ABCDE: Appearance-Based Confidence Detection by Evaluating Gaze Behavior Using Deep Learning

Ankur Bhatt [1], Ko Watanabe [2], Jayasankar Santhosh [3],
Andreas Dengel[4], Shoya Ishimaru [5]
[134] RPTU Kaiserslautern-Landau
[234] German Research Center of Artificial Intelligence (DFKI GmbH)
[5] Osaka Metropolitan University

Abstract

Self-confidence is a pivotal trait that profoundly impacts performance across various life domains. It fosters positive outcomes by facilitating quick decision-making and timely actions. In the context of video-based learning, accurate detection of self-confidence is critical as it enables the provision of personalized feedback, thereby enhancing learners' experiences and improving their confidence levels. This study addresses the challenge of self-confidence detection by evaluating and comparing traditional machine-learning methods with an advanced deep-learning approach using eye-tracking data collected through two distinct modalities: an eye-tracker and an appearance-based model. Our experimental setup involved fourteen participants, each of whom viewed eight distinct videos and provided corresponding responses. To analyze and assess the collected data, we implemented and compared five different algorithms: Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), Extreme Gradient Boosting (XGBoost), and a deep-learning based 1D Convolutional Neural Network (1D CNN) and Transformer models. The 1D CNN model achieved the highest macro F1-scores using leave-one-participant-out cross-validation (LOPOCV), with performances of 0.662 on eye-tracking data and 0.635 on appearance-based data. In contrast, under leave-one-question-out cross-validation (LOQOCV), Logistic Regression demonstrated superior performance for eye-tracking data (F1-score: 0.560), while Transformer-based models yielded the highest F1-score (0.616) for appearance-based data. These findings underscore the

[1]ankur.bhatt@dfki.de
[2]ko.watanabe@dfki.de
[3]jayasankar.santhosh@dfki.de
[4]andreas.dengel@dfki.de
[5]ishimaru@omu.ac.jp

effectiveness of deep learning in capturing complex gaze behavior patterns, thereby providing a robust framework for estimating self-confidence in video-based learning environments.

1 Introduction

In the era of digital education, quantified learning has emerged as a promising solution to address the challenges of online learning environments. By monitoring learning behaviors, quantified learning provides actionable feedback to students and educators. Smart sensors integrated into devices such as computers, smartphones [1], chairs [2], and eyeglasses [3] capture insights into students' physical and cognitive states during learning activities. Physical states are reflected in non-verbal cues like utterance rates [4, 5], nodding [6, 7], and smiling [8–10], while cognitive states include processes such as engagement [11, 12], boredom [13, 14], and self-confidence [15–17].

Self-confidence is a critical factor in academic success, with studies showing its strong connection to learning outcomes [18, 19]. Enhancing self-confidence has been shown to improve academic performance [20] significantly. Quantified learning bridges traditional and online education gaps by enabling personalized interventions to boost confidence through tailored feedback. However, the relationship between non-verbal cues, such as eye movements, and self-confidence levels in digital education remains underexplored, particularly as telepresence technologies continue to grow.

Eye gaze provides critical insights into human attention and cognitive states. By analyzing gaze patterns, researchers can infer intentions [21] and better understand social interactions [22]. Gaze behavior—comprising fixation patterns, saccadic movements, and gaze aversion—offers valuable data about learners' attention, focus, and self-confidence. Eye-tracking has thus emerged as a non-invasive tool for analyzing these states. Two primary methods for eye-tracking are hardware-based systems, which offer high precision but rely on costly equipment, and software-based solutions, which use webcams to provide a scalable and cost-effective alternative for educational applications.

Over the years, gaze estimation methods have been categorized into three main approaches [23]:

1. **Appearance-Based Methods:** Rely on visual cues, such as the appearance of eyes, to estimate gaze direction. They are hardware-independent and use webcams, which makes them scalable.
2. **2D Eye Feature Regression Methods:** Use geometric features, such as pupil center location, to predict the point of gaze (PoG) without requiring calibration. However, they often depend on specialized equipment.
3. **3D Eye Model Recovery Methods:** Involve creating geometric models of the eye to estimate gaze direction, typically requiring person-specific calibration and specialized tools like infrared cameras, limiting scalability.

Our findings demonstrate the advantages of deep learning over traditional feature extraction-based, with the 1D CNN outperforming other models in classifying self-confidence. This research bridges a critical gap in practical computing and learning analytics by systematically comparing hardware-based and software-based methods. Through this analysis, we aim to design more sophisticated educational systems that provide tailored feedback and enhance learning outcomes. The contributions of the research are listed as follows:

- C1 We propose a novel framework for estimating self-confidence in video-based learning using gaze data from both hardware-based (eye-tracker) and software-based (appearance-based) modalities.
- C2 We systematically compare hand-crafted feature-based with deep learning methods.
- C3 We highlight the feasibility of webcam-based gaze estimation for scalable and cost-effective confidence assessment, bridging the gap between quantified learning and adaptive feedback in education.

2 Related Work

In this section, we focus on introducing existing related work on self-confidence estimation, eye-tracking methods, and appearance-based gaze estimation.

2.1 Self-Confidence Estimation

The critical role of confidence in influencing neurocognitive states has been extensively studied across domains, including standardized learning environments [24], cognitive assessments [25], and skill acquisition in fields like culinary arts [26]. Forbes-Riley and Litman [27] demonstrated that integrating confidence into tutoring systems enhances learning pace and user satisfaction. Similarly, individuals who advocate for themselves often achieve greater success, as positive reinforcement and acknowledgment of their performance bolster confidence.

Confidence is particularly crucial in addressing misconceptions during learning. Sun and Yeh [28] emphasized that fostering self-confidence enables students to identify and correct misunderstandings, promoting more accurate self-assessment. Additionally, Roderer and Roebers [29] noted age-related differences in confidence, with younger individuals typically exhibiting higher levels than older counterparts.

In neuroscience, the link between physiological measures, such as electroencephalography (EEG), and self-efficacy—the belief in one's ability to perform tasks—is well-documented [28, 30]. However, traditional EEG-based methods are often intrusive, causing discomfort and reduced engagement. In contrast, eye-tracking offers a non-intrusive, naturalistic alternative for monitoring cognitive and emotional states. For example, Maruichi et al. [31] proposed estimating self-confidence through stroke-level handwriting behavior, providing personalized feedback to help learners address knowledge gaps and improve outcomes efficiently.

Building on this, Bruhin et al. [32] conducted a laboratory experiment examining self-confidence's influence on teamwork. Their study manipulated participants' confidence through general knowledge quizzes of varying difficulty and revealed that overconfidence increased individual effort, reduced free-riding, and boosted team performance and revenue. These findings highlight the synergistic effects of self-confidence in collaborative settings, where balancing ability and effort drives collective success.

