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

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

Abstract

We introduce AVCAffe, the first *A*udio-*V*isual dataset consisting of *C*ognitive load and *Affe*ct attributes. We record AVCAffe by simulating *remote work* scenarios over a video-conferencing platform, where subjects collaborate to complete a number of cognitively engaging tasks. AVCAffe is the largest originally collected (not collected from the Internet) affective dataset in English language. We recruit 106 participants from 18 different countries of origin, spanning an age range of 18 to 57 years old, with a balanced male-female ratio. AVCAffe comprises a total of 108 hours of video, equivalent to more than $58,000$ clips along with task-based self-reported ground truth labels for arousal, valence, and cognitive load attributes such as mental demand, temporal demand, effort, and a few others. We believe AVCAffe would be a challenging benchmark for the deep learning research community given the inherent difficulty of classifying affect and cognitive load in particular. Moreover, our dataset fills an existing timely gap by facilitating the creation of learning systems for better self-management of remote work meetings, and further study of hypotheses regarding the impact of remote work on cognitive load and affective states.

1 Introduction

Remote work, also referred to as 'work from home', has recently become the predominant employment paradigm for many individuals in different sectors. Dictated partially by the recent COVID-19 pandemic, and facilitated through advancements in connectivity, business communication chat platforms, and video-conferencing tools, remote work is the new reality of work for millions across the world. While this new paradigm of work has a number of advantages such as enabling social distancing and flexible hours, it brings about a number of challenges that were less common in in-person work environments. For instance, studies have shown that remote work settings could contribute to increased cognitive load and fatigues in individuals due to the following (Bennett et al. 2021; Fauville et al. 2021; Riedl 2021): (*i*) back-to-back work-related meetings with minimal physical mobility in-between, (*ii*) the inability to effectively perceive and transmit non-verbal expressive cues, (*iii*) the need to apply intense focus on the screen with minimal variation.

Figure 1: Sample representative frames of AVCAffe during different tasks.

In order to better understand and manage the impact of remote work meetings on individuals, it is necessary to design and develop tools capable of quantifying factors such as cognitive load and affect in relevant settings. A key ingredient for developing such systems is the availability of related *datasets* along with *ground-truth information*, which could be used by machine learning and deep learning algorithms for training purposes. Throughout the literature, a large number of *affective computing* datasets (Kossaifi et al. 2019; Busso et al. 2008; Ringeval et al. 2013; Correa et al. 2018; Valstar et al. 2013) have been made available, which can indeed be used to partially address the need in this area by capturing arousal, valence, and other emotion-related factors. Affective computing (Picard 2000), which is an area that aims to investigate methods and algorithms for detection (Sarkar et al. 2019; Kollias et al. 2017; Sarkar and Etemad 2020a,b), quantification (Kollias and Zafeiriou 2020, 2021), and generation of emotions (Kollias and Zafeiriou 2018; Kollias et al. 2020), has witnessed a surge in terms of methods and performances as a direct result of the progress in deep neural networks (Sarkar et al. 2019; Sarkar and Etemad 2020a,b; Kollias et al. 2017; Kollias and Zafeiriou 2020, 2021). Nonetheless, there are currently no available datasets directed toward understanding users in *remote work* settings. Moreover, there are currently no audio-visual datasets that target *cognitive load* along side affective states.

In this paper we attempt to tackle this shortcoming by de-

veloping a large-scale **A**udio-**V**isual Dataset of **C**ognitive Load and **Affe**ct (**AVCAffe**) for remote work. We procure this dataset by designing a study in which participants aim to perform a number of cognitively engaging tasks collaboratively over a video-conferencing platform. These tasks are designed, in consultations with psychologists and related literature, to elicit different cognitive and affective responses with varying intensities. In particular, our study design consists of 7 different tasks, ranging from casual conversations, sharing jokes, finding small and precise differences between two almost similar pictures, decision making, email writing, multi-tasking and a few others. We recruit a total of 106 participants to partake in one-on-one meetings and carry out the designated tasks together. AVCAffe consists of diverse group of participants in terms of age group (18 to 57 years), ethnicity (spread over 18 different countries of origin), profession (engineer, scientist, nurse, student, and lawyer, among several others), daily usage of online communication tools, and so on. Finally, the audio and video are recorded throughout the session, along with self-reported ground truths on a number of cognitive load and affect attributes at the end of each task.

In summary, we make the following contributions:

- We create the first audio-visual dataset that includes '*cognitive load*' attributes. Moreover, to the best of our knowledge, this is the first large-scale audio-visual dataset that focuses on '*remote work*' settings in the context of understanding affect and cognitive load.
- AVCAffe is the largest originally recorded audio-visual affective computing dataset (not obtained from the Internet) in English language. It consists of 106 participants, over 108 hours of videos, and over 58K clips, with self-reported ground truths for affect (*arousal* and *valence*) and cognitive load (*mental demand*, *temporal demand*, *effort*, *performance*, *physical demand*, and *frustration*).
- We perform extensive analyses to validate our study design. We provide extensive deep learning baselines for estimating cognitive load and affect on AVCAffe, in both uni-modal and multi-modal setups.

The details of the dataset availability are further mentioned in Appendix A. We believe AVCAffe would be a valuable and challenging benchmark for the deep learning and affective computing research communities to accurately model cognitive load and affect, especially considering the timely context of remote work. The dataset, codes, and supplementary material are made freely available on the project website[1] to contribute to the field.

2 Related Work

In recent years, there has been a growing interest towards investigating human emotions and affective states across different experimental setups (Martin et al. 2006; Kossaifi et al. 2019; Correa et al. 2018; Abadi et al. 2015; Healey and Picard 2005; Sarkar et al. 2019), such as watching videos for mood elicitation (Correa et al. 2018; Kossaifi et al. 2019; Phinnemore et al. 2021), dyadic conversations (Busso et al. 2016, 2008), and others. In this section, we summarize some of the popular affective audio-visual datasets available in the literature. Additionally, a brief overview of the existing public datasets is presented in Table 1.

