Annotation Consensus Networks for Improving Machine Learning with Human Annotated Data

by

Ibrahim Shoer

A Dissertation Submitted to the
Graduate School of Sciences and Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Doctor of Philosophy

in

Electrical and Electronics Engineering

KOÇ ÜNİVERSİTESİ

June 17, 2025

Annotation Consensus Networks for Improving Machine Learning with Human Annotated Data

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a doctoral dissertation by

Ibrahim Shoer

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

_____

Prof. Engin Erzin

_____

Prof. Yücel Yemez

_____

Prof. Çiğdem Eroğlu Erdem

_____

Prof. Bilge Günsel

_____

Asst. Prof. Zafer Doğan

Date: _____

To my beloved parents, whose unwavering love and sacrifices laid the foundation for all my achievements. To my dear wife, for her endless patience, encouragement, and faith in me throughout this journey. To my son, a source of constant joy and motivation. And to my entire family, for their support and belief in me every step of the way. This work is for you.

ABSTRACT

Annotation Consensus Networks for Improving Machine Learning with Human Annotated Data

Ibrahim Shoer

Doctor of Philosophy in Electrical and Electronics Engineering

June 17, 2025

In deep learning applications, particularly in tasks involving subjective judgments such as emotion recognition, medical diagnosis, or content moderation, it is common to collect annotations from multiple human annotators for the same data instances. This practice captures diverse perspectives but often introduces variations or disagreements due to differences in interpretation, expertise, or individual bias. Effectively modeling and reconciling these inconsistencies is a key challenge in building robust and reliable machine learning models. Annotations are typically aggregated using simple methods such as averaging or majority voting. However, these conventional methods do not effectively capture common patterns or collective annotator behavior, and instead dilute valuable information about shared understanding among annotators.

To address this problem, this thesis introduces a novel Annotation Consensus Network (ACN) that explicitly models and leverages annotator consensus as a more reliable training signal. ACN extracts a Learned Annotation Consensus (LAC) that aligns with all human annotations, providing improved supervision for machine learning models within an end-to-end training framework. This consensus reduces label variance and enables backbone networks to achieve more accurate and stable learning. In this thesis, ACN is integrated into two widely studied tasks, video summarization and continuous emotion recognition from speech, both of which rely on human-annotated ground truth targets. Extensive experiments demonstrate that ACN significantly improves the quality of training annotations. Models trained with ACN consistently outperform those using traditional annotation aggregation methods, achieving higher accuracy, better generalization, and increased robustness to annotator variability.

This thesis highlights that explicitly modeling consensus among annotators is critical for improving deep learning performance. By learning a common representation from multiple annotators, ACN effectively harnesses collective expertise, providing a strong foundation for more accurate and reliable AI systems. This approach is beneficial for any domain facing annotation disagreements, enabling models to better understand and represent common human judgments rather than isolated individual perspectives.

ÖZETÇE

Etiketlenmiş Verilerle Makine Öğrenimini İyileştirmek için Etiket Uzlaşı Ağları

Ibrahim Shoer

Elektrik ve Elektronik Mühendisliği, Doktora

17 Haziran 2025

Derin öğrenme uygulamalarında, özellikle duygu tanıma, tıbbi teşhis veya içerik denetleme gibi öznel yargılar içeren görevlerde, aynı veri örnekleri için birden fazla insan etiketleyiciden değerlendirmeler toplamak yaygın bir pratiktir. Bu yöntem, farklı bakış açılarını yakalamaya olanak tanırken, yorumlama biçimi, uzmanlık düzeyi veya bireysel önyargılardan kaynaklanan varyasyonlara ya da fikir ayrılıklarına da neden olabilir. Bu tutarsızlıkları etkili bir şekilde modellemek ve uzlaştırmak, güçlü ve güvenilir makine öğrenimi modelleri geliştirmede önemli bir zorluktur. Etiketler genellikle ortalama alma veya çoğunluk oyu gibi basit yöntemlerle birleştirilir. Ancak bu geleneksel yöntemler, ortak desenleri ya da toplu etiketleme davranışlarını yeterince yansıtamaz ve etiketleyiciler arasındaki ortak anlayışa dair değerli bilgileri seyrelterek kayba uğratır.

Bu problemi ele almak amacıyla, bu tez, etiketleyiciler arasındaki uzlaşıyı daha güvenilir bir öğrenme sinyali olarak açıkça modelleyen ve kullanan Etiket Uzlaşı Ağı (Annotation Consensus Network (ACN)) yapısını yeni bir yaklaşım olarak önermektedir. ACN, tüm etiketlemelerle uyumlu olan Öğrenilmiş Etiket Uzlaşısı'nı (Learned Annotation Consensus (LAC)) çıkararak, uçtan uca bir eğitim çerçevesinde makine öğrenimi modelleri için daha gelişmiş bir denetim sağlar. Bu uzlaşı, etiket varyansını azaltır ve derin ağlarının daha doğru ve kararlı öğrenme gerçekleştirmesini mümkün kılar. Bu tezde, ACN iki yaygın görev olan video özetleme ve konuşmadan sürekli duygu tanıma problemlerine entegre edilmiştir; her iki görev de insan değerlendirmelerine dayalı etiketleme değer hedefleriyle çalışmaktadır. Kapsamlı deneyler, ACN'nin eğitim etiketlerinin kalitesini önemli ölçüde artırdığını göstermektedir. ACN ile eğitilen modeller, geleneksel etiket birleştirme yöntemlerini kullanan modellere kıyasla tutarlı biçimde daha yüksek doğruluk, daha iyi genelleme ve etiketlemeler arası

farklılıklara karşı daha fazla dayanıklılık sergilemektedir.

Bu tez, etiketleyiciler arasındaki uzlaşıyı açıkça modellemenin derin öğrenme başarımını artırmada kritik bir rol oynadığını vurgulamaktadır. ACN, birden fazla etiketleyiciden ortak bir temsil öğrenerek kolektif uzmanlığı etkili bir şekilde kullanmakta ve daha doğru ve güvenilir yapay zekâ sistemleri için sağlam bir temel sunmaktadır. Bu yaklaşım, etikeleme farklılıklarının yaşandığı tüm alanlar için faydalıdır; modellerin bireysel görüşler yerine ortak insan yargılarını daha iyi anlamasını ve temsil etmesini sağlar.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my advisor, Prof. Engin Erzin, for their invaluable guidance, patience, and dedication throughout the course of my doctoral studies. Your mentorship has profoundly shaped my academic and professional growth.

I also extend my sincere thanks to the KUIS AI Lab for providing a stimulating research environment, resources, and consistent support that made this work possible.

Finally, I am grateful to everyone—colleagues, friends, and mentors—who offered insights, encouragement, and collaboration during this journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

ACN	Annotation Consensus Network
LAC	Learned Annotation Consensus
GNN	Graph Neural Network
GTN	Graph Transformer Network
MLP	Multi-Layer Perceptron
AC	Annotation Consensus
ER	Emotion Recognition
SER	Speech Emotion Recognition
CER	Continuous Emotion Recognition
HRI	Human-Robot Interaction
AI	Artificial Intelligence
NLP	Natural Language Processing
CNN	Convolutional Neural Network
FC	Fully Connected Layer
RNN	Recurrent Neural Network
DQN	Deep Q-Network
W2V	Word2Vec
RECOLA	Remote Collaborative and Affective Interactions
COGNIMUSE	Cognitive Museum of Human Emotions
JESTKOD	Jest-Konuşma-Duygu-Durum Veritabanı
MSE	Mean Squared Error
ML	Machine Learning
MSVA	Multi-Source Visual Attention
PGL-SUM	Positional encoding-based Global and Local attention Summarization
RNC	Rank-N-Contrast

Chapter 1

INTRODUCTION

Recent advancements in deep learning have led to substantial breakthroughs in tasks that inherently involve subjective judgments, such as emotion recognition, video summarization, medical diagnosis, and content moderation. These domains critically depend on human annotations, as automated systems must learn from nuanced human insights to accurately reflect complex real-world interpretations [Apostolidis et al., 2021a, Poria et al., 2017, Booth and Narayanan, 2024].

Due to intrinsic human subjectivity, annotations gathered from multiple annotators frequently demonstrate considerable variation, driven by factors such as annotator expertise, cultural background, perceptual differences, and individual biases. Although multiple annotations enhance data richness by encompassing diverse human perspectives, the resulting variability introduces significant challenges for conventional machine learning frameworks, which typically assume the existence of clearly defined and consistent ground-truth labels [Cheplygina and Pluim, 2018]. Effectively handling these annotation discrepancies is therefore essential to ensure that models are robust, reliable, and generalizable in real-world scenarios.

Conventionally, annotations from multiple annotators are combined using straightforward aggregation methods like averaging, majority voting, or median-based strategies [Khetan et al., 2017, Sheshadri and Lease, 2013]. While these methods are simple and computationally efficient, they often neglect the nuanced patterns present within the annotations. Averaging, for example, may obscure informative disagreements by merging diverse perspectives into a diluted representation, potentially losing critical insights. Majority voting similarly risks amplifying dominant biases, potentially marginalizing minority views that could reflect subtle yet meaningful

insights or alternative interpretations. Such oversimplification limits the potential of machine learning models by reducing the complexity and richness inherent in human annotations [Cheplygina and Pluim, 2018]. Consequently, it is crucial to move beyond simplistic aggregation towards methods that explicitly model annotator behavior and consensus.

To address these fundamental challenges, we introduce the *Annotation Consensus Network* (ACN), a specialized neural network framework explicitly designed to model, interpret, and reconcile discrepancies among annotations from multiple evaluators. The primary innovation of ACN lies in its ability to learn a unified representation, referred to as the *Learned Annotation Consensus* (LAC), which effectively captures the shared annotator behavior and common interpretations underlying the diverse annotation sets.

By explicitly modeling consensus, ACN enhances annotation reliability, producing refined ground truths that are less sensitive to individual biases and more representative of collective judgment [Tanno et al., 2019]. Unlike traditional aggregation methods, ACN identifies and preserves valuable commonalities while effectively mitigating noise and disagreement, thereby providing a robust foundation for training more reliable and accurate machine learning models.

We apply the ACN to two deep learning tasks that are especially impacted by annotation subjectivity: *video summarization* and *continuous emotion recognition*. Both tasks require interpretation of complex content, where human annotators naturally differ in their assessments. These differences are not merely noise but reflect diverse cognitive and emotional perspectives that must be taken into account to build systems aligned with human expectations.

In both cases, ACN acts as an intermediate layer that learns from the full set of annotator inputs, modeling their areas of agreement and disagreement. The resulting Learned Annotation Consensus (LAC) serves as an improved supervisory signal for training, better capturing the human intent embedded in the annotations and preserving the expressive richness of subjective data.

The following sections explore these two domains in more depth, detailing the

role of ACN in managing annotation subjectivity and enhancing consensus-driven learning.

1.1 Video Summarization (VSum)

In the domain of *video summarization*, annotation variability presents a central challenge. The task involves condensing lengthy video content into shorter summaries that retain the most meaningful, informative, or emotionally salient segments, enabling efficient viewing and comprehension [Apostolidis et al., 2021a, Lu and Grauman, 2013]. Deep learning-based summarization systems have become essential tools for video analysis in domains such as surveillance, education, content recommendation, and media production.

Yet, due to the inherently subjective nature of summarization, annotators often differ in their judgments about which video segments are most relevant. These differences stem from diverse interpretations of narrative structure, visual importance, emotional salience, and contextual relevance [Ji et al., 2019]. The resulting annotations are often inconsistent, with significant variance in segment-level importance scores. This inconsistency complicates both model training and evaluation, as conventional supervised learning frameworks typically assume a single, objective ground truth.

While earlier efforts have addressed the lack of annotated data and variability through data augmentation, transfer learning, and the use of more sophisticated neural architectures [Kadam et al., 2022], these strategies do not fully resolve the issue of subjective disagreement among annotators. Traditional approaches for resolving annotation discrepancies, such as averaging or majority voting, are insufficient, as they often obscure important minority opinions or produce diluted supervision signals that fail to capture the underlying narrative intent [Cheplygina and Pluim, 2018].

To address this core issue, we propose integrating the *Annotation Consensus Network* (ACN) into the video summarization learning process. ACN models the full set of annotations, learning a *Learned Annotation Consensus* (LAC) that reflects the

shared judgment across annotators while minimizing the impact of individual biases and inconsistencies. By doing so, ACN provides a more stable and representative supervision signal that guides the training of summarization models in a consensus-driven manner.

Additionally, to enhance model sensitivity to inter-segment relationships, we introduce a *ranking loss* that emphasizes the relative importance of segments based on the consensus scores. Unlike conventional losses such as Mean Squared Error (MSE), which treat each frame independently, the ranking loss encourages the model to prioritize segments that are more widely agreed upon across annotators [Zhang et al., 2016, Zhu et al., 2022].