2.2 Eye-tracking in Action

Eye behaviors and body language are key indicators of self-confidence. Individuals with low self-confidence often exhibit prolonged revisiting and re-evaluating of questions or choices, reflecting indecisiveness and uncertainty [33]. Eye-tracking has proven to be a valuable tool for understanding these behaviors and enhancing the learning experience. For instance, Okoso et al. [34] demonstrated how eye-tracking can identify specific sections of a text that learners struggle to comprehend, enabling targeted attention to those areas.

Eye-tracking has shown broader applications in educational contexts. Lee et al. [35] found a positive correlation between sustained eye contact with virtual tutors and improved learning outcomes, highlighting the role of gaze in fostering engagement. Similarly, Augereau et al. [36] used eye movement patterns to estimate English language proficiency during testing, achieving high accuracy and minimal errors. In problem-solving tasks, Yamada et al. [17] pioneered the use of eye-tracking to automatically assess self-confidence levels through eye movement analysis. Yamada et al. proposed a confidence-aware system that utilizes eye-tracking data to analyze reading and answering behaviors, extract gaze-based features, and predict confidence levels [37].

Expanding on these concepts, Ishimaru et al. [16] introduced a confidence-aware learning assistant that uses eye-tracking to detect students' self-confidence while answering multiple-choice questions. The findings demonstrated that gaze behavior is closely linked to self-confidence and can be effectively used for real-time confidence estimation in educational settings. By building on these outcomes, we have further validated the feasibility of gaze as a metric for confidence detection in our study. The system adapts the review process based on the estimated confidence levels, providing personalized feedback and targeted interventions to enhance learning.

These studies collectively highlight the potential of eye-tracking to provide insights into learners' cognitive processes, enabling tailored feedback and reinforcement strategies to support individual learning needs.

2.3 Appearance-based Gaze Estimation

Numerous techniques have been developed to enhance accuracy in gaze estimation, utilizing both traditional machine learning and deep-learning approaches. For example, Lu et al. proposed dividing eye images into 15 subregions and using the summed pixel intensities as features [38]. However, traditional appearance-based methods often require manual calibration, which is time-consuming and complicates the collection of user-

specific training samples under controlled settings. To address this, Williams et al. [39] introduced a semi-supervised Gaussian process regression approach, reducing the necessary training samples. Although such techniques improve efficiency, they are limited to controlled environments (e.g., fixed headgear or specific users) and struggle in more challenging scenarios. Deep-learning-based methods have emerged as a robust alternative. They automatically extract features from eye images and overcome the limitations of traditional approaches. These methods are trained using various paradigms, including supervised, semi-supervised, self-supervised, and unsupervised learning.

Supervised learning has driven significant advancements in gaze estimation. Zhang et al. [40] introduced one of the first methods using a Convolutional Neural Network (CNN) to compute gaze directions, surpassing many traditional appearance-based approaches. Krafka et al. [41] developed a gaze estimation architecture leveraging face images. The iTracker model proposed by Krafka et al. [42] combines left and right eye images, face images, and face grid information to infer gaze location. This model uses inputs such as the face image, its location in the frame (face grid), and eye images, estimating head pose relative to the camera and eye pose relative to the head. Its architecture is based on AlexNet [43]. Similarly, Bao et al. [44] proposed AFF-Net, which extracts features from both the face and rectangular regions using convolutional and fully connected layers. This architecture integrates left and horizontally flipped right eye images through a novel stacking mechanism and Squeeze-and-Excitation (SE) layers, with Adaptive Group Normalization (AdaGN) recalibrating eye features based on facial appearance characteristics, while Bhatt et al. [45] employed two methods: one with a single feature extractor and another with four, processing different inputs (e.g., original frame, face, left eye, and right eye) using backbones like VGG16, ResNet50, and EfficientNetB7.

Video-based gaze estimation has also gained attention, leveraging the additional information in video sequences compared to static images [46]. In these methods, static features are extracted from individual frames using a CNN, and temporal dependencies are captured using a Recurrent Neural Network (RNN), enhancing gaze estimation accuracy.

Beyond supervised learning, other training paradigms have been explored. In the semi-supervised domain, Wang et al. [47] introduced an adversarial learning approach to improve model performance for target subjects or datasets. For self-supervised methods, Cheng et al. [48] proposed an asymmetry regression network comprising a regression component to estimate gaze directions and an evaluation component to assess prediction reliability for both eyes. In unsupervised learning, Yu and Odobez [49] employed a CNN to extract 2D features from eye images. A gaze redirection network generates the corresponding image for the other eye by analyzing the feature differences between paired images, enabling gaze estimation without labeled data.

These diverse methods showcase the evolution of gaze estimation techniques from traditional and supervised learning to more innovative approaches like self-supervised and unsupervised learning, pushing the boundaries of accuracy and adaptability in real-world applications.

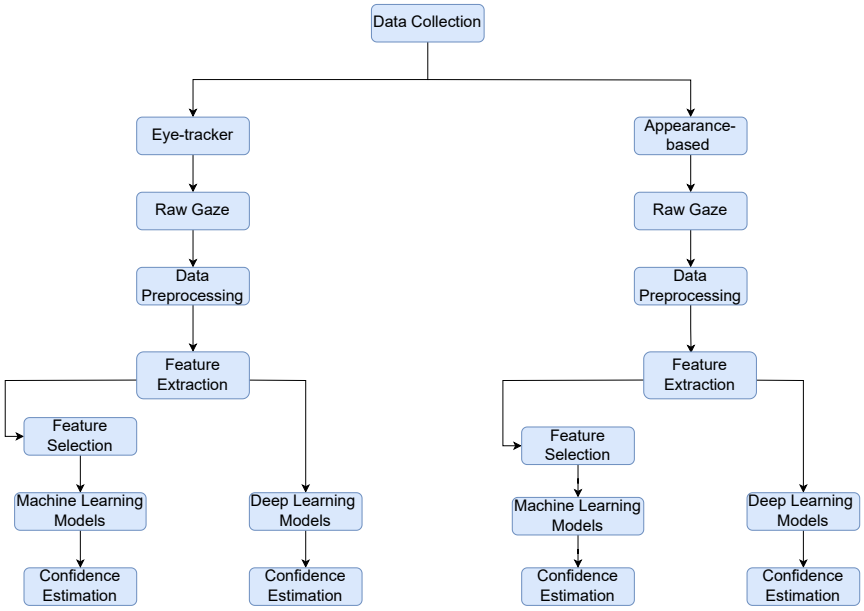


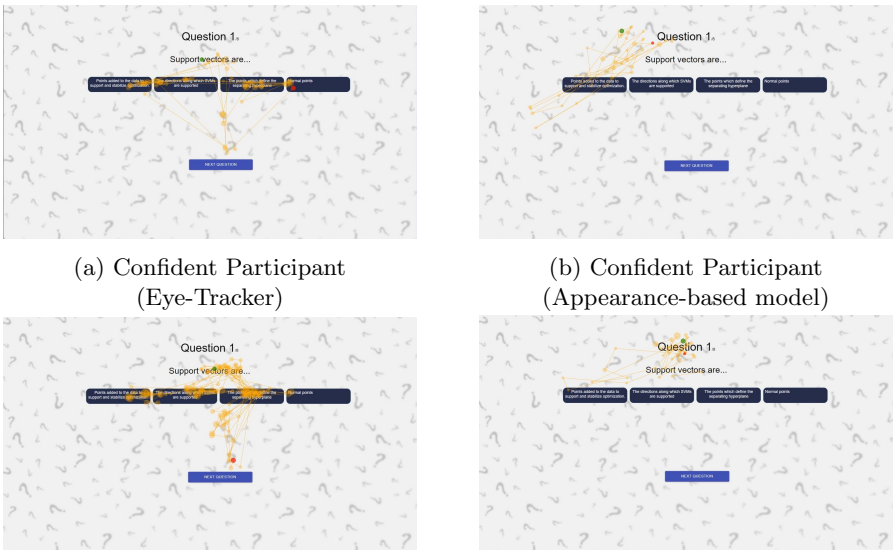
Figure 1: Overview of our proposed method.

