extractor simultaneously tries to fool the domain classifier by producing domain-invariant features.

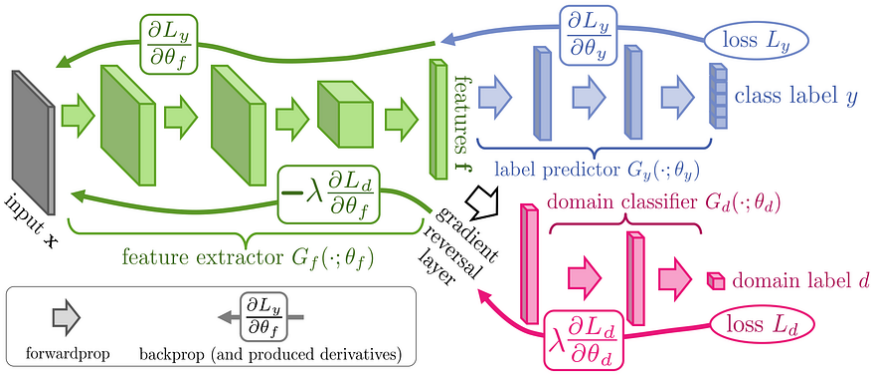

Figure 2.9: DANN

Note: It is a feature-based adaptation method that learns a domain-invariant representation through adversarial training.

The key innovation enabling this adversarial objective is the Gradient Reversal Layer (GRL), which implements a simple yet elegant mechanism during backpropagation. During forward propagation, the GRL acts as an identity function, passing features unchanged to the domain classifier. During

backward propagation, however, the GRL multiplies gradients by a negative constant before passing them to the feature extractor. This gradient reversal has a profound effect: gradients that would normally update the feature extractor to improve domain classification are instead reversed, encouraging the feature extractor to produce features that maximize domain classifier confusion—precisely the desired domain-invariant property.

Formally, the domain adversarial objective can be expressed as a minimax game. Let f_θ denote the feature extractor with parameters θ, g_y the emotion classifier, and g_d the domain classifier. The training objective becomes:

$$\min_{\theta, g_y} \max_{g_d} \mathcal{L}_y(f_\theta(x), y) - \lambda \mathcal{L}_d(g_d\big(f_\theta(x)\big), d) \tag{2.30}$$

where $\mathcal{L}_y$ is the emotion classification loss, $\mathcal{L}_d$ is the domain classification loss, and λcontrols the trade-off between task performance and domain invariance. The feature extractor f_θ minimizes emotion loss while maximizing domain loss (via gradient reversal), whereas the domain classifier g_d maximizes its own accuracy, creating an adversarial dynamic that drives the learning of domain-invariant yet emotion-discriminative representations [72].

The hyperparameter λ, often referred to as the domain adaptation strength or adversarial coefficient, critically influences training dynamics. Setting λ too high causes the model to prioritize domain confusion over emotion discrimination, resulting in representations that are domain-invariant but insufficiently expressive for accurate emotion classification—source domain accuracy may remain acceptable, but the representations lack the discriminative power needed for fine-grained emotion distinctions [72]. Conversely, setting λ too low provides insufficient pressure toward domain invariance, allowing the model to overfit to source domain-specific patterns that fail to transfer.

Recent work has explored adaptive scheduling of λ throughout training rather than using a fixed value. A common approach gradually increases λ from near-zero to a maximum value as training progresses, following schedules such as:

$$\lambda_p = \frac{2}{1 + \exp(-\gamma \cdot p)} - 1 \tag{2.31}$$

where $p$represents training progress (epoch / total_epochs) and γ controls the rate of increase [72]. This progressive schedule allows the model to first learn task-relevant features from the source domain before gradually introducing domain invariance pressure, preventing the adversarial objective from disrupting the initial learning of emotion-discriminative patterns.

However, in multimodal emotion recognition contexts, standard DANN faces unique challenges [73]. The architecture must learn not only domain-invariant unimodal features but also cross-modal interactions that generalize across domains [69]. A model trained on RAVDESS's synchronized, high-quality audio-visual data may learn cross-modal attention patterns that break down when applied to CREMA-D's more variable synchronization and quality characteristics. Simply applying domain adversarial training to the final fused representation may fail to address domain shift at the individual modality level or in cross-modal alignment mechanisms.

2.5.3 Multimodal Domain Adaptation with Progressive Fusion

Extending domain adaptation to multimodal fusion architectures requires careful consideration of where and how to apply adversarial training. A naive approach would apply a single domain classifier to the final fused representation, but this fails to ensure that individual modality representations are domain-invariant before fusion. If audio features remain domain-specific while video features achieve domain invariance, the fusion mechanism itself must compensate for this asymmetry—a difficult learning problem that may result in suboptimal cross-modal integration.

More sophisticated multimodal domain adaptation architectures employ multiple domain classifiers operating at different levels of the fusion hierarchy [74]. For a progressive fusion architecture with multiple cross-attention stages, domain classifiers can be attached to (1) individual modality representations before fusion, (2) intermediate representations after each fusion stage, and (3) the final fused representation. Each domain classifier receives gradient reversal signals, encouraging domain invariance at its respective level. This hierarchical domain confusion ensures that domain adaptation occurs not only in the final task-relevant representation but throughout the entire multimodal processing pipeline.

The integration of progressive cross-modal fusion with domain adversarial training presents both opportunities and challenges. On one hand, the multi-stage refinement process provides natural insertion points for domain classifiers at different abstraction levels, enabling fine-grained control over where domain invariance is enforced. The residual connections and layer normalization inherent in progressive fusion architectures also facilitate stable gradient flow even with multiple adversarial objectives. On the other hand, the increased model complexity introduces additional hyperparameters—each domain classifier may require its own λvalue, and the relative weighting among domain classifiers at different levels must be carefully tuned.

A critical practical consideration in multimodal domain adaptation for emotion recognition is computational efficiency. Training domain adversarial networks typically requires iterating over both source and target domain data, and the need to balance source and target batch sizes to prevent domain classifier bias can significantly increase training time [75]. When the target domain dataset (e.g., CREMA-D with 7,442 samples) is substantially larger than necessary for effective adaptation, using only a strategically selected subset of target data can reduce computational costs by 3-5x while maintaining adaptation effectiveness. Techniques such as uncertainty-based sampling or diversity-based selection can identify the most informative target samples for domain alignment, avoiding redundant processing of highly similar target examples.

The label space mismatch between RAVDESS (8 classes) and CREMA-D (6 classes) necessitates additional architectural considerations. One approach maintains separate task classifiers for source and target domains with different output dimensions, sharing only the feature extractor and domain classifier. During source domain training, the 8-class classifier provides supervision, while during target domain fine-tuning (if target labels are available), the 6-class classifier provides supervision. The domain classifier operates on shared features, encouraging alignment despite the different task structures. Alternatively, a unified taxonomy can be created by mapping RAVDESS labels to CREMA-D's 6-class space (e.g., merging "calm" into "neutral" and "surprised" into "happy"), simplifying the architecture at the cost of discarding potentially useful source domain distinctions.

Conservative scheduling of the adversarial coefficient λ becomes especially important in cross-corpus emotion adaptation where source and target domains exhibit substantial distributional differences. Aggressive domain confusion (high λ) early in training may prevent the model from learning sufficient source domain patterns before attempting transfer. A conservative schedule maintains very low λ values (e.g., 0.0001) for an extended warmup period spanning 40-50% of total training, allowing the model to achieve strong source domain performance, then gradually increases λ at a measured pace to a moderate maximum (e.g., 0.02-0.1) rather than the aggressive values (0.5-1.0) sometimes used in other domain adaptation contexts [75]. This conservative approach reflects the reality that emotion recognition requires learning subtle, high-dimensional patterns where excessive domain confusion can discard task-relevant information along with domain-specific artifacts.

Empirical results from cross-corpus emotion recognition studies demonstrate that well-designed multimodal domain adaptation can substantially reduce the performance gap between source and target domains [68], [74]. While standard models may exhibit a 30-40% accuracy drop from source to target, domain-adaptive architectures can reduce this gap to 10-15%

through effective invariant feature learning. However, complete elimination of the domain gap remains elusive—subtle differences in emotional expression authenticity, cultural norms, and recording conditions introduce irreducible distribution shifts that cannot be fully compensated through representation learning alone. Continued research into more sophisticated alignment mechanisms, particularly those that respect the hierarchical and multimodal nature of emotion expression, remains an active area of investigation.

Chapter 3

Methodology

This chapter presents the comprehensive methodological framework employed to address cross-dataset emotion recognition through multimodal fusion and domain adaptation. The methodology encompasses systematic data preprocessing, architecture design, training procedures, and evaluation protocols developed to enable effective transfer learning from professionally recorded studio data (RAVDESS) to diverse real-world recordings (CREMA-D).

