Predicting Player Engagement in *Tom Clancy's The Division 2*: A Multimodal Approach via Pixels and Gamepad Actions

Kosmas Pinitas
Institute of Digital Games
University of Malta
Msida, Malta
kosmas.pinitas@um.edu.mt

David Renaudie
Massive Entertainment
Ubisoft
Malmö, Sweeden
david.renaudie@massive.se

Mike Thomsen
Massive Entertainment
Ubisoft
Malmö, Sweeden
mike.thomsen@massive.se

Matthew Barthet
Institute of Digital Games
University of Malta
Msida, Malta
matthew.barthet@um.edu.mt

Konstantinos Makantasis
Institute of Digital Games
University of Malta
Msida, Malta
konstantinos.makantasis@um.edu.mt

Antonios Liapis
Institute of Digital Games
University of Malta
Msida, Malta
antonios.liapis@um.edu.mt

Georgios N. Yannakakis
Institute of Digital Games
University of Malta
Msida, Malta
georgios.yannakakis@um.edu.mt

ABSTRACT

This paper introduces a large scale multimodal corpus collected for the purpose of analysing and predicting player engagement in commercial-standard games. The corpus is solicited from 25 players of the action role-playing game *Tom Clancy's The Division 2*, who annotated their level of engagement using a time-continuous annotation tool. The cleaned and processed corpus presented in this paper consists of nearly 20 hours of annotated gameplay videos accompanied by logged gamepad actions. We report preliminary results on predicting long-term player engagement based on in-game footage and game controller actions using Convolutional Neural Network architectures. Results obtained suggest we can predict the player engagement with up to 72% accuracy on average (88% at best) when we fuse information from the game footage and the player's controller input. Our findings validate the hypothesis that long-term (i.e. 1 hour of play) engagement can be predicted efficiently solely from pixels and gamepad actions.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; • **Applied computing** → **Computer games**; • **Computing methodologies** → **Machine learning algorithms**; Computer vision.

KEYWORDS

datasets, convolutional neural networks, affect modelling, engagement modelling, digital games

ACM Reference Format:
Kosmas Pinitas, David Renaudie, Mike Thomsen, Matthew Barthet, Konstantinos Makantasis, Antonios Liapis, and Georgios N. Yannakakis. 2023. Predicting Player Engagement in *Tom Clancy's The Division 2*: A Multimodal Approach via Pixels and Gamepad Actions. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '23), October 09–13, 2023, Paris, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3577190.3614203

ICMI '23, October 09–13, 2023, Paris, France

ACM ISBN 979-8-4007-0055-2/23/10…$15.00
https://doi.org/10.1145/3577190.3614203

1 INTRODUCTION

A largely unresolved challenge in the field of Affective Computing (AC) is the task of modelling affect over long periods of time. To examine the degree to which reliable long-term computational models of affect can be constructed, it is imperative to have access to corpora containing affect responses and annotations over extended time periods. The most widely used affect datasets [19, 21, 35, 44], however, contain sessions that last up to a few minutes, at most.

Motivated by the lack of multimodal corpora for the study of long-term affect modelling, this paper introduces a game affect corpus consisting of 1-hour long interactive gameplay sessions. The introduced dataset contains data from 20 participants who played one hour of *Tom Clancy's The Division 2* (Ubisoft, 2019)—*The Division 2* for short—and annotated their own gameplay videos in terms of engagement using the PAGAN annotation tool [34]. Apart from the in-game footage modality, the presented version of *The Division 2* dataset also features the player's inputs on the game controller (gamepad). The long-term interactive nature of *The Division 2* as an elicitor offers a unique contextual environment for modelling affect over extended periods of time, thereby broadening the research horizons of AC per se. The features of *The Division 2*

also capture the multimodal interaction capacities on popular and commercial-standard applications such as games.

To validate our hypothesis that in-game footage and gamepad actions can reliably predict player affect for extended periods of time, we train a number of Convolutional Neural Network (CNN) architectures on nearly 20 hours of frame and player input data of *The Division 2* corpus. Inspired by [26, 27], we assume that rich and sufficient affect information is existent (and interwoven) within the pixels and the actions of the gameplay, and thus that a predictive model will be able to capture it effectively. This initial study tests the efficiency of deep learning models (relying on these two modalities) on predicting long-term player engagement represented as a binary classification task (i.e. high vs. low engagement). Importantly for the nature of this study we also investigate the impact of time conditioning on the predictive capacity of long-term affect models. Our key results indicate that we can predict high and low engagement states from long-term affect stimuli with high accuracy, particularly when the fusion of the two modalities (game footage and gamepad actions) is time-conditioned.

This paper is novel in several ways. First, to the best of our knowledge, this is the first time a commercial-standard game such as *The Division 2* is used for the study of long-term player experience and affect manifestations at large. Second, the paper presents a generic methodology for modelling affect by fusing pixel information with gamepad actions in this corpus. Third, we introduce the concept of *time conditioning* for the purpose of modelling long-term engagement in games and beyond. Finally, the initial results presented in this paper serve as the baseline for this new multimodal corpus.

2 BACKGROUND

This section reviews related work on affect and engagement modelling from pixels and other modalities (Sections 2.1- 2.2) and moves on to survey literature on affect corpora (Section 2.3).

2.1 Affect Modelling

Affective computing is a multidisciplinary field that studies the expression of emotions and aims to develop models that can computationally capture such manifestations [41]. As videos and images can elicit emotion, it comes as no surprise that affect modelling from visual cues is gaining ground. Before the advent of deep learning, the dominant approach involved the use of domain knowledge and high-level hand-crafted visual features [8, 60]. Although such approaches are memory efficient and allow for real-time emotion recognition, the development of large-scale affect datasets [20, 44] and the gradual advancement of deep learning led to significant breakthroughs in affect modelling [30] and multimodal deep fusion [31]. Indicatively, Breuer and Kimmer [6] employed CNNs for various facial expression recognition tasks whereas Ng et al. [38] used CNNs pretrained on ImageNet [45] to perform emotion recognition on small datasets. Assuming that games can be effective elicitors of affect, Makantasis et al. [26] trained CNNs to map gameplay footage to arousal while Pinitas et al. [42] evolved parameters of a preference learner to predict arousal in gameplay videos.

Other modalities such as physiology and speech (audio) have also been extensively used for modeling affect, either individually or in a multimodal setting [1, 2, 24, 47]. Notably, Martinez et al. [30] were the first to apply CNNs for detecting affect via physiological signals. Makantasis et al. [27] employed CNNs and modelled arousal from raw gameplay footage and sound. Zhang et al. [59] used a Convolutional LSTM and a 1D-CNN to extract spatio-temporal facial and bio-sensing features, respectively. Recently, Pinitas et al. [43] employed Supervised Contrastive Learning on audiovisual and physiological data to model arousal. Unlike the aforementioned studies, this paper presents preliminary findings regarding *long-term* player engagement prediction from in-game footage and game controller input using CNNs.

