Multimodal Technologies and Interaction

MDPI

Review

From Cues to Engagement: A Comprehensive Survey and Holistic Architecture for Computer Vision-Based Audience Analysis in Live Events

Marco Lemos *,† **, Pedro J. S. Cardoso** † **and João M. F. Rodrigues** †

NOVA LINCS and ISE, Universidade do Algarve, 8005-139 Faro, Portugal; pcardoso@ualg.pt (P.J.S.C.); jrodrig@ualg.pt (J.M.F.R.)

* Correspondence: mmlemos@ualg.pt

† These authors contributed equally to this work.

Abstract

The accurate measurement of audience engagement in real-world live events remains a significant challenge, with the majority of existing research confined to controlled environments like classrooms. This paper presents a comprehensive survey of Computer Vision AI-driven methods for real-time audience engagement monitoring and proposes a novel, holistic architecture to address this gap, with this architecture being the main contribution of the paper. The paper identifies and defines five core constructs essential for a robust analysis: Attention, Emotion and Sentiment, Body Language, Scene Dynamics, and Behaviours. Through a selective review of state-of-the-art techniques for each construct, the necessity of a multimodal approach that surpasses the limitations of isolated indicators is highlighted. The work synthesises a fragmented field into a unified taxonomy and introduces a modular architecture that integrates these constructs with practical, business-oriented metrics such as Commitment, Conversion, and Retention. Finally, by integrating cognitive, affective, and behavioural signals, this work provides a roadmap for developing operational systems that can transform live event experience and management through data-driven, real-time analytics.

Keywords: affective computing; crowd engagement; HCI; real-time engagement; real-time analytics; computer vision; emotion recognition; crowd behaviour; event monitoring

1. Introduction

Artificial intelligence (AI) and big data analysis are two examples of digital technologies that can enhance the user experience at events and provide audience engagement feedback to promoters, enabling them to adjust the event to attendees' expectations. This facilitates personalised experiences for events and other activities and improves resource management.

Engagement can be defined in various ways; the emphasis here is on in-person event participation, rather than virtual or online interactions. This is commonly referred to as audience engagement, which describes an event's capacity to maintain attendees' interest and encourage their involvement. Engaging real-world events are designed to be captivating and compelling, meaning they should capture the audience's attention immediately and keep it throughout the event, or at least during key moments.

It is important to define two concepts, crowds and groups [1]. (i) A group is a collection of individuals, ranging in size from two to hundreds, who are present together at any given

Academic Editor: Qianling Jiang

Received: 10 November 2025
Revised: 28 December 2025
Accepted: 6 January 2026
Published: 8 January 2026

time and engaging in social contact. Its members move in a similar direction and at a similar speed, which brings them closer to one another. Multiple groups can cohabitate during an event. Conversely, a (ii) crowd (or mass) is a special, huge gathering of people who are physically present in the same place. It typically arises when individuals who share a common objective unite as a single entity, losing their individuality and assuming the characteristics of the crowd entity.

It is crucial to emphasise that while wearable biosensors, such as electroencephalography (EEG) devices, electrodermal activity (EDA) equipment, or heart rate monitors, are often regarded as the gold standard for measuring physiological correlates of engagement and emotional arousal, their deployment in real-world live events presents significant practical limitations. These devices are inherently intrusive, requiring physical attachment to attendees, which can alter natural behaviour, induce discomfort, and scale poorly in large, dynamic crowds. Moreover, logistical challenges such as device distribution, maintenance, data synchronisation, and participant compliance render biosensors infeasible for most real-world event settings, where minimal disruption and seamless integration are paramount.

In contrast, Computer Vision-based methods offer a non-intrusive and scalable alternative, capable of capturing behavioural and affective cues from a distance without direct attendee contact. Although vision-based indicators, such as facial expressions, head pose, and body movement, are proxies for internal states and may not match the precision of direct physiological measurement, they provide a practical and ethically preferable means of inferring engagement in ecologically valid environments. The trade-off lies in accepting a slight reduction in individual-level physiological accuracy in exchange for the ability to monitor large audiences in real time, across diverse and unconstrained settings. This approach aligns with developing deployable systems that strike a balance between scientific rigour and real-world applicability, while adhering to privacy-preserving and minimally invasive design principles. Following the above, this paper focuses primarily on Computer Vision methods and models, while psychological theories of crowd behaviour are beyond its scope. For that, the reader is referred to, e.g., Varghese and Thampi's work [2]. Additionally, in the remainder of the text, the word crowd will be used generically to refer to crowds and groups until otherwise specified.

The main contribution of this paper is the introduction of a unified multimodal engagement architecture, grounded in business-oriented engagement metrics. This architecture comprises five key constructs: Attention, Emotion and Sentiment, Body Language, Scene Dynamics, and Behaviours. Together, these constructs offer a holistic and modular approach that surpasses the limitations of isolated indicators commonly employed in previous studies. In addition, the architecture integrates practical metrics such as Commitment, Conversion, Retention, and Feedback, thereby aligning engagement analysis with the strategic goals of event management and marketing alignment that are notably absent from existing academically focused methodologies. A complementary contribution is the comprehensive survey of recent advancements in AI-driven Computer Vision for audience engagement monitoring. This survey represents a structured review of methods applied to real-world live events, addressing complex challenges such as crowd dynamics and occlusion.

In a nutshell, the paper delivers a unified five-construct engagement architecture with mathematical formulations for real-time metric computation, the association with the Binary Engagement Model proof-of-concept (see [3]), and a prioritised roadmap addressing critical implementation gaps in standardised datasets, occlusion handling, and privacy-preserving methods for operationalising the architecture in live group- and crowd-based events.

Following this section's initial introduction, Section 2 provides a comprehensive survey of AI-driven Computer Vision approaches to crowd engagement monitoring in

real-world events. Section 3 summarises some models and methods to implement the standard constructs identified in the previous section and ends with the authors' proposed constructs with the identified blocks needed to implement a complete audience engagement architecture. The paper is concluded with a discussion and a look to future work in Section 4.

2. Selective Review of Recent Advancements in AI-Driven Audience Engagement Monitoring

This section is divided into three subsections. The first focuses on the review methodology used, the second on the existing state-of-the-art audience/crowd engagement detection systems, and the last provides a brief list of datasets related to engagement detection.

2.1. Review Methodology

This study presents a selective review of automatic (AI-based) engagement detection methods/models/frameworks in audiences (crowds) in real-world environments, focusing on non-intrusive methods. Given the practical constraints of real-world events, the comprehensive survey emphasises on Computer Vision (complemented by sound analysis) techniques, while excluding less feasible approaches such as wearables or expensive biosensors (e.g., those embedded in bracelets). The study also includes the main list of datasets available for this task.

It is important to clarify that this work does not aim to conduct a systematic literature review. *Systematic reviews* follow a rigorous, protocol-driven methodology designed to answer a specific research question by identifying, critically evaluating, and synthesising all relevant evidence—often incorporating statistical techniques such as meta-analysis. In contrast, the objective here is to carry out a *comprehensive survey*: a broad, descriptive overview of the field that summarises existing studies, trends, and developments. This approach does not require a formal assessment of study quality or adherence to a strictly predefined and reproducible search protocol.

In essence, while a systematic review seeks a definitive, evidence-based conclusion by treating existing studies as data, a comprehensive survey aims to map the research landscape and provide a narrative synthesis of the current state of knowledge. This is the scope and intention of the present paper. Following the above, by evaluating current methodologies, this study aims to provide insights into the state of the field and highlight viable solutions for real-world applications. The papers selected for analysis in this study met the following the criteria:

(i) Related to real-world engagement applications (not virtual/online applications);
(ii) Published in a journal or a conference;
(iii) Written in English;
(iv) Accessible online;
(v) Provide systematic information on computing methods;
(vi) Published between 2023 and October 2025.

The information was primarily gathered from Google Scholar, with additional references from, in order, Scopus, IEEE Xplore, and the ACM Digital Library. The following combination of keywords was used:

(i) Crowd engagement detection;
(ii) Automatic engagement detection AND (computer vision OR sound analysis) AND (non-intrusive OR contactless);
(iii) Crowd engagement AND (real-world OR in-the-wild) AND (facial expression OR gaze tracking OR acoustic sensing);

(iv) Audience attention monitoring AND (vision-based OR audio-based) NOT (wearable OR biosensor);
(v) Engagement recognition in crowds AND (dataset OR benchmark).

Following the initial results obtained through the selected keyword combination, a second (subjective) filtering stage was applied. This involved excluding publications that, based on the authors' experience in research and development, were deemed unlikely to be applicable or feasible in real-world scenarios.

As a small final note, it is important to clarify the references to sound or audio-related terms which appear in the search. Despite the paper's explicit focus on Computer Vision-based engagement estimation, this reflects the intentionally broad scope of our literature search, which was designed to capture relevant multimodal research where audio and visual analysis are integrated or co-developed. Including such studies provides an extra guarantee that all vision-based studies were screened.

2.2. Audience/Crowd Engagement Detection Systems

There is a notable gap in the literature on engagement detection in real-world events; most of it is focused on the detection of engagement in classrooms. Booth et al. [4] present a tutorial on engagement detection and its applications in learning. They define engagement as a multicomponent construct involving affective, cognitive, and behavioural components, influenced by context and time. The paper discusses traditional methods (self-reports, observer-based measures) and automated methods (sensors, machine learning) for measuring engagement, highlighting challenges like validity, scalability, and ethical concerns. It explores proactive designs (emotional, cognitive, and behavioural) to enhance engagement by optimising learning experiences and reactive designs that adapt in real time based on detected engagement levels. Examples include adaptive feedback systems and dynamic content adjustments. The paper emphasises the importance of interdisciplinary approaches and future research directions, such as integrating heterogeneous measures, leveraging wearable technologies, and blending proactive and reactive designs to improve engagement and learning outcomes.

On the specific topic of engagement in classrooms, several papers were selected, as most research to date (as mentioned above) has centred on classroom environments. To illustrate current approaches and methodologies, we selected a representative set of classroom-focused studies. The study by Lasri et al. [5] presents an approach to detecting the engagement levels of deaf and hard-of-hearing students in a classroom environment through facial expression recognition (FER) using a Deep Convolutional Neural Network (DCNN) model, specifically a fine-tuned VGG-16 [6] architecture. The engagement index (EI) was computed by multiplying the dominant emotions' probabilities, derived from the SoftMax output of the model, by their corresponding emotion weights, resulting in a concentration index (CI) that quantifies the students' engagement levels.

The ICAPD framework uses simAM-YOLOv8n model for detecting student cognitive engagement in classrooms [7]; it categorises engagement into five levels based on visual behaviours. The simAM-YOLOv8n model, enhanced with an attention mechanism, improves detection accuracy in dense classroom scenes. The study uses annotated classroom videos to train the model, demonstrating superior performance compared to other methods. This approach aims to help teachers analyse and adjust instruction based on real-time engagement data, offering a promising solution for enhancing classroom learning.

Sumer et al. [8] investigate student engagement using audiovisual recordings in real classroom settings. The authors used Computer Vision to classify engagement based on attentional (head pose) and affective (facial expressions) features, achieving AUCs (Area Under the Curve of the Receiver Operating Characteristic curve) of 0.620 and 0.720 for grades

8 and 12, respectively. Attention features outperformed affective ones, and combining both slightly improved performance. Personalising models with minimal person-specific data significantly enhanced accuracy. The study highlights the potential of automated engagement analysis but notes limitations like sample size and ethical considerations for real-world applications.