Section 3.1 details the data acquisition and preprocessing pipeline for both audio and video modalities. For audio, we describe mel-spectrogram extraction with voice activity detection, temporal normalization, and instance-level feature standardization to ensure robust representations across different recording conditions. For video, we present face detection, temporal sampling, resolution adaptation, and storage optimization strategies that balance recognition performance with computational feasibility.

Section 3.2 examines the model architecture design, consisting of three principal components: EmoCatcher for audio encoding, TimeSformer for video encoding, and progressive fusion for multimodal integration. EmoCatcher employs convolutional feature extraction with bidirectional recurrent temporal modeling, capturing emotion-relevant prosodic patterns from mel-spectrograms. TimeSformer extends Vision Transformers to video understanding through divided space-time attention, adapted for facial expression analysis via Low-Rank Adaptation (LoRA). The progressive fusion architecture enables hierarchical cross-modal integration through multiple attention stages. For domain adaptation, we extend this base architecture with Domain-Adversarial Neural Networks (DANN), incorporating gradient reversal layers and domain discriminators to learn features that maintain

discriminative power for emotion recognition while achieving invariance to dataset-specific characteristics.

	RAVDESS	CREMA-D
Samples (Total)	1440	7442
Actors	24	91
Gender	12M+12F	48M+43F
Age Range	21-33	20-74
Ethnicity	Homogeneous	Diverse
Emotion Classes	8	6
Sentences	2	12
Recording Type	Studio	Consumer
Expression Type	Acted	Semi-natural
Audio Quality	48 kHz	Variable
Video Quality	720p/30fps	Variable

Table 3.1: Dataset Characteristics and Statistics

3.1 Data Acquisition and Processing

3.1.1 Datasets

RAVDESS The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [70] was collected at Ryerson University between 2012 and 2015. Professional actors were recruited through auditions requiring demonstrated vocal control and emotional expressiveness. The final dataset includes recordings from 24 actors (12 female, 12 male) aged 21 to 33 years, all native English speakers with neutral North American accents. Actors performed two semantically neutral statements—"Kids are talking by the door" and "Dogs are sitting by the door"—with specific instructions to convey eight distinct emotions: neutral, calm, happy, sad, angry, fearful, disgust, and surprised. Each emotion was expressed at two intensity levels (normal and strong), resulting in multiple takes per actor-emotion-sentence combination. The lexically-matched statements control for linguistic content variation, isolating emotional prosody and facial expression as the primary sources of affective information. Recording sessions occurred in a professional studio environment with acoustic treatment (anechoic foam padding, controlled

reverberation time < 0.3 seconds). Audio was captured using a Neumann TLM 102 condenser microphone at 48 kHz sampling rate with 24-bit depth, positioned 30 cm from the speaker. Video was recorded simultaneously using a Canon Vixia HF G20 camera at 720p resolution (1280×720 pixels) and 29.97 fps. Consistent lighting conditions (5600K color temperature, three-point lighting setup) minimized shadows and maintained uniform facial illumination across all recordings.

For this study, we utilize only the speech portion of RAVDESS, excluding song recordings which exhibit fundamentally different acoustic characteristics (sustained vowels, wider pitch range, formalized phrasing). After filtering, this yields 1,440 samples distributed across eight emotion categories. Table 3.1 presents the complete breakdown by emotion, gender, and intensity level. The professional production quality of RAVDESS offers several advantages for source domain training. First, high signal-to-noise ratio (>40 dB) ensures emotion-relevant features dominate over recording artifacts. Second, professional actors produce clear, prototypical expressions closely aligned with Ekman's basic emotion theory—anger features raised voice and furrowed brows, sadness exhibits lowered pitch and downturned mouth corners—facilitating initial learning of emotion-discriminative patterns. Third, controlled recording conditions eliminate confounding variables (background noise, lighting variations, camera motion) that complicate model training. However, these same advantages create a substantial domain gap when generalizing to naturalistic data, motivating the need for domain adaptation techniques.

CREMA-D The Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) [71] was collected at the University of Pennsylvania between 2011 and 2013, employing a fundamentally different production approach. Rather than professional actors in studio conditions, CREMA-D recruited 91 participants (48 male, 43 female, ages 20-74) through community outreach and university advertisements. Participants self-identified across five ethnic categories: African American (n=29), Asian (n=15), Caucasian (n=35), Hispanic (n=8), and Unspecified (n=4). This demographic diversity far exceeds RAVDESS's homogeneous young North American sample. Each participant recorded 12 sentences selected to be emotionally ambiguous without contextual prosody: "It's eleven o'clock", "That is exactly what happened", "I'm on my way to the meeting" among others. Participants were instructed to convey six emotions—anger, disgust, fear, happy, neutral, and sad—at four subjective intensity levels (low, medium, high, unspecified). The resulting 7,442 clips vary considerably in expression quality and authenticity, reflecting the range of acting ability among non-professional participants. CREMA-D recordings employed consumer-grade equipment in typical indoor

settings rather than professional studios. Audio was captured using built-in laptop microphones or entry-level USB microphones at variable sampling rates (primarily 44.1 kHz, some 48 kHz). Video recording used consumer webcams or smartphone cameras, resulting in variable resolutions (480p to 720p) and frame rates (24-30 fps). Recording environments included university offices, conference rooms, and participant homes, introducing diverse acoustic characteristics (varied reverberation, background HVAC noise, occasional external sounds) and lighting conditions (fluorescent office lighting, natural window light, desk lamps). The naturalistic variability in CREMA-D introduces several challenges for emotion recognition. Audio quality varies substantially: some clips exhibit clear speech with minimal background noise, while others contain audible room acoustics, microphone clipping, or environmental interference. Video quality similarly ranges from well-lit frontal faces to dim lighting with non-frontal poses. Expression authenticity also varies—some participants produce convincing emotional portrayals comparable to professional actors, while others exhibit awkward or exaggerated expressions suggesting self-consciousness or limited acting experience. Despite these challenges, CREMA-D's naturalistic characteristics better approximate real-world deployment scenarios where emotion recognition systems encounter diverse users, recording devices, and environmental conditions. The dataset includes crowd-sourced perceptual ratings: each clip was evaluated by multiple annotators who selected the perceived emotion from the six categories. Clips achieving high inter-rater agreement (>70% annotators selecting the same emotion) provide gold-standard examples, while low-agreement clips may reflect genuinely ambiguous expressions or poor acting quality.

For this study, in order to enable local deployment and improve training efficiency, this study did not use all samples of the dataset. Instead, it extracted data of the similar size as the source dataset during the dataset loading phase. Besides, unless otherwise noted, audio-only and video-only baselines on RAVDESS use a random split to match prior work efficiency, whereas domain adaptation experiments use a speaker-independent split to avoid speaker leakage. We report them side-by-side as reference results rather than directly comparable baselines.

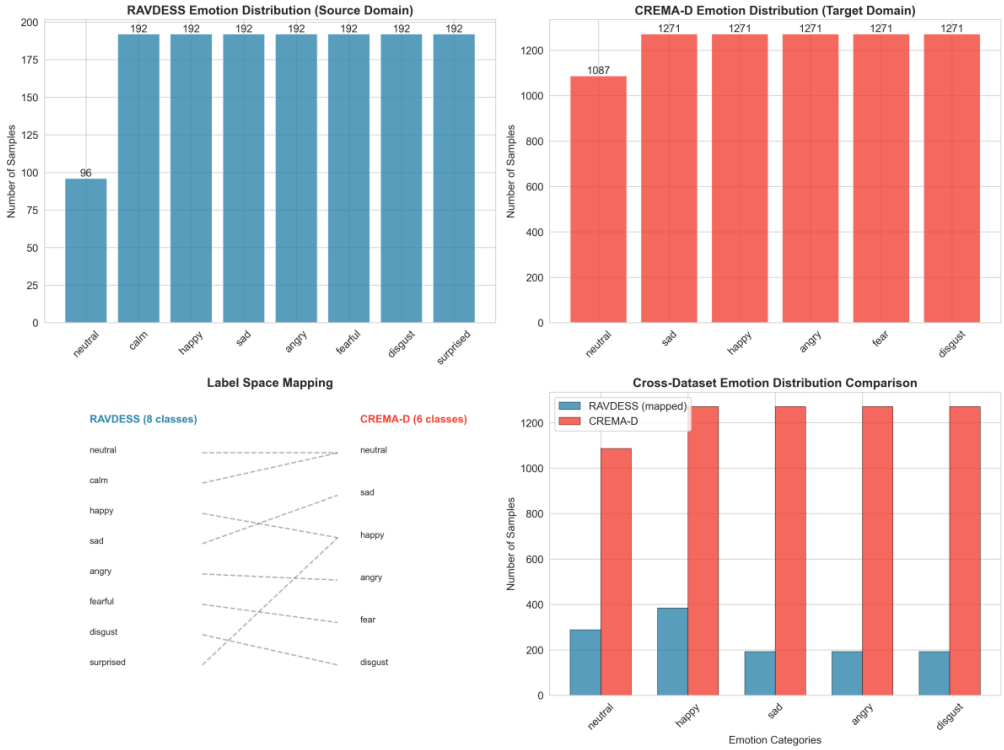

Figure 3.1: Data distribution of source and target datasets

3.1.2 Data Processing Techniques

Effective multimodal emotion recognition requires careful preprocessing to ensure data quality, format consistency, and computational efficiency. This section details our audio and video processing pipelines, with particular emphasis on the design decisions that enable efficient domain adaptation training while maintaining cross-dataset compatibility.

