Reading your mind: Camera-based emotion recognition through human-environment gaze interaction

Anonymous authors

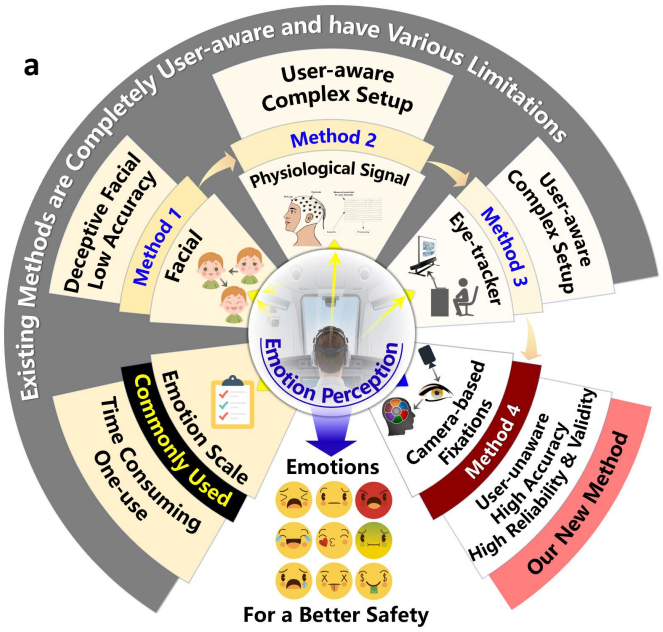
ABSTRACT

Emotion recognition represents a form of mind-reading that has long challenged researchers, with traditional methods facing significant limitations: physiological approaches require complex setups and make users conscious of monitoring, facial recognition struggles with deceptive expressions, and existing gaze-based methods lack precision with subtle emotional states. Our research achieves a breakthrough by effectively "reading minds" through an unobtrusive camera-based system that analyzes how humans visually interact with their environment. By integrating gaze fixation patterns with environmental semantics and temporal dynamics, we reveal that emotions are not merely physiological responses but complex outcomes of human-environment interactions. Our novel semantic interactive orders and EmoGazeNet model decode the intricate language of gaze behavior, transforming readily available standard HD camera input into a powerful window to internal emotional states — all while subjects remain completely unaware they are being monitored, eliminating the need for specialized equipment or user cooperation. Experimental results validate this mind-reading capability, demonstrating a 13% accuracy improvement over traditional gaze-based approaches and surpassing even physiological signal-based methods by 2.7% in detecting subtle emotional shifts. This technology represents a paradigm shift in emotion recognition — offering a scalable, cost-effective solution for applications ranging from driver monitoring to security surveillance — while proving that reading minds is possible through careful observation of how we visually engage with our world. The codes, datasets and results are publically available at this link.

Emotions[1–5] play a crucial role in human interactions, profoundly influencing our daily lives. Reading these internal emotional states — essentially "mind reading" — represents one of the most challenging yet valuable capabilities for human-computer interaction systems. Emotion recognition[6,7] focuses on developing systems that can automatically identify and interpret emotions, which can be mainly divided into three categories according to the data source: physiological signal, behavioral (such as facial expression), and eye movement signal (i.e., gaze). Applications of emotion recognition span various fields, such as human-computer interaction, healthcare, public security, education, and entertainment, offering significant benefits for user experience[8], emotional well-being[9–12], and interpersonal communication[13,14]. However, traditional emotion recognition methods[15–17], despite their accuracy, have limitations, as shown in Figure 1-**a** and Figure 1-**b**. These limitations have prevented the development of truly unintrusive "mind-reading" technologies capable of understanding emotions in everyday settings.

Current emotion recognition methods are divided into three levels based on the depth of emotional understanding they provide and the complexity of their setup (Figure 1-**a**). At the foundation are emotion scales[18–20], considered the most commonly used due to their accuracy, as they rely on individuals self-reporting their feelings. However, these scales are limited to one-time use, are time-consuming, and unsuitable for real-time or continuous monitoring. Building on this, facial-based methods[21–26] (**Method 1**), rely on externally observable cues, such as facial expressions[27–30], to infer emotions. These methods are "user-unaware" (meaning users are unaware that their data is being monitored) and easy to implement. However, they often struggle with interpreting ambiguous or deceptive emotions and are influenced by individual differences in expression, which limits their reliability and accuracy. Advancing to a deeper layer, physiological signal-based methods[31,32] (**Method 2**), capture internal responses like EEG[33–37] or heart rate to provide more direct insights into emotional states. While these methods offer objective and quantifiable data, they require wearable devices, making them "user-aware" (meaning users are aware that they are being monitored) and often uncomfortable. Additionally, they are costly, time-consuming to set up, and impractical for use in dynamic or open environments. Gaze (eye movement)-based methods[38,39] (**Method 3**) take a unique approach by focusing on eye movements, such as fixation metrics, pupil size, and gaze distribution, using eye trackers. These methods are non-intrusive and user-friendly, enabling natural and comfortable emotion detection without requiring direct user engagement. However, they are limited by the constraints of eye movement patterns, providing less emotional insight and resulting in reduced accuracy. Moreover, although these methods are considered "user-aware," they can be time-consuming to set up in certain circumstances. This highlights the challenge of balancing simplicity, user comfort, and the depth of emotional understanding, which is essential for creating effective emotion recognition systems that are both robust and adaptable for real-world applications.

Our groundbreaking contribution to the field is a new paradigm that transforms ordinary HD cameras into mind-reading devices. To address the challenges in emotion recognition, we propose a method that combines the strengths of existing approaches while overcoming their limitations. Ordinary HD cameras, which are inexpensive, easy to deploy, and widely available, could offer a promising alternative. Thus, we envision that by enabling users to interact freely with their environment, can we leverage external human-environment interactions to gain insights into internal emotions (Figure 2-(1))? To implement this, we must first identify what information an HD camera can capture from the user. While facial expressions are often unreliable for interpreting ambiguous or deceptive emotions, gaze information presents a valuable alternative. Studies have shown that gaze patterns are closely linked to emotional states, revealing indicators such as interest, stress, and cognitive engagement. This makes gaze a useful, though indirect, signal for understanding emotions. Despite its advantages — particularly its "user-unaware" nature, meaning no wearable devices or active user participation are required — gaze (eye movement)-based emotion recognition has limitations. Gaze patterns primarily capture the user's line of sight and fixation behavior, but they do not directly reflect deeper emotional responses or motivations. For instance, the duration or frequency of gazing at an object may not accurately correlate with emotional intensity, as cognitive processes and external factors can influence gaze. Consequently, the relationship between gaze and emotion remains indirect, making it challenging to draw precise emotional conclusions from gaze data alone.



b

Methods / Attributes	Acquisition Equipment	User Experience	Density	Need Wearable Device	Performance Characteristic
Physiological Signal	EEG, EMG, GSR	Sensitive	Once	Yes	High precision in **controlled environments**
Micro-expression	HD Camera	Insensitive	Intermittent	No	Effective in **specific contexts**
Gait	HD Camera	Insensitive	Intermittent	No	**Limited scalability** in complex environments
Body Surface Sensor	Wristband, Watch	Sensitive	Continuous	Yes	**High adaptability** to real-world scenarios
Emotion Scale	Scale	Sensitive	Once	No	**Subjective assessment** with potential for user bias
Eye Movement	Eye Tracker	Sensitive	Continuous	Yes	High precision in **controlled environments**
Eye Gaze&Environment (Context Gaze)	HD Camera	Insensitive	Continuous	No	**High reliability, validity and adaptability to real-world scenarios**

Figure 1. Comparison of existing emotion recognition methods. a Based on the depth of emotional understanding they provide and the complexity of their setup, emotion recognition methods are divided into four methods: **Method 1** (facial-based) has deceptive facial and low accuracy; **Method 2** physiological signal-based) offers deeper insights but requires complex setups and is "user-aware"; **Method 3** (eye-tracker/gaze-based) is also complex and "user-aware". **Method 4** (our new method) uses camera-based fixations for high-accuracy, user-unaware recognition, providing a simple, efficient solution that overcomes the limitations of traditional methods. **b** Comparative Analysis of Emotion Evaluation Methods and Their Attributes. This analysis highlights the advantages of the Eye Gaze & Environment method, which offers high accuracy and continuous data collection while minimizing user sensitivity, making it particularly effective for real-time emotion monitoring in diverse environments. "User Experience" describes the impact of different emotion evaluation methods on the user. "Density" describes the frequency of emotional data collection.

To further substantiate the innovation of combining gaze data with environmental context, it is beneficial to elaborate on the theoretical foundation and psychological evidence behind this approach. Research indicates that emotional responses can significantly influence how individuals allocate their attention to environmental stimuli, with different emotional states triggering distinct gaze patterns[40,41]. This aligns with the context-dependent emotional information processing theory, which suggests that emotions are not only shaped by internal cognitive processes but also by external environmental factors[42]. Specifically, gaze patterns reflect selective attention to contextually relevant objects, influenced by emotional states. Importantly, while gaze patterns can vary between individuals, the emotional context determines which environmental stimuli become more salient. This supports the idea that people with similar emotional states may direct their gaze toward different objects or in distinct sequences based on their personal experiences and cognitive tendencies. The integration of gaze with contextual information provides a richer, more nuanced understanding of emotional states, which differs from traditional gaze (eye movement)-based methods by adding a layer of semantic relevance. Our method uniquely captures this dynamic interplay between gaze and context, addressing limitations of prior gaze (eye movement)-based approaches that typically overlook the temporal and semantic dimensions of gaze behavior. Additionally, studies suggest that brain areas like the hippocampus and amygdala play critical roles in encoding emotional memories tied to contextual information[43], further supporting the idea that emotional states influence gaze patterns in a context-dependent manner. This theoretical grounding reinforces the potential of our approach to enhance emotion recognition by more accurately reflecting the intricate relationship between emotions, attention, and environmental stimuli.

However, further research[40,41] has shown that when gaze data is combined with the environmental context, it offers a much more accurate reflection of psychological states. This critical insight — that mind-reading can be achieved through analyzing how humans visually interact with their environment — forms the theoretical foundation of our approach. This aligns with the context-dependent emotional information processing theory[42], which posits that emotions are shaped not only by internal cognitive processes but also by the external environment. In the context of gaze-environment interactions, gaze patterns dynamically reflect an individual's selective attention to salient environmental stimuli, revealing how emotions guide cognitive resources and attention. From a cognitive neuroscience perspective, this process involves context-dependent attentional allocation, where emotional responses to stimuli activate the frontoparietal network, directing gaze to contextually relevant targets. Studies also suggest that the hippocampus and amygdala encode emotional memories in context[43], as gaze patterns shift with context, reflecting how the brain integrates situational information with emotions. This interaction between gaze and context, integrating situational information with emotions, could improve interpretation and provide deeper insights into how emotions are shaped by environments.

Building on these theoretical insights, we present a revolutionary emotion recognition framework that effectively "reads minds" through camera-based fixations that combines eye gaze patterns with environmental context (Figure 1-**a**-**Method 4**). This approach leverages the dynamic interplay between a user's visual attention and their surroundings to provide deeper insights into emotional states. To implement this, we developed a "user-unaware (users are not required to wear any devices and remain unaware that their data is being collected or that they are being monitored)" gaze tracking method that eliminates the need for specialized eye-tracking devices (Figure 2-(2)). Using commonly available HD cameras, this method captures gaze points in natural, unconstrained settings and maps them onto a gaze fixation scanpath through multi-angle observations of eye appearance and