Also in a classroom environment, the model presented by Zhao et al. [9] is a lightweight facial expression recognition architecture specifically designed to detect student engagement through facial expressions in real-time classroom settings. It features several key components, including two standard convolution layers followed by batch normalisation and a Mish activation function, which enhance the model's ability to learn complex features. Central to its design are four group-in-bottleneck residual blocks that facilitate efficient feature extraction while keeping the parameter count low, complemented by attention mechanisms such as the Convolution Block Attention Module (CBAM) and Channel Attention Module (CAM) to improve feature representation. Engagement levels are computed by mapping the detected emotions to three categories—high, medium, and low engagement—based on the established relationship between specific facial expressions and corresponding engagement levels.

Vrochidis et al. [10] propose a six-layer deep learning framework to monitor audience engagement in online video events by jointly analysing video and audio streams. For video analysis, they employed HopeNet [11] for head pose estimation and JAA-Net [12] for emotion recognition, complemented by a RetinaFace [13]-based face detection module. In the audio domain, they used a DenseNet-121 [14] architecture trained on a custom dataset for event detection, including clapping, speech, and pauses. Their system operates on keyframes extracted from event recordings, detecting participants' head orientations, emotional states, and sound events to infer engagement levels minute-by-minute.

The study done at King Faisal University [15] uses machine learning and Computer Vision to monitor student engagement in real-time. By analysing visual cues such as body language and pose estimation, the system classifies students into three categories: Engaged, Not Engaged, and Partially Engaged. The system demonstrated high accuracy, providing valuable insights for educators to enhance teaching strategies. Ethical considerations such as privacy and bias are addressed. Qarbal et al. [16] present a two-level verification approach that combines head pose estimation with facial expression analysis. The first level uses a Convolutional Neural Network (CNN) trained on the Head Pose Image Database to classify head poses into four categories, indicating visual attention. The second level utilises pre-trained models, ResNet50 and EfficientNet, to analyse facial expressions using the DAISEE (Dataset for Affective States in E-Environments) dataset.

Teotia et al. [17] present another very interesting study that focuses on the effectiveness of Vision Language Models (VLMs) in detecting classroom-specific emotions like engagement and distraction. Through empirical studies using classroom behaviour and emotion recognition datasets, it is found that VLMs struggle with engagement detection due to their nuanced and context-dependent nature. Nevertheless, the study highlights the need for more classroom-specific training data and common-sense reasoning frameworks to improve VLM performance in this area. The research addresses a gap in the existing literature and suggests potential avenues for enhancing VLM capabilities in educational settings.

Although it is not the primary topic, Human–Robot Interaction (HRI) deserves a quick mention, as its methods are relevant to this area of research. For this, in this context, Sorrentino et al. [18] provide a systematic study that examines HRI, stressing its ambiguity and the difficulties in defining and evaluating it. The review analyses 28 studies, identifying common definitions, methods, and features used for automatic engagement detection. It emphasises the need for a clear definition of engagement, better annotation procedures,

and the integration of multimodal features. The review also discusses the limitations of current engagement prediction models and suggests improvements, such as incorporating temporal information and deploying models in real-time interactions. The authors call for more robust, context-aware, and personalised engagement assessment frameworks to enhance the interaction quality between humans and robots. Ravandi et al. [19] examine deep learning methods for detecting user engagement in HRI, finding that CNNs are the most common approach, primarily using visual cues like facial features and pose. The review identifies key gaps, including the need for context-specific datasets and more research on temporal dynamics and non-social robots.

Going back to student engagement detection based on Computer Vision, [20] presents a systematic literature review. The authors examine and categorise the types of student engagement detected, the learning contexts in which detection occurs, and the methods employed for such detection; they also analyse the types of data sources, datasets, and features used, as well as the preprocessing and feature engineering techniques applied to enhance model accuracy. Bei et al. [21] introduce a transformer-based model for engagement recognition in videos. The model uses three independent class tokens to extract features from ocular, head, and trunk regions. Inspired by human observation, it also performs disengagement behaviour recognition at frame and video levels, which guides the model and improves interpretability. The model achieves state-of-the-art performance on DAiSEE and on EmotiW-EP datasets.

Despite producing an article that focuses on online environments (the only one presented in the paper), Wang et al. [22] present a real-time low-cost framework for monitoring student engagement in E-learning environments. Utilising the MediaPipe library, their model extracts and analyses multi-dimensional facial features, head pose, eye blinking, gaze direction, and smiles to assess behavioural and emotional engagement. The proposed model includes a dynamic threshold for blink detection and achieves high accuracy and computational efficiency, outperforming several complex deep learning and traditional methods. This enables the system to run effectively on low-resource edge devices without specialised hardware, making it a practical and scalable solution for providing immediate feedback to educators and enhancing the online learning experience. Nevertheless, as with many (most, if not all) online applied systems, it only works with a "frontal/detailed" image of the (body and/or) face.

To summarise, between classroom-focused engagement detection systems and those for real-world live events, what is transferable is the core multimodal idea, i.e., using Computer Vision cues like facial expressions and head pose/gaze for attention, and body posture to infer affective, cognitive, and behavioural engagement, often via deep learning models originally developed and validated in educational or HRI contexts. Classroom work provides reusable building blocks such as FER-based engagement indices, head-pose-based attention estimation, and multimodal fusion strategies that can be adapted as components of larger architectures. However, much is not directly transferable, as classroom systems assume relatively small groups, frontal views, limited occlusions, stable lighting, and individual-level labels, whereas live events involve large crowds, severe occlusion, moving and distant targets, heterogeneous behaviours, and a lack of fine-grained ground truth, demanding crowd-centric construct dimensions like density, collective motion, and high-level behaviours (commitment, conversion, retention, feedback, anomalies) plus specialised tracking and crowd-analysis methods that go beyond what classroom-focused models are designed to handle.

2.3. Crowd Engagement Datasets

A critical prerequisite for this research is the validation of suitable, publicly available datasets. Although numerous datasets exist for video-based behaviour analysis, none, to the best of our knowledge, are fully dedicated to engagement, except the OUC Classroom Group Engagement Dataset (OUC-CGE) [23]. Nevertheless, this dataset does not fit the purpose of engagement detection in "live events". Despite the above-mentioned scarcity, several datasets are available for tasks such as emotion and sentiment classification in images and video. Notable examples include the HAPPEI (HAPpy PEople Images) dataset [24] and the MED (Motion Emotion Dataset) [25]. The authors also published a comprehensive survey on affective computing databases in 2024 (see [26]). Within this specific domain, it is important to highlight a major bottleneck: the limited availability of public datasets containing videos of groups or crowds with high-quality emotional or sentiment annotations. Although datasets such as HAPPEI and MED address this gap to some extent, they remain among the few exceptions.

Turning to datasets related to activity detection and behavioural analysis, Table 1 presents a selection of prominent examples. While additional datasets exist, the ones included are, in the authors' view, the most suitable for supporting the development of the proposed engagement architecture, although it is important to note that none of them are specifically designed for engagement detection. Table 1 is structured across five dimensions: domain/scenes (describing the context captured), dataset size, type of annotation available, and relevant notes/comments.

Table 1. Summary of key datasets used in activity detection and behaviour analysis.

Dataset	Domain/Scenes	Size	Labels/Annotations	Evaluation Protocol and Metrics	Notes
UCSD Anomaly Detection [27]	Pedestrian walkway videos	34 train/36 test	Frame-level anomaly labels; pixel-level masks (subset)	Frame-level ROC/AUC; pixel-level AUC; EER	Classic small-scale benchmark; anomalies include bikes, skaters, vehicles
CUHK Avenue [28]	Campus avenue videos	16 train/21 test; 47 abnormal events	Frame labels + bounding boxes	Frame-level AUC; IoU-based localisation (VOC)	Provides rectangles for localisation tasks
UMN Unusual Crowd Activity [29]	Indoor/outdoor crowd panic scenario videos	3 sequences (1453, 4144, 2144 frames)	Frame-level normal to abnormal segments	Frame-level ROC/AUC; per-scene AUC	Sudden panic run abnormality
ShanghaiTech Campus [30]	Campus surveillance videos	270k training frames; 130 events	Pixel-level masks + frame labels	Frame-level AUC; pixel-level AUC	Large modern benchmark with diverse scenes
UT-Interaction [31]	Human–human interaction indoor/outdoor videos	20 sequences (1 min each), 6 classes	Temporal intervals and bounding boxes	Classification accuracy with LOSO CV	Interaction recognition dataset
UCF-Crime [32]	Real-world surveillance videos	1900 videos; 13 anomaly types	Video-level anomaly labels; some frame-level GT	Frame-level ROC/AUC; PR curves	Weakly labelled large-scale anomaly dataset
GENKI-4K [33]	Natural, unconstrained face images (4000 images)	Binary smile label (smiling and non-smiling)	5-fold cross-validation	Smaller dataset, focused specifically on smile detection (ollected from real-world scenarios)	
CelebA [34]	Celebrity face images (web-sourced, 200k images of 10k identities)	40 binary attributes (e.g., smiling, glasses, hair colour), identity labels, bounding boxes, 5 landmark locations; 162k train/20k val/20k test	Often used for attribute classification, face detection, landmark localisation	Large-scale, diverse, multiple tasks possible (attributes, recognition, landmark detection)	

In brief, despite the variety of existing datasets, there is a notable absence of datasets specifically dedicated to *engagement detection* and, crucially, to *quantifying the level of engagement*.

2.4. Discussion

The analysis in the present work, and others, reflects a multimodal architecture for an engagement model by integrating three core constructs: (i) *Emotion*, (ii) *Attention*, (iii) *Body Language, Scene Dynamics, and Behaviours*. The review of the existing literature establishes that a holistic assessment cannot rely on a single constructor; therefore, the model should capture cognitive focus through visual attention and gaze patterns, while concurrently analysing affective states via sentiment and emotion recognition from verbal and non-verbal cues. Furthermore, it incorporates kinetic features such as posture, gesture, and fidgeting to interpret body language, and it contextualises these individual signals within broader Scene Dynamics, including group interactions and environmental factors. Finally, these low-level features are synthesised into interpretable behaviours such as task persistence or social initiation, which serve as the direct, observable manifestations of engagement, thereby operationalising the abstract construct into measurable components for analysis.

3. Audience Engagement Model's Architectural Elements

Following the principle of the comprehensive survey, and having now defined the main constructs for the architecture, some existing methods and models are briefly described to extract the information for each of the constructs (again, focusing on Computer Vision AI-driven techniques).

3.1. Emotion

Emotion is a fundamental aspect of human behaviour and interactions as it can be used to detect and understand the level of engagement of an audience.

A reliable method for continuous affect recognition and prediction based on deep learning is presented by Stephen et al. [35]. Their method is unusual as it transfers the continuous valence–arousal dimensional space into discrete labels and learned facial expression representation. This deep learning-based method is structured with processes including regression model prediction, CNN feature extraction, and image scaling.

Andrey [36] proposes a comprehensive approach for facial emotion analysis in videos, comprising facial expression recognition, action unit (AU) analysis, and valence–arousal prediction. The author utilises pre-training lightweight CNNs on large facial datasets like VGGFace2 [37] and fine-tunes them on emotional labels from AffectNet. The paper by Nguyen et al. [38] presents an approach to affective behaviour analysis, focusing on valence–arousal prediction within the ABAW3 challenge framework. Leveraging deep learning techniques, the authors propose a two-stage model for continuous emotion estimation. In the first stage, feature extraction is performed using the RegNet architecture [39], followed by multimodal fusion with GRUs [40] and transformers [41]. The second stage involves temporal learning with GRU blocks and the application of local attention mechanisms to enhance model performance.