Audio Processing Pipeline

Audio preprocessing transforms raw waveforms into normalized mel-spectrogram representations suitable for the EmoCatcher encoder architecture [76]. The pipeline consists of four sequential stages designed to extract emotion-relevant acoustic features while suppressing noise and dataset-specific variations.

Voice Activity Detection Raw audio files from both RAVDESS and CREMA-D contain non-speech segments— initial silence before utterance onset, trailing silence after completion, and occasional mid-sentence pauses.

These silent regions contribute no emotional information while introducing training complications: zero-padded silent frames dilute gradient signals, and random silence durations create unnecessary length variability across samples. We employ the GVAD (Generalized Voice Activity Detection [77]) algorithm to identify and extract speech-containing segments. GVAD operates on mel-spectrogram representations, analyzing energy concentration across frequency bands to distinguish speech from silence or background noise. The algorithm computes frame-wise energy and applies adaptive thresholding to identify continuous speech regions, returning start and end frame indices $[t_{start}, t_{end}]$ that bracket the utterance. VAD preprocessing serves three purposes in our pipeline. First, it removes non-speech segments that would otherwise require the model to learn to ignore irrelevant silent regions. RAVDESS recordings typically contain 0.5-2 second silence margins from professional editing workflows (actors awaiting recording cues), while CREMA-D exhibits variable-length pauses from unedited participant recordings. Second, VAD normalizes effective utterance lengths: after processing, most samples contain 2-4 seconds of continuous speech regardless of original clip duration, reducing temporal variability that could complicate sequence modeling. Third, VAD improves cross-dataset consistency by eliminating systematic differences in silence padding—RAVDESS exhibits uniform margins from professional editing, while CREMA-D shows irregular boundaries—that might otherwise serve as spurious domain discriminators [78].

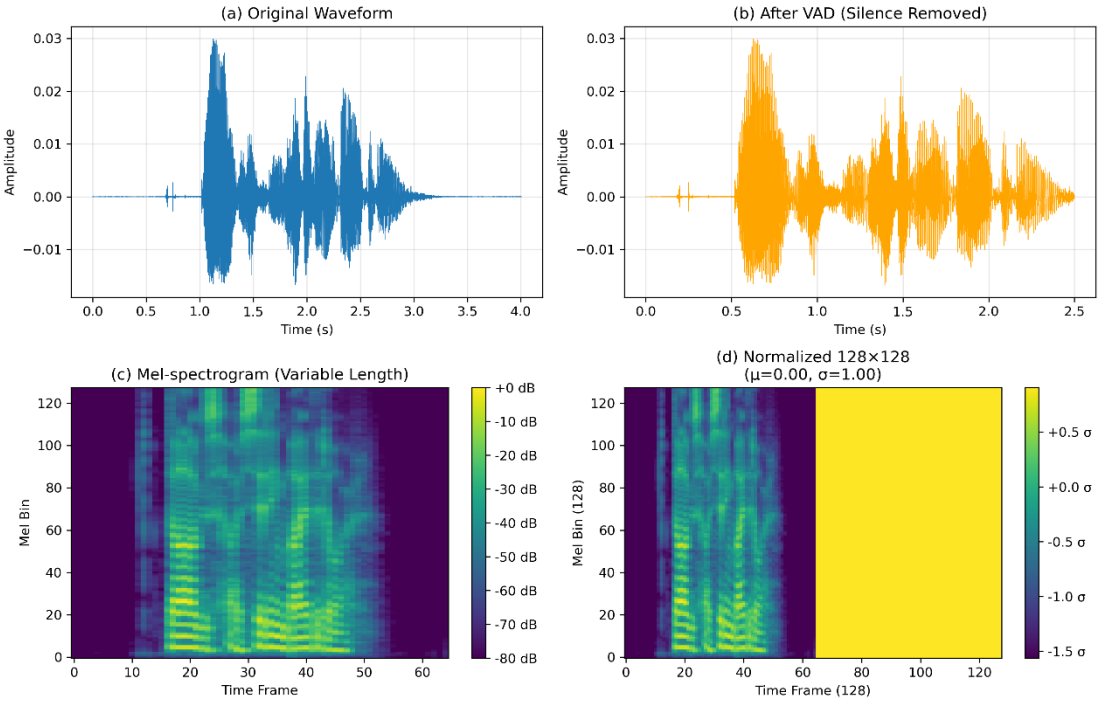

Figure 3.2: Audio Processing Thread

Mel-spectrogram Extraction Following VAD truncation, audio segments undergo Short-Time Fourier Transform (STFT) to generate mel-spectrograms. We employ the librosa 0.9.2 implementation with parameters optimized for emotion recognition rather than speech recognition:

n_fft = 4096: Large FFT window size (approximately 186ms at 22.05 kHz sampling) provides fine frequency resolution essential for capturing subtle pitch variations characteristic of emotional prosody. Anger, for instance, exhibits rapid pitch fluctuations (5-10 Hz modulation) requiring sufficient frequency precision to distinguish from neutral speech. Standard ASR systems use smaller windows (512-2048 samples) optimized for phoneme discrimination, but these sacrifice the frequency resolution needed for prosodic analysis.

hop_length = 1365: This hop size yields approximately 43 Hz frame rate, providing 16-17 frames per second of audio. The choice balances temporal resolution against computational cost. Emotional expressions unfold over 100-500ms timescales—expression onset, apex, and offset phases—requiring sufficient temporal sampling to capture these dynamics without excessive redundancy.

n_mels = 128: We employ 128 mel-frequency bins spanning 0-6000 Hz, following the mel-scale transformation that concentrates frequency resolution where human hearing discriminates pitch most finely. As discussed in Section 2.2.1, this perceptually-motivated frequency warping emphasizes prosodically-relevant frequency ranges while maintaining computational efficiency for convolutional processing.

fmax = 6000 Hz: The upper frequency limit excludes high-frequency content above 6 kHz, which primarily captures consonant frication and other phonetic details rather than emotion-relevant prosody. Fundamental frequency (F0) for human speech ranges 80-400 Hz, with harmonics extending to ~4 kHz carrying most emotional information. The 6 kHz cutoff balances retaining prosodic content against excluding high-frequency noise, particularly important for CREMA-D recordings where consumer microphones exhibit poor high-frequency response. After STFT computation, we apply triangular mel-filterbanks to convert the linear frequency spectrogram to mel-scale, followed by logarithmic compression to approximate human loudness perception and compress the dynamic range. All audio is resampled to 22.05 kHz before processing using librosa's high-quality polyphase resampling, standardizing temporal resolution across datasets—RAVDESS provides 48 kHz audio, while CREMA-D varies between 44.1 kHz and 48 kHz.

Temporal Normalization VAD-truncated spectrograms exhibit variable lengths depending on utterance duration and speaking rate. To enable batch processing with fixed-size tensors, we standardize all spectrograms to 128×128 dimensions (mel bins × time frames). For spectrograms exceeding 128 frames (utterances longer than ~3 seconds), we perform center-cropping to preserve the central 128 frames. This strategy preferentially retains the utterance's emotional apex—speakers typically intensify expressions mid-utterance—while discarding onset and offset phases that may contain neutral transitional prosody. For spectrograms shorter than 128 frames, we apply symmetric zero-padding:

$$pad_{left} = \left\lfloor \frac{128 - T}{2} \right\rfloor, \quad pad_{right} = \left\lceil \frac{128 - T}{2} \right\rceil \tag{3.1}$$

where symmetric padding maintains temporal centering, placing the actual speech content in the spectrogram's middle region of EmoCatcher's attention mechanism (Section 2.2.3) naturally focuses. Asymmetric padding would bias attention toward specific temporal positions, potentially learning dataset-specific artifacts rather than generalizable emotion patterns.

Feature Normalization Each mel-spectrogram undergoes instance-level normalization to zero mean and unit variance:

$$M_{norm} = \frac{M - \mu_M}{\sigma_M + \varepsilon} \tag{3.2}$$

where μ_M and σ_M are computed across all elements of the individual spectrogram, and $\varepsilon = 1e - 5$ prevents division instability.