In an earlier work, an acted audio-visual dataset named IEMOCAP is created by employing 10 skilled actors in an experimental setup of dyadic interactions (Busso et al. 2008). Similar to IEMOCAP, MSP-IMPROV (Busso et al. 2016) is an acted audio-visual database consisting of 12 participants' audio-visual recordings, focused on understanding emotional behaviours during dyadic conversations. Other popular affective datasets include SEMAINE (McKeown et al. 2011) and AVEC-13 (Valstar et al. 2013) which are collected in human-machine interaction experiment setups. SEMAINE consists of a total of 959 conversations between a 'human' and an 'operator' targeting 7 basic emotional states, while AVEC-13 (Valstar et al. 2013) consists of audio-visual recordings and self-assessed subjective depression scores from 292 subjects.

In addition to purely affective audio-visual datasets, there are a few affective datasets such as MAHNOB-HCI (Soleymani et al. 2011), DECAF (Abadi et al. 2015), and AMIGOS (Correa et al. 2018) which are comprised of physiological signals (electrocardiogram, galvanic skin response, electroencephalogram, and others) along with the participants' audio-visual recordings. MAHNOB-HCI (Soleymani et al. 2011) and DECAF (Abadi et al. 2015) are collected in an experimental setup where participants are asked to watch videos in order to elicit affect states. AMIGOS (Correa et al. 2018) explores understanding of people's emotions, personality, and mood while in groups, as well as in individual settings. In a recent work, K-EmoCon (Park et al. 2020), recordings are acquired from a total of 32 subjects who participated in debates on a social issue. Finally, SWEA (Kossaifi et al. 2019) is an affective dataset of spontaneous behaviors where participants are asked to watch video clips in order to elicit their mental states which is followed by a discussion on the watched clips.

For the sake of completeness, we further briefly mention some of the popular affective datasets Aff-Wild (Zafeiriou et al. 2017), CMU-MOSI (Zadeh et al. 2018), Liris-Accede (Baveye et al. 2015), which are created by scraping videos from YouTube or other similar sources, unlike the datasets discussed earlier which are mostly collected in laboratory setups. Aff-Wild consists of a total of 298 video clips of 200 unique subjects, with a total duration of 30 hours of data. CMU-MOSI is a collection of $23,500$ sentence utterance videos from more than 1000 YouTube speakers, with a total duration of approx 66 hours of data. On the other hand, Liris-Accede consists of a total of $9,800$ movie excerpts (8 to 12 seconds long), with a total duration of approximately 27 hours of audio-visual content.

Distinctions from our work. We find major differences between AVCAffe and earlier works. First, none of the prior works study 'cognitive load' in audio-visual modalities. To the best of our knowledge, AVCAffe is the first audio-visual dataset with cognitive load annotations. Specifically, we cast the problem in the context of 'remote work', which has been largely under-served despite its relevance and timeliness in recent times. Moreover, our dataset captures data in scenar-

[1] https://pritamqu.github.io/AVCAffe

Database	#Sub	Annotations			Size (hrs)	Elicitation
		V	A	CL		
IEMOCAP (2008)	10	✓	✓	✗	12	Dyadic conversation
MAHNOB-HCI (2011)	27	✓	✓	✗	11	Watching videos
SEMAINE (2011)	150	✓	✓	✗	80	Human machine interaction
DECAF (2015)	30	✓	✓	✗	N/A	Watching movies and music videos
MSP-IMPROV (2016)	12	✓	✗	✗	18	Dyadic conversation
HUMAINE (2007)	4	✓	✓	✗	4	Natural and induced conversations
RECOLA (2013)	46	✓	✓	✗	3.5	Online dyadic interactions
SEWA (2019)	398	✓	✓	✗	44	Watching videos and discussion
AMIGOS (2018)	40	✓	✓	✗	N/A	Watching videos and conversation
K-EmoCon (2020)	32	✓	✓	✗	4	Debates
Aff-Wild (2017)	200	✓	✓	✗	30	Collected from YouTube
MOSEI (2018)	1000	✓	✓	✗	66	Collected from YouTube
AVCAffe	**106**	✓	✓	✓	**108**	**Remote work**

Table 1: A brief summary of existing public datasets are presented. Here, V: Valence, A: Arousal, CL: Cognitive Load.

ios that are much closer to in-the-wild settings than pre-planned laboratory setups. Participants frequently exhibit 'spontaneous' behaviour and work 'collaboratively', making AVCaffe unique amongst other datasets in the field.

Amongst all the prior works, we find RECOLA (Ringeval et al. 2013) to be slightly close to our approach. It attempts to investigate human emotions during online dyadic conversations, where participants perform a survival tasks in a given time period. We would like to highlight that unlike RECOLA, which consists of just one task, the experiment design of AVCAffe consists of several tasks with varied levels of difficulty to create a more realistic remote work environment. Moreover, during data collection of RECOLA, video clips are shown to participants to manipulate the states of the participants, whereas we do not perform such steps and rather aim to elicit spontaneous behaviors by the participants. Additionally, AVCAffe (108 hours) is a much larger dataset than RECOLA (3.5 hours). Lastly, we focus on investigating both affect and cognitive states, whereas RECOLA only consists of arousal and valence.

3 AVCAffe

3.1 Study Design

We conduct this study in an online setup, where participants and researchers join the data collection sessions through the Zoom video conferencing platform. We recruit participants for our study based on two inclusion criteria: (*i*) being within the standard age range for employment, i.e., 18 to 60 years old; (*ii*) the ability to converse in English. Additionally, we request participants not to consume alcohol, marijuana, or other substances that may severely alter their affective/cognitive states in the 12-hour window leading to the study or during the study. To capture facial expressions or visual appearance of the participants, we recommend that participants use a computer with a webcam and a stable Internet connection. Additionally, we ask participants not to use any filters or virtual backgrounds during the sessions as it may deteriorate the video quality as well as alter the facial expressions. To collect high quality audio recordings, we recommend that participants use a headphone and avoid using any voice modulation software during the session.