The study by Gupta et al. [42] introduces a deep learning-based engagement detection system that deploys a real-time learner leveraging FER. Addressing the challenges of online education, especially post-COVID-19, it measures student engagement by analysing facial expressions captured via webcams during online-only sessions. The FER process involves a multistage pipeline: facial detection using Faster R-CNN [43], extraction of 470 key facial points using a custom face-point extractor (MFACEXTOR), and emotion classification through CNNs. The emotions happy, sad, angry, neutral, afraid, and surprised are categorised as engaged or disengaged states using a calculated EI. This score is generated by

the deep learning model during facial emotion classification. Each emotion (e.g., happy, neutral, or angry) has an associated EP value based on the likelihood of its presence in the observed facial expression and weight of emotion (WE), which is a predefined weight assigned to each emotion, reflecting its impact on engagement. Positive emotions (e.g., happy or neutral) are given higher weights as they are more indicative of engagement, while negative emotions (e.g., sad or angry) are assigned lower weights.

Juliette et al. [44] employ video measures in desktop settings to track arousal and valence levels through facial expressions and physiological reactions to stressors. To generalise arousal and valence detection, the models underwent training on a variety of tasks. SMOTE was used to balance the classes, and grid search was used to improve the hyperparameters. MRMR [45] was used to reduce features, and several classification models were used. The training/validation and testing sets of the dataset comprised video measures that recorded individuals' physiological reactions and facial expressions in response to stress.

The study by Lorenzo et al. [46] explores valence and arousal estimation from neuromorphic vision data using event cameras, which excel at capturing subtle and rapid facial micro-movements. Training was performed on simulated neuromorphic data derived from the AFEW-VA [47] dataset, where RGB videos annotated with valence and arousal values were converted into event streams using the V2E [48] simulator. Manojkumar and Helen [49] introduce a computational model for crowd emotion analysis based on valence–arousal dimensions. It extracts features like crowd density, motion variance, and a novel enthalpy metric from video surveillance. These inputs are processed through a fuzzy inference system to classify collective emotional states. Implemented with a YOLOv8 DarkNet framework, the model demonstrates enhanced accuracy in identifying high-arousal, negative-valence scenarios indicative of potential threats, outperforming traditional methods.

3.2. Attention

Attention is a fundamental aspect of human behaviour and interactions, and it can be used to detect and understand the level of engagement of an audience. Some challenges in this area are the need to identify the specific target of attention, particularly when that target is moving (e.g., a speaker walking across a stage). This subsection addresses these challenges.

Attention can be assessed through head pose estimation, which determines whether a person is looking at a given target (object or individual). It also requires identifying, often in real time, what is being given attention. This subsection addresses both issues.

Hempel et al. [50] introduce a novel method for head pose estimation from single images using a continuous 6D rotation matrix representation. Traditional methods for head pose estimation often use discrete binning or quaternion representations, which can lead to ambiguity and reduced performance. The proposed method avoids these pitfalls by representing rotations with a 6D vector that is subsequently mapped to a 3×3 rotation matrix, ensuring orthogonality and reducing prediction errors. In Hempel et al. [51], the same authors continue a previous project [50], proposing a novel method for robust and unconstrained head pose estimation, addressing full-range rotation challenges. The model uses a geodesic loss function within the Special Orthogonal Group to stabilise learning and ensure precise predictions.

Focusing on location estimation, Reich [52] proposes a monocular 3D multi-object tracking framework that uses an Extended Kalman Filter to estimate object trajectories in a 3D space based on monocular video input. Unlike typical monocular approaches that operate solely in the 2D image domain, the method focuses on recovering accurate 3D coordinates for each object, which are critical for physically plausible tracking. Track

initialisation is performed using 3D bounding boxes generated by a monocular 3D object detector (based on CenterTrack [53]), which provides object distance, image coordinates, dimensions, and orientation. These measurements are transformed from polar to Cartesian coordinates to initialise a full 3D state vector including position (x, y, z), size, orientation, and kinematic variables like velocity and acceleration. The EKF is then used to propagate and update these states using 2D bounding boxes in subsequent frames, avoiding temporal correlation issues common in monocular 3D detections.

Hossain et al. [54] present a lightweight and efficient system for estimating the 3D coordinates of objects using image or video input. The system uses a stereo camera setup with two parallel cameras to capture depth information, allowing accurate spatial localisation of objects. Object detection is performed using the YOLOv3 model, trained on the COCO dataset [55], which provides bounding boxes for over 80 object classes. The system calculates vertical and horizontal angles from both camera views and applies trigonometric formulas to estimate the 3D position of each object relative to the primary camera. Yang et al. [56] present an enhanced monocular depth estimation framework. Their approach abandons real labelled images in favour of synthetic data with highly accurate depth annotations, thus avoiding common label noise found in traditional datasets. They train a powerful teacher model based on the DINOv2-G [57] encoder solely on synthetic data, and then generate pseudo-labels for a large-scale corpus of 62 million unlabelled real images. Student models (ViT-S, ViT-B, ViT-L, ViT-G) are subsequently trained on these pseudo-labelled images, allowing for excellent generalisation and efficient performance across model sizes ranging from 25 M to 1.3 B parameters.

3.3. Body Language, Scene Dynamics, and Behaviours

Body Language, Scene Dynamics, and Behaviours are essential for understanding human behaviour and interactions. These three constructs provide valuable insights into the dynamics of an audience and the interactions between the audience and, e.g., a speaker.

Sundararaman et al. [58] introduce a new benchmark for pedestrian tracking in dense crowds by focusing on head detection rather than full-body detection, which becomes unreliable due to occlusion at high densities. They present the Crowd of Heads Dataset (CroHD), a large-scale dataset consisting of 11,463 frames with over 2.2 million annotated head instances across nine scenes. To effectively detect small, occluded heads, they develop a novel head detector named HeadHunter, which leverages a ResNet-50 [59] backbone with a Feature Pyramid Network [60,61] and context-sensitive modules [62]. For tracking, they extend this detector into HeadHunter-T, which integrates a Particle Filter-based motion model and a colour histogram re-identification module. They also propose a new evaluation metric, IDEucl, that measures how consistently a tracker maintains identity across a trajectory in image space.

Deb et al. [63] address the complex task of tracking dispersed human crowd groups from aerial perspectives using a single bounding box, a challenge due to group reformation, occlusion, and shape variation. They introduce a new photorealistic dataset, the Unreal UAV Crowd Tracking (UUCT) dataset, designed in Unreal Engine and Airsim [64], featuring 70 long sequences across seven attributes with RGB, segmentation, and depth data. To overcome the limitations of existing single-object trackers (SOTs) when applied to dynamic crowd formations, the authors propose the Hybrid Motion Pooling (HyMP) architecture. HyMP augments DiMP [65] with spatial and temporal graph learning to capture human-to-human interactions and motion continuity, using Graph Convolutional Networks (GCNs) [66] and low-rank bilinear pooling.

Yeh et al. [67] propose a deep learning-based system for aerial crowd-flow analysis using stationary drones. The system integrates YOLOv5 [68] for real-time object detection

and StrongSort [69] with OSNet [70] for pedestrian tracking and re-identification. The approach enables the generation of three key visual outputs: crowd trajectory maps, hotspot maps, and flow direction maps, aiding in the assessment of crowd density, movement patterns, and congestion points. For dataset preparation, they employ LabelGo [71] for semi-automated annotation and utilise data from the MOT Challenge, Pexels, and iStock. Li et al. [72] propose a framework for video crowd localisation called GNANet, which introduces a multi-focus Gaussian Neighbourhood Attention (GNA) mechanism to effectively model spatial–temporal dependencies in surveillance footage. GNANet is designed to locate human head centres in crowded scenes by leveraging both scene modelling and context cross-attention modules. The GNA module efficiently captures long-range dependencies while preserving spatial topology, and the multi-focus mechanism enhances robustness to scale variations in head size due to perspective effects.

Ekanayake et al. [73] propose a deep learning model, the Multi-Column Multistage Bilinear Convolution Attention Network (MCMS-BCNN-Attention), for crowd density estimation and anomaly detection in surveillance video data. The architecture integrates three parallel feature extraction modules: a bilinear CNN attention mechanism based on transfer learning with DenseNet121 [14] and EfficientNetV2 [74], a multi-column [75] CNN, and a multistage CNN, enabling the fusion of deep spatial features for robust classification. The model classifies crowd density into five levels (very low to very high) and can also perform binary classification for anomaly detection. Evaluations on public datasets including UCSD [27] Ped1 and Ped2, PETS2009 [76], and UMN [77] Plaza1 and Plaza2 demonstrated superior performance, with the model achieving up to 99.10% accuracy and AUC scores of 1.00 for anomaly detection scenarios. The architecture benefits from enhanced spatial feature representation, efficient training convergence, and robustness across varied lighting and occlusion conditions, outperforming several state-of-the-art baselines in both multiclass density estimation and binary anomaly detection tasks.

Liu et al. [78] propose a crowd counting approach called the Dynamic-Refined Density Map Network, which addresses the inaccuracies in traditional density map generation due to uniform or heuristically derived ground truths (GTs). Their method consists of two primary components: Refine Net and Counting Net, which are jointly trained. Refine Net utilises a U-shaped architecture with a Regional Attention Module (RAM) to adaptively generate refined GTs that better represent real head sizes in varying contexts. Counting Net employs a multi-column architecture with co-prime dilation rate convolution groups, improving feature extraction continuity while mitigating information loss common in standard dilated convolutions.

Lin et al. [79] introduced the Multifaceted Attention Network (MAN), a crowd counting model designed to address large-scale variations in crowded scenes by enhancing transformer-based architectures. The model incorporates three attention mechanisms: Learnable Region Attention (LRA) for spatially adaptive local context, Local Attention Regularisation (LAR) to enforce consistency among local attention regions, and Instance Attention Loss (IAL) to reduce the impact of annotation noise. The architecture utilises a VGG-19 [6] backbone to extract features, followed by a transformer encoder enhanced with LRA, and a regression decoder to estimate the density map.

Wang et al. [80] propose the Hybrid Attention Network (HANet) for crowd counting, addressing both background noise suppression and scale variation adaptation simultaneously. The model architecture is composed of three main modules: a backbone (the first ten layers of a VGG16-BN [6] network pre-trained on ImageNet), a hybrid attention module (HAM), and a backend for density map generation. The hybrid attention module features parallel spatial and channel attention streams that progressively embed multi-scale context information across different cascaded levels. Qi et al. [81] propose a novel

multi-object tracking system called TraPeHat, designed to address the challenge of severe occlusion in densely crowded scenes. Recognising that the head is typically the most visible and least occluded part of the human body, the authors build their system around head tracking as a foundation for more reliable pedestrian tracking. The method operates online and starts by detecting and tracking pedestrian heads using a specially designed head detector and tracker. At the same time, full-body bounding boxes are detected using an enhanced Faster R-CNN-based [43] body detector, which includes a fusion strategy, patched non-maximum suppression, and an optional refinement module for improved robustness.

Pan et al. [82] propose a novel method for detecting pedestrians in densely populated and occluded environments using deep learning techniques. The approach integrates multi-scale feature maps and an occlusion detection module to improve robustness and accuracy. The occlusion detection component is composed of sub-modules that first identify occluded regions and then detect pedestrians within those regions. A combination of collaborative deep learning and dynamic part modelling is employed to extract pedestrian features from various angles and handle partial occlusions effectively.