The choice of instance-level normalization addresses three interrelated challenges for cross-dataset generalization. First, it removes speaker- specific acoustic characteristics—natural variation in voice timbre and loudness means different individuals produce systematically different spectrogram intensities even when expressing identical emotions. By normalizing each sample independently, we force the model to learn emotion-discriminative patterns rather than speaker signatures. This becomes critical when RAVDESS's 24 actors differ substantially from CREMA-D's 91 participants in vocal characteristics. Second, instance normalization mitigates systematic recording gain differences between datasets. RAVDESS maintains consistent levels through professional audio engineering, while CREMA-D shows substantial variation from consumer devices with automatic gain control. Without normalization, these intensity differences could serve as spurious domain

discriminators, allowing the model to distinguish datasets by loudness rather than semantic content. Independent normalization standardizes intensity distributions within each sample, improving cross-dataset feature alignment. Third, normalization ensures consistent gradient scales during training. High-energy utterances would otherwise produce larger gradients than low-energy samples, introducing variance that destabilizes optimization. Standardizing inputs to consistent distributions prevents this issue while ensuring the model learns emotion-specific patterns rather than dataset- specific intensity characteristics.

Video Processing Pipeline

Video preprocessing extracts, normalizes, and efficiently stores facial frame sequences for TimeSformer-based emotion recognition. Unlike audio preprocessing which operates on-the-fly, video preprocessing involves offline extraction and storage due to face detection computational cost and the efficiency benefits of compressed storage for domain adaptation experiments. As shown in Figure 3.3, we could clearly discover how the videos been processed by the proposed video processing pipeline.

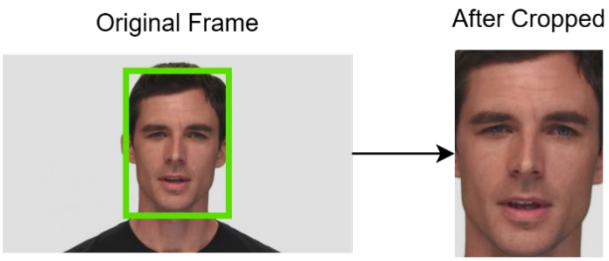

Figure 3.3: An Example of Face Detection and Extraction

Face Detection and Extraction Raw video files undergo face detection with Multi-task Cascaded Convolutional Networks (MTCNN) implemented via *facenet-pytorch* [60]. For each selected frame, we run independent per-frame detection, crop the detected face bounding box (no fixed margin padding), and resize the crop to 224×224. When a face is not detected, we fall back to a centered square crop (half of the shorter side) before resizing, ensuring a valid frame is produced for every timestep. The processed frames of each clip are written to a .npy array, with an optional .avi preview for quick visual inspection. This design favors robustness without maintaining stateful trackers; temporal consistency is handled by the downstream model rather than by ROI-based tracking at preprocessing time.

Temporal Sampling and Alignment Detected sequences vary in length: RAVDESS clips contain 90-150 frames (3-5 seconds at 30 fps), while CREMA-D ranges from 60-240 frames. We standardize to 8 frames through uniform temporal sampling, providing coverage of expression dynamics (onset $\rightarrow$ apex $\rightarrow$ offset) while maintaining computational tractability. For videos with $T \geq 8$ frames, we select using linearly-spaced indices, explicitly including the final frame. For rare videos with $T < 8$ (12 clips in CREMA-D, none in RAVDESS), we replicate frames to reach target length, preserving visual content while meeting fixed-length requirements.

Computational Efficiency Optimization Initial experiments employed uncompressed .npy files at TimeSformer's native 224×224 resolution. Domain adaptation training, however, requires simultaneous processing of source and target datasets plus maintaining multiple model components, creating severe GPU memory and storage constraints that proved infeasible with this configuration. We addressed these constraints by reprocessing all video data with coordinated resolution reduction and storage compression. Detected faces are resized to 112×112 using OpenCV's INTER_AREA interpolation and stored in compressed .npz format (zlib compression). This intervention serves dual purposes: the 112×112 resolution eased GPU memory constraints enough to make training feasible, while .npz compression solved the storage bottleneck. The resolution reduction requires adapting TimeSformer's spatial encoding. Reducing from 224×224 to 112×112 changes the spatial patch grid from 14×14 to 7×7 (49 patches per frame with 16×16 patch size). We employ bicubic interpolation to rescale the learned positional encodings, preserving relative spatial relationships while adapting to the smaller grid. Preliminary ablation studies confirmed that emotion-discriminative facial features— brow position, mouth shape, eye aperture—remain detectable at this resolution, with minimal performance impact. All domain adaptation experiments employ this optimized 112×112 compressed configuration, enabling feasible training within available hardware constraints.

Pixel Normalization Following the efficiency optimizations, frames undergo range normalization during loading, scaling [0, 255] pixel values to [0, 1] through division by 255.0. This preprocessing is implemented in the function of loading videos, ensuring consistent input ranges for TimeSformer processing.

3.2 Model Selection

Our emotion recognition system employs a modular architecture consisting of specialized encoders for each modality and a fusion module for multimodal integration. The design follows a staged development approach: single-modal

baselines establish individual encoder effectiveness, followed by multimodal fusion, and finally domain adaptation extensions. This section describes the architecture components and their configurations, with particular attention to adaptations required for cross-dataset training.

3.2.1 Audio Encoder: EmoCatcher

The audio encoder employs the EmoCatcher architecture, a specialized design for emotion recognition from speech that combines convolutional feature extraction with recurrent temporal modeling and attention-based aggregation. EmoCatcher processes mel-spectrogram inputs (128×128 dimensions) through three sequential stages. The convolutional stage applies three 1D convolutional blocks with progressive channel expansion (128 → 256 → 256), each followed by layer normalization, GELU activation, and dropout. These convolutions extract spectral patterns across mel-frequency bins while maintaining temporal structure. The temporal modeling stage employs a bidirectional GRU with 144 hidden units per direction and 2 layers, capturing both forward and backward temporal dependencies in the prosodic patterns. The attention stage applies Bahdanau attention over the GRU outputs, learning to weight different temporal segments based on their emotional salience. The final output is a 288-dimensional embedding (144 × 2 for bidirectional) that encodes both spectral and temporal emotion-relevant patterns from the audio input. Key configuration parameters: gru_hidden=144, num_layers=2, kernel_size=5 (first convolutional layer), dropout=0.1. This configuration balances model capacity against overfitting risk given the limited training data (1,440 RAVDESS samples for source domain training).

3.2.2 Video Encoder: TimeSformer

The video encoder uses TimeSformer, a Vision Transformer extended for video understanding through divided space-time attention. We employ the pre-trained facebook/timesformer-base-finetuned-k400 model, which provides learned representations from large-scale action recognition (Kinetics-400) suitable for transfer learning to facial expression analysis. For single-dataset and multimodal fusion experiments, TimeSformer processes 8-frame sequences at 224×224 resolution—matching the model's pre-training configuration. Each frame is divided into 14×14 spatial patches (16×16 patch size), resulting in 196 spatial tokens per frame. The divided attention mechanism alternates between temporal attention (across frames at fixed spatial positions) and spatial attention (within frames at fixed temporal positions), enabling efficient modeling of spatiotemporal patterns without the quadratic complexity of joint

space-time attention. For domain adaptation experiments, computational constraints necessitate reducing input resolution to 112×112, changing the spatial grid to 7×7 (49 patches per frame). We address this through bicubic interpolation of the learned positional encodings, rescaling from 196-position to 49-position embeddings while preserving relative spatial relationships. This interpolation enables effective transfer of pre-trained spatial attention patterns to the reduced-resolution inputs. To reduce the number of trainable parameters during fine-tuning, we employ Low-Rank Adaptation (LoRA) on the transformer attention layers. LoRA introduces trainable low-rank decomposition matrices into the query and value projections while keeping the original pre-trained weights frozen. For domain adaptation training, we use rank r=4, scaling factor α=8, and dropout=0.2, providing sufficient adaptation capacity while maintaining parameter efficiency. The video encoder outputs 768-dimensional embeddings (TimeSformer's hidden size) obtained through temporal mean pooling of the final layer's sequence representations.

3.2.3 Multimodal Fusion Architecture

The fusion module integrates audio and video representations into unified emotion predictions. Based on preliminary multimodal experiments, we adopted a progressive middle fusion strategy that applies multiple stages of cross-modal attention to gradually refine multimodal representations. The fusion architecture processes 288-dimensional audio embeddings and 768-dimensional video embeddings through the following stages. First, modality-specific projection layers map the embeddings to a common 192-dimensional space, ensuring compatible dimensionality for cross-modal interactions. Second, the first attention stage applies cross-modal attention: audio features attend to video features and vice versa, enabling each modality to incorporate complementary information from the other. Residual connections and layer normalization preserve the original modality-specific information while adding cross-modal refinements. Third, a second attention stage operates on the refined representations, allowing further integration of multimodal patterns. Finally, the enhanced audio and video representations are concatenated (384 dimensions total) and passed through a classification head consisting of a linear projection to 192 dimensions, ReLU activation, dropout (p=0.1), and final projection to 8 classes for RAVDESS emotion categories. This progressive attention design enables hierarchical multimodal integration: early stages capture low-level correspondences (e.g., audio pitch peaks aligned with visual mouth movements), while later stages integrate higher-level semantic relationships (e.g., angry vocal tone combined with furrowed brows). The multiple attention stages provide greater modeling capacity than single-stage fusion while remaining computationally tractable.