2.2 Engagement Modelling

It can be argued that *engagement* plays an important role in human-computer interaction (HCI) as a multifaceted construct that encompasses cognitive, affective, and behaviourally characteristics of the user [3, 5]. Given its pivotal role in HCI research, several studies have focused on modelling different aspects of user engagement. Dermouche and Pelachaud [10] developed an LSTM-based model to predict user engagement in real time dyadic interactions based on facial expressions, head movements and gaze. Ting et al. [50] employed Bayesian Networks to model variables of student engagement in virtual learning environments, while Fan et al. [13] presented a robotic coach system based on multi-user engagement.

Engagement modelling has been central to AC research because it facilitates the computational modelling of more complex emotional responses: different levels of engagement correspond to different arousal-valence points on the affective circumplex model [46] Indicatively, Vries et al. [9] propose a methodology for reverse engineering a consumer behaviour model for online customer engagement based on a computational and data-driven perspective. Games have also proven to be an engaging entertainment medium; consequently, it is unsurprising that there is a growing body of research in the field of player engagement modelling. Specifically, Melhart et al. [33] used viewers' chat logs as a proxy for engagement and employed a small neural network to predict moment-to-moment gameplay engagement based solely on game telemetry. Xue et al. [53] proposed a Dynamic Difficulty Adjustment framework to maximise a player's engagement (as stay time). Finally, Huang et al. [16] introduced a two-stage player engagement modeling approach using Hidden Markov Models. In this paper we view engagement via the lens of affect (see Section 3.2.2), and fuse captured gameplay footage and player actions to predict high or low engagement in different time segments of a long gameplay session.

2.3 Affect Corpora

Over the years, affect modelling has relied increasingly on large-scale and data-hungry computational models, which in turn require extensive affect corpora that encompass quantifiable expressions of emotions elicited via appropriate stimuli. A commonly held view is that acquiring annotated data that contain reliable affect information is a fundamental aspect of this endeavour. As this study introduces and builds upon data from a large-scale affect corpus, this section provides an overview of the most commonly employed affect corpora and their characteristics.

A key distinguishing factor among affect datasets is the annotation protocol used. The *first-person annotation* protocol involves

participants performing a task and then annotating their own affect. For instance, MAHNOB-HCI [48] and DEAP [19] databases consist of multiple modalities, such as electroencephalogram, electrodermal activity, facial video, and others, recorded from a first-person perspective and annotated with affect labels via a first-person annotation protocol. However, growing body of work employs a *third-person annotation* protocol, where participants perform a task while a team of annotators (usually experts) annotate the participant's emotions. Indicatively, the RECOLA database [44] includes recordings of online dyadic interactions between participants solving a task in collaboration; a group of six experts provided the socio-affective data annotations at a later stage. A similar annotation protocol has been used in the SEWA database [21], which consists of audio-visual recordings of participants discussing in pairs. While affect corpora in games are often based on first-person annotation of a recent playthrough [35], Mavromoustakos et al. [32] tasked two external experts to annotate tension on a large corpus (over 26 hours of videos) of competitive Hearthstone matches.

Affect corpora tend to rely on audiovisual media such as music videos and movies [20, 37, 58]. It is thus unsurprising that there is growing research attention on both board and video games [11, 32, 56] due to the fact that they are interactive elicitors offering rich affect information. One of the first video game-based affect corpora is the platformer experience dataset [17], a collection of videos of *Super Mario Bros* (Nintendo, 1985) players, facial cues, and gameplay features. The recent AGAIN dataset [35] offers game footage and event logs annotated for arousal in a continuous first-person fashion. The FUNii dataset [4] features multiple recordings of electrocardiogram activity, electrodermal activity, controller input, gaze, and head position, and provides first-person annotations for fun, difficulty, workload, immersion, and user experience.

Unlike earlier multimodal affect corpora, this work analyses affect from a large-scale dataset of 20 gameplay sessions (nearly 1 hour each) annotated for engagement in a first-person manner. Moreover, we employ deep learning algorithms to model long-term engagement relying on users' multimodal signal streams. The dataset covered here contains only gamepad input and captured gameplay footage; however, the extended version of the dataset also features more participants and input modalities including electrodermal activity, photoplethysmography, and eye-tracking signals.

3 TOM CLANCY'S THE DIVISION 2 CORPUS

This section presents the large-scale multimodal corpus of annotated gameplay videos for the action-role playing game *Tom Clancy's The Division 2*. Section 3.1 describes the game, Section 3.2 provides an overview of the data collection protocol and Section 3.3 covers our method for pre-processing the collected data.

3.1 The Game

Tom Clancy's the Division 2 (Ubisoft, 2019) is an online action role-playing third-person shooter developed by Massive Entertainment and published by Ubisoft in 2019 (see Fig. 1). The game, which has sold over 20 million copies worldwide, features both single-player and multi-player gameplay. Players can customise their characters and must scavenge for resources to survive in a challenging setting. The game contains over 30 missions where players must work their way through scripted content alone or in a group with other players. In this corpus, participants played the first mission of the game (*Dawn's Early Light*), which offers a good balance between in-game exploration and combat. In this mission the player's goal is to stop the siege on the White House while securing the area.

Figure 1: In-game image of a player engaged in combat in *The Division 2*.

The Division 2 is a commercial-standard game environment which is ideal for eliciting rich affective responses. Besides high-quality graphics, *The Division 2* has a complex input system and intense action set. Collectively, the properties of this game allow for a meaningful realisation of the affective loop [57]. The rich forms of HCI within the meticulously designed simulated world facilitate user immersion, which is essential for modelling the affective aspect of engagement. In the selected mission, stimuli are varied during a long gameplay session (approximately one hour). The balance between combat and exploration in this mission leads to intense user actions followed by periods of less intense activities.

3.2 The Corpus

Data collection was carried out in two phases. First, 25 participants were asked to play roughly one hour of gameplay of *The Division 2* using an XBOX controller (gamepad). All players played the same mission, *Dawn's Early Light* (see Section 3.1) alone (in single-player mode) until they completed the mission. Following the protocol introduced in [35], participants were then asked to watch the recorded video of their own gameplay and annotate their *engagement* in a continuous manner using the RankTrace [25] annotation tool of the PAGAN platform [34] (see Fig. 2). At the beginning of the experiment, participants filled in a demographic survey. Building on ethical principles of AI and games research [36], care was taken to ensure data was collected and analysed respecting GDPR principles.

