MultiMAE-DER: Multimodal Masked Autoencoder for Dynamic Emotion Recognition

Peihao Xiang, Chaohao Lin, Kaida Wu, Ou Bai

Department of Electrical and Computer Engineering, Florida International University, USA

{pxian001,clin027,kwu020,obai}@fiu.edu

July 9, 2024

Research Background

- Emotion Recognition Modalities: Visual (Facial, Behavior), Audio (Speech, Tone), Text (Semantics, Context).
- Traditional Emotion Recognition: Single-modality, Static, and Context-free.
- Dynamic Emotion Recognition: Multi-modality, Dynamic, and Contextual.
- For the emotion recognition in real-time human-computer interaction environments, dynamic emotion recognition is more suitable than traditional emotion recognition, specially, from videos.
- Purpose: This study will explore the impact of dynamic feature correlation in multimodal data (i.e., audio-visual cross-domain data) on the emotion recognition (not only facial expressions but also speech emotions).

Challenges

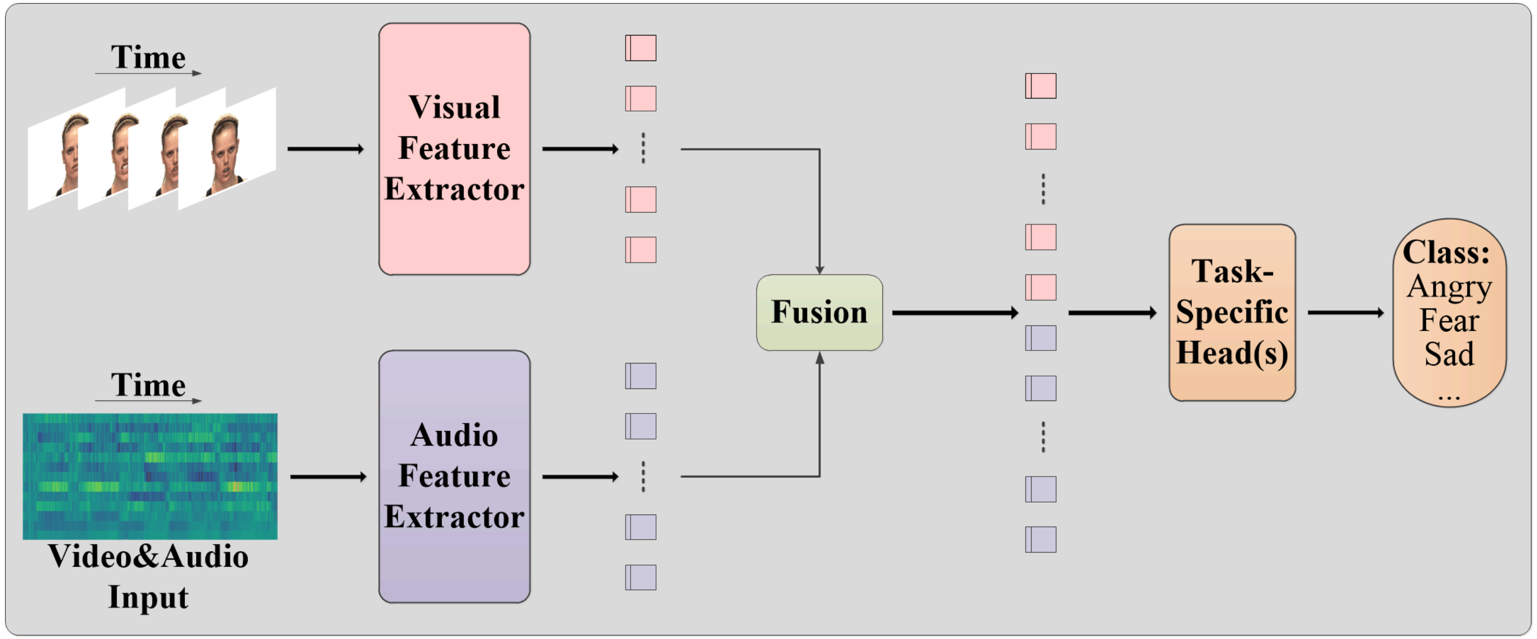
A. Considering that the model for extracting dynamically correlated features generally requires a large amount of training data, state-of-the-art dynamic emotion recognition models that are established on the supervised learning from labeled training data may NOT be robust on dynamic emotion recognition.