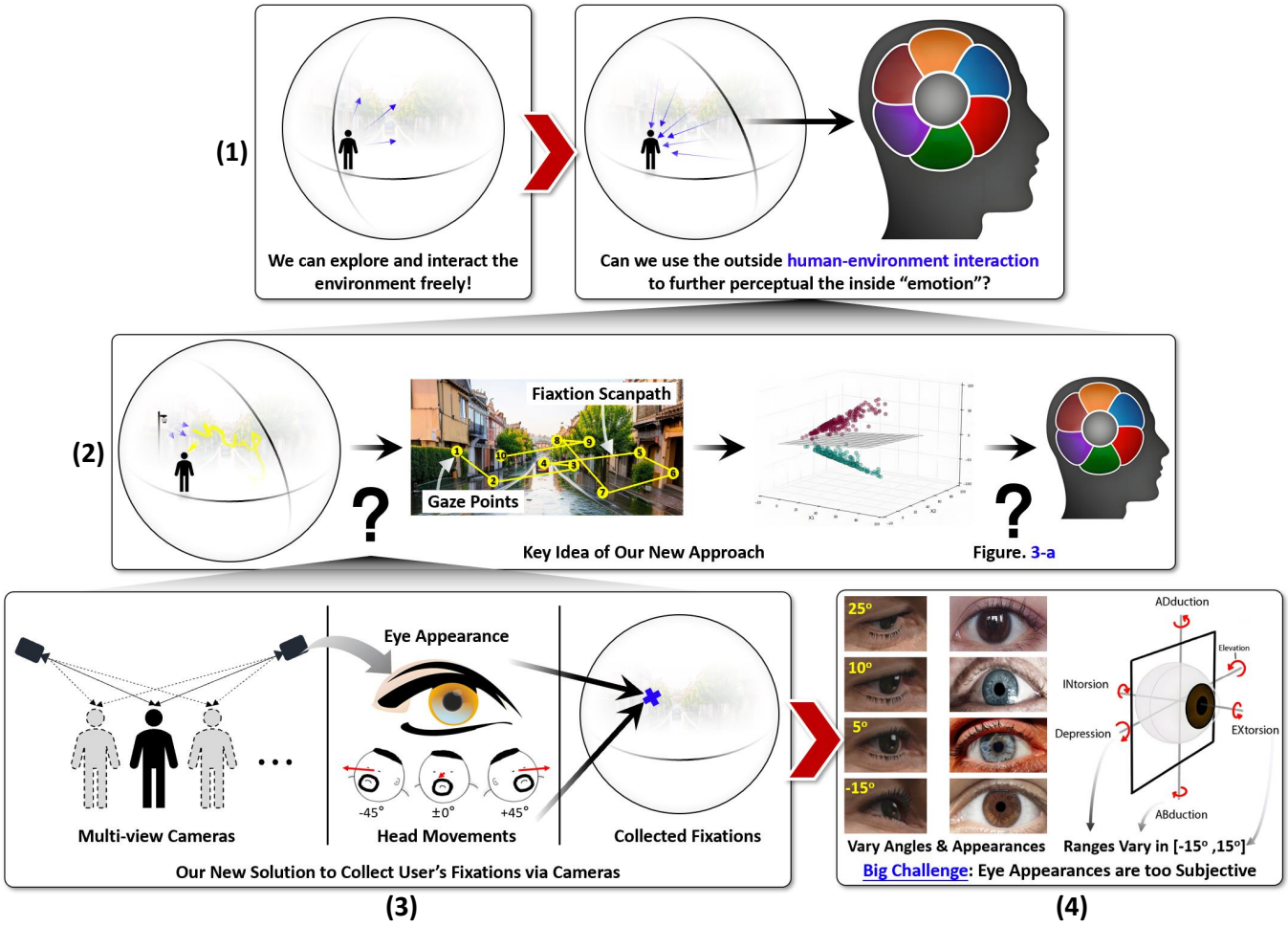


Figure 2. Human-environment interaction for contextual gaze-based emotion recognition. **(1)** illustrates the concept of leveraging human-environment interaction to infer emotions. **(2)** introduces a novel contextual gaze-based approach that combines fixation scanpaths with semantic understanding for deeper emotional insights. **(3)** A multi-camera system captures eye appearances and head movements to enable user-unaware, real-world emotion recognition. **(4)** Key challenges include variability in eye appearances and gaze angles, requiring robust gaze estimation techniques.

head movements (Figure 2-(3)). Crucially, it ensures that users remain unaware of the monitoring process, making it ideal for unobtrusive and continuous emotion monitoring. A major challenge, however, is that eye appearance is highly subjective (Figure 2-(4)). Factors such as variations in gaze angles, individual differences in eye features like sclera visibility and iris size, and the complexity of 3D eye movements make it difficult to achieve consistent and accurate tracking. Our solution to this challenge — an online personalized calibration method — represents a significant advance in making mind-reading technology practical and accurate in real-world environments. We incorporate this approach (see Figure 3-**a** & Methods — Online Personalized Calibration) that integrates subjective fixation (user-specific gaze tendencies) with objective fixation (scene-based salient points). This adaptive approach dynamically adjusts to individual differences, significantly enhancing gaze mapping accuracy and adaptability.

The core innovation of our work is the Semantic Interactive Orders (SIO) framework, which decodes the language of human gaze to reveal internal emotional states. Using the newly developed gaze tracking method (Figure 2-(3)), we collect raw eye fixation points. Unlike existing approaches that rely solely on gaze coordinates, which have demonstrated low accuracy in emotion recognition (Figure 1-**a**-**Method 3** and Figure 3-**b**(2)), SIO combines gaze coordinates with semantic environmental information and temporal dynamics (Figure 3-**b**(3)). By mapping fixation patterns to meaningful objects and their contextual interactions over time, this framework provides a richer representation of gaze behavior, significantly enhancing the accuracy and robustness of emotion recognition. Emphasizing a user-centric approach, the method is designed to ensure data security and anonymity, aligning with privacy regulations and addressing concerns in sensitive contexts.

Our experiments validate that this mind-reading technology significantly outperforms existing approaches, demonstrating that unobtrusive emotion recognition is not only possible but highly effective. Experimental results confirm that this integrated method significantly improves emotion recognition performance, particularly in detecting subtle emotional shifts. Our approach shows a 13% improvement in accuracy compared to traditional gaze (eye movement)-based methods. Furthermore, when compared to more objective and accurate EEG-based methods, our approach still outperforms them by 2.7% in scenarios involving non-extreme emotional changes[1], demonstrating its robustness. This breakthrough proves that accurate mind-reading is achievable through careful observation of how humans visually engage with their environment, offering a paradigm shift in emotion recognition technology. Additionally, this method is cost-effective, scalable, and enables continuous, remote emotion monitoring without the need for specialized sensors or complex setups, making it ideal for real-world applications.

[1] Non-extreme emotional changes refer to subtle or moderate shifts in emotion, such as slight changes in mood or feelings, rather than intense emotional reactions.

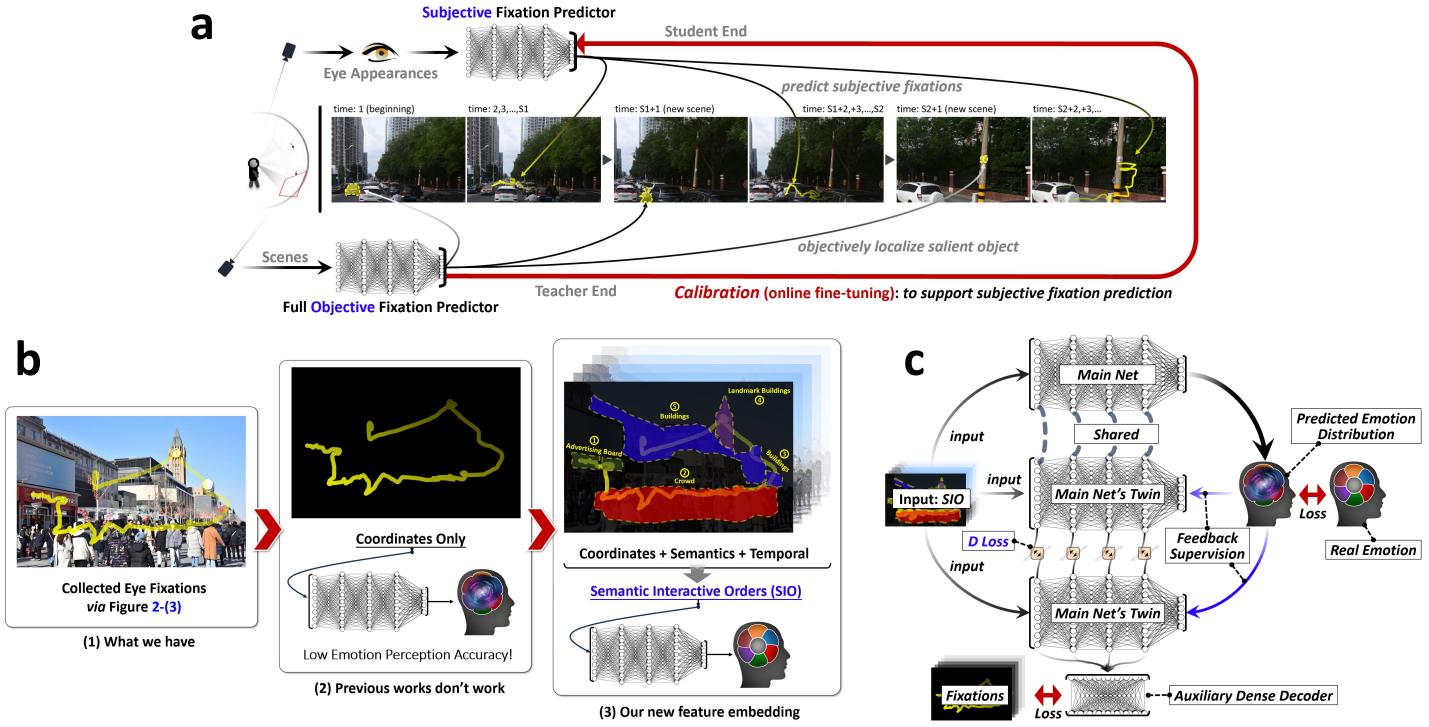


Figure 3. Calibration and semantic-aware modeling for improved contextual gaze-based emotion recognition. a shows an online calibration method that combines subjective (user-specific) and objective (scene-based) fixations to dynamically adapt gaze tracking for personalized emotion recognition. **b** compares traditional gaze-coordinate methods with the proposed Semantic Interactive Orders (SIO) framework, which integrates coordinates, semantics, and temporal dynamics for improved emotion detection. **c** showcases the proposed emotion recognition framework architecture. The model uses SIO feature embeddings as input, combined with a feedback supervision mechanism. A twin neural network structure predicts emotion distributions, and the framework employs auxiliary dense decoders and multiple loss functions to optimize performance, achieving efficient emotion recognition.

Results

A novel paradigm for emotion recognition

We present a novel emotion recognition framework that combines eye fixation patterns with environmental context analysis, addressing a critical gap in understanding how visual attention dynamically shapes emotional responses. Unlike traditional eye-tracking methods constrained by environmental stability requirements and limited capture ranges, our multi-camera system leverages standard HD cameras to remotely track gaze coordinates with millimeter precision, eliminating the need for specialized hardware. This "user-unaware", camera-based approach enables continuous, unobtrusive emotion monitoring, making it both cost-effective and highly scalable for real-world applications.

Camera-based gaze tracking method

Collection setting. We strategically position eight HD cameras around a designated area, ensuring full coverage and eliminating blind spots. This setup allows individuals to move freely within the space while continuously capturing images from all angles (see Figure 3-a). For each position within the space, a segment of the panoramic image corresponding to the individual's field of view (FOV) is projected onto a 2D plane. Detailedly, we use a third-person multi-camera panoramic modeling approach to ensure a "user-unaware" solution by generating a panoramic model of the scene from any location, capturing the user's gaze interaction with the environment (see Methods — Third-person multi-camera panoramic modeling section).

To achieve this, we recruited 30 annotators (18 females and 12 males, aged 18 to 28) to collect data on their eye appearance and regions of focus. Prior to data collection, participants underwent emotion induction through video and image stimuli. Each participant collected data six times in the same scene, under six distinct emotional states. Based on Paul Ekman's basic emotion theory[44], we also categorize emotions into six types, i.e., "Angry", "Disgust", "Fear", "Happy", "Sad", and "Surprised". To avoid memory residual interference (i.e., carryover effects from one emotion affecting the regions of focus in subsequent emotions), there was a two-day interval between each data collection session for different emotions within the same scene. In total, we collected data from one static indoor scenes, two high/low light indoor scene, and one dynamic outdoor scenes.

Gaze mapping. By analyzing the visual appearance of the eyes in this 2D projection, our system predicts the coordinates of the gaze point using existing advanced gaze prediction algorithms (see Supplementary Figure 4). This projection and prediction process occurs at short intervals, resulting in a comprehensive dataset of gaze coordinates mapped onto the 360-degree panoramic image over time. Notably, our findings indicate that a projection interval of 0.1 seconds optimizes the accuracy of gaze point collection (see Supplementary Figure 5). Since the current advanced gaze prediction algorithms learn the mapping between eye appearance and coordinates directly, ignoring the subjectivity of gaze data and the variability in eye appearance across individuals. This results in reduced generalization and accuracy of the mapping. To improve, we then employ an online personalized calibration method to reduce the interference caused by individual differences in eye appearance using our collected data (see Figure 3-**a** and Methods section). Further, due to variations in camera angles, lighting, and other factors, collected eye images may lack clarity, so we enhance them with super-resolution. Additionally, most existing datasets for gaze estimation include only limited head angles, restricting the range of eye appearances. Our method provides greater flexibility in gaze angles, capturing diverse viewpoints. To bridge this gap, we perform 3D reconstruction on the facial data from the existing gaze prediction dataset (e.g., ShanghaiTechGaze[45]), generating eye appearance data from various angles. This data then

retrains the model to better fit our requirements. The details of collection and post-processing of eye appearance data in real-world environments can be seen in Supplementary "Methods — Gaze Point Collection Process" section and Supplementary Algorithm 1 and Algorithm 2.