3.2.1 Modalities of User Input. During the gameplay phase of *The Division 2*, several modalities were collected (see Fig. 2). In this paper, we only process two types of information about the game context and the player behaviour. The *frame* modality consists of a series of high-resolution frames of in-game footage (usually 1280×720 pixels). The *gamepad* modality contains detailed player actions with the game controller. The possible gamepad actions captured in the dataset are 25, and include buttons pressed (e.g. "A button pressed") or other controller interactions (e.g. "left stick up")

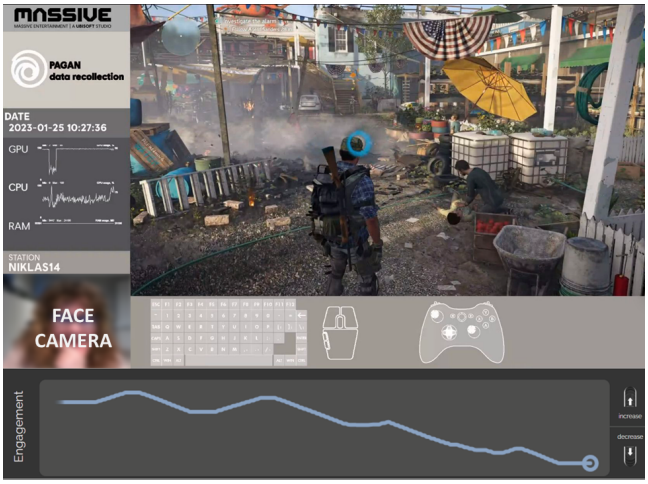

Figure 2: Data collection snapshot for *The Division 2*. Annotated videos contain timestamps for data synchronisation, statistics about compute resources, the ID of the workstation, the face of the participant (blurred out in this paper), eye-tracking data, and the live input on their gamepad. The bottom of the layout visualises the participant's engagement annotation trace using the PAGAN annotation tool.

but do *not* capture in-game events resulting from these inputs (e.g. "player fires weapon"). These actions are stored in string format, and include co-occurring actions (combos). Frames and gamepad actions are stored and processed for our analysis (see Section 3.3).

3.2.2 Annotation. After completing their gameplay sessions, participants were asked to annotate their experience in a first-person manner using the RankTrace [25] annotation tool of the PAGAN platform [34] (see Fig. 2). PAGAN allows users to annotate a singular dimension in a time-continuous fashion while watching a video (in this case, the recorded gameplay data and other modalities as shown in Fig. 2). PAGAN traces are synchronised to the video, and the annotators uses the mouse scroll wheel to increase or decrease the intensity of the perceived affect dimension. Resulting traces are unbounded, and thus annotators can keep increasing or decreasing this value as new stimuli are shown. The web interface of PAGAN shows the entirety of the trace so far, scaling the x-axis of the shown trace (Fig. 2) as the video progresses.

Unlike the majority of continuously annotated affect corpora (see Section 2.3), this corpus tasked participants to annotate player experience with *engagement* traces, instead of e.g. arousal or valence [46]. Participants were given the following definition of engagement prior to their annotation task: "*Engagement refers to the level of attention; a high level of engagement is associated with a feeling of tension, excitement, and readiness while a low level of engagement is associated with boredom, low interest, and disassociation with the game.*" Although this definition of engagement closely resembles the definition of arousal, the latter refers to physiological activation and, consequently, it can not fully capture the complex and multifaceted nature of the gaming experience [23, 41]. Additionally, engagement also captures aspects of valence since it can be a result of both positive and negative emotions such as happiness, terror, and anger [22]. Based on the circumplex model of affect, we can argue that high engagement maps to high valence and high arousal while low engagement represents emotional states that are closer to the low arousal low valence quadrant of affect [5]. Ultimately, we selected the annotation label of *engagement* as it is highly representative of the gameplay context provided to our annotators whilst being related to the core affect dimensions of arousal and valence.

To maximize the reliability and consistency of engagement annotations, participants were required to watch their gameplay at double speed (i.e. videos of 30 minutes) to minimize any effects caused by long-term annotation fatigue. The annotation trace is rescaled to the video duration before processing (see Section 3.3).

Table 1: *The Division 2* Corpus Properties

Property	Raw	Clean
Number of Participants	25	20
Number of Gameplay Videos	25	20
Number of Gamepad logs	25	20
Number of Annotated Video logs	24	20
Video database size	24 hours	18.8 hours
Number of Elicitors	1 game	
Gameplay video duration	53 to 65 minutes	
Annotation Perspective	First-person	
Annotation Type	Continuous unbounded	
Affect Labels	Engagement	

3.2.3 Dataset and Participants. The raw dataset consists of 24 hours of gameplay (57.65 minutes per participant). However, only data from 20 participants are included in the clean dataset used here, due to inconsistencies on the annotation timestamps and missing data. This first iteration of *The Division 2* engagement corpus includes the participants' controller inputs and gameplay frames (see Section 3.2.1). Following [35], we summarise the properties of *The Division 2* corpus on Table 1.

Participants' ages ranged between 18 and 35, forming a diverse mix of individuals within the young adult category. Geographically, all participants are residents of Malmö, providing a localized perspective on the data. To collect precise data from electrodermal activity and eye-tracking, participants must not suffer from any skin condition or astigmatism. In terms of gaming experience, we aimed for participants that have not played the game before to ensure that they approached the study with a fresh perspective. However, a certain level of familiarity with the primary input method (XBOX controller) and other shooter games was required to acquire gameplay data of high and comparable quality.

3.3 Data Pre-Processing

As this study aims to model long-term engagement via players' multimodal signals, we consider the following data pre-processing method. We split each participant's session (video) into overlapping time windows [29, 43] using a sliding step of 1.5 seconds and a window length of 10 seconds, corresponding to 22,541 samples in the entire clean dataset. The sliding step and window length are essential hyperparameters since they influence, respectively, the

size of the dataset and the information contained in each window. The stimuli-based time windows (frames or gamepad modalities) are shifted by 1 sec to the annotation time window, accounting for the reaction time between stimulus and emotional response and the speed difference between gameplay and annotation [42].

After splitting each session into time windows, each window consists of a sequence of frames and logged gamepad actions. For the frame modality, we keep only 3 frames per second to reduce computational load. The 10 second time window used in this paper therefore consists of 30 RGB images of dimensions 224×224×3 (scaled down from the original high-resolution video). For the gamepad modality, we calculate the number of times the player pressed a specific key on the game controller during this time window, and also include a "no key" input as the number of times no key was pressed. Moreover, we calculate the number of n-button combos with n ranging between 2 and 6. We convert these to input frequencies by dividing by the time window length. We thus collect 31 real-valued features from the gamepad modality (25 keypress frequencies, one "no key" frequency, and 5 combo frequencies), which are used as input to the model (see Section 4.2).