To demonstrate the integration of ACN into practical video summarization systems, we apply it across three representative baseline models:

- **Graph Transformer Network (GTN):** This baseline integrates Graph Neural Networks (GNNs) with convolutional transformer blocks to capture both spatial and temporal dependencies among video segments [Wu et al., 2020b, Zhao et al., 2021b, Shi et al., 2021]. It provides a strong foundation for modeling narrative coherence and temporal dynamics.

- **Multi-Source Visual Attention (MSVA):** MSVA is a supervised framework that combines information from multiple input streams—including RGB content, motion features, and optical flow—via parallel attention mechanisms. This multi-source fusion allows for a rich understanding of frame-level salience [Ghauri et al., 2021].

- **Positional encoding-based Global and Local attention model (PGL-SUM):** PGL-SUM employs both global and local attention mechanisms along with absolute positional encoding to model dependencies at multiple temporal resolutions [Apostolidis et al., 2021b].

These three baselines represent diverse architectural paradigms and serve as testbeds for evaluating the influence of consensus-based supervision in video sum-

marization. The integration of ACN into each of these models enables a consistent and robust learning signal, aligning model predictions more closely with the shared human understanding of video content.

In summary, video summarization serves as a compelling use case for ACN, where modeling annotation consensus is essential to overcoming the challenges posed by subjectivity and achieving more human-aligned summaries.

1.2 Continuous Emotion Recognition (CER)

For *speech emotion recognition* (SER), subjectivity in emotional interpretation presents a fundamental challenge. SER systems aim to identify human emotions by analyzing both what is said (linguistic content) and how it is said (paralinguistic cues such as tone, pitch, and tempo) [Schuller, 2018, Calvo and D'Mello, 2010]. However, human annotators frequently diverge in their emotional assessments, influenced by individual differences in emotional sensitivity, cultural background, perceptual biases, and personal experience [Yannakakis et al., 2018, Ringeval et al., 2013].

The standard approach of averaging emotion labels to derive a single target value simplifies training but removes valuable information about disagreement and emotional nuance [Poria et al., 2017, Triantafyllopoulos et al., 2023]. This simplification risks producing models that fail to capture the full range of emotional expressions or generalize across diverse populations.

To overcome these limitations, we apply the Annotation Consensus Network (ACN) to model inter-annotator consensus within SER datasets. ACN allows the system to learn from all annotator inputs, identifying underlying agreement while preserving the diversity of perspectives. The resulting Learned Annotation Consensus (LAC) offers a more representative emotional label that serves as a robust supervisory signal for emotion recognition models.

Our approach supports the development of fairer and more reliable SER systems, particularly important in cross-cultural or multilingual contexts where affective interpretations can vary significantly. This consensus-driven framework enhances the system's ability to capture subtle emotional expressions and generalize across dif-

ferent speakers and settings.

We evaluate ACN in the context of three widely used continuous emotion recognition datasets:

- **RECOLA** [Ringeval et al., 2013]: A multimodal dataset including continuous annotations of valence and arousal, capturing spontaneous emotional responses in dyadic interactions.

- **COGNIMUSE** [Giannakopoulos et al., 2017]: A multimodal and multilabel dataset including both low-level audiovisual cues and high-level affective annotations, offering a rich resource for studying subjective emotional interpretations.

- **JESTKOD** [Bozkurt et al., 2015]: a Turkish-language dataset comprising dyadic interactions in agreement and disagreement scenarios, capturing natural emotional expression across varied conversational dynamics.

Beyond discrete emotion classification, ACN's capacity to model dynamic emotional states and handle continuous labels makes it suitable for tasks such as *continuous emotion recognition* (CER), which is crucial for emotionally intelligent AI in fields like human-robot interaction (HRI), therapeutic systems, and emotionally aware virtual assistants.

ACN also contributes to fairness in emotion recognition systems by avoiding the dominance of majority-label assumptions. Instead of enforcing uniformity in training labels, it accommodates inter-rater variability, ensuring the model can respond appropriately to a broad range of emotional expressions.

Moreover, emotion recognition technologies benefit from modeling annotation consensus in scenarios beyond SER. For example, valence and arousal signals are key to emotion-based video summarization [Köprü and Erzin, 2023], where affective cues help identify emotionally salient moments for generating narrative-rich summaries. In domains like marketing, accessibility, and content personalization, affect modeling

supported by consensus learning leads to more ethically grounded and user-sensitive AI.

Ultimately, ACN provides a principled approach to addressing annotation subjectivity in emotion recognition tasks. Consensus modeling enables the construction of emotionally intelligent systems that reflect not just an averaged emotional state but a collective understanding drawn from real human judgments.

1.3 Broader Implications and Scope

Although our focus is on video summarization and emotion recognition, the underlying challenge of subjective annotations is pervasive across many fields. Domains such as medical diagnosis [Tanno et al., 2019], sentiment analysis, and large-scale crowdsourced labeling tasks [Cheplygina and Pluim, 2018] routinely encounter annotation variability due to human interpretation, perceptual limits, and domain complexity.

In such settings, traditional aggregation-based methods often fall short of capturing the richness and variability present in human-provided labels. The Annotation Consensus Network (ACN), by explicitly modeling agreement and divergence among annotators, presents a generalizable framework for learning from subjective data. It provides a principled alternative to collapsing diverse annotations into a single label, offering a more faithful representation of collective human judgment.

The ACN approach opens new directions for building artificial intelligence systems that are more interpretable, fair, and grounded in real-world human reasoning. By preserving valuable disagreements and resolving annotation inconsistencies through structured learning, ACN contributes to developing AI technologies that are not only more accurate but also more aligned with human values and perspectives.

1.4 Thesis Contributions

This thesis introduces the **Annotation Consensus Network (ACN)** as a general solution to the problem of annotation subjectivity in machine learning. Our contri-

butions span two major domains—**video summarization** and **emotion recognition**—where the variability of human annotations presents significant challenges for model training and evaluation.

Video Summarization

Our research in video summarization has led to multiple contributions:

- **Audio-Visual Video Summarization Framework**: We developed a novel audio-visual video summarization framework that integrates four methods of audio-visual information fusion using GRU-based and attention-based networks. Additionally, we introduced an explainability methodology employing audio-visual canonical correlation analysis (CCA) to elucidate the role of audio in video summarization tasks. Experimental evaluations on the TVSum and COGNIMUSE datasets demonstrated improvements in F1 score and Kendall's tau, particularly for videos with positively correlated audio-visual content. This work has been published in the proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2023 [Shoer et al., 2023].

- **ACN-Integrated Video Summarization**: Building upon the ACN framework, we integrated it with three representative backbone architectures, MSVA, PGL-SUM, and GTN, operating at both frame and segment levels. By training summarization models to align with the learned annotation consensus rather than individual annotations, we achieved consistent improvements in summarization accuracy across benchmark datasets. The introduction of the Rank-N-Contrast (RNC) loss further enhanced model performance by aligning intermediate embeddings with the continuous structure of the consensus labels [Shoer and Erzin, 2025a].

Emotion Recognition

In the domain of affective computing, we proposed the **W2V-CER-ACN** model, which combines wav2vec 2.0 audio embeddings with an ACN module to predict emotional states (arousal and valence) from speech signals. By learning consensus labels across multiple annotators and aligning model outputs with this consensus, we mitigated the effects of annotation subjectivity. The use of the Concordance Correlation Coefficient (CCC) loss ensured that predictions were both accurate and precise with respect to the ground truth. Experimental evaluations on the RECOLA and COGNIMUSE datasets demonstrated that the W2V-CER-ACN model consistently outperformed baseline models, achieving higher CCC scores for both valence and arousal predictions [Shoer and Erzin, 2025b]. These contributions underscore the value of consensus modeling in scenarios where ground truth is inherently subjective or noisy, offering a framework that is broadly applicable to a range of real-world problems.

Chapter 2

LITERATURE REVIEW

2.1 Annotation Consensus Literature

Annotation consensus strategies aim to integrate multiple human annotations into a stable and informative supervisory signal. In subjective machine learning tasks, the variability across annotators necessitates careful consensus modeling to ensure the quality and interpretability of learned models. The literature on consensus modeling spans from early heuristic aggregation methods to more recent learning-based formulations that explicitly model annotator behavior.

2.1.1 Heuristic and Statistical Aggregation Methods

Initial approaches to consensus construction employed simple statistical operations such as averaging, majority voting, and median filtering to combine multiple annotations [Cheplygina and Pluim, 2018]. These methods are attractive due to their simplicity and computational efficiency. However, they implicitly assume annotation errors are independent and symmetrically distributed, which is often violated in practice.

More refined techniques introduced preprocessing stages, including annotator filtering, annotation smoothing, and temporal alignment, especially in time-series or video tasks. Works such as [Lopes et al., 2017, Mariooryad and Busso, 2014, Metallinou and Narayanan, 2013] emphasized correcting for annotator reaction lags and interface-induced inconsistencies. Other studies proposed selection or weighting schemes based on inter-annotator agreement, correlation, or reliability metrics [Booth et al., 2018, Booth and Narayanan, 2020].

While these methods improved consistency, they still operated under the as-

sumption that variability among annotations is undesirable noise to be suppressed or averaged out—an assumption that has been increasingly questioned.

2.1.2 Modeling Structured Annotation Variability

Recent findings indicate that annotation disagreements are often systematic rather than random. For example, Booth et al. [Booth and Narayanan, 2020] demonstrated that annotator disagreements exhibit persistent patterns related to perception and interpretative bias. This insight has led to a paradigm shift from purely statistical aggregation toward modeling the structure of disagreement itself.

Tanno et al. [Tanno et al., 2019] introduced a deep probabilistic model for learning consensus in medical imaging, capturing label uncertainty while improving diagnostic accuracy. Khetan et al. [Khetan et al., 2017] proposed a framework to infer true labels while learning annotator-specific reliability, showing that jointly modeling label and annotator leads to superior performance in crowdsourced tasks.

These learning-based models represent a growing trend in subjective data modeling: treating each annotator as an information source with systematic characteristics, rather than as a noisy copy of ground truth. By capturing the latent consensus while accounting for structured noise, these methods offer more robust and interpretable training signals.

While prior work has shown the effectiveness of learning-based consensus in domains like medical diagnosis and crowd labeling, relatively few studies have applied similar strategies in the context of temporally continuous, perceptually rich domains such as video summarization or emotion recognition. These tasks are especially vulnerable to annotation subjectivity, yet still frequently rely on simplistic aggregation.

In this thesis, we build upon the structured annotation modeling perspective by introducing a domain-agnostic consensus learning framework. Our goal is to bridge the gap between existing theory and its practical application in complex perceptual tasks, where modeling consensus is not only beneficial but necessary.

2.2 Video Summarization Literature

Video summarization seeks to condense long-form video content into shorter summaries that retain key information, reducing viewer time while preserving narrative coherence [Apostolidis et al., 2021a]. The literature spans unsupervised, supervised, attention-based, ranking-based, and graph-based approaches, each offering trade-offs in accuracy, interpretability, and computational efficiency.

2.2.1 Evolution of VSum Techniques

Early approaches relied heavily on unsupervised heuristics such as clustering or saliency detection to identify keyframes or segments [De Avila et al., 2011, Ejaz et al., 2014, Mahmoud et al., 2013]. While simple, these methods lacked semantic awareness and generalizability across video types.

Supervised methods improved upon this by using annotated summaries to guide learning. Gong et al. [Gong et al., 2014] employed document summarization principles, framing summary generation as a greedy selection problem to optimize F1 overlap with human annotations. These approaches benefit from direct human supervision but are limited by annotation quality and consistency.

2.2.2 Attention-Based Methods

With the advent of self-attention, new models emerged that better captured long-range dependencies in video sequences. Zhong et al. [Ji et al., 2019] introduced self-attention encoders to model global temporal relations. Zhao et al. [Zhao et al., 2021a] proposed dual attention mechanisms for both inter-frame and intra-frame dependencies. Later works, including [Liu et al., 2020, Xiao et al., 2020, Apostolidis et al., 2021b], extended this idea using hierarchical and multi-level attention modules to account for both local and global context in segment importance scoring.

2.2.3 Ranking-Based Methods

Several works have framed summarization as a ranking problem. Hoai et al. [Hoai and Zisserman, 2015] developed rank-based models for activity recognition, demonstrating their utility in prioritizing temporal segments. Bojanowski et al. [Bojanowski et al., 2014] incorporated label ordering constraints in clustering for better temporal alignment. Zha et al. [Zha et al., 2023] proposed Rank-N-Contrast (RNC), which learns segment relationships via a regression-based objective, improving label consistency and segment ordering.