3 Methodology

This section provides an overview of the entire process, starting with the data collection procedure, data preprocessing, feature extraction, model architecture, and evaluation strategy. Figure 1 shows the overall flow of our approach. First, we detail the dataset's preparation process, including the steps to ensure the data was suitable for analysis. This is followed by a description of the feature extraction methodology, where we outline how meaningful features were derived from the raw data to feed into the models. Next, we discuss the machine-learning and deep-learning models utilized in this study, explaining their structure and how they were trained. Finally, we present the evaluation protocol, which includes cross-validation techniques used to assess the effectiveness of the models.

3.1 Data Preprocessing

Facial videos were collected through a webcam during the experiment. In this study, the model introduced by Bhatt et al. was employed to generate 2D gaze points [45] to extract gaze information from these videos. This method enabled the tracking of gaze behavior, which was essential for analyzing confidence-related variations in eye movement patterns. Eye-tracking data, however, can be influenced by various noise sources, such as blinks and head movements. To mitigate these challenges, noise-reduction techniques were applied to eliminate distortions and enhance the reliabil-



(a) Confident Participant (Eye-Tracker)

(b) Confident Participant (Appearance-based model)

(c) Not Confident Participant (Eye-Tracker)

(d) Not Confident Participant (Appearance-based model)

Figure 2: Comparison of eye gaze patterns for confident and non-confident participants captured using an eye-tracker and appearance-based model during question responses.

ity of the eye movement measurements. These techniques allowed for a more precise examination of eye movement metrics, including fixations (sustained gazes at a specific location) and saccades (rapid transitions between fixations). The method introduced by Buscher et al. was used to analyze the proposed model to detect fixations and saccades [50]. This approach is particularly effective in mitigating noise caused by factors such as blinks, head movements, and other distortions that may impact the accuracy of gaze measurement. In this context, a fixation is defined as maintaining a gaze on a specific location for a brief period (typically less than one second), while a saccade refers to a rapid eye movement between fixation points. Rather than exporting the absolute coordinates of fixations, the differential coordinates were extracted to capture positional changes between consecutive fixations. These differential eye gaze patterns highlight variations in gaze behavior when answering questions with and without confidence, as depicted in Figure 2 illustrates the differences in eye gaze patterns observed when solving questions with and without confidence, using two distinct eye-tracking methods: hardware-based eye trackers and appearance-based approaches.

3.2 Features Extraction

Table 1 shows a list of features used in this study. Our methodology integrates traditional feature engineering techniques with state-of-the-art

Table 1: The list of features

No	Feature
1-2	Fixation duration {mean, std}
3-4	Saccade length {mean, std}
5-6	Saccade angle {mean, std}
7-8	Saccade speeds {mean, std}

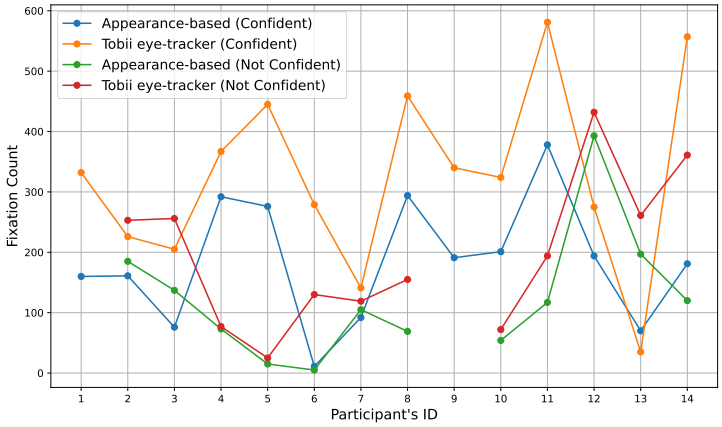


Figure 3: Comparison of fixation counts for confident and not-confident labels using the Eye-Tracker and Appearance-Based model across participants.

deep-learning approaches to achieve a robust evaluation of self-confidence estimation. We derived a comprehensive set of handcrafted features from the gaze data from the eye-tracker and appearance-based model, including fixation duration, saccade length, angle, and velocity. By transforming the raw data into these high-level descriptors, we ensure a precise and contextually relevant representation of how audiovisual stimuli modulate eye movement patterns and cognitive engagement. This fusion of handcrafted and learned features enables a more nuanced understanding of participants' self-confidence dynamics. Figure 3 illustrates a comparative analysis of fixation counts for confident and non-confident responses across participants using the Eye-Tracker and the Appearance-Based model.

3.3 Model Architecture

This study investigates the efficacy of combining conventional hand-crafted feature extraction techniques with advanced machine learning and deep-learning models for predicting self-confidence in video-based learning environments. Our approach adopts a dual strategy:

1. We explored traditional machine learning methods by utilizing handcrafted features derived from eye-tracking data. These features were

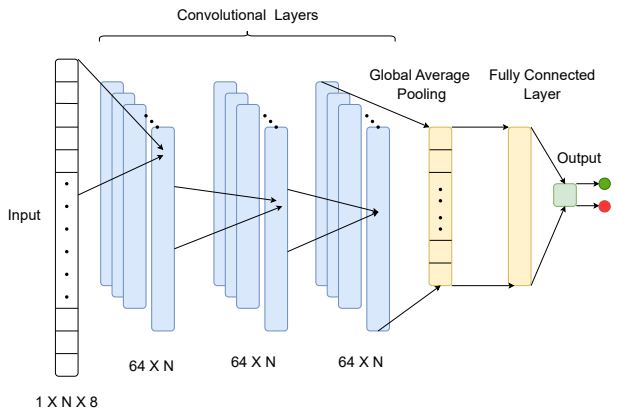


Figure 4: Proposed deep-learning based 1D CNN architecture.

input into a suite of machine learning algorithms, including Support Vector Machines (SVM), Random Forest (RF), Logistic Regression (LR), and Extreme Gradient Boosting (XGBoost).