When it comes to the engagement traces, we perform a min-max normalization, transforming the unbounded engagement values to a value range of $[0, 1]$ on a per-trace basis (see Fig. 3). Similarly to the frames, we process the affect traces into time windows of 10 seconds. Finally, the average value of each time window provides a single engagement value per time window (see Section 4.1).

4 MODELLING ENGAGEMENT

We present our methodology for modelling engagement below. Section 4.1 outlines the learning paradigm used to model engagement in *The Division 2*, Section 4.2 outlines the CNN architectures used for the unimodal (frames and gamepad actions) and multimodal network, while Section 4.3 presents different time-conditioning strategies explored.

4.1 Learning Paradigms

Arguably one of the most crucial steps in affect modelling is the choice of the supervised learning paradigm under which the mapping between multimodal user signals and affect labels will be inferred. When analysing an entirely new affect corpus such as *The Division 2*, it is useful for all possible learning paradigms to be explored—including regression, classification and ordinal learning [55]. In this initial study of *The Division 2*, we focus on affect classification since it is one of the most commonly used learning paradigms in player and affect modelling [26–29, 43].

In this first experiment, we follow the paradigm used in short-term time-continuous affect annotation traces [26, 27] and classify time windows based on the average trends of the entire 1-hour normalised trace (see Fig. 3). Specifically, we select classes based on the average affect value of the entire trace (μ_i) of each participant (i) which acts as the class splitting criterion. For participant i, a time window t is labelled as *high engagement* when $e_{i,t} > \mu_i + \epsilon$ and as *low engagement* when $e_{i,t} < \mu_i - \epsilon$; $e_{i,t}$ is the average normalised engagement value within the time window t of participant i (sampled at 30 Hz). It should be noted that the threshold ϵ is used to eliminate windows with ambiguous affect annotation values close to μ_i, which may deteriorate the stability of the models. Following best practices from [26, 27] and preliminary tests with this corpus, we set $\epsilon = 0.05$ for all experiments.

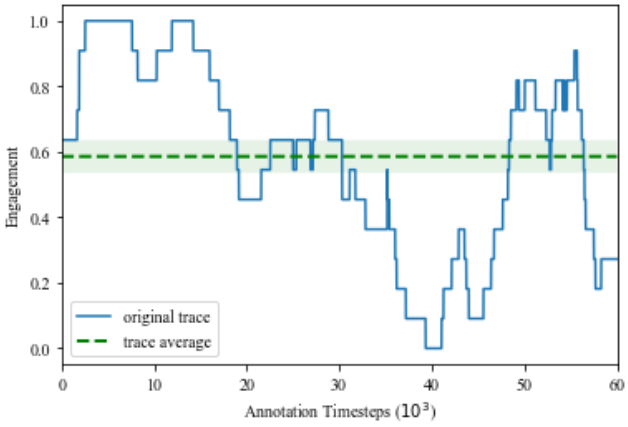

Figure 3: The normalised engagement trace (blue) of participant i. The green line represents the mean value (μ_i) of the trace and shaded area represents the ambiguity area $[\mu_i - \epsilon, \mu_i + \epsilon]$. Time windows with mean values above this shaded area correspond to high engagement; time windows below this shaded area correspond to low engagement.

4.2 Model Architecture

As mentioned in Section 3, this paper considers two gameplay modalities of *The Division 2*: (a) the player's controller input (i.e. *gamepad* modality) and (b) the in-game footage (i.e. *frame* modality). Since we treat the long-term engagement traces as a classification task (see Section 4.1), all architectures end with a 2-neuron softmax-activated layer that predicts low or high engagement.

The architecture used for the *gamepad* modality is visualised in Figure 4a. A network takes the 31 inputs from gamepad actions (see Section 3.3) and processes them via a Gelu-activated [39] fully connected layer of 30 neurons, followed by a 2-neuron softmax-activated layer.

The architecture used to predict engagement from the *frame* modality is visualised in Figure 4b. This network accepts 30 scaled-down RGB images as input (i.e. a tensor of $224 \times 224 \times 3 \times 30$ based on Section 3.3) and processes them via a ResNet18 architecture that outputs 512 feature maps of dimensions 7×7 per input frame, corresponding to a $30 \times 512 \times 7 \times 7$ feature tensor. This ResNet18 architecture is pre-trained on ImageNet [45], similar to an abundance of previous work [38], and its weights are frozen during this training process. The ResNet18 output passes through a spatial max pooling layer, reducing the dimensionality to 30×512, and a temporal average pooling returning a 1D vector of 512 features. The last vector is then fed into two consecutive Gelu-activated fully connected layers of 128 and 30 neurons respectively, each followed by a 0.1 dropout layer. Similar to the gamepad architecture, the last layer is a 2-neuron softmax-activated layer.

The fusion architecture considers both modalities (frames and gamepad actions) and is illustrated in Fig. 4c. Following a *late fusion*

