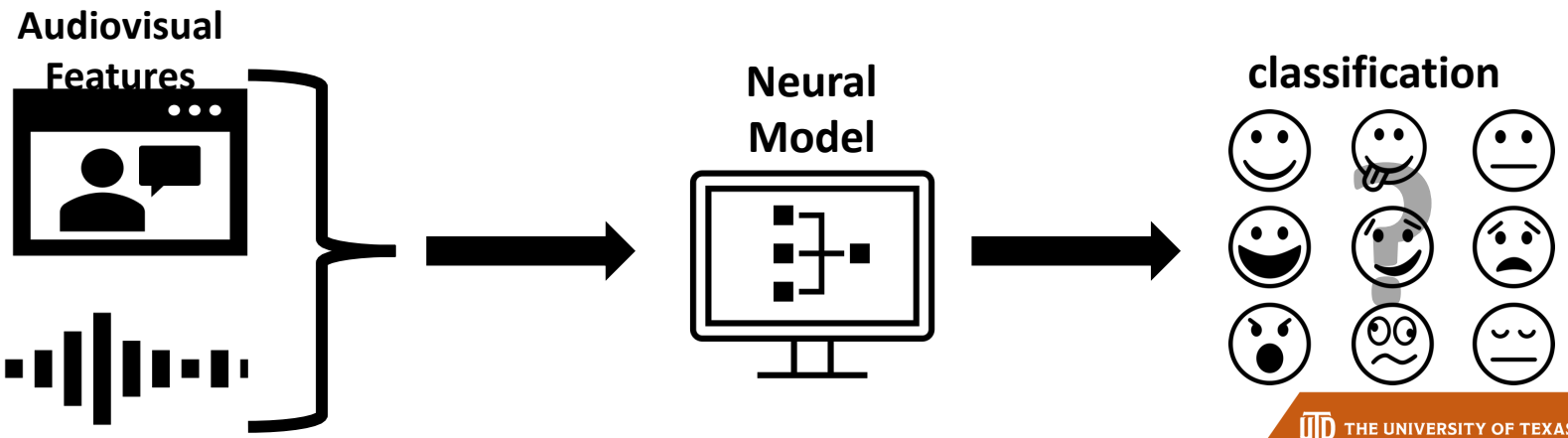
Learning Cross-modal Audiovisual Representations with Ladder Networks for Emotion Recognition