This innovative approach allows for precise and continuous tracking of eye gaze across a wide area, overcoming the limitations and discomfort associated with traditional eye tracking devices. The collected gaze coordinates are then integrated with the 360-degree panoramic images and fed into our proposed emotion recognition model, EmoGazeNet, enabling robust and accurate emotion detection. Furthre, to assess the accuracy of gaze point collection accuracy, we have introduced an object-box-based evaluation metric: if the gaze coordinates fall within the object box, it is considered accurate; otherwise, it is deemed to have a significant error. This metric provides a straightforward way to evaluate the precision of gaze estimation systems, ensuring that the predicted gaze points are closely aligned with the actual areas of interest within the environment (see Methods section).

Context Gaze-based deep model

Directly mapping eye appearance, gaze coordinates, and environmental context can lead the model to learn superficial visual patterns, associating environment with emotional states without understanding deeper gaze-related correlations. To address this, we introduce EmoGazeNet, a novel GAN-based model designed to capture meaningful, context-aware interactions between gaze behavior, emotional states, and environmental cues. EmoGazeNet takes two primary inputs: (1) a panoramic environment represented by an ERP (Equirectangular Projection) image, and (2) sequential gaze coordinates reflecting human-environment interactions. These gaze coordinates form a scanpath used to segment the ERP image into distinct object patches, ordered according to the sequence of viewing. Different arrangements of these patches effectively reflect variations in emotional states.

As shown in Figure 3-**c** & Supplementary Figure 1, EmoGazeNet, based on Generative Adversarial Networks (GANs), is designed with two main components: the Generator (Main Net) and the Discriminator (Main Net's Twins). The Generator is responsible for generating the probability distribution of emotion categories, while the Discriminator's task is to distinguish between real and generated data. Through adversarial training, both components continuously improve, with the Generator becoming better at generating realistic emotional state predictions, and the Discriminator sharpening its ability to differentiate between true and synthesized data. This enables the model to not only learn the direct relationships between gaze and emotional states but also refine its understanding of the deeper, context-aware correlations between eye movements and the surrounding environment. Details of EmoGazeNet are shown in the Supplementary "Methods - EmoGazeNet model achitecture".

Performance evaluation metric

To evaluate the performance of EmoGazeNet, we report three key metrics: Accuracy (Acc), F1 score, and Contextual Attention Weighted F1 Score (*caw*F1). Acc measures the overall correctness of the model's predictions, calculated as the ratio of correct predictions (both true positives and true negatives) to the total number of predictions made. F1 score provides a balanced measure of precision and recall, which is particularly useful in scenarios with imbalanced class distributions. Contextual Attention Weighted F1 Score (*caw*F1) is our proposed evaluation metric tailored specifically for emotion recognition tasks. This metric not only assesses classification performance but also incorporates fixation-context consistency. By doing so, *caw*F1 evaluates the model's ability to correctly classify emotions while simultaneously understanding the relationship between eye fixations and the environmental context, thereby providing a more comprehensive assessment of the model's performance.

Dataset Statistics for Validating the Proposed Method's Effectiveness

We evaluate the performance of our proposed method on three datasets.

Our 2D and 360-degree screen datasets. The variability in modalities across current datasets poses challenges for direct performance comparisons. To address this, we developed a new dataset EmoGaze2D-50 inspired by the SEED-IV dataset[46], which includes EEG and eye movement data, but with an expanded focus. In our dataset, we captured multimodal information — EEG, facial expressions, eye movement data, precise fixation points, and environmental context — as participants watched 50 videos under different emotional states.

Since our proposed model, EmoGazeNet, processes Equirectangular Projection (ERP) images from panoramic environments, the EmoGaze2D-50 dataset, which contains only 2D images, is not suitable. Therefore, we created an another new dataset, EmoGaze360-1K, specifically for EmoGazeNet. This dataset comprises 1,000 panoramic images — 800 indoor and 200 outdoor — spanning 52 categories, sourced from platforms like YouTube and Vimeo. EmoGaze360-1K also includes emotional annotations for six emotional states across various modalities, including EEG, facial expressions, eye movement, precise fixation points, and environmental context. Six distinct emotions were recorded in both genuine and deceptive conditions, following strict modality-specific standards to ensure consistency and high fidelity. The dataset is split into training and testing sets with a 70/30 ratio, supporting robust ten-fold cross-validation. This dual-context approach enables fairer comparisons across emotion recognition models and offers deeper insights into how emotions are expressed when authentic or intentionally concealed (see Supplementary "Methods - EmoGaze360-1K" section — EmoGaze360-1K and EmoGaze2D-50 datasets construction). Our experiments show that training on the entire EmoGaze360-1K dataset yields the best performance (Supplementary Figure 6).

Our 360-degree real-world indoor scene dataset. To evaluate the performance of our gaze acquisition method, we collected eye appearance data from subjects using our proposed "user-unaware" gaze tracking method across four distinct indoor environments, one public outdoor scene, and one driving scenario (called "Real360"). Utilizing sophisticated post-processing techniques mentioned before, we first predict eye gaze coordinates, then transform these eye gaze coordinates into object-level regions. By analyzing the objects and corresponding environments, we can infer emotional states with greater accuracy.

Performance comparison between proposed deep model EmoGazeNet with existing emotion recognition methods on 2D screen dataset

Most mainstream emotion recognition methods are trained on 2D videos where emotions are induced in participants. To ensure consistency and fairness in evaluating our proposed deep model, we used the comprehensive multimodal EmoGaze2D-50 dataset. This dataset, also based on 2D videos, incorporates a diverse array of data types, including EEG readings, facial expressions, eye movement data, precise visual fixation points, and contextual environmental information. This dataset is particularly distinctive as it encompasses six emotional states under both deceptive and real emotional states. We conducted a comparative analysis against several state-of-the-art methods, including electroencephalography (EEG)-based method (ACTNN[47]), facial-based method (Toisoul[48]), gaze-based method (CCER[49]), and our context gaze-based method EmoGazeNet.

Based on the data shown in Figure 4-**a**, we can clearly see the significant advantage of the context gaze-based method in distinguishing between real and deceptive emotions. The ECG-based method performed well in real emotion detection (90.74%) with an overall accuracy of 88.94%, though its accuracy in deceptive

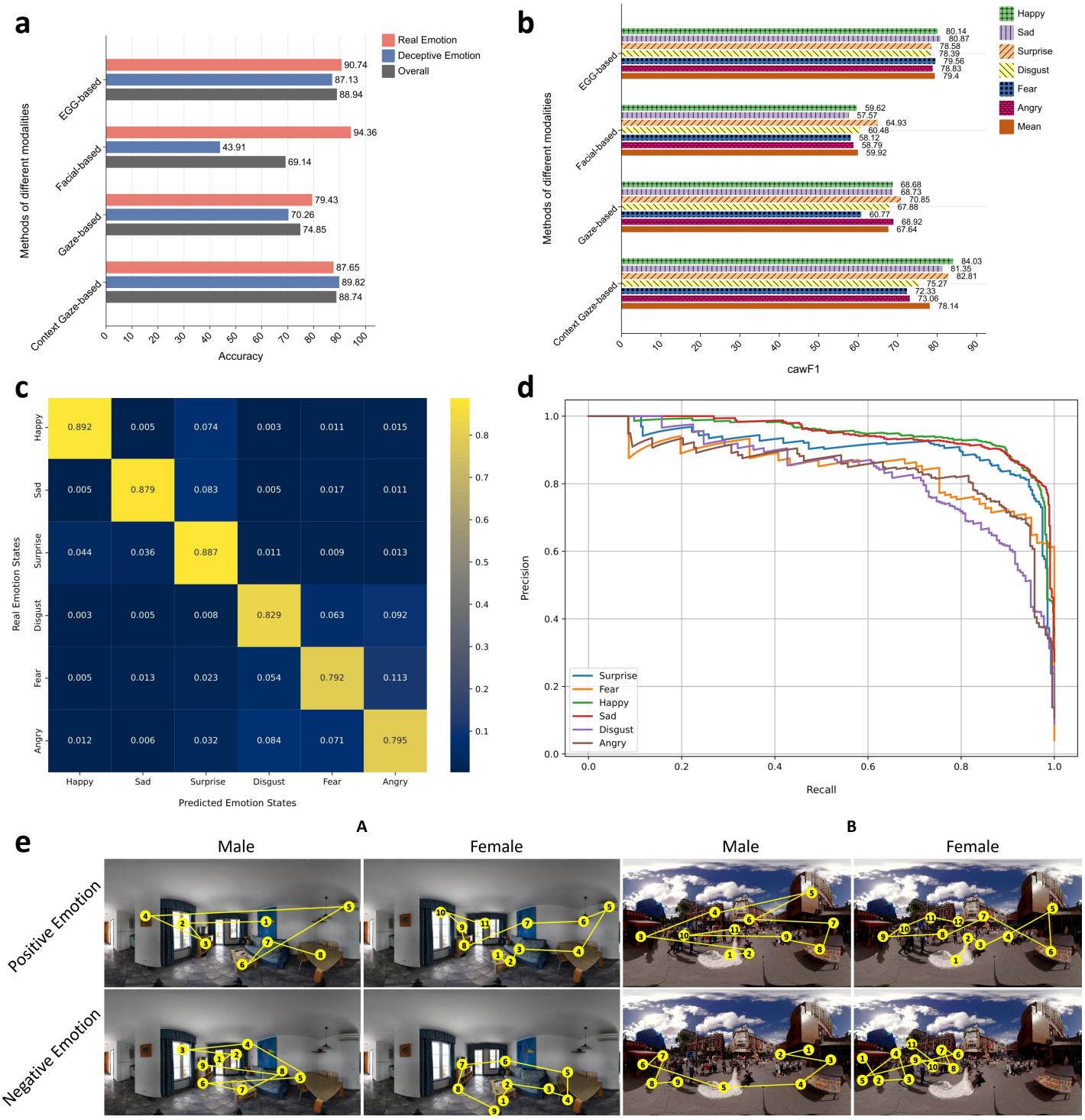


Figure 4. Experimental validation of our proposed method against other methods and its performance on screen scenes. Quantitative comparisons between our context gaze-based method and physiological signal (such as EGG)-based (ACTNN), facial expression-based (Toisoul), gaze (such eye movement)-based (CCER) methods. **a** In EmoGaze2D-50 dataset regarding Accuracy metric, when the users conceal their emotions, our approach can strike a balance between emotion recognition accuracy and user comfort. The facial expression-based method performs the worst. **b** In EmoGaze360-1K regarding our newly-proposed *caw*F1 metric, our proposed emotion recognition model, EmoGazeNet, outperforms methods that rely solely on facial and eye movement analysis. Comparatively, EmoGazeNet's overall performance is only marginally less than that of EEG-based approaches, with a slight 1.26% deficit. However, when it comes to recognizing the emotions of "Happy", "Surprise", and"Sad", our method actually surpasses EEG techniques in terms of *caw*F1 performance. **c** The confusion matrix reflects the high accuracy of emotion state prediction, with minimal misclassification across different emotional states. **d** PR curve shows that Happy and Sad have higher precision maintained even at higher recall levels, indicating that the model performs better on these two emotions compared to others. **e** Scanpath visualization of different genders. "Positive Emotion": happy, surprise; "Negative Emotion": fear, sad, disgust and angry.

emotion detection dropped slightly to 87.13%. The facial-based method achieved the highest accuracy in real emotion detection

(94.36%), but dropped sharply to 43.91% for deceptive emotions, highlighting its limitation in handling deceptive states. In comparison, the gaze-based method achieved accuracies of 79.43% and 70.26% in real and deceptive emotion detection, respectively, with an overall accuracy of 74.85%. It showed balanced performance but still lagged behind other methods. Our context gaze-based method showed significant improvement over the traditional gaze-based method, with an accuracy of 89.82% for deceptive emotions and 87.65% for real emotions, achieving an overall accuracy of 88.74%, close to the ECG-based method (88.94%). Additionally, in deceptive emotion detection, our method even surpassed the ECG-based method, demonstrating its robustness and reliability across different emotional contexts. A two-sample t-test was conducted to compare the accuracy of our method and the traditional EEG-based method, facial-based method, Gaze-based method in deceptive emotion detection. The calculated p-value was 0.031, 0.023, 0.035 ($p < 0.05$), indicating a statistically significant difference, which strongly validates the superiority of our method in this aspect.

Fine-grained performance comparison between proposed deep model EmoGazeNet with existing emotion recognition methods on 360-degree screen dataset

We also assessed our proposed emotion recognition model EmoGazeNet on the comprehensive multimodal 360-degree screen image dataset, known as EmoGaze360-1K. We reported the *caw*F1 performance of various emotion recognition methods — physiological signal (EGG)-based, facial-based, gaze (eye movement)-based, and our fixation-environment integration method — across six fine-grained emotion states. As illustrated in Figure 4-**b**, "Happy" emotion achieved the highest *caw*F1 score, with the context gaze-based and EEG-based methods scoring 84.03% and 80.14%, respectively. This high score may result from the distinct and consistent patterns associated with happiness, which makes classification easier. The "Sad" emotions followed with slightly lower *caw*F1 score across the context gaze-based method, but is still higher than physiological signal (EGG)-based methods. The "Fear" emotion showed the lowest recognition scores, especially in the facial-based (58.12%) and gaze (eye movement)-based (60.77%) methods. The context gaze-based method demonstrated the second highest overall performance, with an average *caw*F1 score of 78.14%, surpassing both the facial-based and gaze (eye movement)-based methods, and only 1.26% lower in overall performance compared to the EEG-based approach. This suggests that the context gaze-based method is highly effective in recognizing a range of emotional states. A one-way ANOVA test confirmed that the differences between EmoGazeNet and physiological signal (EGG)-based method, facial-based method, gaze (eye movement)-based method were statistically significant ($p = 0.015, 0.038, 0.027 < 0.05$).