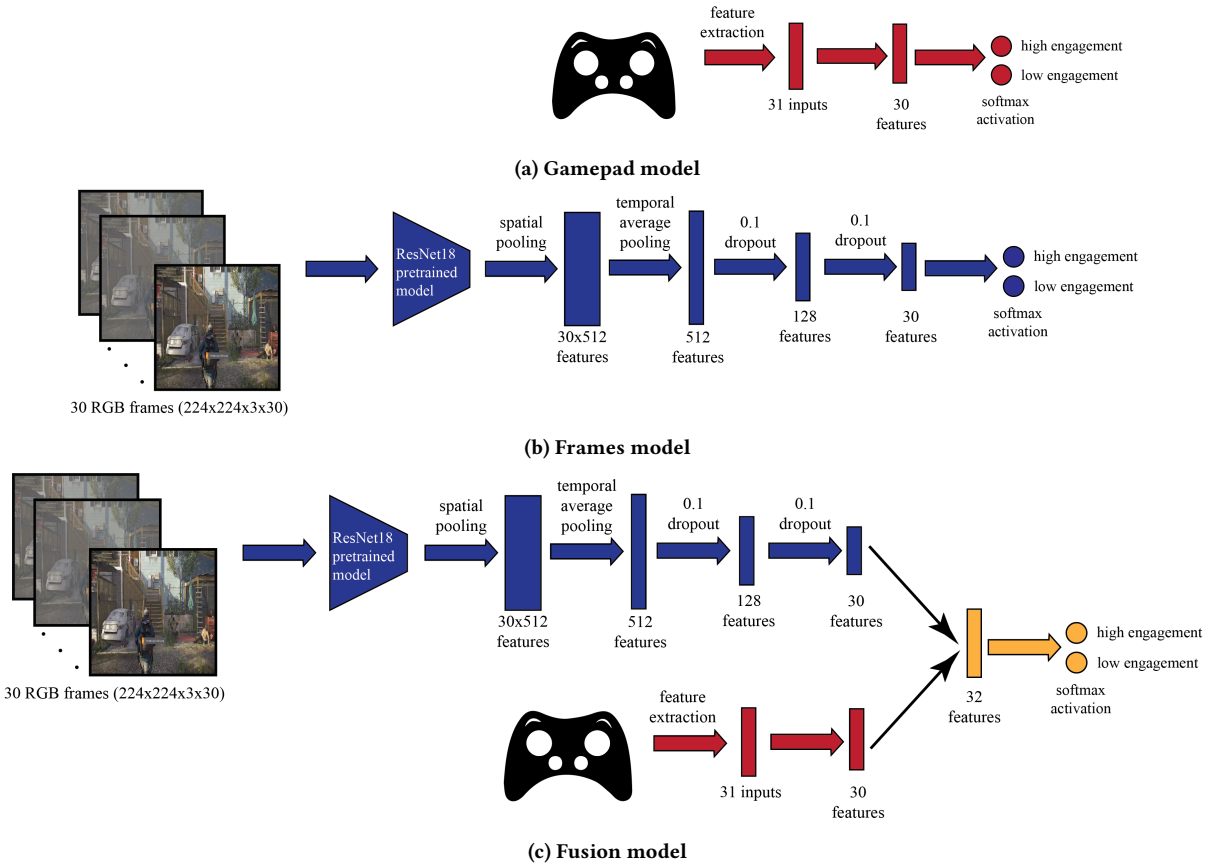

Figure 4: The three model architectures employed for binary classification between high engagement and low engagement, using pixel information, gamepad actions, or both modalities through late fusion.

approach [27], the above unimodal architectures are combined by concatenating their latent representations to form a 1D vector of 60 features which is then fed into a Gelu-activated fully connected layer of 32 neurons followed by a 2-neuron softmax-activated layer.

4.3 Conditioning on Time

Time conditioning refers to the practice of incorporating temporal information into machine learning models to improve their performance. When it comes to affect modelling, time conditioning can be of great value as the relationship between input variables (i.e. affect stimuli) and output variables (i.e. affect labels) changes over time—a phenomenon also known as *concept drift* [14]. In order to validate our hypothesis that incorporating time features into an affect model can result in algorithms capable of capturing the dynamics of emotion over time, we explore three time-conditioning strategies that are detailed in the remainder of this section.

The first step of time conditioning involves transforming the scalar timestep value into a D-dimensional vector $e(t)$, which was first introduced by Transformers [51]. This vector offers a unique, deterministic and bounded encoding for each timestep while ensuring that the distance between any two timesteps is consistent across samples. The $e(t)$ vector is fed into a learnable linear down-projection layer that facilitates the injection of time in the models regardless of the size of the input modalities. Following preliminary tests, in this paper we use a vector of 512 features for all conditioning strategies, and treat the timestep at high granularity (increments of 20 minutes). The timestep t_L can take three possible values depending on the time window's start time t_w, i.e. $t_L = 1$ for $t_w \in [0, 20)$ minutes, $t_L = 2$ for $t_w \in [20, 40)$ mins, and $t_L = 3$ for $t_w \in [40, \infty)$ mins. The implementation of the 512-dimensional encoding is provided by the FAIRSEQ library [40] via Eq. (1):

$$e(t) = \left[\ldots, \cos\left(t_L \cdot c^{-\frac{2d}{D}}\right), \sin\left(t_L \cdot c^{-\frac{2d}{D}}\right), \ldots \right]^T \quad (1)$$

where $d = 1 \ldots D/2$ ($D = 512$ in this paper), $c = 10000$, and t_L takes the values of 1, 2, or 3 depending on which 20-minute increment the time window belongs to.

4.3.1 Shift Last Hidden Layer (M_{SLL}). In this case, we follow the conditioning process employed in Decision Transformers [7]. The sinusoidal embedding is constructed via Eq. (1) and is then down-projected linearly in order to match the dimensionality of the input of the model's last hidden layer. For the frames model (Fig. 4b), the last hidden layer is 30 features and $e(t)$ is down-projected to the

previous layer (128 neurons). For a network with H hidden layers the conditioned output is constructed as follows:

$$O_{c,H-1} = O_{H-1} + s_{H-1} \tag{2}$$

where H is the last hidden (non-output) layer of the network, O_{H-1} and $O_{c,H-1}$ is the output of the previous hidden layer $(H-1)$, before and after conditioning, respectively; s_{H-1} is the linear projection of the sinusoidal time embedding of the previous hidden layer.

4.3.2 Scale and Shift Last Hidden Layer (M_{SSLL}). This method uses the same steps as in Section 4.3.1, but instead of learning a linear down-projection of size n matching the dimensions of the penultimate hidden layer (e.g. $n = 128$ in the frames model), we employ a linear projection of $2n$ neurons such that:

$$O_{c,H-1} = (l_{H-1} + 1) \cdot O_{H-1} + s_{H-1} \tag{3}$$

where l_{H-1} and s_{H-1} are, respectively, the first and last n elements of the linear time embedding projection of the penultimate hidden layer $(H-1)$; remaining notations are the same as in Eq. (2).

4.3.3 Scale and Shift All Layers (M_{SSAL}). Following [52], we explore the case of time-conditioning each layer as long as it is one-dimensional: e.g. in the frame model (Fig. 4b) the layers with 512, 126, 30, and 2 neurons are scaled and shifted. For each layer i of n neurons, we learn a linear projection of $2n$ neurons such that:

$$O_{c,i-1} = (l_i + 1) \cdot O_{i-1} + s_i \tag{4}$$

where l_i and s_i are, respectively, the first and last n elements of the linear time embedding projection for layer i; O_{i-1} and $O_{c,i-1}$ are the outputs of the (previous) $i-1$ layer, respectively, before and after conditioning.

5 RESULTS

This section first outlines the experimental protocol we use to evaluate the algorithms and then presents the key results of the initial round of experiments performed with *The Division 2* corpus.

5.1 Experimental Protocol

We test the capacity of the proposed modelling approaches to predict engagement in *The Division 2*. The model is trained to classify frames and/or gamepad inputs within a time window as low or high engagement. Models in this paper are trained via the Adam optimiser with learning rate of 0.005 and batch size of 256. Moreover, we ensure that the same training, validation and test data are used for all models, promoting a fair comparison.