Lucas Goncalves and Carlos Busso

Grant IIS-1718944

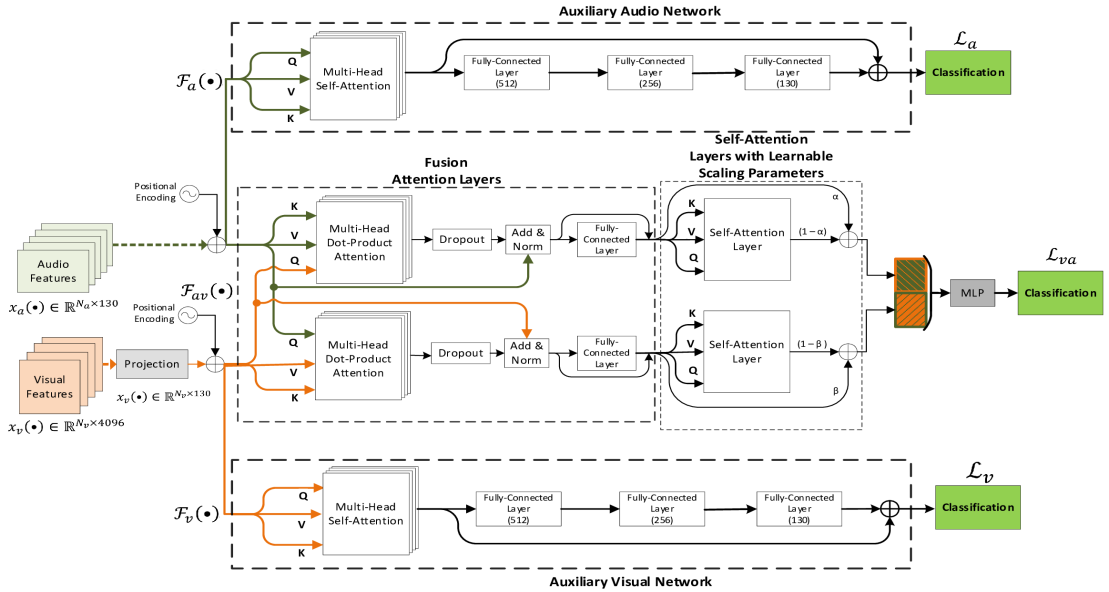
Motivation

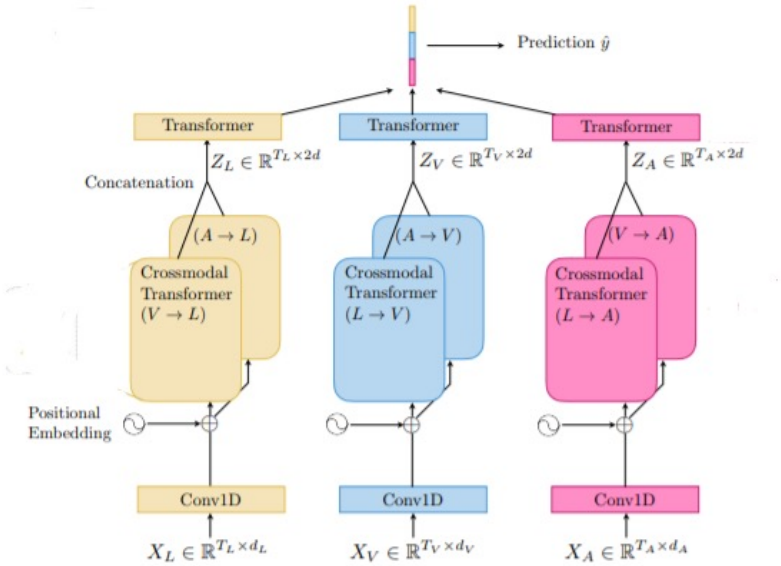
- **Application of audiovisual emotion recognition**
  - Enhancing human-computer interaction
    - Entertainment, education, wearable devices, … etc.
- **Audiovisual Emotion Recognition Problem**
  - Models have to process data points coming from heterogeneous sources
  - Capture modality-specific information while building strong cross-modal representations