Pai et al. [83] propose a method for classifying crowded scenes based on global motion patterns using a feature called the Histogram of Angular Deviations (HAD). The authors focus on distinguishing three types of crowd scenes—structured, semi-structured, and unstructured—based on the directional coherence of trajectories extracted from video frames. Motion trajectories are obtained using a generalised Kanade–Lucas–Tomasi [84] (gKLT) tracker, and average angular orientation features are computed for each trajectory. Zhang et al. [85] propose a novel method, CDEM-M, for crowd density estimation and geographic mapping by integrating surveillance video with GIS technologies. They developed a Crowd Semantic Segmentation Model (CSSM) based on the DeepLabv3+ [86] network and a Crowd Denoising Model (CDM) using Convolutional Neural Networks to accurately extract crowd features, especially in high-altitude and large-scene surveillance. The spatial mapping of crowds from video to geographic space was achieved using a homography matrix. To estimate crowd numbers, they constructed a back propagation (BP) [87] neural network optimised by a Genetic Algorithm, which used features like distance to camera, camera inclination, and geographic area of the crowd polygons.

Zhang et al. [88] propose JointTrack, an anchor-free joint head–body detection framework designed to improve pedestrian detection and tracking in extremely crowded scenes, where heavy occlusion is common. The core innovation lies in predicting heads and bodies simultaneously using a modified YOLOX [89] architecture, enhanced with a novel Joint SimOTA module that dynamically learns the spatial relationship between head and body regions without relying on fixed ratios. The authors train their model on a composite dataset derived from Crowdhuman [90], MOT20 [91], and HT21 [58], employing data augmentation techniques such as Mosaic and MixUp.

An approach to crowd density estimation using a Global Measuring Crowd Collectiveness (GMCC) metric is proposed by Mei et al. [92], who integrate intra-crowd and inter-crowd collectiveness to assess collective motion in drone-captured videos. The method builds on a global energy spread process derived from optical flow fields, modelling individuals as particles in a dynamic system. Intra-crowd collectiveness is quantified through velocity magnitude and directional consistency, while inter-crowd collectiveness measures cohesion among different crowd clusters. The model is robust to illumination changes due to an illumination-invariance strategy integrated into the optical flow estimation. Ranasinghe et al. [93] propose a novel crowd counting framework that leverages denoising diffusion probabilistic models [94,95] (DDPMs) to generate high-fidelity crowd density maps. Unlike traditional regression-based methods, which suffer from background noise and density loss due to the use of broad Gaussian kernels, CrowdDiff employs a

narrow kernel to improve the quality and accuracy of density maps. The model architecture consists of a conditional diffusion model trained to denoise corrupted density maps, with an auxiliary regression branch used only during training to enhance feature learning.

Akpulat et al. [96] propose a novel anomaly detection method for crowd scenes using a trajectory-based approach rooted in finite-time braid entropy (FTBE). The authors extract motion trajectories from optical flow using a particle advection method and cluster them via mean-shift based on endpoint positions. Each cluster is analysed through FTBE, a topological complexity metric derived from braid theory, along with motion vector magnitudes to form a feature vector. A fully connected deep neural network is trained to detect local anomalies, while a step braid entropy score (SBES) aggregates these features to identify global abnormalities. This method avoids the limitations of region-based and traditional tracking approaches by offering a holistic, location-independent view of crowd behaviour.

A deep learning method for detecting and tracking individuals in dense crowds was proposed by Badauraudine et al. [97], who aimed to enhance automated surveillance and mitigate crowd-related disasters. Using the PRISMA methodology, the authors analysed 4384 papers across five major databases, ultimately narrowing their focus to 13 primary studies published between 2019 and 2024. Finally, most recently, Martins et al. (2025) [98] provided a comprehensive analysis of the emotional body gesture recognition landscape by systematically reviewing extant literature surveys. Their work synthesises findings across numerous systematic reviews and surveys (SRoSs), critically examining published research on key aspects such as databases, methodological approaches, and overarching perspectives within the field.

3.4. Discussion and Audience Engagement Architecture Constructs

Based on the methods surveyed in Sections 2.2, 2.3, and 3.1–3.3, complemented by the authors' earlier papers [3,99], it is evident that a robust audience engagement framework necessitates a multimodal and multi-construct approach, as no single data stream is sufficient for a holistic assessment. The literature reveals a clear trend towards integrating complementary modalities: Emotion Level and Sentiment State are addressed through advanced deep learning techniques that predict valence and arousal from facial expressions, either directly from discrete labels or in a continuous space, often leveraging pre-trained CNNs fine-tuned on affective computing datasets. Concurrently, Attention Direction and Level are tackled via sophisticated head pose estimation models that provide a continuous 6D representation for robust, unconstrained gaze tracking, coupled with monocular 3D tracking frameworks to spatially contextualise a subject's focus within a dynamic scene. These constructs form a core cognitive–affective duo, quantifying internal states of focus and emotion.

Furthermore, a future architecture should be expanded to incorporate broader behavioural and contextual cues. Body Language analysis extends beyond the face to interpret intent through posture and gesture, while Scene Dynamics provide the macro-context, utilising advanced crowd analysis techniques such as head detection in dense environments (HeadHunter), trajectory-based motion modelling (HyMP, HAD), and density estimation (MCMS-BCNN-Attention, diffusion models) to understand group cohesion, motion patterns, and overall crowd density. Finally, these low-level signals are synthesised into interpretable Behaviours such as commitment, social interaction, or anomaly detection, which serve as the ultimate, observable manifestations of engagement. The discussed methods collectively demonstrate that while the field leverages powerful, often lightweight models for real-time applicability, the integration of these distinct constructs is paramount. This synergy allows a future Computer Vision-based architecture to move from isolated

feature extraction to a comprehensive, multi-faceted interpretation of engagement, essential for accurately capturing the complex phenomenon in real-world event settings.

In accordance with this, the previous three constructs were complemented and reorganised into five constructs, each of which must be implemented with the following (sub-)blocks (dimensions) in order to create a (fully) integrated audience engagement Computer Vision-based architecture:

(a) **Attention Direction and Level**. (a.1) Focus: Track the individual or crowd and determine whether their head orientation (gaze) is directed toward the event's primary focus, object, or person. (a.2) Heatmaps (a.2.1): The aggregated attention over a period of time and (a.2.2) the cumulative number of persons in a specific region close/or related to the object/person that is being engaged.

(b) **Emotion Level and Sentiment State**. (b.1) Emotion determines the degree and level of emotion by calculating arousal and valence, which is combined with the (b.2) Sentiment State, namely, negative, neutral, and positive.

(c) **Body Language**. (c.1) Actions and gestures convey intent and emotion, bridging language barriers to foster genuine connection (ex., clapping hands). (c.2) Pose communicates openness, confidence, and cultural respect, directly influencing rapport and the perceived sincerity of engagement. Together, they are the non-verbal foundation of trust and collaborative spirit in any event.

(d) **Scene Dynamics**. The computation includes the following: (d.1) Density: This represents the density of persons by section in the event. (d.2) Count: This tracks the number of persons assisting the event. (d.3) Motion patterns: Different motion patterns are tracked during the event, and the model also creates a heatmap with the motion patterns of groups and respective chronologies of how groups move inside the event.

(e) **Behaviours** represent observable actions and patterns that indicate a person's level of interest, attention, and interaction with a product, service, content, or another person. They can be divided into the following: (e.1) Commitment: The architecture monitors how many times a group or the crowd interacts with a product, service, or event. (e.2) Conversion: It monitors how many groups complete, or if the crowd completes, a pre-defined action/path. (e.3) Retention: It monitors how many groups return, or if the crowd returns, to a product, service, or event sector after their first visit. (e.4) Feedback: It monitors how many times a type of feedback was presented by a group or the crowd (e.g., waving or clapping hands). (e.5) Social: It monitors how many different groups interact with each other: (e.6) Odd Behaviours/Alerts: It can detect a subset or group of attendees whose behaviour deviates significantly from other groups or event sectors or indicates abnormal crowd activity.

Figure 1 provides an illustration and summary of the constructs along with their respective blocks, which are essential for constructing the overall architecture. It is important to mention that some of the constructs are more dedicated to the analysis of small groups or close views of the subjects of analysis, namely *Emotion Level and Sentiment State*, *Attention Direction and Level*, and *Body Language*. In those cases, it is expected that the main source of information will be the individuals' facial features. Similarly, *Scene Dynamics*, *Behaviours*, and, again, *Body Language* are more related to crowds and a global view of the attendees, where the main information comes from the "mass" (all persons as a single entity), instead of coming from the individuals' faces.

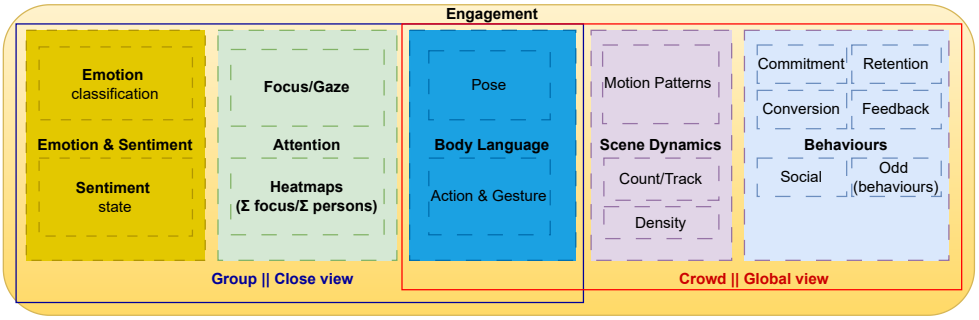

Figure 1. Illustration of the constructs required to implement the proposed engagement model.

The construct *Body Language* applies to both group and crowd analysis, but it is important to stress that the source of the information to build these respective blocks is different. At the group level, this requires face and pose tracking per individual; at the crowd level, *Body Language* is inferred from collective synchronisation or aggregate posture, where individuality is lost due to occlusion.

Building on the above, for small groups (close-view analysis), the engagement pipeline begins with the detection and tracking of each individual. From the tracked facial data, deep learning models are employed to extract emotional and sentiment-related features, estimating parameters such as valence, arousal, and sentiment polarity. Simultaneously, attention direction and intensity are inferred through head pose and gaze tracking, generating indicators of focus around the event's central object or person. These affective and attentional signals are further enriched with Body Language cues including gestures and posture, which provide additional evidence of individual intent and engagement. This process is applied to each person individually and subsequently integrated (ensembled) to produce a collective engagement profile for the group.

The pipeline for crowd-level or global (macro) analysis, begins with large-scale detection and tracking of individuals across the scene. Aggregated features are used to compute Scene Dynamics, including density maps, crowd counts, and group-level motion patterns, which are visualised through spatio-temporal heatmaps. In parallel, collective body language, such as synchronised movements or postural alignment, adds interpretative depth to the analysis of group intent. These global and low-level signals are then synthesised into high-level behavioural indicators, capturing phenomena such as crowd commitment, inter-group social interactions, and anomalies in movement or feedback patterns.

It is important to emphasise that, although two distinct pipelines are presented, in certain event formats (such as stage presentations involving a single speaker) a simplified pipeline may be sufficient. However, in many other contexts, particularly pavilion-based exhibitions, both pipelines operate in parallel to capture and integrate engagement across multiple point of interest (PoIs). This dual approach enables a more comprehensive and context-sensitive analysis of audience engagement.

Table 2 consolidates the methods previously discussed in the Sections 2.2 and 3 and reorganises them according to the proposed architecture constructs, highlighting their dual role as both state-of-the-art approaches and building blocks for engagement analysis.

Table 2. Crowd/scene-level methods mentioned in the manuscript.

