Towards stereoscopic vision

Attention-guided gaze estimation with EEG in 3D space

Qin, Dantong; Long, Yang; Zhang, Xun; Zhou, Zhibin; Jin, Yuting; Wang, Pan

DOI
10.1016/j.neucom.2025.130577

Publication date
2025

Document Version
Final published version

Published in
Neurocomputing

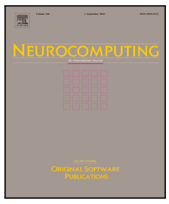
Citation (APA)
Qin, D., Long, Y., Zhang, X., Zhou, Z., Jin, Y., & Wang, P. (2025). Towards stereoscopic vision: Attention-guided gaze estimation with EEG in 3D space. *Neurocomputing*, *648*, Article 130577. https://doi.org/10.1016/j.neucom.2025.130577

Neurocomputing 648 (2025) 130577

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



Towards stereoscopic vision: Attention-guided gaze estimation with EEG in 3D space☆

Dantong Qin [a,b], Yang Long [a], Xun Zhang [c], Zhibin Zhou [b], Yuting Jin [c], Pan Wang [c,*]

[a] Durham University, Durham, DH1 3LE, United Kingdom
[b] The Hong Kong Polytechnic University, Kowloon, Hong Kong Special Administrative Region of China
[c] Delft University of Technology, Delft, 2600 AA, Netherlands

ARTICLE INFO

Communicated by Z. Cao

Keywords:
Gaze estimation
3D
Brain–computer interface
Spatial attention prediction
Virtual reality (VR)

ABSTRACT

Since traditional gaze-tracking methods rely on line-of-sight estimation, spatial attention modeling from neural activity offers an alternative perspective to gaze estimation. This paper presents a proof-of-concept study on attention-guided gaze estimation with Electroencephalography (EEG), investigating whether brain signals can be leveraged to estimate attentional focus within a controlled 3D environment. We first conducted a preliminary survey to gather public opinions, revealing a generally positive attitude towards EEG-driven gaze tracking. Building on this insight, we collected an EEG dataset in VR, where participants engaged with stimuli presented at predefined spatial locations. We introduce a deep learning model that estimates the relative saliency of candidate positions, enabling gaze estimation through optimization within the learned representation. Our results demonstrate that attentional focus was successfully mapped in a 3D coordinate space from 5 participants, and low-frequency oscillations contributed more significantly to predictive performance. The model achieved robust accuracy in distinguishing gaze locations, highlighting the potential of EEG-based gaze estimation for attention tracking in 3D environments.

1. Introduction

Understanding human behavior [1–5] is crucial for human–computer interaction (HCI), where gaze estimation [6–8] serves as a key modality, enabling applications in assistive technology, virtual reality, autonomous systems, and behavioral biometrics [9–14]. Traditional gaze estimation methods primarily rely on appearance-based approaches, which predict gaze direction by analyzing eye region images [15,16], or geometry-based models, which estimate gaze positions through eye pose and head orientation calibration [4,17–19]. While these methods have achieved high accuracy in 2D settings, they face limitations in "depth perception and robustness" in 3D environments.

In three-dimensional spaces, gaze estimation is further complicated by visual attention dynamics, where gaze direction alone does not always correspond to the point of cognitive focus. Eye-tracking methods infer gaze depth through vergence-based calculations [20,21] or scene geometry assumptions [17–19], which are effective in controlled environments but may not generalize well to more complex or real-world scenarios [22–24]. Additionally, gaze estimation is influenced by physiological and demographic factors, including anatomical variations that affect gaze calibration [25,26]. These challenges raise the question of whether alternative modalities beyond eye movement tracking could contribute to understanding gaze behavior in 3D space.

One possible direction is leveraging neural activity to infer gaze positions, as the brain encodes both overt and covert attention mechanisms (see Fig. 1). Neural-based approaches, such as EEG, have the potential to complement existing gaze tracking methods by offering insights into cognitive attention states that are not always reflected in eye movements. Additionally, previous research has shown that EEG oscillations are correlated with visuospatial attention [27,28], and neural responses vary with stimulus distance [29–31]. These findings suggest that EEG signals may carry useful information for estimating gaze locations, particularly in 3D settings where eye-tracking methods face limitations.

Recent advances in brain decoding research have demonstrated the potential of neural signals in visual cognition, particularly for object recognition [32–35] and stimulus reconstruction [36–39]. However, most existing studies focus on semantic decoding from 2D screen-based stimuli, with far less emphasis on spatial perception and depth-related

☆ The study protocol was approved by the PolyU Institutional Review Board for research involving human participants, under reference number HSEARS20220906006.

* Corresponding author.
E-mail address: P.Wang-2@tudelft.nl (P. Wang).

https://doi.org/10.1016/j.neucom.2025.130577
Received 31 January 2025; Received in revised form 11 May 2025; Accepted 21 May 2025
Available online 6 June 2025

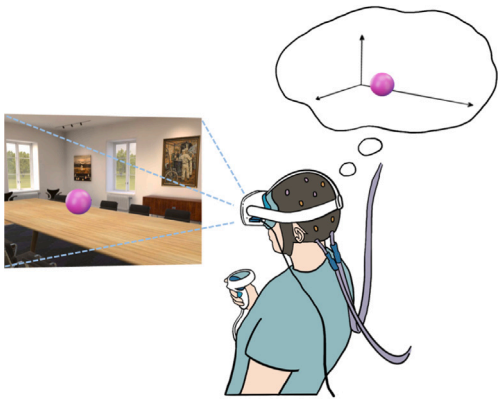


Fig. 1. Conceptual overview of our approach using brain–computer interface (BCI) technology for intuitive 3D gaze estimation, offering an alternative to eye-tracking methods by decoding visual attention directly from brain signals in a virtual environment.

attention mechanisms. While EEG-based methods have been applied to 2D gaze classification, direct mapping from EEG to 3D gaze remains largely unexplored.

This work presents a proof-of-concept investigation into whether EEG signals can be used for attention-guided gaze estimation in 3D space. We propose a model that decodes visuospatial attention from neural activity, estimating the most salient point of regard (PoR) among a set of discrete, predefined spatial targets within a 3D environment. Although the model outputs lie in continuous space, the stimuli in this study were restricted to four fixed positions. This setting enables structured decoding of spatial attention and provides a basis for future work toward denser or continuous spatial layouts. Our model consists of an EEG encoder and a position-aware regression module. The Transformer-based [40] EEG encoder captures global dependencies across unordered EEG channels, offering an advantage over conventional CNN-based feature extraction. Given the low spatial resolution of EEG, we avoid generative models that require dense training data and instead employ a lightweight MLP-based regressor to assign saliency scores to specific spatial locations. This approach prioritizes task-relevant neural features while maintaining computational efficiency. During inference, gaze predictions are refined iteratively by adjusting an initial candidate point based on the learned neural attention representation, allowing the model to converge toward the most likely PoR in the given spatial environment.

Since this task differs from traditional 2D gaze estimation, no established datasets exist for direct benchmarking. To address this, we conducted a controlled EEG experiment in virtual reality, where participants engaged with stereoscopic stimuli positioned at four spatial orientations.

The results demonstrate that our method successfully differentiate spatial locations based on EEG activity, achieving an average classification accuracy of 80.1% on five participants, and a gaze localization within a 3D coordinate space. To further understand the neural basis of the models predictions, we also conducted a series of analyses to examine the spatial and temporal characteristics of EEG responses across gaze conditions, including ERP dynamics and channel-wise model attribution. By leveraging neural signals for spatial attention decoding, our approach offers a complementary perspective on gaze estimation, with the potential to enhance multimodal integration with existing visual tracking systems or support alternative solutions for individuals with specific needs [41].

Based on this study, we highlight three key contributions:

1. We demonstrate the feasibility of EEG-based gaze localization in 3D space, extending neural decoding beyond conventional 2D paradigms.
2. We introduce a Transformer-based model with iterative refinement to estimate the most salient point of regard (PoR) in predefined spatial locations.
3. We contribute a EEG dataset collected in a VR environment with stereoscopic stimuli, enabling future research on neural gaze decoding and attention modeling in immersive settings.

2. Related work

2.1. Exploration of stereo vision in BCIs

Estimating gaze in stereoscopic environments remains an open challenge, with most prior studies focusing on neural responses to depth perception rather than explicit gaze prediction. While neuroscience research has explored how the brain processes stereoscopic depth cues using EEG and fMRI, the application of deep learning to directly predict gaze within 3D space is largely unexplored.

Depth Representation in the Visual Cortex. Early fMRI studies primarily modeled visual attention in the cortex using two-dimensional (2D) visual field mappings, such as population receptive field (pRF) estimations [42]. Subsequent studies have investigated how neural signals encode depth-related information. EEG and fMRI research has shown that the brain exhibits distinct responses to depth variations, with spatial attention modulating position selectivity in the visual cortex [43,44], and classification methods, such as SVM, have been employed to differentiate between depth levels in visual stimuli [31, 45].

EEG-based Analysis of 3D Stimuli. Beyond cortical mapping, researchers have investigated how EEG signals respond to depth-rich stimuli. Research using 2D vs. 3D stimuli has shown that depth cues influence cognitive load and neural activity [29,30]. SSVEP-based EEG studies have further demonstrated that stimuli at different distances elicit distinguishable neural patterns, even when presented at the same visual angle [46]. More recently, VR-based studies have introduced 3D stimuli scattered across different positions to investigate anticipatory potentials related to target selection [47]. Instead of performing gaze estimation, these studies explored anticipatory potentials associated with target selection, providing insights into EEG-based interaction models for 3D environments.

2.2. EEG deep learning frameworks

Deep learning-based EEG signal decoding has garnered significant attention due to its ability to extract complex patterns from EEG signals. Among various approaches, convolutional neural networks (CNNs) have been widely used for their effectiveness in learning spatial features. EEGNet [48], a compact CNN-based model, demonstrates strong generalization across multiple BCI paradigms. Similarly, EEG-Inception [49], built on the InceptionTime network, enhances temporal feature extraction for classification tasks. To address CNNs' limitation of learning only local features due to constrained convolutional kernel sizes, recurrent neural networks (RNNs) [50] have been introduced as an alternative. Specifically, LSTM-based models have proven effective in capturing long-range dependencies in EEG signals, outperforming conventional CNNs in decoding performance [51–53]. More recently, transformer-based models, leveraging attention mechanisms, have gained traction in various domains [4,54–56]. Their ability to model global dependencies without recurrent structures makes them particularly effective for sequential data. Given these advantages, transformers have been successfully applied to EEG decoding [57–59], yielding promising results.