The confusion matrix in Figure 4-**c** illustrates the classification performance of our model on screen panorama data across various emotional categories. The values on the diagonal represent the model's accuracy in correctly classifying each emotion, indicating strong performance in identifying "Happy" (0.892), "Sad" (0.879), "Surprise" (0.887), "Disgust" (0.829), "Fear" (0.792), and "Angry" (0.795). "Happy" is primarily misclassified as "Surprise" (0.074), suggesting occasional confusion between these emotions, possibly due to similar facial expressions. Similarly, "Sad" tends to be misclassified as "Surprise" (0.083), reflecting challenges in distinguishing subtle emotional expressions. For the "Surprise" emotion, the model occasionally misclassifies it as "Happy" (0.044) and "Sad" (0.036), indicating some difficulty in distinguishing between neutral and related emotions. "Disgust" is more likely to be confused with "Angry" (0.092) and "Fear" (0.063), while "Fear" is primarily misclassified as "Angry" (0.113), likely due to overlapping facial cues among these emotions. Finally, "Angry" is mainly confused with "Fear" (0.071) and "Disgust" (0.084), which is consistent with the misclassification patterns observed for "Fear".

We also provided PR (Precision-Recall) curve to evaluate the performance of a model. In a PR curve, the closer the curve is to the top right corner, the better the model's performance. As shown in Fig 4-**d**, the PR curves for Fear and Disgust show high precision at low recall levels (close to 0.2 to 0.6), but precision quickly decreases as recall increases (0.6). This suggests that the model might struggle with these two emotions. The PR curves for Surprise and Angry are more balanced, with no distinct areas of high precision, but overall, there is a good balance between precision and recall, and the curves are relatively smooth. Happy and Sad have more prominent PR curves, with higher precision maintained even at higher recall levels, indicating that the model performs better on these two emotions compared to others.

For 360-degree indoor scenes (Figure 4-**e**(A)), under positive emotions, males tend to focus first on salient objects like the artwork on the blue wall, then quickly glance at windows, and finally notice the sofa and table. Females start with details on the coffee table, gradually expanding their focus to the sofa and table, and eventually to the view outside the window or door. Under negative emotions, males usually focus on windows or doors first to seek a sense of security. Then they will assess the layout of the room and pay a little attention to some bright areas. Finally, they will fix their eyes on items like sofas and coffee tables. Females begin by focusing on dark corners, then notice details on the sofa and coffee table, with their gaze primarily staying in the dark areas. For outdoor scenes (Figure 4-e(B)), under positive emotions, males first focus on the wedding scene, then quickly scan surrounding pedestrians and buildings. Females tend to first pay attention to details like the wedding dress and the expressions of the newlyweds, before expanding their focus to the entire scene, including pedestrians and buildings. Under negative emotions, males' gaze tends to be continuous, while females' line of sight is relatively shorter. Males' gaze often focuses directly on darker buildings or crowded areas in the background, and they have a shorter line of sight. Females first notice shadowy areas or details in the crowd and pay less attention to the bright wedding dress.

Fine-grained emotion recognition based on our eye gaze collection method in real-world indoor scene

We have analyzed the performance of our proposed methodology in the domain of fine-grained emotion recognition, focusing on its *caw*F1 score across a spectrum of real-world indoor scenarios dataset Real360 that encompass diverse lighting conditions and dynamic environments. As shown in Figure 5-**a**, "Happy" has the highest *caw*F1 score in all scenarios, especially in the dynamic scenario where it reaches 80.24%, indicating robust recognition across different conditions. In contrast, "Angry" and "Fear" show relatively lower *caw*F1 score, with "Angry" in the low-light scenario being the lowest (59.85%), suggesting that lighting conditions significantly impact the recognition of these emotions. The high-light scenario has a higher overall *caw*F1 score (average 73.86%), while the low-light scenario shows lower values (average 65.96%), highlighting that sufficient lighting aids emotion recognition. Notably, the *caw*F1 score for "Surprise" in the dynamic scenario (78.53%) exceeds that in the static scenario (74.84%), which may indicate that dynamic environments facilitate better recognition of certain emotions like "Surprise", suggesting that scene dynamics contribute positively to the prediction of specific emotions. A paired t-test for the comparison between high-light and low-light scenarios showed a significant difference ($p = 0.042 < 0.05$). We have also provided scanpaths of different genders in the four real-world scenarios under different emotion states (Supplementary Figure 7 & 8).

Figure 5-**c** illustrates the performance improvements of our proposed eye gaze collection method (based on the static scene) as three key limitations are progressively resolved, i.e., (1) low-quality eye appearance images, (2) eye appearance variations in different users, and (3) limited angles of eye appearance images. The baseline version (i.e., with the three key limitations), repre-

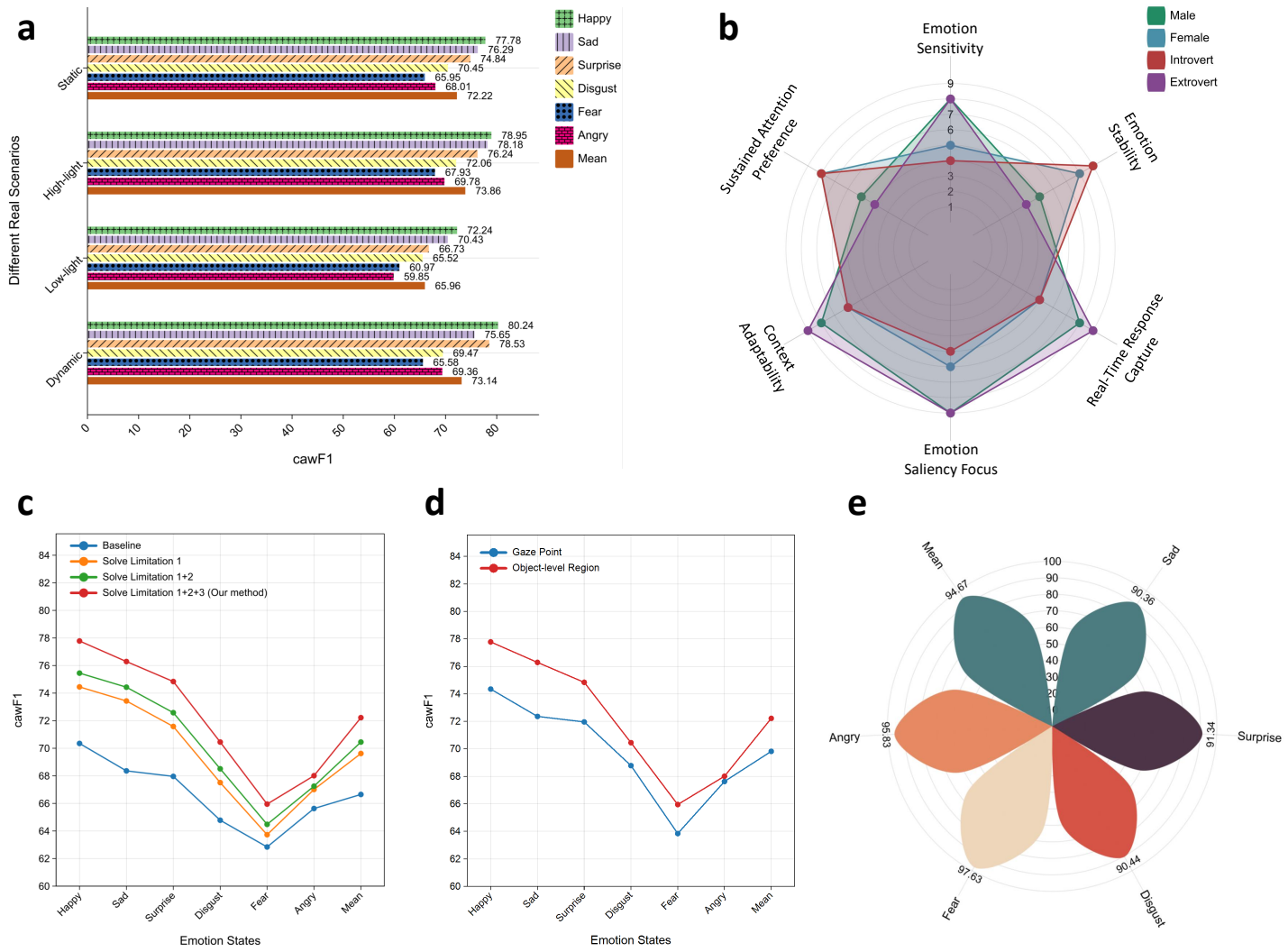


Figure 5. Experimental validation of our proposed method on different settings and its performance on real scenes. a For real scenes (four conditions), our method achieves the best performance in high-light scene and the worst performance in low-light scene. **b** This radar chart illustrates that females and extroverts, with high emotional sensitivity and adaptability, are well-suited for dynamic tasks requiring quick responses, whereas males and introverts, characterized by greater emotional stability and sustained attention, are better equipped for long-term monitoring in stable environments. **c** Utilizing an improved eye gaze collection method, we've resolved three key limitations, which has significantly enhanced the performance of the naive version. For the first limitation (limitation 1) — low quality eye appearance images — we've employed super-resolution and object-level regions to enhance clarity. To address the eye appearance variations in different users (limitation 2), we've integrated a online personality calibration process. Additionally, for the limited angles of eye appearance images (limitation 3), we developed a 3D reconstruction method to generate eye images from various perspectives. **d** Compared to segmenting the entire image based on gaze coordinates, mapping these coordinates to object-level regions and leveraging the sequence of these regions for emotion recognition yields better results. **e** Our research shows that using object-level regions for emotion recognition based on gaze coordinates is more effective than traditional methods. The average accuracy of correct correspondence between gaze coordinates and objects is over 93%, with the "Fear" emotion achieving the highest accuracy at 97.63%.

sented by the blue line, shows the lowest *caw*F1 scores across all emotions. After addressing limitation 1 (low-quality eye appearance images), shown by the yellow line, the performance improves slightly. Further enhancements are seen with the resolution of limitation 2 (eye appearance variations in different users), indicated by the green line. Finally, resolving all three limitations, including limitation 3 (limited angles of eye appearance images), results in the highest performance across all emotions, as depicted by the red line. The improved version consistently achieves better scores, particularly for "Happy" and "Sad" emotions, with a notable improvement for "Disgust" and "Angry".

We further compared two methods for eye gaze collection: gaze point-based (red line) and object-level region-based (blue line) based on the static scene. As shown in Figure 5-**d**, the object-level region method consistently outperforms the gaze point method across all emotions, particularly for "Happy", "Sad", and "Surprise" emotions, where it achieves higher *caw*F1 scores. Both methods perform the worst on "Fear", with object-level regions still providing better results. Overall, object-level region mapping shows superior performance, aligning with the study's finding that accurate gaze-to-object correspondence (i.e., the gaze coordinates are precisely within the area occupied by the object) significantly improves emotion recognition, with an average correctness above 94.67%, peaking at 97.63% for "Fear" (Figure 5-**e**). A paired-samples t-test was carried out to compare the *caw*F1 scores of the object-level region-based method and the gaze point-based method. The p-value was calculated to be 0.018 ($p < 0.05$), demonstrating a significant difference in performance between the two methods, and validating the superiority of the object-level region-based method.

