VEATIC: Video-based Emotion and Affect Tracking in Context Dataset

Zhihang Ren*[1], Jefferson Ortega*[1], Yifan Wang*[1], Zhimin Chen[1], Yunhui Guo[2],
Stella X. Yu[1,3], David Whitney[1]
[1]University of California, Berkeley, [2]University of Texas at Dallas,
[3]University of Michigan, Ann Arbor
[1]{peter.zhren,jefferson_ortega,wyf020803,zhimin,dwhitney}@berkeley.edu,
[2]yunhui.guo@utdallas.edu,[3]stellayu@umich.edu

Abstract

*Human affect recognition has been a significant topic in psychophysics and computer vision. However, the currently published datasets have many limitations. For example, most datasets contain frames that contain only information about facial expressions. Due to the limitations of previous datasets, it is very hard to either understand the mechanisms for affect recognition of humans or generalize well on common cases for computer vision models trained on those datasets. In this work, we introduce a brand new large dataset, the Video-based Emotion and Affect Tracking in Context Dataset (**VEATIC**), that can conquer the limitations of the previous datasets. VEATIC has* 124 *video clips from Hollywood movies, documentaries, and home videos with continuous valence and arousal ratings of each frame via real-time annotation. Along with the dataset, we propose a new computer vision task to infer the affect of the selected character via both context and character information in each video frame. Additionally, we propose a simple model to benchmark this new computer vision task. We also compare the performance of the pretrained model using our dataset with other similar datasets. Experiments show the competing results of our pretrained model via VEATIC, indicating the generalizability of VEATIC. Our dataset is available at https://veatic.github.io.*

1. Introduction

Recognizing human affect is of vital importance in our daily life. We can infer people's feelings and predict their subsequent reactions based on their facial expressions, interactions with other people, and the context of the scene. It is an invaluable part of our communication. Thus, many studies are devoted to understanding the mechanism of affect recognition. With the emergence of Artificial Intelligence (AI), many studies have also proposed algorithms to automatically perceive and interpret human affect, with the potential implication that systems like robots and virtual humans may interact with people in a naturalistic way.

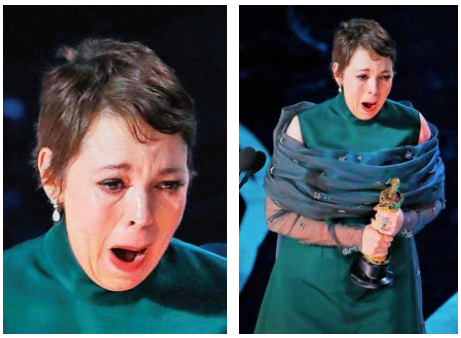

Figure 1. Importance of context in emotion recognition. How does she feel? Look at the woman in picture (a). If you had to guess her emotion, you might say that she is sad or in grief. However, picture (b) reveals the context of the scene allowing us to correctly observe that she is very happy or excited.

When tasked with emotion recognition in the real world, humans have access to much more information than just facial expressions. Despite this, many studies that investigate emotion recognition often use static stimuli of facial expressions that are isolated from context, especially in assessments of psychological disorders [3, 18] and in computer vision models [60, 62]. Additionally, while previous studies continue to investigate the process by which humans perceive emotion, many of these studies fail to probe how emotion recognition is influenced by contextual factors like the visual scene, background information, body movements, other faces, and even our beliefs, desires, and conceptual processing [4, 34, 8, 42, 44]. Interestingly, visual contextual information has been found to be automatically and effortlessly integrated with facial expressions [2]. It can also override facial cues during emotional

*These authors contributed equally to this work.

judgments [26](Figure 1), and can even influence emotion perception at the early stages of visual processing [7]. In fact, contextual information is often just as valuable to understand a person's emotion as the face itself [8, 9, 10]. The growing evidence of the importance of contextual information in emotion recognition [4] demands that researchers reevaluate the experimental paradigms in which they investigate human emotion recognition. For example, to better understand the mechanisms and processes that lead to human emotion recognition during everyday social interactions, the generalizability of research studies should be seriously considered. Most importantly, datasets for emotion and affect tracking should not only contain faces or isolated specific characters, but contextual factors such as background visual scene information, and interactions between characters should also be included.

In order to represent the emotional state of humans, numerous studies in Psychology and Neuroscience have proposed methods to quantify humans' emotional state which include both categorical and continuous models of emotion. The most famous and dominant categorical theory of emotion is the theory of basic emotions which states that certain emotions are universally recognized across cultures (anger, fear, happiness, etc.) and that all emotions differ in their behavioral and physiological response, their appraisal, and in expression [16]. Alternatively, the circumplex model of affect, a continuous model of emotion, proposes that all affective states arise from two neurophysiological systems related to valence and arousal and all emotions can be described by a linear combination of these two dimensions [52, 47, 53]. Valence indicates whether an emotion is negative or positive, while arousal indicates the intensity of an emotion. Another model of emotion recognition, the Facial Action Coding System model, states that all facial expressions can be broken down into the core components of muscle movements called Action Units [17]. Previous emotion recognition models have been built with these different models in mind [61, 63, 41]. However, few models focus on measuring affect using continuous dimensions, an unfortunate product of the dearth of annotated databases available for affective computing.

Based on the aforementioned emotion metrics, many emotion recognition datasets have been developed. Early datasets, such as SAL [15], SEMAINE [39], Belfast induced [58], DEAP [28], and MAHNOB-HCI [59] are collected under highly controlled lab settings and are usually small in data size. These previous datasets lack diversity in terms of characters, motions, scene illumination, and backgrounds. Moreover, the representations in early datasets are usually discrete. Recent datasets, like RECOLA [49], MELD [46], OMG-emotion dataset [5], Aff-Wild [69], and Aff-Wild2 [29, 30], start to collect emotional states via continuous ratings and utilize videos on the internet or called "in-the-wild". However, these datasets lack contextual information and focus solely on facial expressions. The frames are dominated by characters or particular faces. Furthermore, the aforementioned datasets have limited annotators (usually less than 10). As human observers have strong individual differences and suffer from many biases [12, 45, 48], limited annotators can lead to substantial annotation biases.