Background

- **Forward Methods**



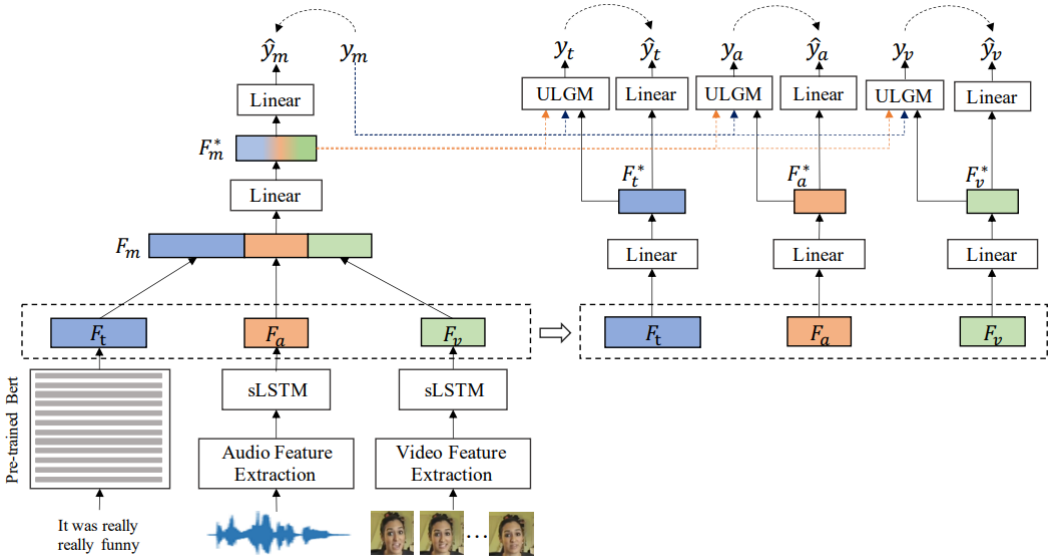


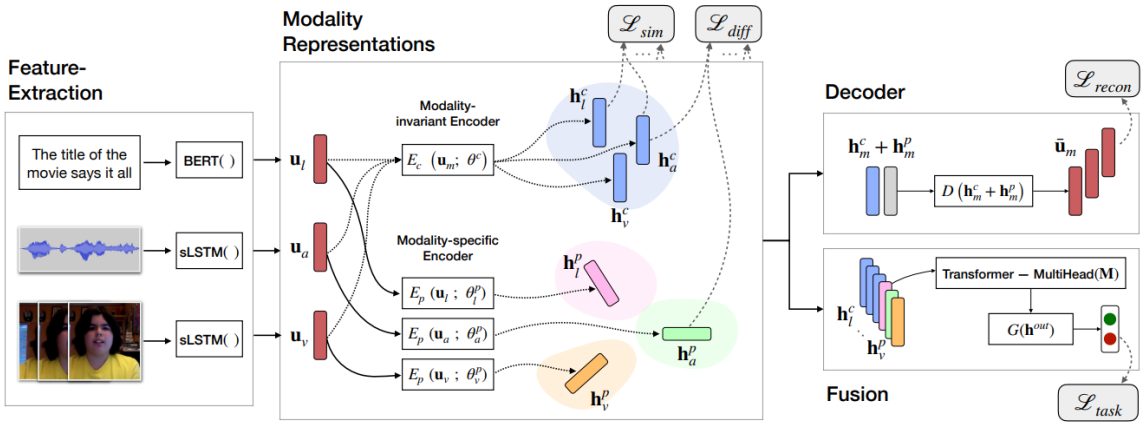
Lucas Goncalves and Carlos Busso, "AuxFormer: Robust approach to audiovisual emotion recognition," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022), Singapore, May 2022

Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, “Multimodal transformer for unaligned multimodal language sequences,” in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics