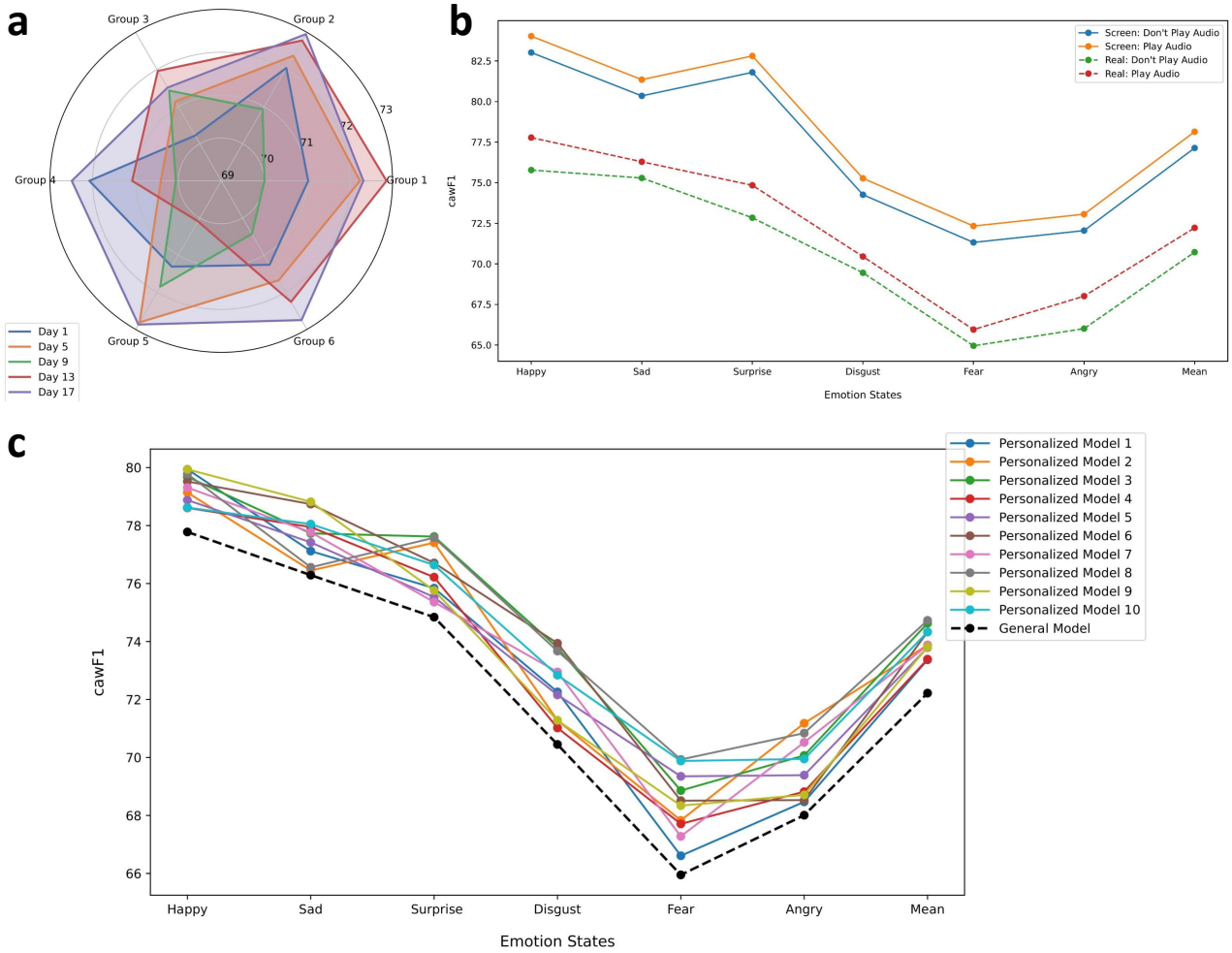


Figure 6. Robustness validation of our proposed method and component evaluation. a Long-term stability monitoring experiment indicates that during the long-term monitoring period, the proposed method can continuously provide a relatively consistent level of accuracy in emotional recognition. **b** Emotional enhancement experiment shows that playing corresponding emotional audio on emotion recognition when collecting gaze points in real scenes and when viewing 360-degree images on a screen are efficient to promote emotion recognition accuracy. **c** Personalized models have higher average accuracies than the general model, indicating that personalized training can enhance the performance of emotion recognition.

Emotion-environment interactions across various gender and personality

We also explored the differences in emotion-environment interactions across various gender and personality types in Real360 dataset. We quantified the performance of participants across six indicators — emotion sensitivity, emotion stability, real-time response capture, emotion salience focus, context adaptability, and sustained attention preference — using a 1–10 rating scale (1 being low, 10 being high). Detailed explanations of these six indicators can be found in the Supplementary "Methods". The experimental results are shown in Figure 5-**b**.

Males scored high in emotion sensitivity (8) and emotion saliency focus (9), making them responsive to emotionally rich or significant cues and well-suited for complex, dynamic environments due to high context adaptability (8). However, with lower scores in emotion stability and sustained attention preference (5), they show more frequent emotional shifts and shorter focus spans. Females, with high stability and sustained attention (8), exhibit steady emotions and prolonged focus, ideal for stable, low-dynamic contexts. Lower sensitivity and real-time response scores (5) indicate slower reactions to diverse, emotionally intense situations.

Introverts excel in emotion stability (9) and sustained attention (8), focusing well in stable, low-stimulus settings but scoring lower in sensitivity (4) and salience focus (5), making them less responsive to subtle emotional cues in dynamic environments. Extroverts score high in sensitivity (8), real-time response (9), and salience focus (9), making them quick to engage with emotional cues in varied settings. Their high adaptability (9) contrasts with lower stability and sustained attention scores (4), favoring frequent shifts in focus and making them ideal for fast-paced, interactive contexts.

This analysis shows that males and extroverts, with their strong emotional sensitivity and adaptability, are well-suited for dynamic emotion recognition tasks, while females and introverts, with greater stability and focus, are better suited for steady, long-term monitoring in single-context environments. These insights support security and driver monitoring. In security, males and extroverts' heightened emotional responsiveness and adaptability aid in identifying sudden changes in high-risk individuals, while females and introverts' stability helps in detecting abnormal behavior over longer periods. For driver monitoring, extroverts benefit from real-time alerts in complex conditions to maintain focus, whereas introverts and stable drivers, more prone to fatigue in extended sessions, can be monitored for declining attention. This approach enhances both public safety and driver security by integrating emotional and environmental interactions.

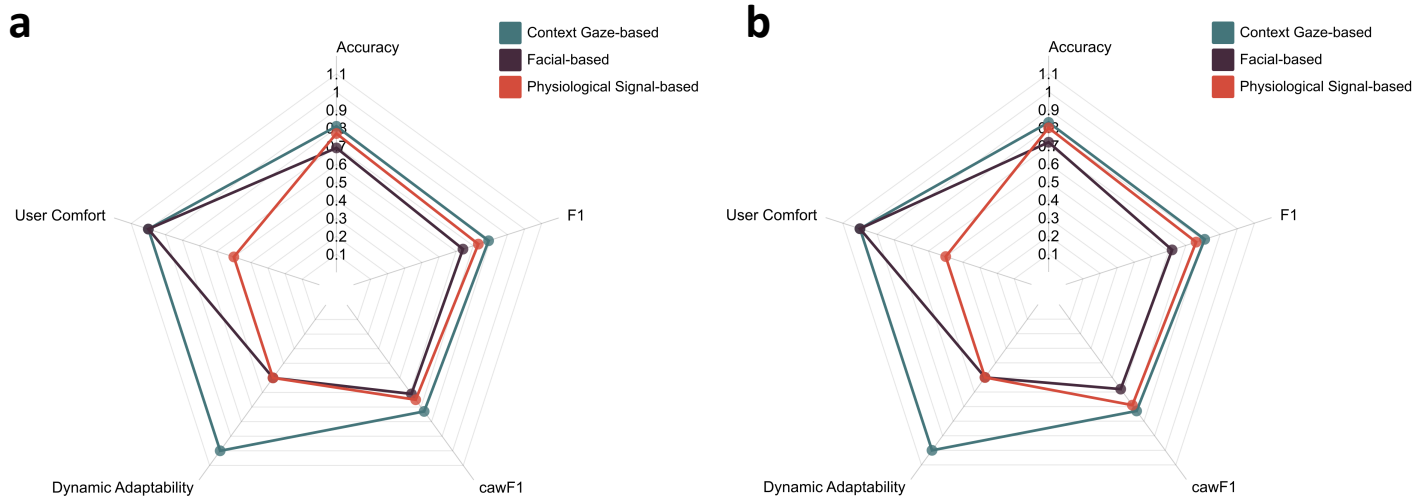


Figure 7. Field experiment to evaluate the practical application of our eye gaze collection method. **a** In the campus scenario, the context gaze-based method performed best in terms of accuracy, dynamic adaptability, and user comfort, demonstrating more stable emotion recognition compared to facial and physiological signal-based methods, while also enhancing user comfort by not requiring additional equipment. **b** In the driving simulator scenario, the context gaze-based method also showed the best performance in accuracy, F1 score, user comfort, and dynamic adaptability, while the facial-based method was significantly affected by angles and lighting in the dynamic environment, leading to the lowest performance. Dynamic Adaptability: the system's ability to flexibly adjust its behavior in response to varying contextual changes; User Comfort: how easy and comfortable it is for users to interact with the system, especially without extra devices. We normalize the results of dynamic adaptability and user comfort to the range of 0-1: dynamic adaptability (high=1, medium=0.5, low=0), user comfort (high=1, medium=0.5, low=0).

Field experiment to evaluate the practical application of our eye gaze collection method

To evaluate the practical application of the proposed eye gaze collection method, we designed two field experiment scenarios: a campus environment and a driving simulator environment. These scenarios, with their unique environmental characteristics and emotional demands, provide a comprehensive assessment of the method's robustness and effectiveness in real-world settings.

Campus Scenario Experiment Setup. The experiment was conducted in an open campus area (e.g., a campus square) to simulate a dynamic and varied social and natural environment. Fifteen university students (6 females, 9 males, aged 22-28) participated and were asked to simulate six different emotional states (such as Happy, Angry, and Surprise) induced by videos or images. During the experiment, eight high-definition cameras captured participants' eye appearance from different angles, which was then combined with gaze-environment interactions for emotion recognition. Additionally, participants wore Apple Watches to monitor skin conductance response (GSR) and had their facial expressions recorded.

Results in Figure 7-**a** showed that in the campus scenario, our context gaze-based method, which incorporates gaze-environment interactions, outperformed the other methods, achieving an accuracy of 81.45%, F1 score of 0.81, and *caw*F1 score of 0.73. This was significantly higher than the facial-based method (71.42%, 0.7, and 0.62, respectively) and the physiological signal-based method (79.62%, 0.74, and 0.65). One-way ANOVA was performed to compare the accuracy of the three methods. The p-value was 0.025 ($p < 0.05$), indicating significant differences among the methods. Post-hoc tests showed that the differences between our context gaze-based method and the facial-based method, as well as the physiological signal-based method, were statistically significant ($p < 0.05$ for both comparisons). Our context gaze-based method also scored highest in dynamic adaptability and user comfort (both 1), demonstrating its suitability for complex, open environments without requiring additional wearable equipment. In contrast, the facial-based and physiological signal-based methods showed lower dynamic adaptability and physiological signal-based methods showed lower user comfort, with facial expression recognition especially affected by lighting and angle, leading to less stable results.

Driving Simulator Scenario Experiment Setup. To simulate driving conditions and collect relevant data, we used a driving simulator that presented various traffic scenarios (e.g., emergency braking, traffic congestion). Ten drivers (7 males, 3 females, aged 20-29) participated and were tasked with handling different driving challenges and emotional stimuli. The experiment used three screens to simulate a realistic driving view, displaying front, left, and right window perspectives to replicate real driving conditions. Only one high-definition camera was used in front of the simulator to capture eye appearance, and facial expressions and GSR data were recorded simultaneously.

In the driving scenario, as shown in Figure 7-**b**, our context gaze-based method also outperformed the other methods, with an accuracy of 83.85%, F1 score of 0.81, and *caw*F1 score of 0.73. By comparison, the facial-based method achieved lower scores (72.62%, 0.62, and 0.58), while the physiological signal-based method scored slightly higher than the facial-based method (82.15%, 0.79, and 0.71). A one-way ANOVA was performed to compare the accuracy of the three methods. The calculated p-value was 0.032 ($p < 0.05$), indicating significant differences among the methods. Post-hoc tests further revealed that the differences between our context gaze-based method and the facial-based method were significant with a p-value of 0.021 ($p < 0.05$), and the difference between our method and the physiological signal-based method also reached statistical significance with a p-value of 0.045 ($p < 0.05$). In terms of dynamic adaptability and user comfort, our context gaze-based method scored the highest (both 1). While the facial-based method had high user comfort (1), it showed lower adaptability (0.5), indicating it was more affected by head movements and changes in lighting within the vehicle. The physiological signal-based method scored low in both dynamic adaptability and user comfort (0.5).

Overall, our context gaze-based method incorporating gaze-environment interactions demonstrated higher accuracy, F1 scores, and adaptability in both the campus and driving simulator scenarios, enabling more accurate emotion recognition in complex environments. Additionally, it provided superior user comfort by

avoiding the need for additional wearable devices. These results suggest that the gaze-based method is particularly advantageous for emotion recognition tasks requiring high adaptability and user comfort.

Future Research Directions in Mental Health and Security. The positive results from our experiments suggest several promising future research areas — 1) Psychological State Assessment: Our method could be used for continuous, non-intrusive monitoring of psychological states, helping track emotional shifts in real-time, especially in clinical or high-stress environments; 2) Early Psychological Disorder Screening: This approach may aid in early detection of mental health disorders like anxiety or depression, identifying subtle emotional cues before they fully develop; 3) Public Safety and Security: The method could enhance public safety by monitoring emotional reactions in real-time at events or security checkpoints, helping detect signs of distress or aggression. These directions highlight the potential of gaze-based emotion recognition in mental health and security applications.

Robustness evaluation

To prove the robustness of the proposed method, we conducted three experiments.