2.2.4 Graph-Based Models

Graph neural networks (GNNs) have been increasingly applied to model spatial-temporal relationships in videos. Wu et al. [Wu et al., 2020a] used dynamic graphs to represent segment transitions across large collections. Zhao et al. [Zhao et al., 2021b] and Zhu et al. [Zhu et al., 2022] proposed hierarchical and dual-graph architectures to jointly model intra-segment semantics and inter-segment structure, boosting narrative continuity.

Some recent models incorporate graph-transformer hybrids, leveraging GNNs for structure and transformers for sequence modeling. These architectures often utilize segment-level embeddings and edge similarity scores to preserve narrative flow, as seen in work inspired by ViViT backbones [Arnab et al., 2021].

2.2.5 Segment-Level vs. Frame-Level Summarization

Frame-level summarization methods assign importance scores to individual frames, then aggregate them post hoc [Kadam et al., 2022]. Although computationally simpler, they often miss higher-order temporal dependencies. Segment-level methods directly evaluate predefined segments, inherently modeling inter-segment context and maintaining narrative structure [Zhao et al., 2021b, Arnab et al., 2021]. Recent literature emphasizes segment-level modeling as more aligned with human perception and better suited to subjective tasks.

2.2.6 Annotation Variability in Summarization

One persistent challenge in supervised summarization is annotation variability. Human annotators differ in segment selection due to subjective preferences, cultural background, or task interpretation. Prior methods attempted to mitigate this by optimizing greedy F1-based objectives or constructing "oracle summaries" [Gong et al., 2014]. However, these approaches generally treat disagreement as noise, overlooking structured patterns of annotator behavior [Booth and Narayanan, 2024, Booth et al., 2018, Booth and Narayanan, 2020].

Efforts to improve consensus reliability have included annotation alignment and agreement modeling [Lopes et al., 2017, Mariooryad and Busso, 2014, Metallinou and Narayanan, 2013]. Yet, most prior works rely on static aggregation, lacking mechanisms to dynamically adapt supervision based on annotator consistency. This limitation motivates the need for learning-based consensus models in summarization, which can reflect both shared and divergent perspectives more effectively.

2.2.7 Evaluation Methodologies

Summarization quality is commonly assessed through both user studies and objective metrics. Traditional metrics include frame-level precision, recall, and F1 [Mahmoud et al., 2013, Chasanis et al., 2008], while more recent studies prefer segment-level evaluation to better capture continuity and coherence [Apostolidis et al., 2021a]. Correlation-based metrics such as Kendall's τ and Spearman's ρ are also used to assess temporal ranking consistency in predicted importance scores [Gygli et al., 2014, Apostolidis et al., 2020]. Segment-level F1, in particular, offers a more robust reflection of narrative fidelity and temporal alignment, especially in settings with variable human annotations.

2.3 Continuous Emotion Recognition Literature

Continuous emotion recognition (CER) focuses on modeling affective states—such as valence and arousal—over time, rather than assigning discrete emotion categories.

Unlike categorical emotion recognition, CER captures the dynamic and fluid nature of emotional expression, making it more suitable for applications in human-computer interaction, affective multimedia, and behavior analysis [Picard, 1997].

2.3.1 Modeling Approaches

Early CER systems relied on handcrafted features extracted from facial expressions, vocal prosody, or physiological signals. These features were typically processed by support vector regression or hidden Markov models to track emotional state trajectories [Ringeval et al., 2015].

More recent work has adopted deep learning methods to model temporal dependencies in affective signals. Recurrent neural networks (RNNs), particularly LSTMs and GRUs, have been widely applied due to their ability to capture time-series dynamics. Tzirakis et al. [Tzirakis et al., 2017] proposed an end-to-end multimodal architecture using convolutional and recurrent layers, achieving competitive performance on RECOLA.

Transformer-based models have also been introduced to address the limitations of RNNs in capturing long-range dependencies. Wagner et al. [Wagner et al., 2023] explored transformer architectures in speech emotion recognition, highlighting their capacity to improve valence modeling using paralinguistic cues alone.

Multimodal fusion has become a core research area in CER, as different modalities contribute complementary information. Attention-based fusion techniques enable models to dynamically weigh each modality's contribution based on its contextual relevance [Pan et al., 2020].

2.3.2 Challenges of Annotation Disagreement

A key bottleneck in CER remains the inconsistency among annotators, particularly in subjective dimensions like valence. Studies consistently report lower inter-annotator agreement on valence than on arousal, reflecting the difficulty in interpreting subtle emotional nuances, especially from speech alone [Ringeval et al., 2013].

COGNIMUSE further reveals that emotion perception varies depending on genre,

context, and viewer background, leading to substantial variance in annotated emotional trajectories [Giannakopoulos et al., 2017]. These disagreements pose challenges not only in model training but also in evaluation, as the reliability of "ground truth" labels is inherently limited.

2.3.3 Consensus Modeling in Emotion Datasets

Several works have explored methods for reconciling annotation disagreement in affective data. Grimm et al. [Grimm et al., 2007] proposed weighting annotators based on reliability, while Hayat et al. [Hayat et al., 2021] advocated for multitask learning frameworks that treat each annotator as a separate output stream. These methods acknowledge disagreement as meaningful and aim to incorporate rather than suppress it.

Despite these efforts, many current models still rely on simple averaging to define consensus, discarding variance information. This not only reduces annotation richness but can introduce bias, particularly in populations with underrepresented perspectives.

Recent trends in affective computing advocate for multi-annotator learning frameworks that treat all annotations as valid and aim to extract a latent consensus through structured learning. This approach aligns with findings in other domains (e.g., medical imaging, crowdsourcing) that demonstrate the benefits of learning from disagreement [Tanno et al., 2019, Khetan et al., 2017].

Although consensus modeling has gained traction in categorical affect recognition and static image tasks, its application to continuous, multimodal, and temporally aligned emotion data remains underexplored. Our work addresses this gap by introducing a consensus learning framework capable of modeling annotation variability in continuous emotion labels across time.

This approach preserves the temporal structure of emotional expression while extracting a robust, human-aligned supervisory signal—an essential capability for building socially aware, emotionally intelligent systems.

Chapter 3

METHODOLOGY

This chapter presents a unified framework based on the Annotation Consensus Network (ACN), applied to two tasks: video summarization and continuous emotion recognition. While the core idea, learning a robust consensus from multiple annotations, remains consistent, the implementation details differ in terms of architecture, input modality, and loss formulations.

3.1 Annotation Consensus Network (ACN)

ACN is designed as a general framework for mitigating annotation subjectivity in supervised learning tasks. It is model-agnostic and can be integrated as a plug-in module into various deep learning models. At its core, ACN learns a consensus signal $\bar{\mathbf{y}}$ from multiple annotators' labels, which can then be used to supervise the primary task model more reliably. ACN guided learning is illustrated in Figure 3.1

In this work, we explore two ACN variants: an MLP-based ACN and a convolutional ACN.

MLP-based ACN: In this variant, the ACN is implemented as a multilayer perceptron (MLP) ACN_{MLP}. Each input is a matrix of annotations of size $M \times U$, where M is the number of time steps and U is the number of annotators. The MLP variant flattens the $M \times U$ annotation matrix into a single input vector and passes it through several fully connected layers (FC) with ReLU activations. These layers progressively transform the raw annotations into a latent space from which consensus labels $\bar{\mathbf{y}}$ are predicted through a final linear projection layer. **Convolutional ACN:** The convolutional variant of ACN (ACN_{CNN}) is tailored for settings where local temporal or spatial patterns among annotations may carry useful information. It applies a series of 1D convolutional layers across the temporal dimension

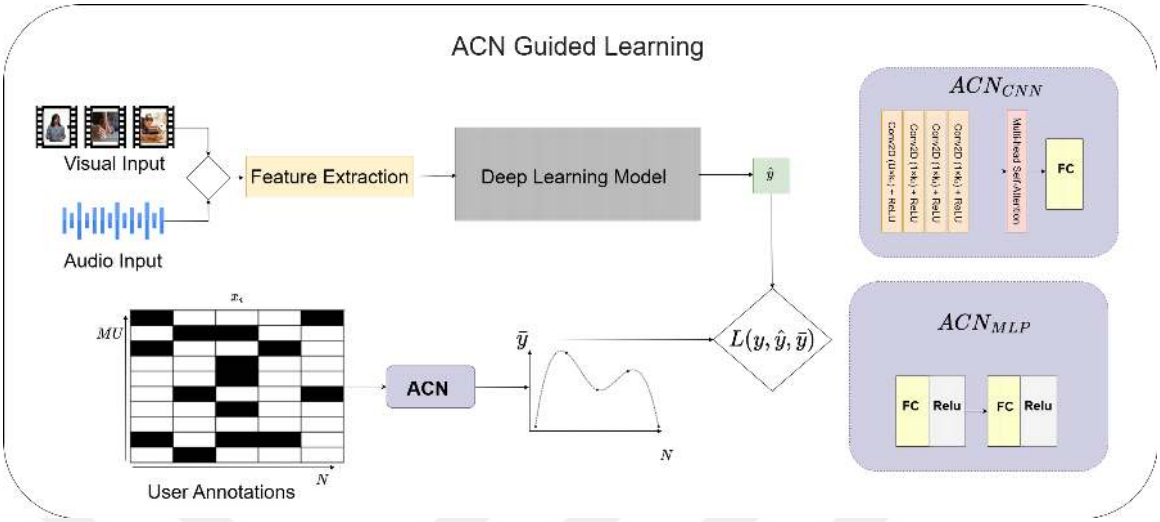

Figure 3.1: The proposed ACN guided Learning

(and optionally the annotator dimension). Each convolutional layer is typically followed by batch normalization, a ReLU activation, and dropout for regularization, as shown in Figure 3.1. The architecture may include multiple convolutional blocks, each comprising convolution, normalization, and activation, to progressively extract hierarchical features from the annotations.

The convolutional ACN architecture depicted in Figure 3.1 starts by reshaping the $M \times U$ annotation matrix into a 2D input, followed by 1D convolutions to reduce the annotator dimension. Three temporal convolution layers with kernel sizes of 9, 7, and 5 are then applied, each followed by ReLU activations. A learnable CLS token is introduced and concatenated to the output. This sequence is passed through a multi-head self-attention mechanism, where the CLS token attends to all positions and captures a condensed embedding of the annotation dynamics.

Finally, a fully connected layer maps the CLS token embedding to the consensus output $\bar{y}_i$. This process is repeated for each of the N video units to produce $\bar{\mathbf{y}} \in \mathbb{R}^N$.

These two variants demonstrate the flexibility of ACN's architecture. Once defined, they can be embedded within downstream learning pipelines for many tasks. In the following task-specific sections, we integrate these ACN modules accordingly.

3.2 VSum with ACN

We propose an Annotation Consensus Network (ACN) guided video summarization framework that enhances traditional summarization models by refining human annotations through a learned consensus representation. The pipeline consists of two parallel streams: one for visual content processing using a backbone summarization model, and another for annotation processing via the ACN. As illustrated in Figure 3.2, extracted visual features are input to the summarization backbone to produce importance scores $\hat{y}$, while annotation vectors from all human annotators are processed by the ACN to generate consensus labels $\bar{y}$. These outputs are optimized jointly through a hybrid loss function incorporating both regression and contrastive components. Consequently, the learned annotation consensus $\bar{y}$ guides the video summarization model by reducing the effects of annotation variability arising from human subjectivity. We present the ACN guided video summarization as

3.2.1 Extraction of Visual Features

We employ two strategies for extracting visual features depending on the input unit granularity used by the summarization model. In the first strategy, units are sampled uniformly from the video (e.g., at 2 frames per second), and visual features are extracted using the *pool5* layer of a pre-trained GoogLeNet [Szegedy et al., 2015], resulting in a feature vector $\mathbf{f}_i \in \mathbb{R}^D$ for each sampled unit.

In the second strategy, videos are first temporally segmented using Kernel Temporal Segmentation (KTS). Frames within each segment are passed through the ViViT transformer [Arnab et al., 2021], and their outputs are aggregated into a segment-level feature vector $\mathbf{f}_i \in \mathbb{R}^D$. This representation encapsulates both spatial and temporal dynamics across the segment, providing a high-level abstraction for summarization.

3.2.2 Loss Functions

We utilize a hybrid loss formulation that includes both standard regression and ranking components to improve both label fidelity and embedding structure.

We train our summarization models using the Mean Squared Error (MSE) loss, defined as:

$$\mathcal{L}_{\text{MSE}}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2, \tag{3.1}$$

where $\mathbf{y}$ and $\hat{\mathbf{y}}$ represent the ground truth and predicted scores, respectively, for each segment or frame.

Additionally, we employ the Rank-N-Contrast (RNC) loss [Zha et al., 2023], which enhances the supervision signal by enforcing alignment between embedding distances and label distances in a continuous space.

