Master's Thesis Presentation for the 143th Dissertation Evaluation

A Study on Robust Audio-Visual Fusion for Emotion Recognition based on Cross-Modal Learning under Noisy Conditions

교차 모달리티 학습 기반 노이즈에 강건한 오디오-비주얼 감정인식에 관한 연구

Seungyeon Jeong

jq3210@gmail.com

8th April, 2025

Contents

Introduction

Audio-Visual Fusion for Emotion Recognition

- The goal of this task through audio-visual multimodality, which is essential for emotion recognition in real-world applications, is useful in various fields such as driver fatigue assessment, healthcare, and marketing [1, 2].
- As emotion recognition technology is applied in marketing and healthcare, it faces challenges from noisy and uncontrolled videos, like face occlusion and varying conditions, making accurate analysis more difficult than in controlled settings [3].

Affwild2

SEWA

Recent emotion recognition datasets conducted in real-world situations that have context mismatch data(occlusion, noisy background music etc..)

VT-KFER

CK+

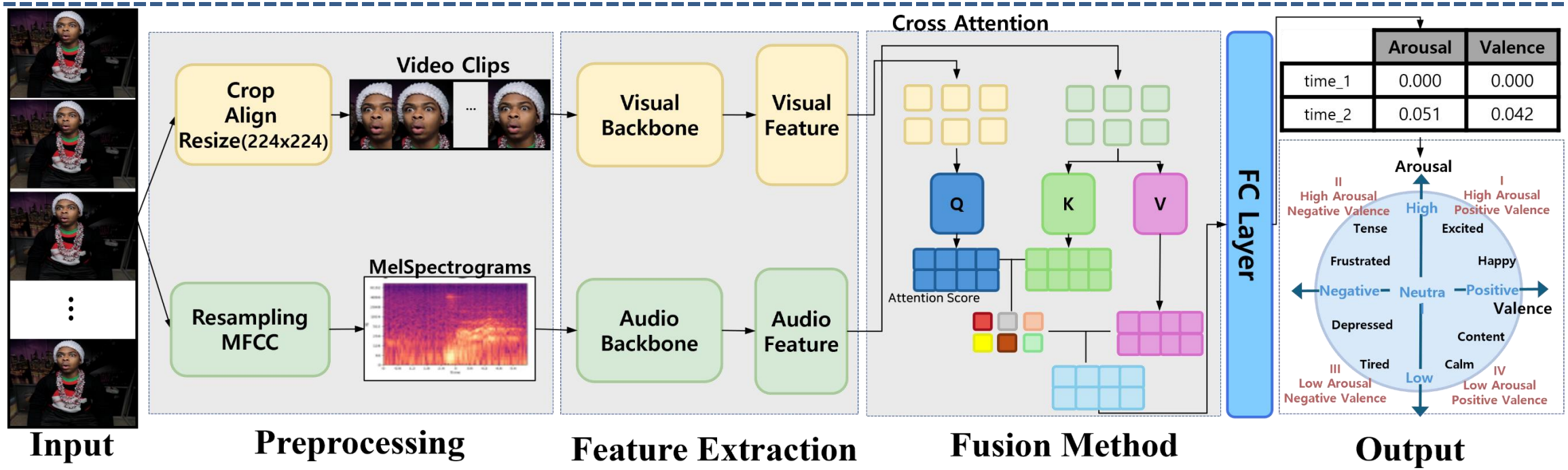
Traditional emotion recognition datasets conducted in a controlled environment.

[1] Praveen, R. Gnana, et al. "A joint cross-attention model for audio-visual fusion in dimensional emotion recognition." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.
[2] Praveen, R. Gnana, Eric Granger, and Patrick Cardinal. "Cross attentional audio-visual fusion for dimensional emotion recognition." *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. IEEE, 2021.
[3] Pei, Ercheng, Dongmei Jiang, and Hichem Sahli. "An efficient model-level fusion approach for continuous affect recognition from audiovisual signals." *Neurocomputing* 376 (2020): 42-53.

Introduction

Objective and Standard Procedure

- Audio-Visual Emotion Recognition involves predicting the valence-arousal state from real-world videos, producing frame-by-frame valence-arousal scores that human emotion states represented continuous data.
- The standard procedure for this task involves taking a video as input, splitting it into visual and audio data, processing each modality to extract features, fusing the features through fusion method, and predicting the valence and arousal states for each frame as output.



