The City College of New York

Nonverbal Communication Cue Recognition: A Pathway to More Accessible Communication

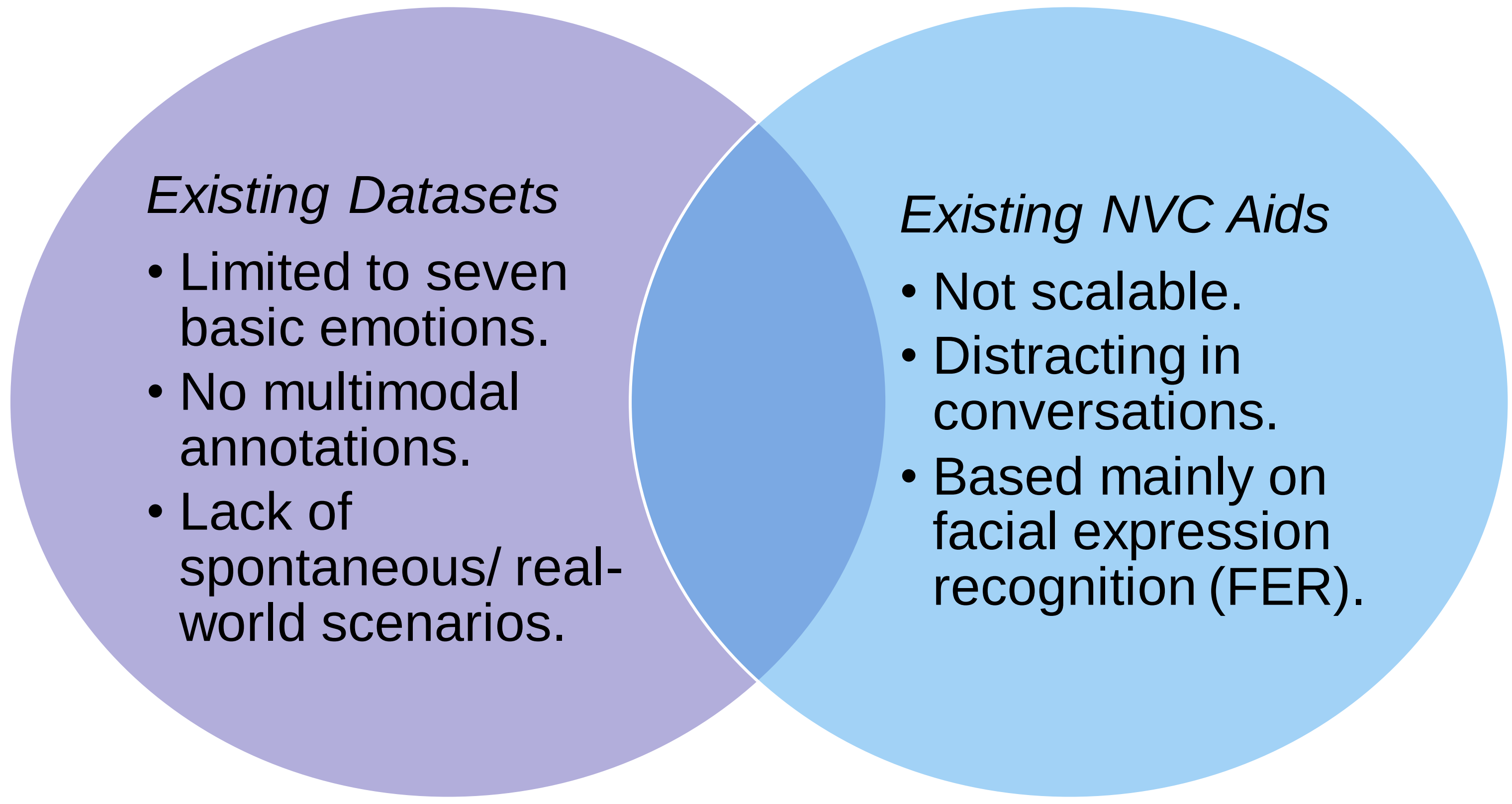
Zoya Shafique, Haiyan Wang, Yingli Tian

Women in Computer Vision

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Motivations

- **2.2 billion** people worldwide have some form of vision impairment [1].
- Body language makes up approximately **55%** of the information communicated during conversations [2].
- The blind and low vision (BLV) community understand other people's intentions, feelings, and beliefs differently than sighted people as they cannot perceive nonverbal cues (NVCs) [3].
- To contribute to the development of better NVC recognition aids, we are building the CCNY NVC Dataset and creating a multimodal action recognition model for NVC recognition in videos.