2. We employed a one-dimensional Convolutional Neural Network (1D-CNN) and Transformer network, deep-learning models particularly adept at capturing intricate temporal dependencies and patterns in sequential data.

Feature-extraction based Model

Our hand-crafted feature-based methodology utilizes four well-established machine learning algorithms for self-confidence estimation: Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), and Extreme Gradient Boosting (XGBoost). We employed the Radial Basis Function (RBF) kernel for SVM, which is particularly effective in capturing non-linear relationships in the feature space. Using a grid search, we optimized the hyperparameters, identifying the best values as $C = 1$ *and* $\gamma = 0.125$. The RF algorithm was configured with `n_estimators = 100` and `criterion = gini` leveraging the ensemble of decision trees to provide robust and accurate predictions. For LR, we applied `class_weight = balanced` to handle class imbalance and set `max_iter = 1000` to ensure convergence during training. This algorithm provided a strong baseline for binary classification tasks, relying on the probabilistic interpretation of feature contributions. To address the class imbalance, the XGBoost model incorporated `scale_pos_weight` calculated as the ratio of negative to positive samples. Additionally, we used `eval_metric = logloss` for evaluation and fine-tuned its parameters to maximize performance. These settings allowed XGBoost to model complex feature interactions and improve predictive accuracy efficiently.

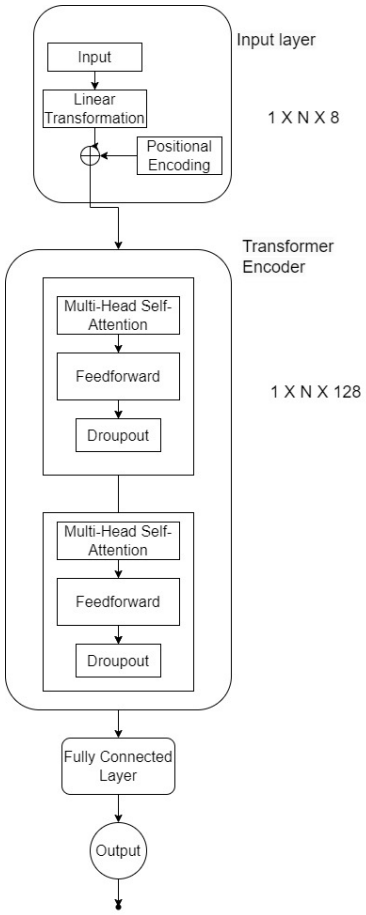


Figure 5: Proposed Transformer architecture.

Deep-Learning based Model

Our deep-learning model utilized a 1D CNN architecture, as illustrated in Figure 5. We applied padding to preprocess the input data to ensure consistent sequence lengths across all samples, allowing the model to handle varying input sizes. The input data consisted of sequences with a maximum length of 308 and 297 data points generated through an eye-tracker and an appearance-based model, respectively. The 1D CNN architecture featured three convolutional layers, each with `filters = 64` and a `kernel_size = 3`, followed by batch normalization and dropout with a rate of 0.1 to regularize the network and mitigate overfitting. After the convolutional layers, a global average pooling layer was used to reduce the sequence dimensions, yielding a fixed-length feature vector of size 64. This feature vector was then passed through a fully connected layer for binary classification. The model was optimized using the Adam optimizer

with an initial learning rate of 0.001, and the Binary Cross-Entropy loss function was used as the objective function for training. Class weights were incorporated into the loss function to address the class imbalance, ensuring that the model appropriately emphasized the minority class.

On the other hand, the Transformer model architecture consisted of an input embedding layer, which mapped the input features of size 8 to a hidden dimension of 128 through a linear transformation. Positional embeddings were added to the input embeddings to encode positional information in the sequences. These positional embeddings were initialized using Xavier uniform initialization to ensure stable convergence. The core of the architecture featured a Transformer encoder comprising 2 layers. Each encoder layer included a multi-head self-attention mechanism with 4 attention heads and a feedforward layer with a hidden size of 512 (4 times the hidden dimension). Dropout with a rate of 0.1 was applied to the encoder layers for regularization to mitigate overfitting. The final encoded sequence representation was aggregated using mean pooling along the sequence dimension to generate a fixed-length feature vector. This vector was passed through a fully connected layer for binary classification. The Adam optimizer optimized the model with an initial learning rate of 0.0001 and weight decay set to 0.0001. The Binary Cross-Entropy with Logits Loss (BCEWithLogitsLoss) function was employed as the objective function. Class weights were incorporated into the loss function to address the class imbalance, ensuring the model appropriately emphasized the minority class during training. A learning rate scheduler, *StepLR*, was applied with a step size of 10 and a $\gamma = 0.1$ to reduce the learning rate during training, helping in convergence.

3.4 Evaluation Strategy

Leave-One-Participant-Out-Cross-Validation (LOPOCV)

The LOPOCV method assessed the model's generalizability in predicting self-confidence for unseen participants. In this approach, data from one participant were excluded from the training set during each iteration and used solely for testing. This process was repeated until data from every participant had been excluded and tested.

Leave-One-Question-Out Cross-Validation (LOQOCV)

The LOQOCV approach evaluated the model's performance by considering "solving" activities together. In each iteration of LOQOCV, one question was excluded from the training dataset and used as the test set. The model was trained on the remaining categories, and this process was repeated until each category had been excluded and tested once. This approach allowed us to assess the model's ability to predict self-confidence across different activity categories. The model's overall accuracy was calculated by averaging the accuracy scores obtained from all iterations, providing a comprehensive performance metric across categories.

The importance of evaluating model performance using standard training-testing splits is well-recognized in deep learning research. While LOPOCV

and LOQOCV were chosen to ensure robust generalizability across unseen participants and questions, incorporating standard training- testing splits could provide additional insights into model performance under more conventional evaluation settings.

4 Data Collection

This section details the gaze data collection process, including an experimental setup and workflow overview. The following subsections provide further information about participant demographics and the data collection protocol.

4.1 Participants

The participant pool consisted of 14 university students with diverse academic and cultural backgrounds. Among them, ten were male, and four were female. The participants were recruited from Computer Science, Cognitive Science, Architecture, and Mechanical Engineering, ensuring a representative sample for evaluating the proposed framework. The group included twelve master's students and two bachelor's students, aged 23 to 30 years. Participants were from India, Russia, Iran, and Germany, reflecting an international cohort. Additionally, 53% of the participants reported consuming caffeine before the experiment.

4.2 Experimental Protocol

The study took place in a controlled lab setting using a Tobii 4C remote eye-tracker with a pro license and a web camera to reduce distractions and significantly influence the results. The videos used to track eye movements, and focus spanned subjects like logic, literature, computer science, and medicine to cover a range of themes effectively. These specific topics were chosen thoughtfully to guarantee representation and significance. The data collection process involved well-defined steps to ensure consistency and reliability. These steps are detailed below:

1. Each participant received a detailed briefing from the experiment conductor about the purpose of the study, the data collection process, and the overall expectations for their participation.
2. Participants reviewed and signed an informed consent form, confirming their understanding and voluntary agreement to participate in the study.
3. Participants completed a demographic questionnaire to provide relevant background information.
4. Participants were seated in front of a computer screen and instructed to attentively watch a series of video stimuli lasting approximately one to two minutes each. They were explicitly asked to follow the experimental guidelines of carefully watching the videos.