To evaluate model performance, we use a leave-2-participants-out cross-validation method. This method is a variant of the popular leave-one-participant-out cross-validation method [18], where two participants are used for the test set and another two participants are used for the validation set (for the purposes of early stopping). Data for training originates from 16 players, ensuring that data in each set belong to different participants and thus resulting in non-overlapping datasets. The models are trained for 50 epochs, but stop training after 5 epochs without a validation metric improvement; in all experiments in this paper, the maximum number of epochs was never reached. We split the dataset of 20 participants into 10 sets (with all participants becoming part of the 2-participants test fold) and calculate classification metrics on the test set averaged from 10 training runs, one per set. We randomise participant order and initial network weights 4 times, thus ensuring different participant pairings in the test set each time, and report results averaged across these train/test setups (i.e. 40 folds).

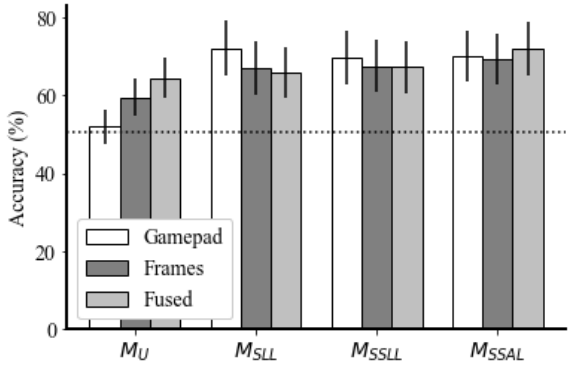

Figure 5: Engagement classification in *The Division 2*. The graph shows average (leave-2-participant-out) test accuracy values and corresponding 95% confidence intervals. The test accuracy of the naive baseline is shown as a dotted line.

We benchmark our models based on the traditional *accuracy score*, as the selected thresholding criterion ($\epsilon = 0.05$) ensures that the dataset is somewhat balanced. A naive *baseline* uses the majority class in the training set and predicts the same class in the test set with an average test accuracy of 51% across all folds (with a 95% confidence interval of 0.28% and a best-fold accuracy of 51.9%). Statistical significance, when reported, refers to two-tailed paired Wilcoxon Signed-Rank Test with $p < 0.05$, where data is matched on the same 2-participants' test folds. When multiple comparisons are performed, the Bonferroni correction is applied [12].

5.2 Engagement Models without Time Context

We report the average test accuracies from 10 cross-validation tests (via leave-2-participants-out) repeated 4 times in Figure 5 under M_U. It is obvious that models trained on the gamepad modality alone perform poorly, with an average test accuracy of 52% (best-fold accuracy of 68%) which is very close to the baseline (no significant differences). Models trained on frames alone perform significantly better than the baseline and the gamepad models, with an average accuracy of 59.4% (best-fold accuracy of 80%). While gamepad data seems insufficient on their own, when fused with pixel information the trained models improve: the fusion model has an average accuracy of 64.5% (best-fold accuracy of 82.4%) which is significantly higher than all other models. As expected, late fusion of multiple modalities seems beneficial when it comes to engagement modelling, although accuracies remain low overall. This has been validated in previous work when fusing pixel and sound data in games [27], but not for controller input.

5.3 Influence of Time Context

The test accuracies of all three conditioning strategies described in Section 4.3 are shown in Fig. 5, with unconditioned versions denoted as M_U. Surprisingly, the best accuracy for the gamepad

modality is achieved with M_{SLL} (average accuracy of 72%). For the same conditioning strategy frames and fusion models perform significantly worse. A possible reason for this behaviour is the simpler architecture of the gamepad model (with only one hidden layer of 30 neurons), as conditioning applied only on the last hidden layer seems very effective. In comparison, scaling and shifting all hidden layers (M_{SSAL}) works better for the larger architectures, especially the fusion model which reaches accuracies of 72% on average (best-fold accuracy of 87.7%) and significantly outperforms the frames model on the same conditioning. Evidently, time conditioning is beneficial regardless of strategy applied: all conditioned models of gamepad or frames modalities perform significantly better than the unconditioned version for their respective modality.

6 DISCUSSION

This paper introduced a novel and long-term player engagement multimodal corpus. The context of the multimodal interaction is the popular game *Tom Clancy's The Division 2*. Findings of an initial engagement modelling experiment on this new dataset reveal that pixel information from the game footage can form efficient predictors of long-term player engagement. Without time embeddings, pixel information can be a strong predictor that is enhanced through fusion with gamepad actions to produce the best models, yet gamepad actions alone do not seem to be good predictors. One reason for poor gamepad models' performance could be the limited input size. Moreover, gamepad action logs lack the in-game context of the keypresses' effect on the game: for example, while gamepad inputs measure how often the player pressed the A button, the in-game effect of such an action may be very different depending on e.g. the avatar's currently held weapon. However, collecting in-game events in commercial games requires access to the game engine which may be unavailable due to intellectual property concerns. Therefore, the current experiment serves another purpose: to gauge to which degree data from player actions that respect current industry practices can be useful for affect modelling tasks.

The classification approaches presented in this initial study reveal that time embeddings are particularly efficient at capturing the long-term effects of player engagement with an average classification accuracy of 72%. We see three promising future research directions here. First, we plan to study and compare alternative learning paradigms such as preference learning which may capture informative local patterns—or changes [54, 55]—of engagement given the user modalities considered. Second, future studies will focus on different direct or indirect methods for integrating time within our multimodal models, including variants of LSTMs [15] and autoregressive models found in decision transformers [7]. Third, exploring other time embeddings with more granular time partitions (compared to the current 20-minute increments) may lead to breakthroughs in time conditioning for long-term affect prediction.

A core limitation of our first engagement modelling experiments is the baseline method we employ to derive the ground truth labels for the time windows of gameplay. Following current approaches in processing shorter gameplay sessions [26, 27], we use the mean annotation value of a 1-hour trace to split windows into low or high engagement and leave ambiguous ones too close to the mean out of the train/test data. This approach is beneficial as it produces an almost equal split between class labels. However, summarising an entire 1-hour annotation session into one mean value overlooks possible habituation effects and the inherent subjectivity biases of human annotators [55], among many other factors. Furthermore, annotating lengthy audiovisual content can cause cognitive fatigue due to the mental exhaustion of the annotators [49], which in turn can affect the quality of the resulting trace. In future work, more nuanced ways of deriving classes should be explored, e.g., via a dynamically adjusted mean value derived from a moving time window of the trace. Initial experiments with a dynamic splitting criterion resulted in unbalanced datasets which in turn caused predictive models to underperform. Future work should explore signal processing approaches for deriving a more nuanced ground truth as well as improving algorithmic processes for modelling it. Another direction for future work that would address this issue is eschewing engagement classification altogether and treating consequent time windows in an *ordinal* fashion [54, 55]. In such a treatment, the goal is to predict only whether the mean engagement between consequent time windows is (sufficiently) different, i.e. escalating or deescalating, which would discount for any long-term habituation or anchoring effects. We foresee several future directions for improving input data or engagement trace processing.