Study/Authors	Methods	Overview	Applied to
Booth et al. [4]	Defines engagement as a multicomponent construct involving affective, cognitive, and behavioural components	Tutorial for engagement detection and its applications in learning	Engagement detection
Lasri et al. [5]	Facial expression recognition (FER) using a fine-tuned VGG-16 [6]	Engagement index (CI) computation	Engagement detection
Xu et al. [7]	YOLOv8n with SimAM attention	Analyse and adjust instruction based on real-time engagement data	Engagement detection
Sumer et al. [8]	Head pose and FER (vision + audio)	Investigate student engagement using audiovisual recordings in real classroom settings	Engagement detection
Zhao et al. [9]	Lightweight CNN, residual blocks, and CBAM/CAM	Real-time lightweight FER model for classroom use	Engagement detection
Vrochidis et al. [10]	HopeNet [11] (pose), JAA-Net [12] (AU), RetinaFace [13], DenseNet [14]	Six-layer deep framework fusing audio and video for online audience engagement	Engagement detection
Albohamood et al. [15]	YOLO-based pose and visual cues	Classifies students into Engaged, Not Engaged, and Partially Engaged.	Engagement detection.
Qarbal et al. [16]	Head pose estimation and facial expression analysis	Two-level verification of head pose and facial expression to strengthen engagement classification	Engagement detection
Teotia et al. [17]	Vision Language Models (VLMs)	Detection of classroom-specific emotions like engagement and distraction	Classroom-specific emotion detection
Sorrentino et al. [18]	-	Systematic review of HRI engagement.	Systematic review
Ravandi et al. [19]	Deep learning, primarily CNNs	Deep learning, mainly CNNs, detects engagement via visual cues in HRI	Engagement detection.
Qarbal et al. [20]	-	Engagement detection based on Computer Vision in learning environment	Systematic review
Bei et al. [21]	Region-focused behaviour capture transformer	Transformer model uses regional body features and disengagement behaviours to recognise video engagement	Engagement detection
Wang et al. [22]	Development of lightweight models that extract 3D facial landmarks using the MediaPipe library	Low-cost learner engagement detection framework for E-learning, combining behavioural (head posture, gaze, blinks) and emotional (smile detection) cues	Engagement detection.
Stephen et al. [35]	Regression model prediction, CNN feature extraction, and image scaling	Transfers continuous valence–arousal dimensional space to discrete labels	Valence and arousal estimation

Table 2. *Cont.*

Study/Authors	Methods	Overview	Applied to
Andrey [36]	Facial expression recognition, action unit (AU) analysis, and valence–arousal prediction	Combines FER, AU, and VA for fine-grained facial emotion analysis	Emotion analysis.
Nguyen et al. [38]	RegNet [39], GRU [40], and transformers [41]	Multimodal temporal fusion for continuous behaviour analysis	Behaviour analysis
Gupta et al. [42]	Faster R-CNN [43] and CNN FER pipeline	Online FER-based engagement index (EI) for E-learning.	Engagement detection
Juliette et al. [44]	SMOTE, MRMR [45], and grid search over multiple classifiers	Valence/arousal from facial and physiological responses across tasks	Valence and arousal estimation
Lorenzo et al. [46]	Neuromorphic event-based VA estimation	Uses event cameras (micro-movements) for valence/arousal estimation	Valence and arousal estimation
Manojkumar and Helen [49]	Fuzzy inference on crowd features, YOLOv8, DarkNet	Fuzzy inference model analyses crowd video to detect collective emotions via valence and arousal	Valence and arousal estimation
Hempel et al. [50]	Continuous 6D rotation matrix representation	Head pose estimation from single images	Head pose estimation.
Hempel et al. [51]	Unconstrained HPE with geodesic loss (SO(3))	Robust and unconstrained head pose estimation, addressing full-range rotation challenges	Head pose estimation
Reich [52]	Extended Kalman Filter	Recovers 3D trajectories from mono input for physically plausible tracking	3D multi-object tracking
Hossain et al. [54]	YOLOv3 and triangulation	Stereo angles and trigonometry to estimate 3D object positions	3D coordinate estimation
Yang et al. [56]	Depth Anything v2 (DINOv2-G [57] teacher; ViT S/B/L/G students)	Synthetic-to-real distillation for monocular depth	Monocular depth estimation
Sundararaman et al. [58]	Novel head detector named HeadHunter, which leverages ResNet-50 [59], Feature Pyramid Network [60,61], and context [62]	Head detection/tracking in dense crowds	Pedestrian tracking in dense crowds
Deb et al. [63]	HyMP: DiMP [65], GCNs [66], and bilinear pooling	Aerial crowd group tracking via hybrid motion pooling	Tracking dispersed human crowds
Yeh et al. [67]	YOLOv5 [68], StrongSort [69], and OSNet [70]	Aerial crowd flow: Trajectories, hotspots, flow maps from drones	Aerial crowd-flow analysis
Li et al. [72]	Gaussian Neighbourhood Attention and context cross-attention	Video crowd localisation via multi-focus attention (head centres)	Model spatial–temporal dependencies in surveillance footage

Table 2. *Cont.*

Study/Authors	Methods	Overview	Applied to
Ekanayake et al. [73]	DenseNet121 [14] / EfficientNetV2 [74], multi-column CNN [75], and multistage CNN	Classifies crowd density performs binary classification for anomaly detection	Crowd density estimation
Liu et al. [78]	Refine Net and Counting Net	Crowd counting approach which addresses inaccuracies in traditional density map generation	Crowd counting
Lin et al. [79]	VGG-19 backbone [6], transformer encoder, and LRA/LAR/IAL	Multifaceted attention for large scale variation counting	Crowd counting
Wang et al. [80]	VGG16-BN [6] and Hybrid Attention Module	Addresses both background noise suppression and scale variation adaptation simultaneously for crowd counting	Crowd counting
Qi et al. [81]	Head-first tracking and enhanced Faster R-CNN [43] body detector	Robust multi-object tracking under severe occlusion (head to body)	Multi-object tracking system
Pan et al. [82]	Multi-scale feature and occlusion detection module	Pedestrian detection under heavy occlusion via dynamic part modelling	Pedestrian detection
Pai et al. [83]	gKLT trajectories [84] and Histogram of Angular Deviations	Classifies crowd scenes by motion coherence	Classify crowded scenes
Zhang et al. [85]	DeepLabv3+ [86] CSSM, CNN denoising, and GA-optimised BP	Crowd density estimation and geographic mapping by integrating surveillance video with GIS technologies	Crowd density estimation
Zhang et al. [88]	Anchor-free joint head–body (YOLOX [89]) and Joint SimOTA	Joint head/body detection improves tracking under occlusion	Head–body detection
Mei et al. [92]	Global collectiveness via optical flow (illum.-invariant)	Crowd collectiveness metric across intra-/inter-crowd coherence	Measure crowd collectiveness
Ranasinghe et al. [93]	Diffusion model [94,95], narrow kernels, and SSIM-based fusion	Denoising diffusion for high-fidelity density maps	Crowd counting
Akpulat et al. [96]	Trajectories, finite-time braid entropy, and DNN classifier	Topological complexity (braid entropy) for local/global anomalies	Anomaly detection
Badauraudine et al. [97]	-	Systematic survey of dense-crowd detection/tracking methods	Systematic review
Martins et al. [98]	-	Emotional body gesture recognition	Systematic review
Binary Engagement Model [3]	YOLOv4 (heads), FaceNet, WHENet (pose), EVAm (VA)	Identity-aware binary engagement model (engaged/not, positive/negative); supports multi-cam, real-world use	Engagement detection

For a global mathematical representation, let us consider the constructs bifurcated into group (g) and crowd (c). As mentioned above, Attention (At) is composed of Focus (At_F), which within groups is denoted by $At^n_{F,g}$, where n is the group identification number, and is denoted at the crowd level by $At_{F,c}$, being $At_F = \mathbb{C}\left\{\bigoplus_n At^n_{F,g}, At_{F,c}\right\}$, where $\bigoplus_n$ denotes the combination of information from different groups and $\mathbb{C}$ represents the combined information from different constructs. Heatmaps are denoted by $At_{hm} = \mathbb{C}\left\{\bigoplus_n At^n_{hm,g}, At_{hm,c}\right\}$, an event ($e$) is defined by the combination of group and crowd information; consequentially, event attention is $At_e = \mathbb{C}\{At_F, At_{hm}\}$.

Let us now consider Emotion (E) to be composed of valence (E_V), arousal (E_A,) and valence–arousal plain (E_P). Within groups these are denoted by $E^n_{V,g}$, $E^n_{A,g}$, and $E^n_{P,g}$. At the crowd level, they are denoted by $E_{V,c}$, $E_{A,c}$, and $E_{P,c}$, respectively. At the event level, they are $E_{V,e} = \mathbb{C}\left\{\bigoplus_n E^n_{V,g}, E_{V,c}\right\}$, $E_{A,e} = \mathbb{C}\left\{\bigoplus_n E^n_{A,g}, E_{A,c}\right\}$, and $E_{P,e} = \mathbb{C}\left\{\bigoplus_n E^n_{P,g}, E_{P,c}\right\}$. The emotional engagement for the event is, then, $E_e = \mathbb{C}\{E_{V,e}, E_{A,e}, E_{P,e}\}$. The Sentiment ($S$) is defined as S^n_g and S_c at the group and crowd levels, respectively. For the event, $S_e = \mathbb{C}\left\{\bigoplus_n S^n_g, S_c\right\}$.

Body Language (L) is composed of Action and Gestures, $L_{ag} = \mathbb{C}\left\{\bigoplus_n L^n_{ag,g}, L_{ag,c}\right\}$, and Poses, $L_{po} = \mathbb{C}\left\{\bigoplus_n L^n_{po,g}, L_{po,c}\right\}$. At the event level it is denoted by $L_e = \mathbb{C}\{L_{ag}, L_{po}\}$. Scene Dynamics ($SD$) include Density, $SD_d = \mathbb{C}\left\{\bigoplus_n SD^n_{d,g}, SD_{d,c}\right\}$, and the density map is denoted by $SD_{dm} = \mathbb{C}\left\{\bigoplus_n SD^n_{dm,g}, SD_{dm,c}\right\}$. It also includes Count, denoted by SD_k, and Motion Patterns, denoted by $SD_m = \mathbb{C}\left\{\bigoplus_n SD^n_{m,g}, SD_{m,c}\right\}$, including the motion heatmap $SD_{hm} = \mathbb{C}\left\{\bigoplus_n SD^n_{hm,g}, SD_{hm,c}\right\}$. For an event, Scene Dynamics are mathematically represented by $SD_e = \mathbb{C}\{SD_d, SD_{dm}, SD_k, SD_m, SD_{hm}\}$.

Behaviour (B) is represented by Commitment (ct)—$B_{ct} = \mathbb{C}\left\{\bigoplus_n B^n_{ct,g}, B_{ct,c}\right\}$; Conversion (cv)—$B_{cv} = \mathbb{C}\left\{\bigoplus_n B^n_{cv,g}, B_{cv,c}\right\}$; Retention (rt)—$B_{rt} = \mathbb{C}\left\{\bigoplus_n B^n_{rt,g}, B_{rt,c}\right\}$; Feedback (fb)—$B_{fb,t} = \mathbb{C}\left\{\bigoplus_n B^n_{fb,t,g}, B_{fb,t,c}\right\}$; Social (s)—$B_s = \mathbb{C}\left\{\bigoplus_{i,j} B_{s,i,j}\right\}$, computed as the ratio between the number of interactions between groups i and the even duration j; and Odd Behaviours—$B_{ob} = \mathbb{C}\left\{\bigoplus_n B^n_{ob,g}, B_{ob,c}\right\}$. The event behaviour is, then, $B_e = \mathbb{C}\{B_{ct}, B_{cv}, B_{rt}, B_{fb,t}, B_s, B_{ob}\}$.