We conducted a long-term stability monitoring experiment designed to track the emotional recognition results of the same group of users over different time periods, e.g., from day to day in a real-world static scenario, where emotional recognition tests are conducted every six hours, twice a day, over a span of multiple days to sufficiently capture the daily changes in the user's emotional state. Specifically, the experiment design involves testing with two people per group, which may help reduce the impact of individual differences on the results. As shown in Figure 6-**a**, the data shows the fluctuation in *caw*F1 scores. The data displays the *caw*F1 score of emotional recognition from Day 1 to Day 17. The *caw*F1 score fluctuate around the 70% to 73% mark, indicating that the proposed method maintains relatively stable performance over the long-term monitoring period. Despite daily fluctuations, there is no significant overall decline or upward trend in *caw*F1 scores, suggesting that the proposed method has good long-term stability. On Day 9 in Group 1, the *caw*F1 score reached its lowest point at 70.01%, while on Day 13, it reached its highest point at 72.86%. These peaks and troughs may be related to various factors, such as changes in the user's state, environmental factors, or minor changes in test conditions. In consecutive tests, the *caw*F1 score showed only minor changes. From the 1st to 5th day in Group 3, it rose slightly from 70.22% to 71.14%. From the 13th to 17th day, it dropped marginally from 71.96% to 71.51%, indicating overall stability. Such short-term fluctuations may be due to random errors or minor changes in the user's emotional state. Looking at the overall data from Day 1 to Day 17, the *caw*F1 score seems to fluctuate around a central value of approximately 71.5% to 72%. This indicates that during the long-term monitoring period, the proposed method can continuously provide a relatively consistent level of accuracy in emotional recognition.

We implemented emotional enhancement experiment aimed to compare the impact of playing or not playing corresponding emotional audio on emotion recognition when collecting gaze points in real static scenes and when viewing 360-degree images on a screen. As shown in Figure 6-**b**, for viewing 360 images on a screen, without playing emotional audio, the average *caw*F1 score is 77.14%; with playing emotional audio, the average *caw*F1 score increases to 78.14%. For collecting gaze points in real scenes, without playing emotional audio, the average *caw*F1 score is 70.72%; with playing emotional audio, the average *caw*F1 score increases to 72.22%. Playing emotional audio had a certain enhancing effect on the recognition *caw*F1 score of most of emotions, although the extent of improvement varied.

We conducted personalized model training experiment aimed to train an individualized emotion recognition model for each user and assess the effectiveness of personalized models in enhancing emotion recognition for individual users based on the static scene in real world. The experiment also compared the performance differences between personalized and general models. Data was collected every two hours, seven times a day, for the training of the personalized model. As shown in Figure 6-**c**, personalized models show varying average *caw*F1 scores across different users (numbered 1 to 10), roughly ranging from 73.37% to 74.73%. The general model has an average *caw*F1 scores of 72.22%, serving as the baseline for comparison with the performance of personalized models. In most cases, personalized models have higher average *caw*F1 score than the general model, indicating that personalized training can enhance the performance of emotion recognition.

Component evaluation and proposal evaluation metric

We compared the impact of different visual stimuli on emotion recognition by examining 2D images, 360-degree panoramic images, and 360-degree panoramic videos on Real360 dataset As shown in Figure 8-**a**, 360 videos demonstrated superior performance across most emotional dimensions. Specifically, 360 videos achieved the highest recognition rates in "Surprise" (76.81%), "Disgust" (71.27%), "Fear" (66.33%), and "Angry" (69.06%), and also attained the highest overall mean score of 72.64%. On the other hand, 360 images excelled in conveying "Happy" (77.78%) and "Sad" (76.29%) emotions, outperforming both 360 videos and 2D images. This indicates that 360 images have a distinct advantage in eliciting both positive and negative emotions. In contrast, 2D images showed slightly lower scores across all emotional dimensions, with "Happy" at 76.62% and "Sad" at 75.19%. However, they still maintained relatively high recognition rates in "Surprise" (75.87%) and an overall mean of 71.39%.

To validate our method, we compared gaze coordinates collected by our approach and an eye tracker using 360 images on Real360 dataset. As shown in Figure 8-**b**, the eye tracker showed slightly higher accuracy (83.79%) compared to our method (80.22%), with an F1 score of 80.65% for the tracker and 78.86% for our approach. The eye tracker also had a higher *caw*F1 score (73.69% v.s. 72.22%). These results suggest that while the eye tracker is more accurate in collecting gaze data, our method performs comparably and is practical for resource-limited or rapid deployment scenarios.

Figure 8-**c** shows the performance variations when using different number of HD cameras during eye appearance acquisition. The use of 8 cameras appears to strike an optimal balance between capturing comprehensive eye appearance data and managing the computational complexity and potential data redundancy. Eight cameras provide sufficient coverage of the eye region to capture the necessary details for accurate gaze prediction without overwhelming the system with excessive data. This balance helps maintain high accuracy and F1 scores, as the model can efficiently process the captured data without being hindered by unnecessary information.

We also evaluated each component of the proposed emotion recognition model EmoGazeNet (see Supplementary Table 1). The evaluation matrix clearly demonstrates that the system reaches its peak performance in terms of ACC (80.22%), F1 Score (78.86%), and *caw*F1 (72.22%) when all key components, such as Scanpath-guided Region Generation, Primary Classification Branch, Auxiliary Classification Branch, and Scanpath-guided Classification Branch, are engaged, highlighting the synergistic impact of these elements on overall system performance. We also compared different scanpath prediction methods and different choices of base encoder in the Generator (Supplementary Figure 2 & 3).

We conducted extensive experiments on EmoGaze360-1K dataset to demonstrate that the proposed metric, *caw*F1, is more effective than existing metrics like accuracy and F1 score. As shown in Figure 9-**a**, the quantitative results reveal that *caw*F1 scores are generally lower than the other two metrics. For instance, in the "Fear" emotional state, participants' gaze tended to focus more on the safe bed, a specific area the model failed to capture accurately,

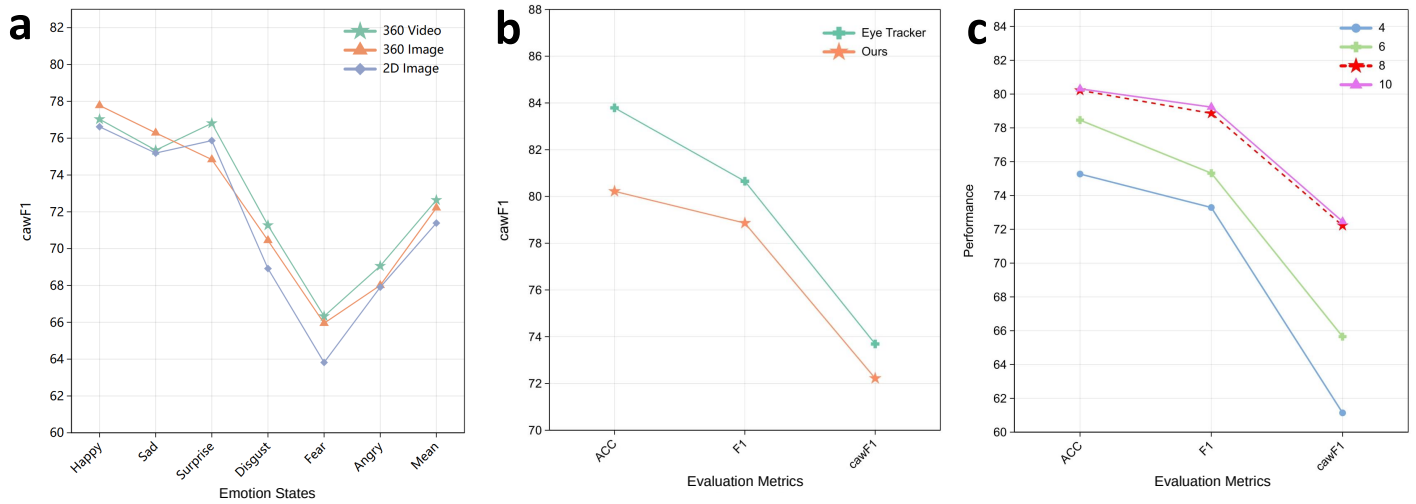


Figure 8. Ablation studies on Real360 dataset. a When collecting users' gaze points in the real-world scene of the Real360 dataset, while 360-degree videos generate better user responses than 360-degree images, the overall differences between 2D images, 360-degree images, and 360-degree videos are minimal. **b** While eye trackers offer a more precise way to capture user's gaze points, our proposed method stands out for its exceptional effectiveness in monitoring emotions in real-time across a variety of settings. **c** The performance is highest when using 8 HD cameras for eye appearance acquisition.

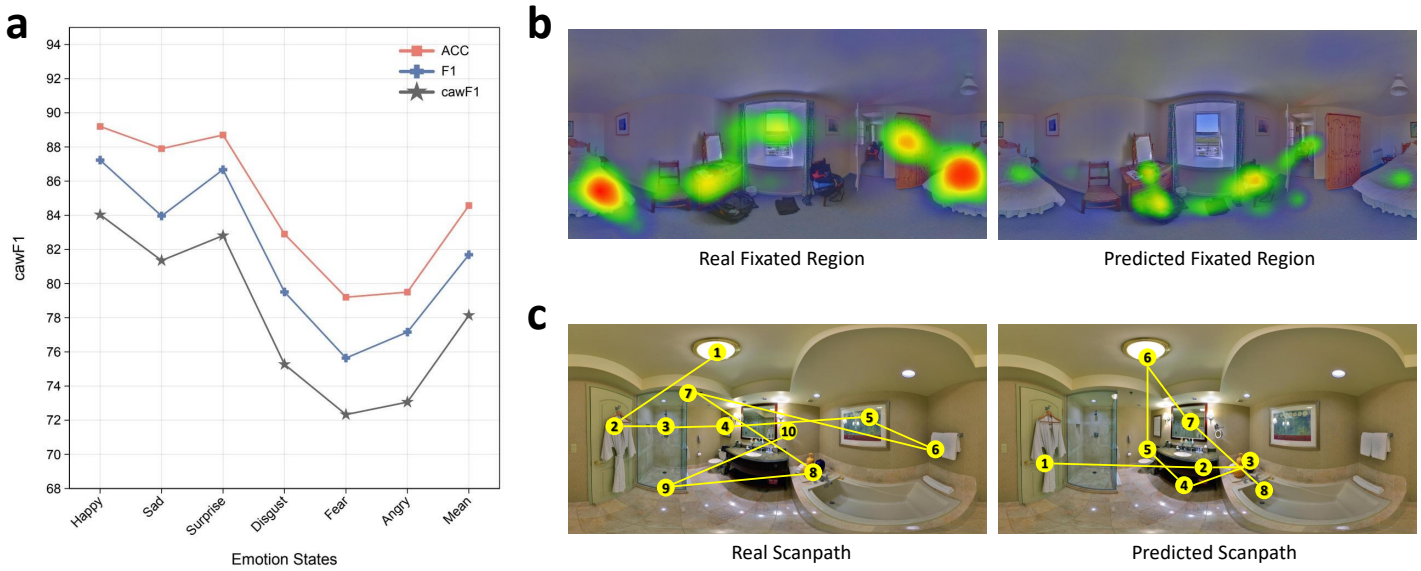


Figure 9. Experimental validation of our proposed evaluation metric *caw*F1 **on EmoGaze360-1K dataset. a** Quantitative illustration of our proposed method on EmoGaze360-1K dataset regarding Accuracy (ACC), F1 and *caw*F1 metrics. The rigor of the new metric is reflected in the fact that it requires the model to not only recognize emotions, but also accurately determine which areas of the picture participants are focusing on based on their emotional states. The low score of the new metric reflects a gap in the model's fine-grained understanding of emotions. **b** In the "Fear" state, the participant's gaze may have been more focused on the safe bed, and the model did not capture this particular area of attention well. The new metric *caw*F1 therefore scored low. **c** The model could correctly predict emotional states but failed to capture the differences in gaze scanpath, thus, the new metric *caw*F1 scored low.

leading to a lower *caw*F1 score (Figure 9-**b**). Additionally, when recording participants' gaze scanpaths while they viewed complex scene images under different emotional states, the model successfully predicted the emotional states but struggled to account for the variations in gaze scanpaths. Consequently, *caw*F1 scored lower, highlighting its stricter evaluation metric (Figure 9-**c**).

Discussion

This study proposes a novel emotion recognition method that combines eye gaze patterns with environmental context, aiming to improve the accuracy and adaptability of emotion recognition by capturing the dynamic interaction between the user's visual attention and their surroundings. Unlike traditional methods, our system does not rely on specialized equipment but instead uses standard HD cameras to collect eye gaze data, ensuring that the monitoring process is unobtrusive to users and does not require complex hardware. The "user-unaware" nature of our approach makes it highly practical for applications such as public safety monitoring and intelligent driving.

To address the challenges posed by individual differences and gaze subjectivity in emotion recognition, we introduce a "user-unaware" gaze tracking method that integrates gaze trajectories with environmental information. This method uses online personalized calibration, leveraging physiological characteristics from the gaze-environment interaction to improve accuracy and mitigate the interference caused by individual differences, enhancing both

the objectivity and precision of emotion recognition.