In this study, we introduce the Video-based Emotion and Affect Tracking in Context Dataset (**VEATIC**, /veˈætɪc/), a large dataset that can be beneficial to both Psychology and computer vision groups. The dataset includes 124 video clips from Hollywood movies, documentaries, and home videos with continuous valence and arousal ratings of each frame via real-time annotation. We also recruited a large number of participants to annotate the data. Based on this dataset, we propose a new computer vision task, i.e., automatically inferring the affect of the selected character via both context and character information in each video frame. In this study, we also provide a simple solution to this task. Experiments show the effectiveness of the method as well as the benefits of the proposed VEATIC dataset. In a nutshell, the main contributions of this work are:

- We build the first large video dataset, **VEATIC**, for emotion and affect tracking that contains both facial features and contextual factors. The dataset has continuous valence and arousal ratings for each frame.
- In order to alleviate the biases from annotators, we recruited a large set of annotators (192 in total) to annotate the dataset compared to previous datasets (usually less than 10).
- We provide a baseline model to predict the arousal and valence of the selected character from each frame using both character information and contextual factors.

2. Related Work

Recently, there have been several datasets that provide frames with both facial and context information, like CAER [33] and EMOTIC [32]. CAER [33] is a video-based dataset that contains categorical labels of each video frame, and EMOTIC [32] is an image-based dataset containing both categorical expression labels and continuous valence-arousal-dominance ratings. Unlike these datasets, our dataset is video-based and contains continuous valence and arousal ratings. A detailed comparison between our dataset with previous datasets can be found in Table 1.

Based on various emotion datasets, studies have started to focus on how to infer emotion automatically. Human affect can be inferred from many modalities, such as audio [70, 68, 65], visual [40, 54, 55, 37], and text [68, 22].

Database	Annotation Type	Condition	# videos	Length of Videos	Total # of Annotators	Context
SAL [15]	Valence-Arousal (Continuous)	Controlled	23	SAL 0: 5min SAL 1: 30min	4	×
SEMAINE [39]	Mixed*	Controlled	\	Total: 6.5hours	6-8	×
SEND [43]	valence	Controlled	193	135s	700	×
Belfast induced [58]	Mixed	Controlled	37	5-60s	6-258	×
MAHNOB-HCI [59]	Mixed	Controlled	20	34.9-117s	50	×
MELD [46]	7 Emotion Categories	In-the-Wild	1,433	3.59s	3	×
OMG Emotion [5]	Mixed	In-the-Wild	567	1min	5	×
RECOLA [49]	Valence-Arousal (Continuous)	Controlled	46	5min	6	×
AFEW [13]	7 Basic Facial Expression	In-the-Wild	1,809	0.3s - 5.4s	3	×
AFEW-VA [31]	Valence-Arousal (Discrete)	In-the-Wild	600	0.5s - 4s	2	×
Aff-Wild [69]	Valence-Arousal (Continuous)	In-the-Wild	298	6s - 14min28s	8	×
Aff-Wild2 [29, 30]	Valence-Arousal (Continuous)	In-the-Wild	260	4s - 15min4s	4	×
AM-FED [38]	12 Action Units	In-the-Wild	242	49.69s	\	×
DEAP [28]	Valence-Arousal -Dominance (Discrete)	Music Videos	120	1min	14-16	✓
CAER [33]	7 Emotion Categories	In-the-Wild	13,201	1s - 5s	6	✓
CAER-S [33]	7 Emotion Categories	In-the-Wild Image-based	\	70,000 images	6	✓
EMOTIC [32]	Mixed	In-the-Wild Image-based	\	18,316 images	3-5	✓
VEATIC (ours)	Valence-Arousal (Continuous)	In-the-Wild	124	10s - 2min37s	192	✓

Table 1. Comparison of the VEATIC dataset with existing emotion recognition datasets. VEATIC contains a large amount of video clips and a long video total duration. It is the first large context-aware emotion recognition video dataset with continuous valence and arousal annotations. VEATIC also has many more annotators compared to other context-aware emotion recognition video datasets. (*: Mixed means containing both continuous and categorical annotations.)

For visual inputs, in particular, there are three major tasks. The valence-arousal estimation task aims to predict the valence and arousal of each image/frame [71, 69, 29, 30]; the expression recognition task focuses on classifying emotional categories of each image/frame [66, 57, 67]; and the action unit (AU) detection task intends to detect facial muscle actions from the faces of each image/frame [25, 56, 35, 64]. Currently, most proposed methods rely highly on the facial area to infer the emotional state. Indeed, the facial area contains rich information about the human emotional state. However, contextual factors also provide essential information that is necessary for humans to correctly infer and perceive the emotional states of others [8, 9, 10]. Several studies [33, 32, 40] have started to incorporate context information as a source of affect inference. In this study, we also adopted both facial and context information to achieve the new task, i.e., to infer the valence and arousal for each video frame.

To infer the affect of a person, we usually need to deal with temporal information of either audio segments, video frames, or words. Many studies [68, 69, 29, 30] started to utilize long short term memory (LSTM) [23], gated recurrent unit (GRU) [11], or recurrent neural network (RNN) [24, 50] to process the temporal information. With the emergence of the visual transformer (ViT) [14], attention has been shifted. Many video understanding tasks [19, 1, 36] have utilized ViT for temporal information understanding and achieving state-of-the-art performance. Our baseline method also adopted ViT as a tool to process the temporal information in video clips.