The *instantaneous engagement*, can now be computed, and represents the engagement at each instance/time t: $E_t = \mathbb{C}\{At_e, E_e, S_e, L_e, SD_e, B_e\}$. *Period engagement* is the engagement for a limited period of time: for $t \in [t_i, t_f]$ (with $t_i < t_f$), it is: $E_p = \bigoplus_{t\in[t_i,t_f]} E_t$. and *Overall event engagement*, which corresponds to the engagement over the entire duration of the event (interval I), is $E = \bigoplus_{t\in I} E_t$.

Following the above, as a proof of concept, the authors present the Binary Engagement Model (BEM) in [3], which is a microscopic, modular, and scalable architecture for detecting and classifying engagement in individuals and groups in real time during events. The model integrates only two dimensions—*Emotion* (characterised by valence and arousal) and *Attention* (based on gaze direction)—to estimate engagement at three temporal levels: instantaneous, periodic, and overall event engagement.

3.5. Ethical and Data Protection Implications

The Computer Vision-based audience analysis architecture proposed operates within a complex ethical and regulatory landscape, particularly in light of the European Union's Artificial Intelligence Act (EU AI Act). The system's reliance on the non-intrusive processing of facial images, gaze, and body language for real-time analytics can classify it as a high-risk AI system under the Act. This classification necessitates stringent requirements for risk

management, data governance, technical robustness, and human oversight. Furthermore, the use of affective computing to infer emotions and sentiment can raise significant ethical concerns regarding potential manipulation and the preservation of human autonomy, especially in contexts where such insights could be used to dynamically influence audience behaviour without their explicit consent. A principled approach to development must therefore prioritise transparency, ensuring that event organisers and attendees are aware of the system's operation and capabilities, and incorporate fundamental rights impact assessments to mitigate potential harms.

Regarding data protection, the use of biometric and behavioural data constitutes large-scale processing of special category data per the General Data Protection Regulation (GDPR). While event organisers may pursue legitimate interests, it must be weighed against the data subjects' rights and expectations, and the highly invasive nature of the processing requires strong data protection by design and by default. This involves the use of real-time anonymisation or aggregation of data at the edge as a means of mitigating privacy intrusions, strong data minimisation principles by only processing what is strictly necessary to calculate engagement metrics, and defining data retention policies. Ultimately, offering real, meaningful choice and control, perhaps through clear signage and the existence of opt-out zones in a venue, is paramount to building trust and ensuring that the use of such technologies is a reasonable exercise of the right to privacy.

4. Discussion

This comprehensive survey and the proposed architeture make significant contributions to the field of Computer Vision AI-driven audience engagement monitoring by synthesising a fragmented body of research and proposing a unified, multi-construct approach. The findings and their interpretation can be discussed through several key lenses.

The central hypothesis underpinning the proposed architeture is that engagement is a complex, multicomponent construct that cannot be accurately captured by a single metric. This is strongly supported by the reviewed literature; for instance, [4] explicitly defines engagement as comprising affective, cognitive, and behavioural components, a trichotomy that aligns perfectly with the architecture's integration of Emotion (affective), Attention (cognitive), and Body Language/Behaviour. The classroom studies (e.g., [16]) demonstrate that combining features (e.g., head pose and facial expressions) yields better performance than either feature alone. This empirically validates the architecture's design choice to fuse multiple data streams rather than rely on a single indicator.

The proposed architecture advances this concept by moving beyond a simple feature fusion to a structured integration of five distinct constructs, providing a more nuanced and holistic quantification of engagement. A critical finding of the survey is the stark disparity in research focus: the vast majority of automated engagement detection literature is centred on educational environments (classrooms, online learning). Studies like [9] are tailored for this context. The proposed architecture successfully generalises and adapts these concepts for the broader, more dynamic context of live events (e.g., conferences, concerts, grand activation, or exhibitions).

It does this by expanding the definition of behaviour from simple attentiveness to include metrics highly relevant to event promoters and organisers, such as Commitment, Conversion, Retention, Feedback, and Social Interaction. This shift in perspective is a key implication, suggesting that the value of engagement analytics extends beyond pedagogical feedback to business intelligence and resource optimisation for the events industry.

For practitioners, these metrics can translate directly into actionable insights. For instance, high commitment (e.g., repeated interactions with a booth or exhibit) may inform stage or space design, encouraging longer dwell times. Conversion rates can help promoters

evaluate the effectiveness of calls-to-action, such as whether a sessions structure motivates sign-ups for workshops or product trials. Retention patterns reveal whether attendees return after breaks or diversions, which could guide adjustments to session length, pacing, or scheduling to reduce drop-off. Feedback signals, such as applause intensity or gestural responses, can support real-time adaptation (e.g., extending a Q&A when enthusiasm is high, or altering delivery when engagement wanes). Finally, social interaction patterns, such as group discussions or peer-to-peer exchanges, can indicate the effectiveness of engagement strategies and inform interventions to foster collaboration or adjust group dynamics. In educational contexts, the same metrics may help instructors tailor lesson pacing, vary delivery methods, or identify which activities sustain student attention over time. These concrete applications make the architecture not only theoretically robust but also directly useful for optimising audience experience and event outcomes.

This review further highlights a significant challenge that currently limits progress in the field: the scarcity of high-quality, publicly available datasets for crowd engagement, annotated with emotional and behavioural labels. While numerous datasets exist for generic crowd behaviour (e.g., pedestrian datasets or UCF-Crime), they lack the specific annotations needed for engagement analysis (including valence, arousal, and sentiment labels tied to individuals/groups). This finding underscores a significant implication: progress in this field is contingent upon the creation of new, rich, and diverse benchmarks. The frameworks and models proposed, including the authors' Binary Engagement Model [3], can only be robustly validated and compared if such datasets become available. This calls for a collaborative effort within the research community to gather and annotate real-world event data.

It is important to note that group-level and crowd-level engagement pose distinct analytical challenges. Group-level analysis is primarily concerned with identity tracking, facial occlusions, and per-individual emotion variability, factors that require high-resolution, close-range data. In contrast, crowd-level analysis must contend with scale (hundreds or thousands of participants), heavy occlusion, and the absence of per-individual ground truth, necessitating the use of aggregate indicators such as density, collective motion, and synchronised behaviours.

It is also important to stress that the datasets from Table 1 can be strategically mapped to pre-train specific blocks of the proposed architecture, though they primarily serve as foundational models rather than direct engagement datasets. ShanghaiTech Campus is directly applicable for pre-training Scene Dynamics (Density/Count) blocks, while GENRI-4K and CelebA, with their smile attributes, are suitable for Sentiment State models. Most others, like UCSD, CUHK Avenue, and UMN, are excellent for foundational pre-training in anomaly detection and motion pattern analysis, which underpin the Odd Behaviours/Alerts and Motion Patterns blocks. UT-Interaction provides valuable data for Social Interaction recognition, and BOSS offers action labels relevant to Gestures and anomaly detection.

However, critical gaps remain for several core engagement constructs. No dataset in Table 1 provides annotations for Attention Direction (gaze-to-target), or macro-level behavioural metrics like Commitment, Conversion, and Retention. Therefore, while these datasets are invaluable for pre-training general Computer Vision backbones (e.g., for density estimation, action recognition, or anomaly detection), a final operational system will require subsequent fine-tuning or transfer learning on bespoke, engagement-annotated data to bridge the gap between general scene understanding and specific engagement inference. These distinctions reinforce the need for differentiated datasets and methods, as techniques effective for small group engagement cannot be directly scaled to large crowds without significant loss of resolution or interpretability. In addition, the architeture remains

sensitive to practical deployment constraints. Performance can degrade under poor or varying lighting conditions (e.g., outdoor or night events), where body cues are harder to capture. Similarly, camera placement poses challenges: off-axis angles, long distances, or obstructed views can reduce accuracy and exacerbate occlusion effects. Addressing these issues requires multimodal fusion (e.g., combining vision with audio) and adaptive camera setups, which we identify as future research directions.

Finally, by organising the architeture based on the five constructs (Attention Direction and Level, Emotion Level and Sentiment State, Body Language, Scene Dynamics, and Behaviours), with its respective blocks/dimensions, the paper provides a much-needed taxonomy and roadmap for future research. It allows researchers to pinpoint which specific component of the engagement puzzle they are addressing and how it fits into the larger picture. The paper successfully maps the landscape of Computer Vision AI-driven audience engagement, confirms the necessity of a multimodal approach, and provides a robust framework to guide future innovation. By addressing the challenges of data scarcity, computational complexity, and ethical implementation, this field holds immense potential to transform the experience and management of live events.

In conclusion, by formalising a multimodal architecture that fuses Emotion, Attention, Body Language, Scene Dynamics, and Behaviour metrics, this work establishes a scalable foundation for real-time engagement analytics, enabling more adaptive and data-driven strategies for live event management.

Future Research Directions

Based on the discussed findings and limitations, several promising future research directions emerge:

- Development of standardised benchmarks: The highest priority should be the creation and public release of comprehensive datasets filmed in real-world event settings, annotated with ground truth for different levels (and types) of engagement, valence, arousal, head poses, behavioural metrics (e.g., clapping, cheering, or leaving), etc. Group- and crowd-based datasets, as shown in Table 1, provide shared benchmarks that facilitate reproducibility and promote adoption within industry.
- Advanced occlusion handling and high-density analysis: Research must focus on novel Computer Vision techniques (e.g., 3D reconstruction, transformer-based models, neuromorphic vision as in Lorenzo et al. [46]) to make the architecture robust in highly occluded, dense crowd scenarios, which are common in large events. Techniques for dense, occluded crowds are directly relevant for large festivals or sporting events, where visibility is poor but safety monitoring is critical.
- Adaptive camera setups and smart placement: Work should explore adaptive camera configurations, such as PTZ systems or multi-camera arrays, that can dynamically adjust placement, angle, and zoom to improve coverage in complex venues. Combined with 3D cameras (e.g., stereo rigs, structured light, or LiDAR), these setups can provide distance-aware information, helping to disambiguate overlapping individuals and estimate the distance between the audience and the target (stage, speaker, or point of interest). This additional spatial context not only mitigates occlusion but also enables richer engagement metrics by linking gaze and body orientation to the audience's physical relation to the event focus.
- Multimodal fusion architectures: Future work should move beyond simple feature concatenation to explore sophisticated late-fusion and attention-based fusion models that can dynamically weight the importance of each construct. In conferences, fusing gaze and applause detection could inform real-time adjustments to presentations; in concerts, body language may outweigh facial cues due to lighting or distance.

- Context-aware and personalised models: Future systems should adapt to the contexts of events and learn personalised baselines for audience segments to improve interpretation. A jazz concert audience expresses engagement differently than a tech keynote audience; tailoring models to these contexts supports both cultural event organisers and corporate planners.
- Ethical AI and privacy-preserving techniques: As these systems deploy, research must integrate privacy-by-design principles. This includes exploring federated learning, on-device processing, and techniques that use low-resolution or abstracted feature data to protect individual identities while still extracting useful crowd-level insights.
- Integration with subjective measures: To validate automated metrics, future work should develop methods to seamlessly integrate sparse subjective feedback with continuous AI-derived data, creating a hybrid validation model. Hybrid models could combine real-time analytics with feedback buttons at conferences or mobile surveys at festivals, enhancing actionable insights for organisers.
- Scaling to denser crowds: Future work should investigate benchmarking computational load across hardware tiers and exploring neuromorphic vision sensors and transformer-based backbones as promising directions for scaling to denser crowds.