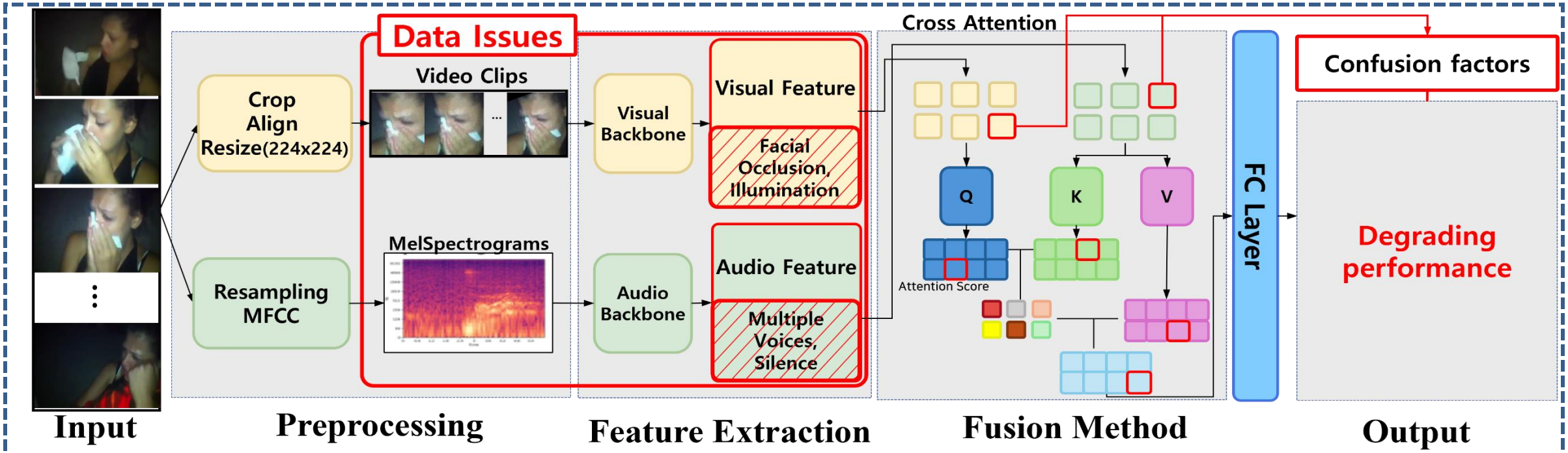
Standard Procedure of Audio-Visual Emotion Recognition

[5] Praveen, R. Gnana, Patrick Cardinal, and Eric Granger. "Audio–visual fusion for emotion recognition in the valence–arousal space using joint cross-attention." *IEEE Transactions on Biometrics, Behavior, and Identity Science* 5.3 (2023): 360-373.

Introduction

Problem and Strategy

- The problem of previous studies is that data issues such as not capturing facial expressions accurately or having mixed voices from multiple speakers can cause confusion in the model, degrading overall performance in real-world scenarios.
- The strategy of this study aims to design a deep neural network to be robust against confusion caused by context mismatch when fusing audio and visual information.

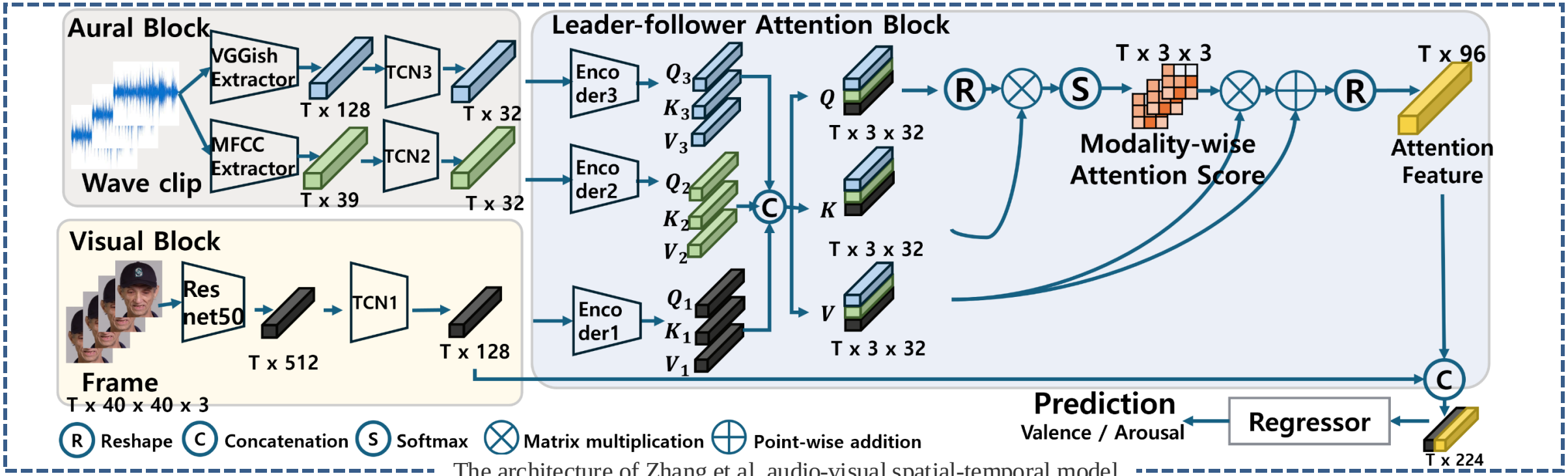


Problem of previous studies

Related Work

Continuous Emotion Recognition with Audio-Visual Leader-follower Attentive Fusion [4]

- This study proposes a spatiotemporal deep neural network using leader-follower attentive fusion for continuous audio-visual emotion recognition and emphasizes the prioritization of visual modality through cross-attentional mechanisms.
- This study suggests that emphasizing the visual modality can reduce issues where it does not contextually align with audio, but emphasis visual information through skip connections degrades performance due to data that fails to capture faces.