Background

- **Backward Methods**



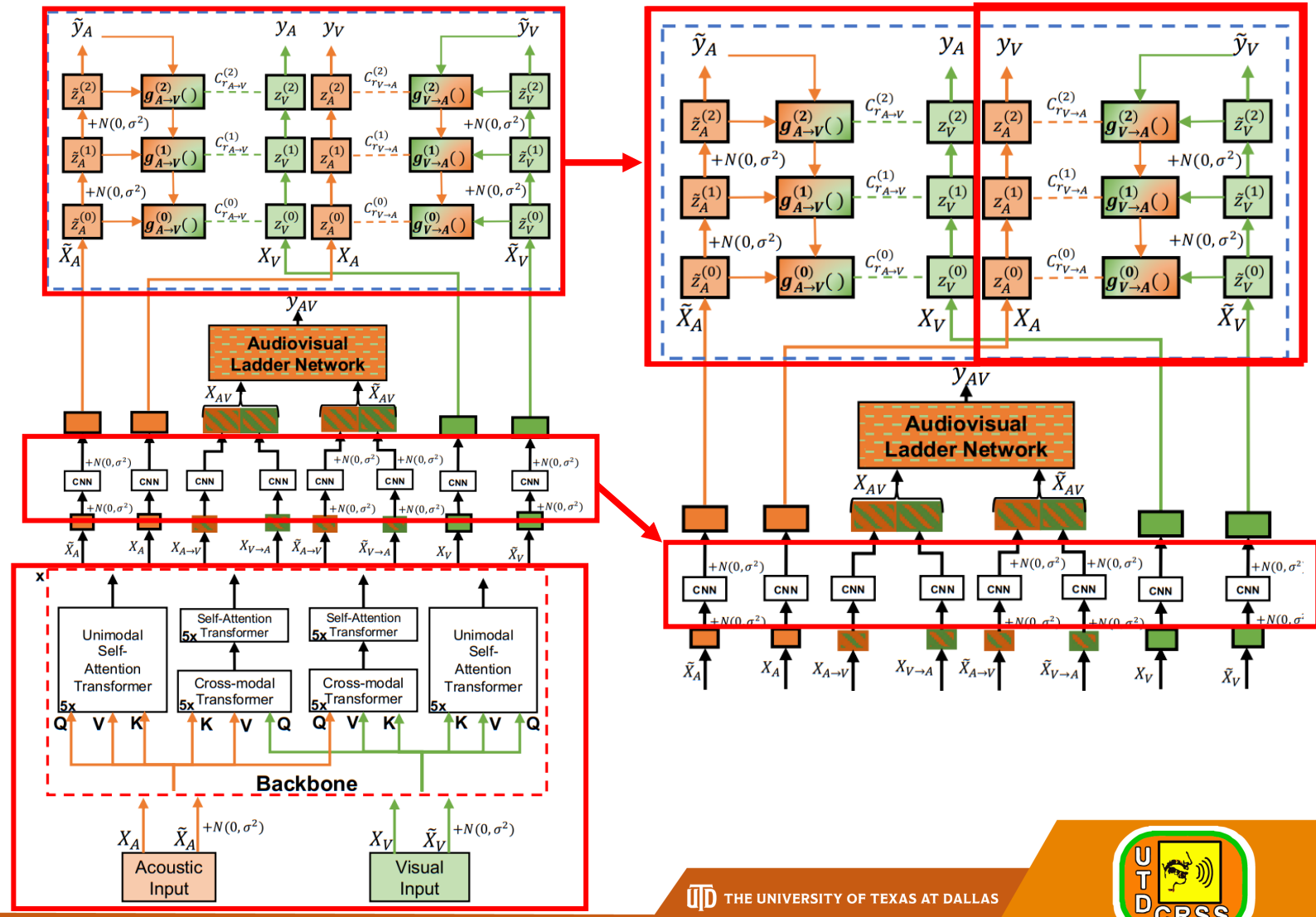


Yu, H. Xu, Z. Yuan, and J. Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," in AAAI Conference on Artificial Intelligence (AAAI 2021), Virtual Conference, February 2021, vol. 35, pp. 10790–10797

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis. In Proceedings of the 28th ACM International Conference on Multimedia (MM '20). Association for Computing Machinery, New York, NY, USA, 1122–1131.

Proposed Framework

- **Components:**
  - Representations from a backbone network
  - Unsupervised auxiliary tasks with multimodal ladder networks
  - Cross-modal skip connections between the encoder the decoder