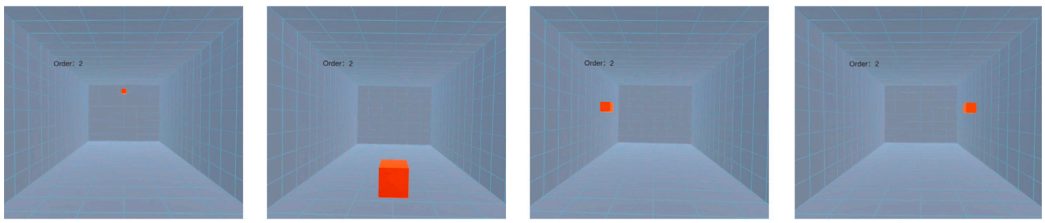


Fig. 2. Virtual environment(VE) and the presentation of stimuli. (a) Location 1: rear center, upper height. (b) Location 2: front center, lower height. (c) Location 3: center left, medium height. (d) Location 4: center right, medium height.

3. Preliminary survey

To assess public acceptance of BCI-based approach, we conducted a survey with 128 valid responses. The participants, aged 18 to 60, included 47 females and 81 males from 30 countries, all with at least a high school education.

Despite more than half of respondents being unfamiliar with BCI technology, the overall perception was positive, with 65% viewing BCIs as an advanced tool for understanding brain activity. About 57% found BCIs intuitive and straightforward, appreciating their ability to directly extract user intent from brain signals. When asked to compare BCIs and eye trackers, 71.1% considered BCIs more comfortable and user-friendly, noting that having a camera directly in front of their face could be distracting. Furthermore, a majority of respondents believed that BCI technology could be used for gaze tracking, considering it either a complementary option (71.1%) or a potential replacement (64.1%) for eye trackers. On the other hand, 7% expressed concerns about the ethical implications of BCIs.

Regarding potential applications, we asked participants to select the three areas where they see the most promise for BCI-based gaze estimation techniques. The top choices were “Game & Immersive Experience”, “Rehabilitation”, and “Clinical Medicine”, with about three-quarters of respondents voting for these fields. “Robotics & Automation” was also favored by 58% of the participants.

4. Experiments and data preprocessing

4.1. Experiment design

We employed virtual reality (VR) technology for stimulus presentation. The Oculus Quest 2 headset (Meta Platforms) was used to display 3D stimuli at four predefined spatial coordinates: Location 1 (rear center, upper height), Location 2 (front center, lower height), Location 3 (center left, medium height), and Location 4 (center right, medium height), as shown in Fig. 2. To ensure efficient neural response capture, we adopted the rapid serial visual presentation (RSVP) paradigm, a widely used approach in EEG-based visual tasks. RSVP enables rapid stimulus presentation while minimizing prolonged cognitive engagement and emotional processing, making it particularly effective for capturing visual information within brief time frames [60–62]. As illustrated in Fig. 3, each session lasts 180 s and begins with a brief 0.5-s period for participants to focus. Following this, stimuli are randomly presented at one of the four predefined locations for 1 s, followed by a 0.5-s blank interval (depicted as an empty virtual room) to mitigate residual visual effects. The stimulus onset interval is 1.5 s, ensuring equal presentation probability across all locations. Each participant completes 50 sessions, with 120 trials per session.

To enhance 3D perception within the virtual environment, we applied shading and texture principles as outlined in [63]. The stimulus consists of an orange cube, providing strong contrast against a gray cement room with subtle grid lines in the background. This design choice optimizes depth perception, aiding participants in discerning spatial relationships within the virtual environment.

4.2. Subjects

Six participants were recruited for the study; however, data from one individual, identified as stereoblind or BCI blind, was excluded from the analysis. The final dataset consisted of five participants (two females, three males), aged 21 to 29 years, all of whom were right-handed with either normal or corrected-to-normal vision. Participants were fully briefed on the study’s purpose and provided written informed consent before the experiment. To ensure optimal cognitive performance, all participants were required to be well-rested and mentally prepared prior to the session.

Each participant’s VR headset was carefully adjusted to optimize face-fitting position and interpupillary distance, ensuring both comfort and proper focus. During the experiment, participants were instructed to remain stationary, minimize body and head movements, and rely solely on eye movements to track the presented stimuli.

4.3. Data preprocessing

As shown in Fig. 4, participants were seated in a soundproof room, equipped with an EEG cap and VR headset. EEG signals were continuously recorded using a 128-channel Quik-Cap at a 1 kHz sampling rate, following the international 10–20 electrode placement system. To ensure high-quality signal acquisition, electrode impedance was maintained below 10 kΩ, with most electrodes achieving impedance levels below 0.1 kΩ.

Epochs containing muscle artifacts or excessive noise were excluded, resulting in a final dataset of 27,960 EEG trials, each corresponding to one of the four directional stimuli. The retained signals were band-pass filtered between 1–30 Hz, and average referencing was applied. To minimize signal loss, independent component analysis (ICA) was performed to remove eye movement artifacts, while extensive data cleaning was avoided to preserve neural activity. This process identified 18 to 35 ICA components, with ocular and cardiac artifacts removed based on empirical analysis and scalp topographies. An example of ICA-based artifact removal is shown in Fig. 5, where ICA000 was identified and removed.

To account for potential lingering effects in stereoscopic perception, decoding was performed across 1500 time points, spanning from stimulus onset (0 ms) to 1500 ms, covering both the stimulus display period and the subsequent break period. Ultimately, the interval [500,1500]ms was selected for training purposes. Details can be found in Section 6.4. Additionally, we conducted further analyses to identify peak neural activity during 3D perception and determine the EEG frequency bands most relevant to 3D cognitive processing, as discussed in Section 6.4.

5. Methodology

This section details the methods employed to model spatial attention distribution and estimate the PoR. The overview framework of the model is shown in Fig. 6.

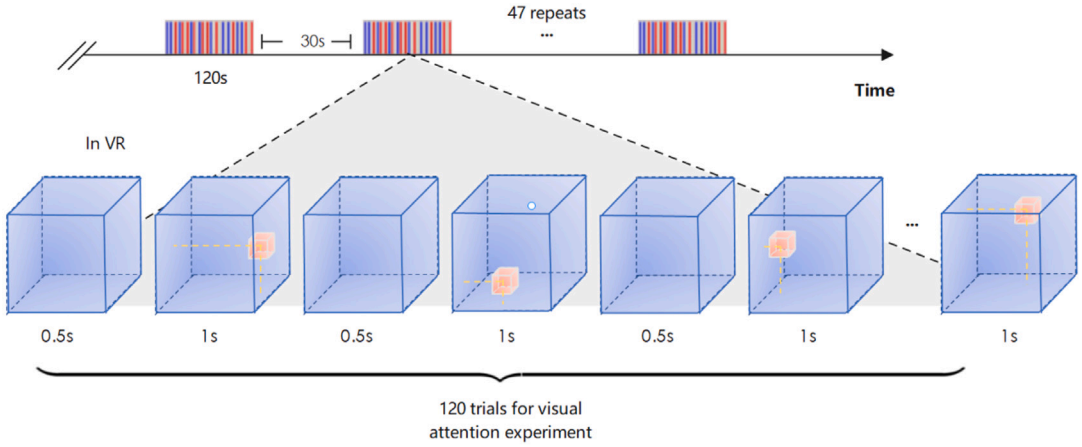


Fig. 3. A schematic view of our experiment design. The yellow dotted line is just for showing the location. It will not appear in the VE.

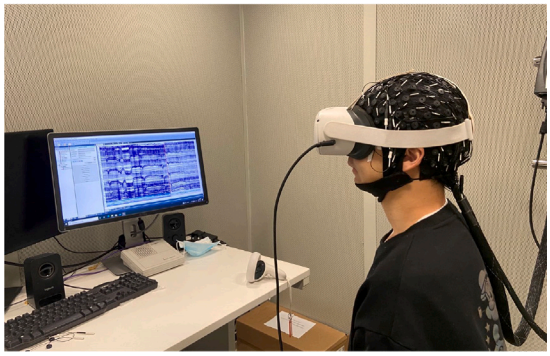


Fig. 4. The EEG collection environment. The EEG data was collected in a soundproof room. Subjects wore a cap fitted with 128 electrodes and put on a VR headset at the same time.

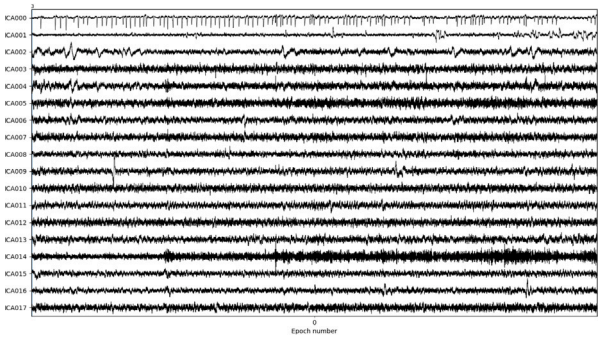


Fig. 5. 18 components ICA result of a sampled segmented EEG. The first component containing artifacts was removed in the preprocessing.

5.1. Encoder architecture

The EEG encoder transforms bioelectrical signals into a compact feature representation that encodes visual information. It consists of three key components: a temporal processing block, a spatial interaction block, and a Transformer-based feature extractor.

temporal block. Given the high temporal resolution of EEG, this block processes signals along the time axis, reducing dimensionality to L_t through a sequence of fully connected layers.

Spatial Block. This block captures cross-channel dependencies while maintaining channel-wise integrity. we employ order-agnostic fully connected layers to learn interactions among EEG channels, preserving the original input dimensionality C.