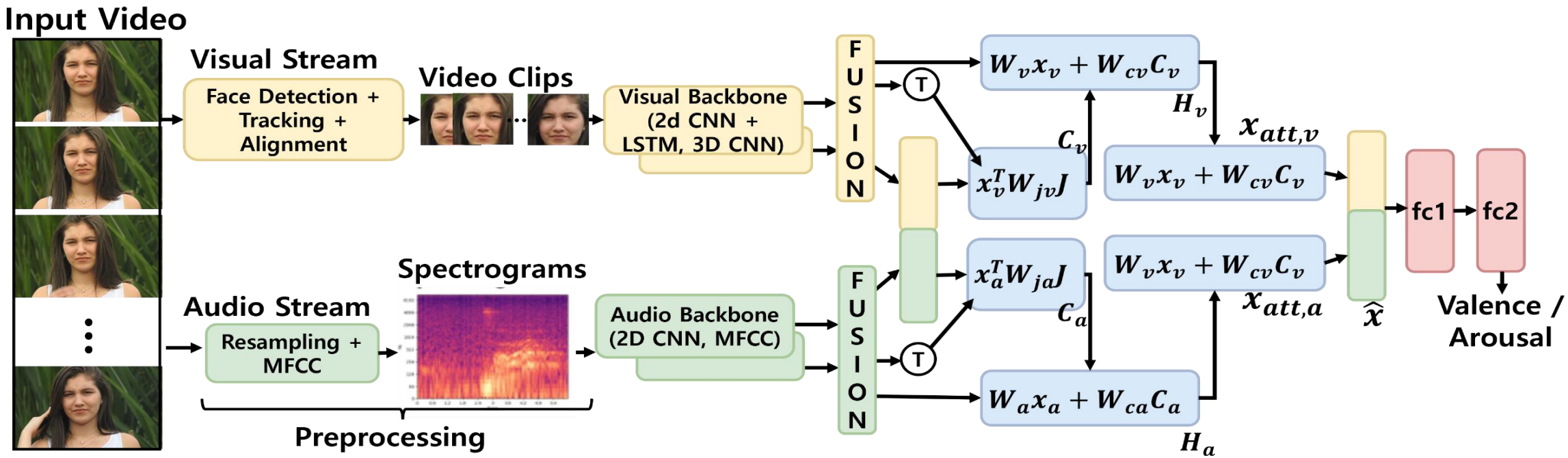
The architecture of Zhang et al. audio-visual spatial-temporal model

[4] Zhang, Su, et al. "Continuous emotion recognition with audio-visual leader-follower attentive fusion." *Proceedings of the IEEE/CVF international conference on computer vision*. 2021.

Related Work

Audio–Visual Fusion for Emotion Recognition in the Valence–Arousal Space Using Joint Cross-Attention [5]

- This study proposes a model using joint cross-attention mechanisms for emotion recognition in the valence-arousal space from audio-visual data, effectively leveraging inter-modal interactions to improve emotion recognition performance.
- This study suggests that joint representation through modalities integration is robust to contextual differences between the two modalities, but including contextually misaligned information in the joint representation degrades performance.



The architecture of Praveen et al. Joint cross-attention model for audio-visual fusion.

[5] Praveen, R. Gnana, Patrick Cardinal, and Eric Granger. "Audio–visual fusion for emotion recognition in the valence–arousal space using joint cross-attention." *IEEE Transactions on Biometrics, Behavior, and Identity Science* 5.3 (2023): 360-373.

Related Work

Common Drawback and Solution

- A common drawback of previous studies is that contextual mismatch is not sufficiently handled during independent feature learning in each modality, leading to contextual mismatches when combined and ultimately degrading performance.
- TCN captures local dependencies with causal and dilated convolutions, while LSTM captures long-term context through sequential processing, reducing mismatches and enhancing audio-visual correlations for improved performance.