To validate the effectiveness of our method, we conducted a series of experiments on screen datasets. First, we tested on the 2D image dataset (EmoGaze2D-50) and found that our gaze-based context method outperformed traditional emotion recognition methods, with an accuracy of 88.74%, comparable to EEG methods (88.94%) and surpassing EEG in detecting deceptive emotions (Figure 4-**a**). In contrast, facial expression-based methods showed weaker performance in recognizing deceptive emotions, highlighting their limitations in handling complex emotional states. On the 360-degree panoramic image dataset (EmoGaze360-1K), our gaze-based context method also achieved the best *caw*F1 score in recognizing "Happy" and "Sad" emotions, surpassing both EEG and facial expression methods (Figure 4-**b**).

In real-world experiments, we further verified our method's performance under varying lighting conditions and dynamic scenes (Figure 5-**a**). In high-light conditions, emotion recognition performed well, especially for "Happy" and "Sad" emotions, with a mean *caw*F1 score of 73.86%. However, in low-light conditions, accuracy dropped significantly, particularly for "Anger", which reached a low of 59.85%. This highlights the significant impact of lighting conditions on recognition accuracy, as poor lighting degrades the quality of gaze data. Additionally, dynamic scenes performed better than static scenes, especially for recognizing "Surprise" with a *caw*F1 score of 78.53%, outperforming static scenes at 74.84%. This suggests that dynamic environments provide more emotional cues, facilitating more accurate emotion recognition.

Our method also demonstrated outstanding performance in two real-world application environments: the open campus and driving simulation environments. In the campus environment (Figure 7-**a**), the accuracy was 81.45%, with an F1 score of 0.81 and a *caw*F1 of 0.73, outperforming facial expression-based (71.42%/0.62/0.62) and physiological signal methods (79.62%/0.74/0.65). In the driving simulation (Figure 7-**b**), the gaze-based method achieved 83.85% accuracy, F1 score of 0.81, and *caw*F1 of 0.73, significantly outperforming facial expression-based (72.62%/0.62/0.58) and physiological signal methods (82.15%/0.79/0.71). In both environments, the gaze-based method scored a perfect 1 in dynamic adaptability and user comfort, demonstrating excellent environmental adaptability.

Overall, our method not only excels in laboratory environments but also demonstrates strong ecological validity in real-world applications. Traditional emotion recognition techniques are often confined to controlled settings, which limits their generalizability to more dynamic, real-world conditions such as changing lighting or fast-moving scenes. In contrast, our gaze-based context method provides a more adaptable and robust solution. By offering superior user comfort, adaptability, and accuracy, our approach outperforms conventional methods, making it well-suited for practical applications such as public safety monitoring and intelligent driving. Furthermore, by eliminating the need for specialized equipment and ensuring continuous, unobtrusive monitoring, our system offers a scalable, real-world solution that maintains high recognition accuracy while minimizing user discomfort.

Experiment results also revealed significant differences in emotion recognition performance based on gender and personality type (Figure 5-**b**). Males and extroverts tended to perform better in emotional sensitivity and adaptability, making them more suitable for dynamic environments. In contrast, Females and introverts showed stronger emotional stability and sustained attention, making them better suited for long-duration emotion monitoring in stable environments. These findings provide valuable insights for the design of personalized emotion monitoring systems.

In long-term monitoring experiments (Figure 6-**a**), our method showed stable performance, with *caw*F1 scores fluctuating between 70%-73%, indicating its suitability for continuous emotion monitoring tasks. Personalized models further improved accuracy, with *caw*F1 scores ranging from 73.37% to 74.73%, surpassing the generic model's score of 72.22% (Figure 6-**c**).

Evaluation of the system components showed that using eight HD cameras for data collection provided sufficient coverage while maintaining high accuracy and computational efficiency (*caw*F1 = 72.22%), avoiding redundant data (Figure 8-**c**). Although traditional eye-tracking devices slightly outperformed in precision (*caw*F1 = 73.69%), our method remains highly practical in resource-constrained or rapid deployment scenarios (Figure 8-**b**). Furthermore, the results demonstrated that when collecting users' gaze points in the real-world scene of the Real360 dataset, the overall differences among viewing 2D images, 360-degree images, and 360-degree videos were relatively small (Figure 8-**a**).

Despite its advantages, our approach has limitations in low-light and dynamic scenarios (Figure 5-a). For instance, in low-light conditions, the *caw*F1 score averaged 65.96%, notably lower than in high-light conditions (73.86%). Furthermore, gaze patterns vary considerably across individuals, and although the "online personalized calibration" mitigates some of this variability, the model's generalization ability could be further improved. Additionally, gaze tracking accuracy in dynamic scenes remains challenging, especially in fast-moving environments.

Future research should enhance environment modeling and gaze mapping in challenging conditions while developing dynamic adaptation mechanisms for personalized emotion recognition. This technology promises applications in psychological assessment, mental health screening, and non-intrusive public safety monitoring, offering real-time insights into emotional states across diverse contexts.

Methods

Online personalized calibration

Gaze data is inherently subjective, as individuals exhibit widely varying gaze patterns in identical situations. For example, when observing the same artwork, some may focus on the main character, while others may be drawn to background details or color contrasts. These variations stem from personal interests, preferences, and observation habits, complicating the adaptability and generalizability of emotion recognition models. Furthermore, the mapping between eye appearance and gaze coordinates varies significantly among individuals, making it challenging for traditional regression models to achieve high accuracy in gaze tracking. Given that gaze is a fine-grained external expression, even small tracking errors can significantly interfere with emotion detection, emphasizing the need for high precision.

To address these challenges, we propose an online personalized calibration method that integrates subjective fixation (user-specific gaze tendencies) and objective fixation (scene-based salient points) to enhance gaze mapping accuracy and adaptability (Figure 2-**(3)** and Figure 3-**a**). This method focuses on leveraging two key factors influencing gaze behavior: head motion and gaze state transitions. First, when the head is stationary, head movement data has minimal influence on gaze accuracy. However, during head movement initiation or cessation, visual inertia causes the gaze to align roughly with the head's direction, offering a valuable reference point. A multi-camera system captures images from multiple angles, and head pose estimation algorithms calculate pitch, yaw, and roll. By monitoring changes in head angles, the system identifies movement start and stop points. At these moments, a "strong hint" mechanism provides an initial gaze range, reducing errors caused by individual differences. Second, gaze transitions between two states: "scanning" and "fixation". In the scanning state, the gaze moves rapidly over a wide area, while in the fixation state, it focuses on a specific object. Distinguishing these states enables more precise gaze tracking. An initial gaze mapping model, combined with head pose data, estimates the approximate gaze position. For static objects, saliency detection identifies the most prominent object as the gaze coordinate, aligning with the objective fixation. For dynamic objects, motion

detection techniques like optical flow pinpoint the movement's starting point as the precise gaze coordinate. This integration of head motion and gaze states ensures robust, individualized gaze tracking across diverse scenarios.

Building on this foundation, the calibration process adapts dynamically to user behavior and environmental changes, starting with a global initialization and continuing through ongoing fine-tuning.

At system initialization (timestamp t=1), a global calibration process aligns subjective and objective fixation (Figure 3-**a**). During this phase, the system collects eye appearance data (e.g., pupil shape, gaze direction) and head movement data (pitch, yaw, roll) using a multi-camera setup. This data forms the basis for aligning the subjective and objective gaze references. To achieve this, a teacher-student model framework is used, inspired by knowledge distillation. The teacher model analyzes the scene to identify salient objects, such as static targets (e.g., cars, trees) or dynamic movement starting points, establishing an objective fixation reference. The student model, which is personalized to the user, predicts gaze points based on subjective fixation tendencies and compares them with the teacher's outputs. This comparison serves to refine the student model through continuous learning, gradually adjusting it to better align with both the scene's characteristics and the user's preferences, enhancing the system's adaptability and accuracy over time.

During significant head movements or scene transitions (e.g., yaw or pitch exceeding thresholds at timestamps t=S1+1 and t=S2+1), dynamic calibration is triggered to adjust gaze predictions. Visual inertia temporarily aligns gaze with head direction, allowing the "strong hint" mechanism to narrow the gaze range. Simultaneously, the teacher model updates salient object detection, particularly for dynamic regions, and the student model is fine-tuned by integrating motion starting points and salient targets. This process distinguishes between scanning and fixation states, offering broad gaze ranges during scanning and precise targets during fixation.

Finally, during scene transitions, the system performs online fine-tuning within the first 200–300 milliseconds — a critical window when visual attention is primarily driven by objective saliency rather than cognitive or emotional factors[50,51]. The system reacquires eye appearance data and head movement updates, while the teacher model reanalyzes salient objects in the new scene. This enables rapid calibration of the student model to reflect new scene characteristics. By aligning subjective fixation with the most prominent static or dynamic features (objective fixation), the system ensures precise gaze tracking even in complex and dynamic environments.

Third-person multi-camera panoramic modeling

Gaze often exhibits distinct "first-person" characteristics [2] in its interaction with the environment. To represent this perspective, conventional methods typically rely on wearable devices, such as head-mounted cameras. However, while these devices can effectively capture the user's field of view, they also increase the user's burden and reduce the overall user experience. To solve this challenge, we propose a "third-person multi-camera panoramic modeling" approach, using a multi-camera approach from a third-person perspective to generate a panoramic model of the scene from any location, ensuring a "user-unaware" solution (see Figure 2-**(3)** & Supplementary "Methods — Third-Person Multi-Camera Panoramic Modeling").

We adopt a problem decomposition strategy, breaking down the complex panoramic sphere generation task into three subproblems: static background fine reconstruction, local foreground object appearance generation, and foreground-background high-realism fusion. Since the static background is relatively stable, we use computationally expensive methods (such as ReconFusion) to reconstruct the base background of the panoramic sphere. For dynamic foreground objects, a "lightweight" approach is needed for high-quality generation and fusion. The specific solution consists of two parts:

First, the foreground semantic skeleton captures key information about movable objects in the scene, such as spatial coordinates, size, appearance, and semantics, using multiple complementary camera views. Due to the differences in object representation across various viewpoints, we need to achieve "common alignment" and "differential complement" of object-level information in a lightweight manner through a "weakly supervised" model. Specifically, a subspace clustering approach is used to establish initial mappings of foreground objects from different camera angles, and through self-iteration, the local structure matching is optimized to create a "sparsely structured" and "semantically rich" foreground semantic skeleton. This method significantly reduces the complexity of panoramic sphere generation and meets real-time requirements.

Second, to reduce panoramic sphere generation's computational overhead, we simplify processing by utilizing pre-reconstructed backgrounds and foreground semantic skeletons. Our approach generates target object appearances from desired viewpoints using semantic skeletons, then fuses them with backgrounds. Camera parameters and object poses from the semantic skeleton enable efficient local-to-global fusion with enhanced realism. We further optimize through "weight-sharing, alternating training", using a single model for both viewpoint generation and fusion, improving quality without additional computational costs.

The advantage of this method lies in problem decomposition, ensuring high-quality generation while reducing the demand for computational resources. By using multi-camera joint generation to create high-quality "first-person perspective" panoramic spheres, we can represent the interaction between the viewpoint and the environment in a "user-unaware" manner, laying a crucial foundation for subsequent research.

Object-box-based evaluation metric for gaze point collection accuracy

To evaluate the accuracy of the collected gaze coordinates, we propose a method based on the object's bounding box. This method assumes that every object in the scene is labeled with a bounding box, and the model's predicted gaze coordinates should fall within the bounding box of an object. We judge the accuracy of the prediction based on whether the gaze coordinates fall within the bounding box. If the coordinates fall inside the object's bounding box, the prediction is considered accurate; if they fall outside the bounding box, the prediction is considered to have a large error.

To describe this process specifically, we assume that the model's predicted gaze coordinates are (x_p, y_p), while the bounding box of the closest object is represented by the coordinates of its top-left and bottom-right corners, (x_{min}, y_{min}) and (x_{max}, y_{max}), respectively.

First, we check if the gaze coordinates satisfy the following conditions to confirm whether they fall inside the object's bounding box:

$$x_{min} \leq x_p \leq x_{max} \quad \text{and} \quad y_{min} \leq y_p \leq y_{max}, \tag{1}$$

if these conditions hold, the gaze coordinates (x_p, y_p) are within the object's bounding box, and the prediction is considered accurate.

Second, if the gaze coordinates do not satisfy the above conditions, i.e.:

$$x_p < x_{\min} \quad \text{or} \quad x_p > x_{\max} \quad \text{or} \quad y_p < y_{\min} \quad \text{or} \quad y_p > y_{\max}, \tag{2}$$

[2]"First-person" characteristics refer to the unique perspective where the user's gaze directly aligns with their view of the environment, making it challenging to capture and interpret objectively.

then the gaze coordinates (x_p, y_p) are outside the object's bounding box, and the prediction is considered to have a large error.