3. VEATIC Dataset

In this section, we introduce the Video-based Emotion and Affect Tracking in Context Dataset (**VEATIC**). First, we describe how we obtained all the video clips. Next, we illustrate the data annotation procedures and pre-processing process. Finally, we report important dataset statistics and visualize data analysis results.

3.1. Video Clips Acquisition

All video clips used in the dataset were acquired from an online video-sharing website (YouTube) and video clips

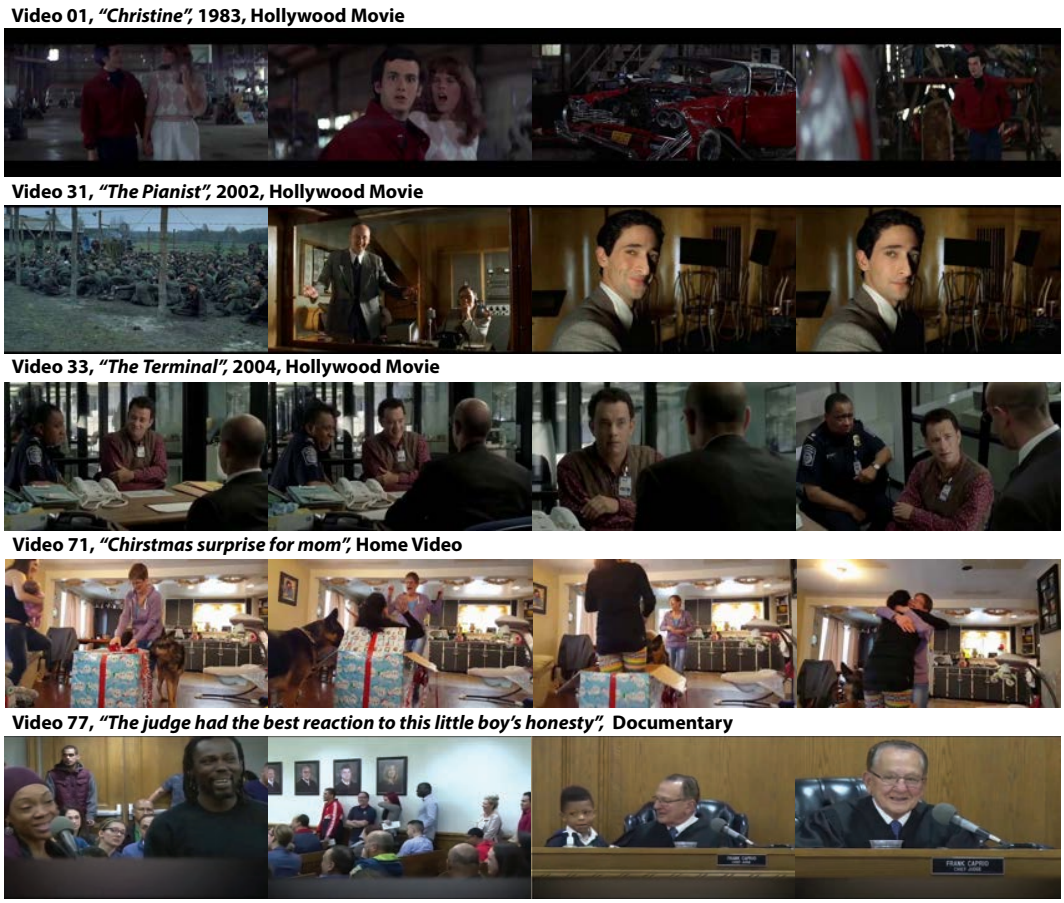

Figure 2. Overview of video frames in VEATIC. We sampled 4 key frames from 5 videos in our dataset. Unlike other datasets where the source of video clips is unique, video clips of VEATIC come from different sources. They include Hollywood movies, documentaries, and homemade videos. Thus, it would make the model trained on our dataset have more generalization ability. For the visual input, VEATIC contains various context information, including different backgrounds, lighting conditions, character interactions, etc. It makes the dataset more representative of our daily life. At last, the emotion/affect of the selected character varies a lot in each video clip, making modeling the character's affect in VEATIC more challenging.

were selected on the basis that the emotions/affect of the characters in the clips should vary across time. In total, the VEATIC dataset contains 124 video clips, 104 clips from Hollywood movies, 15 clips from home videos, and 5 clips from documentaries or reality TV shows. Sample frames from the VEATIC dataset are shown in (Figure 2). These videos contain zero to multiple interacting characters. While audio is an important contextual cue, our current dataset focused on the importance of spatial (setting and background) and temporal context (information across time). Thus, all sound was removed from the videos so observers only had access to visual information when tracking the emotion of the target character.

3.2. Data Annotation and Pre-processing

In total, we had 192 observers who participated in the annotation of the videos in the dataset. All participants were students at UC Berkeley participating for course credit and provided signed consent in accordance with the guidelines and regulations of the UC Berkeley Institutional Review Board and all experimental procedures were approved.