5. While participants viewed each task video, their facial expressions were recorded using a webcam to capture their gaze movements, and eye movements were recorded using a remote eye tracker.
6. As participants responded to the questionnaire, their facial expressions and eye movements were continuously recorded using the same equipment.
7. After answering each question, participants were asked to indicate their confidence level in their responses by selecting "Yes" or "No". This self-reported confidence served as a benchmark to evaluate the system's performance in assessing participant self-assurance.
8. Upon completing the confidence assessment for each question, the following video was automatically played, and the process (Steps 5–8) was repeated for all videos in the series.
9. After the experiment, participants were thanked for their time and effort and received a 10 Euro gift card as a token of appreciation.

The experiment lasted 30 minutes. During the trial, the desktop used for the study was stabilized to prevent movement or shaking. The experimental environment was carefully arranged to eliminate potential sources of interference, such as debris or exposure to other devices, ensuring optimal conditions for gaze data recording. Additionally, the system's volume was standardized across all participants, and the screen brightness was uniformly set for all trials. These measures were implemented to create a controlled and standardized environment, minimizing external influences that could negatively impact data collection or compromise the reliability of the results.

5 Results and Discussion

This section presents the results of a comprehensive comparison between hand-crafted feature-based methods and a deep-learning-based approach for estimating self-confidence from eye movements. The objective is to provide an in-depth analysis of each methodology's strengths and limitations within the study's context.

Table 2 presents the performance metrics for various machine learning and deep-learning models, including SVM, LR, RF, XGBoost, 1D CNN, and Transformers, trained on eye-tracking data obtained from both an eye-tracker and an appearance-based model under the LOPOCV framework. Similarly, Table 3 provides the evaluation metrics for the same set of models trained on the same data but evaluated using the LOQOCV framework. The reported metrics include precision, recall, F1-score, and macro averages for the classes: "Not Confident" and "Confident". These metrics enable a comprehensive comparison of model performance across the two distinct validation strategies.

To ensure a comprehensive and objective assessment of model performance, we employed standard classification metrics, including precision, recall, and F1-score, alongside macro F1-score for overall evaluation. These metrics were chosen because self-confidence detection is inherently

Table 2: Performance evaluation of machine learning and deep-learning models using eye-tracking data and an appearance-based approach under the Leave-One-Participant-Out Cross-Validation (LOPOCV) framework.

Device	Model	Class Weights	Macro	Class	Precision	Recall	F1-score
Eye-tracker	SVM	–	0.525	Not Confident	0.370	0.455	0.408
				Confident	0.684	0.605	0.642
	LR	–	0.557	Not Confident	0.406	0.549	0.467
				Confident	0.719	0.590	0.648
	RF	–	0.465	Not Confident	0.307	0.142	0.194
				Confident	0.655	0.835	0.735
	XGBoost	–	0.507	Not Confident	0.349	0.324	0.336
				Confident	0.666	0.691	0.678
	1D CNN	No	**0.662**	Not Confident	0.687	0.354	0.468
				Confident	0.789	0.937	0.857
		Yes	0.559	Not Confident	0.392	0.677	0.471
				Confident	0.811	0.537	0.646
	Transformer	No	0.418	Not Confident	0.000	0.000	0.000
				Confident	0.720	1.000	0.837
		Yes	0.418	Not Confident	0.000	0.000	0.000
				Confident	0.720	1.000	0.837
Appearance-based	SVM	–	0.520	Not Confident	0.393	0.549	0.458
				Confident	0.667	0.515	0.582
	LR	–	0.530	Not Confident	0.400	0.510	0.448
				Confident	0.669	0.564	0.612
	RF	–	0.452	Not Confident	0.296	0.182	0.226
				Confident	0.617	0.752	0.678
	XGBoost	–	0.468	Not Confident	0.322	0.312	0.317
				Confident	0.614	0.624	0.619
	1D CNN	No	**0.635**	Not Confident	0.888	0.266	0.410
				Confident	0.763	0.986	0.860
		Yes	0.504	Not Confident	0.353	0.935	0.513
				Confident	0.931	0.337	0.495
	Transformer	No	0.448	Not Confident	1.000	0.033	0.064
				Confident	0.712	1.000	0.832
		Yes	0.413	Not Confident	0.000	0.000	0.000
				Confident	0.705	1.000	0.827

imbalanced, with a tendency for models to favor the majority class. Furthermore, the macro F1-score was used to account for class imbalances by averaging F1-scores across both confidence levels, ensuring that the model's performance is not disproportionately influenced by a dominant class. By integrating these metrics, we aimed to capture the nuanced trade-offs between different models and cross-validation strategies, offering a robust framework for evaluating self-confidence detection accuracy.

Table 3: Baseline performance of machine learning and deep-learning models using eye-tracking data and an appearance-based approach under the Leave-One-Question-Out Cross-Validation (LOQOCV) framework.

Device	Model	Class Weights	Macro	Class	Precision	Recall	F1-score
Eye-tracker	SVM	–	0.529	Not Confident	0.377	0.534	0.442
				Confident	0.697	0.549	0.614
	LR	–	**0.560**	Not Confident	0.409	0.579	0.48
				Confident	0.727	0.573	0.641
	RF	–	0.473	Not Confident	0.308	0.189	0.234
				Confident	0.653	0.782	0.712
	XGBoost	–	0.536	Not Confident	0.389	0.385	0.387
				Confident	0.683	0.687	0.685
	1D CNN	No	0.492	Not Confident	0.428	0.096	0.157
				Confident	0.730	0.950	0.826
		Yes	0.504	Not Confident	0.353	0.935	0.513
				Confident	0.931	0.337	0.495
	Transformer	No	0.514	Not Confident	0.300	0.290	0.295
				Confident	0.728	0.737	0.732
		Yes	0.559	Not Confident	0.358	0.612	0.452
				Confident	0.793	0.575	0.666
Appearance-based	SVM	–	0.509	Not Confident	0.379	0.497	0.430
				Confident	0.651	0.535	0.588
	LR	–	0.538	Not Confident	0.408	0.511	0.454
				Confident	0.674	0.577	0.622
	RF	–	0.436	Not Confident	0.266	0.161	0.200
				Confident	0.609	0.746	0.671
	XGBoost	–	0.536	Not Confident	0.389	0.385	0.387
				Confident	0.683	0.687	0.685
	1D CNN	No	0.520	Not Confident	0.416	0.166	0.238
				Confident	0.722	0.902	0.802
		Yes	0.431	Not Confident	0.31	0.766	0.442
				Confident	0.750	0.291	0.420
	Transformer	No	**0.616**	Not Confident	0.435	0.566	0.492
				Confident	0.793	0.694	0.740
		Yes	0.445	Not Confident	0.290	0.600	0.391
				Confident	0.700	0.388	0.500

5.1 LOPOCV Analysis

Table 2 presents the results for Leave-One-Participant-Out Cross-Validation (LOPOCV) across different models. For eye-tracker-based models, the SVM achieved a macro F1-score of 0.525, demonstrating moderate performance. It excelled in recall for the "Confident" class but struggled with "Not Confident" predictions due to lower precision and recall for that class. Logistic Regression (LR) slightly outperformed the SVM with a macro F1-score of 0.557, exhibiting similar behavior by favoring the "Confident" class.