Proposed Framework

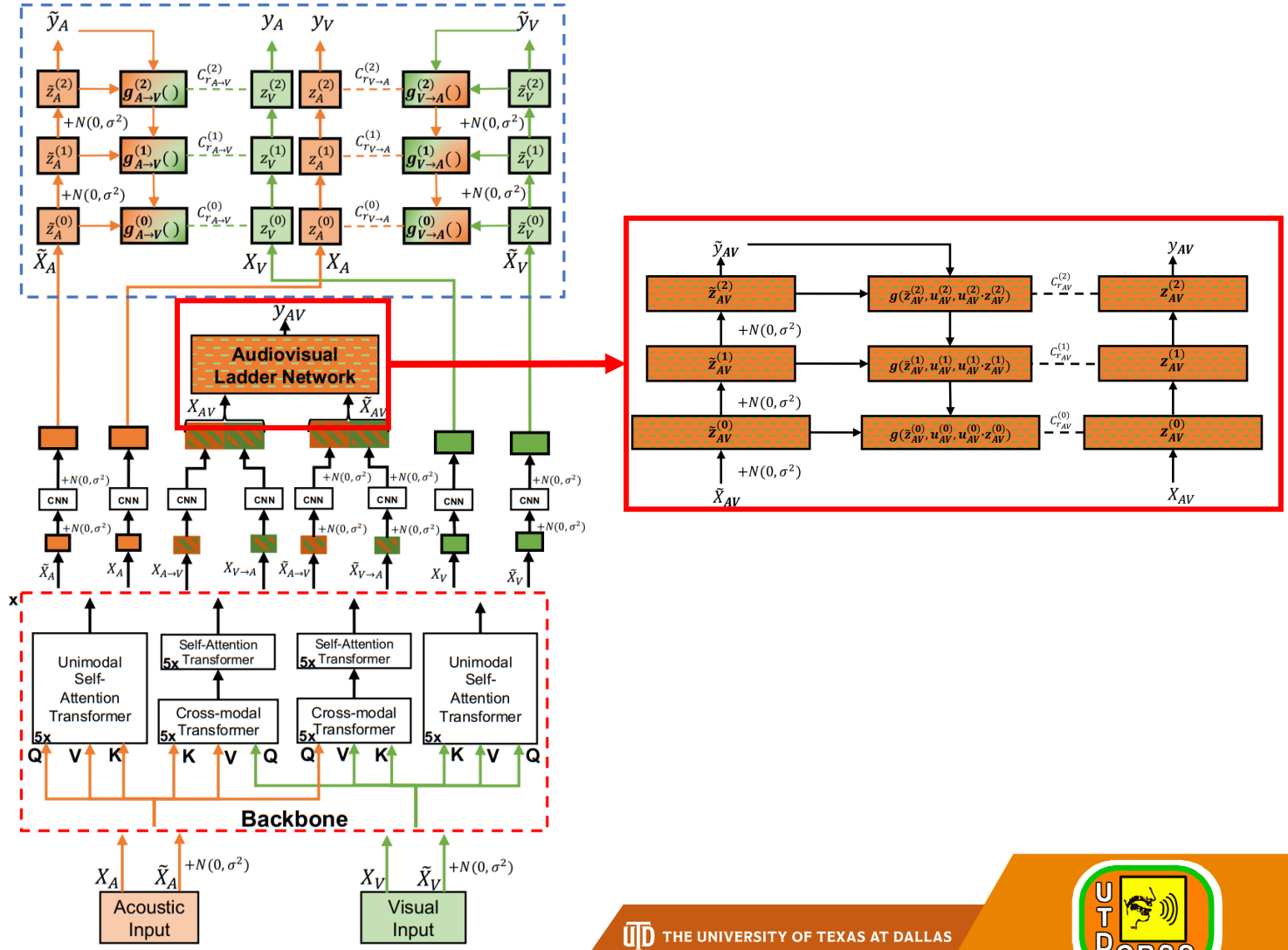
- **Components:**
  - The audiovisual ladder network takes as input X_{AV} and $\tilde{X}_{AV}$, which are processed by the cross-modal transformer
  - Ensures that cross-modal audiovisual representations are adequately captured by the system

