AVCAffe: A Large Scale Audio-Visual Dataset of Cognitive Load and Affect for Remote Work (Supplementary Material)

Pritam Sarkar[1, 2] **Aaron Posen**[1] **Ali Etemad**[1]

[1] Queen's University, Canada [2] Vector Institute

{pritam.sarkar, jordan.posen, ali.etemad}@queensu.ca

https://pritamqu.github.io/AVCAffe

The organization of the supplementary material is as follows:

- Appendix A: Availability;
- Appendix B: Questionnaire;
- Appendix C: Details of Baselines;
- Appendix D: Inter-label relationships;
- Appendix E Broader Impact;
- Appendix F: Additional Representative Frames.

A Availability

This dataset is made freely available for the research community, which can be accessed from the project website. The initial release of the dataset includes:

- Short video segments (average duration of 6 seconds).
- Face-crops corresponding to short video segments.
- Full length videos of each participant per each task (video length of 2.5-10 mins.).
- Self-reported ground truths for affect and cognitive load.
- Outcome of the pre-study questionnaire.
- Dataloader codes for easy and efficient use, written in PyTorch.

B Questionnaire

The questionnaire used to collect the self-reported ground truths are presented in Table S1. Please note, these questions are directly obtained from the original paper where NASA-TLX (Hart 2006) and SAM (Bradley and Lang 1994) are introduced for measuring cognitive load and affect respectively.

C Details of Baselines

To create the baselines, we use the official Pytorch (Paszke et al. 2019) implementations for all the audio visual backbones. The MLP head for the multi-modal fusion networks consist of 2 fully connected layers of hidden dimension 4096 followed by ReLU (Nair and Hinton 2010) activation and dropout (Hinton et al. 2012). Next, to train the networks, we downsample the visual stream at 8 frames per second. Following, the facial crops are extracted using FaceNet (Schroff, Kalenichenko, and Philbin 2015) to effectively classify affect and cognitive load states from the visual streams. Moreover, we resize the frames to a spatial resolution of 112^2 and feed 2 seconds of visual input to the visual encoder with a final input size of $3 \times 16 \times 112^2$. Next, we downsample the audio stream at 16 KHz, and use 2 seconds of audio segments to the audio encoder. We transform the audio segments to mel-spectrograms using 80 mel filters, set the hop size to 10 milliseconds, and use an FFT window length of 1024. Thus, the final audio input dimension becomes 80×200.

Following, we apply standard augmentations on both audio and visual streams during the training. In particular, we apply Multi-scale Crop, Random Horizontal Flip, and Color Jitter on the visual segments. We then simply apply Volume Jitter on the audio waveforms. We train the baseline models with an Adam (Kingma and Ba 2015) optimizer for 20 epochs using a warm-up multi-step learning rate scheduler with a batch size of 64. Moreover, to tackle overfitting, we apply weight decay, dropout, and early stopping. To provide a wide range of baselines using different backbones, we sweep a range of hyper parameters and report the performance of the best models. Specifically, we try with learning rates {0.00001, 0.00003, 0.00007, 0.00005, 0.0001}, learning rate decays {0.1, 0.5, 0.7}, learning rate milestones {(5, 10), (5, 15)}, dropouts {0.0, 0.5}, and weight decays {0.0, 1e-4, 1e-5, 1e-6}. The uni-modal variants are trained on a single NVIDIA RTX6000 24 GB GPU, whereas the multi-modal variants are trained using 2 GPUs in parallel.

D Inter-label relationships

As discussed earlier, we explore the relationships between five output categories of affect and cognitive load attributes namely arousal, valence, mental demand, effort, and temporal demand. To provide a quantitative analysis between these output categories, we perform the Pearson Correlation test (Kowalski 1972) using the normalized self-reported scores. The results are presented in Table S2 confirms our earlier qualitative findings, as it shows high correlations between effort and mental demand as well as effort and temporal demand, indicating that with increasing amounts of effort, participants experience higher mental and temporal load or vice-versa. Moreover, our statistical tests do not show any correlations between the cognitive load and affect attributes, which further confirms that our dataset has been able to successfully capture unique information beyond the common

Cognitive Load: Rate your response on a scale of 0-21
Mental Demand: How mentally demanding was the task? **Physical Demand**: How physically demanding was the task? **Temporal Demand**: How rushed was the pace of the task? **Performance**: How successful were you in accomplishing what you were asked to do? **Effort**: How hard did you have to work to accomplish your level of performance? **Frustration**: How insecure, discouraged, irritated, stressed, or annoyed were you?
Affect: Choose the image in each category that best describes your state.