Method	Publication	Summary	Drawback
Leader-Follower Fusion	IEEE/CVF	This study proposes a model using a leader-follower attention mechanism, prioritizing visual data as the leader and using audio data as the follower. The model shows improvement in emotion recognition performance on the Aff-Wild2 dataset.	Emphasizing the visual modality can reduce issues where it does not contextually align with audio
JCA	IEEE Transactions on Biometrics, Behavior, and Identity Science	This study proposes a joint cross-attention model that enhances emotion recognition by leveraging both inter-modal and intra-modal relationships, combining facial and vocal modalities to improve performance on the Aff-Wild2 datasets	Including contextually misaligned information in the joint representation degrades performance

Table 1. Comparison of methods between different models

Proposed Method

Notation and Definition

- Feature representations for video and audio modalities are extracted using pretrained backbone networks. These features are denoted as X_{visual} for video and X_{audio} for audio, each capturing distinct spatiotemporal patterns.
- The visual feature representation is $X_{visual} = \{x_v^1, x_v^2, \cdots, x_v^l\} \in R^{d_v \times L}$, while the audio feature representation is $X_{audio} = \{x_a^1, x_a^2, \cdots, x_a^l\} \in R^{d_a \times L}$. d_v and d_a represent the feature dimensions for video and audio, respectively, and L is the number of clips.

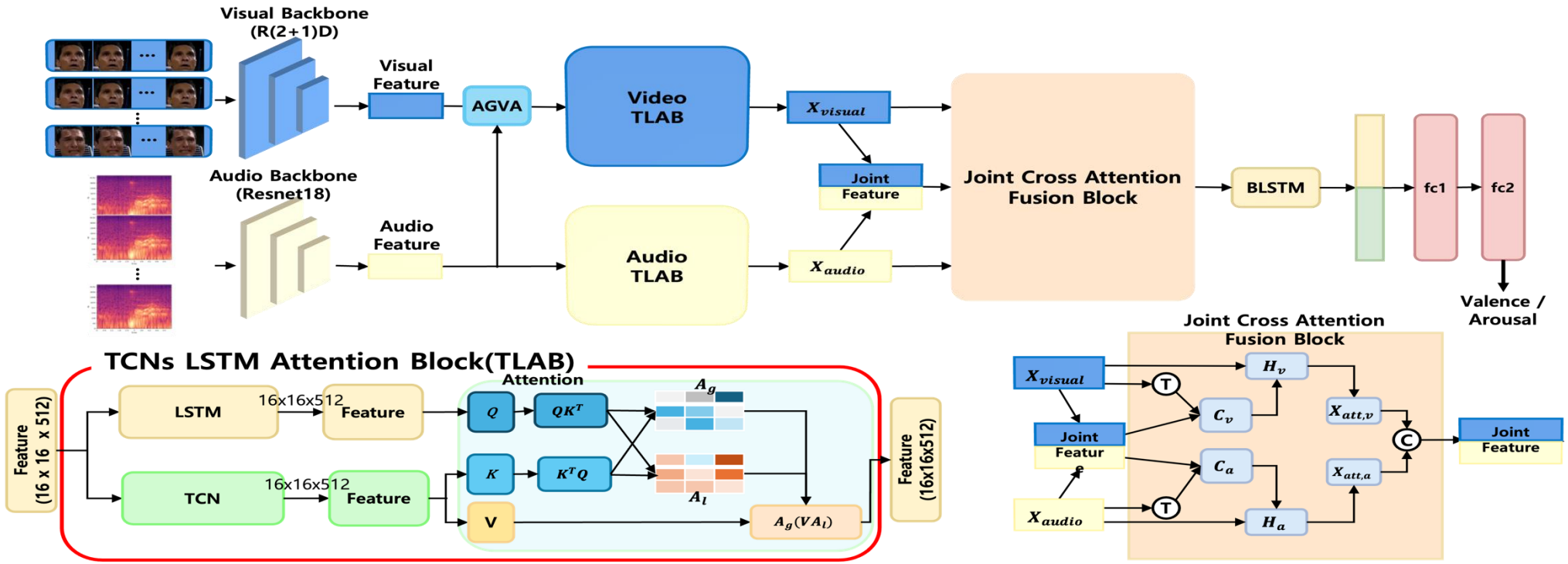
Notation	Description	Usage
X_{visual}, X_{audio}	Video feature, Audio feature	Video and audio features extracted from the backbone model
AGVA	Audio Guided Visual Attention	Enhances feature alignment by leveraging audio cues to refine visual representations
T_v, T_a, L_v, L_a	TCN, LSTM video, audio feature	Video and audio features processed through TCN, LSTM to capture local, global information
A_{G_v}, A_{G_a}	Global Attention Weight	Global attention weights, computed using LSTM-based Query and TCN-based Key
A_{L_v}, A_{L_a}	Local Attention Weight	Local attention weights, computed using TCN-based Key and LSTM-based Query
$TLAB_v, TLAB_a$	output TCN_LSTM_Attention_Block	Output of the TCN_LSTM_Attention_Block
J_{feat}	Joint feature	Concatenated feature representation of TLABs audio and video outputs
C_v, C_a	Joint Cross Attention Matrices	Used to learn intra-modal and inter-modal relationships within and between modalities
H_v, H_a	Attention Maps for Audio and Video	Computed by combining each modalities features with the joint correlation matrix to extract refined features
$x_{att,v}, x_{att,a}$	Audio-Visual Features	Refined feature vectors processed through the JCA block
BLSTM	Bi-LSTM	Captures sequential dependencies from fused audio-visual features bidirectionally
$\hat{X}$	Final Fused Representation	Final representation from JCA block, refined by BLSTM for sequential modeling