Participants watched and rated a total of 124 videos in the dataset. To prevent observers from getting fatigued, we split the annotation procedure into two 1-hour and 30-minute annotation sessions. Before participants were able to annotate any videos, they were shown a printed version of the valence-arousal affect rating grid with example emotions labeled in different locations of the grid according to the ratings provided by Bradley and Lang (1999) [6]. Importantly, we chose to use valence and arousal instead of emotional categories for emotion annotation to avoid constraining observers emotion judgements to a limited set of emotions. Annotators were instructed to familiarize themselves with the dimensions and the sample word locations which they would later utilize in the annotation process. After participants familiarized themselves with the affect

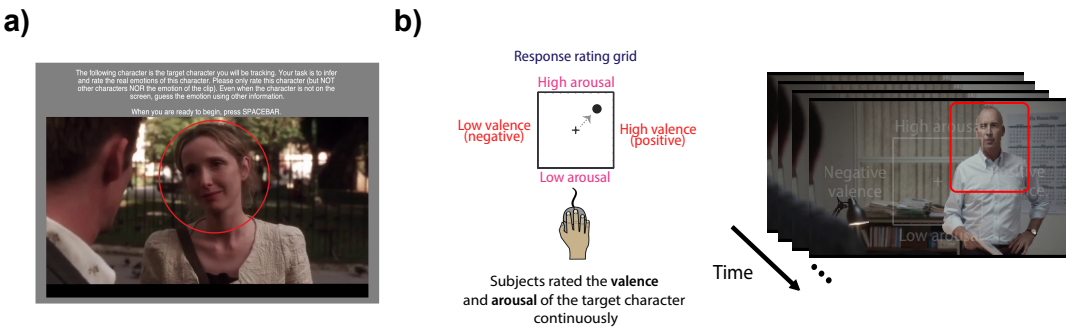

Figure 3. User interface used for video annotation. a) Participants were first shown the target character and were reminded of the task instructions before the start of each video. b) The overlayed valence and arousal grid that was present while observers annotated the videos. Observers were instructed to continuously rate the emotion of the target character in the video in real-time. If observers did not move their mouse for more than 10 seconds, the response rating grid would flash to remind the observer to continuously rate the emotion.

rating grid, they then completed a two-minute practice annotation where they continuously tracked the valence and arousal of a target character in a video (Figure 3b). Annotators were instructed to track the valence and arousal of the target character in the video by continuously moving their mouse pointer in real time within the 2D valence-arousal grid. The grid would map to their valence and arousal ratings in the range of $[-1, 1]$. To control for potential motor biases, we counterbalanced the valence-arousal dimensions between participants where half of the annotators had valence on the x-axis and arousal on the y-axis and the other half had the dimensions flipped so that arousal was on the x-axis and valence was on the y-axis. Once observers finished the practice annotation session, they then started annotating the videos in the dataset.

Before participants started the annotations, they were shown an image with the target character circled (Figure 3a) which informs the participants which character they will track when the video begins. Then, they annotated the video clips in real time. At the end of each video annotation, participants reported their familiarity with the video clip using a 1-5 discrete Likert scale that ranged from "Not familiar", "Slightly familiar", "Somewhat familiar", "Moderately familiar", and "Extremely familiar". Participants were also asked about their level of enjoyment while watching the clip which was rated using a 1-9 discrete Likert scale that ranged from 1 (Not Enjoyable) to 9 (Extremely Enjoyable). Additionally, in order to not make participants feel bored, all 124 video clips were split into two sessions. Participants rated the video clips in two sessions separately.

During each trial, we assessed whether participants were not paying attention by tracking the duration that they kept the mouse pointer at any single location. If the duration was longer than 10 seconds, the affect rating grid would start to fluctuate which reminded participants to continue tracking the emotion of the target character. In order to assess whether there were any noisy annotators in our dataset, we computed each individual annotator's agreement with

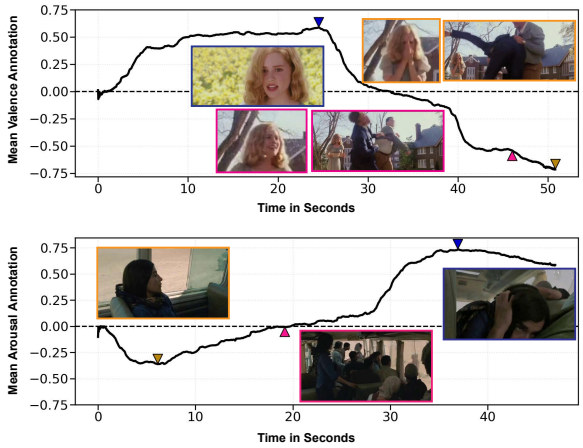

Figure 4. Visualization of sample mean ratings of valence and arousal for specific video clips with the zoom-in view of the selected character. We show key frames related to specific mean ratings of valence and arousal. Corresponding frames and ratings are marked the same color.

the consensus by calculating the Pearson correlation between each annotator and the leave-one-out consensus (aggregate of responses except for the current annotator) for each video. We found that only one annotator had a correlation lower than .2 across all videos with the leave-one-out consensus. Since only one annotator fell below our threshold, we decided to keep the annotator in the dataset in order to not remove any important alternative annotations to the videos.

3.3. Visualization and Data Analysis

Figure 4 shows sample mean ratings and key frames in 2 different video clips. Clearly, both the valence and arousal here have a wide range of ratings. Moreover, it shows that context information, either spatial and/or temporal, plays an important role in emotion recognition tasks. In the valence example (upper figure), without the temporal and/or

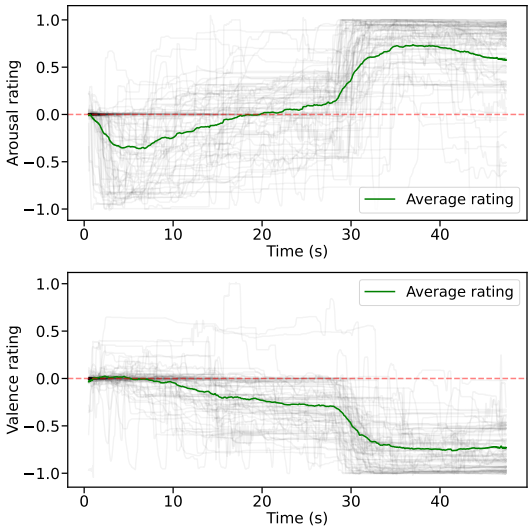

Figure 5. Example valence and arousal ratings for a single video (video 47). Transparent gray lines indicate individual subject ratings and the green line is the average rating across participants.