Transformer Feature Extractor. To enhance feature refinement, we leverage self-attention mechanisms from the Transformer architecture [40]. The feature $x \in \mathbb{R}^{C\times L_t}$, processed by the temporal and spatial blocks, is passed through two Transformer layers. In the attention layer, each channel $X = \{x_i \in \mathbb{R}^{1\times L_t} | i = 1, \ldots, C\}$ generates query q_i, key k_i, and a value v_i using shared weight matrices:

$$q_i = x_i W^q, \quad k_i = x_i W^k, \quad v_i = x_i W^v \tag{1}$$

The attention weights are computed via scaled dot-product attention:

$$\hat{a}_{i,j} = Softmax(q_i \cdot k_j^T / \sqrt{d}), \tag{2}$$

where d is the key dimension, and $\hat{a}_{i,j}$ represents the attention score between EEG channels i and j. The output is computed as a weighted sum of values:

$$b_i = \sum_{j=1}^{N} \hat{a}_{i,j} \times v_j. \tag{3}$$

The multi-head attention mechanism captures diverse activation patterns, and the outputs are concatenated before passing through a point-wise feed-forward network to apply independent transformations across all feature dimensions. This self-attention framework effectively models interchannel dependencies, crucial for extracting features from brain signals.

To establish that EEG signals encode visual spatial information, we first train the encoder with a classification objective, assessing its ability to distinguish between different stimuli locations. The trained encoder is then utilized as a pretrained feature extractor, facilitating a more efficient training process when integrated with the decoder.

5.2. EEG insights into 3D visual localization

Rather than directly regressing stimulus coordinates from EEG features, we incorporate spatial priors during training to guide the model in learning attention distributions. The following sections detail this approach.

5.2.1. Definition of the space

The space $\mathbb{S}$ refers to the visual space experienced by the subject in VR. It is defined with dimensions of $30 \times 23.82 \times 60.5$, where length, width, and height are given in arbitrary units, as absolute measurements hold no physical significance in the virtual environment. For spatial reference, we define the front of the space as the plane formed by the X- and Y-axes in Fig. 7, with the origin O positioned at the left-front-bottom corner of the space. The user's viewpoint is denoted by an eye icon, indicated with a yellow dashed line in the figure. The coordinates are transformed from Unity's world coordinate

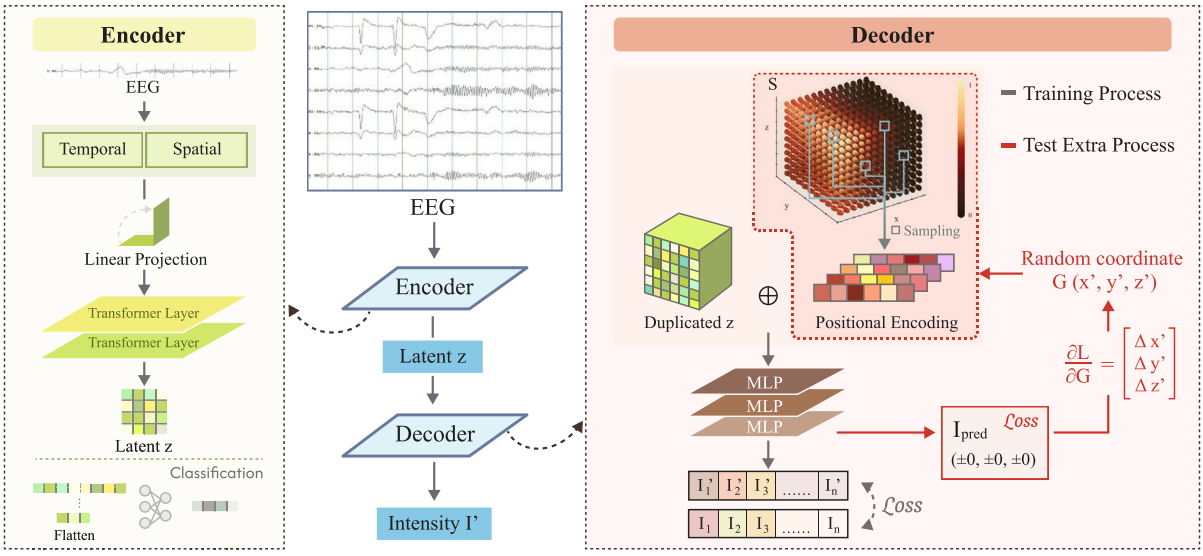


Fig. 6. The model architecture. After EEG feature extraction in the encoder, the data is transformed into a latent space representation, denoted as latent z. In the decoder, Z is then fused with the positional encoding of points sampled from the space. Subsequently, the fused information is passed through an MLP to predict the attention strength at that point. The red line represents the test routine. In the test, the data with new randomly sampled points are sent to the well-trained model. The visual focus of the data was identified using gradient descent optimization.

Table 1
The detailed location information in our experiment. World coordinate is the coordinates in unity software during modeling, and converted coordinate is converted to a field space with O as the origin, whose length, width, and height are 30, 23.82, and 60.5 respectively. Walls refer to 6 boundary surfaces in space. Location 1, 2, 3, and 4 correspond to the locations in Fig. 7.

Stimulus location	World coordinate	Converted coordinate	Distances to the walls					
			Front	Back	Top	Bottom	Left	Right
User Viewpoint	(0, −1, −19.5)	(15, 9.91, 0)	0	60.5	13.91	9.91	15	15
Location 1	(0, 10, 40)	(15, 20.91, 59.5)	59.5	1	2.91	20.91	15	15
Location 2	(0, −4, −10)	(15, 6.91, 9.5)	9.5	51	16.91	6.91	15	15
Location 3	(−10, 2, 10)	(5, 12.91, 29.5)	29.5	31	10.91	12.91	5	25
Location 4	(10, 2, 10)	(25, 12.91, 29.5)	29.5	31	10.91	12.91	25	5

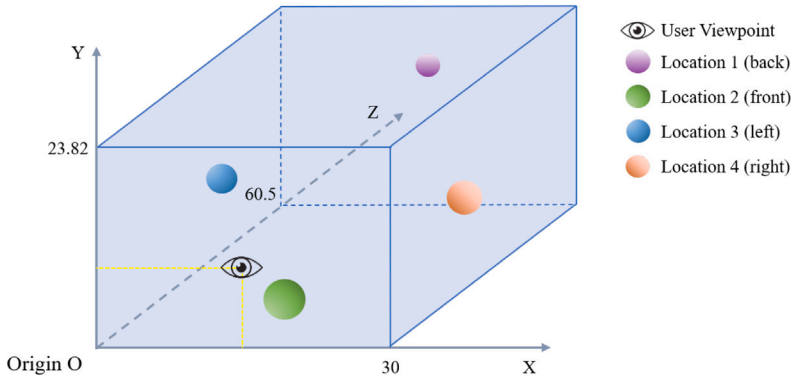


Fig. 7. The demonstration of the space. The sphere indicates the relative positions of the four stimulus locations in space. Note that although the four positions are simply referred to as front, back, left, and right in the illustration, there are differences in height.

system to align with O as the origin. The stimulus cube, which has a side length of 2, is simplified to a point representation, with its coordinates corresponding to its center in our model. The 3D positions and labels of all stimuli are summarized in Table 1.

5.2.2. Decoding 3D visual attention

To predict attention intensity, we incorporate positional encoding during decoding, embedding EEG features within a spatial context. Given an expanded feature with batch size $x \in \mathbb{R}^{I \times C \times L_t}$, we randomly sample n points from the spatial domain S to enhance spatial representation. Since raw (x, y, z) coordinates may not provide sufficient expressiveness for learning fine-grained spatial variations, we follow the approach in [64], projecting these points into a higher-dimensional space using a set of sinusoidal functions:

$$\gamma(p) = (\sin(2^0 \pi p), \cos(2^0 \pi p), \ldots, \sin(2^{L-1} \pi p), \cos(2^{L-1} \pi p)). \quad (4)$$

We set $L = 10$, encoding each coordinate as a 60-dimensional vector, which is concatenated with its original value, yielding a 63-dimensional representation per point. These positional encodings, akin to a queries, are then concatenated with the duplicated feature x_d, which acts like keys, to form the position-encoded EEG feature $\hat{x}_d \in \mathbb{R}^{n \times C \times L_t} \oplus code$. This combination can be seen as setting up a call to the decoder, where the enriched feature representation allows the model to compute the attention for these points. This formulation enhances the model's

ability to resolve spatial structure in EEG data, enabling denser spatial sampling to mitigate the constraints of limited EEG data collection.

We employ a 4-layer MLP decoder to predict attention intensity. Instead of generating a saliency map, we use a direct regression approach, mapping position-encoded EEG features to attention values. The attention intensity is defined as $I = (\pm p_x, \pm p_y, \pm p_z)$, where each component represents the distance from the stimulus along the respective axis. The sign indicates direction, with positive values denoting greater displacement from the stimulus in that direction.

Given that EEG data was recorded under four distinct visual conditions, where attention was directed to one of four stimulus locations, the resulting attention distributions vary across conditions. A single spatial point may exhibit different intensity values depending on the condition. We model attention as an isotropic field centered at the stimulus, where intensity decays linearly with increasing distance. While this model does not directly mimic biological attention mechanisms, it captures the abstract dynamics of attention distribution.

5.3. Gaze point prediction

5.3.1. Test method regression

Following the training process, the model learns to represent a user's spatial attention distribution from EEG data. At any point within S, the model predicts the relative intensity and direction of attention as inferred from EEG signals. To determine the visual focus for each test EEG sample, the task is to identify the spatial coordinates with the highest saliency value. For each input EEG x, we initialize a randomly selected point G with coordinates (x', y', z'). With the model weights frozen, we iteratively refine G through gradient descent to locate the point where relative attention displacement is minimized. The optimization objective is defined as:

$$G' = \arg\min_{G} \|\mathscr{M}_{frozen}(x, G) - I_{tgt}\|_2, \tag{5}$$

where $\mathscr{M}_{frozen}$ denotes the trained model with fixed weights. Here, $I_{tgt} = (\pm 0, \pm 0, \pm 0)$ represents the reference point in the attention field, indicating no relative displacement from the stimulus location. Through iterative optimization, G gradually moves toward the most likely visual focus, ultimately converging to G', the predicted PoR. The full testing procedure is detailed in Algorithm 1.