B. Further, the current dynamic emotion recognition models only consider the correlation between feature sequences of single modality, either spatially or temporally, while ignoring the advantage of multimodal sequence fusion, i.e., the correlation features between dynamic cross-domain data from the perspective of overall spatiotemporal sequence.

FIU

HCPS Laboratory

Related Work



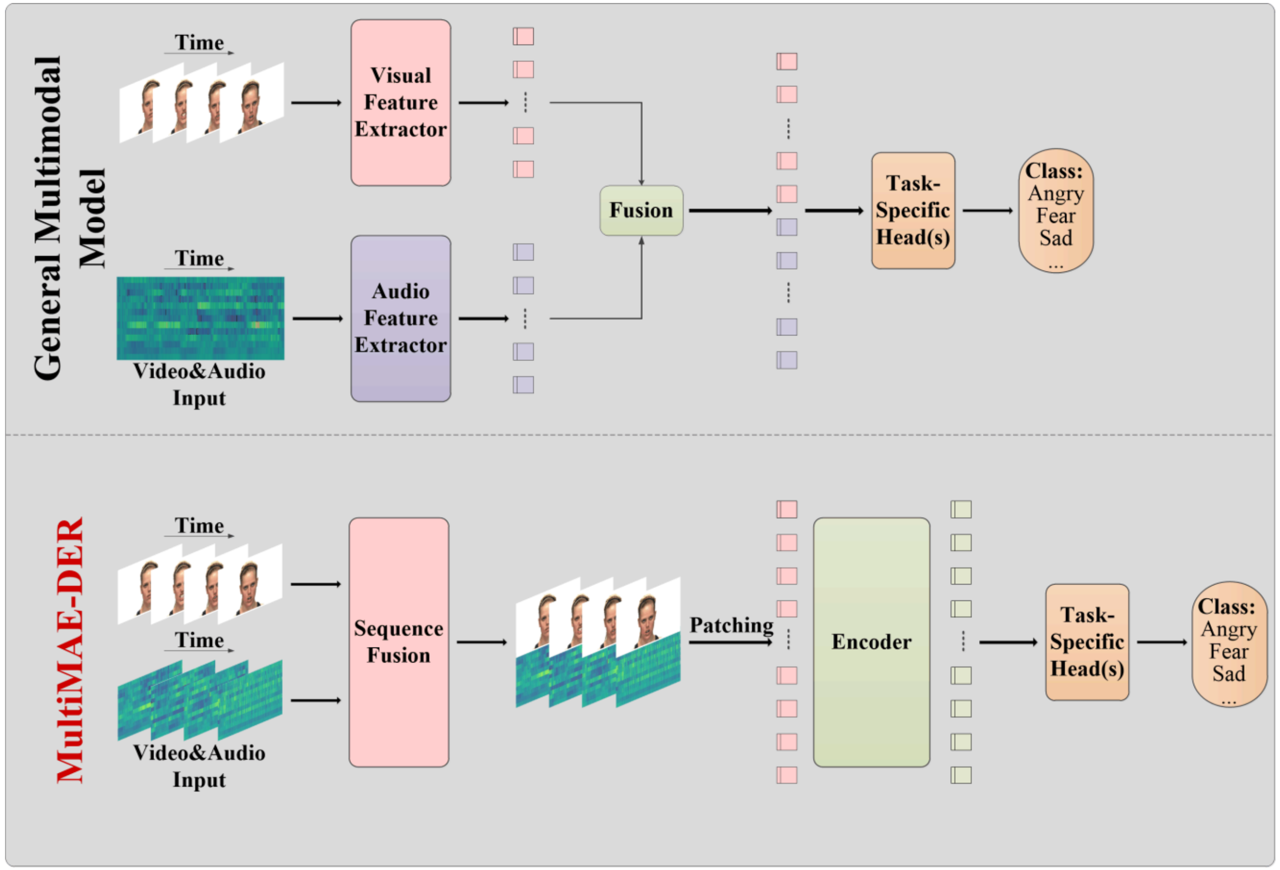
- State-of-the-Art Modality Fusion:
  - Single-modal Extract Features
  - Cross-modal Feature Fusion
  - Emotion Task-Head
- For example, AVT [1] and VQ-MAE-AV [2]