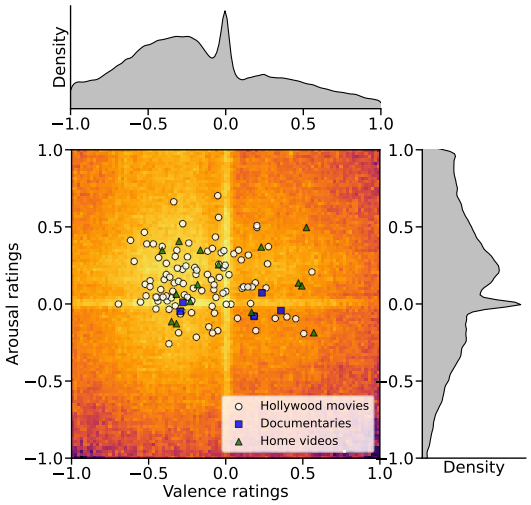

Figure 6. Distribution of valence and arousal ratings across participants. Individual white dots represent the average valence and arousal of the continuous ratings for each video clip for Hollywood movies. Blue squares and green triangles represent the average valence and arousal for documentaries and home videos, respectively. Ratings were binned into 0.02 intervals and the total number of data points was counted within each bin.

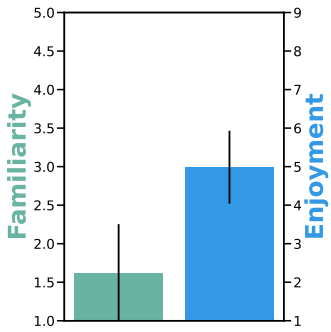

Figure 7. Familiarity and enjoyment ratings across videos for video IDs 0-97. Vertical black lines indicate 1 SD.

spatial context information of the fighting, it would be hard to recognize whether the character (the woman) in the last frame (yellow) is surprisingly happy or astonished. In the arousal example (lower figure), even without the selected character's face, observers can easily and consistently infer the character's arousal via the intense context.

Figure 5 illustrates sample valence and arousal ratings of all participants for a single video in our dataset. Individual subject's ratings (gray lines) followed the consensus ratings across participants (green line) for both valence and arousal ratings. The dense gray line overlapping around the green consensus line indicates agreements between a wide range of observers. Additionally, We investigated how observers' responses varied across videos by calculating the standard deviation across observers for each video. We found that the variance between observers for both valence and arousal dimensions was small with valence having an average standard deviation of $\mu = 0.248$ and a median of 0.222 and arousal having an average standard deviation of $\mu = 0.248$ and a median of 0.244, which are comparable with the valence and arousal rating variance from EMOTIC [32].

The distribution of the valence and arousal ratings across all of our videos is shown in Figure 6. We found that individual participant ratings were distributed fully across both valence and arousal dimensions which highlights the diversity of the VEATIC dataset. We also collected familiarity and enjoyment ratings for each video across participants (shown in Figure 7). We found that observers were unfamiliar with the videos used in the dataset as the average familiarity rating was 1.61 for video IDs 0-97. Additionally, observers rated their enjoyment while watching the videos as an average of 4.98 for video IDs 0-97 indicating that observers moderately enjoyed watching and annotating the video clips. Familiarity and enjoyment ratings were not collected for video IDs 98-123 as the annotations for these videos were collected at an earlier time point during data collection which did not include these ratings.

Table 2 below summarizes the basic statistics of the VEATIC dataset. In a nutshell, VEATIC has a long total video clip duration and a variety of video sources that cover a wide range of contexts and emotional conditions. Moreover, compared to previous datasets, we recruited far more participants to annotate the ratings.

4. Experiments

In this study, we propose a new emotion recognition in context task, i.e. to infer the valence and arousal of the selected character via both context and character informa-

Attribute	Description
No. of Frames	257,601
No. of Videos	124
Total No. of Annotators	192
Avg. No. of Annotators per video	60
Length of Videos	10 s - 2 min 37s
Mean Image Resolution	854 × 480
Hollywood movies	104
Documentaries	5
Home videos	15

Table 2. Statistics of VEATIC Dataset.

tion in each video frame. Here, we propose a simple baseline model to benchmark the new emotion recognition in context task. The pipeline of the model is shown in Figure 8. We adopted two simple submodules: a convolutional neural network (CNN) module for feature extraction and a visual transformer module for temporal information processing. The CNN module structure is adopted from Resnet50 [21]. Unlike CAER [33] and EMOTIC [32], where facial/character and context features are extracted separately and merged later, we directly encode the fully informed frame. For a single prediction, consecutive N video frames are encoded independently. Then, the feature vectors of consecutive frames are first position embedded and fed into the transformer encoder containing L sets of attention modules. At last, the prediction of arousal and valence is accomplished by a multilayer perceptron (MLP) head.

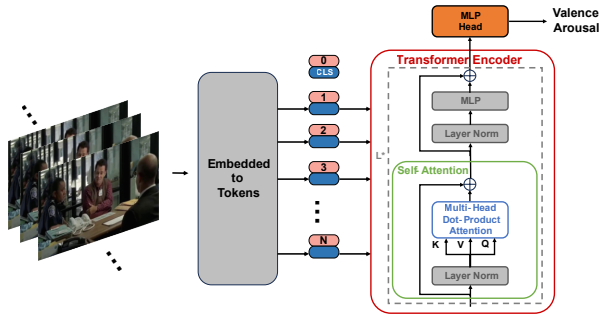

Figure 8. The architecture of the benchmark model for emotion and affect tracking in context task. The model consists of a CNN feature extraction module and a visual transformer for combining temporal information of consecutive frames.