The Random Forest (RF) model performs lowest with a macro F1-

score of 0.465, highlighting its difficulty handling class imbalance, especially for "Not Confident" predictions. XGBoost improved over RF, achieving a macro F1-score of 0.507, but it remained less effective than SVM and LR.

The 1D CNN, trained without class weights, achieved the best performance for eye-tracker data with a macro F1-score of 0.662. It balanced predictions well across both classes, demonstrating high precision for "Confident" and reasonable performance for "Not Confident". However, when class weights were introduced, its performance dropped to 0.559 due to a trade-off: recall for "Not Confident" improved significantly (from 0.354 to 0.677), but precision for the same class fell sharply (from 0.687 to 0.392), reducing overall effectiveness.

The Transformer model performed the worst among all eye-tracker-based models, with a macro F1-score of 0.418. It consistently struggled with "Not Confident" predictions, and adding class weights failed to improve its performance, underscoring its difficulty with imbalanced data.

The SVM achieved a macro F1-score of 0.520 for appearance-based models, showing moderate and balanced performance across classes, similar to its performance on eye-tracker data. LR slightly surpassed SVM with a macro F1-score of 0.530, improving recall for "Confident.". RF underperformed with a macro F1-score of 0.452, reflecting its inability to effectively classify "Not Confident" instances. XGBoost slightly outperformed RF with a macro F1-score of 0.468 but still lagged behind SVM and LR.

The 1D CNN, without weights, also achieved the best performance for appearance-based models, with a macro F1-score of 0.635. It performed well for both classes, excelling in identifying "Confident" instances. However, introducing class weights reduced its macro F1-score to 0.504. Like the eye-tracker results, weights improved recall for "Not Confident" (from 0.266 to 0.935) but caused a significant drop in precision (from 0.888 to 0.353), leading to an overall decline. The Transformer model again performed the worst, achieving a macro F1-score of 0.448 without weights and 0.413 with weights, struggling consistently with imbalanced data and showing minimal improvement.

In conclusion, the 1D CNN without class weights achieved the highest macro F1-score across both device types, with 0.662 for eye-tracker data and 0.635 for appearance-based data, making it the most robust model. While class weights improved recall for the minority class ("Not Confident"), the precision and overall performance trade-offs limited generalization. These results highlight the ability of deep-learning models, particularly the 1D CNN, to effectively learn complex patterns in the data, even without explicit adjustments for class imbalance. Therefore, the 1D CNN without weights is recommended as the best-performing model.

5.2 LOQOCV Analysis

Table 3 summarizes the performance of various models under the Leave-One-Question-Out Cross-Validation (LOQOCV) framework. For eye-tracker-based models under the Leave-One-Question-Out Cross-Validation (LOQOCV) framework, the SVM achieved a macro F1-score of 0.529, show-

ing moderate performance with strong recall for the "Confident" class but limited precision for the "Not Confident" class, indicating an imbalance. Logistic Regression (LR) slightly outperformed SVM with a macro F1-score of 0.560, providing better overall performance while still favoring recall for the class "Confident" over precision "Not Confident". Random Forest (RF) scored lower, with a macro F1-score of 0.473, reflecting poor handling of the imbalance, particularly for the class "Not Confident". XGBoost achieved a macro F1-score of 0.536, performing better than RF but less effectively than LR. The 1D CNN, without class weights, reached a macro F1-score of 0.492, with acceptable performance for the class "Confident" but poor recall for class "Not Confident" (0.096), lowering overall effectiveness. Adding class weights improved the 1D CNN's score to 0.504 by significantly increasing recall for class "Not Confident" (0.935) but reducing its precision. The Transformer model achieved 0.514 without weights, showing modestly balanced results. With class weights, its score improved to 0.559, achieving a strong balance between recall and precision for both classes.

For appearance-based models, the SVM achieved a macro F1-score of 0.509, showing moderate performance with a slight bias toward the class "Confident". LR performed better, with a Macro F1-score of 0.538, improving recall for the class "Confident" but struggling with precision for the class "Not Confident". RF had the lowest score (0.436), indicating poor handling of class imbalance. XGBoost matched SVM with a macro F1-score of 0.537, showing moderate improvements over RF but no significant edge. The 1D CNN, without class weights, achieved 0.520, with high recall for class "Confident" (0.902) but low precision for class "Not Confident" (0.417), impacting the overall balance. Introducing class weights reduced its performance to 0.431, as precision for the class "Not Confident" dropped sharply (0.310) despite improved recall. The Transformer model achieved the highest macro F1-score (0.616) without weights, balancing recall, and precision effectively for both classes. Applying weights reduced its performance to 0.445, as focusing on improving recall for the class "Not Confident" negatively affected precision.

In conclusion, the Transformer model without weights achieved the best performance for appearance-based data with a macro F1-score of 0.616, while Logistic Regression performed best for eye-tracker data with a macro F1-score of 0.560 across all models under the LOQOCV framework.

The 1D CNN without class weights emerged as the most effective model in our case due to its ability to capture temporal dependencies in gaze behavior, such as fixation durations and saccade transitions, through its convolutional layers. The local receptive fields of CNNs enable the model to extract meaningful gaze movement patterns while maintaining robustness to noise. Moreover, techniques like batch normalization and dropout help to mitigate overfitting, allowing the model to generalize effectively despite the relatively small size of the dataset. In contrast, the Transformer model, which relies on self-attention mechanisms to learn global dependencies, did not perform as well. A significant factor in this is the dataset size—Transformers typically require larger datasets to accurately learn attention distributions. Given the limited size of our dataset, the model faced challenges in effectively capturing relevant gaze-

based confidence patterns. Additionally, the class imbalance present in the dataset led the Transformer to overfit to the majority class ("Confident"), which impacted its ability to generalize to the minority class ("Not Confident"). Even when class weights were introduced, the model prioritized recall, sometimes at the expense of precision, which limited its overall performance. Therefore, in our specific case, the 1D CNN's ability to learn local temporal features with fewer parameters and regularization techniques proved to be more suitable for gaze-based confidence estimation. On the other hand, the Transformer's reliance on large datasets and global attention mechanisms presented challenges when applied to our dataset.