Table 2. Definitions of Key Notations

Proposed Method

Procedural Overview

- Video and audio features are extracted using R(2+1)D and ResNet18. AVGA refines spatial video features, while TLAB combines TCN for local dependencies and LSTM for long-term temporal modeling.
- The TLAB outputs from both modalities are concatenated and processed through the Joint Cross Attention Fusion Block (JCA). The fused features pass through BLSTM and fully connected layers to predict valence and arousal.

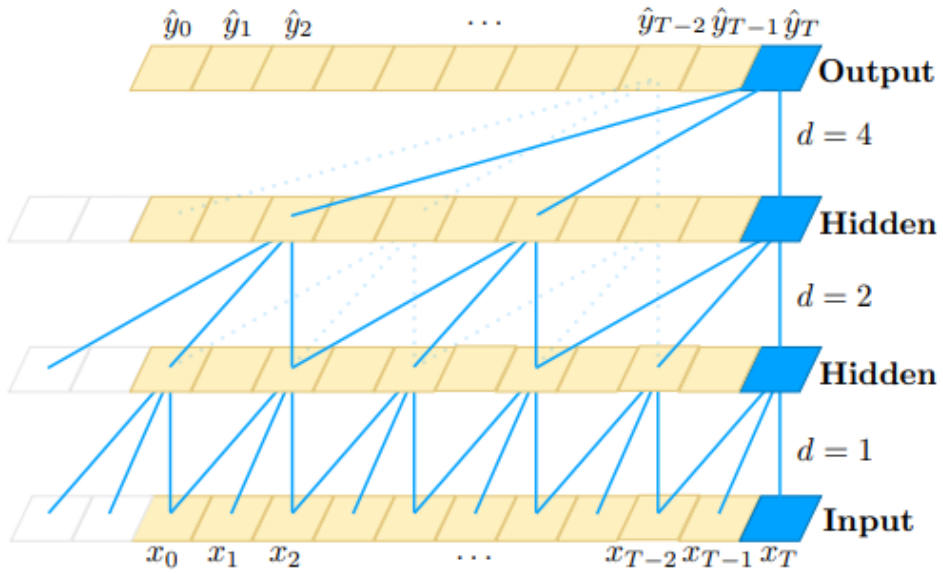


Proposed Method Architecture

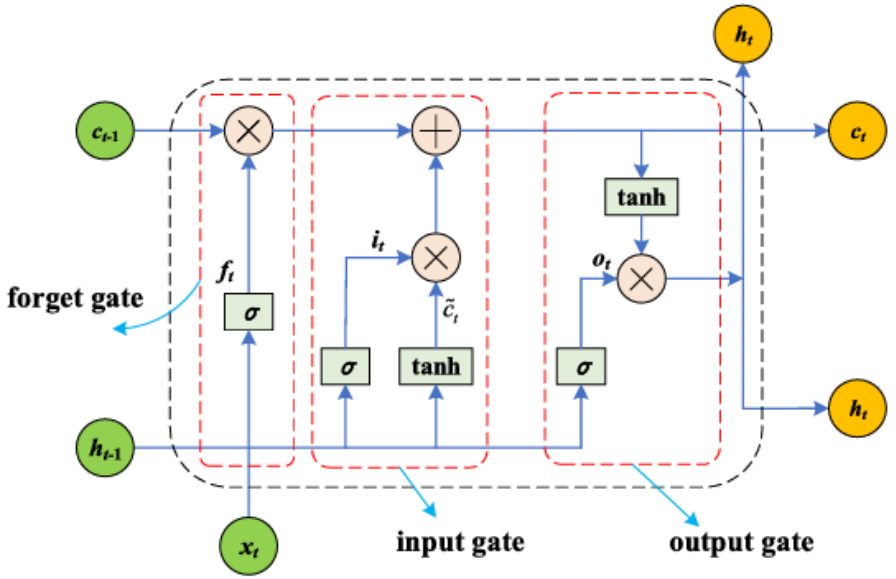
Proposed Method

Details

- Both TCN and LSTM capture long-term temporal information; TCN emphasizes localized context through causal and dilated convolutions, while LSTM preserves subtle sequential nuances via its recurrent processing, complementing each other [7].
- To leverage both strengths, an attention mechanism aligns local TCN-based patterns with global LSTM-based representations, ensuring a balanced integration of fine-grained details and long-term temporal structure for improved multimodal fusion.