Difference	AVT	VQ-MAE-AV
Feature Extractor	Traditional 3DCNN Model	Self-supervised MAE Pre-trained Model
Fusion Strategy	Self-attention	Cross-attention
RAVDESS Dataset Result	79.20%	83.20%

Motivation

Explore efficient methods to extract dynamic feature correlations across cross-domain data from:

- Spatial-only sequence
- Temporal-only sequence
- Spatiotemporal sequence



Contribution

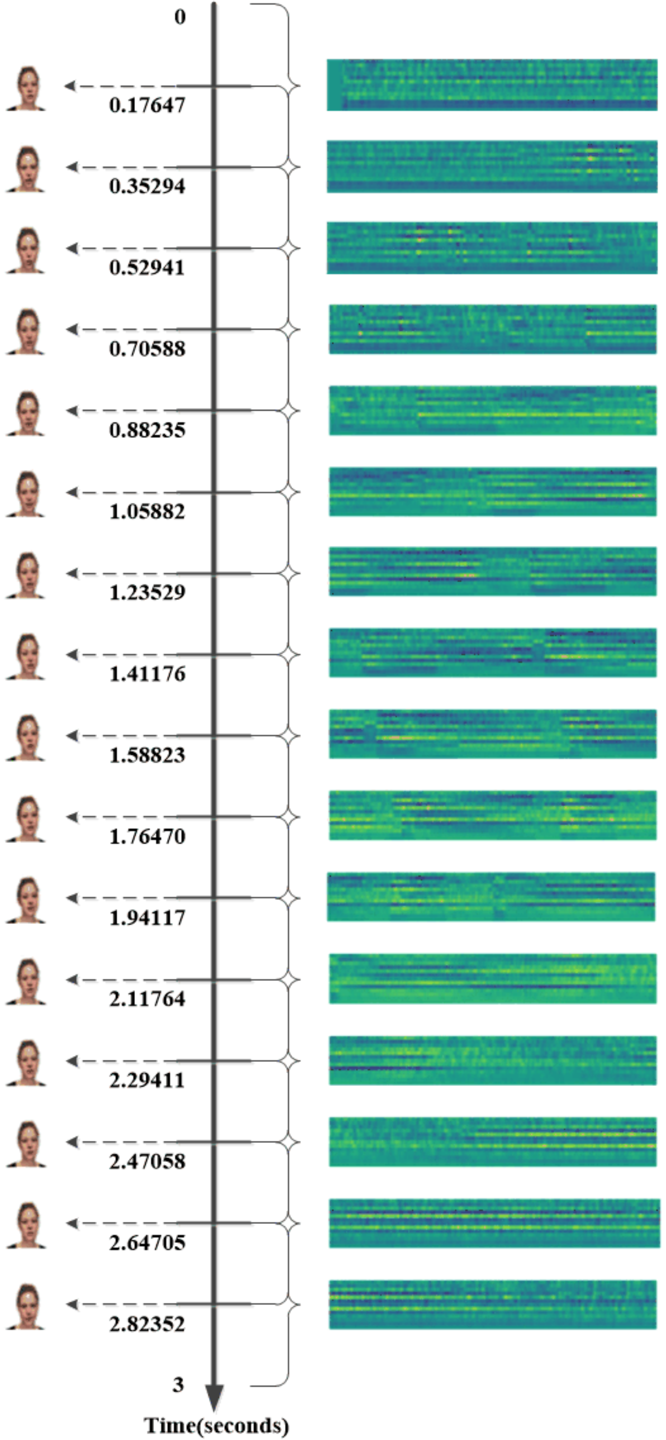
- A New Framework for dynamic emotion recognition that extending the conventional approach using single-modal input to multimodal input encompassing both visual and audio elements.