6 Limitations and Future Work

This study on self-confidence estimation using machine learning and deep learning provides valuable insights. However, it also highlights several limitations and challenges that need to be addressed. A key limitation is the potential for data collection bias, as the model's performance is highly dependent on the quality, size, and diversity of the gaze datasets used for training. Insufficient demographic representation and limited contextual variability in the dataset may constrain the model's ability to generalize its findings to broader populations.

Another critical issue is gaze noise, which arises from head pose variations, inaccuracies in eye-tracking devices, and fluctuations in environmental lighting. These factors introduce significant variability in the raw gaze data, potentially undermining the precision of confidence detection.

Moreover, generalization across individuals presents a substantial challenge due to the inherently subjective nature of gaze patterns, which can vary widely based on personal habits, cognitive styles, and task-specific influences. Such individual variability complicates the model's ability to identify universal patterns for confidence estimation.

Future work should focus on improving the robustness of appearance-based eye-tracking models. These models are currently sensitive to head movements, changing lighting conditions, and diverse user postures, which can affect the accuracy of point-of-gaze estimation. Developing more advanced architectures that incorporate pose-invariant gaze estimation techniques and domain adaptation strategies can help ensure reliable performance across dynamic and uncontrolled environments.

Additionally, expanding data collection to include multimodal inputs—such as speech cues, facial expressions, and physiological signals—could significantly enhance the accuracy of self-confidence detection. A multimodal approach would enable the model to capture richer contextual information, facilitating a more comprehensive assessment of learner confidence. This approach would also improve generalization, making the system applicable to diverse educational settings and learner profiles.

Finally, future efforts should aim at developing personalized self-confidence estimation frameworks that account for individual differences in gaze behavior. By integrating user-specific calibration or adaptive learning mechanisms, models could better capture subjective variations in confidence

expression. Such a personalized approach would enhance the accuracy and reliability of confidence estimation, further advancing adaptive learning systems in education.

7 Conclusion

We propose a novel method for estimating self-confidence using eye-tracking data obtained through eye-tracker-based and appearance-based approaches. The method integrates both hand-crafted machine-learning techniques and a deep-learning framework. In an experimental setup, participants watched videos and answered questions based on the content while their eye movement behavior was recorded and analyzed. The comparison of methods reveals that the deep-learning-based 1D CNN model outperforms the Transformer and traditional machine learning approaches in confidence estimation under the leave-one-participant-out cross-validation framework. In contrast, under the leave-one-question-out cross-validation framework, Logistic Regression demonstrated the best performance for eye-tracker data, with a macro F1-score of 0.560, while the Transformer excelled for appearance-based data, achieving a macro F1-score of 0.622.

Acknowledgment

This work is partially supported by JSPS KAKENHI Grant Number 23KK0188.

References

[1] Riku Higashimura, Ko Watanabe, Andrew Vargo, Motoi Iwata, Andreas Dengel, and Koichi Kise. Estimating unknown english words from user smartphone reading behaviors. *IEEE Access*, pages 1–1, 2024. doi: 10.1109/ACCESS.2024.3457510.

[2] Teruhiro Mizumoto, Yasuhiro Otoda, Chihiro Nakajima, Mitsuhiro Kohana, Motohiro Uenishi, Keiichi Yasumoto, and Yutaka Arakawa. Design and implementation of sensor-embedded chair for continuous sitting posture recognition. *IEICE TRANSACTIONS on Information and Systems*, 103(5):1067–1077, 2020.

[3] Lik-Hang Lee and Pan Hui. Interaction methods for smart glasses: A survey. *IEEE access*, 6:28712–28732, 2018.

[4] Haruki Suzawa, Ko Watanabe, Masakazu Iwamura, Koichi Kise, Andreas Dengel, and Shoya Ishimaru. Supporting smooth interruption in a video conference by dynamically changing background music depending on the amount of utterance. In *Adjunct Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2022 ACM International Symposium on Wearable Computers*, UbiComp/ISWC '22 Adjunct, page 299–302,

New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394239. doi: 10.1145/3544793.3560384. URL https://doi.org/10.1145/3544793.3560384.

[5] Dagen Wang and Shrikanth S. Narayanan. Robust speech rate estimation for spontaneous speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(8):2190–2201, 2007. doi: 10.1109/TASL.2007.905178.

[6] Ko Watanabe, Yusuke Soneda, Yuki Matsuda, Yugo Nakamura, Yutaka Arakawa, Andreas Dengel, and Shoya Ishimaru. Discaas: Micro behavior analysis on discussion by camera as a sensor. *Sensors*, 21 (17):5719, 2021.

[7] Toshiki Hayashida, Yugo Nakamura, Hyuckjin Choi, and Yutaka Arakawa. Privacy-aware quantitative measurement of psychological state in meetings based on non-verbal cues. In *2024 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, pages 433–436, 2024. doi: 10.1109/PerComWorkshops59983.2024.10502817.

[8] Jacob Whitehill, Gwen Littlewort, Ian Fasel, Marian Bartlett, and Javier Movellan. Toward practical smile detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):2106–2111, 2009. doi: 10.1109/TPAMI.2009.42.

[9] Chenhao Chen, Yutaka Arakawa, Ko Watanabe, and Shoya Ishimaru. Quantitative evaluation system for online meetings based on multimodal microbehavior analysis. *Sensors and Materials*, 34(8):3017–3027, 2022.

[10] Hong Liu, Yuan Gao, and Pinging Wu. Smile detection in unconstrained scenarios using self-similarity of gradients features. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 1455–1459, 2014. doi: 10.1109/ICIP.2014.7025291.

[11] Ko Watanabe, Tanuja Sathyanarayana, Andreas Dengel, and Shoya Ishimaru. Engauge: Engagement gauge of meeting participants estimated by facial expression and deep neural network. *IEEE Access*, 11:52886–52898, 2023. doi: 10.1109/ACCESS.2023.3279428.

[12] Ko Watanabe, Andreas Dengel, and Shoya Ishimaru. Metacognition-engauge: Real-time augmentation of self-and-group engagement levels understanding by gauge interface in online meetings. In *Proceedings of the Augmented Humans International Conference 2024*, AHs '24, page 301–303, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400709807. doi: 10.1145/3652920.3653054. URL https://doi.org/10.1145/3652920.3653054.

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

[14] Dimitris Giakoumis, Dimitrios Tzovaras, Konstantinos Moustakas, and George Hassapis. Automatic recognition of boredom in video games using novel biosignal moment-based features. *IEEE Transactions on Affective Computing*, 2(3):119–133, 2011. doi: 10.1109/T-AFFC.2011.4.

[15] Rafael A Calvo and Sidney D'Mello. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on affective computing*, 1(1):18–37, 2010.

[16] Shoya Ishimaru, Takanori Maruichi, Andreas Dengel, and Koichi Kise. Confidence-aware learning assistant, 2021. URL `https://arxiv.org/abs/2102.07312`.

[17] Kento Yamada, Koichi Kise, and Olivier Augereau. Estimation of confidence based on eye gaze: an application to multiple-choice questions. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*, pages 217–220, 2017.