4.1. Loss Function and Training Setup

The loss function of our baseline model is a weighted combination of two separate losses. The MSE loss regularizes the local alignment of the ground truth of ratings and the model predictions. In order to guarantee the alignment of the ratings and predictions on a larger scale, such as learning the temporal statistics of the emotional ratings, we also utilize the concordance correlation coefficient (CCC) as a regularization. This coefficient is defined as follows,

$$\rho_c = \frac{2s_{xy}}{s_x^2 + s_y^2 + (\bar{x} - \bar{y})^2} \tag{1}$$

where s_x and s_y are the variances of the ground truth and predicted values in a training batch respectively, $\bar{x}$ and $\bar{y}$ are the corresponding mean values, and s_{xy} is the respective covariance value. After computing the CCC of both arousal and valence, the CCC loss is computed as follows,

$$\mathcal{L}_{CCC} = 1 - \frac{\rho_a + \rho_v}{2} \tag{2}$$

where ρ_a and ρ_v are the concordance correlation coefficient (CCC) for the arousal and valence, respectively. Together, our final training loss is defined as,

$$\mathcal{L} = \mathcal{L}_{CCC} + \lambda \mathcal{L}_{MSE} \tag{3}$$

During training, λ is set to 0.1. The sliding window size N is 5 and the depth of the visual transformer L is set to 6. We train our emotion and affect tracking baseline model end-to-end with Adam optimizer [27], where $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate is set to 0.0005 with a cosine annealing schedule. The first CNN feature extraction module is initialized using a pretrained Resnet50 [21] on Imagenet [51] while the visual transformer is initialized using Kaiming initialization [20]. The batch size is set to 20, and we utilize 4 GPUs for training. For each video, the first 70% of the frames are for training and the rest 30% of the frames are for testing.

4.2. Evaluation Metrics

When testing our model's performance, we utilize the concordance correlation coefficient (CCC, ρ_c) as one of the evaluation metrics. In addition to CCC, we also utilize the Pearson correlation coefficient (PCC), the root mean square error (RMSE), and the sign agreement (SAGR) to evaluate the model's performance. The SAGR metric is defined as follows.

$$SAGR(X, Y) = \frac{1}{N} \sum_{i=1}^{N} \delta(sign(x_i), sign(y_i)) \tag{4}$$

where $\bar{x}$ and $\bar{y}$ are the means of sample X and Y, and $\delta(x, y)$ denotes the Kronecker delta $\delta(x, y) = 1$ if $x = y$; Otherwise $\delta(x, y) = 0$.

The SAGR measures how much the signs of the individual values of two vectors X and Y match. It takes on values in [0, 1], where 1 represents the complete agreement and 0 represents a complete contradiction. The SAGR metric can capture additional performance information than others. For example, given a valence ground truth of 0.2, predictions of 0.7 and -0.3 will lead to the same RMSE value. But clearly, 0.7 is better suited because it is a positive valence.

Frame Type	Valence				Arousal			
	CCC↑	PCC↑	RMSE↓	SAGR↑	CCC↑	PCC↑	RMSE↓	SAGR↑
Fully Informed	**0.6678**	**0.6967**	**0.3084**	**0.8149**	**0.6296**	**0.6584**	**0.2410**	0.7637
Character Only	0.5116	0.5609	0.3776	0.7451	0.5725	0.6247	0.2333	0.7497
Context Only	0.6185	0.6567	0.3245	0.8071	0.6088	0.6181	0.2416	**0.7828**

Table 3. Performance of our proposed model on fully-informed, character-only, and context-only conditions. Inference via both character and context information, the model performs the best. It shows the importance of both context and character information in emotion and affect tracking tasks.

Method	RMSE↓			Method	ACC↑
	Valence	Arousal	Overall		
EMOTIC	1.1730	1.2900	1.2315	**CAER-NET-S**	0.7351
Ours	1.2151	1.3213	1.2682	**Ours**	0.6904

Table 4. Comparison of our fine-tuned proposed method with EMOTIC and CAER-S pretrained model on their Datasets. Our simple model achieves competitive results, indicating the generalizability of VEATIC.

4.3. Benchmark Results

We benchmark the new emotion recognition in context task using the aforementioned 4 metrics, CCC, PCC, RMSE, and SAGR. Results are shown in Table 3. Compared to other datasets, our proposed simple method is on par with state-of-the-art methods on their datasets.

We also investigate the importance of context and character information in emotion recognition tasks by feeding the context-only and character-only frames into the pretrained model on fully-informed frames. In order to obtain fair comparisons and exclude the influence of frame pixel distribution differences, we also fine-tune the pretrained model on the context-only and character-only frames. The corresponding results are shown in Table 3 as well. Without full information, the model performances drop for both context-only and character-only conditions.

In order to show the effectiveness of the VEATIC dataset, we utilized our pretrained model on VEATIC, fine-tuned it on other datasets, and tested its performance. We only tested for EMOTIC [32] and CAER-S [33] given the simplicity of our model and the similarity of our model to the models proposed in other dataset papers. The results are shown in Table 4. Our pretrained model performs on par with the proposed methods in EMOTIC [32] and CAER-S [33]. Thus, it shows the effectiveness of our proposed VEATIC dataset.

5. Discussion

Understanding how humans infer the emotions of others is essential for researchers understanding of social cognition. While psychophysicists conduct experiments, they need specific stimulus sets to design experiments. However, among published datasets, there is currently no context-based video dataset that contains continuous valence and arousal ratings. The lack of this kind of datasets also prevents researchers from developing computer vision algorithms for the corresponding tasks. Our proposed VEATIC dataset fills in this important gap in the field of computer vision and will be beneficial for psychophysical studies in understanding emotion recognition.