5.3.2. Test method bruteforce

As an alternative to the regression-based approach, we also implemented a brute-force testing method that directly predicts field strength for a large number of points, identifying the gaze point as the one closest to $(\pm 0, \pm 0, \pm 0)$. For each test instance, we sampled 10,000 points within the 3D space and computed the attention predictions. We then analyzed the results by examining the top 0.2%, 1%, and 5% of points that were closest to the ground truth (GT). The proximity of these points to the GT serves as an indicator of the model's ability to accurately capture the attention field. This method complements the regression-based testing, especially since regression may not always converge to the point of minimum loss and does not allow for visualization of the attention distribution.

6. Implementation and results

In this section, the environmental settings, predicted outcomes, analysis for feature classification, and evaluation of training are discussed sequentially.

Algorithm 1 Estimating Predicted Stimulus Location from EEG

Freeze model weights for EEGEncoder and MLP.
$\mathbf{f} \leftarrow$ EEGEncoder(eeg) ▷ Extract EEG features
for each $\mathbf{e} \in$ EEG signals **do**
  Randomly initialize $\mathscr{G} = (x', y', z')$
  $\mathbf{p} \leftarrow$ PositionalEncoding($\mathscr{G}$)
  $\mathbf{c} \leftarrow$ Combine($\mathbf{f}, \mathbf{p}$)
  for $t = 1$ **to** T **do** ▷ Optimization iterations
    $I \leftarrow$ MLP($\mathbf{c}$) ▷ Predict intensity
    $\mathscr{L} \leftarrow$ MSE($I, (\pm 0, \pm 0, \pm 0)$) ▷ Loss function
    Compute gradient:

$$\frac{\partial \mathscr{L}}{\partial \mathscr{G}} = \begin{bmatrix} \triangle\, x' \\ \triangle\, y' \\ \triangle\, z' \end{bmatrix}$$

    Update $\mathscr{G}$
  end for
end for

6.1. Implementation details

The model was implemented in Python 3.8 using Anaconda, with training performed on 4 V100 GPUs. PyTorch 1.12 was used as the deep learning framework. For data preprocessing, we utilized EEGLAB in Matlab and the MNE Python package. Our dataset, consisting of over 27,600 shuffled samples, was split into 80% for training and 20% for testing, with standardization applied along the channels. The EEG encoder consists of a temporal block with 3 layers, a spatial block with 2 layers, and a Transformer module with 6 attention heads and 3 Transformer blocks. We trained the model using a batch size of 2000. A dropout rate of 0.5 was applied in the fully connected layers, except for the final layer. Both the encoder and decoder were trained for up to 2000 epochs, with early stopping if the loss did not improve for 50 epochs. The Adam optimizer was used with a learning rate of 0.0002, $\beta 1 = 0.9$, and $\beta 2 = 0.999$.

6.2. Evaluation metric

To evaluate the accuracy of the predicted gaze points, we used the L_2 distance to measure the spatial error between the predicted and ground truth points. This metric effectively penalizes larger deviations, providing a clear assessment of prediction accuracy in 3D space. Additionally, we used the dispersion metric σ_p to evaluate the spread of the predicted gaze points. Both metrics are defined as:

$$\|L\|_2 = \sqrt{\sum_{i=1}^{n} (P_{GT_i} - P_{pred_i})^2} \quad \text{and} \quad \sigma_p = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (P_{pred_i} - \bar{P}_{pred})^2} \tag{6}$$

where $P_{GT\,i}$ and $P_{pred\,i}$ represent the coordinates of the ground truth and predicted gaze points, respectively, with all dimensions scaled to the range [−1, 1].

6.3. Results

6.3.1. Test method regression

Table 2 presents the average L_2 distances for gaze point predictions across the four stimulus locations for each of the five subjects. We found that prediction errors at Location 1 were slightly higher, especially for subjects 4 and 5. Participant feedback suggests that this may be due to the stimulus at Location 1, positioned at the upper rear in the VR setup, appearing more distant with blurred edges, making intuitive position judgment more difficult. Conversely, the average errors at Locations 3 and 4 were smaller, possibly because stimuli at Locations 1 and 2 were directly in the line of sight, showing only two sides of the cube

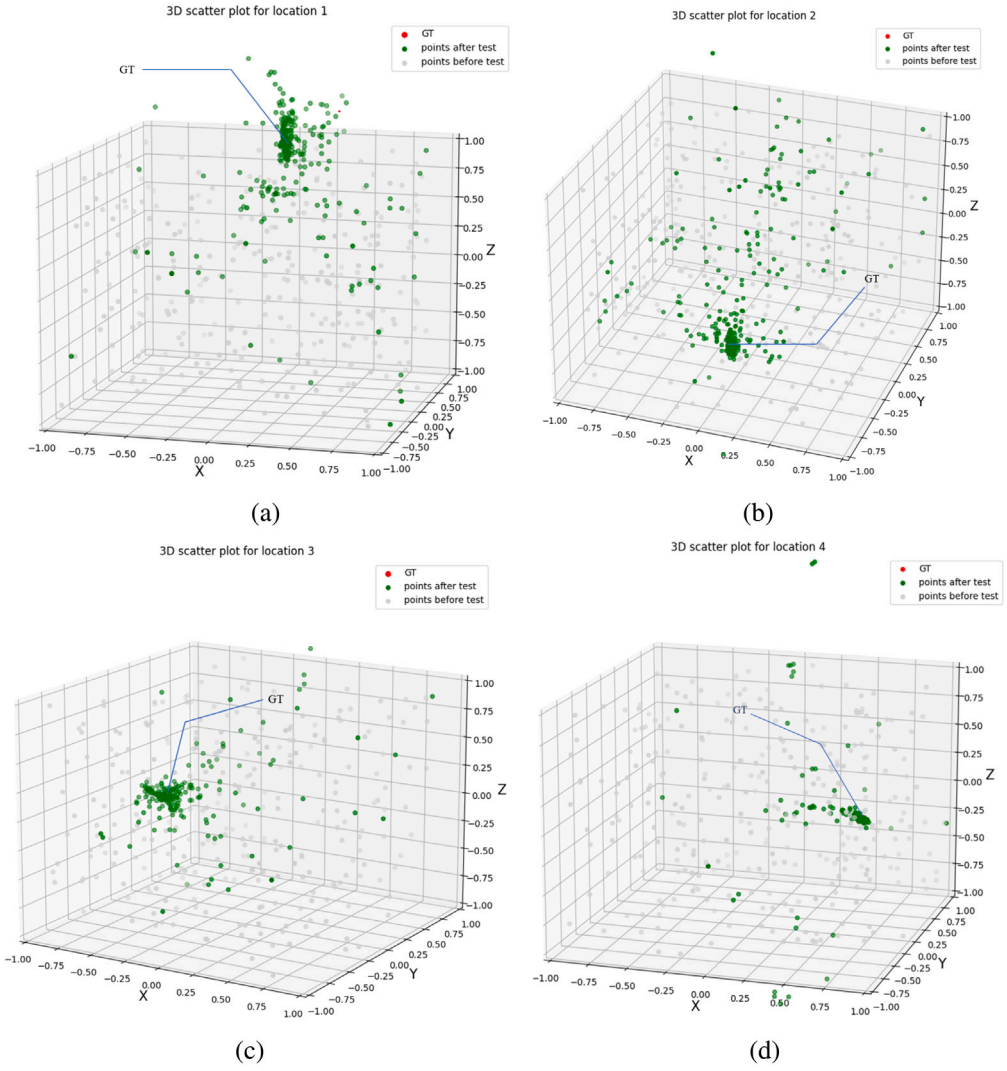


Fig. 8. Visualization of the 3D gaze prediction results for subject 3. (a) Location 1. (b) Location 2. (c) Location 3. (d) Location 4. The gray dots are randomized points, the green dots represent the positions to which the randomized points moved after testing, and the red dots represent the ground truth stimulus locations.

Table 2
The error distances $\|L\|_2$. The values are obtained by calculating the average of all predicted values for each location across individuals.

Subject ID	L_2 Distance $\downarrow$ $_{+\sigma_p}$			
	Location 1 (0, 10, 40)	Location 2 (0, −4, −10)	Location 3 (−10, 2, 10)	Location 4 (10, 2, 10)
Subject1	$0.39_{+0.48}$	$0.46_{+0.42}$	$0.272_{+0.39}$	$0.398_{+0.4}$
Subject2	$0.474_{+0.45}$	$0.448_{+0.4}$	$0.366_{+0.41}$	$0.264_{+0.37}$
Subject3	$0.376_{+0.46}$	$0.352_{+0.41}$	$0.224_{+0.39}$	$0.208_{+0.37}$
Subject4	$0.6_{+0.39}$	$0.484_{+0.4}$	$0.372_{+0.38}$	$0.268_{+0.28}$
Subject5	$0.771_{+0.33}$	$0.664_{+0.36}$	$0.481_{+0.37}$	$0.485_{+0.27}$
Avg.	0.520	0.485	0.343	0.325

and complicating accurate position judgment. In addition to the L_2 distances, the dispersion metric σ_p was also analyzed to evaluate the consistency of predictions. When L_2 distances are smaller, a lower σ_p indicates precise predictions. On the other hand, small σ_p values with greater L_2, as seen for subject 5, may indicate an overall shift from the ground truth.

Fig. 8 visualize the results from subject 3. Gray points, representing random pre-test positions, are initially scattered throughout the field. During testing, these points iteratively gravitate towards the actual stimulus locations (red points) and converge at the green points. Most points effectively moved to the GT or nearby regions, forming highly concentrated clusters near the PoR. Meanwhile, there are notable individual differences. Fig. 9 shows additional results for other subjects. The first row presents subject 2, whose results, while not as precise as those of subject 3, with several points outside the GT area, still demonstrate acceptable accuracy. The second row shows results for subject 5, who had the least favorable outcome, with points more dispersed after movement. This variability in model performance across subjects likely reflects individual brain functional connectivity differences, leading to unique neural signal patterns.