- Optimizing Visual-Audio sequence fusion strategies.

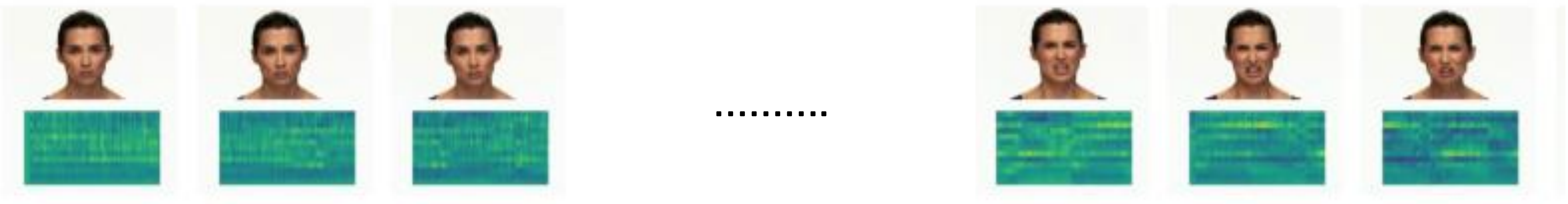
Map of Visual vs. Audio

- Visual: Facial expression image
  - Process: 16 frames down-sampled from 90 frames (3 seconds)
  - Data: $V \in R^{16\times224\times224}$

- Audio: Speech spectrogram
  - Process: 16 spectrogram from Mel-Frequency Cepstrum (MFCC)
  - Data: $A \in R^{16\times224\times224}$



Multimodal Sequence Strategy



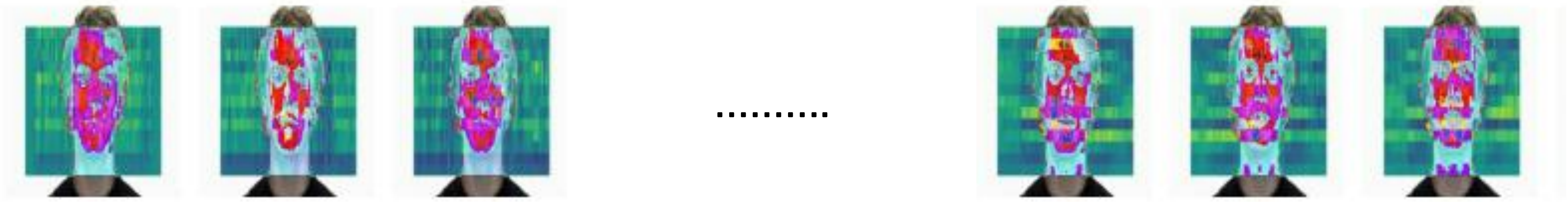
Strategy 1: Combine of Facial and Spectrogram (CFAS)

$$\mathbf{X}_i = \mathrm{Concat}(\mathbf{V}_i, \mathbf{A}_i) \tag{1}$$

Where i is the time index, $i \in [1,16]$.

- ❖ Strategy: Concatenation
- ❖ Reason: Planar spatial integrity on dynamic emotion recognition.

Multimodal Sequence Strategy



Strategy 2: Sum of Facial and Spectrogram (SFAS)

$$\mathbf{X}_i = \mathrm{Add}(\mathrm{Norm}(\mathbf{V}_i), \mathrm{Norm}(\mathbf{A}_i)) \qquad (2)$$

Where i is the time index, $i \in [1,16]$.

- Strategy: Superposition
- Reason: Depth spatial integrity on dynamic emotion recognition.

FIU

HCPS Laboratory

Multimodal Sequence Strategy

Strategy 3: First Facial Later Spectrogram (FFLS)

$$\mathbf{X} = \mathrm{Seq}(\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_8, \mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_8) \qquad (3)$$

Where V_i and A_i are the video and audio time sequences, $i \in [1,8]$.