Third, we can define an accuracy evaluation function Accuracy, which takes a value of 1 (accurate) or 0 (inaccurate), using the following formula:

$$\mathbf{A} = \begin{cases} 1, & \text{if } x_{\min} \le x_p \le x_{\max} \quad \text{and} \quad y_{\min} \le y_p \le y_{\max}, \\ 0, & \text{otherwise .} \end{cases} \tag{3}$$

Here, $\mathbf{A} = 1$ indicates that the prediction is accurate, meaning the gaze coordinates fall within the object's bounding box; $\mathbf{A} = 0$ indicates that the prediction is inaccurate, meaning the gaze coordinates are outside the bounding box.

Proposed evaluation metric

Traditional multi-classification evaluation metrics such as precision, recall, and F1 score are usually used for basic performance measurement, but they may not be sufficient to capture the complexity of emotion recognition tasks, particularly when considering the interaction between emotional states and visual attention. These conventional metrics focus solely on the accuracy of emotion classification without taking into account the specific areas of the environment that individuals might focus on under different emotional states. As a result, models evaluated using these metrics might appear to perform well, even when they fail to accurately predict the gaze patterns or fixation points that are crucial for understanding the emotional context.

Targeting at this issue, we propose a comprehensive evaluation metric that addresses these limitations by integrating both classification performance and fixation-context consistency into a single evaluation metric — Contextual Attention Weighted F1 Score (*caw*F1). Unlike traditional metrics, *caw*F1 not only assesses the model's ability to correctly classify emotions but also evaluates how well the model can predict the areas of the environment that are most relevant to the observed emotional state. This makes the metric more rigorous and reflective of the model's true understanding of the interplay between emotion and attention. By incorporating gaze patterns into the evaluation, *caw*F1 ensures that models are held to a higher standard, where successful emotion recognition is closely tied to accurate environmental context interpretation. The metric can be defined as:

$$caw\text{F1} = \frac{\sum_{i=1}^{n} \text{FCC}_i \cdot b\text{F1}_i}{\sum_{i=1}^{n} \text{FCC}_i}, \tag{4}$$

where n is the number of samples, $b\text{F1}_i$ is the balanced F1 score for the i-th sample. FCC_i is the fixation-context consistency score for the i-th sample, used to measure the consistency of the model between the detected viewpoints and the context of the environment. FCC can be calculated by:

$$\text{FCC} = \frac{1}{n} \sum_{i=1}^{n} (\alpha \cdot \text{Sim}(v_i^{local}, e_i^{local}) + \beta \cdot \text{Sim}(v_i^{global}, e_i^{global})), \tag{5}$$

where n is the number of samples, v_i^{local} and v_i^{global} are the local and global fixation feature vectors of the i-th sample, e_i^{local} and e_i^{global} is the local and global environment context feature vectors of the i-th sample. α and β are the weight parameters which satisfy $\alpha + \beta = 1$. Sim is used to compute the cosine similarity between fixation features and environmental context features.

The features of fixation and environment context regarding local and global conditions can be extracted by pre-trained convolutional neural networks (e.g., ResNet, VGG, etc.) For local features v_i^{local} and e_i^{local}, we extract features within a certain area around the gaze point. For example, features within a fixed-size window around the point of gaze and corresponding environmental information and may be extracted. For global features v_i^{global} and e_i^{global}, we extract global features from the entire image to capture overall fixations and environmental information.

Implementation Details

The EmoGazeNet model is developed and implemented using PyTorch in Python with CUDA. Model training is performed on an NVIDIA Geforce RTX 3090 graphics processing unit (GPU). We use the Adam optimizer with the learning rate of 0.001 to train the EmoGazeNet model for 1000 epochs with batch size of 16. The complement training process takes around 17 hours. The model has 50 GFLOPs and 17.84 million parameters.

References

1. Goel, S., Jara-Ettinger, J., Ong, D. C. & Gendron, M. Face and context integration in emotion inference is limited and variable across categories and individuals. *Nat. Commun.* **15** (2024).
2. Blouin, A. M. *et al.* Human hypocretin and melanin-concentrating hormone levels are linked to emotion and social interaction. *Nat. Commun.* **4**, 1547 (2013).
3. McClay, M., Sachs, M. E. & Clewett, D. Dynamic emotional states shape the episodic structure of memory. *Nat. Commun.* **14**, 6533 (2023).
4. Zhao, S. *et al.* Affective image content analysis: Two decades review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 6729–6751 (2021).
5. Zhao, S., Jia, G., Yang, J., Ding, G. & Keutzer, K. Emotion recognition from multiple modalities: Fundamentals and methodologies. *IEEE Signal Process. Mag.* **38**, 59–73 (2021).
6. Amin, M. M., Mao, R., Cambria, E. & Schuller, B. W. A wide evaluation of chatgpt on affective computing tasks. *IEEE Trans. Affect. Comput.* (2024).
7. Zhao, S., Yao, H., Gao, Y., Ding, G. & Chua, T.-S. Predicting personalized image emotion perceptions in social networks. *IEEE Trans. Affect. Comput.* **9**, 526–540 (2018).
8. Zhang, J., Yin, Z., Chen, P. & Nichele, S. Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. *Inf. Fusion* **59**, 103–126 (2020).
9. Awais, M. *et al.* Lstm-based emotion detection using physiological signals: Iot framework for healthcare and distance learning in covid-19. *IEEE Internet Things J.* **8**, 16863–16871 (2020).
10. Vine, V., Boyd, R. L. & Pennebaker, J. W. Natural emotion vocabularies as windows on distress and well-being. *Nat. Commun.* **11**, 4525 (2020).
11. Schaare, H. L. *et al.* Associations between mental health, blood pressure and the development of hypertension. *Nat. Commun.* **14**, 1953 (2023).
12. Jiang, R. *et al.* The brain structure, inflammatory, and genetic mechanisms mediate the association between physical frailty and depression. *Nat. Commun.* **15**, 4411 (2024).
13. Zhao, S. *et al.* Curriculum cyclegan for textual sentiment domain adaptation with multiple sources. In *Proceedings of the Web Conference 2021*, 541–552 (2021).
14. Zhao, S. *et al.* An end-to-end visual-audio attention network for emotion recognition in user-generated videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 303–311 (2020).
15. Hsu, Y.-L., Wang, J.-S., Chiang, W.-C. & Hung, C.-H. Automatic ecg-based emotion recognition in music listening. *IEEE Trans. Affect. Comput.* **11**, 85–99 (2020).
16. Skaramagkas, V. *et al.* esee-d: Emotional state estimation based on eye-tracking dataset. *Brain Sci.* **13**, 589 (2023).

17. Zhao, S., Hong, X., Yang, J., Zhao, Y. & Ding, G. Toward label-efficient emotion and sentiment analysis. *Proc. IEEE* **111**, 1159–1197 (2023).

18. Pekrun, R., Vogl, E., Muis, K. R. & Sinatra, G. M. Measuring emotions during epistemic activities: the epistemically-related emotion scales. *Cogn. Emot.* **31**, 1268–1276 (2017).

19. Russell, J. A. Measures of emotion. In *The Measurement of Emotions*, 83–111 (Elsevier, 1989).

20. Nelis, S. M., Rae, G. & Liddell, C. The level of expressed emotion scale: A useful measure of expressed emotion in adolescents? *J. Adolesc.* **34**, 311–318 (2011).

21. Zhu, T., Li, L., Yang, J., Zhao, S. & Xiao, X. Multimodal emotion classification with multi-level semantic reasoning network. *IEEE Trans. Multimed.* **25**, 6868–6880 (2022).

22. Yang, J., Li, J., Li, L., Wang, X. & Gao, X. A circular-structured representation for visual emotion distribution learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4237–4246 (2021).

23. Xu, Z. & Wang, S. Emotional attention detection and correlation exploration for image emotion distribution learning. *IEEE Trans. Affect. Comput.* **14**, 357–369 (2023).

24. Pan, J. & Wang, S. Progressive visual content understanding network for image emotion classification. In *Proceedings of the 31st ACM International Conference on Multimedia*, 6034–6044 (2023).

25. Zhang, Z., Wang, L. & Yang, J. Weakly supervised video emotion detection and prediction via cross-modal temporal erasing network. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18888–18897 (2023).

26. Chen, H., Shi, H., Liu, X., Li, X. & Zhao, G. Smg: A micro-gesture dataset towards spontaneous body gestures for emotional stress state analysis. *Int. J. Comput. Vis.* **131**, 1346–1366 (2023).

27. Zou, B., Wang, Y., Zhang, X., Lyu, X. & Ma, H. Concordance between facial micro-expressions and physiological signals under emotion elicitation. *Pattern Recognit. Lett.* **164**, 200–209 (2022).

28. Wang, L., Jia, G., Jiang, N., Wu, H. & Yang, J. Ease: Robust facial expression recognition via emotion ambiguity-sensitive cooperative networks. In *Proceedings of the 30th ACM International Conference on Multimedia*, 218–227 (2022).

29. Zhai, Y. *et al.* Looking into gait for perceiving emotions via bilateral posture and movement graph convolutional networks. *IEEE Trans. Affect. Comput.* (2024).

30. Jia, G. & Yang, J. S^2-ver: Semi-supervised visual emotion recognition. In *European Conference on Computer Vision*, 493–509 (2022).

31. Song, T., Zheng, W., Song, P. & Cui, Z. Eeg emotion recognition using dynamical graph convolutional neural networks. *IEEE Trans. Affect. Comput.* **11**, 532–541 (2020).

32. Li, X. *et al.* Exploring eeg features in cross-subject emotion recognition. *Front. Neurosci.* **12** (2018).

33. Cao, M. *et al.* Virtual intracranial eeg signals reconstructed from meg with potential for epilepsy surgery. *Nat. Commun.* **13**, 994 (2022).

34. Kaveh, R., Schwendeman, C., Pu, L., Arias, A. C. & Muller, R. Wireless ear eeg to monitor drowsiness. *Nat. Commun.* **15**, 6520 (2024).

35. Tan, H. *et al.* Intracranial eeg signals disentangle multi-areal neural dynamics of vicarious pain perception. *Nat. Commun.* **15**, 5203 (2024).

36. Song, T. *et al.* Variational instance-adaptive graph for eeg emotion recognition. *IEEE Trans. Affect. Comput.* **14**, 343–356 (2023).

37. Zhang, G. *et al.* Sparsedgcnn: Recognizing emotion from multichannel eeg signals. *IEEE Trans. Affect. Comput.* **14**, 537–548 (2023).

38. Tabbaa, L. *et al.* Vreed: Virtual reality emotion recognition dataset using eye tracking & physiological measures. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, 1–20 (2021).

39. Sharma, P. *et al.* Student engagement detection using emotion analysis, eye tracking and head movement with machine learning. In *International Conference on Technology and Innovation in Learning, Teaching and Education*, 52–68 (2022).

40. Lin, J.-S. C. & Liang, H.-Y. The influence of service environments on customer emotion and service outcomes. *Manag. Serv. Qual. An Int. J.* **21**, 350–372 (2011).

41. Farshchi, M. A. & Fisher, N. Emotion and the environment: the forgotten dimension. In *Creating the Productive Workplace*, 73–92 (CRC Press, 1999).

42. Kashdan, T. B., Volkmann, J. R., Breen, W. E. & Han, S. Social anxiety and romantic relationships: The costs and benefits of negative emotion expression are context-dependent. *J. Anxiety Disord.* **21**, 475–492 (2007).

43. Richardson, M. P., Strange, B. A. & Dolan, R. J. Encoding of emotional memories depends on amygdala and hippocampus and their interactions. *Nat. Neurosci.* **7**, 278–285 (2004).

44. Ekman, P. & Friesen, W. V. Constants across cultures in the face and emotion. *J. Pers. Soc. Psychol.* **17**, 124 (1971).

45. Lian, D. *et al.* Multiview multitask gaze estimation with deep convolutional neural networks. *IEEE TNNLS* **30**, 3010–3023 (2018).

46. Zheng, W.-L. & Lu, B.-L. Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks. *IEEE Trans. Auton. Ment. Dev.* **7**, 162–175 (2015).

47. Gong, L., Li, M., Zhang, T. & Chen, W. Eeg emotion recognition using attention-based convolutional transformer neural network. *Biomed. Signal Process. Control.* **84** (2023).

48. Toisoul, A., Kossaifi, J., Bulat, A., Tzimiropoulos, G. & Pantic, M. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nat. Mach. Intell.* **3**, 42–50 (2021).

49. Gong, X., Chen, C. P. & Zhang, T. Cross-cultural emotion recognition with eeg and eye movement signals based on multiple stacked broad learning system. *IEEE Trans. Comput. Soc. Syst.* (2023).

50. Awh, E., Belopolsky, A. V. & Theeuwes, J. Top-down versus bottom-up attentional control: A failed theoretical dichotomy. *Trends Cogn. Sci.* **16**, 437–443 (2012).

51. Theeuwes, J. Top–down and bottom–up control of visual selection. *Acta Psychol* **135**, 77–99 (2010).

Competing interests

The authors declare no competing interests.