Let $x_i \in \mathcal{X}$ be an input (e.g., a video shot), $y_i \in \mathbb{R}$ its continuous label (e.g., an importance score), and $f_\theta(x_i) = \mathbf{v}_i \in \mathbb{R}^d$ its ℓ_2-normalized embedding. Cosine similarity is defined as $\text{sim}(\mathbf{v}_i, \mathbf{v}_j) = \mathbf{v}_i^\top \mathbf{v}_j$. A temperature parameter $\tau > 0$ is used to control the contrastive sharpness.

The higher-rank set is defined as:

$$S_{i,j} = \{\mathbf{v}_k \,|\, k \neq i,\ |y_i - y_k| \geq |y_i - y_j|\}\,, \tag{3.2}$$

which includes all embeddings $\mathbf{v}_k$ that are farther in label space from y_i than y_j is.

The RNC loss objective is:

$$\mathcal{L}_{\text{RNC}}(\mathbf{y}, \mathbf{v}) = \frac{-1}{N(N-1)} \sum_{i \neq j} \log \frac{e^{\text{sim}(\mathbf{v}_i, \mathbf{v}_j)/\tau}}{\sum_{\mathbf{v}_k \in S_{i,j}} e^{\text{sim}(\mathbf{v}_i, \mathbf{v}_k)/\tau}}. \tag{3.3}$$

This loss encourages embedding pairs with similar labels to be close in embedding space while pushing dissimilar ones apart in proportion to their label differences.

3.2.3 VSum Models

We evaluate three video summarization models: MSVA [Ghauri et al., 2021], PGL-SUM [Apostolidis et al., 2021b], and GTN [Shi et al., 2021]. These models differ in their architecture and their input granularity (frame vs. segment-level).

MSVA and PGL (Frame-level Models)

MSVA and PGL operate on video frames sampled uniformly at 2 frames per second. Visual features are extracted using GoogLeNet's *pool5* layer. Frame-level annotations from the dataset are used as ground truth $\mathbf{y}$.

MSVA backbone: MSVA is a lightweight frame-level model that uses a pairwise self-attention block followed by a shallow projection head. For each frame feature $x_t \in \mathbb{R}^D$, the model attends to a temporal window $[t-p, t+p]$, scoring pairs via a scaled dot product between learned projections U and V. Attention weights $\alpha_{t,i}$ are computed via softmax and used to form a context vector $\tilde{x}_t$, which is passed through a two-layer MLP and LayerNorm to produce a latent representation z_t. A final linear regressor outputs the importance score $\hat{y}_t \in [0, 1]$.

PGL-SUM backbone: PGL-SUM enriches each frame feature using both global and local self-attention. A global multi-head attention layer (Transformer-style, without $1/\sqrt{d_k}$ scaling) operates over the full sequence, with absolute positional encodings added to the attention logits to produce position-aware descriptors z_t^G. Simultaneously, the sequence is divided into M segments, each processed by a local attention module with reduced query/key/value dimensions $D/(H \cdot M)$, yielding local descriptors z_t^L. These are combined as $z_t' = z_t^G \oplus z_t^L$, added to the original input x_t via a residual connection, and passed through dropout, LayerNorm, and a two-layer regressor to compute $\hat{y}_t$.

GTN (Segment-level Model)

GTN operates on segments obtained via Kernel Temporal Segmentation. For each segment, visual features are extracted using the ViViT model and averaged to form segment-level embeddings.

GTN treats segments as nodes in a graph and constructs edges based on temporal distance, which are later refined using feature similarity. A stack of graph transformer layers propagates information and produces segment-level summarization scores $\hat{y}_i$. Segment-level ground truth $\mathbf{y}$ is the average frame-level annotations across the segment.

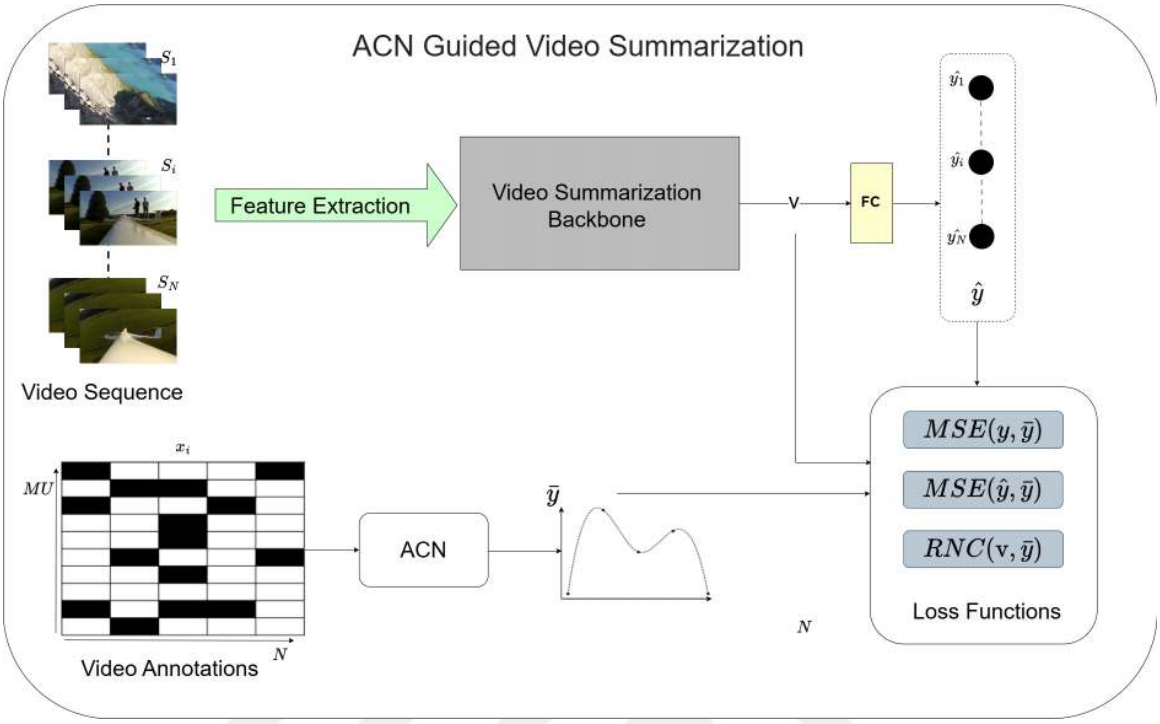

Figure 3.2: The proposed ACN guided video summarization architecture

GTN backbone: The segment-level summarization employs the GTN architecture, illustrated in Figure 3.3. Each segment feature vector $\mathbf{f}_i$ is represented as a node within a graph structure. Initially, edges are weighted inversely proportional to temporal distance, emphasizing stronger relationships between temporally adjacent segments. The GTN comprises multiple graph transformer layers, leveraging convolutional transformer layers and multi-head attention mechanisms to model both local (temporal proximity) and global (semantic similarity) relationships among segments.

Following initial graph construction, segment embeddings ($\mathbf{e}_i$) are refined, and similarity scores (σ_{ij}) between segment pairs are computed to update edge weights, thereby reconfiguring the graph to reflect semantic rather than solely temporal relationships. A subsequent graph transformer layer then outputs segment probability scores $\hat{y}_i$, indicating each segment's relevance for inclusion in the final video summary.

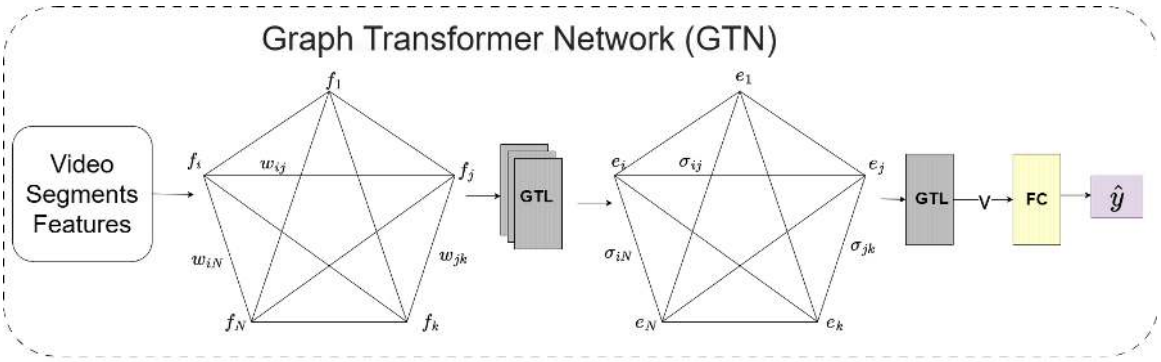

Figure 3.3: Graph Transformer Network (GTN).

Training Configurations

Each model is trained under four configurations:

1. Baseline: Only the backbone summarization model is trained using MSE loss $\mathcal{L}_{\text{MSE}}(\mathbf{y}, \hat{\mathbf{y}})$, where $\mathbf{y}$ is the ground truth annotations, and $\hat{\mathbf{y}}$ is the output of the video summarization backbone.

2. +RNC: Only the backbone summarization model is trained using MSE and RNC loss:

$$\mathcal{L}_{+\text{ANC}} = \alpha \cdot \mathcal{L}_{\text{MSE}}(\mathbf{y}, \hat{\mathbf{y}}) + (1 - \alpha) \cdot \mathcal{L}_{\text{RNC}}(\mathbf{y}, \mathbf{v}), \tag{3.4}$$

where $\mathbf{v}$ is the embedding vector of the video summarization backbone coming before the last fully connected layer, and α is a weighting factor that balances the contributions of both loss functions.

3. +ACN: The backbone and ACN are trained jointly using a consensus-based loss:

$$\mathcal{L}_{+\text{ACN}} = \alpha \cdot \mathcal{L}_{\text{MSE}}(\hat{\mathbf{y}}, \bar{\mathbf{y}}) + (1 - \alpha) \cdot \mathcal{L}_{\text{MSE}}(\mathbf{y}, \bar{\mathbf{y}}), \tag{3.5}$$

where $\bar{\mathbf{y}}$ is the output of the ACN network.

4. +ACN+RNC: We incorporate RNC to regularize the intermediate embeddings v:

$$\mathcal{L}_{+\text{ACN+RNC}} = \beta \cdot \mathcal{L}_{\text{RNC}}(\bar{\mathbf{y}}, \mathbf{v}) + (1 - \beta) \cdot \mathcal{L}_{+\text{ACN}}, \tag{3.6}$$

where β is a weighting factor that balances the contributions of the RNC loss.

Visual features are propagated through the backbone to estimate $\hat{y}_i$. For frame-level models, annotation windows centered around each sampled frame are passed to the ACN to compute $\bar{y}_i$. For segment-level models, M frames are uniformly sampled from each segment, and their corresponding annotations are aggregated and passed to the ACN to compute the consensus score $\bar{y}_i$ for that segment. The RNC loss ranks the embeddings from v according to the learned annotation consensus (LAC).

These experimental setups allow us to investigate the effect of consensus modeling and ranking-based supervision across both frame- and segment-level summarization models.

3.3 CER with ACN

Our primary objective is to predict two core dimensions of affect—arousal and valence—from raw audio signals using the wav2vec 2.0 architecture, which we fine-tune for affective computing tasks. Following recommendations in [Wagner et al., 2023], we freeze the convolutional feature encoder layers while fine-tuning the higher-level transformer blocks. This approach leverages the robust, pre-trained audio representations in the lower layers and focuses the training updates on the transformer layers that are more directly responsible for capturing fine emotional details.

The continuous emotion recognition network, denoted as *W2V-CER*, processes a batch of N audio signals $\{S_1, S_i, ..., S_N\}$ through the frozen CNN layers to obtain latent representations, which are then enriched contextually by the trainable transformer layers. A regression head maps each contextually enriched embedding to predicted outputs $\{\hat{y}_1, \hat{y}_i, ..., \hat{y}_N\}$, assessing the arousal and valence dimensions.