- Strategy: Visual-auditory continuous sequence
- Reason: Visual-auditory spatiotemporal integrity on dynamic emotion recognition.

FIU

HCPS Laboratory

Multimodal Sequence Strategy

Strategy 4: First Spectrogram Later Facial (FSLF)

$$\mathbf{X} = \mathrm{Seq}(\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_8, \mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_8) \tag{4}$$

Where V_i and A_i are the video and audio time sequences, $i \in [1,8]$.

- ❖ Strategy: Audio-visual continuous sequence
- ❖ Reason: Audio-visual spatiotemporal integrity on dynamic emotion recognition.

Multimodal Sequence Strategy

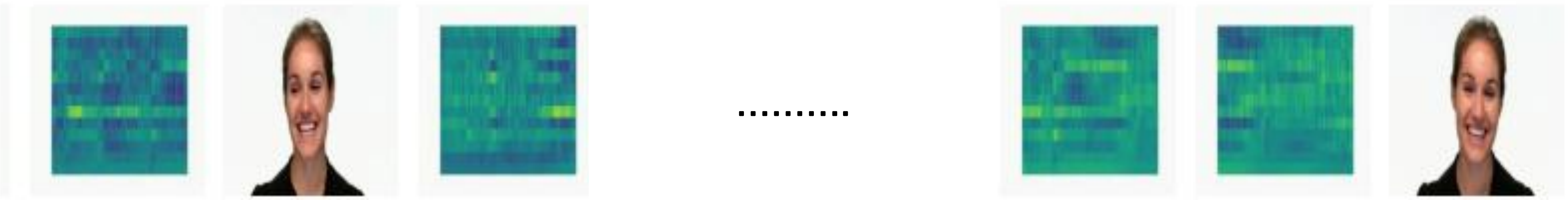
Strategy 5: One Facial One Spectrogram (OFOS)

$$\mathbf{X} = \mathrm{Seq}(\mathbf{V}_1, \mathbf{A}_1, \mathbf{V}_2, \mathbf{A}_2, \ldots, \mathbf{V}_8, \mathbf{A}_8) \tag{5}$$

Where V_i and A_i are the video and audio time sequences, $i \in [1,8]$.

- ❖ Strategy: Discrete sequence
- ❖ Reason: Audio-visual periodic temporal sequences on dynamic emotion recognition.

Multimodal Sequence Strategy



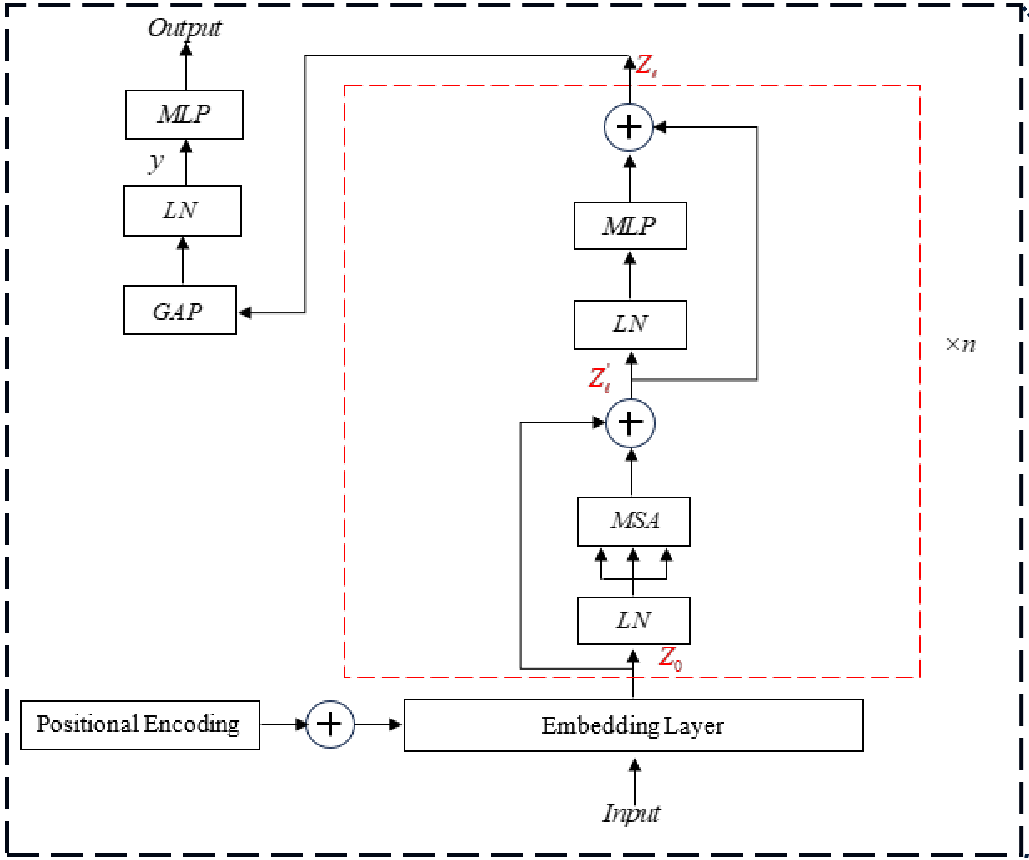
Strategy 6: Random of Facial and Spectrogram (RFAS)