Each session consists of 2 participants, and at-least 1 individual from the research team to moderate and facilitate the session. At the beginning of each session, we play an introductory video for the participants describing an overview and goal of the study. The introductory video includes short descriptions of all the tasks, a brief definition of arousal, valence, and cognitive load, as well as a few additional guidelines regarding the setup as described earlier. Next, participants are asked to fill out a pre-study questionnaire to gather some additional information about the participant pool. Questions include basic demographic information (age, sex, ethnicity), profession, and basic work-related information such as number of working hours per day, number of hours working on a computer, most frequently used mode of communication at work (video, audio, or text), and so on.

In each session, participants complete a number of tasks. One participant is randomly assigned as 'Participant A' and the other as 'Participant B'. Each session typically lasts between 75 to 90 minutes. We record the audio-video of each participant throughout the session. The tasks are implemented using a web-based application, Qualtrics (Qualtrics 2022). At the end of each task, we collect the self-reported ground truths to record the participants' affect and cognitive load (details are provided in Section 3.3).

To conduct this study, we have secured ethics approval from the General Research Ethics Board at Queen's University, Canada. The collected data are stored on secure servers, and each record is identified using an alphanumeric code. Personal information such as first and last name, email address, and others, have been discarded. Participants' consent has been collected using a secure web-based form. We seek

2 types of approval from the participants: (*i*) approval to participate in the study, which should be 'Yes' in order for the study to proceed; (*ii*) approval to use the participants' image/video/audio in articles, publications, or accompanied media content, where participants could opt for 'No', but still participate in our study. Note that participants' photos used in this manuscript are only taken from those participants who have explicitly approved the use of their images in publications.

3.2 Task Design

Our goal is to design a study protocol that closely resemble remote work and meetings. To facilitate this concept, we devise a series of tasks with varied levels of difficulty, eliciting cognitive load and affect in a controlled experimental setup. Our study design requires 2 participants to collaborate and communicate over a video conferencing platform to successfully complete the tasks. Each session is composed of 7 tasks, which are designed for a total duration of approximately 1 hour. A brief description of each task is presented below. Further discussions on the rationale behind choosing each task is discussed later in Section 3.4.

A. Open discussion. In this task, participants discuss a topic of their choice. We provide a set of non-personal potential topics along with a few suggested discussion points for each topic. However, other topics are allowed to be chosen upon agreement between both the participants. The suggested topics include, (*i*) movies and TV series, (*ii*) games or sports, (*iii*) books, (*iv*) work, (*v*) hobbies, passions, or creative interests. A duration of 7.5 minutes is allocated for this task.

B. Lighten the mood. In this stage, participants are asked to share some interesting or humorous incidents, or alternatively tell jokes to the other participant. We provide a few pre-written jokes to each participant in case they can't think of any. Similar to *open discussion*, 7.5 minutes is allocated for this task.

C. Diapix. In this task participants are given two highly similar pictures (Baker and Hazan 2011), in which they need to find and mark the differences. Each participant is only given one of the two pictures and they have to work together to find the differences between the pair of pictures through verbal communication. Participants are asked to find a total of 10 differences and mark them in under 10 minutes.

D. Montclair map. During this task, participants complete a map-matching task (Pardo et al. 2019) with one another. The two participants receive nearly identical maps that have paths drawn on them from a starting point around/through various landmarks leading to a finish point. Without seeing each other's maps, the participants need to communicate with each other to locate the missing landmarks on both maps. This task needs to be completed in under 2.5 minutes, and 2 of such tasks need to solved consecutively.

E. Lost at sea. In this stage, participants are presented with a hypothetical situation (Knox 2009) where participants are on a chartered yacht with friends, crew, and a captain, for a holiday trip. Due to unfortunate circumstances, a fierce fire breaks out on the ship and all the crew members and the captain are lost. The task of the participants is to prioritize 15 items which need to be saved for their survival in the middle of the sea. The items need to be selected unanimously and ranked under 10 minutes.

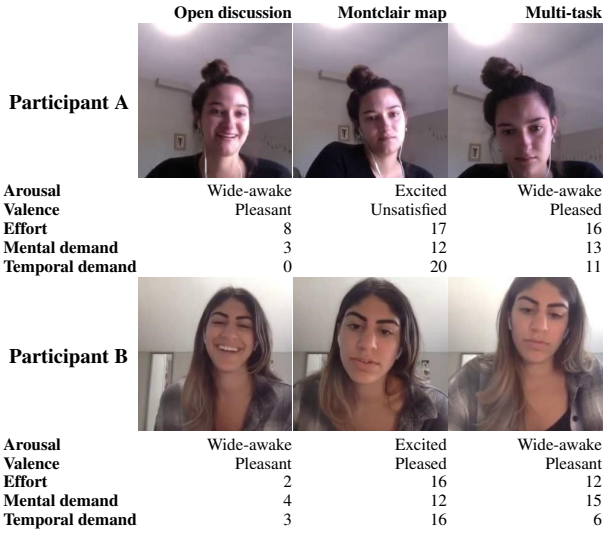

Figure 2: Sample clips along with self-reported affect and cognitive load scores during different task.

F. Reading comprehension. In this task, one participant (active) reads a given a short passage from (Grammar-Bank 2009), followed by answering a few related questions asked by the other participant (passive). The roles of the 2 participants are then reversed. The passages and questions are both provided by us. 5 minutes is allocated to each round, where 2 minutes are to be used for reading and 3 minutes are allocated for answering.

G. Multi-task. The last component of the session is a multi-task scenario, as often encountered in our day-to-day work. In this task, both participants write an email based on a few given points. They are both tasked with interrupting each other by asking a few questions unrelated to the email. The questions are provided to the participants at regular intervals using popup messages. The interrupting questions are simple arithmetic problems (e.g.,, summation, multiplication, or division of 2 numbers), as well as synonyms/antonyms questions related to common English words.