Although the above directions are framed as technical challenges, they directly align with practical needs in real-time scenarios, such as improving audience experience at cultural and corporate events.

Author Contributions: Conceptualisation, M.L. and J.M.F.R.; methodology, M.L., P.J.S.C. and J.M.F.R.; software, M.L.; validation, M.L., P.J.S.C. and J.M.F.R.; formal analysis, P.J.S.C.; investigation, M.L., P.J.S.C. and J.M.F.R.; resources, J.M.F.R.; writing—original draft preparation, M.L., P.J.S.C. and J.M.F.R.; writing—review and editing, M.L., P.J.S.C. and J.M.F.R.; supervision, J.M.F.R.; project administration, J.M.F.R.; funding acquisition, J.M.F.R. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the UID/04516/NOVA Laboratory for Computer Science and Informatics (NOVA LINCS) with the financial support of FCT.IP, and by the project AI.EVENT: Monitor Live Audience with AI (ALGARVE-FEDER-01180500, Ref. 17325), co-financed by ALGARVE 2030, Portugal 2030, and the European Union.

Data Availability Statement: No new data were created or analysed in this study. Data sharing is not applicable to this article.

Acknowledgments: Foundation for Science and Technology, ALGARVE 2030, Portugal 2030.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
AU	Action Unit
AUC	Area Under the Curve
BEM	Binary Engagement Model
BP	Back Propagation
CI	Concentration Index
CNN	Convolutional Neural Network
DCNN	Deep Convolutional Neural Network
DNN	Deep Neural Network
EI	Engagement Index
FER	Facial Expression Recognition

GIS	Geographic Information System
GRU	Gated Recurrent Unit
HAD	Histogram of Angular Deviations
HPE	Head Pose Estimation
HRI	Human–Robot Interaction
IE	Instantaneous Engagement
MAE	Mean Absolute Error
PE	Period Engagement
PoI	Point of Interest
R&D	Research and Development
RMSE	Root Mean Square Error
VA	Valence–Arousal
VLM	Vision Language Models
YOLO	You Only Look Once

References

1. Sánchez, F.L.; Hupont, I.; Tabik, S.; Herrera, F. Revisiting crowd behaviour analysis through deep learning: Taxonomy, anomaly detection, crowd emotions, datasets, opportunities and prospects. *Inf. Fusion* **2020**, *64*, 318–335. [CrossRef] [PubMed]
2. Varghese, E.B.; Thampi, S.M. Towards the cognitive and psychological perspectives of crowd behaviour: A vision-based analysis. *Connect. Sci.* **2021**, *33*, 380–405. [CrossRef]
3. Lemos, M.; Cardoso, P.J.S.; Rodrigues, J.M.F. Microscopic Binary Engagement Model. In Proceedings of the Computational Science—ICCS 2025, Singapore, 7–9 July 2025; Springer Nature: Cham, Switzerland, 2025; pp. 119–134. .. [CrossRef]
4. Booth, B.M.; Bosch, N.; DMello, S.K. Engagement Detection and Its Applications in Learning: A Tutorial and Selective Review. *Proc. IEEE* **2023**, *111*, 1398–1422. [CrossRef]
5. Lasri, I.; Riadsolh, A.; Elbelkacemi, M. Facial emotion recognition of deaf and hard-of-hearing students for engagement detection using deep learning. *Educ. Inf. Technol.* **2023**, *28*, 4069–4092. [CrossRef]
6. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks For Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556. [CrossRef]
7. Xu, Q.; Wei, Y.; Gao, J.; Yao, H.; Liu, Q. ICAPD Framework and simAM-YOLOv8n for Student Cognitive Engagement Detection in Classroom. *IEEE Access* **2023**, *11*, 136063–136076. [CrossRef]
8. Sumer, O.; Goldberg, P.; DMello, S.; Gerjets, P.; Trautwein, U.; Kasneci, E. Multimodal Engagement Analysis From Facial Videos in the Classroom. *IEEE Trans. Affect. Comput.* **2023**, *14*, 1012–1027. [CrossRef]
9. Zhao, Z.; Li, Y.; Yang, J.; Ma, Y. A lightweight facial expression recognition model for automated engagement detection. *Signal Image Video Process.* **2024**, *18*, 3553–3563. [CrossRef]
10. Vrochidis, A.; Dimitriou, N.; Krinidis, S.; Panagiotidis, S.; Parcharidis, S.; Tzovaras, D. A Deep Learning Framework for Monitoring Audience Engagement in Online Video Events. *Int. J. Comput. Intell. Syst.* **2024**, *17*, 124. [CrossRef]
11. Ruiz, N.; Chong, E.; Rehg, J.M. Fine-grained head pose estimation without keypoints. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2074–2083. [CrossRef]
12. Shao, Z.; Liu, Z.; Cai, J.; Ma, L. JAA-Net: Joint facial action unit detection and face alignment via adaptive attention. *Int. J. Comput. Vis.* **2021**, *129*, 321–340. [CrossRef]
13. Deng, J.; Guo, J.; Zhou, Y.; Yu, J.; Kotsia, I.; Zafeiriou, S. Retinaface: Single-stage dense face localisation in the wild. *arXiv* **2019**, arXiv:1905.00641. [CrossRef]
14. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708. [CrossRef]
15. habib Albohamood, A.; Shaker Alqattan, M.; Padua Vizcarra, C. Real-time Student Engagement Monitoring in Classroom Environments using Machine Learning and Computer Vision. In Proceedings of the 2025 4th International Conference on Computing and Information Technology (ICCIT), Tabuk, Saudi Arabia, 13–14 April 2025; pp. 420–424. [CrossRef]
16. Qarbal, I.; Sael, N.; Ouahabi, S. Student Engagement Detection Based on Head Pose Estimation and Facial Expressions Using Transfer Learning. In Proceedings of the International Conference on Smart City Applications, Tangier, Morocco, 1–3 October 2024; Springer: Berlin/Heidelberg, Germany, 2025; pp. 246–255. [CrossRef]
17. Teotia, J.; Zhang, X.; Mao, R.; Cambria, E. Evaluating Vision Language Models in Detecting Learning Engagement. In Proceedings of the 2024 IEEE International Conference on Data Mining Workshops (ICDMW), Abu Dhabi, United Arab Emirates, 9–12 December 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 496–502. [CrossRef]
18. Sorrentino, A.; Fiorini, L.; Cavallo, F. From the definition to the automatic assessment of engagement in human–robot interaction: A systematic review. *Int. J. Soc. Robot.* **2024**, *16*, 1641–1663. [CrossRef]

19. Ravandi, B.S.; Khan, I.; Markelius, A.; Bergström, M.; Gander, P.; Erzin, E.; Lowe, R. Exploring task and social engagement in companion social robots: A comparative analysis of feedback types. *Adv. Robot.* **2025**, *39*, 884–899. [CrossRef]
20. Qarbal, I.; Sael, N.; Ouahabi, S. Students Engagement Detection Based on Computer Vision: A Systematic Literature Review. *IEEE Access* **2025**, *13*, 140519–140545. [CrossRef]
21. Bei, Y.; Guo, S.; Gao, K.; Feng, Z. Behavior capture guided engagement recognition. *Pattern Recognit.* **2025**, *164*, 111534. [CrossRef]
22. Wang, J.; Yuan, S.; Lu, T.; Zhao, H.; Zhao, Y. Video-based real-time monitoring of engagement in E-learning using MediaPipe through multi-feature analysis. *Expert Syst. Appl.* **2025**, *288*, 128239. [CrossRef]
23. Lu, W.; Yang, Y.; Song, R.; Chen, Y.; Wang, T.; Bian, C. A Video Dataset for Classroom Group Engagement Recognition. *Sci. Data* **2025**, *12*, 644. [CrossRef]
24. HAPPEI Dataset. Available online: https://users.cecs.anu.edu.au/~few_group/Group.htm (accessed on 28 August 2025).
25. MED: Multimodal Event Dataset. Available online: https://github.com/hosseinm/med?tab=readme-ov-file (accessed on 28 August 2025).
26. Vaz, P.J.; Rodrigues, J.M.F.; Cardoso, P.J.S. Affective Computing Databases: In-Depth Analysis of Systematic Reviews and Surveys. *IEEE Trans. Affect. Comput.* **2025**, *16*, 537–554. [CrossRef]
27. UCSD Anomaly Detection Dataset. Available online: http://www.svcl.ucsd.edu/projects/anomaly/dataset.htm (accessed on 28 August 2025).
28. CUHK Avenue Dataset. Available online: https://www.cse.cuhk.edu.hk/leojia/projects/detectabnormal/dataset.html (accessed on 28 August 2025).
29. UMN Crowd Dataset. Available online: https://mha.cs.umn.edu/proj_events.shtml (accessed on 28 August 2025).
30. ShanghaiTech Campus Dataset. Available online: https://svip-lab.github.io/dataset/campus_dataset.html (accessed on 28 August 2025).
31. UT-Interaction Dataset. Available online: https://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html (accessed on 28 August 2025).
32. UCF-Crime Dataset. Available online: https://www.crcv.ucf.edu/research/real-world-anomaly-detection-in-surveillance-videos/ (accessed on 28 August 2025).
33. Gao, Y.; Liu, H.; Wu, P.; Wang, C. A new descriptor of gradients self-similarity for smile detection in unconstrained scenarios. *Neurocomputing* **2016**, *174*, 1077–1086. [CrossRef]
34. Lingenfelter, B.; Davis, S.R.; Hand, E.M. A quantitative analysis of labeling issues in the celeba dataset. In Proceedings of the International Symposium on Visual Computing, San Diego, CA, USA, 3–5 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 129–141. [CrossRef]
35. Hwooi, S.K.W.; Othmani, A.; Sabri, A.Q.M. Deep learning-based approach for continuous affect prediction from facial expression images in valence-arousal space. *IEEE Access* **2022**, *10*, 96053–96065. [CrossRef]
36. Savchenko, A.V. Frame-level prediction of facial expressions, valence, arousal and action units for mobile devices. *arXiv* **2022**, arXiv:2203.13436. [CrossRef]
37. Cao, Q.; Shen, L.; Xie, W.; Parkhi, O.M.; Zisserman, A. Vggface2: A dataset for recognising faces across pose and age. In Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 67–74. [CrossRef]
38. Nguyen, H.H.; Huynh, V.T.; Kim, S.H. An ensemble approach for facial expression analysis in video. *arXiv* **2022**, arXiv:2203.12891. [CrossRef]
39. Radosavovic, I.; Kosaraju, R.P.; Girshick, R.; He, K.; Dollár, P. Designing network design spaces. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10428–10436. [CrossRef]
40. Cho, K. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv* **2014**, arXiv:1409.1259. [CrossRef]
41. Devlin, J. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805. [CrossRef]
42. Gupta, S.; Kumar, P.; Tekchandani, R.K. Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models. *Multimed. Tools Appl.* **2023**, *82*, 11365–11394. [CrossRef]
43. Hai, L.; Guo, H. Face detection with improved face r-CNN training method. In Proceedings of the 3rd International Conference on Control and Computer Vision, Macau, China, 23–25 August 2020; pp. 22–25. [CrossRef]
44. Bruin, J.; Stuldreher, I.V.; Perone, P.; Hogenelst, K.; Naber, M.; Kamphuis, W.; Brouwer, A.M. Detection of arousal and valence from facial expressions and physiological responses evoked by different types of stressors. *Front. Neuroergonomics* **2024**, *5*, 1338243. [CrossRef]
45. Ding, C.; Peng, H. Minimum redundancy feature selection from microarray gene expression data. *J. Bioinform. Comput. Biol.* **2005**, *3*, 185–205. [CrossRef] [PubMed]
46. Berlincioni, L.; Cultrera, L.; Becattini, F.; Bimbo, A.D. Neuromorphic valence and arousal estimation. *J. Ambient Intell. Humaniz. Comput.* **2024**, 1–11. [CrossRef]