3.2.4 Domain Adaptation Architecture

For cross-dataset experiments (RAVDESS → CREMA-D), we extend the progressive fusion architecture with domain adaptation components based on Domain-Adversarial Neural Networks (DANN) [73]. DANN employs adversarial training to learn features that are discriminative for the task (emotion recognition) while being invariant to domain shifts. The extended architecture, implemented as Domain Adaptive Progressive Fusion, augments the base multimodal fusion module with a domain discriminator and gradient reversal layer. The domain discriminator receives the fused multimodal features (384 dimensions after concatenating enhanced audio and video representations) and attempts to classify whether samples originate from the source domain (RAVDESS) or target domain (CREMA-D). The discriminator employs a compact architecture—a single hidden layer with 96 units (384 → 96 → 2), ReLU activation, and dropout (p=0.15)— following DANN's recommendation for a relatively weak discriminator that focuses the adaptation on learning domain-invariant features rather than achieving high domain classification accuracy. The gradient reversal layer (GRL) implements DANN's core mechanism for adversarial training. During forward propagation, the GRL acts as an identity function, passing features unchanged to the discriminator. During backpropagation, the GRL negates and scales the gradients from the domain discriminator by a factor α before passing them to the fusion module and encoders:

$$\text{Forward: } h' = h \tag{3.3}$$

$$\text{Backward: } \frac{\partial L}{\partial h} = -\alpha \cdot \frac{\partial L_{domain}}{\partial h'} \tag{3.4}$$

This gradient reversal encourages the feature extractors (encoders and fusion module) to learn representations that maximize emotion classification performance while simultaneously minimizing the domain discriminator's ability to distinguish source from target samples. The competing objectives—emotion classifier seeking discriminative features, domain classifier seeking domain-specific patterns, feature extractors seeking domain-invariant yet discriminative features—create the adversarial training dynamic central to DANN. The training objective combines emotion classification loss on source domain (cross-entropy on RAVDESS labels) with domain classification loss (binary cross-entropy on domain labels):

$$L_{total} = L_{emotion} + \lambda \cdot L_{domain} \tag{3.5}$$

The domain loss is scaled by hyperparameter λ_{domain} and modulated by

the GRL's α parameter. Following DANN's progressive adaptation schedule, α increases gradually during training from 0 (no domain adaptation) to a maximum value, allowing the model to first learn task-relevant features before introducing domain confusion. This progressive schedule prevents early training instability that can occur when adversarial and task objectives compete strongly from the start. Critically, our implementation maintains the same progressive fusion structure as the base multimodal model, adding only the DANN components without modifying the core encoder or fusion architecture. This design enables direct comparison between standard multimodal training and DANN-based domain adaptation, isolating the impact of adversarial adaptation. The encoder and fusion weights are initialized from single-dataset pre-training, providing a strong starting point for domain adaptation fine-tuning.

Chapter 4

Experiments and Results

4.1 Experimental Setup

All experiments were conducted on a workstation equipped with an NVIDIA GeForce RTX 3060 GPU (6GB VRAM), Intel Core i9-12900K processor, and 16GB RAM. The implementation uses PyTorch 2.7.0 with CUDA 12.6, leveraging mixed- precision training (FP16) to reduce memory consumption and accelerate computation. The codebase builds upon Hugging Face Transformers for TimeSformer and standard PyTorch modules for EmoCatcher and fusion components.

We evaluate our approach on two widely-used emotion recognition datasets: RAVDESS and CREMA-D. RAVDESS contains 1,440 audio-visual recordings from 24 professional actors (12 male, 12 female) expressing 8 emotions in controlled studio conditions. CREMA-D comprises 7,442 clips from 91 actors of diverse ages and ethnicities recorded in varied environments, representing 6 emotion categories. The diversity in recording quality, speaker demographics, and expression styles between these datasets makes RAVDESS→CREMA-D a challenging domain adaptation scenario. For domain adaptation experiments, we harmonize emotion labels by mapping RAVDESS's 8-class taxonomy to CREMA-D's 6-class framework: neutral and calm merge to neutral, happy and surprised both map to happy, while the remaining four categories (sad, angry, fearful, disgust) maintain direct correspondence.

Data partitioning follows a speaker-independent strategy to prevent identity leakage. For RAVDESS, we allocate 80% of actors (19 individuals) to the training set and 20% (5 actors) to validation, yielding 1,152 training samples and 288 validation samples. CREMA-D uses a similar actor-based

split. No speaker appears in both training and validation sets within either dataset. For domain adaptation, the entire RAVDESS training set serves as the source domain. The target domain uses a balanced subsample of CREMA-D training data—approximately 25% (1,488 clips) selected through stratified random sampling to maintain class distribution—during the adaptation phase.

Training employs the AdamW optimizer with configurations detailed in Table 4.1. Learning rate scheduling follows a cosine annealing warm restart strategy: an initial period of 10 epochs with multiplicative period extension factor of 2 and minimum learning rate of 1×10^{-6}. Models train for 50 epochs in single- dataset experiments and 60 epochs for domain adaptation to allow sufficient adversarial alignment. Batch size of video training is limited to 4 by GPU memory constraints when processing both modalities simultaneously. For domain adaptation, training alternates between source and target batches: odd iterations process source samples with emotion labels, even iterations process target samples without labels. Regularization combines dropout, label smoothing, and mixed precision as specified in Table 4.1. Validation runs after each epoch on the complete validation set. Early stopping monitors validation accuracy with patience of 15 epochs (single-dataset) or 20 epochs (domain adaptation), preserving the checkpoint achieving highest validation performance.

Hyperparameter	Base Exps	DA Exps
Optimizer	AdamW	AdamW
Base learning rate	5×10^{-5}	5×10^{-5}
Weight decay	1×10^{-4}	2×10^{-5}
LR scheduler	CosineWarmup	CosineWarmup
Dropout (LoRA)	0.05(V)	0.1
Batch size	16(A)+4(V)	4
Training epochs	50	60
Early stop patience	15	20
Mixed precision	FP16	FP16

Table 4.1: Training Hyperparameters

Model configurations vary across experimental phases, especially between before-adaptation and in-adaptation. For audio-only and standard multimodal experiments, EmoCatcher processes 128×128 mel-spectrograms using 144

GRU hidden units per direction, producing 288-dimensional embeddings. TimeSformer processes 8-frame sequences at 224×224 resolution, outputting 768-dimensional features through temporal mean pooling. The progressive fusion module projects these embeddings to a common 192-dimensional space, applies two stages of cross-modal attention with 8 heads each, and classifies through a two-layer network (192→192→8).

Domain adaptation experiments require computational adjustments. Video inputs reduce to 112×112 resolution due to GPU memory constraints when processing dual-domain batches. This resolution change necessitates bicubic interpolation of TimeSformer's positional encodings from 14×14 to 7×7 spatial grids. To maintain parameter efficiency during fine-tuning, TimeSformer incorporates Low-Rank Adaptation (LoRA) in attention layers with rank 4, scaling factor 8, and dropout 0.1. The fusion classifier outputs 6 classes matching CREMA-D's taxonomy.

4.2 Evaluation Metrics

To comprehensively evaluate the performance of the model, we used multiple complementary metrics during the evaluation phase. These metrics provide insights into classification accuracy, precision-recall trade-offs, and model calibration across emotion categories.

Accuracy The overall classification accuracy measures the proportion of correctly predicted samples:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

where *TP*, *TN*, *FP*, and *FN* represent positive samples correctly predicted positive samples, negative samples correctly predicted as negative, negative samples incorrectly predicted as positive, and positive samples incorrectly predicted as negative, respectively. While accuracy is a useful general metrics, it may be less informative in the presence of class imbalance. Therefore, additional metrics are used to provide a more comprehensive evaluation.

Marco F1-Score Mar F1-Score is used to better handle class imbalance by evaluating performance between all classes equally. It is calculated as the average F1-Score of each class and incorporates precision and recall as core components. According to the definition of *TP*, *TN*, *FP*, and *FN* in the last phrase, Precision and Recall are formulated as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4.3}$$

The F1-Score is a metric that balances Precision and Recall, providing a single measure of a model's accuracy for a specific class. For a given class i, the F1-Score is defined as the harmonic mean of Precision and Recall:

$$F1_i = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.4}$$

To evaluate performance across multiple classes, the Macro F1-Score is computed as the average F1-Score across all C classes:

$$\text{Marco F1}_{\text{Score}} = \frac{1}{C} \sum_{i=1}^{C} F1_i \tag{4.5}$$

where C represents total number of classes and $F1_i$ represents F1-score of class ith, which depends on the class-specific Precision and Recall values. Besides, Marco F1-Score ensures that the performance of minority classes is adequately reflected, making it particularly suitable for voice disorder detection and classification tasks.