Existing Datasets

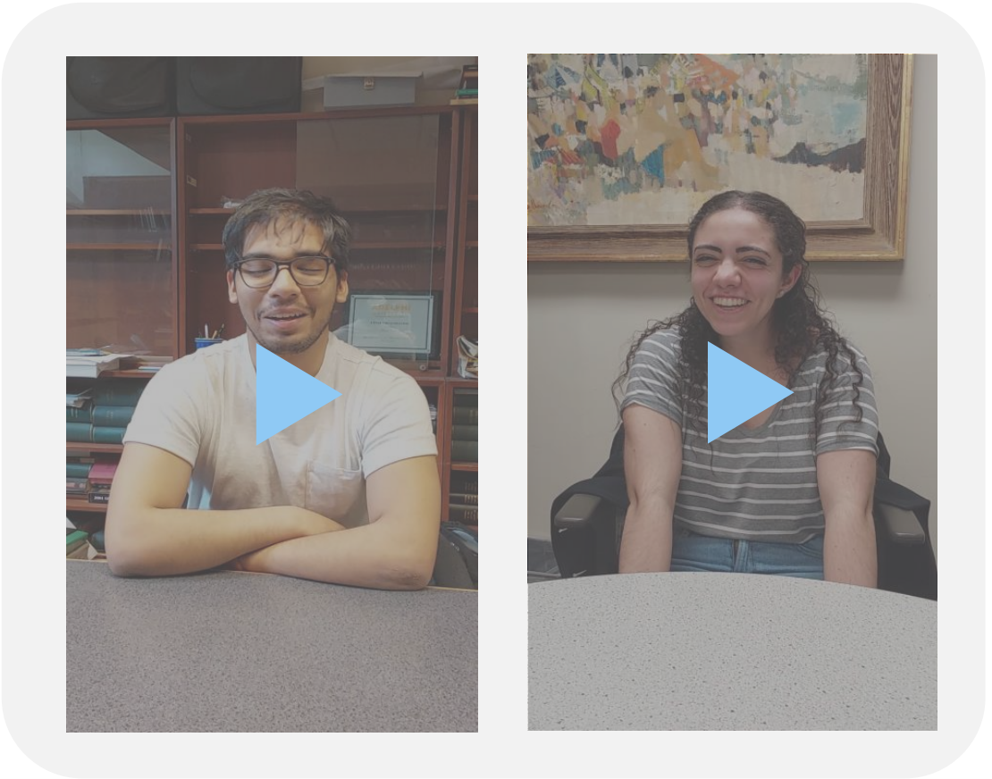
- Limited to seven basic emotions.
- No multimodal annotations.
- Lack of spontaneous/ real-world scenarios.