In addition to the main CER model, we introduce the *Annotators Consensus Network (ACN)* to address the inherent subjectivity in emotional annotations from multiple annotators. For a given audio segment annotated by U annotators, each providing M temporal annotations, the resulting inputs are represented as $\{A_1, A_i, ..., A_N\}$ with a shape of $M \times U$ for each segment. The ACN processes these annotations through an MLP to produce consensus outputs $\{\bar{y}_1, \bar{y}_i, ..., \bar{y}_N\}$ on the

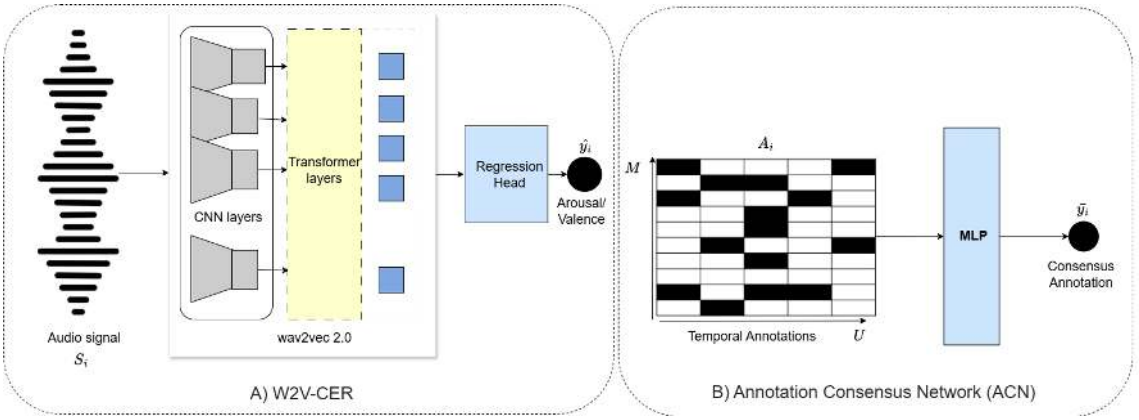

Figure 3.4: The architecture of the continuous emotion recognition network via learning annotation consensus consists of two integrated components: A) the baseline W2V-CER network, and B) the annotation consensus network (ACN), which incorporates multi-annotator learning.

emotional states, reducing variability and improving the generalizability of the predictions. The architecture of both networks is depicted in Figure 3.4. The models are trained for either arousal or valence. We tried to jointly train arousal and valence, an improvement was only noticed on the Recola dataset for valence, we report the results of the joint training on Recola in the results section 4.2.

3.3.1 Loss Function

To train our models, we employed the Concordance Correlation Coefficient (CCC) loss, which is particularly well-suited for tasks where maintaining agreement between predicted and actual values is crucial. The CCC loss is formulated as follows:

$$L_{CCC}(x, y) = 1 - \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \tag{3.7}$$

where ρ is the Pearson correlation coefficient between the predictions and the ground truth, providing a robust metric that combines measures of accuracy and precision.

3.3.2 CER Models

The baseline CER model, W2V-CER, is trained with the $L_{CCC}(y, \hat{y})$ loss function that aligns the W2V-CER predictions with the ground truth annotations.

Our proposed CER model, W2V-CER-ACN, is a combination of W2V-CER and ACN networks that are jointly trained with the weighted loss, $L_{CER-ACN}$, given as

$$L_{CER-ACN} = \alpha \cdot L_{CCC}(y, \bar{y}) + (1 - \alpha) \cdot L_{CCC}(\bar{y}, \hat{y}), \tag{3.8}$$

where α a hyperparameter that balance the influence of two loss functions. The loss function $L_{CCC}(y, \bar{y})$ measures the discrepancy between the ground truth annotations and the learned consensus from the ACN. The second loss, $L_{CCC}(\bar{y}, \hat{y})$, evaluates the alignment between the learned consensus and the predictions of the W2V-CER model. This dual-loss framework ensures that the model accurately captures true annotations while maintaining consistency with the consensus derived from multiple annotators' perspectives.

By integrating both the W2V-CER and the ACN in this manner, our system is robust to the subjectivity of individual annotators and reduces model complexity by minimizing the number of trainable parameters. This configuration leads to stable convergence and lower computational overhead, which is particularly advantageous in large-scale applications.

Chapter 4

EXPERIMENTAL EVALUATIONS

In this chapter, we present a comprehensive evaluation of our proposed Annotation Consensus Network (ACN) framework across two distinct tasks: video summarization and continuous emotion recognition. The experiments are designed to assess the effectiveness of the ACN in mitigating annotator subjectivity, enhancing supervision quality, and improving model performance. We begin by detailing the datasets used for each task, followed by descriptions of baseline models and training configurations. Quantitative results are reported using standard evaluation metrics, while ablation studies highlight the contributions of individual components such as consensus learning and contrastive loss. Finally, we analyze qualitative outputs to provide deeper insights into model behavior under various annotation conditions.

4.1 Video Summarization Evaluations

4.1.1 Datasets

In this study, we use two prominent datasets, TVSum [Song et al., 2015] and SumMe [Zhang et al., 2016], for training and evaluating our video summarization models. The TVSum dataset, sourced from YouTube, features a wide variety of content spread across 10 categories, including culinary tutorials and automotive maintenance, encompassing a total of 50 videos. In contrast, the SumMe dataset consists of user-generated videos and professional footage that capture various real-life activities, totaling 25 videos. This combination of datasets provides a comprehensive range of video types and contexts, enabling robust training and performance evaluation of our summarization models.

For TVSum, the total number of annotators is $M_{TV} = 20$. For SumMe, where

the number of annotators varies between 15 and 18, we set the number of annotators as $M_{SM} = 18$ for simplicity, with the missing users represented by the average of the available annotations.

4.1.2 Training

In the experimental evaluations, we use a five-fold cross-validation approach with ordered, non-overlapping splits. Each model is trained for up to 300 epochs, incorporating early stopping [Ghauri et al., 2021]. We train all methods under identical conditions to ensure a fair comparison. The performance of these models is thoroughly analyzed and discussed in Section 4.1.5.

By default, the average annotation (AA) scores are used as the training targets for the GTN model, while the frame-level annotations provided in the dataset serve as training targets for both MSVA and PGL-SUM.

All our experiments were conducted using an NVIDIA V100 GPU, leveraging the PyTorch framework along with PyTorch Geometric for graph-based computations and modeling.

4.1.3 Evaluation Metrics

The primary metric for evaluating video summarization models is the F1 score, which is defined as the harmonic mean of recall and precision. The F1 score is calculated as

$$\mathrm{F1} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}, \tag{4.1}$$

where TP, FP, and FN refer to true positives, false positives, and false negatives, respectively. True positives represent instances where both the model and the ground truth agree on including a segment in the summary, reflecting the model's accuracy. False positives occur when the model incorrectly includes a segment not present in the ground truth, indicating a tendency to overestimate its importance and thereby compromising precision. False negatives represent segments missed by the model that should have been included according to the ground truth.

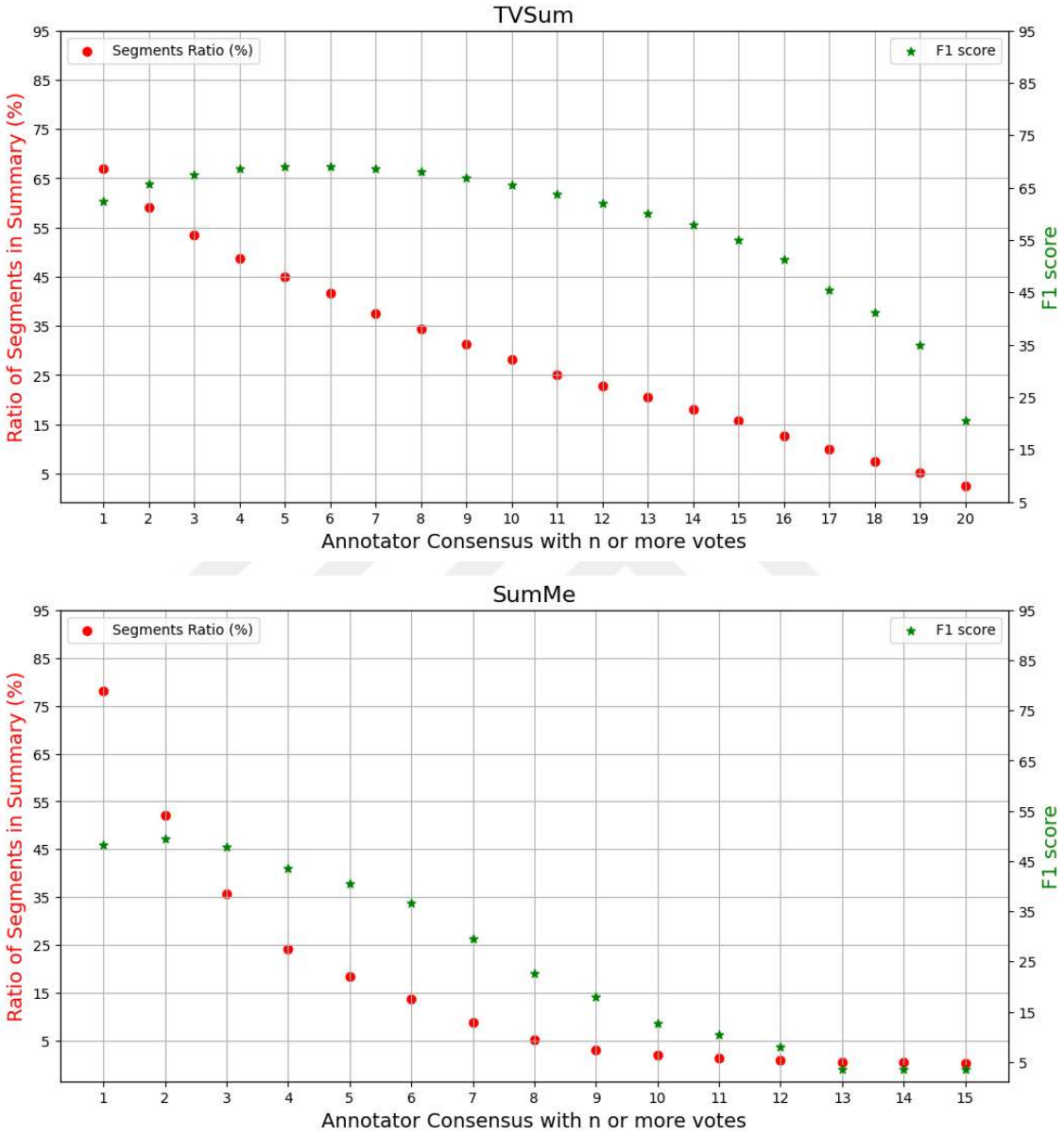

Figure 4.1: Graphs showing the ratio of segments in summary (red circles, left axis) and F1 scores (green diamonds, right axis) against annotator consensus, based on segments receiving votes from n or more annotators for TVSum (Top) and SumMe (Bottom) videos

4.1.4 Ground Truth Characterizations

Another key performance-related parameter is the selection of segments included in the video summary, which serves as the ground truth for evaluation. The variability in annotations significantly impacts the effectiveness of video summarization models. Typically, a segment is included in the summary based on annotations from M annotators, where $M > 15$ for the datasets used, as defined in 4.1.1. The diversity of annotations makes it challenging to establish a definitive rule for segment inclusion. Few, if any, segments are unanimously selected by all annotators. It is common for a segment to be chosen by at least one annotator, highlighting the subjective nature of the task.

While average annotation (AA) serves as a viable ground truth in many studies, we also consider a new annotator consensus-based approach for ground truth generation. A consensus-based ground truth can eliminate outliers and provide an alternative deterministic baseline for comparison with the learned LAC ground truth. For this purpose, we analyze summarization annotations based on the number of votes each segment received from n-or-more annotators. The number of votes for segment i is defined as

$$n_i = \left\lfloor \frac{1}{M} \sum_{m=1}^{M} \sum_{u=1}^{U} \mathbf{x}_i(m,u) \right\rfloor, \tag{4.2}$$

where $\mathbf{x}_i(m,u)$ is the annotation vote of annotator u at frame m in the i-th segment. Figure 4.1 illustrates the ratio of segments included in the summary (left axis) and the corresponding F1 scores (right axis) for varying minimum consensus values of n, which represent the number of annotators who selected the segments in the ground-truth summaries for the TVSum and SumMe datasets. The F1 scores evaluate annotator consensus by comparing each annotator's segment selections with the ground truth summaries defined by the n-or-more consensus.

The desired ground truth for the video summary aims to achieve a relatively high F1 score while including a low ratio of segments in the summary. Figure 4.1 shows that segments with a consensus of 10 or more votes achieve a 65% F1 score for the TVSum dataset while limiting the summary to approximately 30% of the total

segments. Similarly, segments with a consensus of 4 or more votes achieve a 45% F1 score while limiting the summary to approximately 34% of the total segments for the SumMe dataset. Based on this analysis, we define a new annotation consensus (AC) ground truth, where a segment is considered part of the summary if it receives votes from at least 10 annotators in the TVSum dataset and at least 4 annotators in the SumMe dataset.

Eventually, we have two ground truth sets, AC and AA, derived from the annotations and the learned segment-level ground truth, LAC.

Note that while AA and LAC ground truth sets have soft scores, original annotations, and the AC ground truth have hard binary scores. The knapsack algorithm [Song et al., 2015] is used in F1 score calculations with soft-score ground truths and soft-score network outputs at both the frame and segment levels. The knapsack algorithm evaluates each segment based on two factors: the score, which represents its relevance or importance, and the number of frames in the segment, which reflects its "cost" in terms of duration. By balancing these factors, the algorithm determines the combination of segments that maximizes the total score while targeting a 15% frame ratio constraint. This approach ensures the selection of the most representative and concise segments, prioritizing those with higher scores and shorter durations, thereby producing efficient and meaningful summaries.