While the full extent of *The Division 2* corpus offers access to more modalities (including physiological signals and eye tracking), this initial study only focused on frame and gamepad action modalities. Inspired by earlier work [26, 27] we assume that in-game footage pixels combined with in-game actions would provide sufficient information for a model to predict player engagement accurately. While findings do corroborate our assumptions, including more modalities will likely improve the models' predictive power.

While *The Division 2* dataset is not currently accessible, our short-term plan is to release the dataset and therefore encourage more research on the study of player engagement modelling via multimodal signals in a commercial-standard game environment.

7 CONCLUSIONS

The purpose of this paper is two-fold: (a) to introduce a novel dataset of long-term gameplay affect annotation traces that contains multiple modalities, and (b) to offer some initial suggestions and experiments on how such long-term affect traces can be processed and modelled. The extensive dataset analysed in this paper leverages two modalities—gameplay image frames and players' interaction data—but future work can explore more modalities already available in *The Division 2* corpus. Experiments used a simple splitting criterion from the literature [35] to turn time-continuous annotations into binary classes, and demonstrated that gameplay frames can be good affect predictors as indicated in earlier studies [26, 27]. Our core findings suggest that long-term affect prediction is possible with high degrees of accuracy when time embeddings are injected to the model. The methods introduced here are generic and applicable to any study investigating long-term affect modelling.

ACKNOWLEDGMENTS

This project has received funding from the European Union's Horizon 2020 programme under grant agreement No 951911.

REFERENCES

[1] Babak Joze Abbaschian, Daniel Sierra-Sosa, and Adel Elmaghraby. 2021. Deep learning techniques for speech emotion recognition, from databases to models. *Sensors* 21, 4 (2021), 1249.

[2] Sharmeen M Saleem Abdullah Abdullah, Siddeeq Y Ameen Ameen, Mohammed AM Sadeeq, and Subhi Zeebaree. 2021. Multimodal emotion recognition using deep learning. *Journal of Applied Science and Technology Trends* 2, 02 (2021), 52–58.

[3] James J Appleton, Sandra L Christenson, Dongjin Kim, and Amy L Reschly. 2006. Measuring cognitive and psychological engagement: Validation of the Student Engagement Instrument. *Journal of school psychology* 44, 5 (2006), 427–445.

[4] Nicolas Beaudoin-Gagnon, Alexis Fortin-Côté, Cindy Chamberland, Ludovic Lefebvre, Jérémy Bergeron-Boucher, Alexandre Campeau-Lecours, Sébastien Tremblay, and Philip L Jackson. 2019. The FUNii database: A physiological, behavioral, demographic and subjective video game database for affective gaming and player experience research. In *Proceedings of the 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 1–7.

[5] Uta K Bindl and Sharon K Parker. 2010. 32 Feeling good and performing well? Psychological engagement and positive behaviors at work. *Handbook of employee engagement: Perspectives, issues, research and practice* 385 (2010).

[6] Ran Breuer and Ron Kimmel. 2017. A deep learning perspective on the origin of facial expressions. *arXiv preprint arXiv:1705.01842* (2017).

[7] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. 2021. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems* 34 (2021), 15084–15097.

[8] Mohamed Dahmane and Jean Meunier. 2011. Emotion recognition using dynamic grid-based HoG features. In *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. IEEE, 884–888.

[9] Natalie Jane de Vries, Jamie Carlson, and Pablo Moscato. 2014. A data-driven approach to reverse engineering customer engagement models: Towards functional constructs. *PLOS ONE* 9, 7 (2014).

[10] Soumia Dermouche and Catherine Pelachaud. 2019. Engagement modeling in dyadic interaction. In *Proceedings of the International Conference on Multimodal Interaction*. 440–445.

[11] Metehan Doyran, Arjan Schimmel, Pınar Baki, Kübra Ergin, Batıkan Türkmen, Almıla Akdağ Salah, Sander CJ Bakkes, Heysem Kaya, Ronald Poppe, and Albert Ali Salah. 2021. MUMBAI: multi-person, multimodal board game affect and interaction analysis dataset. *Journal on Multimodal User Interfaces* (2021), 1–19.

[12] Oliver Dunn. 2012. Multiple comparisons among means. *J. Amer. Statist. Assoc.* 56 (2012), 52–64. Issue 293.

[13] Jing Fan, Dayi Bian, Zhi Zheng, Linda Beuscher, Paul A Newhouse, Lorraine C Mion, and Nilanjan Sarkar. 2016. A robotic coach architecture for elder care (ROCARE) based on multi-user engagement models. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 25, 8 (2016), 1153–1163.

[14] João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. 2014. A survey on concept drift adaptation. *ACM computing surveys* 46, 4 (2014), 1–37.

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

[16] Yan Huang, Stefanus Jasin, and Puneet Manchanda. 2019. "Level up": Leveraging skill and engagement to maximize player game-play in online video games. *Information Systems Research* 30, 3 (2019), 927–947.

[17] Kostas Karpouzis, Georgios N Yannakakis, Noor Shaker, and Stylianos Asteriadis. 2015. The platformer experience dataset. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII)*. 712–718.

[18] Michael Kearns and Dana Ron. 1999. Algorithmic Stability and Sanity-Check Bounds for Leave-One-out Cross-Validation. *Neural Computation* 11, 6 (1999).

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

[20] Jean Kossaifi, Georgios Tzimiropoulos, Sinisa Todorovic, and Maja Pantic. 2017. AFEW-VA database for valence and arousal estimation in-the-wild. *Image and Vision Computing* 65 (2017), 23–36.

[21] Jean Kossaifi, Robert Walecki, Yannis Panagakis, Jie Shen, Maximilian Schmitt, Fabien Ringeval, Jing Han, Vedhas Pandit, Antoine Toisoul, Björn Schuller, et al. 2019. Sewa db: A rich database for audio-visual emotion and sentiment research in the wild. *IEEE transactions on pattern analysis and machine intelligence* 43, 3 (2019), 1022–1040.

[22] Celine Latulipe, Erin A Carroll, and Danielle Lottridge. 2011. Love, hate, arousal and engagement: exploring audience responses to performing arts. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1845–1854.

[23] Joseph LeDoux. 1998. *The emotional brain: The mysterious underpinnings of emotional life*. Simon and Schuster.

[24] Eva Lieskovská, Maroš Jakubec, Roman Jarina, and Michal Chmulík. 2021. A review on speech emotion recognition using deep learning and attention mechanism. *Electronics* 10, 10 (2021), 1163.