Existing NVC Aids

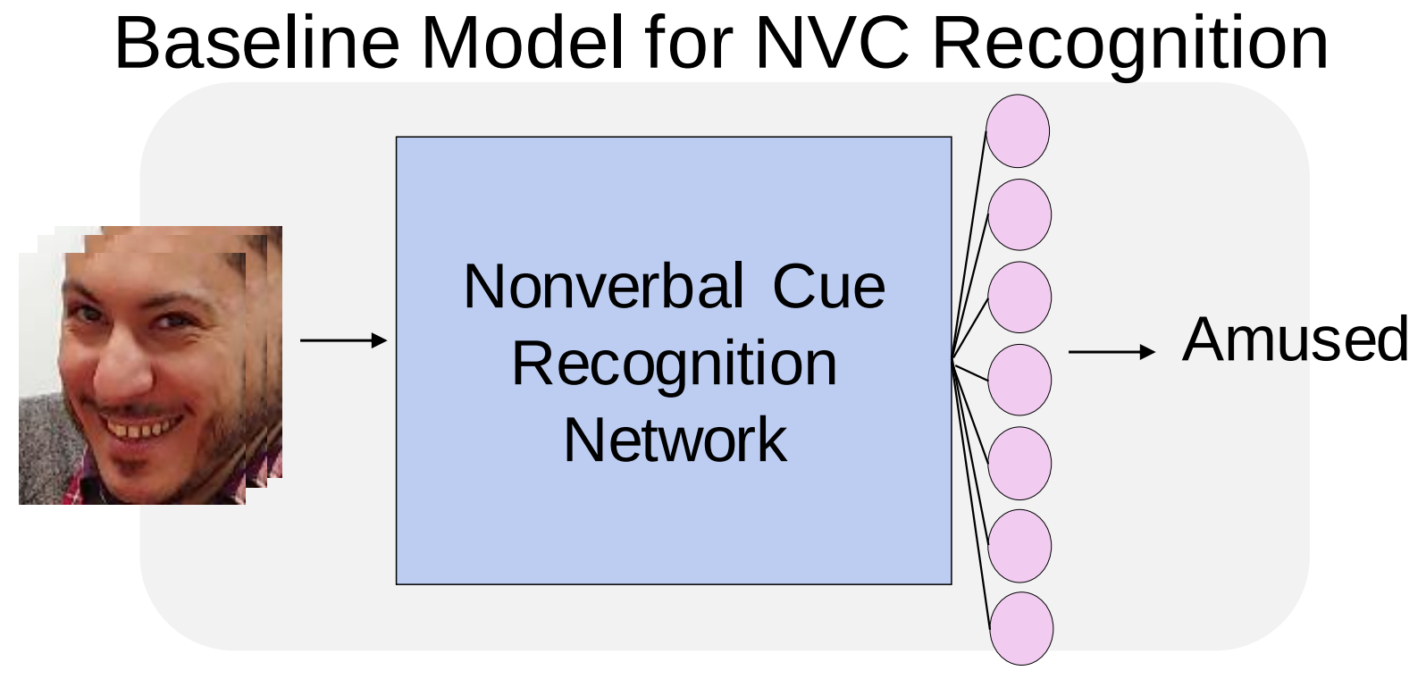
- Not scalable.
- Distracting in conversations.
- Based mainly on facial expression recognition (FER).

Key Contributions

CCNY NVC Dataset



Baseline Model for NVC Recognition



Nonverbal Cue Recognition Network

Amused

Introduced an in progress multi-modal dataset with both high-level emotion and fine-grained action annotations.

Achieved comparable results to previous SOTA methods on the Aff-Wild2 Dataset [4] with the proposed 3D-ResNet [7] for FER.