3.3 Ground Truths

We use the NASA Task Load Index (NASA-TLX) (Hart 2006) and Self Assessment Manikin (SAM) (Bradley and Lang 1994) scales to collect cognitive load and affect scores, respectively. Following the protocols laid out in (Hart 2006), cognitive load scores are collected on a scale of 0-21, across mental demand, physical demand, temporal demand, performance, effort, and frustration categories. Additionally, participants report their arousal and valence scores (Bradley and Lang 1994) on a scale of 0-4. For the sake of completeness, we present the questionnaire used to collect the responses in the Appendix B. We present sample ground truths reported by two of the participants in Figure 2.

We note that it has become common practice (Zadeh et al. 2018; Kossaifi et al. 2019; Baveye et al. 2015) to annotate

arousal and valence through 'external annotators' either via crowdsourcing or experts, especially when there is no access to self-reported ground truths, e.g., when videos are scraped from the Internet or movie excerpts. However, we find two issues with this approach. First, cognitive load attributes such as mental demand, effort, and temporal demand, are highly complex mental states, and we could not come across any literature to support that cognitive load can be annotated through external annotators, at least to a reasonable degree of confidence. Second, it has been pointed out (Baveye et al. 2015; Zeng, Shan, and Chen 2018) that this approach suffers from poor intra-annotator (ratings from the same annotator) as well as inter-annotator (ratings from different annotators) inconsistency due to the subjective nature of *emotion*. Therefore, we intentionally collect and rely on self-reported scores for both the affect and cognitive load ground truth values, and do not employ external annotators to annotate any part of the data.

3.4 Design Considerations

We include the *open discussion* and *lighten the mood* activities to induce positive emotions and enable participants to become comfortable with each other, which in turn could enable more effective collaborations. We carefully choose different tasks to target varying levels of cognitive and affective states at different stages of the experiment. For instance the *Montclair map* task is specifically aimed at higher temporal demand given the short amount of allotted time, while the *reading comprehension* and *multi-task* activities are included to induce higher mental demand. Our analysis provided in Section 4.1 demonstrates that various types and levels of cognitive and affective states have indeed been induced at different stages of the experiment. It should be noted that the tasks appear in the same order as mentioned in Section 3.2 for all the participants to induce incremental cognitive load in a controlled manner. We therefore do not shuffle the order of the tasks for different participant pairs.

As mentioned earlier, we collect self-reported ground truths from the participants at the end of each task. We note that this setup results in sparse annotations for the corresponding audio-visual recordings. A potential alternative would have been to collect ground truth values at shorter interval, in between tasks. However, we noticed in our dry runs that more frequent interruptions would be too disruptive to the flow of tasks and distract the participants from the actual problems designed to be solved.

3.5 Data Pre-processing

First, using the raw videos collected at the end of each session, we prepare a separate video recording of each participant for each task. These *long videos* are typically between 2.5 minutes to 10 minutes in length, based on the duration of the tasks. To locate affect and cognitive load attributes at a more temporally granular scale, we further segment the *long videos* into *short video segments* with average duration of 6 seconds based on the speaker's utterances, using the silence detection algorithm (Pydub 2022). Finally, we resize the frame resolution to their shorter side at 256 pixels. We set the video sampling rate at 25 FPS and the audio sampling rate at 44.1 kHz.

# Subjects. 106	**Duration.** 108 hrs. **# Clips.** 58,112
Video. 456×256 at 25fps; **Audio.** 44.1KHz.	
Gender. Male: 52, Female: 53, Non-Binary: 1.	
Age: 18 to 20 : 8; 21 to 30 : 75; 31 to 40 : 17; 41 to 50 : 2; 51 to 60 : 4.	
Countries of origin: Bangladesh(1), Brazil(2), Canada(67), China(3), Ecuador(1), Egypt(1), Germany(1), Hong Kong(1), India(11), Iran(4), Ireland(1), Jordan(1), Mexico(4), Nigeria(2), Pakistan(2), Sweden(1), USA(2), Vietnam(1)	
Ground truths. Arousal, Valence, Mental Demand, Temporal Demand, Effort, Physical Demand, Performance, and Frustration	

Table 2: Key statistics of AVCAffe.

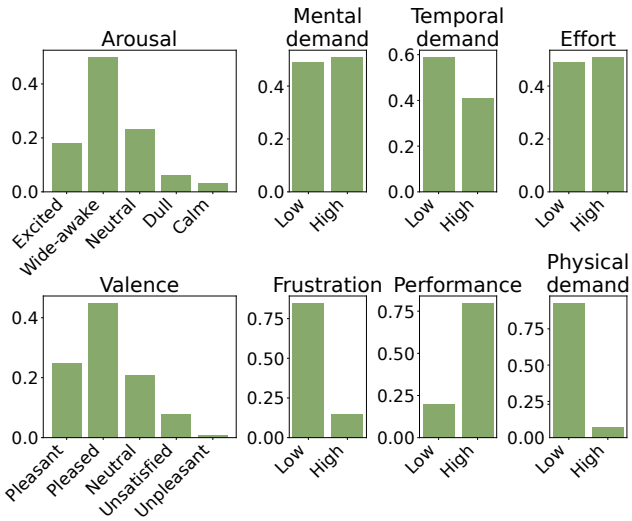

Figure 3: Class distributions of AVCAffe.

3.6 Data Split

We separate the samples into train and validation/test splits maintaining even distributions of age, gender, and ethnicity. We set aside 20 participants for validation and rest of the 86 participants are used for training. Moreover, to avoid any information leakage between the train and validation splits, participants from the same session always reside in the same split, (either in train or validation).

3.7 Statistics

In this study, we recruit a total of 106 participants, spread over 18 different countries. Our participant pool consists of 49% male, 50% female, and 1% non-binary participants. Out of the total participants, around 60% are from North America, and the rest of the population belongs to India, Iran, and several other countries. Additionally, the data population is also spread over a wide range of age groups, specifically between 18-57 years. Lastly, AVCAffe consists of approximately 58K video clips, equivalent to 108 hours of video

recordings and their corresponding affect and cognitive load labels. The key statistics are highlighted in Table 2.