- **Training losses:**

$$\mathcal{L}_{sup} = \frac{1}{3}(C_{s_{AV}} + C_{s_A} + C_{s_V})$$

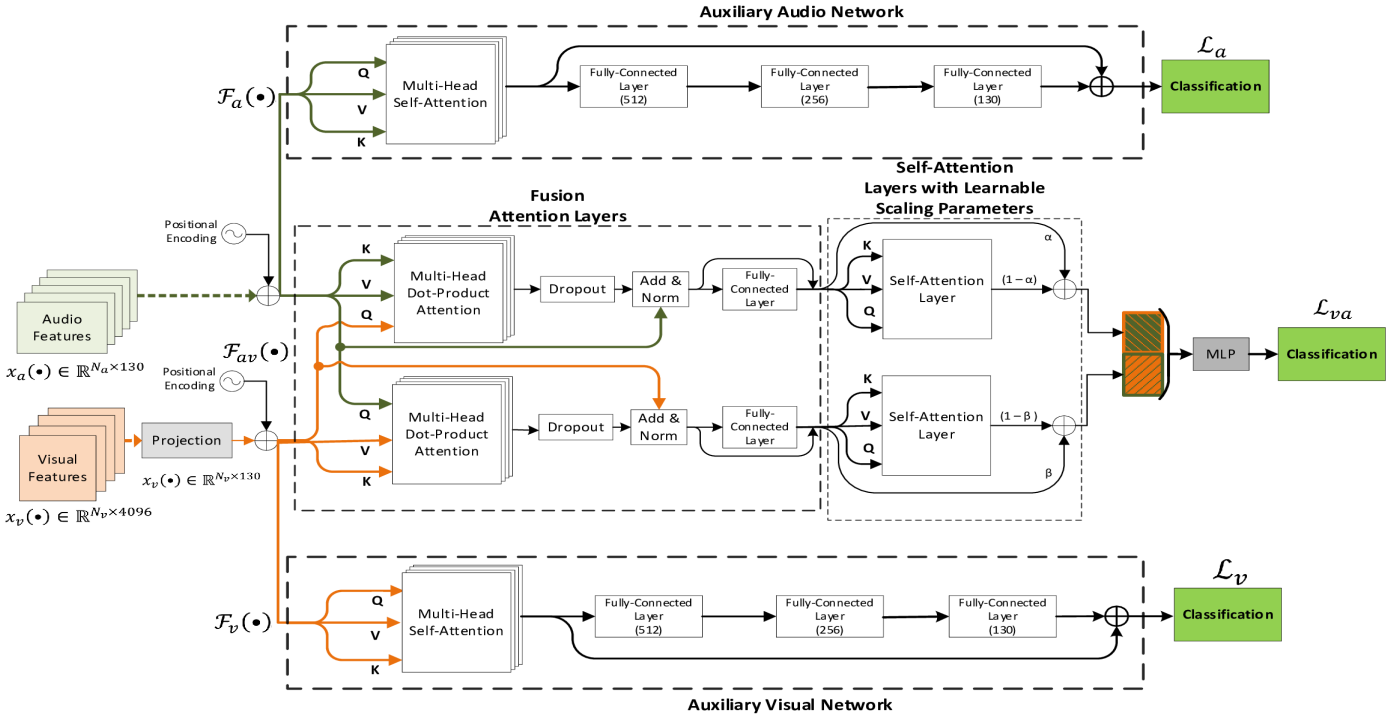
$$\mathcal{L}_{uns} = \lambda_l \left(\sum_l C^{(l)}_{r_{AV}} + \frac{1}{2} \left(\sum_l C^{(l)}_{r_{A\to V}} + C^{(l)}_{r_{V\to A}} \right) \right)$$