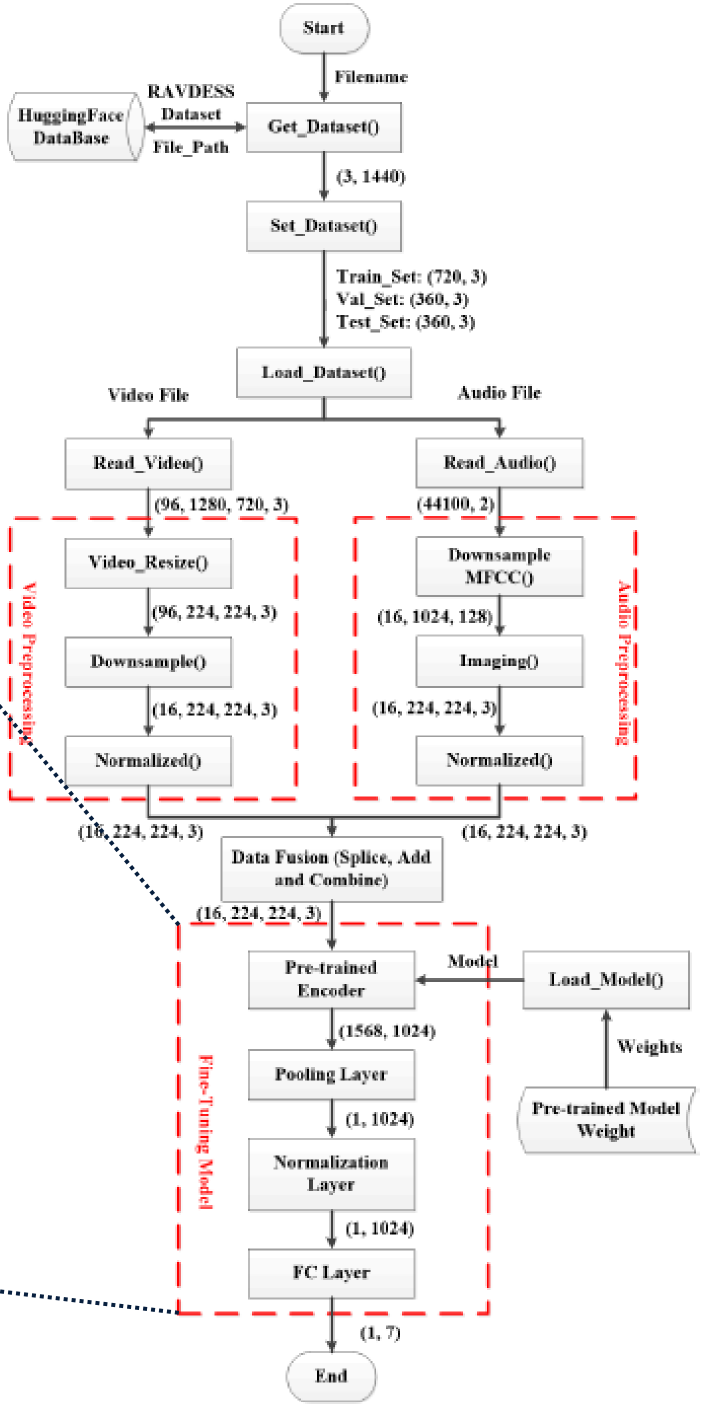
$$\mathbf{X} = \mathrm{Rand}(\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_8, \mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_8) \qquad (6)$$