We present the class distributions of all the ground truth labels in Figure 3. We notice that 'wide-awake' and 'pleased' which are the second-highest choices of arousal and valence respectively, are the dominating selections. We observe more or less balanced distributions for mental demand, temporal demand, and effort. However, in case of frustration and physical demand, the majority ($\geq 80\%$) of the videos are categorized as 'low', which is expected as our tasks are neither meant for high physical demand nor meant to elicit high levels of frustration. Similarly, we find 'high' as the dominating class for performance, which is because participants are mostly successful in finishing these tasks.

4 Analysis and Evaluation

4.1 Analysis

We perform an in-depth analysis of self-reported ground truths and present the results in Figures 4 and 5. Our analysis helps us validate the success of our study design, indicating that different tasks are able to induce varying degrees of cognitive and affective states amongst the participants throughout the session. Our detailed findings are as follows.

Cognitive Load. While analyzing cognitive states, we present the density plots of each cognitive load attribute such as, mental demand, effort, and temporal demand in Figure 4. We find distinct shifts in cognitive load over time, specifically for *mental demand*, *effort*, and *temporal demand*. We notice that the participants' *mental demand* is fairly low during the *open discussion* and *lighten the mood*. On the other hand, participants show higher mental demand during tasks like *multi-task* and *reading comprehension (active)*. Moreover, participants show moderate mental demand during other tasks like *diapix*, *montclair map*, and *lost at the sea*. It should be noted that, these outcomes are completely in line with our intended design considerations. Next, while analysing temporal demand, we find that participants experience higher temporal demand during *montclair map*; moderate temporal demand during *multi-task* and *reading comprehension*; and low temporal demand during tasks like *open discussion* and *lighten the mood*. Participants report low effort required to complete tasks like *open discussion* and *reading comprehension (passive)*, whereas high effort scores are reported for *multi-task* and *reading comprehension (active)*. In case of other cognitive load attributes such as *physical demand* and *frustration*, we do not notice any considerable shifts across different tasks, and participants always report low score on these attributes. These findings coincide with our intended design as our study is not meant to evoke *frustration*. Moreover, due to the fact that all of these tasks are computer-based, minimal *physical demand* is required. Lastly, in case of *performance*, we notice minimal shift in self-reported scores across different tasks. On average, participants report moderate to high in terms of successful completion of the tasks, i.e., *performance*.

Affect. To analyse the self-reported affect scores, we project arousal and valence responses onto a 3D valence-arousal space, presented in Figure 4. We notice considerable shifts in self-reported responses across different stages of the experiment. For example, at the beginning of the sessions, i.e., the *Prestudy*, the majority of the participants report 'neutral' in terms of both arousal and valence. Next, in case of *Open discussion* and *Lighten the mood*, a clear shift in arousal and valence is noticed from the center to the first quadrant of the valence-arousal space. Specifically during *Lighten the mood*, participants report high arousal ('excited') and high valence ('pleasant'). Interestingly, during the next task (*Diapix*), we notice a downward shift for both arousal and valence, i.e., 'excited' to 'wide-awake' for arousal, and 'pleasant' to 'pleased' for valence. Moreover, during *Montclair map*, the majority of the responses remain the same as the previous task. However, we notice some participants experience 'neutral' in the case of arousal, and 'unsatisfied' in the case of valence. Similarly, during *reading comprehension (active)* and *multi-task*, we observe further shifts in affect response towards the third ('dull' and 'unsatisfied') and second ('wide-awake' and 'unsatisfied') quadrants of the valence-arousal space. Such prominent shifts in valence-arousal space during different tasks indicate that our study design successfully targets different affective states at different stages of the session.

Inter-label relationships. Lastly, we aim to explore the relationships between the prominent affect and cognitive load attributes, namely arousal, valence, mental demand, effort, and temporal demand. To investigate this, we project the normalized self-reported ground-truths onto a 3D space as shown in Figure 5. We observe a strong positive correlation between effort and mental demand, indicating that with increasing amounts of effort, participants experience higher mental demand, and vice-versa. Additionally, some degree of positive correlation is noticed between temporal demand and effort, as well as between temporal demand and mental demand. Interestingly, we do not observe strong correlations between cognitive load and affect attributes. This indicates that our dataset has been able to successfully capture unique information beyond the arousal and valence classes. We further provide a quantitative analysis in the Appendix D.

4.2 Benchmark Evaluation

Training. We present exhaustive deep-learning baselines on AVCAffe in both uni-modal and multi-modal setups. We use a wide variety of architectures as video and audio backbones. In particular, we use R(2+1)D (Tran et al. 2018), ResNet3D (Hara, Kataoka, and Satoh 2017), and MC3 (Tran et al. 2018) as the video backbones. Additionally, ResNet (He et al. 2016) and VGG (Simonyan and Zisserman 2014) are used as the audio backbones. In case of uni-modal training, we use the embeddings obtained from the backbones and pass them through a fully-connected layer. In case of multi-modal training, we perform late fusion (Zhang et al. 2016) by concatenating the embeddings obtained from audio and video backbones, followed by passing them through an MLP head. As mentioned earlier, the self-reported arousal and valence scores are originally collected on 5-point scales which correspond to 5 distinct classes. Therefore, we for-