Table 4.1 presents average frame-level F1 scores for selected annotation ground truths against the original annotations from the annotators (A). The first row evaluates the annotations of each annotator against all other annotators, hence it is A vs A. The variability in video summarization across annotators poses a significant challenge for model evaluation. The second and third rows in Table 4.1 evaluate the fitness of annotations against the AC and AA ground truths, respectively. For both datasets, the F1 scores are higher than those for A, indicating that the annotations align more closely with AC and AA. This suggests that AC and AA provide more reliable ground truths than individual annotators. Specifically, the AC evaluation yields a higher F1 score, particularly for the SumMe dataset, demonstrating strong alignment with the annotators' segment decisions.

The last row in Table 4.1 presents the LAC F1 scores, which are higher than the A scores but remain below those of AC and AA. This suggests that LAC provides a reliable ground truth, with annotations aligning closely to it. However, LAC does not fully converge with either the AC or AA ground truths. This divergence is expected, as the learning of annotation consensus in LAC is guided by a dual optimization objective: aligning with the AA ground truth and simultaneously enhancing the performance of the video summarization network. This two-fold optimization distinguishes LAC from both AA and AC, while still supporting improved summarization performance by the backbone model.

Table 4.1: Average frame-level F1 scores for the ground truths against the original annotations from the annotators (A)

GT	SumMe	TVSum
A	30.76	53.83
AC	43.47	65.47
AA	35.19	66.08
LAC	35.11	62.49

4.1.5 VSum Results

We evaluate MSVA, PGL-SUM, and GTN on the SumMe and TVSum benchmarks, reporting average *frame-level* F1 scores in Table 4.2 and *segment-level* F1 scores in Table 4.3. Each backbone model is trained under four different variants:

Baseline MSE loss on the original annotations,

+RNC Baseline augmented with the RNC loss,

+ACN Baseline guided with the ACN. Typically ACN is the convolutional based ACN, we also show results using the MLP based ACN as ACN_{MLP}.

+ACN+RNC ACN and RNC applied jointly.

Table 4.2: Frame-level average F1 (%) against all annotators (**A**) on SumMe and TVSum. Best result for each backbone–dataset pair is in bold.

Model	**Variant**	**SumMe**	**TVSum**
MSVA	Baseline	21.15	60.29
	+RNC	23.61	60.33
	+ACN	22.75	61.05
	+ACN$_{MLP}$	22.35	**61.23**
	+ACN+RNC	**24.20**	61.08
GTN	Baseline	22.99	57.14
	+RNC	23.46	57.88
	+ACN	23.27	59.09
	+ACN$_{MLP}$	22.89	56.28
	+ACN+RNC	**24.17**	**59.38**
PGL	Baseline	23.47	59.71
	+RNC	23.62	59.45
	+ACN	24.97	61.11
	+ACN$_{MLP}$	23.93	60.27
	+ACN+RNC	**25.35**	**61.25**

Frame-level performance

Introducing RNC alone already boosts performance on the noisier SumMe dataset: MSVA improves by +2.46 F1 (21.15 → 23.61), GTN by +0.47, and PGL-SUM by +0.15. ACN, on the other hand, yields larger gains on TVSum, where annotator agreement is higher: GTN improves by +1.95 F1 (57.14 → 59.09) and PGL-SUM by +1.40. Combining both mechanisms yields the strongest performance across both datasets: MSVA reaches 24.20 F1 on SumMe (+3.05 over Baseline), while PGL-SUM achieves a leading 61.25 F1 on TVSum. These results confirm that (i) denoising labels via consensus estimation and (ii) regularizing embeddings to reflect

Table 4.3: Segment-level average F1 (%) against AC ground truth on SumMe and TVSum.

Model	**Variant**	**SumMe**	**TVSum**
MSVA	Baseline	35.91	80.32
	+RNC	40.63	79.83
	+ACN	39.67	80.97
	+ACN$_{MLP}$	39.08	**81.49**
	+ACN+RNC	**42.34**	81.10
GTN	Baseline	41.55	75.30
	+RNC	**43.78**	76.08
	+ACN	41.92	77.54
	+ACN$_{MLP}$	41.61	73.20
	+ACN+RNC	42.77	**79.06**
PGL	Baseline	39.09	78.04
	+RNC	40.43	78.14
	+ACN	42.53	81.45
	+ACN$_{MLP}$	40.06	80.6
	+ACN+RNC	**44.02**	**81.47**

that consensus are complementary strategies.

Segment-level performance

Using coarser temporal units increases absolute F1 scores across all methods, though the performance trends remain consistent with the frame-level results. On SumMe, RNC alone yields substantial improvements for MSVA (+4.72 F1) and GTN (+2.23 F1); ACN contributes additional gains, and the full **ACN+RNC** combination sets new benchmarks: 42.34 F1 for MSVA, 42.77 F1 for GTN, and a dataset-best 44.02 F1 for PGL-SUM. On TVSum, the improvements are more modest but still consistent,

with PGL-SUM+ACN+RNC achieving the highest score (81.47 F1), closely followed by MSVA at 81.10.

Backbone comparison

Before applying consensus-aware learning, GTN's graph-transformer is strongest on SumMe, whereas PGL's dual-path attention favors TVSum. After adding ACN and RNC, the gaps narrow or even reverse, with PGL+ACN+RNC emerging as the top model on both datasets at both granularities. This suggests that careful treatment of annotation noise and embedding structure can outweigh architectural differences in the underlying summarization networks.

ACN variant comparison

We additionally compare the performance of MLP-based and convolutional ACN variants. Across the board, the convolutional ACN tends to outperform the MLP-based version, especially on SumMe. For example, PGL achieves 42.53 F1 with convolutional ACN versus 40.06 F1 with MLP-ACN on segment-level SumMe. This is likely because the convolutional variant better captures local temporal dependencies in the annotation structure—an important characteristic for video summarization, where annotations are densely aligned in time. On TVSum, however, MLP-ACN can occasionally match or slightly outperform the convolutional variant, such as in the MSVA results. Overall, the convolutional ACN demonstrates stronger consistency across architectures and datasets, making it more suitable for video summarization tasks.

In practice, ACN adds negligible inference overhead, and RNC incurs only a modest cost during training. Together, they form an inexpensive, general plug-in that consistently improves video summarization backbones by first aligning predictions with a learned human consensus and then enforcing that consensus within the latent space.

Qualitative results

Figure 4.2 compares three scoring signals on two TVSum videos (top rows) and two SumMe videos (bottom rows). The top (green) curve (**AA**) represents the frame-level importance obtained by averaging the raw human scores in each dataset. The middle (red) curve (**LAC**) is produced by the ACN module and reflects the learned annotation consensus, which refines AA by amplifying frames that consistently attract human attention and suppressing isolated fluctuations. The bottom (blue) curve (**PGL-SUM+ACN+RNC**) is the soft importance score generated by the full summarization model after joint training with the LAC signal. A 0/1-knapsack optimizer [Song et al., 2015] converts this soft curve into the dashed hard intervals that constitute the final summary; each thumbnail corresponding to a frame within a hard interval is outlined with a green border.

In the tyre-replacement tutorial *(98MoyGZKHXc)*, annotators agree on shots that show the wheel, the tools, and the verbal instructions. LAC marks these moments, and the **PGL-SUM+ACN+RNC** network score aligns with them, so the resulting summary retains only the relevant instructional segments. The second TVSum clip *(sTEELN-vY30)* depicts an accident in a rail yard; both LAC and the network emphasize the collision and rescue scenes, and the knapsack step selects exactly those intervals.

SumMe exhibits lower annotator consistency. In *Bus_in_Rock_Tunnel*, two pronounced peaks correspond to the bus emerging from the rock tunnel, while several smaller peaks capture less critical commentary. LAC lowers the minor spikes, enabling the network to focus on the release event. In *Car_railcrossing*, the crash produces the highest peak in all three signals; the network additionally highlights the rescue activity that follows. The knapsack layer, therefore, includes both the collision and the immediate assistance sequence.

Across all four videos, the learned annotation consensus suppresses noise (for example, repeated rock close-ups) and shifts the network's attention toward semantically important events, leading to shorter and more informative summaries. The uniform green borders confirm that every hard interval is represented by an accom-

panying thumbnail.

4.2 Continuous Emotion Recognition Evaluations

4.2.1 Datasets

We utilized the RECOLA, COGNIMUSE, and JESTKOD datasets, each of which provides continuous annotations for valence and arousal by multiple annotators, making them well-suited for studying annotation consensus in continuous emotion recognition.

The **RECOLA** dataset consists of 9.5 hours of multimodal recordings from 46 French-speaking participants and includes annotations for arousal and valence made by three males and three females, providing a balanced perspective on affective states. This dataset is widely used in the research community, particularly highlighted in the AVEC 2016 challenge that featured 9 subjects for training and 9 for validation. For our validation, we adopted the gold standard annotations from AVEC 2018 to ensure comparability and consistency with previous studies. RECOLA also offers a balanced gender distribution and continuous annotations for valence and arousal at 25 Hz frequency.

The **COGNIMUSE** dataset features 30-minute segments from 7 Academy Award-winning films, annotated for both intended (used as the ground truth) and experienced emotions (used as input to the ACN), providing a unique dataset for evaluating models on both predicted and perceived emotional states. Each of the 7 participants attended two annotation sessions, enhancing the dataset with varied emotional insights. Annotations in COGNIMUSE are similarly detailed, provided at 25 Hz frequency, which facilitates a fine-grained analysis of emotional trajectories over time.

The **JESTKOD** dataset contains Turkish-language dyadic interactions focused on agreement and disagreement discussions. It comprises recordings from 10 participants (4 female, 6 male), aged 20 to 25, across 5 sessions. Each session includes 19–23 video clips lasting 2–4 minutes, with participants discussing predefined top-

ics. The total recording time is 259 minutes. Unlike scripted datasets, JESTKOD captures spontaneous emotional expression in real-time social interactions, offering an important testbed for consensus modeling in conversational affect analysis.

4.2.2 Training

Our training protocol was designed to evaluate the performance differences between the baseline W2V-CER and the proposed W2V-CER-ACN models. We conducted training over 15 epochs, employing a learning rate of 5×10^{-4} and a batch size of 32.

In terms of temporal resolution: Data from **RECOLA** was segmented into 3-second windows with a 0.4-second shift, allowing us to capture transient emotional expressions effectively. For **COGNIMUSE**, we used 5-second windows with a 3-second shift to accommodate the longer-duration emotional turns typical of cinematic content. For **JESTKOD**, we adopted 5-second windows with a 5-second shift, reflecting the conversational nature of the data and ensuring clear segmentation of agreement/disagreement dynamics.

To ensure rigorous and comprehensive testing: - We implemented a 7-fold cross-validation strategy on COGNIMUSE, testing the model on each film individually while training on the remaining films. - For JESTKOD, we employed a 5-fold cross-validation scheme, partitioning the dyadic sessions to ensure speaker-independent validation and robust performance assessment.

During the joint training of our networks, we set both hyperparameters α and β to 0.5 to equally balance the importance of aligning predictions with the ground truth annotations and achieving consensus among the annotators' experienced emotions.

4.2.3 CER Results

Our experimental evaluations on the RECOLA, COGNIMUSE, and Jestkod datasets aimed to assess the impact of integrating the Annotators Consensus Network (ACN) on model performance, with a specific focus on the prediction accuracy of arousal and valence. We measured performance using the Concordance Correlation Coeffi-

cient (CCC), where higher values indicate better alignment with the ground truth annotations. In contrast to the video summarization task, the default ACN used in the W2V-CER and W2V-CER-ACN models is the MLP-based variant.

Table 4.4: CCC performance of continuous emotion recognition on the COGNIMUSE dataset for seven individual films and the average across all films.

Task	Model	DEP	BMI	CHI	CRA	FNE	GLA	LOR	**Average**
Valence	W2V-CER	**0.128**	0.155	0.128	-0.064	0.237	0.169	**0.343**	0.157
	W2V-CER-ACN	0.109	**0.425**	**0.326**	**-0.056**	**0.296**	**0.193**	0.257	**0.221**
Arousal	W2V-CER	**0.467**	**0.557**	0.601	0.654	0.545	0.510	**0.608**	0.563
	W2V-CER-ACN	0.426	0.550	**0.635**	**0.713**	**0.558**	**0.569**	0.597	**0.579**

The results from the COGNIMUSE dataset, as shown in Table 4.4, provide valuable insights on the performance improvements of the ACN. For valence prediction, the W2V-CER-ACN model achieved an average CCC score of **0.221**, outperforming the baseline W2V-CER model, which had an average CCC score of 0.157. The W2V-CER-ACN model showed significant improvements in predicting valence for films like BMI, CHI, CRA, FNE, and GLA, highlighting the ACN's effectiveness in scenarios where emotional variance is subtle and complex.