[18] Ian Sadler. The role of self-confidence in learning to teach in higher education. *Innovations in Education and Teaching International*, 50 (2):157–166, 2013. doi: 10.1080/14703297.2012.760777. URL `https://doi.org/10.1080/14703297.2012.760777`.

[19] Juan Carlos Ortiz-Ordoñez, Friederike Stoller, and Bernd Remmele. Promoting self-confidence, motivation and sustainable learning skills in basic education. *Procedia - Social and Behavioral Sciences*, 171: 982–986, 2015. ISSN 1877-0428. doi: https://doi.org/10.1016/j.sbspro.2015.01.205. 5th ICEEPSY International Conference on Education Educational Psychology.

[20] Elizabeth A Linnenbrink and Paul R Pintrich. The role of self-efficacy beliefs instudent engagement and learning intheclassroom. *Reading &Writing Quarterly*, 19(2):119–137, 2003.

[21] Yiannis Demiris. Prediction of intent in robotics and multi-agent systems. *Cognitive processing*, 8(3):151–158, 2007.

[22] Hyun Soo Park, Eakta Jain, and Yaser Sheikh. Predicting primary gaze behavior using social saliency fields. In *2013 IEEE International Conference on Computer Vision*, pages 3503–3510, 2013. doi: 10.1109/ICCV.2013.435.

[23] Yihua Cheng, Haofei Wang, Yiwei Bao, and Feng Lu. Appearance-based gaze estimation with deep learning: A review and benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–20, 2024. doi: 10.1109/TPAMI.2024.3393571.

[24] Radka Jersakova, Richard J Allen, Jonathan Booth, Céline Souchay, and Akira R O'Connor. Understanding metacognitive confidence: Insights from judgment-of-learning justifications. *Journal of Memory and Language*, 97:187–207, 2017.

[25] Sabina Kleitman and Jennifer Gibson. Metacognitive beliefs, self-confidence and primary learning environment of sixth grade students. *Learning and Individual Differences*, 21(6):728–735, 2011.

[26] Jennifer A Pooler, Ruth E Morgan, Karen Wong, Margaret K Wilkin, and Jonathan L Blitstein. Cooking matters for adults improves food resource management skills and self-confidence among low-income participants. *Journal of nutrition education and behavior*, 49(7):545–553, 2017.

[27] Katherine Forbes-Riley and Diane J Litman. Adapting to student uncertainty improves tutoring dialogues. In *AIED*, pages 33–40, 2009.

[28] Jerry Chih-Yuan Sun and Katherine Pin-Chen Yeh. The effects of attention monitoring with eeg biofeedback on university students' attention and self-efficacy: The case of anti-phishing instructional materials. *Computers & Education*, 106:73–82, 2017.

[29] Thomas Roderer and Claudia M Roebers. Can you see me thinking (about my answers)? using eye-tracking to illuminate developmental differences in monitoring and control skills and their relation to performance. *Metacognition and learning*, 9:1–23, 2014.

[30] Lu-Ho Hsia, Iwen Huang, and Gwo-Jen Hwang. Effects of different online peer-feedback approaches on students' performance skills, motivation and self-efficacy in a dance course. *Computers & Education*, 96:55–71, 2016.

[31] Takanori Maruichi, Shoya Ishimaru, and Koichi Kise. Self-confidence estimation on vocabulary tests with stroke-level handwriting logs. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 3, pages 18–22. IEEE, 2019.

[32] Adrian Bruhin, Fidel Petros, and Luís Santos-Pinto. The role of self-confidence in teamwork: Experimental evidence. *Experimental Economics*, pages 1–26, 2024.

[33] Kazuaki Kojima, Keiich Muramatsu, and Tatsunori Matsui. Experimental study toward estimation of a learner mental state from processes of solving multiple choice problems based on eye movements. In *20th International Conference on Computers in Education, ICCE 2012*, 2012.

[34] Ayano Okoso, Takumi Toyama, Kai Kunze, Joachim Folz, Marcus Liwicki, and Koichi Kise. Towards extraction of subjective reading incomprehension: Analysis of eye gaze features. In *Proceedings of the 33rd annual acm conference extended abstracts on human factors in computing systems*, pages 1325–1330, 2015.

[35] Hanju Lee, Yasuhiro Kanakogi, and Kazuo Hiraki. Building a responsive teacher: how temporal contingency of gaze interaction influences word learning with virtual tutors. *Royal Society open science*, 2(1): 140361, 2015.

[36] Olivier Augereau, Hiroki Fujiyoshi, and Koichi Kise. Towards an automated estimation of english skill via toeic score based on reading analysis. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 1285–1290. IEEE, 2016.

[37] Kento Yamada, Koichi Kise, and Olivier Augereau. Estimation of confidence based on eye gaze: an application to multiple-choice questions. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*, pages 217–220, 2017.

[38] Feng Lu, Yusuke Sugano, Takahiro Okabe, and Yoichi Sato. Adaptive linear regression for appearance-based gaze estimation. *IEEE transactions on pattern analysis and machine intelligence*, 36(10): 2033–2046, 2014.

[39] Oliver Williams, Andrew Blake, and Roberto Cipolla. Sparse and semi-supervised visual mapping with the s^ 3gp. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 230–237. IEEE, 2006.

[40] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4511–4520, 2015. doi: 10.1109/CVPR.2015.7299081.

[41] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra M. Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye tracking for everyone. *CoRR*, abs/1606.05814, 2016. URL `http://arxiv.org/abs/1606.05814`.

[42] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye tracking for everyone, 2016.

[43] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, may 2017. ISSN 0001-0782. doi: 10.1145/3065386. URL `https://doi.org/10.1145/3065386`.

[44] Yiwei Bao, Yihua Cheng, Yunfei Liu, and Feng Lu. Adaptive feature fusion network for gaze tracking in mobile tablets, 2021.

[45] Ankur Bhatt, Ko Watanabe, Andreas Dengel, and Shoya Ishimaru. Appearance-based gaze estimation with deep neural networks: From data collection to evaluation. *International Journal of Activity and Behavior Computing*, 2024(1):1–15, 2024.

[46] Petr Kellnhofer, Adrià Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. *CoRR*, abs/1910.10088, 2019. URL `http://arxiv.org/abs/1910.10088`.

[47] Kang Wang, Rui Zhao, Hui Su, and Qiang Ji. Generalizing eye tracking with bayesian adversarial learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11907–11916, 2019.

[48] Yihua Cheng, Feng Lu, and Xucong Zhang. Appearance-based gaze estimation via evaluation-guided asymmetric regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[49] Yu Yu and Jean-Marc Odobez. Unsupervised representation learning for gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7314–7324, 2020.

[50] Georg Buscher, Andreas Dengel, and Ludger van Elst. Eye movements as implicit relevance feedback. In *CHI'08 extended abstracts on Human factors in computing systems*, pages 2991–2996. 2008.