This method allows for the visualization of all test data results with the same GT on a single plot. Overall, for all subjects, the predictions are clearly concentrated near the GT. However, we observed some

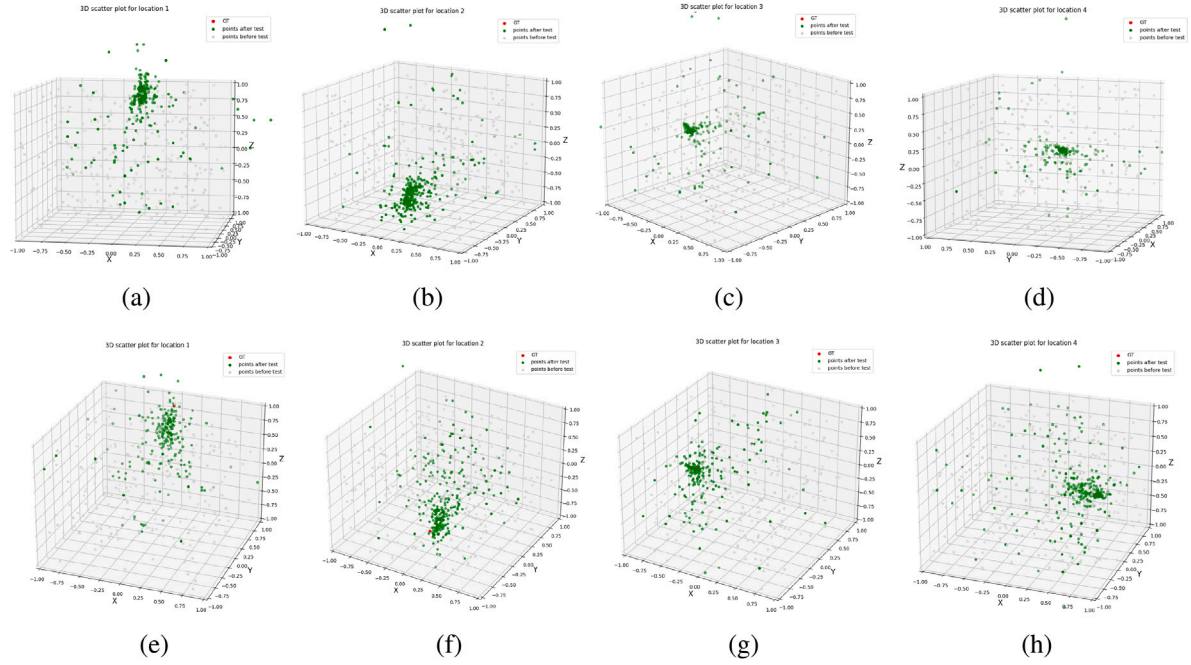


Fig. 9. Visualization of the 3D gaze prediction results for more subjects. The first row presents the results for subject 2, while the second row shows those for subject 5, whose predictions exhibited the worst performance on our model. The figures arranged from left to right, correspond to Locations 1, 2, 3, and 4, respectively.

Table 3
The error distances $\|L\|_2$ of top 0.2%, 1%, and 5% of points at the four stimulus locations under test method bruteforce. The values are obtained by calculating the average of all predicted values for each location across individuals.

Subject ID	L_2 Distance ↓											
	0.2% predicted closest points				1% predicted closest points				5% predicted closest points			
	L1	L2	L3	L4	L1	L2	L3	L4	L1	L2	L3	L4
Subject1	0.227	0.235	0.217	0.212	0.305	0.293	0.284	0.276	0.484	0.419	0.413	0.407
Subject2	0.402	0.377	0.261	0.247	0.449	0.417	0.315	0.303	0.576	0.515	0.432	0.424
Subject3	0.298	0.251	0.186	0.171	0.357	0.304	0.249	0.234	0.511	0.427	0.379	0.367
Subject4	0.506	0.323	0.347	0.291	0.538	0.374	0.396	0.340	0.631	0.488	0.500	0.450
Subject5	0.708	0.534	0.453	0.483	0.732	0.568	0.487	0.510	0.794	0.652	0.567	0.583
Avg.	**0.4282**	**0.344**	**0.293**	**0.281**	0.476	0.391	0.346	0.333	0.599	0.500	0.458	0.446

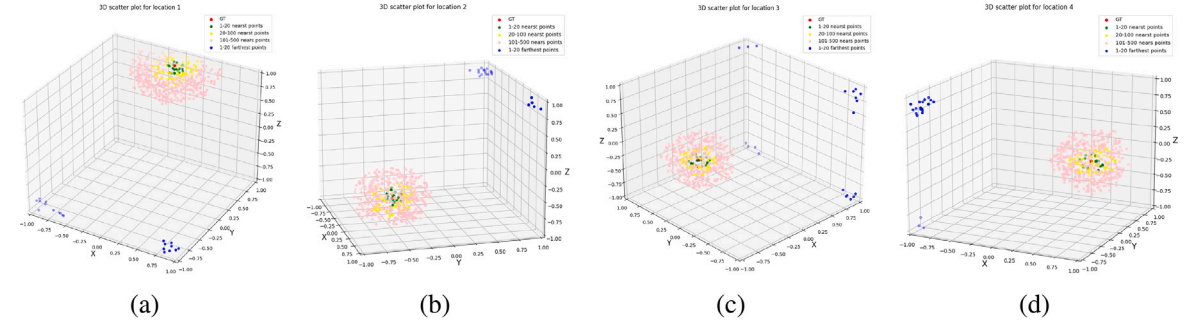


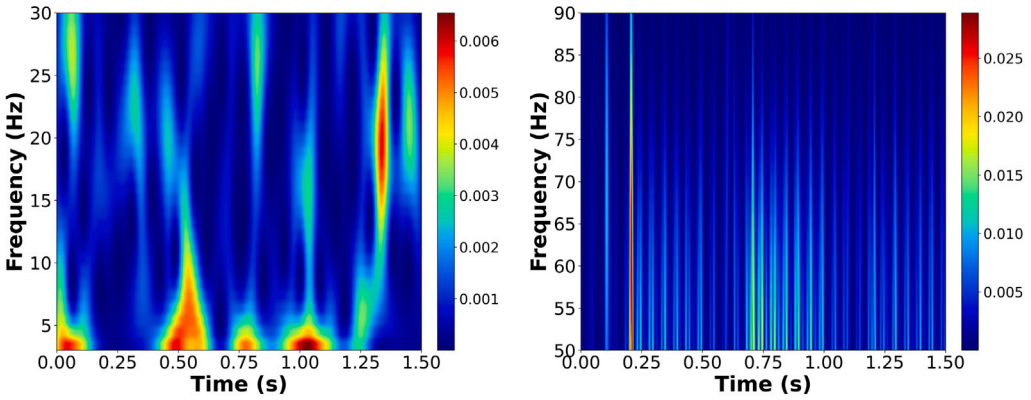
Fig. 10. The visualization of gaze localization using "Bruteforce". The green dots represent the top 20 points with the highest predicted intensity, closely surrounding the Ground Truth (GT). Points with predicted intensities in the ranges of 21–100 (yellow) and 101–500 (pink) also extend uniformly around the GT. The blue dots signify the 20 points with the lowest prediction intensity.

outliers forming a long tail. These outliers may be due to the model's prediction inaccuracies and insufficient movement of points that were initially sampled too far from the GT.

6.3.2. Test method bruteforce

As discussed in Section 5.3, the 'Bruteforce' method serves as a supplementary approach to the 'Regression' method. It predicts attention intensity across a large number of points for each test sample, allowing us to: (1) identify the point of highest intensity as the predicted gaze focus, and (2) evaluate the model's representation of the attention distribution. For each test data, we sampled 10,000 points and analyzed the top 20, 100, and 500 points, corresponding to the top 0.2%, 1%, and 5% in terms of predicted intensity.

Table 3 shows the average L_2 distance of these points from the ground truth (GT). The error of the top 20 points is roughly considered as the prediction error. The results align closely with those in Table 2, showing similar and slightly lower errors for Locations 2, 3, and 4 compared to Location 1, and poorer performance for subject 5. Notably,



(a) 3-30 Hz power changes (b) 50-90 Hz power changes

Fig. 11. Time–frequency representation of EEG activity averaged across five subjects and all 122 channels. Morlet wavelet transform was applied using 28 frequencies from 3 to 30 Hz and 25 from 50 to 90 Hz, with 1–5 and 1–3 cycles respectively. Power was z-scored within subject and then averaged across trials and subjects. No baseline correction was applied.

Table 4
Classification results on frequency bands.

Band	Frequency-Hz	Accuracy
Delta, theta, alpha and beta	**1–30**	**0.805**
Low frequency gamma	30–50	0.318
High frequency gamma	50–100	0.303
Full gamma	30–100	0.286
All frequencies	1–100	0.763

Table 5
Classification results on time blocks.

Time block-ms	Accuracy
0–250	0.423
0–500	0.581
250–750	0.642
500–750	0.647
500–1000	0.717
0–1000	0.729
750–1250	**0.750**
1000–1500	0.733
500–1500	**0.805**

the average L_2 distance for all locations in the 'Bruteforce' method is lower than in 'Regression', alleviating the issue of insufficient point movement distance.

Additionally, when analyzing the top 100 and 500 points, we observed that the error increases as prediction intensity decreases, indicating that the model effectively learns the overall attention distribution. Fig. 10 visualizes the positions of these points for several examples. The top 20, 100, and 500 points form progressively larger clusters around the GT, enveloping it from the inside out. The figure also includes blue points representing the lowest 0.2% in intensity, which are dispersed at the farthest corners of the space.

6.4. Feature classification

To identify the most informative neural signals involved in stereoscopic vision, we conducted a series of classification experiments across different EEG frequency bands and temporal segments. Given the four-category classification task, the chance level was 0.25.

Frequency bands. Since signals above 100 Hz rarely penetrate the skull, we evaluated classification rates across several frequency bands: Delta (1–4 Hz), Theta (4–8 Hz), Alpha (8–12 Hz), Beta (12–30 Hz), and Gamma (30–100 Hz), following guidelines from a previous study on visual decoding with EEG [65]. Frequencies at 50 Hz and 100 Hz were filtered to remove power line noise. The results are presented in Table 4. We found the highest accuracy in the 1–30 Hz range, with diminished performance in the Gamma band. To further understand why the 1–30 Hz range yielded the highest score, we conducted a time–frequency analysis of neural activity during the first 1500 ms after stimulus onset in Fig. 11. Combining our results with existing literature, we observed distinct power modulations in the Delta, Theta, and Alpha bands, suggesting that multiple cognitive processes could be engaged in this task. Prior research has established strong links between these frequency bands and various perceptual and cognitive operations [66–68]. Specifically, Delta and Theta oscillations are associated with working memory. Delta activity has been linked to sustained attention and the processing of perceptual cues [69], while Theta rhythms are widely implicated in spatial navigation and cognitive control [70,71]. Additionally, Alpha oscillations contribute to visuospatial attention by modulating cortical excitability and suppressing task-irrelevant information [72–75]. In contrast, the 30–100 Hz Gamma band exhibited minimal power changes throughout the trial, which may explain its weaker classification performance. Gamma oscillations are typically associated with sustained cognitive engagement and higher-order processing [76,77], which may be less prominent given the brief nature of our stimulus presentation. These findings provide further insight into the role of low-frequency oscillations in spatial perception.