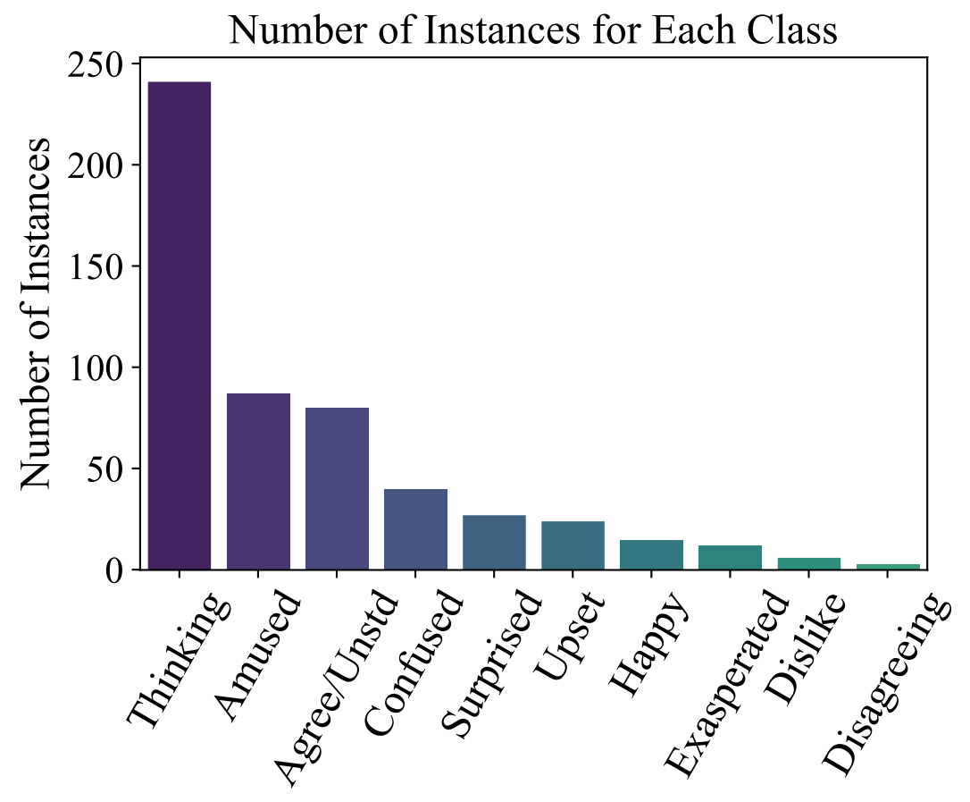
CCNY NVC Dataset

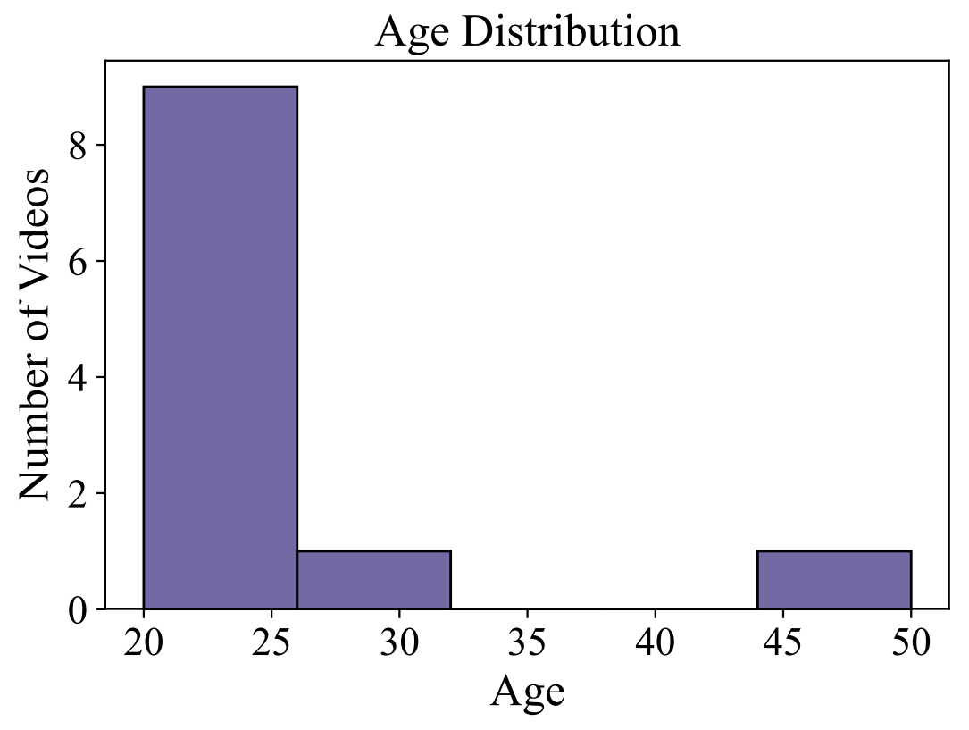
Dataset Statistics

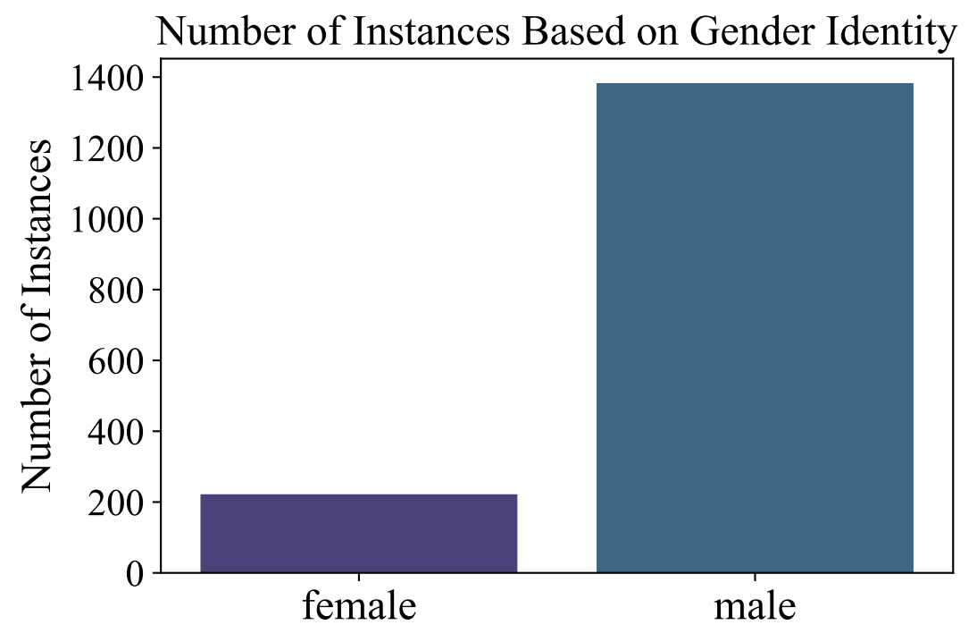
128 videos

12 subjects

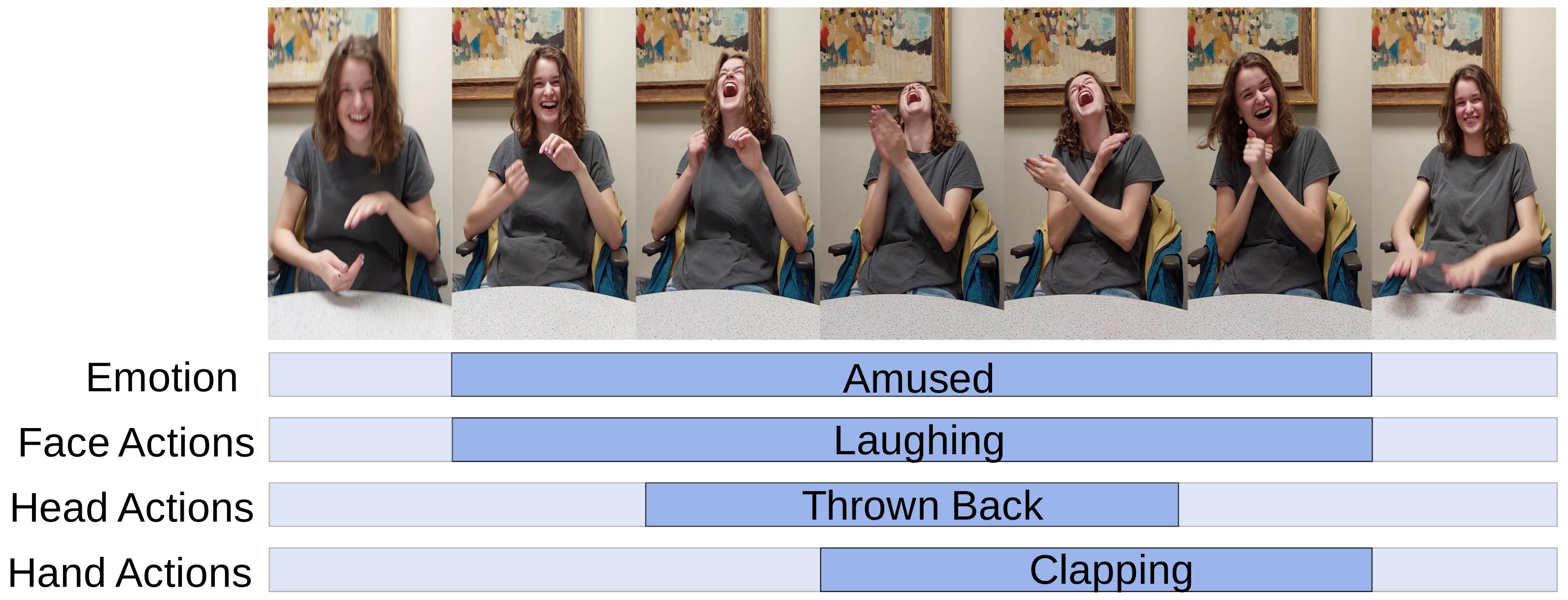
10 classes







Annotation Structure