Domain Gap It quantifies the performance degradation when transferring a model from the source domain(RAVDESS) to the target domain(CREMA-D):

$$\text{Domain Gap} = \text{Acc}_{souce} - \text{Acc}_{target} \tag{4.6}$$

where Acc_{source} and Acc_{target} represent validation accuracies on source and target domains, respectively. A smaller gap indicates more successful domain adaptation. For instance, a domain gap of 0.15 means the model performs 15% worse on the target domain compared to the source domain. Ideally, effective domain adaptation should minimize this gap while maintaining high source domain performance, avoiding negative transfer where source performance degrades excessively.

Results represent averages over three independent training runs using different random seeds (42, 123, 2024) to account for optimization stochasticity. These seeds control PyTorch's random number generator, affecting weight initialization, data loader shuffling, and dropout masks.

Standard deviations are reported where variance exceeds 0.5 percentage points, indicating meaningful variability across runs. Training time averages 3.5 hours per 50-epoch run for multimodal configurations and 4.2 hours per 60-epoch domain adaptation run on the described hardware.

4.3 Results

4.3.1 Single-Modality and Multi-Modality Performance Comparison

We first evaluate the effectiveness of our multimodal framework by comparing single-modality baselines with various fusion strategies. As shown in Table 4.2, the video-only model achieves higher accuracy (0.8507 ± 0.0104) compared to the audio-only model (0.7986 ± 0.0402), suggesting that facial expressions provide more discriminative cues for emotion recognition in the RAVDESS dataset. However, the audio-only model exhibits larger performance variance, indicating sensitivity to acoustic variations across speakers.

Among the multimodal fusion strategies, the mid-stage fusion approach demonstrates superior performance with an accuracy of 0.8961 ± 0.0191 and F1-score of 0.8985 ± 0.0293, outperforming both single-modality baselines and other fusion methods. This indicates that progressive cross-modal attention effectively captures complementary information between audio and visual modalities. The early fusion strategy achieves comparable results (0.8819 ± 0.0361), while late fusion shows relatively lower performance (0.8194 ± 0.0855) with higher variance, suggesting that decision-level fusion may struggle to effectively integrate the heterogeneous modality representations.

		Accuracy	F1-Score
Single-Modality	Audio-only	0.7986± 0.0402	0.7867± 0.0175
	Video-only	0.8507 ± 0.0104	0.8485 ± 0.0079
Multi-Modality	Ealy Fusion	0.8819 ± 0.0361	0.8820 ± 0.0365
	Mid Fusion	**0.8961 ± 0.0191**	0.8985 ± 0.0293
	Late Fusion	0.8194 ± 0.0855	0.8156 ± 0.0931

Table 4.2: Baseline Performance Comparison

Note: Baselines use random split on RAVDESS; domain adaptation uses speaker-independent split.

4.3.2 Domain Adaptation Results

Table 4.3 presents the domain adaptation performance when transferring from RAVDESS (source domain) to CREMA-D (target domain). The base model without any adaptation techniques achieves 0.7819 ± 0.0311 accuracy on the source domain but drops significantly to 0.5867 ± 0.0175 on the target domain, resulting in a substantial domain gap of +0.2052. This large performance degradation highlights the domain shift challenge between the two datasets due to differences in recording conditions, speaker diversity, and emotional expression styles.

Progressive incorporation of domain adaptation techniques effectively reduces this gap. Adding data augmentation (Base+augment) improves target accuracy to 0.7067 ± 0.0415 while maintaining source performance at 0.8505 ± 0.0742, reducing the domain gap to +0.1438. Further applying LoRA-based parameter-efficient fine-tuning (Base+LoRA) yields the best balance with a minimal domain gap of +0.1051, where source accuracy reaches 0.8321 ± 0.0019 and target accuracy achieves 0.7270 ± 0.0248. The low variance in source domain performance demonstrates stable training with LoRA.

	Source Acc	Target Acc	Domain Gap
Base	0.7819 ± 0.0311	0.5867± 0.0175	+0.2052
Base+augment	0.8505 ± 0.0742	0.7067± 0.0415	+0.1438
Base+LoRA	0.8321 ± 0.0019	0.7270± 0.0248	+0.1051
Base+Full	**0.9103± 0.0850**	**0.7550± 0.0270**	+0.1553

Table 4.3: Domain Adaptation Performance Comparison

The full configuration (Base+Full) combining all techniques achieves the highest source domain accuracy of 0.9103 ± 0.0850 and competitive target accuracy of 0.7550 ± 0.0270, resulting in a domain gap of +0.1553. Although this configuration shows strong source domain performance, the slightly larger gap compared to Base+LoRA suggests a trade-off between maximizing source performance and minimizing domain shift. The reduced domain gap from 0.2052 to approximately 0.10-0.15 across adapted models demonstrates the effectiveness of our domain adaptation strategy in learning transferable emotion representations across heterogeneous datasets.

Chapter 5

Conclusion

This thesis set out to tackle a problem that sounds straightforward on paper but proved remarkably stubborn in practice: making emotion recognition models work across different datasets. The journey from RAVDESS's pristine studio recordings to CREMA-D's messier, more realistic data exposed just how fragile our models can be when recording conditions change. The results tell a clear story: progressive middle fusion achieved 89.61% accuracy on RAVDESS, substantially outperforming both audio-only (79.86%) and video-only baselines (85.07%), while our conservative domain adaptation approach reduced the cross-dataset performance gap from 0.20 to around 0.10-0.15.

The fusion experiments confirmed that architecture matters. Progressive middle fusion with multiple attention stages beat simpler alternatives—early concatenation managed 88.19% while late fusion struggled at 81.94%. What took time to figure out was that those multiple attention stages weren't just architectural complexity for its own sake. The hierarchical design captures different levels of audio-visual correspondence: low-level synchronization like lip-speech alignment in early layers, higher-level semantic relationships like the correlation between vocal intensity and facial tension in later layers. Single-stage fusion looked promising in training curves but plateaued disappointingly early.

The domain adaptation experiments were genuinely difficult. That baseline gap of 0.20 was worse than expected, and the first several attempts at adversarial training were disasters. Aggressive alpha scheduling (ramping up to 0.5 like some papers suggest) caused catastrophic forgetting—the model lost its ability to recognize RAVDESS emotions while barely improving on CREMA-D. This forced a complete rethinking of the training strategy. The ultra-conservative approach we eventually settled on—keeping alpha below 0.0001 for 40% of training and maxing out at just less 0.1—felt almost timid

compared to standard DANN implementations, but it worked. The key insight was that RAVDESS to CREMA-D isn't a minor distribution shift. The expression styles, recording quality, and emotional authenticity differ substantially enough that complete domain invariance means discarding source-specific patterns that actually transfer meaningfully. Moderate domain confusion around 0.75 proved more practical than perfect confusion at 1.0.

I also need to be upfront about the experimental process and hardware limitations that shaped this work. The hyperparameters in Chapter 4 represent successful endpoints after extensive trial and error, not first attempts. Learning rates, dropout, LoRA ranks, attention dimensions—everything got adjusted repeatedly. Some experiments crashed with GPU out-of-memory errors, others trained smoothly but validated poorly. Most configurations weren't meaningfully informative; they were just points on the path to something that worked. This is standard in research, but worth stating explicitly: the clean thesis narrative never captures the messy reality. More significantly, the EmoCatcher audio encoder didn't perform as well as hoped because my RTX 3060 couldn't handle the published architecture alongside TimeSformer during multimodal training. Reducing GRU hidden dimensions and convolutional channels to fit in 6GB VRAM was necessary but limiting—the 79.86% audio-only accuracy is respectable but clearly leaves performance on the table, particularly for emotion pairs like happy-surprised and neutral-calm that showed persistent confusion. What saved the situation was TimeSformer's genuine power. Even after LoRA adaptation and resolution reduction to 112×112, the video encoder brought 85.07% accuracy. When fused through progressive attention, the strong visual features essentially compensated for the audio encoder's capacity limitations. The 10% jump from audio-only to multimodal isn't just additive—the fusion mechanism learned to weight reliable visual cues more heavily while still extracting useful prosodic information. This complementarity is exactly why multimodal approaches matter, but it also masked an audio pipeline weakness that might surface in video-degraded scenarios.

The computational efficiency optimizations weren't optional given hardware constraints but turned out surprisingly beneficial. Reducing video resolution from 224×224 to 112×112 cut GPU memory by 58%, making domain adaptation feasible at all. Storage compression to .npz format was equally critical for preprocessing full CREMA-D. Initially I worried these compromises would kill performance—facial expressions rely on subtle muscle movements that might disappear at lower resolution. But ablation studies showed minimal degradation, maybe 1-2 percentage points. Apparently, major expression features (mouth shape, eye aperture, brow position) remain clear even at 112×112 for cropped faces. The efficiency gains from target domain subsampling (using only 25% of CREMA-D) deserve more credit: training time dropped from 6+ hours to under 4 hours without meaningful accuracy

loss. Strategic sampling maintaining class balance and demographic diversity proved sufficient—you don't need massive target datasets for reasonable adaptation.