Temporal segments. Temporal decoding was performed over 1500 time points (0–1500 ms post-stimulus onset), encompassing both the stimulus presentation and the subsequent break period. The classification results for different time intervals are presented in Table 5. Accuracy increased with longer time windows, rising from 250 ms to 500 ms and further to 1000 ms, indicating progressive information accumulation over time. Lower accuracy in the 0–250 ms window suggests a weaker initial neural response to the stimulus, while a steady increase in classification performance — from 0.581 in the 0–500 ms range to a peak of 0.750 in the 750–1250 ms range — suggests enhanced cognitive processing that extends beyond the stimulus display period.

Quantitative evaluation. We compare our approach to the work by Montenegro and Argyriou [78] and Kastrati et al. [79], both of which focus on EEG-based gaze classification. The results are shown in Table 6. All works significantly surpass chance levels, showcasing the effectiveness of EEG-based gaze estimation. While our approach achieves a higher F1 score but lower accuracy, the difference in task objectives and contexts makes a direct numerical comparison only partially indicative of overall performance. Montenegro and Argyriou's work highlights 9-class 2D gaze decoding on a 2D planer, while Kastrati et al. emphasize

Table 6
EEG classification comparison with relevant works.

		Task	Accuracy	Precision	Recall	F1	Chance
Montenegro and Argyriou [78]	qA_KNN	2D gaze classification	–	0.7901	0.7407	0.7400	$\frac{1}{9}$
	qA_RF		–	0.7031	0.6790	0.6698	
	qB_SVM		–	0.6821	0.6667	0.6474	
	qB_Adab		–	0.5679	0.5556	0.5426	
Kastrati et al. [79]		2D Left–Right classification	0.981	–	–	–	$\frac{1}{2}$
Ours		3D gaze estimation	0.8052	0.8096	0.8070	0.8078	$\frac{1}{4}$

Table 7
Performance comparison of different models used as encoders. We evaluate each model on both EEG classification and gaze estimation tasks. GFLOPs and Params are computed based on the gaze prediction network.

Encoder	Classification Accuracy (%)	Regression (L_2)				GFLOPs	Params
		L1	L2	L3	L4		
Logistic regression	66.3	–				–	–
SVM	76.5	–				–	–
Random Forest	76.1	–				–	–
EEGNet [48]	79.3	0.54	0.50	0.36	0.36	0.07	0.08 MB
EEGWaveNet [80]	74.6	0.82	0.76	0.50	0.49	0.01	0.14 MB
Ours-w/o transformer	77.2	0.67	0.60	0.46	0.47	0.09	4.71 MB
Ours	**80.1**	**0.52**	**0.49**	**0.34**	**0.33**	0.11	5.06 MB

Table 8
Evaluation on different numbers of attention heads.

# Attention head →	**2**	**4**	**6**
Acc.	79.8	78.7	80.1

Table 9
Evaluation on different numbers of transformer blocks.

# Transformer blocks →	**1**	**3**	**5**
Acc.	79.4	80.1	79.3

hardware optimization through electrode clustering and preprocessing techniques. In contrast, our goal is to guide the model in learning EEG representations through classification, tailored for spatial decoding in virtual environments.

6.5. Ablation studies

To evaluate the impact of different encoder architectures, we conducted an ablation study by replacing our Transformer-based encoder with alternative models. The results, presented in Table 7, demonstrate the impact of different encoders on both classification and gaze estimation performance. We first replace our encoder with EEGNet [48] and EEGWaveNet [80], two CNN-based architectures commonly used for EEG signal processing. EEGNet is a lightweight model optimized for efficient spatial and temporal feature extraction, while EEGWaveNet leverages dilated causal convolutions to capture hierarchical temporal dependencies. Despite its compact design, EEGNet achieves a classification accuracy comparable to our full model (79.3% vs. 80.1%), suggesting that a well-designed CNN can effectively extract meaningful EEG representations. However, our approach yields more consistent improvements across metrics.

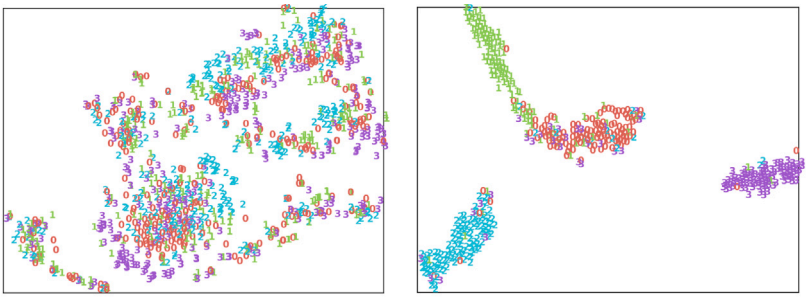
To further analyze the contribution of our Transformer module, we remove it while retaining only the spatial and temporal components. This variant, denoted as Ours-w/o Transformer, results in a slight drop in classification accuracy and gaze estimation performance, reinforcing the importance of long-range dependency modeling. Sensitivity analysis on key Transformer hyperparameters shown in Tables 8 and 9 indicates that a shallow Transformer architecture is sufficient to achieve strong classification performance. Additionally, we investigate three traditional machine learning models: Logistic Regression, SVM, and Random Forest. While SVM and Random Forest demonstrate strong classification performance, their lack of hierarchical feature learning limits their applicability to more complex tasks such as gaze estimation.

6.6. Pattern analysis

Latent Feature Structure Visualization. In Fig. 12, we visualize the EEG feature representations learned by our full model using t-distributed stochastic neighbor embedding (t-SNE). The unprocessed EEG data exhibits an unstructured distribution, whereas after training, the extracted features form well-separated clusters, indicating that the model effectively captures discriminative patterns in the EEG signals.

Location-Specific Neural Response Patterns. We examine whether distinct gaze positions elicit differentiable neural activity patterns by analyzing EEG responses from a representative subject (Subject 2). Fig. 13 shows scalp ERP topographies across four gaze conditions (L1–L4) and four key time points. The time points 0.0 s and 1.0 s correspond to stimulus onset and offset, respectively. At 0.5 s and 1.0 s, widespread activation is present in posterior regions. Additionally, the L3 and L4 conditions exhibit lateralized polarity pattern, which may relate to the brain's response to spatially lateralized stimuli. To further illustrate the temporal evolution of these responses, Fig. 14 shows stimulus-locked ERP waveforms (0–1500 ms) in Oz and POz channels. Each gaze position elicited a distinct temporal profile, including differences in early (0–500 ms) and late (>600 ms) components. Although the model was trained on data from 500 to 1500 ms, we present the full post-stimulus window here. Polarity differences across conditions were also apparent. These structured temporal variations likely serve as informative features for the model's classification. While such patterns may also support extension to continuous-space decoding, doing so would likely require the model to generalize from discrete patterns to a continuous embedding of spatial attention. Such generalization would further require access to denser and more diverse spatial sampling for robust learning.

Model-Based Channel Attribution. To better understand the spatial contribution of EEG channels to the model performance, we performed a gradient-based saliency analysis on the trained classification models. Specifically, we computed the absolute gradient of the classification loss with respect to each input EEG channel and averaged it across the temporal dimension to derive a per-channel importance score. This



(a) Raw EEG data (b) EEG features

Fig. 12. t-SNE results from one subject. (a) Raw EEG features and (b) EEG features after training.

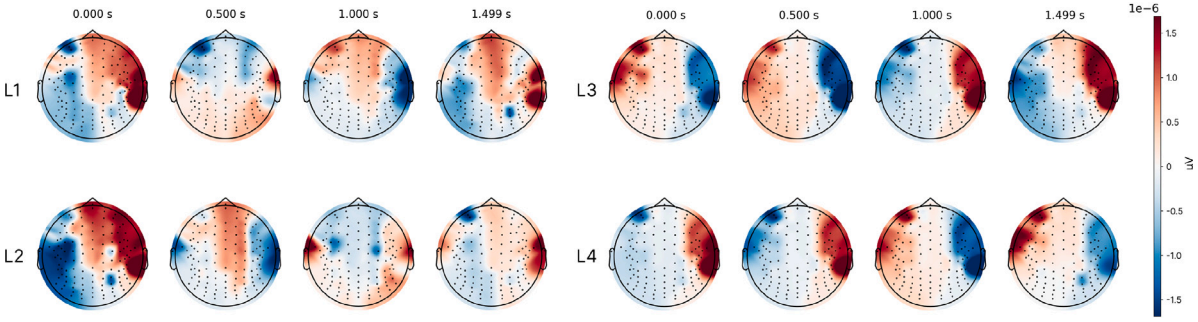


Fig. 13. Scalp topographic maps of event-related potentials (ERPs) for Subject 2 across four gaze conditions (L1–L4) and four time points (0.000 s, 0.500 s, 1.000 s, and 1.499 s). Color values represent voltage (μV) at each channel.

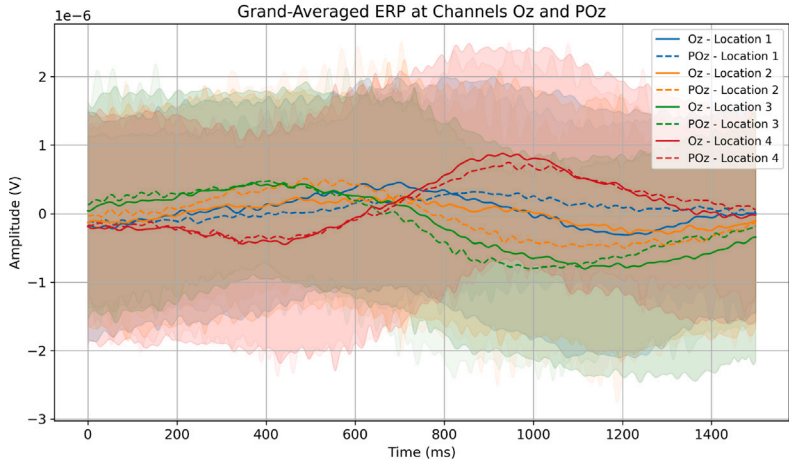


Fig. 14. Stimulus-locked ERP waveforms (0–1500 ms) for Subject 2 at occipital (Oz) and parietal (POz) electrodes, averaged across all trials per spatial condition.

procedure was performed on the three subjects with the highest classification performance using individually trained models, and the resulting saliency maps were averaged to obtain a cross-subject topographic visualization (Fig. 15, left).