TCN



LSTM

TCN for local patterns, LSTM for long-term dependencies, and attention for multimodal fusion

[7] Wang, Xiaoming, et al. "Learning an Enhanced TCN-LSTM Network for Temperature Process Modeling in Rotary Kilns." IEEE Transactions on Automation Science and Engineering (2024).

Proposed Method

Details

$X_{video}, X_{audio} = Features\ extracted\ from\ backbone$

$X_{video} = AVGA(X_{video})$ # Audio Visual Guided Attention: reshape visual shape (16, 16, 25088) -> (16, 16, 512)

[TCN-LSTM Attention Block]

$T_v, T_a, L_v, L_a = TCN(X_{video}), TCN(X_{audio}), \text{LSTM}(X_{video}), LSTM(X_{audio})$ # (16, 16, 512)

$A_g,\ A_l = Q_{LSTM} \times K_{TCN}^T,\ K_{TCN}^T \times Q_{LSTM}$ # Attention Global, Attention Local

$TLAB_v, TLAB_a = A_g(VA_l)_{video}, A_g(VA_l)_{audio}$ # Output of TCN-LSTM Attention Block

$J = concat(TLAB_v, TLAB_a)$ # Concatenate audio, visual features from extracted TLAB

[Joint Cross Attention Fusion Block]

$C_v, C_a = \frac{\tanh(TLAB_v^T W_{JV} J)}{\sqrt{d}}, \frac{\tanh(TLAB_a^T W_{JA} J)}{\sqrt{d}}$

$H_v, H_a = ReLU(W_v TLAB_v + W_{c_v} C_v^T), ReLU(W_a TLAB_a + W_{c_a} C_a^T)$

$X_{att,v}, X_{att,a} = W_{H_v} H_v + x_v,\ W_{H_a} H_a + x_a$

$\hat{X} = BLSTM([X_{att,v}; X_{att,a}])$

$return\ \hat{X}$

Experimental Settings

Dataset

- Aff-Wild2 is an emotion recognition dataset that includes videos collected from YouTube, with each frame annotated for valence and arousal values.
- This dataset, used in the Affective Behavior Analysis in-the-wild (ABAW) competition, is a reliable source for emotion analysis with expert-annotated valence and arousal labels.

	Train	Valid	Test	Total
Raw video	475	73	50	598
Annotation	356	76	X	432
Input data	351	73	X	424

Provided data from database

Cross Validation

- Number of Iterations: 10
- Split Input data(424) to (Train: Validation: Test) = 0.6 : 0.2: 0.2

Experimental Settings

Hyperparameter Settings

- Hyperparameter settings were referenced from [2, 3] except for number of epochs and learning rate to enable practical experiments in our setup.

Method	Batch Size	Max Epochs	Loss Function	Optimizer	Learning Rate
Leader-Follower Fusion	2	100	CCC loss	Adam	1e-5
JCA	16	100	CCC loss	Adam	1e-3
Our Experiment	16	30	CCC loss	Adam	1e-4

Evaluation Measures

Measures	Formula	Description
CCC *(Concordance Coefficient Correlation)*	$\frac{2 * Corr(True, Pred) * \sigma_{True} * \sigma_{pred}}{\sigma_{True}^2 + \sigma_{pred}^2 + (\mu_{True} - \mu_{Pred})^2}$	• $Corr(True, Pred)$: Pearson correlation coefficient between the true values (True) and the predicted values (Pred). • Concordance between two variables (True, Pred). • CCC value ranges from -1 to 1, with a value closer to 1 indicating a higher concordance between the two variables. • CCC value close to 0 indicates low concordance.

Experimental Results

Performance Comparison

- The proposed method outperforms baseline models in all metrics, achieving the highest CCC scores for both valence and arousal.
- While previous studies often performed better on arousal, the proposed method shows a clear advantage in valence prediction.

Method	CCC_Valence	CCC_Arousal	CCC_Mean
Ours	**0.643 ± 0.010**	**0.597 ± 0.025**	**0.620 ± 0.008**
Leader-Follower Fusion	0.376 ± 0.021	0.436 ± 0.016	0.406 ± 0.018
JCA	0.452 ± 0.128	0.528 ± 0.091	0.490 ± 0.050