Emotion: Amused

Face Actions: Laughing

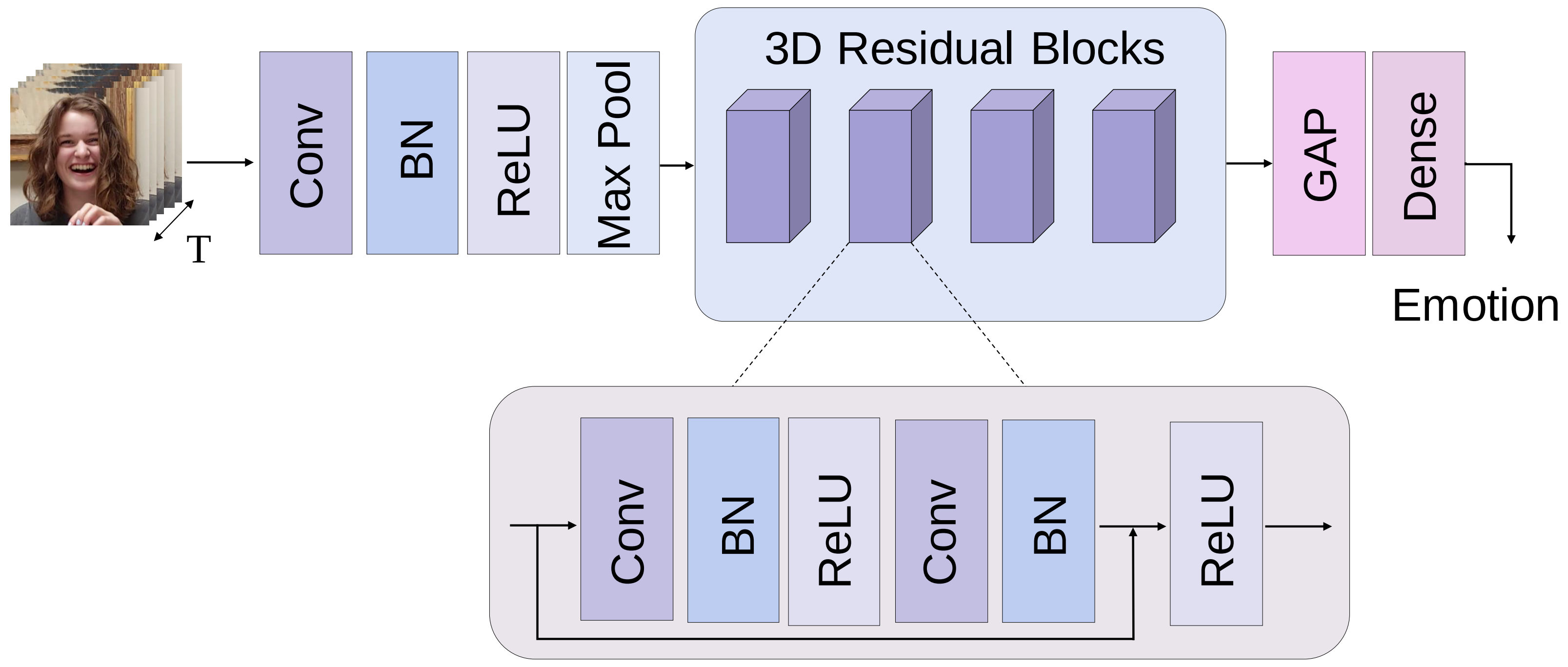
Head Actions: Thrown Back

Hand Actions: Clapping

Additional Information

- Captured using a Samsung Galaxy S7 FE 12.4".
- Videos of casual conversations in first person point of view.
- Large intra-class variance as shown on the right.

Facial NVC Recognition



T

Conv

BN

ReLU

Max Pool

3D Residual Blocks

GAP

Dense

Emotion

Conv

BN

ReLU

Conv

BN

ReLU

- Trained using weighted sampling of classes, weight decay, and focal loss with the Adam optimizer.