The resulting scalp distribution shows that both occipital and parietal regions (e.g., POz, Pz, Oz) exhibited moderate to high saliency, suggesting a contribution from posterior visual areas to the decoding of 3D spatial attention. Meanwhile, several frontal-temporal electrodes (e.g., FT9, T10, FT8) — often associated with ocular activity — also showed strong saliency. This likely reflects residual eye-movement signals despite ICA-based preprocessing, and highlights the susceptibility of EEG decoding models to non-neural influences. Our findings suggest that the model may be leveraging a combination of neural and non-neural features. In real-world gaze estimation, components traditionally treated as artifacts could actually improve performance and robustness. Future work may benefit from explicitly comparing neural-only models with hybrid approaches that incorporate ocular signals (e.g., EOG) to enhance generalization and ecological validity.

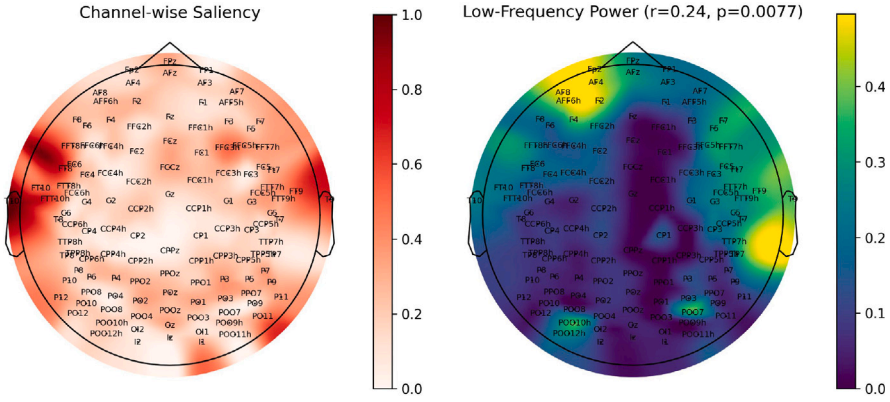


Fig. 15. Comparison between gradient-based saliency (left) and low-frequency (1–30 Hz) power distribution (right) across EEG channels, averaged across subjects 1, 2, and 3. Warmer colors in the saliency map indicate higher channel-level contributions to the model's predictions. The right panel shows normalized power, revealing spatially localized spectral activity. The two maps show a spatial correlation ($r = 0.24$, $p = 0.0077$), based on Pearson's coefficient across channels.

In the right panel of Fig. 15, we visualized the topographic distribution of low-frequency (1–30 Hz) power, which showed a positive Pearson correlation with the saliency map ($r = 0.24$, $p = 0.0077$). Some regions with elevated power likely reflecting residual noise, did not align with high saliency, suggesting they did not drive the model's predictions.

7. Discussion

In this paper, we present a novel approach for 3D gaze estimation using BCI technology as an alternative to traditional eye-tracking methods. This section discusses the reflections and limitations of our approach, as well as potential directions for future research.

7.1. Reflections on the use of BCIs

Providing a More Natural User Experience: Feedback from 5 participants indicated that our method felt more intuitive and provided a more natural user experience. They reported no noticeable mental burden throughout the experiment process and found it easy to judge the positions of the stimuli. Additionally, participants expressed a willingness to adopt this technology in practical applications, particularly excited about its potential use in gaming. These reflections suggest that BCI-based gaze estimation could offer a more comfortable and less intrusive alternative to traditional eye-tracking methods, which sometimes involve discomfort or visual interference from cameras.

Multimodal Integration and Flexibility: BCI technology can be integrated with other visual methods to enhance the accuracy of gaze estimation in challenging environments. For example, BCI can be combined with traditional visual sensors to improve the reliability of gaze point predictions where lighting or visual obstructions are present. This flexibility gives BCI technology broader potential for application across various scenarios.

7.2. Limitations

Limited Number and Position of Stimuli. The current experiment only used four stimuli at fixed positions. While effective for initial testing, this setup restricts the diversity of the gaze estimation in a more complex environment. Future research could benefit from increasing the number of stimuli and varying their positions more dynamically within the 3D space to better simulate real-world scenarios.

Small Sample Size and Individual Variability. The relatively small sample size of five participants limits the generalizability of our findings. While small sample sizes are common in foundational neural decoding studies, inter-individual variability in brain signals — such as differences in functional connectivity and neural response patterns — can introduce inconsistencies in model performance. These variations may lead to discrepancies in gaze prediction accuracy across participants, highlighting the need for further investigation into subject-specific adaptations. Future studies should aim to address these challenges by expanding the participant pool to include individuals with diverse demographics, cognitive profiles, and visual characteristics. Additionally, integrating personalized calibration techniques or adaptive neural representations may help mitigate individual variability, improving model robustness and generalizability.

Regression and Brute-Force Methods. The regression and brute-force methods currently used in this study can be computationally intensive and may require optimization for applications needing rapid, real-time processing. As the search space expands or becomes more complex, these methods may experience slower convergence rates. To enhance computational efficiency and scalability, future research could explore more optimization techniques, such as non-inversion methods or evolutionary algorithms. Alternatively, integrating generative models that allow for multimodal prediction could provide a more flexible and robust framework, combining different data types and learning paradigms to better capture the complexities of real-world scenarios.

Practical Application. As an exploratory study, Our EEG data collection involves extensive preparation, including gel-based electrodes, system calibration, and manual adjustments, making it impractical for real-time deployment. To facilitate practical adoption, future efforts should focus on lightweight EEG headsets with fewer electrodes, particularly dry-electrode systems, which can improve wearability, reduce setup time, and enhance user comfort. Integrating portable and user-friendly EEG solutions will be essential to maintaining long-term usability while ensuring stable signal acquisition in diverse environments. Addressing these challenges will be critical for bridging the gap between research prototypes and real-world applications.

Data Ethics. The use of EEG for gaze estimation raises ethical considerations. In our preliminary survey, 7% of respondents expressed concerns about ethical issues, highlighting the importance of addressing data security and transparency. EEG signals may contain unintended cognitive or emotional information, necessitating data anonymization, encryption, and secure storage to protect participant privacy. Additionally, participant consent should explicitly clarify how EEG data is collected, processed, and shared, ensuring transparency in research applications.

8. Conclusion

This study investigates the use of brain–computer interface (BCI) technology for 3D gaze estimation, offering a novel alternative to

traditional eye-tracking methods. Our approach leverages EEG signals to directly decode spatial attention. We collected an EEG dataset in a virtual reality setting, exposing participants to various 3D stimuli to train and evaluate our model. The results demonstrate that BCI-based gaze estimation can effectively map neural activity to visual attention, achieving reasonable PoR predictions. Additionally, the survey indicates positive public perception of BCI technology, suggesting its promise for future applications. While our results are promising for initial testing, further research is needed to refine the methodology, expand the variety of stimuli, and explore more dynamic environments to fully realize the potential of BCI-driven gaze estimation.

CRediT authorship contribution statement

Dantong Qin: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Data curation, Conceptualization. **Yang Long:** Writing – review & editing, Supervision, Methodology. **Xun Zhang:** Methodology, Data curation. **Zhibin Zhou:** Funding acquisition. **Pan Wang:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Data curation, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Pan Wang reports financial support was provided by The Hong Kong Polytechnic University. Yang Long, one of the editors of this journal, is one of the authors of this manuscript. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The research presented in this article was partially funded by grants from the Hong Kong Polytechnic University [Project No. P0042736].

Data availability

Data will be made available on request.