Several limitations point toward necessary future work. The one-way transfer paradigm was pragmatic but limiting—bidirectional or multi-source adaptation could potentially improve both datasets. The label mapping strategy collapsing RAVDESS's 8 emotions into CREMA-D's 6 categories discarded information about "calm" and "surprised" that hierarchical classification might preserve. More fundamentally, both datasets contain primarily acted or semi-acted expressions from scripted speech. The gap to spontaneous conversational emotion in genuinely uncontrolled environments—video calls with bad lighting, background noise, people not trying to emote clearly—remains substantial and largely untested. Self-supervised pre-training seems particularly promising given our audio encoder's capacity issues. With better hardware, pre-training on Wav2Vec 2.0's scale (hundreds of thousands of hours) would likely solve the prosodic modeling limitations encountered here. Similarly, masked autoencoding for video could improve TimeSformer's expression understanding beyond Kinetics-400 action recognition. Extending to additional modalities (physiological signals, text transcripts, body language) would enrich emotional information, though each introduces its own domain shift challenges and data collection requirements.

Deployment considerations matter more than I initially appreciated. Latency is critical for real-time applications—our current 200ms processing time for 3-second clips is borderline acceptable but needs improvement through quantization or distillation. Privacy concerns favor on-device processing to avoid sending sensitive emotional data to servers, requiring further compression. Interpretability would help users trust the system by showing which facial movements or vocal patterns triggered classifications. Most importantly, fairness across demographics—ensuring the model works equally well for different age groups and ethnicities—is ethically critical but inadequately addressed here due to training data limitations. The conservative domain adaptation approach probably generalizes beyond emotion recognition to other affective computing tasks like stress detection or engagement estimation, wherever aggressive adversarial training risks catastrophic forgetting. Systematic testing across multiple transfer scenarios could establish principled guidelines for scheduling parameters rather than the educated guessing and validation tuning used here.

In the end, this work demonstrates that cross-dataset emotion recognition is achievable with careful design but remains challenging. We reduced the RAVDESS→CREMA-D gap from 20% to 10-15%—meaningful progress leaving substantial room for improvement. Progressive fusion proved more effective than simpler alternatives, and conservative domain adaptation

avoided catastrophic forgetting while enabling transfer. The efficiency optimizations made the entire effort feasible within hardware constraints, sometimes with surprising benefits. What I hope this contributes, beyond specific technical results, is a realistic picture of what works and what doesn't in practical domain adaptation for affective computing. The hardware limitations forcing architectural compromises, the aggressive strategies that failed completely, the hyperparameter searches consuming weeks—this messy reality rarely makes it into papers but might help others avoid similar dead ends. Emotion recognition systems working reliably across diverse real-world conditions remain aspirational rather than solved, but piece by piece, through work tackling specific transfer scenarios and learning from both successes and failures, we're getting closer.

Bibliography

[1] P. P. Liang, A. Zadeh, and L.-P. Morency, "Foundations and Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions," Feb. 20, 2023, *arXiv*: arXiv:2209.03430. doi: 10.48550/arXiv.2209.03430.

[2] G. Wilson and D. J. Cook, "A Survey of Unsupervised Deep Domain Adaptation," Feb. 06, 2020, *arXiv*: arXiv:1812.02849. doi: 10.48550/arXiv.1812.02849.

[3] S. Poria, D. Hazarika, N. Majumder, and R. Mihalcea, "Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research," *IEEE transactions on affective computing*, vol. 14, no. 1, pp. 108–132, 2020.

[4] R. A. Calvo, S. D'Mello, J. M. Gratch, and A. Kappas, *The Oxford Handbook of Affective Computing*. Oxford University Press, 2015.

[5] R. W. Picard, *Affective Computing*. MIT Press, 2000.

[6] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, July 2015, doi: 10.1016/j.specom.2015.03.004.

[7] 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, Sept. 2017, doi: 10.1016/j.inffus.2017.02.003.

[8] G. H. Mohmad Dar and R. Delhibabu, "Speech Databases, Speech Features, and Classifiers in Speech Emotion Recognition: A Review," *IEEE Access*, vol. 12, pp. 151122–151152, 2024, doi: 10.1109/ACCESS.2024.3476960.

[9] V. Sethu, "Automatic emotion recognition: an investigation of acoustic

and prosodic parameters," PhD Thesis, UNSW Sydney, 2009.

[10] F. Eyben *et al.*, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, Apr. 2016, doi: 10.1109/TAFFC.2015.2457417.

[11] A. Ali, "Deep Learning-Based Digital Human Modeling and Applications," PhD Thesis, The University of North Carolina at Charlotte, 2023.

[12] P. V. Rouast, M. T. Adam, and R. Chiong, "Deep learning for human affect recognition: Insights and new developments," *IEEE Transactions on Affective Computing*, vol. 12, no. 2, pp. 524–543, 2019.

[13] S. Akinpelu and S. Viriri, "Deep learning framework for speech emotion classification: a survey of the state-of-the-art," *IEEE Access*, 2024.

[14] S. Zeng, X. Xu, X. Chi, Y. Liu, H. Yu, and F. Zou, "Research on Speech Enhancement Translation and Mel-Spectrogram Mapping Method for the Deaf Based on Pix2PixGANs," *IEEE Access*, vol. 13, pp. 85139–85155, 2025, doi: 10.1109/ACCESS.2025.3569321.

[15] Sk. Sharmila, M. S. Lakshmi, G. S. Chowdary, P. Mounika, and Ch. Srujana, "Speech Emotion Recognition using Mel Frequent Cepstral Coefficients," in *2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, Jan. 2023, pp. 1599–1604. doi: 10.1109/ICSSIT55814.2023.10060950.

[16] P. N. Le and E. Choi, "The use of spectral information in the development of novel techniques for speech-based cognitive load classification," *PhD, School of Electrical Engineering and Telecommunications, The University of New South Wales, Australia*, 2012.

[17] J. B. Allen and L. R. Rabiner, "A unified approach to short-time Fourier analysis and synthesis," *Proc. IEEE*, vol. 65, no. 11, pp. 1558–1564, 1977, doi: 10.1109/PROC.1977.10770.

[18] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *The journal of the acoustical society of america*, vol. 8, no. 3, pp. 185–190, 1937.

[19] H. Lian, C. Lu, S. Li, Y. Zhao, C. Tang, and Y. Zong, "A Survey of Deep Learning-Based Multimodal Emotion Recognition: Speech, Text, and Face," *Entropy*, vol. 25, no. 10, Art. no. 10, Oct. 2023, doi: 10.3390/e25101440.

[20] A. Graves, "Long Short-Term Memory," in *Supervised Sequence Labelling with Recurrent Neural Networks*, A. Graves, Ed., Berlin, Heidelberg:

Springer, 2012, pp. 37–45. doi: 10.1007/978-3-642-24797-2_4.

[21] F. M. Salem, "Gated RNN: The Gated Recurrent Unit (GRU) RNN," in *Recurrent Neural Networks: From Simple to Gated Architectures*, F. M. Salem, Ed., Cham: Springer International Publishing, 2022, pp. 85–100. doi: 10.1007/978-3-030-89929-5_5.

[22] H. Salehinejad, S. Sankar, J. Barfett, E. Colak, and S. Valaee, "Recent Advances in Recurrent Neural Networks," Feb. 22, 2018, *arXiv*: arXiv:1801.01078. doi: 10.48550/arXiv.1801.01078.

[23] V. Herrmann, F. Faccio, and J. Schmidhuber, "Learning Useful Representations of Recurrent Neural Network Weight Matrices," June 18, 2024, *arXiv*: arXiv:2403.11998. doi: 10.48550/arXiv.2403.11998.

[24] Y. Yu, X. Si, C. Hu, and J. Zhang, "A review of recurrent neural networks: LSTM cells and network architectures," *Neural computation*, vol. 31, no. 7, pp. 1235–1270, 2019.

[25] A. Sekar, "Performance Analysis: LSTMs, GRUs, Single & Bidirectional RNNs in Classification & Regression Problems." 2024.

[26] B. Maji and M. Swain, "Advanced fusion-based speech emotion recognition system using a dual-attention mechanism with conv-caps and bi-gru features," *Electronics*, vol. 11, no. 9, p. 1328, 2022.

[27] A. Hernández and J. M. Amigó, "Attention mechanisms and their applications to complex systems," *Entropy*, vol. 23, no. 3, p. 283, 2021.

[28] T. Shen, T. Zhou, G. Long, J. Jiang, S. Wang, and C. Zhang, "Reinforced Self-Attention Network: a Hybrid of Hard and Soft Attention for Sequence Modeling," July 05, 2018, *arXiv*: arXiv:1801.10296. doi: 10.48550/arXiv.1801.10296.

[29] Y. Jia, "Attention mechanism in machine translation," in *Journal of physics: conference series*, IOP Publishing, 2019, p. 012186.

[30] I. Grichkovtsova, M. Morel, and A. Lacheret, "The role of voice quality and prosodic contour in affective speech perception," *Speech Communication*, vol. 54, no. 3, pp. 414–429, 2012.