Experimental Results

Paired t-test

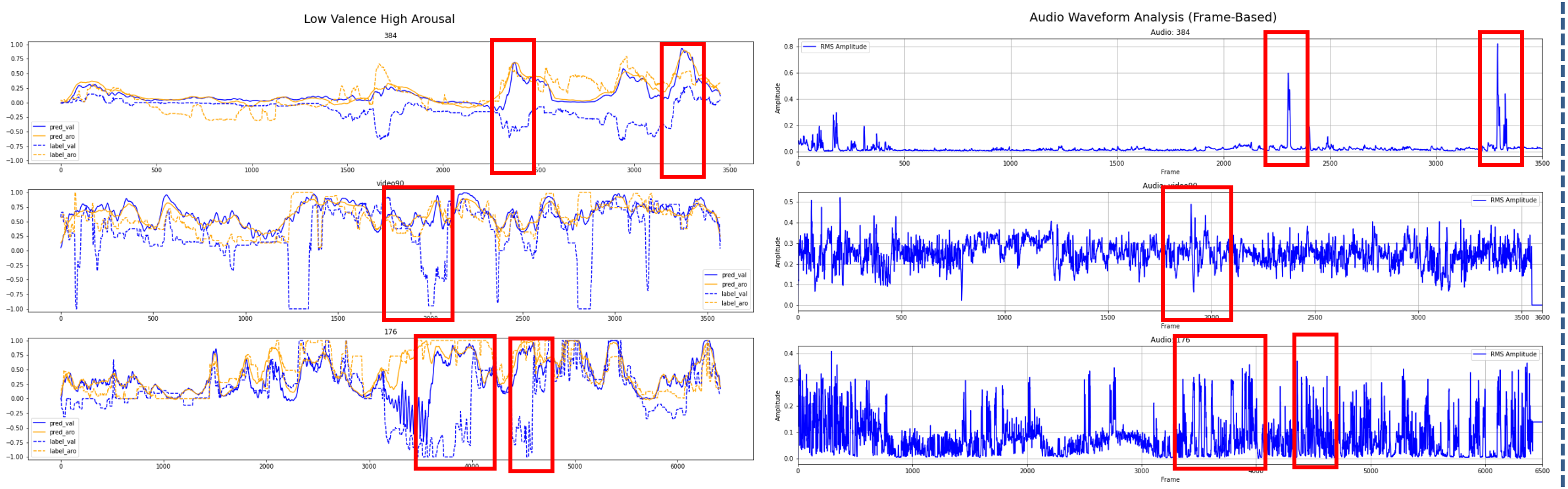
- Paired t-tests confirm that the proposed method significantly outperforms Leader-Follower across all metrics ($p < .01$).
- Compared to JCA, improvements in valence and mean CCC are also statistically significant ($p < .01$ and $p < .05$, respectively).

Method	CCC_Valence	CCC_Arousal	CCC_Mean
Ours	**0.643 ± 0.010**	**0.597 ± 0.025**	**0.620 ± 0.008**
Leader-Follower Fusion	0.376 ± 0.021**	0.436 ± 0.016**	0.406 ± 0.018**
JCA	0.452 ± 0.128**	0.528 ± 0.091	0.490 ± 0.050*

Experimental Results

In-depth Analysis

- The predicted valence (solid blue line) and label valence (dashed blue line) show contrasting trends across multiple segments , with distinct shifts occurring at specific time points.
- Significant gaps between predicted and label valence consistently appear when the audio volume rises sharply, showing a clear link between sound intensity and prediction errors.



JCA Bad CCC Score Videos

Experimental Results

In-depth Analysis

- This table shows cases where valence predictions do not match the labels across different videos. In horror movie reactions, co-driver emotion analysis, and other reaction videos, the model often struggles to track emotional changes accurately.
- When valence predictions differ from the labels, noticeable changes in audio levels are often present. Loud sounds, such as sudden effects in horror movies or car noises, seem to affect the model predictions, sometimes leading to incorrect results.