47. Kossaifi, J.; Tzimiropoulos, G.; Todorovic, S.; Pantic, M. AFEW-VA database for valence and arousal estimation in-the-wild. *Image Vis. Comput.* **2017**, *65*, 23–36. [CrossRef]
48. Hu, Y.; Liu, S.C.; Delbruck, T. v2e: From video frames to realistic DVS events. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 1312–1321. [CrossRef]
49. Manojkumar, K.; Helen, L.S. Monitoring the crowd emotion using valence and arousal of crowd based on prominent features of crowd. *Signal Image Video Process.* **2025**, *19*, 519. [CrossRef]
50. Hempel, T.; Abdelrahman, A.A.; Al-Hamadi, A. 6d rotation representation for unconstrained head pose estimation. In Proceedings of the 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 16–19 October 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 2496–2500. [CrossRef]
51. Hempel, T.; Abdelrahman, A.A.; Al-Hamadi, A. Toward Robust and Unconstrained Full Range of Rotation Head Pose Estimation. *IEEE Trans. Image Process.* **2024**, *33*, 2377–2387. [CrossRef]
52. Reich, A.; Wuensche, H.J. Monocular 3d multi-object tracking with an ekf approach for long-term stable tracks. In Proceedings of the 2021 IEEE 24th International Conference on Information Fusion (FUSION), Sun City, South Africa, 1–4 November 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–7. [CrossRef]
53. Zhou, X.; Koltun, V.; Krähenbühl, P. Tracking objects as points. In Proceedings of the European Conference on Computer Vision, Online, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 474–490. [CrossRef]
54. Hossain, M.R.; Rahman, M.M.; Karim, M.R.; Al Amin, M.J.; Bepery, C. Determination of 3D Coordinates of Objects from Image with Deep Learning Model. In Proceedings of the 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 26–29 January 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 25–30. [CrossRef]
55. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755. [CrossRef]
56. Yang, L.; Kang, B.; Huang, Z.; Zhao, Z.; Xu, X.; Feng, J.; Zhao, H. Depth anything v2. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 21875–21911. [CrossRef]
57. Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. DINOv2: Learning Robust Visual Features without Supervision. *arXiv* **2024**, arXiv:2304.07193v2. [CrossRef]
58. Sundararaman, R.; De Almeida Braga, C.; Marchand, E.; Pettre, J. Tracking pedestrian heads in dense crowd. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 3865–3875. [CrossRef]
59. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
60. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125. [CrossRef]
61. Zhu, C.; Tao, R.; Luu, K.; Savvides, M. Seeing small faces from robust anchor's perspective. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5127–5136. [CrossRef]
62. Tang, X.; Du, D.K.; He, Z.; Liu, J. Pyramidbox: A context-assisted single shot face detector. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 797–813. [CrossRef]
63. Deb, T.; Rahmun, M.; Bijoy, S.A.; Raha, M.H.; Khan, M.A. UUCT-HyMP: Towards tracking dispersed crowd groups from UAVs. In Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China, 18–22 July 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–8. [CrossRef]
64. Shah, S.; Dey, D.; Lovett, C.; Kapoor, A. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and Service Robotics: Results of the 11th International Conference*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 621–635. [CrossRef]
65. Bhat, G.; Danelljan, M.; Gool, L.V.; Timofte, R. Learning discriminative model prediction for tracking. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6182–6191. [CrossRef]
66. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv* **2016**, arXiv:1609.02907. [CrossRef]
67. Yeh, K.H.; Hsu, I.C.; Chou, Y.Z.; Chen, G.Y.; Tsai, Y.S. An aerial crowd-flow analyzing system for drone under YOLOv5 and StrongSort. In Proceedings of the 2022 International Automatic Control Conference (CACS), Kaohsiung, Taiwan, 3–6 November 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–6. [CrossRef]
68. Jocher, G.; Stoken, A.; Borovec, J.; Changyu, L.; Hogan, A.; Diaconu, L.; Ingham, F.; Poznanski, J.; Fang, J.; Yu, L.; et al. ultralytics/yolov5: V3. 1-bug fixes and performance improvements. *Zenodo* **2020**. [CrossRef]
69. Du, Y.; Zhao, Z.; Song, Y.; Zhao, Y.; Su, F.; Gong, T.; Meng, H. Strongsort: Make deepsort great again. *IEEE Trans. Multimed.* **2023**, *25*, 8725–8737. [CrossRef]

70. Zhou, K.; Yang, Y.; Cavallaro, A.; Xiang, T. Omni-scale feature learning for person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republci of Korea, 27 October–2 November 2019; pp. 3702–3712. [CrossRef]
71. Real-Time Multi-Camera Multi-Object Tracker Using YOLOv5 and StrongSORT with OSNet. Available online: https://github.com/mikel-brostrom/Yolov5_StrongSORT_OSNet (accessed on 28 August 2025).
72. Li, H.; Liu, L.; Yang, K.; Liu, S.; Gao, J.; Zhao, B.; Zhang, R.; Hou, J. Video crowd localization with multifocus gaussian neighborhood attention and a large-scale benchmark. *IEEE Trans. Image Process.* **2022**, *31*, 6032–6047. [CrossRef]
73. Ekanayake, E.; Lei, Y.; Li, C. Crowd density level estimation and anomaly detection using multicolumn multistage bilinear convolution attention network (MCMS-BCNN-Attention). *Appl. Sci.* **2022**, *13*, 248. [CrossRef]
74. Tan, M.; Le, Q. Efficientnetv2: Smaller models and faster training. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; PMLR: New York, NY, USA, 2021; pp. 10096–10106. [CrossRef]
75. Zhang, Y.; Zhou, D.; Chen, S.; Gao, S.; Ma, Y. Single-image crowd counting via multi-column convolutional neural network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26–30 June 2016; pp. 589–597. [CrossRef]
76. Ferryman, J.; Shahrokni, A. Pets2009: Dataset and challenge. In Proceedings of the 2009 Twelfth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, Snowbird, UT, USA, 7–9 December 2009; IEEE: Piscataway, NJ, USA, 2009; pp. 1–6. [CrossRef]
77. Mehran, R.; Oyama, A.; Shah, M. Abnormal crowd behavior detection using social force model. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; IEEE: Piscataway, NJ, USA, 2009; pp. 935–942. [CrossRef]
78. Liu, Y.; Cao, G.; Ge, Z.; Hu, Y. Crowd counting method via a dynamic-refined density map network. *Neurocomputing* **2022**, *497*, 191–203. [CrossRef]
79. Lin, H.; Ma, Z.; Ji, R.; Wang, Y.; Hong, X. Boosting crowd counting via multifaceted attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 19628–19637. [CrossRef]
80. Wang, F.; Sang, J.; Wu, Z.; Liu, Q.; Sang, N. Hybrid attention network based on progressive embedding scale-context for crowd counting. *Inf. Sci.* **2022**, *591*, 306–318. [CrossRef]
81. Qi, Z.; Zhou, M.; Zhu, G.; Xue, Y. Multiple pedestrian tracking in dense crowds combined with head tracking. *Appl. Sci.* **2022**, *13*, 440. [CrossRef]
82. Pan, Z. Multi-Scale Occluded Pedestrian Detection Based on Deep Learning. In Proceedings of the 2023 International Conference on Evolutionary Algorithms and Soft Computing Techniques (EASCT), Bengaluru, India, 20–21 October 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–6. [CrossRef]
83. Pai, A.K.; Chandrahasan, P.; Raghavendra, U.; Karunakar, A.K. Motion pattern-based crowd scene classification using histogram of angular deviations of trajectories. *Vis. Comput.* **2023**, *39*, 557–567. [CrossRef]
84. Zhou, B.; Tang, X.; Wang, X. Measuring crowd collectiveness. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 3049–3056. [CrossRef]
85. Zhang, X.; Sun, Y.; Li, Q.; Li, X.; Shi, X. Crowd density estimation and mapping method based on surveillance video and GIS. *ISPRS Int. J. Geo-Inf.* **2023**, *12*, 56. [CrossRef]
86. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818. [CrossRef]
87. Qiu, W.; Wen, G.; Liu, H. A back-propagation neural network model based on genetic algorithm for prediction of build-up rate in drilling process. *Arab. J. Sci. Eng.* **2022**, *47*, 11089–11099. [CrossRef]
88. Zhang, Y.; Chen, H.; Lai, Z.; Zhang, Z.; Yuan, D. Handling heavy occlusion in dense crowd tracking by focusing on the heads. In Proceedings of the Australasian Joint Conference on Artificial Intelligence, Brisbane, Australia, 18 November–1 December 2023; Springer: Berlin/Heidelberg, Germany, 2023; pp. 79–90. [CrossRef]
89. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430. [CrossRef]
90. Shao, S.; Zhao, Z.; Li, B.; Xiao, T.; Yu, G.; Zhang, X.; Sun, J. CrowdHuman: A Benchmark for Detecting Human in a Crowd. *arXiv* **2018**, arXiv:1805.00123. [CrossRef]
91. Dendorfer, P.; Rezatofighi, H.; Milan, A.; Shi, J.; Cremers, D.; Reid, I.; Roth, S.; Schindler, K.; Leal-Taixé, L. Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv* **2020**, arXiv:2003.09003. [CrossRef]
92. Mei, L.; Yu, M.; Jia, L.; Fu, M. Crowd Density Estimation via Global Crowd Collectiveness Metric. *Drones* **2024**, *8*, 616. [CrossRef]
93. Ranasinghe, Y.; Nair, N.G.; Bandara, W.G.C.; Patel, V.M. CrowdDiff: Multi-hypothesis crowd density estimation using diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 12809–12819. [CrossRef]
94. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6840–6851. [CrossRef]

95. Nichol, A.Q.; Dhariwal, P. Improved denoising diffusion probabilistic models. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; PMLR: New York, NY, USA, 2021; pp. 8162–8171. [CrossRef]
96. Akpulat, M.; Ekinci, M. Anomaly detection in crowd scenes via cross trajectories. *Appl. Intell.* **2025**, *55*, 525. [CrossRef]
97. Badauraudine, M.F.M.; Noor, M.N.M.M.; Othman, M.S.; Nasir, H.B.M. Detection and Tracking of People in a Dense Crowd through Deep Learning Approach-A Systematic Literature Review. *Inf. Res. Commun.* **2025**, *1*, 65–73. [CrossRef]
98. Martins, P.V.C.; Cardoso, P.J.; Rodrigues, J.M. Affective Computing Emotional Body Gesture Recognition: Evolution and the Cream of the Crop. *IEEE Access* **2025**, *13*, 192871–192890. [CrossRef]
99. Rodrigues, J.M.F.; Cardoso, P.J.S.; Lemos, M.; Cherniavska, O.; Bica, P. Engagement Monitorization in Crowded Environments: A Conceptual Framework. In Proceedings of the 11th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-Exclusion, New York, NY, USA, 13–15 November 2025; DSAI '24; p. 815. [CrossRef]