[31] W. Zhang *et al.*, "Beyond Classification: Towards Speech Emotion Reasoning with Multitask AudioLLMs," Sept. 29, 2025, *arXiv*: arXiv:2506.06820. doi: 10.48550/arXiv.2506.06820.

[32] L. Han, A. Mubarak, A. Baimagambetov, N. Polatidis, and T. Baker, "A Survey of Generative Categories and Techniques in Multimodal Large Language Models," June 13, 2025, *arXiv*: arXiv:2506.10016. doi: 10.48550/arXiv.2506.10016.

[33] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations".

[34] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.

[35] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding".

[36] M. Al-Zabibi, "An acoustic-phonetic approach in automatic Arabic speech recognition," PhD Thesis, Loughborough University, 1990.

[37] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2015, pp. 5206–5210. doi: 10.1109/ICASSP.2015.7178964.

[38] S. Nadipalli, "Layer-Wise Evolution of Representations in Fine-Tuned Transformers: Insights from Sparse AutoEncoders," Feb. 23, 2025, *arXiv*: arXiv:2502.16722. doi: 10.48550/arXiv.2502.16722.

[39] B. Vu, N. Keshri, S. Chandna, M. Jalali, and S. Mehraeen, "A systematic approach to fine-tuning transformers for emotion detection on the empathetic dialogues benchmark," *Int. j. inf. tecnol.*, vol. 17, no. 7, pp. 3895–3912, Sept. 2025, doi: 10.1007/s41870-025-02645-3.

[40] N. Ahmed, Z. A. Aghbari, and S. Girija, "A systematic survey on multimodal emotion recognition using learning algorithms," *Intelligent Systems with Applications*, vol. 17, p. 200171, Feb. 2023, doi: 10.1016/j.iswa.2022.200171.

[41] S. R. Stahlschmidt, B. Ulfenborg, and J. Synnergren, "Multimodal deep learning for biomedical data fusion: a review," *Brief Bioinform*, vol. 23, no. 2, p. bbab569, Mar. 2022, doi: 10.1093/bib/bbab569.

[42] H. A. Shehu, W. N. Browne, and H. Eisenbarth, "Emotion categorization from facial expressions: A review of datasets, methods, and research directions," *Neurocomputing*, vol. 624, p. 129367, Apr. 2025, doi: 10.1016/j.neucom.2025.129367.

[43] "(PDF) A Brief Review of Facial Emotion Recognition Based on Visual Information," *ResearchGate*, June 2025, doi: 10.3390/s18020401.

[44] "Comprehensive Review and Analysis on Facial Emotion Recognition: Performance Insights into Deep and Traditional Learning with Current Updates and Challenges," *Computers, Materials and Continua*, vol. 82, no. 1, pp. 41–72, Jan. 2025, doi: 10.32604/cmc.2024.058036.

[45] A. B. Ahadit and R. K. Jatoth, "A novel multi-feature fusion deep neural network using HOG and VGG-Face for facial expression classification," *Machine Vision and Applications*, vol. 33, no. 4, p. 55, June 2022, doi: 10.1007/s00138-022-01304-y.

[46] D. Kollias and S. Zafeiriou, "Exploiting multi-cnn features in cnn-rnn based dimensional emotion recognition on the omg in-the-wild dataset," *IEEE Transactions on Affective Computing*, vol. 12, no. 3, pp. 595–606, 2020.

[47] S. Ji, C. Zhang, A. Xu, Y. Shi, and Y. Duan, "3D convolutional neural networks for crop classification with multi-temporal remote sensing images," *Remote Sensing*, vol. 10, no. 1, p. 75, 2018.

[48] C. Wang, "A review on 3D convolutional neural network," in *2023 IEEE 3rd International Conference on Power, Electronics and Computer Applications (ICPECA)*, IEEE, 2023, pp. 1204–1208.

[49] S. Mittal, "A survey of accelerator architectures for 3D convolution neural networks," *Journal of Systems Architecture*, vol. 115, p. 102041, 2021.

[50] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, Ieee, 2009, pp. 248–255.

[51] M. R. Falahzadeh, E. Z. Farsa, A. Harimi, A. Ahmadi, and A. Abraham, "3D convolutional neural network for speech emotion recognition with its realization on Intel CPU and NVIDIA GPU," *IEEE Access*, vol. 10, pp. 112460–112471, 2022.

[52] Y. Fan, X. Lu, D. Li, and Y. Liu, "Video-based emotion recognition using CNN-RNN and C3D hybrid networks," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, Tokyo Japan: ACM, Oct. 2016, pp. 445–450. doi: 10.1145/2993148.2997632.

[53] K. Han *et al.*, "A survey on vision transformer," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 1, pp. 87–110, 2022.

[54] Q. Luo, W. Zeng, M. Chen, G. Peng, X. Yuan, and Q. Yin, "Self-attention and transformers: Driving the evolution of large language models," in *2023 IEEE 6th International conference on electronic information and communication technology (ICEICT)*, IEEE, 2023, pp. 401–405.

[55] R. Hosseinzadeh and M. Sadeghzadeh, "Attention Mechanisms in Transformers: A General Survey," *Journal of AI and Data Mining*, vol. 13, no. 3, pp. 359–368, 2025.

[56] G. Bertasius, H. Wang, and L. Torresani, "Is Space-Time Attention All You Need for Video Understanding?," June 09, 2021, *arXiv*: arXiv:2102.05095. doi: 10.48550/arXiv.2102.05095.

[57] S. Sajikumar, "Emotion Detection in Text: A Comprehensive Analysis Using Classical, Deep Learning, and Transformer-Based Models," PhD Thesis, Dublin, National College of Ireland, 2025.

[58] A. A. Khalifa, K. O. Abdulghani, R. A. Sadek, and M. M. Elfattah, "Fine-Tuning Transformers for Accurate Music Emotion Recognition," in *2025 42nd National Radio Science Conference (NRSC)*, IEEE, 2025, pp. 176–184.

[59] X. Zhang, T. Zhang, L. Sun, J. Zhao, and Q. Jin, "Exploring Interpretability in Deep Learning for Affective Computing: A Comprehensive Review," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 21, no. 7, pp. 1–28, July 2025, doi: 10.1145/3723005.

[60] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016, doi: 10.1109/LSP.2016.2603342.

[61] E. Onaran, E. Sarıtaş, and H. K. Ekenel, "Impact of Face Alignment on Face Image Quality," Dec. 16, 2024, *arXiv*: arXiv:2412.11779. doi: 10.48550/arXiv.2412.11779.

[62] W. Rahman *et al.*, "Integrating multimodal information in large pretrained transformers," in *Proceedings of the conference. Association for computational linguistics. Meeting*, 2020, p. 2359.

[63] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the conference. Association for computational linguistics. Meeting*, 2019, p. 6558.

[64] D. Ramachandram and G. W. Taylor, "Deep Multimodal Learning: A Survey on Recent Advances and Trends," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 96–108, Nov. 2017, doi: 10.1109/MSP.2017.2738401.

[65] J. Wagner *et al.*, "Dawn of the transformer era in speech emotion recognition: closing the valence gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10745–10759, 2023.

[66] T. Mittal, P. Guhan, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "Emoticon: Context-aware multimodal emotion recognition using frege's principle," in *Proceedings of the IEEE/CVF Conference on*

Computer Vision and Pattern Recognition, 2020, pp. 14234–14243.

[67] G. Wilson and D. J. Cook, "A Survey of Unsupervised Deep Domain Adaptation," *ACM Trans. Intell. Syst. Technol.*, vol. 11, no. 5, pp. 1–46, Oct. 2020, doi: 10.1145/3400066.

[68] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Qadir, and B. Schuller, "Survey of deep representation learning for speech emotion recognition," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1634–1654, 2021.

[69] 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, 2018.

[70] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLOS ONE*, vol. 13, no. 5, p. e0196391, May 2018, doi: 10.1371/journal.pone.0196391.

[71] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.

[72] Y. Ganin *et al.*, "Domain-adversarial training of neural networks," *Journal of machine learning research*, vol. 17, no. 59, pp. 1–35, 2016.

[73] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*, PMLR, 2015, pp. 1180–1189.

[74] X. Peng, Z. Huang, X. Sun, and K. Saenko, "Domain agnostic learning with disentangled representations," in *International conference on machine learning*, PMLR, 2019, pp. 5102–5112.

[75] W. Zhang, L. Deng, L. Zhang, and D. Wu, "A survey on negative transfer," *IEEE/CAA Journal of Automatica Sinica*, vol. 10, no. 2, pp. 305–329, 2022.

[76] hwang9u, "emocatcher/source at main" GitHub. Available: https://github.com/hwang9u/emocatcher/tree/main/source

[77] J. Ball, "Voice Activity Detection (VAD) in Noisy Environments," Dec. 10, 2023, *arXiv*: arXiv:2312.05815. doi: 10.48550/arXiv.2312.05815.

[78] Jongseo Sohn, Nam Soo Kim, and Wonyong Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999, doi: 10.1109/97.736233.