However, it is important to note that valence prediction remains challenging, especially for certain films where negative CCC values were observed, indicating poor performance. This highlights the challenge of valence prediction in scenarios involving long and unrelated movies, where emotional continuity varies significantly, making it difficult for the model to predict intended emotional states.

For arousal prediction, the W2V-CER-ACN model outperformed the baseline, achieving an average CCC score of **0.579** compared to 0.563 for the baseline W2V-CER model. The ACN significantly enhanced the model's sensitivity to dynamic changes in arousal, particularly in films such as CHI, CRA, FNE, and GLA.

The CCC score results on the RECOLA dataset, presented in Table 4.5, show a clear improvement when ACN is used. For valence prediction, W2V-CER-ACN

Table 4.5: CCC performance of continuous emotion recognition on the RECOLA dataset for individually and jointly trained valence and arousal prediction tasks.

Task	W2V-CER	W2V-CER-ACN	W2V-CER-ACN$_{CNN}$
Valence	0.391	0.437	**0.485**
Arousal	0.651	**0.666**	0.652
Valence & Arousal	0.438/0.645	0.482/**0.657**	**0.488**/0.624

improves the CCC from 0.391 to **0.437**, and the CNN-based variant reaches **0.485**. For arousal, W2V-CER-ACN achieves the highest score of **0.666**. When jointly training both tasks, both variants of ACN provide consistent improvements over the baseline, with W2V-CER-ACN reaching **0.657** for arousal and the CNN variant achieving **0.488** for valence.

Table 4.6: CCC performance of continuous emotion recognition on the Jestkod dataset for valence and arousal prediction tasks.

Task	W2V-CER	W2V-CER-ACN	W2V-CER-ACN$_{CNN}$
Valence	0.254	**0.271**	0.252
Arousal	0.440	0.447	**0.466**

Table 4.6 presents the CCC performance of the CER models on the Jestkod dataset. For valence, W2V-CER-ACN achieves the best score of **0.271**, outperforming both the baseline and CNN variant. For arousal, the CNN-based ACN performs best with **0.466**, while W2V-CER-ACN scores 0.447 and the baseline 0.440. These results highlight the effectiveness of ACN in enhancing the model's alignment with consensus annotations, particularly in settings with rich temporal fluctuations and annotation variability.

ACN variant comparison

As demonstrated in Tables 4.5 and 4.6, the convolutional variant (W2V-CER-ACN$_{\text{CNN}}$) shows competitive performance, but does not consistently outperform the MLP-based ACN. Specifically, on the RECOLA dataset, the CNN variant achieves a higher CCC for valence (**0.485** vs. 0.437), while on the Jestkod dataset it slightly underperforms for valence but surpasses the MLP version for arousal (**0.466** vs. 0.447).

These findings suggest that while convolutional ACNs may provide advantages in capturing local temporal dynamics, the simpler MLP-based ACN is better suited for the continuous emotion recognition setting where annotation variability across annotators dominates over temporal-local structure. The MLP variant remains a strong and computationally efficient choice for audio-based emotion recognition tasks involving sequential but semantically diffuse labels.

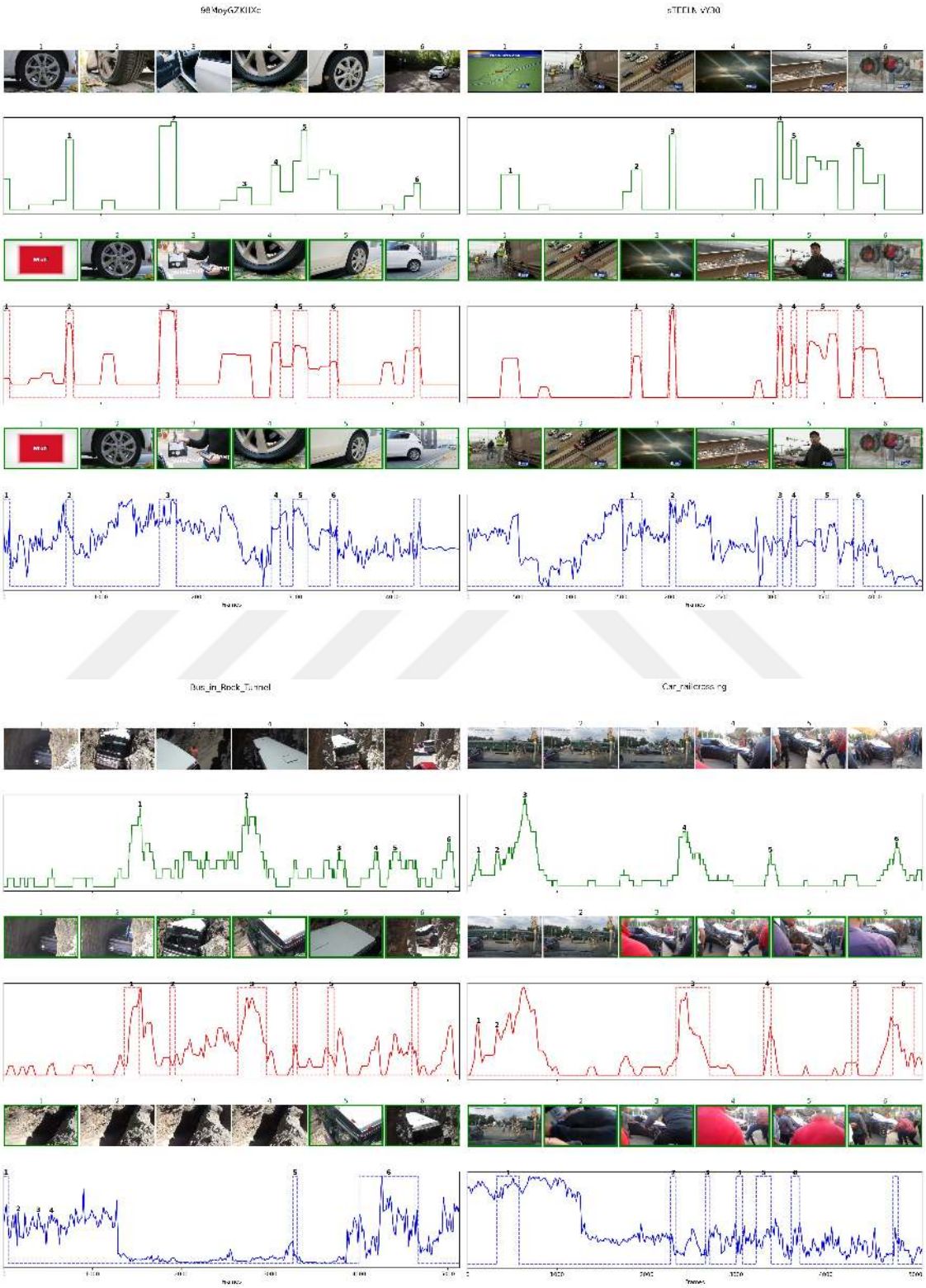

Figure 4.2: Qualitative comparison of frame-level importance (AA) Average frame scores, (LAC)learned annotation consensus, (PGL-SUM+ACN+RNC) output of the video summarization model. Dashed boxes show the hard intervals selected by the knapsack. A green border on a thumbnail indicates the summary intervals.

Chapter 5

CONCLUSION

This thesis presented the *Annotation Consensus Network* (ACN), a general-purpose neural module designed to synthesize multiple subjective annotations into a coherent, learnable consensus. We demonstrated the versatility and effectiveness of ACN in two distinct domains: video summarization and continuous emotion recognition (CER). Across both applications, ACN was shown to improve supervision quality, stabilize model training, and deliver more reliable predictions by explicitly modeling annotator agreement.

5.1 Key Findings

- ACN significantly improves model performance in both video summarization and CER tasks by reducing label variance and refining supervision quality.
- The benefit of ACN is most pronounced in datasets with high annotator disagreement (e.g., SumMe for summarization, and COGNIMUSE for emotion recognition), confirming its utility in handling noisy annotation environments.
- The RNC loss complements consensus learning by enforcing consistency in the embedding space, further improving model generalization.
- ACN introduces no additional inference-time cost, making it a modular and scalable addition to existing architectures.
- Joint modeling of arousal and valence in CER yielded improved results in some datasets (e.g., Recola), although its effectiveness varies with emotional content diversity.

5.2 Limitations

While the ACN framework is flexible and effective, it has certain limitations:

- **Training-Time Requirements:** ACN relies on access to multiple annotations per sample during training. In domains where only single labels are available, its utility is limited unless synthetic or proxy annotations can be generated.

- **Domain-Specific Design Choices:** Architectural choices such as convolutional vs. MLP layers in the ACN are domain- and input-dependent. These components may require careful tuning across tasks.

- **Dataset Sensitivity:** In affective computing, performance gains varied across datasets. For example, valence prediction on the COGNIMUSE dataset showed weaker improvements, indicating that model success may depend on the richness and consistency of the input data.

5.3 Future Work

Building on the findings of this thesis, several directions are worth pursuing:

- **Multimodal Consensus Learning:** Extending ACN to simultaneously process multimodal annotations (e.g., audio, video, and text) could enhance generalization in emotion recognition and other perception-based tasks.

- **Real-Time Applications:** Optimizing ACN for streaming or low-latency applications such as live emotion monitoring or adaptive summarization in media players could enable interactive deployments.

- **Generalization to Other Domains:** The consensus modeling approach is applicable to tasks like medical diagnosis (multi-expert input), collaborative decision making, or crowd-sourced labeling. Exploring these domains would validate the broader utility of ACN.

5.4 Closing Remarks

The Annotation Consensus Network offers a principled solution to the long-standing challenge of learning from subjective human annotations. Its effectiveness across distinct domains shows that learning consensus representations can be both a practical and powerful strategy. We hope that this work not only improves current systems for summarization and affect prediction but also inspires broader applications of consensus learning in machine learning systems that seek to understand, reflect, or reason about human input.

BIBLIOGRAPHY

[Apostolidis et al., 2020] Apostolidis, E., Adamantidou, E., Metsai, A. I., Mezaris, V., and Patras, I. (2020). AC-SUM-GAN: Connecting actor-critic and generative adversarial networks for unsupervised video summarization. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(8):3278–3292.

[Apostolidis et al., 2021a] Apostolidis, E., Adamantidou, E., Metsai, A. I., Mezaris, V., and Patras, I. (2021a). Video summarization using deep neural networks: A survey. *Proceedings of the IEEE*, 109(11):1838–1863.

[Apostolidis et al., 2021b] Apostolidis, E., Balaouras, G., Mezaris, V., and Patras, I. (2021b). Combining global and local attention with positional encoding for video summarization. In *2021 IEEE International Symposium on Multimedia (ISM)*, pages 226–234.

[Arnab et al., 2021] Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., and Schmid, C. (2021). Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6836–6846.

[Bojanowski et al., 2014] Bojanowski, P., Lajugie, R., Bach, F., Laptev, I., Ponce, J., Schmid, C., and Sivic, J. (2014). Weakly supervised action labeling in videos under ordering constraints. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 628–643. Springer.

[Booth et al., 2018] Booth, B. M., Mundnich, K., and Narayanan, S. S. (2018). A novel method for human bias correction of continuous-time annotations. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3091–3095.

[Booth and Narayanan, 2020] Booth, B. M. and Narayanan, S. S. (2020). Fifty shades of green: Towards a robust measure of inter-annotator agreement for continuous signals. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 204–212.

[Booth and Narayanan, 2024] Booth, B. M. and Narayanan, S. S. (2024). People make mistakes: Obtaining accurate ground truth from continuous annotations of subjective constructs. *Behavior Research Methods*, pages 1–17.

[Bozkurt et al., 2015] Bozkurt, E., Khaki, H., Keçeci, S., Türker, B. B., Yemez, Y., and Erzin, E. (2015). Jestkod database: Dyadic interaction analysis. In *2015 23nd Signal Processing and Communications Applications Conference (SIU)*, pages 1374–1377.

[Calvo and D'Mello, 2010] Calvo, R. A. and D'Mello, S. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on affective computing*, 1(1):18–37.

[Chasanis et al., 2008] Chasanis, V., Likas, A., and Galatsanos, N. (2008). Efficient video shot summarization using an enhanced spectral clustering approach. In *Artificial Neural Networks-ICANN 2008: 18th International Conference, Prague, Czech Republic, September 3-6, 2008, Proceedings, Part I 18*, pages 847–856. Springer.

[Cheplygina and Pluim, 2018] Cheplygina, V. and Pluim, J. P. (2018). Crowd disagreement about medical images is informative. In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis: 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3*, pages 105–111. Springer.

[De Avila et al., 2011] De Avila, S. E. F., Lopes, A. P. B., da Luz Jr, A., and de Albuquerque Araújo, A. (2011). VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method. *Pattern Recognition Letters*, 32(1):56–68.