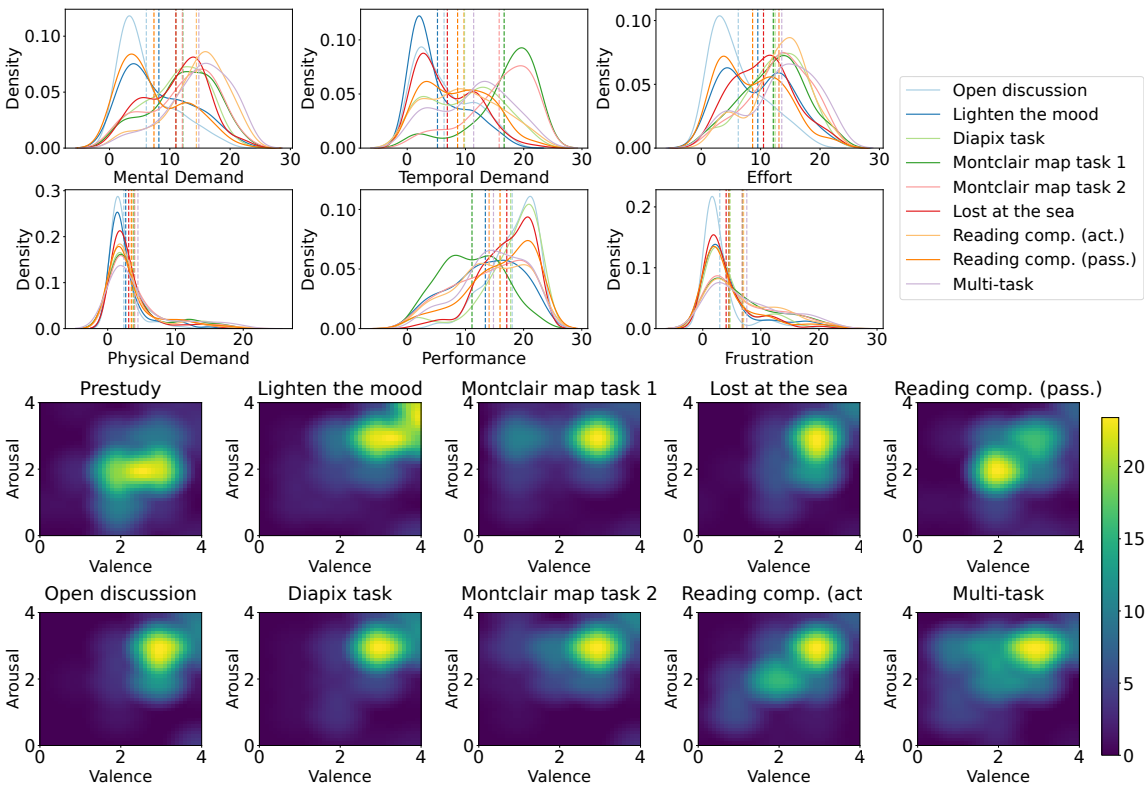

Figure 4: Top: We present the density plots of self-reported cognitive load scores, each color refers to an individual task. Bottom: We present the affect scores projected in a 3-d plane, where the colors denote population density, yellow being the most dense.

mulate this problem as a multi-class classification task. The self-reported scores for cognitive load are originally collected on a scale of 0-21. We empirically find it quite challenging to accurately model such level of granularity for cognitive load due to it's inherent complexity. Hence, to make the task simpler, we formulate this problem as binary classification. Following the NASA-TLX scale (Hart 2006), the self-reported cognitive scores greater than 10 are marked as High and less than or equals to 10 are marked as Low. Moreover, to create the baselines, we do not use *frustration*, *physical demand*, and *performance*, as these three attributes do not show enough variance in our collected dataset (see our discussion in Section 4.1). Next, to train the framework, we feed 2 seconds of audio visual segments to the networks. In particular, for audio segments we generate mel-spectrograms and for visual segments we obtain the face crops before feeding to the network. We use cross-entropy loss to train the network, specifically binary cross-entropy loss for cognitive load and categorical cross-entropy for affect attributes. Additional details on training are provided in the Appendix C.

Evaluation Protocol. We evaluate AVCAffe on a total of 5 affect (arousal and valence) and cognitive load (mental demand, effort, and temporal demand) attributes. We perform multi-class and binary classification for affect and cognitive load respectively. We evaluate the models at two levels, (*i*) *short video segments* (duration of 6 seconds) and (*ii*) *long videos* (duration of 2 to 10 minutes). To evaluate on *short video segments*, we uniformly sample 3 clips per sample and report the F1-score by averaging the clip level predictions. Next, to obtain the prediction correspond to *long videos* we average the predictions correspond to the *short video segments*. Following a standard practice (Busso et al. 2008; Kossaifi et al. 2019; Zafeiriou et al. 2017), we use the weighted F1-measures as the evaluation metric because of it's robustness towards imbalanced class distribution. Please note, we do not apply any augmentations during validation.

Results. The performances of the baselines are presented in Table 3. We notice that in overall multi-modal networks outperform the uni-modal variants in both *short video segments* and *long videos* on all the attributes except *Effort* on *short video segments*. In particular, we notice that visual features work considerably well in predicting *Effort* on *short video segments* and even outperform the multi-modal variants. Additionally, while comparing between audio-only vs. visual-only, we find that in most of the cases visual-only models perform better. We empirically find that using the face-crops instead of the whole frames is helpful to extract better representations from the visual streams. Additionally, we notice that our multi-modal baselines perform relatively better in classifying *affect* attributes on the *short video segments* in comparison to the *long videos*. For example, the

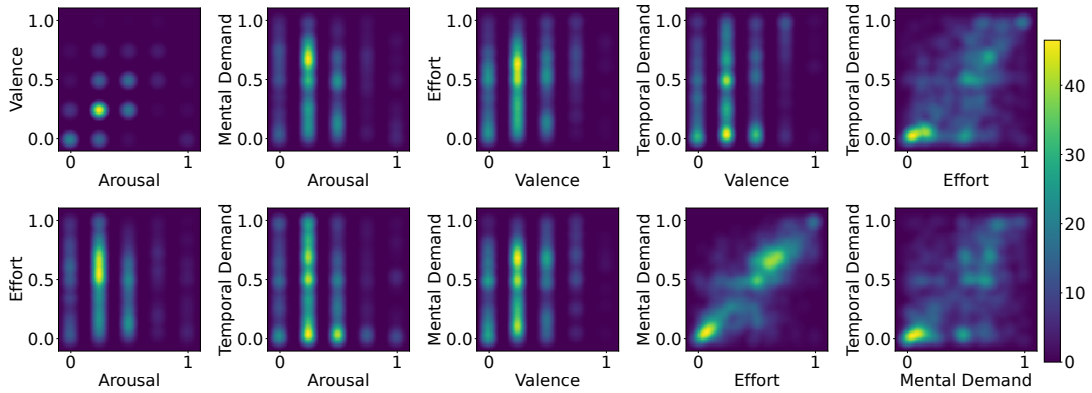

Figure 5: Qualitative analysis on inter-label relationships.