$$\mathcal{L}_{total} = \mathcal{L}_{sup} + \mathcal{L}_{uns}$$



Proposed Framework

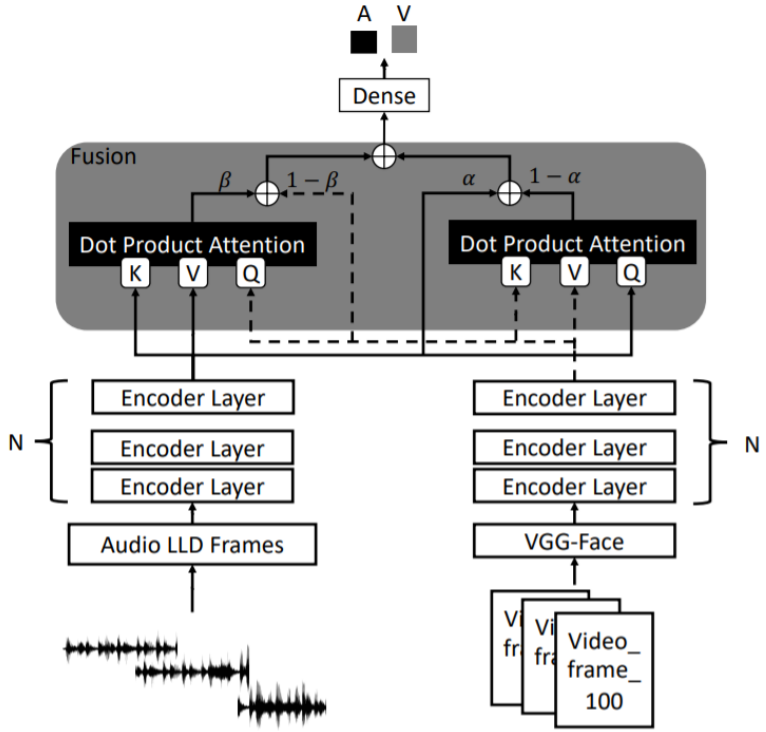
- **Baseline 1**



Lucas Goncalves and Carlos Busso, "AuxFormer: Robust approach to audiovisual emotion recognition," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022), Singapore, May 2022

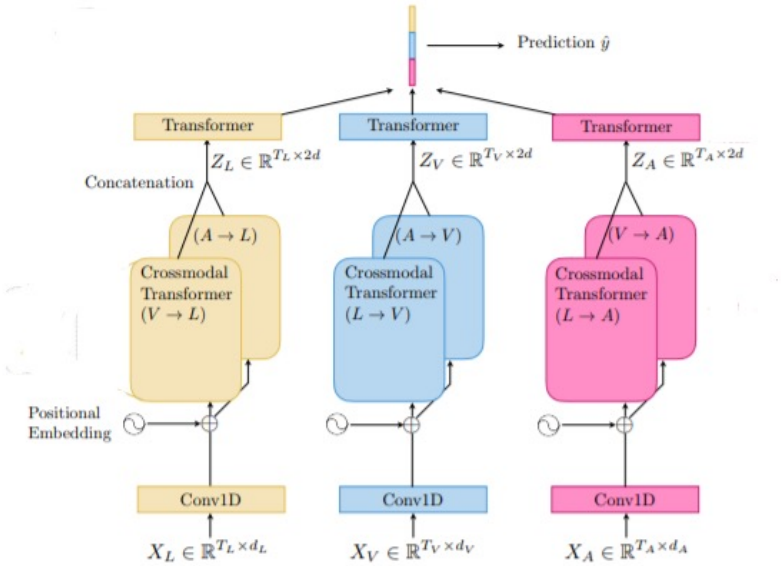
Baseline models

Baseline 2



S. Parthasarathy and S. Sundaram, "Training strategies to handle missing modalities for audio-visual expression recognition," in Companion Publication of the 2020 International Conference on Multimodal Interaction, ser. ICMI '20 Companion.

Baseline 3



Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics

CREMA-D Corpus

- **Videos of subjects saying sentences while displaying pre-defined emotions**
  - Corpus was collected from an ethnically and racially diverse group
  - 91 actors (48 males and 43 females)
  - Contains 7,442 clips
  - 6-class problem: anger, happiness, sadness, fear, disgust, and neutral

neutral happy fear

H. Cao, D. Cooper, M. Keutmann, R. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced emotional multimodal actors dataset, "IEEE Transactions on Affective Computing, vol. 5, no. 4,pp. 377–390, October-December 2014

Experimental Setup

- **Settings**

  Results are reported based on average over 20 trials

  - We compare the results using a one-tailed matched paired t-test over the 20 results with p-value <0.05 to assert statistical significance

- **Randomly split database:**
  - 70% train set
  - 15% development set
  - 15% test set
- **Speaker-independent splits:**
  - No speaker overlap in train, development, and test sets