During data collection, participants continuously tracked and rated the emotions of target characters in the video clips which is different from general psychophysical experiments where responses are collected after a delay. This design in our dataset was vital in order to mimic the real-time emotion processing that occurs when humans process emotions in their everyday lives. Though most of the video clips in VEATIC are from movies which makes the emotion processing easier, it is a necessary step towards understanding real emotion processing in real-life scenarios. Additionally, emotion processing is not an immediate process and it relies heavily on the temporal accumulation of information over time in order to make accurate inferences about the emotions of others.

The strength of the VEATIC dataset is that it mimics how humans perceive emotions in the real world: continuously and in the presence of contextual information both in the temporal and spatial domain. Such a rich dataset is vital for future computer vision models and can push the boundaries of what current models can accomplish. With the creation of more rich datasets like VEATIC, it may be possible for future computer vision models to perceive emotions in real time while interacting with humans.

6. Conclusion

In this study, we proposed the first context-based large video dataset, **VEATIC**, for continuous valence and arousal prediction. Various visualizations show the diversity of our dataset and the consistency of our annotations. We also proposed a simple baseline algorithm to solve this challenge. Empirical results prove the effectiveness of our proposed method and the VEATIC dataset.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021.

[2] Hillel Aviezer, Shlomo Bentin, Veronica Dudarev, and Ran R Hassin. The automaticity of emotional face-context integration. *Emotion*, 11(6):1406, 2011.

[3] Simon Baron-Cohen, Sally Wheelwright, Jacqueline Hill, Yogini Raste, and Ian Plumb. The "reading the mind in the eyes" test revised version: a study with normal adults, and adults with asperger syndrome or high-functioning autism. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, 42(2):241–251, 2001.

[4] Lisa Feldman Barrett and Elizabeth A Kensinger. Context is routinely encoded during emotion perception. *Psychological science*, 21(4):595–599, 2010.

[5] Pablo Barros, Nikhil Churamani, Egor Lakomkin, Henrique Siqueira, Alexander Sutherland, and Stefan Wermter. The omg-emotion behavior dataset. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2018.

[6] Margaret M Bradley and Peter J Lang. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Technical report C-1, the center for research in psychophysiology . . . , 1999.

[7] Marta Calbi, Francesca Siri, Katrin Heimann, Daniel Barratt, Vittorio Gallese, Anna Kolesnikov, and Maria Alessandra Umiltà. How context influences the interpretation of facial expressions: a source localization high-density eeg study on the "kuleshov effect". *Scientific reports*, 9(1):1–16, 2019.

[8] Zhimin Chen and David Whitney. Tracking the affective state of unseen persons. *Proceedings of the National Academy of Sciences*, 116(15):7559–7564, 2019.

[9] Zhimin Chen and David Whitney. Inferential affective tracking reveals the remarkable speed of context-based emotion perception. *Cognition*, 208:104549, 2021.

[10] Zhimin Chen and David Whitney. Inferential emotion tracking (iet) reveals the critical role of context in emotion recognition. *Emotion*, 22(6):1185, 2022.

[11] Kyunghyun Cho, Bart Merrienboer, Caglar Gulcehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *EMNLP*, 2014.

[12] Jules Davidoff. *Differences in visual perception: The individual eye*. Elsevier, 2012.

[13] Abhinav Dhall, Roland Goecke, Simon Lucey, Tom Gedeon, et al. Collecting large, richly annotated facial-expression databases from movies. *IEEE multimedia*, 19(3):34, 2012.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[15] Ellen Douglas-Cowie, Roddy Cowie, Cate Cox, Noam Amir, and Dirk Heylen. The sensitive artificial listener: an induction technique for generating emotionally coloured conversation. In *LREC workshop on corpora for research on emotion and affect*, pages 1–4. ELRA Paris, 2008.

[16] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

[17] Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978.

[18] Zhiyun Gao, Wentao Zhao, Sha Liu, Zhifen Liu, Chengxiang Yang, and Yong Xu. Facial emotion recognition in schizophrenia. *Frontiers in psychiatry*, 12:633717, 2021.

[19] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 244–253, 2019.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[22] Will E Hipson and Saif M Mohammad. Emotion dynamics in movie dialogues. *PloS one*, 16(9):e0256153, 2021.

[23] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[24] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

[25] Zhao Kaili, Wen-Sheng Chu, and Honggang Zhang. Deep region and multi-label learning for facial action unit detection. In *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3391–3399, 2016.

[26] Mary Kayyal, Sherri Widen, and James A Russell. Context is more powerful than we think: contextual cues override facial cues even for valence. *Emotion*, 15(3):287, 2015.

[27] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[28] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing*, 3(1):18–31, 2011.

[29] Dimitrios Kollias and Stefanos Zafeiriou. Aff-wild2: Extending the aff-wild database for affect recognition. *arXiv preprint arXiv:1811.07770*, 2018.

[30] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, 2019.

[31] Jean Kossaifi, Georgios Tzimiropoulos, Sinisa Todorovic, and Maja Pantic. Afew-va database for valence and arousal estimation in-the-wild. *Image and Vision Computing*, 65:23–36, 2017.

[32] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Context based emotion recognition using emotic dataset. *IEEE transactions on pattern analysis and machine intelligence*, 42(11):2755–2766, 2019.

[33] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-aware emotion recognition networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10143–10152, 2019.

[34] Tae-Ho Lee, June-Seek Choi, and Yang Seok Cho. Context modulation of facial emotion perception differed by individual difference. *PLOS one*, 7(3):e32987, 2012.

[35] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Self-supervised representation learning from videos for facial action unit detection. In *Proceedings of the IEEE/CVF Conference on Computer vision and pattern recognition*, pages 10924–10933, 2019.