Audio	Visual	#Params	Mental D.		Effort		Temporal D.		Arousal		Valence	
			Short	Long	Short	Long	Short	Long	Short	Long	Short	Long
Random Classifier		-	51.3	45.6	48.0	43.6	35.1	33.9	32.9	26.2	34.0	30.3
VGG-16	✗	14.7M	58.8	61.2	58.8	62.1	57.9	56.7	38.3	36.1	40.3	39.1
ResNet-18	✗	11.2M	58.2	60.7	57.0	60.8	58.2	54.4	38.1	30.4	39.3	36.3
✗	MC3-18	11.7M	60.4	61.0	61.4	63.8	60.0	59.4	41.4	34.0	42.0	38.8
✗	ResNet3D-18	33.4M	59.3	59.0	61.0	62.7	58.5	57.9	37.8	30.9	41.9	39.5
✗	R(2+1)D-18	31.5M	60.5	59.6	**65.5**	67.7	59.6	54.9	39.7	33.3	38.7	34.9
VGG-16	MC3-18	47.4M	59.4	60.2	59.7	66.2	60.8	61.4	41.3	38.9	40.3	**41.7**
VGG-16	ResNet3D-18	69.1M	**65.0**	**65.6**	59.7	60.5	59.2	60.3	40.7	37.3	**43.9**	39.4
VGG-16	R(2+1)D-18	67.2M	60.1	64.7	59.7	**69.4**	60.4	**66.7**	42.1	**40.5**	41.1	39.5
ResNet-18	MC3-18	43.9M	61.3	60.2	59.4	62.1	58.8	57.7	42.4	36.0	41.4	39.2
ResNet-18	ResNet3D-18	65.6M	58.8	61.2	60.7	64.4	**61.2**	61.7	42.6	35.1	39.8	39.1
ResNet-18	R(2+1D)-18	63.7M	60.4	62.7	60.8	61.1	58.6	59.0	**44.0**	39.5	40.9	37.7

Table 3: Baselines on AVCAffe are presented. The best F1-scores in each subcategory (audio-only/visual-only/audio-visual) are underlined and best scores of each label are highlighted in bold. Here, Random Classifier refers to a randomly initialized classifier with no training which serves as a reference point to understand the performance of different models.

best multi-modal scores achieved on arousal and valence are 44.0 and 43.9 on *short video segments* and 40.5 and 41.7 on *long videos*. However, a different trend is noticed on the *cognitive load* attributes. For example, our best model on *temporal demand* achieves F1-scores of 66.7 on *long videos* and 61.2 on *short video segments*. Similar trend is also noticed in the case of Effort. This interesting finding may open the door to future lines of inquiry into the differences between affect and cognitive load both from a human perception and from an affective computing perspective.

5 Summary

We present a novel audio-visual database of cognitive load and affect collected in a setup resembling 'remote work'. To the best of our knowledge AVCAffe is the first audio-visual dataset comprised of *cognitive load* annotations. Moreover, AVCAffe is the largest affective computing dataset in English language. Additionally, we perform extensive analyses utilizing the self-reported ground-truths, which reveal interesting insights on the cognitive and affective states of participants during the experiment sessions. Given the sparsity of the annotations and the challenging nature of estimating cognitive load and affect, we introduce an interesting challenge to the deep learning research community. Furthermore, we present extensive baselines to provide benchmarks for future works. We believe AVCAffe would be a useful and challenging dataset for the deep learning community.

Acknowledgments

We are grateful to Bank of Montreal and Mitacs for funding this research. We are thankful to SciNet HPC Consortium for helping with the computation resources. We thank Shuvendu Roy, Dept. of Electrical and Computer Engineering, at Queen's University for his collaboration during this study. We would like to further thank Prof. Kevin Munhall, Dept. of Psychology, at Queen's University for his valuable discussions at the study design stage.

References

Abadi, M. K.; Subramanian, R.; Kia, S. M.; Avesani, P.; Patras, I.; and Sebe, N. 2015. DECAF: MEG-based multimodal database for decoding affective physiological re-

sponses. *Transactions on Affective Computing*, 6(3): 209–222.

Baker, R.; and Hazan, V. 2011. DiapixUK: task materials for the elicitation of multiple spontaneous speech dialogs. *Behavior Research Methods*, 43(3): 761–770.

Baveye, Y.; Dellandrea, E.; Chamaret, C.; and Chen, L. 2015. LIRIS-ACCEDE: A video database for affective content analysis. *Transactions on Affective Computing*, 6(1): 43–55.

Bennett, A. A.; Campion, E. D.; Keeler, K. R.; and Keener, S. K. 2021. Videoconference fatigue? Exploring changes in fatigue after videoconference meetings during COVID-19. *Journal of Applied Psychology*, 106(3): 330.

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.

Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J. N.; Lee, S.; and Narayanan, S. S. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4): 335–359.

Busso, C.; Parthasarathy, S.; Burmania, A.; AbdelWahab, M.; Sadoughi, N.; and Provost, E. M. 2016. MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception. *Transactions on Affective Computing*, 8(1): 67–80.

Correa, J. A. M.; Abadi, M. K.; Sebe, N.; and Patras, I. 2018. Amigos: A dataset for affect, personality and mood research on individuals and groups. *Transactions on Affective Computing*.

Douglas-Cowie, E.; Cowie, R.; Sneddon, I.; Cox, C.; Lowry, O.; Mcrorie, M.; Martin, J.-C.; Devillers, L.; Abrilian, S.; Batliner, A.; et al. 2007. The HUMAINE database: Addressing the collection and annotation of naturalistic and induced emotional data. In *International Conference on Affective Computing and Intelligent Interaction*, 488–500. Springer.

Fauville, G.; Luo, M.; Queiroz, A. C.; Bailenson, J. N.; and Hancock, J. 2021. Zoom exhaustion & fatigue scale. *Computers in Human Behavior Reports*, 4: 100119.

Grammar-Bank. 2009. Short Reading Comprehension Passages. https://www.grammarbank.com/short-reading-comprehension-passages.html. Accessed: 2023-02-22.