[Ejaz et al., 2014] Ejaz, N., Mehmood, I., and Baik, S. W. (2014). Feature aggregation based visual attention model for video summarization. *Computers & Electrical Engineering*, 40(3):993–1005.

[Ghauri et al., 2021] Ghauri, J. A., Hakimov, S., and Ewerth, R. (2021). Supervised video summarization via multiple feature sets with parallel attention. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6s.

[Giannakopoulos et al., 2017] Giannakopoulos, T., Makris, A., Kosmopoulos, D., Avrithis, Y., and Theodoridis, S. (2017). Cognimuse: A multimodal video database annotated with saliency, events, semantics and emotion with application to summarization. *EURASIP Journal on Image and Video Processing*, 2017(1):1–23.

[Gong et al., 2014] Gong, B., Chao, W.-L., Grauman, K., and Sha, F. (2014). Diverse sequential subset selection for supervised video summarization. *Advances in Neural Information Processing Systems*, 27.

[Grimm et al., 2007] Grimm, M., Kroschel, K., Mower, E., and Narayanan, S. (2007). Primitives-based evaluation and estimation of emotions in speech. *Speech communication*, 49(10-11):787–800.

[Gygli et al., 2014] Gygli, M., Grabner, H., Riemenschneider, H., and Van Gool, L. (2014). Creating summaries from user videos. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13*, pages 505–520. Springer.

[Hayat et al., 2021] Hayat, H., Ventura, C., and Lapedriza, A. (2021). Recognizing emotions evoked by movies using multitask learning. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE.

[Hoai and Zisserman, 2015] Hoai, M. and Zisserman, A. (2015). Improving human action recognition using score distribution and ranking. In *Computer Vision–ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part V 12*, pages 3–20. Springer.

[Ji et al., 2019] Ji, Z., Xiong, K., Pang, Y., and Li, X. (2019). Video summarization with attention-based encoder–decoder networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(6):1709–1717.

[Kadam et al., 2022] Kadam, P., Vora, D., Mishra, S., Patil, S., Kotecha, K., Abraham, A., and Gabralla, L. A. (2022). Recent challenges and opportunities in video summarization with machine learning algorithms. *IEEE Access*, 10:122762–122785.

[Khetan et al., 2017] Khetan, A., Lipton, Z. C., and Anandkumar, A. (2017). Learning from noisy singly-labeled data. *arXiv preprint arXiv:1712.04577*.

[Köprü and Erzin, 2023] Köprü, B. and Erzin, E. (2023). Use of affective visual information for summarization of human-centric videos. *IEEE Transactions on Affective Computing*, 14(4):3135–3148.

[Liu et al., 2020] Liu, Y.-T., Li, Y.-J., and Wang, Y.-C. F. (2020). Transforming multi-concept attention into video summarization. In *Proceedings of the Asian Conference on Computer Vision*.

[Lopes et al., 2017] Lopes, P., Yannakakis, G. N., and Liapis, A. (2017). Ranktrace: Relative and unbounded affect annotation. In *2017 Seventh International Confer-*

ence on Affective Computing and Intelligent Interaction (ACII), pages 158–163. IEEE.

[Lu and Grauman, 2013] Lu, Z. and Grauman, K. (2013). Story-driven summarization for egocentric video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2714–2721.

[Mahmoud et al., 2013] Mahmoud, K. M., Ghanem, N. M., and Ismail, M. A. (2013). Unsupervised video summarization via dynamic modeling-based hierarchical clustering. In *2013 12th International Conference on Machine Learning and Applications*, volume 2, pages 303–308. IEEE.

[Mariooryad and Busso, 2014] Mariooryad, S. and Busso, C. (2014). Correcting time-continuous emotional labels by modeling the reaction lag of evaluators. *IEEE Transactions on Affective Computing*, 6(2):97–108.

[Metallinou and Narayanan, 2013] Metallinou, A. and Narayanan, S. (2013). Annotation and processing of continuous emotional attributes: Challenges and opportunities. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8. IEEE.

[Pan et al., 2020] Pan, Z., Luo, Z., Yang, J., and Li, H. (2020). Multi-modal attention for speech emotion recognition.

[Picard, 1997] Picard, R. W. (1997). *Affective computing*. MIT press.

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

[Ringeval et al., 2015] Ringeval, F., Schuller, B., Valstar, M., Cowie, R., and Pantic, M. (2015). Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 193–199. IEEE.

[Ringeval et al., 2013] Ringeval, F., Sonderegger, A., Sauer, J., and Lalanne, D. (2013). Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–8. IEEE.

[Schuller, 2018] Schuller, B. W. (2018). Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Communications of the ACM*, 61(5):90–99.

[Sheshadri and Lease, 2013] Sheshadri, A. and Lease, M. (2013). Square: A benchmark for research on computing crowd consensus. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 1, pages 156–164.

[Shi et al., 2021] Shi, Y., Huang, Z., Feng, S., Zhong, H., Wang, W., and Sun, Y. (2021). Masked label prediction: Unified message passing model for semi-supervised classification.

[Shoer and Erzin, 2025a] Shoer, I. and Erzin, E. (2025a). Improving video summarization through learned annotation consensus. *IEEE Transactions on Multimedia*.

[Shoer and Erzin, 2025b] Shoer, I. and Erzin, E. (2025b). Learning annotation consensus for continuous emotion recognition. *IEEE Transactions on Affective Computing*.

[Shoer et al., 2023] Shoer, I., Kopru, B., and Erzin, E. (2023). Role of audio in audio-visual video summarization. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

[Song et al., 2015] Song, Y., Vallmitjana, J., Stent, A., and Jaimes, A. (2015). TVSum: Summarizing web videos using titles. In *CVPR 2015, IEEE Conference on Computer Vision and Pattern Recognition*, pages 5179–5187.

[Szegedy et al., 2015] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions . In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, Los Alamitos, CA, USA. IEEE Computer Society.

[Tanno et al., 2019] Tanno, R., Saeedi, A., Sankaranarayanan, S., Alexander, D. C., and Silberman, N. (2019). Learning from noisy labels by regularized estimation of annotator confusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11244–11253.

[Triantafyllopoulos et al., 2023] Triantafyllopoulos, A., Reichel, U., Liu, S., Huber, S., Eyben, F., and Schuller, B. W. (2023). Multistage linguistic conditioning of convolutional layers for speech emotion recognition. *Frontiers in Computer Science*, 5:1072479.

[Tzirakis et al., 2017] Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B. W., and Zafeiriou, S. (2017). End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of selected topics in signal processing*, 11(8):1301–1309.

[Wagner et al., 2023] Wagner, J., Triantafyllopoulos, A., Wierstorf, H., Schmitt, M., Burkhardt, F., Eyben, F., and Schuller, B. W. (2023). Dawn of the transformer era in speech emotion recognition: closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10745–10759.

[Wu et al., 2020a] Wu, J., Zhong, S.-h., and Liu, Y. (2020a). Dynamic graph convolutional network for multi-video summarization. *Pattern Recognition*, 107:107382.

[Wu et al., 2020b] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Philip, S. Y. (2020b). A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):4–24.

[Xiao et al., 2020] Xiao, S., Zhao, Z., Zhang, Z., Yan, X., and Yang, M. (2020). Convolutional hierarchical attention network for query-focused video summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12426–12433.

[Yannakakis et al., 2018] Yannakakis, G. N., Cowie, R., and Busso, C. (2018). The ordinal nature of emotions: An emerging approach. *IEEE Transactions on Affective Computing*, 12(1):16–35.

[Zha et al., 2023] Zha, K., Cao, P., Son, J., Yang, Y., and Katabi, D. (2023). Rank-n-contrast: Learning continuous representations for regression. In *Thirty-seventh Conference on Neural Information Processing Systems*.

[Zhang et al., 2016] Zhang, K., Chao, W.-L., Sha, F., and Grauman, K. (2016). Video summarization with long short-term memory. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pages 766–782. Springer.

[Zhao et al., 2021a] Zhao, B., Gong, M., and Li, X. (2021a). Audiovisual video summarization. *IEEE Transactions on Neural Networks and Learning Systems*.

[Zhao et al., 2021b] Zhao, B., Li, H., Lu, X., and Li, X. (2021b). Reconstructive sequence-graph network for video summarization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2793–2801.

[Zhu et al., 2022] Zhu, W., Han, Y., Lu, J., and Zhou, J. (2022). Relational reasoning over spatial-temporal graphs for video summarization. *IEEE Transactions on Image Processing*, 31:3017–3031.

Appendix A

A.1 Video Summarization

In this section, we show the training parameters that we used in our experiments and results.

Model	**Variant**	$\boldsymbol{\alpha}^*_{\text{SumMe}}$	$\boldsymbol{\beta}^*_{\text{SumMe}}$	$\boldsymbol{\alpha}^*_{\text{TVSum}}$	$\boldsymbol{\beta}^*_{\text{TVSum}}$
MSVA	+ACN	0.3	–	0.9	–
	+ACN+RNC	0.3	0.4	0.9	0.3
GTN	+ACN	0.6	–	0.9	–
	+ACN+RNC	0.8	0.9	0.8	0.6
PGL	+ACN	0.1	–	0.4	–
	+ACN+RNC	0.1	0.3	0.4	0.5

Table A.1: α and β values for video summarization experiments.

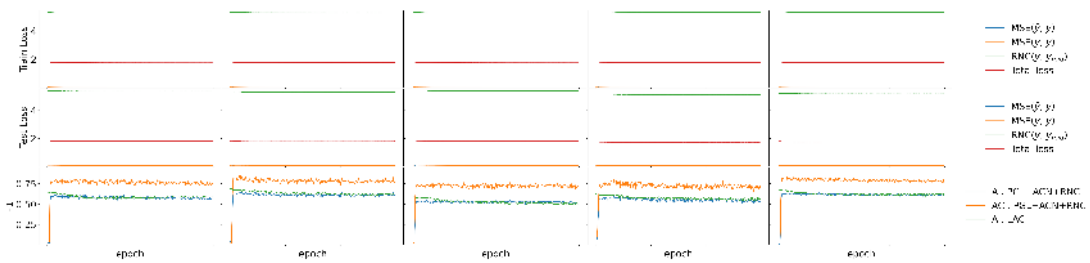

Figure A.1: Learning characterization with different loss function progress over epochs in training (top row), testing (middle row), and with F1 score performance against A at frame-level and AC at the segment-level, and A vs LAC at frame-level (bottom row) for 5 folds (columns) in TVSum

A.2 Continuous Emotion Recognition

session	speaker 1	speaker 2
1	m1	m2
2	f1	f2
3	f3	m3
4	m4	m5
5	f4	m6

Table A.2: Male (m) and female (f) speakers in each session of Jestkod dataset.

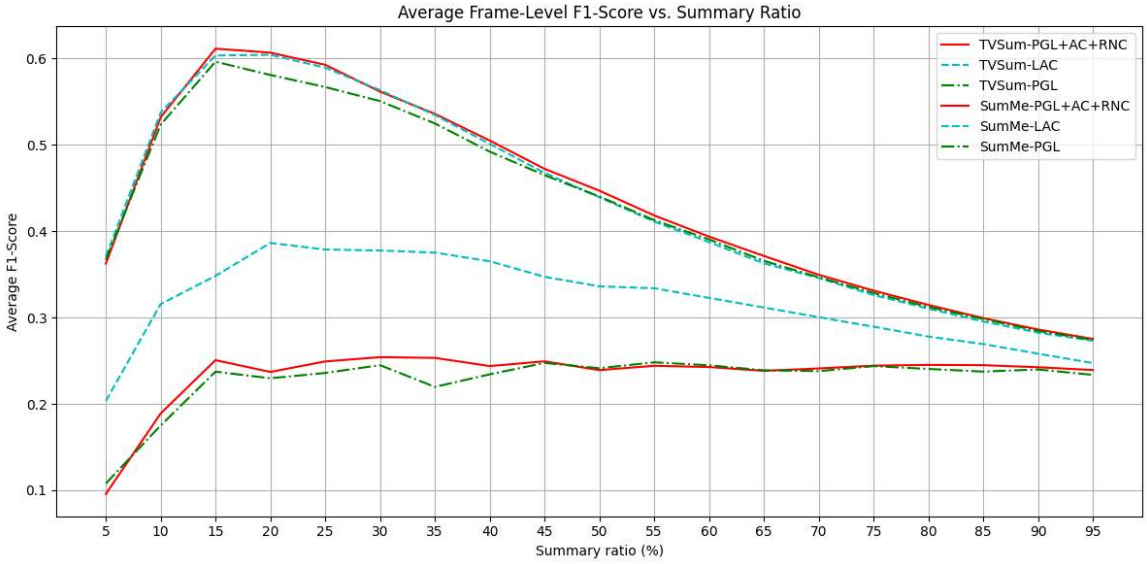

Figure A.2: Average frame-level F1 versus summary ratio (%) with respect to A; top three curves belong to TVSum.