Where V_i and A_i are the video and audio time sequences, $i \in [1,8]$.

- Strategy: Random sequence
- Reason: Audio-visual random spatiotemporal sequences on dynamic emotion recognition.

MultiMAE-DER: Model Structure





FIU

HCPS Laboratory

Process Details

Input:

$$X = [x_1, x_2, ..., x_m]^T \in R^{m \times p^2 \bullet C}$$

Where, the length for any x_i is $p^2 \cdot C$

Encoder Processing Steps:

$$Z_0 = XE + E_{pos}$$

$$Where, E \in R^{p^2 \bullet C \times d}, E_{pos} \in R^{m \times d}$$

$$Z_\ell^{'} = MSA(LN(Z_{\ell-1})) + Z_{\ell-1}$$

$$Where, \ell = 1, 2, ..., n$$

$$Z_\ell = MLP(LN(Z_\ell^{'})) + Z_\ell^{'}$$

$$Where, \ell = 1, 2, ..., n$$

$$y = LN(Average\_pooling(Z_\ell))$$

$$Example: Input.size = 16 * 224 * 224 * 3$$

$$Assume: T = 2, p = 16, d = 1024, class = 7$$

$$Then: m = 8 * 14 * 14, x_i = 16 * 16 * 3$$

$$X.size = m * x_i = 8 * 196 * 768$$

$$Z_0.size = 1568 * 1024$$

$$Z_l'.size = 1568 * 1024$$

$$Z_l.size = 1568 * 1024$$

$$y.size = 1 * 1024 \qquad Output.size = 1 * 7$$

RAVDESS Dataset Evaluation

Method	SSL	Modality	UAR	WAR
AV-LSTM [15]	χ	V+A	–	65.80
AV-Gating [15]	χ	V+A	–	67.70
MCBP [24]	χ	V+A	–	71.32
MMTM [25]	χ	V+A	–	73.12
ERANNs [26]	χ	V+A	–	74.80
MSAF [16]	χ	V+A	–	74.86
SFN-SR [17]	χ	V+A	–	75.76
MATER [27]	χ	V+A	–	76.30
MulT [28]	χ	V+A	–	76.60
AVT [29]	χ	V+A	–	79.20
VQ-MAE-AV [30]	✓	V+A	–	83.20
MultiMAE-DER	✓	V	–	74.13
MultiMAE-DER	✓	A	–	80.55
MultiMAE-DER-RFAS	✓	V+A	75.97	75.44
MultiMAE-DER-SFAS	✓	V+A	75.79	76.94
MultiMAE-DER-OFOS	✓	V+A	77.78	78.61
MultiMAE-DER-CFAS	✓	V+A	80.65	81.39
MultiMAE-DER-FFLS	✓	V+A	82.27	83.56
MultiMAE-DER-FSLF	✓	V+A	**83.23**	**83.61**

Major Findings:

- Best Strategy: MultiMAE-DER-FSLF (strategy 4)
- Outperforms the supervised model AVT by 4.41% (83.61% vs. 79.20%).
- Outperforms the self-supervised model VQ-MAE-AV by 0.41% (83.61% vs. 83.20%).
- Outperforms the visual-only model by 9.48% (83.61% vs. 74.13%).
- Outperforms the audio-only model by 3.06% (83.61% vs. 80.55%).