[36] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022.

[37] Cheng Luo, Siyang Song, Weicheng Xie, Linlin Shen, and Hatice Gunes. Learning multi-dimensional edge feature-based au relation graph for facial action unit recognition. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 1239–1246, 2022.

[38] Daniel McDuff, Rana Kaliouby, Thibaud Senechal, May Amr, Jeffrey Cohn, and Rosalind Picard. Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 881–888, 2013.

[39] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE transactions on affective computing*, 3(1):5–17, 2011.

[40] Trisha Mittal, Pooja Guhan, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. Emoticon: Context-aware multimodal emotion recognition using frege's principle. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14234–14243, 2020.

[41] MA Nasri, Mohamed Amine Hmani, Aymen Mtibaa, Dijana Petrovska-Delacretaz, M Ben Slima, and A Ben Hamida. Face emotion recognition from static image based on convolution neural networks. In *2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, pages 1–6. IEEE, 2020.

[42] Erik C Nook, Kristen A Lindquist, and Jamil Zaki. A new look at emotion perception: Concepts speed and shape facial emotion recognition. *Emotion*, 15(5):569, 2015.

[43] Desmond C Ong, Zhengxuan Wu, Zhi-Xuan Tan, Marianne Reddan, Isabella Kahhale, Alison Mattek, and Jamil Zaki. Modeling emotion in complex stories: the stanford emotional narratives dataset. *IEEE Transactions on Affective Computing*, 12(3):579–594, 2019.

[44] Desmond C Ong, Jamil Zaki, and Noah D Goodman. Computational models of emotion inference in theory of mind: A review and roadmap. *Topics in cognitive science*, 11(2):338–357, 2019.

[45] Timea R Partos, Simon J Cropper, and David Rawlings. You don't see what i see: Individual differences in the perception of meaning from visual stimuli. *PloS one*, 11(3):e0150615, 2016.

[46] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*, 2018.

[47] Jonathan Posner, James A Russell, and Bradley S Peterson. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology*, 17(3):715–734, 2005.

[48] Zhihang Ren, Xinyu Li, Dana Pietralla, Mauro Manassi, and David Whitney. Serial dependence in dermatological judgments. *Diagnostics*, 13(10):1775, 2023.

[49] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–8. IEEE, 2013.

[50] David E Rumelhart, Geoffrey E Hinton, Ronald J Williams, et al. Learning internal representations by error propagation, 1985.

[51] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[52] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[53] James A Russell. dimensional contextual perspective. *The psychology of facial expression*, page 295, 1997.

[54] Andrey V Savchenko. Facial expression and attributes recognition based on multi-task learning of lightweight neural networks. In *2021 IEEE 19th International Symposium on Intelligent Systems and Informatics (SISY)*, pages 119–124. IEEE, 2021.

[55] Andrey V Savchenko, Lyudmila V Savchenko, and Ilya Makarov. Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. *IEEE Transactions on Affective Computing*, 13(4):2132–2143, 2022.

[56] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. Deep adaptive attention for joint facial action unit detection and face alignment. In *Proceedings of the European conference on computer vision (ECCV)*, pages 705–720, 2018.

[57] Jiahui She, Yibo Hu, Hailin Shi, Jun Wang, Qiu Shen, and Tao Mei. Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6248–6257, 2021.

[58] Ian Sneddon, Margaret McRorie, Gary McKeown, and Jennifer Hanratty. The belfast induced natural emotion database. *IEEE Transactions on Affective Computing*, 3(1):32–41, 2011.

[59] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE transactions on affective computing*, 3(1):42–55, 2011.

[60] Paweł Tarnowski, Marcin Kołodziej, Andrzej Majkowski, and Remigiusz J Rak. Emotion recognition using facial expressions. *Procedia Computer Science*, 108:1175–1184, 2017.

[61] Y-I Tian, Takeo Kanade, and Jeffrey F Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 23(2):97–115, 2001.

[62] Vedat Tümen, Ömer Faruk Söylemez, and Burhan Ergen. Facial emotion recognition on a dataset using convolutional neural network. In *2017 International Artificial Intelligence and Data Processing Symposium (IDAP)*, pages 1–5. IEEE, 2017.

[63] Gaetano Valenza, Antonio Lanata, and Enzo Pasquale Scilingo. The role of nonlinear dynamics in affective valence and arousal recognition. *IEEE transactions on affective computing*, 3(2):237–249, 2011.

[64] Raviteja Vemulapalli and Aseem Agarwala. A compact embedding for facial expression similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5683–5692, 2019.

[65] Kannan Venkataramanan and Haresh Rengaraj Rajamohan. Emotion recognition from speech. *arXiv preprint arXiv:1912.10458*, 2019.

[66] Kai Wang, Xiaojiang Peng, Jianfei Yang, Shijian Lu, and Yu Qiao. Suppressing uncertainties for large-scale facial expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6897–6906, 2020.

[67] Fanglei Xue, Zichang Tan, Yu Zhu, Zhongsong Ma, and Guodong Guo. Coarse-to-fine cascaded networks with smooth predicting for video facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2412–2418, 2022.

[68] Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung. Multimodal speech emotion recognition using audio and text. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 112–118. IEEE, 2018.

[69] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: valence and arousal'in-the-wild'challenge. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 34–41, 2017.

[70] Yuanyuan Zhang, Jun Du, Zirui Wang, Jianshu Zhang, and Yanhui Tu. Attention based fully convolutional network for speech emotion recognition. In *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1771–1775. IEEE, 2018.

[71] Yuan-Hang Zhang, Rulin Huang, Jiabei Zeng, and Shiguang Shan. M 3 f: Multi-modal continuous valence-arousal estimation in the wild. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 632–636. IEEE, 2020.