Results on Aff-Wild2 [4]

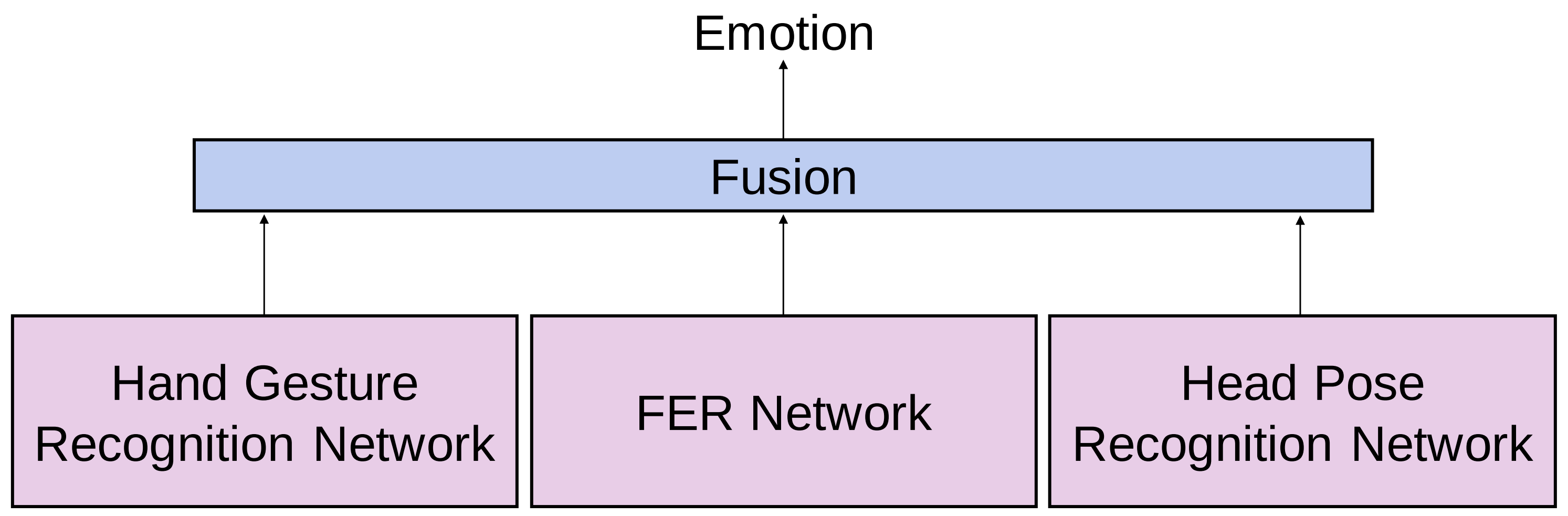
- Our model achieves comparable results with previous SOTA methods on the validation set of the Aff-Wild2 dataset without using any extra data.
- Our method is, to the best of our knowledge, the first to use temporal context for emotion recognition.
- We measure our model's performance using the official evaluation criteria presented in the Aff-Wild2 competition:

$$\epsilon_{total} = 0.67 \times F_1 + 0.33 \times TAcc$$

Method	F1 Score	Accuracy	ABAW2 Metric
Baseline [4]	30	50	36.6
CPIC-FIR2021 [5]	40.2	63	47.7
Netease Fuxi Virtual Human [6]	**75.7**	**85.6**	**79**
Ours	<u>64.3</u>	<u>68.2</u>	<u>65.6</u>

Conclusions & Future Work

- Achieving comparable results on Aff-Wild2 showcases the validity of our model.
- We aim to extend our facial NVC recognition network into a multimodal network for emotion recognition based on nonverbal cues.
- Our end goal is to create a real time NVC recognition aid for the BLV community.
- We are continuously working on the CCNY Dataset to ensure unbiased and balanced representation.



Emotion

Fusion

Hand Gesture Recognition Network

FER Network

Head Pose Recognition Network

References & Acknowledgement

[1] Blindness and vision impairment. https://www.who.int/en/news-room/fact-sheets/detail/blindness-and-visual-impairment.
[2] M. L Knapp, et al. Nonverbal communication in human interaction. Cengage Learning, 2013.
[3] J. Sak-Wernicka. Exploring theory of mind use in blind adults during natural communication. Journal of psycholinguistic research, 2016.
[4] D. Kollias et al. Analysing affective behavior in the second abaw2 competition. ICCV, 2021.
[5] Y. Jin, et al. A multi-modal and multi-task learning method for action unit and expression recognition. arXiv:2107.04187, 2021.
[6] W. Zhang, et al. Prior aided streaming network for multi-task affective recognition at the 2nd abaw2 competition. arXiv:2107.03708, 2021.
[7] L. Jing, er al. Recognizing american sign language manual signs from rgb-d videos. arXiv:1906.02851, 2019.

Acknowledgements: We would like to thank NSF IIS-2041307 for funding this research.