CREMA-D Dataset Evaluation

Method	SSL	Modality	UAR	WAR
EF-GRU [31]	χ	V+A	–	57.06
LF-GRU [31]	χ	V+A	–	58.53
TFN [32]	χ	V+A	–	63.09
MATER [27]	χ	V+A	–	67.20
AuxFormer [33]	χ	V+A	–	71.70
AV-LSTM [15]	χ	V+A	–	72.90
AV-Gating [15]	χ	V+A	–	74.00
RAVER [34]	χ	V+A	–	<u>77.30</u>
VQ-MAE-AV [30]	✓	V+A	–	78.40
MultiMAE-DER	✓	V	–	77.83
MultiMAE-DER	✓	A	–	78.45
MultiMAE-DER-RFAS	✓	V+A	74.62	74.90
MultiMAE-DER-SFAS	✓	V+A	75.73	75.48
MultiMAE-DER-OFOS	✓	V+A	76.88	76.54
MultiMAE-DER-CFAS	✓	V+A	78.24	78.16
MultiMAE-DER-FFLS	✓	V+A	78.59	78.83
MultiMAE-DER-FSLF	✓	V+A	**79.12**	**79.36**

Major Findings:

- Best Strategy: MultiMAE-DER-FSLF (strategy 4)
- Outperforms the supervised model RAVER by 2.06% (79.36% vs. 77.30%).
- Outperforms the self-supervised model VQ-MAE-AV by 0.96% (79.36% vs. 78.40%).
- Outperforms the visual-only model by 1.53% (79.36% vs. 77.83%).
- Outperforms the audio-only model by 0.91% (79.36% vs. 78.45%).

FIU

HCPS Laboratory

IEMOCAP Dataset Evaluation

Method	SSL	Modality	UAR	WAR
AV-HuBERT [35]	✓	V+A	–	46.45
MAViL [36]	✓	V+A	–	54.94
AVBERT [37]	✓	V+A	–	61.87
MultiMAE-DER	✓	V	–	56.13
MultiMAE-DER	✓	A	–	58.69
MultiMAE-DER-RFAS	✓	V+A	58.62	59.98
MultiMAE-DER-SFAS	✓	V+A	60.39	60.17
MultiMAE-DER-OFOS	✓	V+A	61.87	61.12
MultiMAE-DER-CFAS	✓	V+A	61.98	62.25
MultiMAE-DER-FFLS	✓	V+A	62.92	63.43
MultiMAE-DER-FSLF	✓	V+A	**63.21**	**63.73**

Major Findings:

- Best Strategy: MultiMAE-DER-FSLF (strategy 4)
- Outperforms the self-supervised model AVBERT by 1.86% (63.73% vs. 61.87%).
- Outperforms the visual-only model by 7.60% (63.73% vs. 56.13%).
- Outperforms the audio-only model by 5.04% (63.73% vs. 58.69%).

Analysis

Results indicate that fusing multimodal data on spatio-temporal sequences significantly improves the model performance by capturing correlations between cross-domain data.

- Strategy 1 (CFAS) exhibits temporal continuity but lacks spatial continuity.
- Strategy 2 (SFAS) has temporal continuity but disrupts the overall spatial sequence structure.
- Strategies 3 (FFLS) and 4 (FSLF) demonstrate both spatial and temporal continuity, with a high concentration of spatio-temporal correlation.
- Strategy 5 (OFOS) shows spatial continuity but disrupts the overall temporal sequence structure.
- Strategy 6 (RFAS) lacks both spatial and temporal continuity, simultaneously disrupting the overall spatio-temporal correlation.

Conclusion

- An exploration to handle multimodal data for dynamic emotion recognition - introducing a novel framework for the multimodal data integration established on self-supervised learning models.
- Investigation on six different multimodal sequence fusion strategies to explore diverse interpretations of multimodal correlation information extracted from a pre-trained model.

FIU

HCPS Laboratory

Thanks for listening!

https://hcps.fiu.edu/

Human Cyber-Physical Systems Laboratory