[25] Phil Lopes, Georgios N Yannakakis, and Antonios Liapis. 2017. Ranktrace: Relative and unbounded affect annotation. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*. 158–163.

[26] Konstantinos Makantasis, Antonios Liapis, and Georgios N Yannakakis. 2019. From pixels to affect: A study on games and player experience. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*.

[27] Konstantinos Makantasis, Antonios Liapis, and Georgios N Yannakakis. 2021. The pixels and sounds of emotion: General-purpose representations of arousal in games. *IEEE Transactions on Affective Computing* (2021).

[28] Konstantinos Makantasis, David Melhart, Antonios Liapis, and Georgios N Yannakakis. 2021. Privileged information for modeling affect in the wild. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*. IEEE.

[29] Konstantinos Makantasis, Kosmas Pinitas, Antonios Liapis, and Georgios N. Yannakakis. 2022. The Invariant Ground Truth of Affect. In *Proceedings of the ACII Workshop on What's Next in Affect Modeling?*

[30] Hector P Martinez, Yoshua Bengio, and Georgios N Yannakakis. 2013. Learning deep physiological models of affect. *IEEE Computational intelligence magazine* 8, 2 (2013), 20–33.

[31] Héctor P Martínez and Georgios N Yannakakis. 2014. Deep multimodal fusion: Combining discrete events and continuous signals. In *Proceedings of the 16th International conference on multimodal interaction*. 34–41.

[32] Paris Mavromoustakos-Blom, David Melhart, Antonios Liapis, Georgios N. Yannakakis, Sander Bakkes, and Pieter Spronck. 2023. Multiplayer Tension In the Wild: A Hearthstone Case. In *Proceedings of the International Conference on the Foundations of Digital Games*.

[33] David Melhart, Daniele Gravina, and Georgios N Yannakakis. 2020. Moment-to-moment Engagement Prediction through the Eyes of the Observer: PUBG Streaming on Twitch. In *Proceedings of the 15th International Conference on the Foundations of Digital Games*. 1–10.

[34] David Melhart, Antonios Liapis, and Georgios N Yannakakis. 2019. PAGAN: Video affect annotation made easy. In *Proceedings of the 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 130–136.

[35] David Melhart, Antonios Liapis, and Georgios N Yannakakis. 2022. The arousal video game annotation (AGAIN) dataset. *IEEE Transactions on Affective Computing* 13, 4 (2022), 2171–2184.

[36] David Melhart, Julian Togelius, Benedikte Mikkelsen, Christoffer Holmgård, and Georgios N Yannakakis. 2023. The Ethics of AI in Games. *IEEE Transactions on Affective Computing* (2023).

[37] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. 2017. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing* 10, 1 (2017), 18–31.

[38] Hong-Wei Ng, Viet Dung Nguyen, Vassilios Vonikakis, and Stefan Winkler. 2015. Deep learning for emotion recognition on small datasets using transfer learning. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*. 443–449.

[39] Anh Nguyen, Khoa Pham, Dat Ngo, Thanh Ngo, and Lam Pham. 2021. An analysis of state-of-the-art activation functions for supervised deep neural network. In *Proceedings of the International Conference on System Science and Engineering (ICSSE)*. IEEE, 215–220.

[40] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*.

[41] Rosalind W Picard. 2000. *Affective computing*. MIT press.

[42] Kosmas Pinitas, Konstantinos Makantasis, Antonios Liapis, and Georgios N Yannakakis. 2022. RankNEAT: outperforming stochastic gradient search in preference learning tasks. In *Proceedings of the Genetic and Evolutionary Computation Conference*. 1084–1092.

[43] Kosmas Pinitas, Konstantinos Makantasis, Antonios Liapis, and Georgios N Yannakakis. 2022. Supervised contrastive learning for affect modelling. In *Proceedings of the International Conference on Multimodal Interaction*. 531–539.

[44] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. 2013. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *Proceedings of the 10th IEEE International Conference and Workshops on Automatic Face and Gesture recognition (FG)*. 1–8.

[45] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of computer vision* 115, 3 (2015).

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

[47] Nicu Sebe, Ira Cohen, and Thomas S Huang. 2005. Multimodal emotion recognition. In *Handbook of pattern recognition and computer vision*. World Scientific, 387–409.

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

[49] Alexis D. Souchet, Stéphanie Philippe, Domitile Lourdeaux, and Laure Leroy. 2022. Measuring Visual Fatigue and Cognitive Load via Eye Tracking while Learning with Virtual Reality Head-Mounted Displays: A Review. *International Journal of Human–Computer Interaction* 38, 9 (2022), 801–824.

[50] Choo-Yee Ting, Wei-Nam Cheah, and Chiung Ching Ho. 2013. Student engagement modeling using bayesian networks. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*. IEEE, 2939–2944.

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[52] Vikram Voleti, Alexia Jolicoeur-Martineau, and Christopher Pal. 2022. Masked conditional video diffusion for prediction, generation, and interpolation. *arXiv preprint arXiv:2205.09853* (2022).

[53] Su Xue, Meng Wu, John Kolen, Navid Aghdaie, and Kazi A Zaman. 2017. Dynamic difficulty adjustment for maximized engagement in digital games. In *Proceedings of the 26th International Conference on World Wide Web Companion*. 465–471.

[54] Georgios N Yannakakis, Roddy Cowie, and Carlos Busso. 2017. The ordinal nature of emotions. In *Proceedings of the 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 248–255.

[55] Georgios N Yannakakis, Roddy Cowie, and Carlos Busso. 2018. The ordinal nature of emotions: An emerging approach. *IEEE Transactions on Affective Computing* 12, 1 (2018), 16–35.

[56] Georgios N Yannakakis, Héctor P Martínez, and Arnav Jhala. 2010. Towards affective camera control in games. *User Modeling and User-Adapted Interaction* 20 (2010), 313–340.

[57] Georgios N Yannakakis and Ana Paiva. 2014. Emotion in games. *Handbook on affective computing* 2014 (2014), 459–471.

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

[59] Yuhao Zhang, Md Zakir Hossain, and Shafin Rahman. 2021. DeepVANet: A Deep End-to-End Network for Multi-Modal Emotion Recognition. In *Procceedings of the 18th International Conference on Human-Computer Interaction (INTERACT)*. 227–237.

[60] Wenming Zheng, Hao Tang, Zhouchen Lin, and Thomas S. Huang. 2010. Emotion Recognition from Arbitrary View Facial Images. In *Proceedings of the 11th European Conference on Computer Vision: Part VI*. Springer-Verlag, Berlin, Heidelberg, 490–503.