Hara, K.; Kataoka, H.; and Satoh, Y. 2017. Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet? *arXiv preprint*, arXiv:1711.09577.

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.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition*, 770–778.

Healey, J.; and Picard, R. 2005. Detecting stress during real-world driving tasks using physiological sensors. *Transactions on Intelligent Transportation Systems*, 6(2): 156–166.

Knox, G. 2009. Lost at sea. https://insight.typepad.co.uk/insight/2009/02/lost-at-sea-a-team-building-game.html. Accessed: 2023-02-22.

Kollias, D.; Cheng, S.; Ververas, E.; Kotsia, I.; and Zafeiriou, S. 2020. Deep neural network augmentation: Generating faces for affect analysis. *International Journal of Computer Vision*, 128(5): 1455–1484.

Kollias, D.; Nicolaou, M. A.; Kotsia, I.; Zhao, G.; and Zafeiriou, S. 2017. Recognition of affect in the wild using deep neural networks. In *Computer Vision and Pattern Recognition Workshops*, 26–33.

Kollias, D.; and Zafeiriou, S. 2018. A multi-task learning & generation framework: Valence-arousal, action units & primary expressions. *arXiv preprint arXiv:1811.07771*.

Kollias, D.; and Zafeiriou, S. 2020. Exploiting multi-cnn features in cnn-rnn based dimensional emotion recognition on the omg in-the-wild dataset. *Transactions on Affective Computing*, 12(3): 595–606.

Kollias, D.; and Zafeiriou, S. 2021. Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework. *arXiv preprint arXiv:2103.15792*.

Kossaifi, J.; Walecki, R.; Panagakis, Y.; Shen, J.; Schmitt, M.; Ringeval, F.; Han, J.; Pandit, V.; Toisoul, A.; Schuller, B. W.; et al. 2019. Sewa db: A rich database for audio-visual emotion and sentiment research in the wild. *Transactions on Pattern Analysis and Machine Intelligence*.

Martin, O.; Kotsia, I.; Macq, B.; and Pitas, I. 2006. The eNTERFACE' 05 Audio-Visual Emotion Database. In *International Conference on Data Engineering Workshops*, 8–8.

McKeown, G.; Valstar, M.; Cowie, R.; Pantic, M.; and Schroder, M. 2011. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *Transactions on Affective Computing*, 3(1): 5–17.

Pardo, J. S.; Urmanche, A.; Gash, H.; Wiener, J.; Mason, N.; Wilman, S.; Francis, K.; and Decker, A. 2019. The Montclair map task: Balance, efficacy, and efficiency in conversational interaction. *Language and Speech*, 62(2): 378–398.

Park, C. Y.; Cha, N.; Kang, S.; Kim, A.; Khandoker, A. H.; Hadjileontiadis, L.; Oh, A.; Jeong, Y.; and Lee, U. 2020. K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations. *Scientific Data*, 7(1): 1–16.

Phinnemore, R.; Cimolino, G.; Sarkar, P.; Etemad, A.; and Graham, T. N. 2021. Happy Driver: Investigating the Effect of Mood on Preferred Style of Driving in Self-Driving Cars. In *International Conference on Human-Agent Interaction*, 139–147.

Picard, R. W. 2000. *Affective computing*. MIT press.

Pydub. 2022. Pydub. https://github.com/jiaaro/pydub. Accessed: 2023-02-22.

Qualtrics. 2022. Qualtrics. https://www.qualtrics.com/. Accessed: 2023-02-22.

Riedl, R. 2021. On the stress potential of videoconferencing: definition and root causes of Zoom fatigue. *Electronic Markets*, 1–25.

Ringeval, F.; Sonderegger, A.; Sauer, J. S.; and Lalanne, D. 2013. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. *International Conference and Workshops on Automatic Face and Gesture Recognition*, 1–8.

Sarkar, P.; and Etemad, A. 2020a. Self-supervised ECG representation learning for emotion recognition. *Transactions on Affective Computing*.

Sarkar, P.; and Etemad, A. 2020b. Self-supervised learning for ecg-based emotion recognition. In *International Conference on Acoustics, Speech and Signal Processing*, 3217–3221. IEEE.

Sarkar, P.; Ross, K.; Ruberto, A. J.; Rodenbura, D.; Hungler, P.; and Etemad, A. 2019. Classification of cognitive load and expertise for adaptive simulation using deep multitask learning. In *International Conference on Affective Computing and Intelligent Interaction*, 1–7. IEEE.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Soleymani, M.; Lichtenauer, J.; Pun, T.; and Pantic, M. 2011. A multimodal database for affect recognition and implicit tagging. *Transactions on Affective Computing*, 3(1): 42–55.

Tran, D.; Wang, H.; Torresani, L.; Ray, J.; LeCun, Y.; and Paluri, M. 2018. A closer look at spatiotemporal convolutions for action recognition. In *Computer Vision and Pattern Recognition*, 6450–6459.

Valstar, M.; Schuller, B.; Smith, K.; Eyben, F.; Jiang, B.; Bilakhia, S.; Schnieder, S.; Cowie, R.; and Pantic, M. 2013. Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In *International Workshop on Audio/Visual Emotion Challenge*, 3–10.

Zadeh, A.; Liang, P. P.; Poria, S.; Vij, P.; Cambria, E.; and Morency, L.-P. 2018. Multi-attention recurrent network for human communication comprehension. In *AAAI Conference on Artificial Intelligence*.

Zafeiriou, S.; Kollias, D.; Nicolaou, M. A.; Papaioannou, A.; Zhao, G.; and Kotsia, I. 2017. Aff-wild: valence and arousal'In-the-Wild'challenge. In *Computer Vision and Pattern Recognition Workshops*, 34–41.

Zeng, J.; Shan, S.; and Chen, X. 2018. Facial expression recognition with inconsistently annotated datasets. In *European conference on computer vision*, 222–237.

Zhang, S.; Zhang, S.; Huang, T.; and Gao, W. 2016. Multimodal deep convolutional neural network for audio-visual emotion recognition. In *International Conference on Multimedia Retrieval*, 281–284.