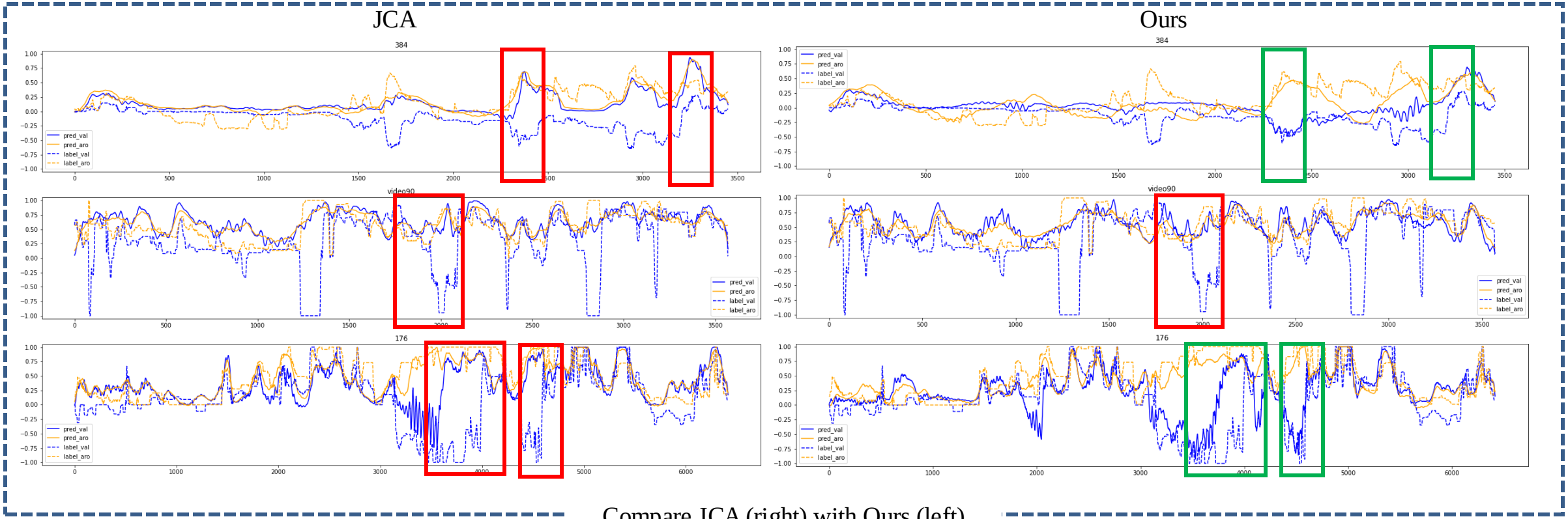
Video	Description	Frame
384	In reaction videos to horror movies, when the sound of the horror movie is loud, the valence prediction tends to differ from the label	
Video90	In videos analyzing the emotions of a co-driver, the valence prediction tends to differ from the label when loud car noises or a woman's scream occur	
176	In reaction videos to specific content, even when the scene is not pleasant and does not evoke positive emotions, the valence prediction tends to differ from the label when loud sounds are present in the video	

JCA Bad CCC Score Videos

Experimental Results

In-depth Analysis

- In videos 384 and 176, JCA struggles to predict valence in regions where loud sounds occur (highlighted in red), whereas our method shows improved alignment with the ground truth in those areas (highlighted in green).
- However, in video 90, mispredictions still persist, suggesting that while our method improves valence prediction in loud sound regions, challenges remain in handling certain complex scenarios.



Compare JCA (right) with Ours (left)

Conclusion and Contribution

Conclusion

- Our study focused on addressing the contextual differences between the two modalities that cause significant confusion in fusion methods within real-life videos.
- The proposed model integrates TCN for local dependencies and LSTM for long-term context, using attention to align their features, enabling balanced multimodal fusion.
- The proposed method is evaluated on the Aff-Wild2 dataset, which collects real-world videos from YouTube and is used in the ABAW Challenges, and our approach outperforms other related methods.

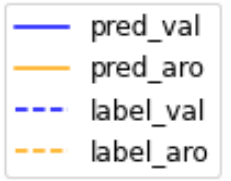
Achievement

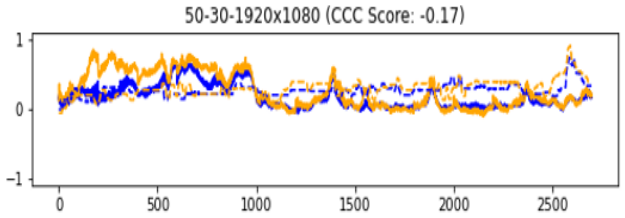
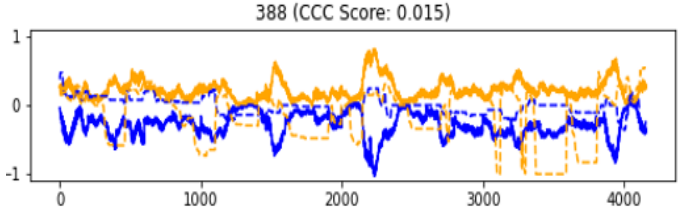
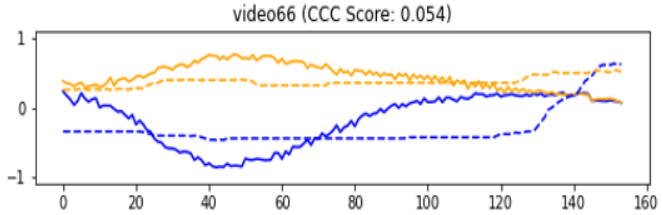
- In Domestic Conferences
  - A Crux on Audio-Visual Continuous Emotion Recognition using Fusion Techniques, ***대한전자공학회 학술대회***, 2024

Thank you

Appendix

Each Studies' Bad CCC Score Videos



Method	Bad CCC Score Videos		
Zhang et al. [4]			
Praveen et al. [5]