Experimental Results

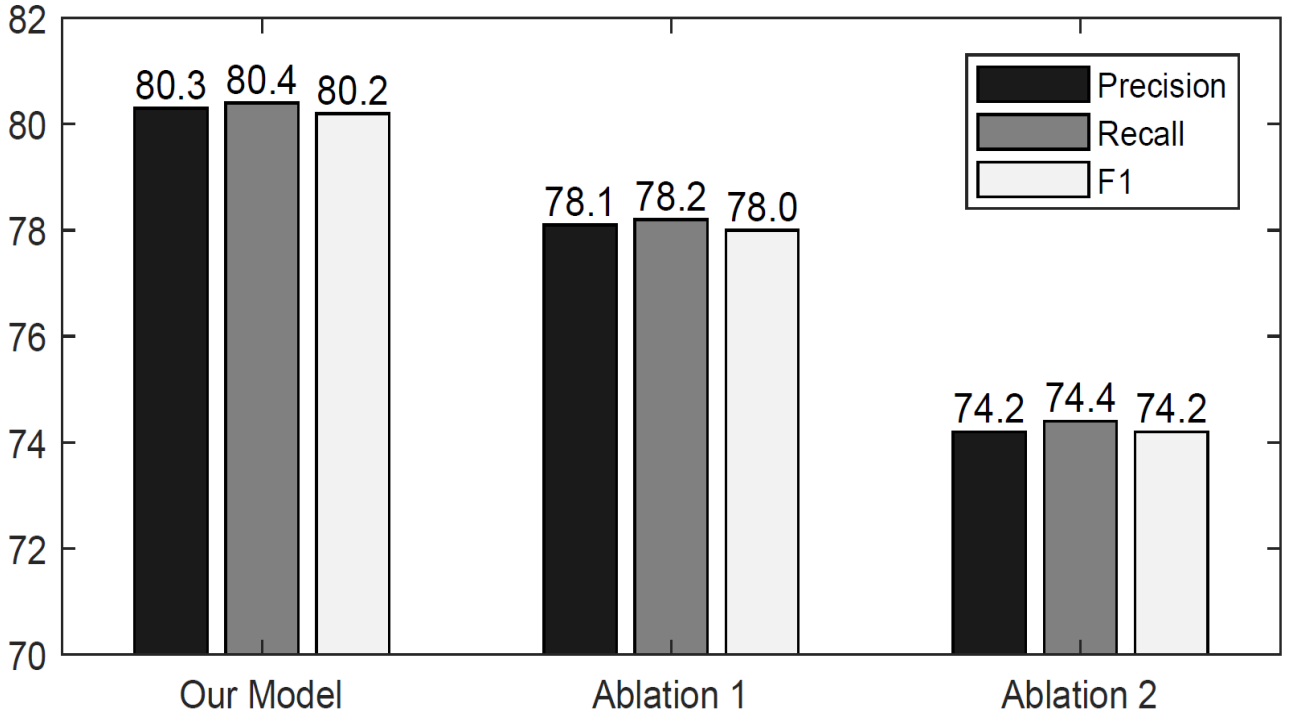
- **Standard Training Performance Details:**
  - Proposed framework achieves the best scores in all cases, significantly outperforming all the baselines
  - Baseline 1 contains mostly the same architecture as our backbone network. By comparing our model's performance with baseline 1, we can quantify the benefits of using the multimodal ladder networks

	Macro			Micro		
Architecture	**Prec.**	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**
Our Model	**80.3**	**80.4**	**80.2**	**80.3**	**80.3**	**80.3**
Baseline 1	76.5	75.7	75.5	75.7	75.7	75.7
Baseline 2	71.6	71.0	70.6	71.0	71.0	71.0
Baseline 3	60.6	57.8	56.3	58.0	58.0	58.0

We compare the results using a one-tailed matched paired t-test over the 20 results with p-value <0.05 to assert statistical significance

Experimental Results

- **Ablation experiments:**
  - Ablation 1 = removal of the unimodal cross-layer ladder network predictions
  - Ablation 2 = removal of the audiovisual ladder network



More Details in our Paper

Work Supported by:

Grant IIS-1718944

Thank you for your attention!

Check out the MSP lab

msp.utdallas.edu

Contact me at:
goncalves@utdallas.edu
linkedin.com/in/ilucasgoncalves
ilucasgoncalves.github.io