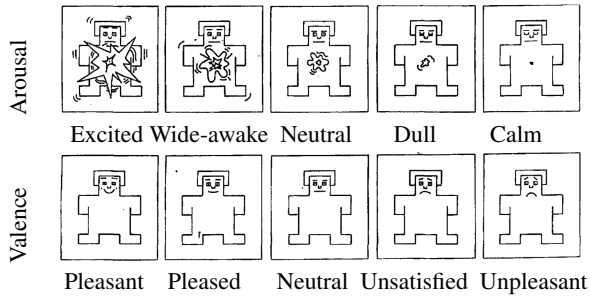

Table S1: Cognitive load and affect questionnaires.

arousal and valence classes.

E Broader Impact

The proposed dataset would be of interest to researchers in both fields of psychology and computer science, to facilitate better understanding of cognitive load, affective states, and broadly human behaviors. The authors do not foresee any negative impacts. We also believe our work is very timely given the rise of remote work as one of the prominent paradigms of work in recent years.

F Additional Representative Frames

We present additional representative frames from different sessions in Figures S1 through S6, showing the diversity of the participant pool in AVCAffe.

Attribute 1	Attribute 2	Correlation
Arousal	Valence	0.321^*
Arousal	Effort	-0.095^*
Arousal	Mental Demand	-0.099^*
Arousal	Temporal Demand	-0.161^*
Valence	Effort	0.179^*
Valence	Mental Demand	0.201^*
Valence	Temporal Demand	0.165^*
Effort	Mental Demand	0.704^*
Effort	Temporal Demand	0.485^*
Mental Demand	Temporal Demand	0.463^*

Table S2: Quantitative analysis on inter-label relationships. We present statistical correlations between different affect and cognitive load attributes. Statistical significance (denoted by *) is considered at $p < 0.01$.

References

Bradley, M. M.; and Lang, P. J. 1994. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1): 49–59.

Hart, S. G. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 50, 904–908. Sage Publications Sage CA: Los Angeles, CA.

Hinton, G. E.; Srivastava, N.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. R. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*.

Kowalski, C. J. 1972. On the effects of non-normality on the distribution of the sample product-moment correlation coefficient. *Journal of the Royal Statistical Society*, 21(1): 1–12.

Nair, V.; and Hinton, G. E. 2010. Rectified linear units improve restricted boltzmann machines. In *International Conference on Machine Learning*.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Neural Information Processing Systems*, 32: 8026–8037.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 815–823.

Figure S1: Sessions of participants from different ethnicity groups. Each row depicts a different session.

Figure S2: Sessions of participants from the same ethnicity groups. Each row depicts a different session.

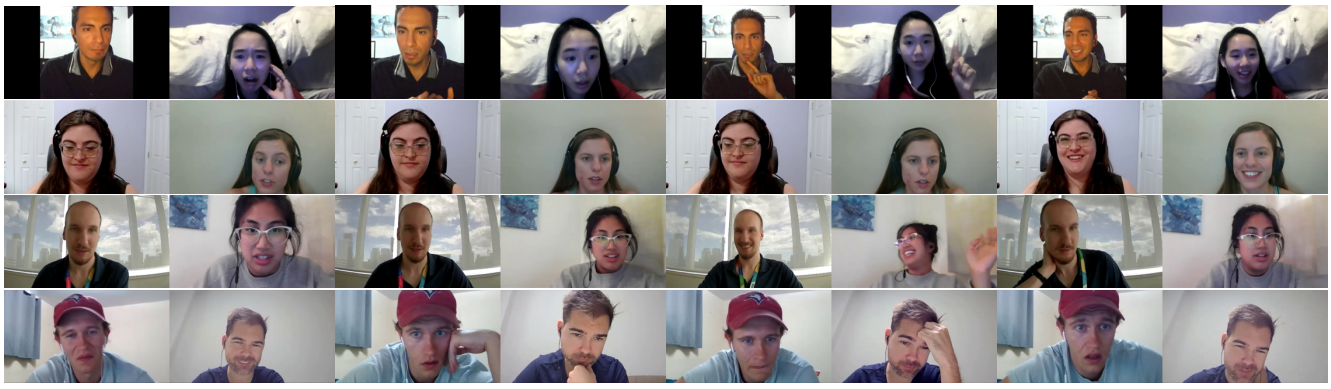

Figure S3: Sessions of participants from different age groups. Each row depicts a different session.

Figure S4: Sessions of participants from the same age groups. Each row depicts a different session.

Figure S5: Sessions of participants of different genders (female-to-male). Each row depicts a different session.

Figure S6: Sessions of participants of the same gender (male-to-male or female-to-female). Each row depicts a different session.