References

[1] T. Liu, H. Liu, B. Yang, Z. Zhang, Ldcnet: limb direction cues-aware network for flexible human pose estimation in industrial behavioral biometrics systems, IEEE Trans. Ind. Inform. (2023).
[2] T. Hempel, A.A. Abdelrahman, A. Al-Hamadi, Toward robust and unconstrained full range of rotation head pose estimation, IEEE Trans. Image Process. 33 (2024) 2377–2387.
[3] H. Liu, T. Liu, Y. Chen, Z. Zhang, Y.-F. Li, EHPE: Skeleton cues-based gaussian coordinate encoding for efficient human pose estimation, IEEE Trans. Multimed. (2022).
[4] H. Liu, C. Zhang, Y. Deng, T. Liu, Z. Zhang, Y.-F. Li, Orientation cues-aware facial relationship representation for head pose estimation via transformer, IEEE Trans. Image Process. 32 (2023) 6289–6302.
[5] H. Liu, T. Liu, Z. Zhang, A.K. Sangaiah, B. Yang, Y. Li, ARHPE: Asymmetric relation-aware representation learning for head pose estimation in industrial human–computer interaction, IEEE Trans. Ind. Inform. 18 (10) (2022) 7107–7117.
[6] P. Pathirana, S. Senarath, D. Meedeniya, S. Jayarathna, Eye gaze estimation: A survey on deep learning-based approaches, Expert Syst. Appl. 199 (2022) 116894.
[7] F. Lu, X. Chen, Person-independent eye gaze prediction from eye images using patch-based features, Neurocomputing 182 (2016) 10–17.
[8] X. Zhao, Y. Huang, Y. Tian, M. Tian, Episode-based personalization network for gaze estimation without calibration, Neurocomputing 513 (2022) 36–45.
[9] M.U. Ghani, S. Chaudhry, M. Sohail, M.N. Geelani, GazePointer: A real time mouse pointer control implementation based on eye gaze tracking, in: INMIC, IEEE, 2013, pp. 154–159.
[10] S. Rivu, Y. Abdrabou, T. Mayer, K. Pfeuffer, F. Alt, GazeButton: enhancing buttons with eye gaze interactions, in: Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications, 2019, pp. 1–7.
[11] U. Masud, N. Almolhis, A. Alhazmi, J. Ramakrishnan, F. Ul Islam, A.R. Farooqi, Smart wheelchair controlled through a vision-based autonomous system, IEEE Access (2024).
[12] B. David-John, C. Peacock, T. Zhang, T.S. Murdison, H. Benko, T.R. Jonker, Towards gaze-based prediction of the intent to interact in virtual reality, in: ACM Symposium on Eye Tracking Research and Applications, 2021, pp. 1–7.
[13] S. Martin, S. Vora, K. Yuen, M.M. Trivedi, Dynamics of driver's gaze: Explorations in behavior modeling and maneuver prediction, IEEE Trans. Intell. Veh. 3 (2) (2018) 141–150.
[14] P. Li, X. Hou, X. Duan, H. Yip, G. Song, Y. Liu, Appearance-based gaze estimator for natural interaction control of surgical robots, IEEE Access 7 (2019) 25095–25110.
[15] Y. Yu, J.-M. Odobez, Unsupervised representation learning for gaze estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 7314–7324.
[16] T. Zhang, Y. Shen, G. Zhao, L. Wang, X. Chen, L. Bai, Y. Zhou, Swift-eye: Towards anti-blink pupil tracking for precise and robust high-frequency near-eye movement analysis with event cameras, IEEE Trans. Vis. Comput. Graphics (2024).
[17] J. Qin, T. Shimoyama, Y. Sugano, Learning-by-novel-view-synthesis for full-face appearance-based 3d gaze estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4981–4991.
[18] S. Nonaka, S. Nobuhara, K. Nishino, Dynamic 3d gaze from afar: Deep gaze estimation from temporal eye-head-body coordination, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 2192–2201.
[19] X. Zhou, J. Lin, Z. Zhang, Z. Shao, S. Chen, H. Liu, Improved itracker combined with bidirectional long short-term memory for 3D gaze estimation using appearance cues, Neurocomputing 390 (2020) 217–225.
[20] Y.-M. Kwon, K.-W. Jeon, J. Ki, Q.M. Shahab, S. Jo, S.-K. Kim, 3D gaze estimation and interaction to stereo display, Int. J. Virtual Real. 5 (3) (2006) 41–45.
[21] L. Sidenmark, C. Clarke, J. Newn, M.N. Lystbæk, K. Pfeuffer, H. Gellersen, Vergence matching: Inferring attention to objects in 3d environments for gaze-assisted selection, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 2023, pp. 1–15.
[22] J. Liu, J. Chi, W. Hu, Z. Wang, 3D model-based gaze tracking via iris features with a single camera and a single light source, IEEE Trans. Hum.- Mach. Syst. 51 (2) (2020) 75–86.
[23] L. Sun, Z. Liu, M.-T. Sun, Real time gaze estimation with a consumer depth camera, Inform. Sci. 320 (2015) 346–360.
[24] D. Kirst, A. Bulling, On the verge: Voluntary convergences for accurate and precise timing of gaze input, in: Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems, 2016, pp. 1519–1525.
[25] P. Blignaut, D. Wium, Eye-tracking data quality as affected by ethnicity and experimental design, Behav. Res. Methods 46 (2014) 67–80.
[26] P. Stawicki, F. Gembler, A. Saboor, I. Volosyak, Comparison of speed, accuracy, and user friendliness between SSVEP-based BCI and eyetracker, in: GBCIC, 2017.
[27] J. Samaha, T.C. Sprague, B.R. Postle, Decoding and reconstructing the focus of spatial attention from the topography of alpha-band oscillations, J. Cogn. Neurosci. 28 (8) (2016) 1090–1097.
[28] P. Sharp, T. Gutteling, D. Melcher, C. Hickey, Spatial attention tunes temporal processing in early visual cortex by speeding and slowing alpha oscillations, J. Neurosci. 42 (41) (2022) 7824–7832.
[29] A. Dan, M. Reiner, EEG-based cognitive load of processing events in 3D virtual worlds is lower than processing events in 2D displays, Int. J. Psychophysiol. 122 (2017) 75–84.
[30] N. Manshouri, T. Kayikcioglu, A comprehensive analysis of 2D&3D video watching of EEG signals by increasing PLSR and SVM classification results, Comput. J. 63 (3) (2020) 425–434.
[31] M.M. Himmelberg, F.G. Segala, R.T. Maloney, J.M. Harris, A.R. Wade, Decoding neural responses to motion-in-depth using EEG, Front. Neurosci. 14 (2020) 581706.
[32] C. Spampinato, S. Palazzo, I. Kavasidis, D. Giordano, N. Souly, M. Shah, Deep learning human mind for automated visual classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6809–6817.
[33] B. Kaneshiro, M. Perreau Guimaraes, H.-S. Kim, A.M. Norcia, P. Suppes, A representational similarity analysis of the dynamics of object processing using single-trial EEG classification, Plos One 10 (8) (2015) e0135697.
[34] Y. Deng, S. Ding, W. Li, Q. Lai, L. Cao, EEG-based visual stimuli classification via reusable LSTM, Biomed. Signal Process. Control. 82 (2023) 104588.
[35] S. Bagchi, D.R. Bathula, EEG-ConvTransformer for single-trial EEG-based visual stimulus classification, Pattern Recognit. 129 (2022) 108757.
[36] P. Tirupattur, Y.S. Rawat, C. Spampinato, M. Shah, Thoughtviz: Visualizing human thoughts using generative adversarial network, in: Proceedings of the 26th ACM International Conference on Multimedia, 2018, pp. 950–958.

[37] I. Kavasidis, S. Palazzo, C. Spampinato, D. Giordano, M. Shah, Brain2image: Converting brain signals into images, in: Proceedings of the 25th ACM International Conference on Multimedia, 2017, pp. 1809–1817.
[38] R. Mishra, K. Sharma, R.R. Jha, A. Bhavsar, NeuroGAN: image reconstruction from EEG signals via an attention-based GAN, Neural Comput. Appl. 35 (12) (2023) 9181–9192.
[39] P. Singh, P. Pandey, K. Miyapuram, S. Raman, EEG2image: image reconstruction from EEG brain signals, in: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2023, pp. 1–5.
[40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).
[41] L. García, R. Ron-Angevin, B. Loubière, L. Renault, G. Le Masson, V. Lespinet-Najib, J.M. André, A comparison of a brain-computer interface and an eye tracker: Is there a more appropriate technology for controlling a virtual keyboard in an als patient? in: Advances in Computational Intelligence: 14th International Work-Conference on Artificial Neural Networks, IWANN 2017, Cadiz, Spain, June 14-16, 2017, Proceedings, Part II 14, Springer, 2017, pp. 464–473.
[42] S.O. Dumoulin, B.A. Wandell, Population receptive field estimates in human visual cortex, Neuroimage 39 (2) (2008) 647–660.
[43] B.P. Klein, B.M. Harvey, S.O. Dumoulin, Attraction of position preference by spatial attention throughout human visual cortex, Neuron 84 (1) (2014) 227–237.
[44] N.J. Finlayson, X. Zhang, J.D. Golomb, Differential patterns of 2D location versus depth decoding along the visual hierarchy, NeuroImage 147 (2017) 507–516.
[45] M. Henderson, V. Vo, C. Chunharas, T. Sprague, J. Serences, Multivariate analysis of BOLD activation patterns recovers graded depth representations in human visual and parietal cortex, Eneuro 6 (4) (2019).
[46] A. Cotrina, A.B. Benevides, J. Castillo-Garcia, A.B. Benevides, D. Rojas-Vigo, A. Ferreira, T.F. Bastos-Filho, A ssvep-bci setup based on depth-of-field, IEEE Trans. Neural Syst. Rehabil. Eng. 25 (7) (2017) 1047–1057.
[47] G.R. Reddy, M.J. Proulx, L. Hirshfield, A. Ries, Towards an eye-brain-computer interface: Combining gaze with the stimulus-preceding negativity for target selections in XR, in: Proceedings of the CHI Conference on Human Factors in Computing Systems, 2024, pp. 1–17.
[48] V.J. Lawhern, A.J. Solon, N.R. Waytowich, S.M. Gordon, C.P. Hung, B.J. Lance, EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces, J. Neural Eng. 15 (5) (2018) 056013.
[49] C. Zhang, Y.-K. Kim, A. Eskandarian, EEG-inception: an accurate and robust end-to-end neural network for EEG-based motor imagery classification, J. Neural Eng. 18 (4) (2021) 046014.
[50] L.R. Medsker, L. Jain, et al., Recurrent neural networks, Des. Appl. 5 (64–67) (2001) 2.
[51] S. Alhagry, A.A. Fahmy, R.A. El-Khoribi, Emotion recognition based on EEG using LSTM recurrent neural network, Int. J. Adv. Comput. Sci. Appl. 8 (10) (2017).
[52] P. Wang, A. Jiang, X. Liu, J. Shang, L. Zhang, LSTM-based EEG classification in motor imagery tasks, IEEE Trans. Neural Syst. Rehabil. Eng. 26 (11) (2018) 2086–2095.
[53] Y. Fang, J. Lei, J. Li, L. Xu, W. Lin, P. Le Callet, Learning visual saliency from human fixations for stereoscopic images, Neurocomputing 266 (2017) 284–292.
[54] H. Liu, C. Zhang, Y. Deng, B. Xie, T. Liu, Y.-F. Li, TransIFC: Invariant cues-aware feature concentration learning for efficient fine-grained bird image classification, IEEE Trans. Multimed. (2023).
[55] S. Mehraban, V. Adeli, B. Taati, Motionagformer: Enhancing 3d human pose estimation with a transformer-gcnformer network, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 6920–6930.
[56] H. Liu, Q. Zhou, C. Zhang, J. Zhu, T. Liu, Z. Zhang, Y.-F. Li, MMATrans: Muscle movement aware representation learning for facial expression recognition via transformers, IEEE Trans. Ind. Inform. (2024).
[57] J. Sun, J. Xie, H. Zhou, EEG classification with transformer-based models, in: 2021 Ieee 3rd Global Conference on Life Sciences and Technologies (Lifetech), IEEE, 2021, pp. 92–93.
[58] Y. Song, Q. Zheng, B. Liu, X. Gao, EEG conformer: Convolutional transformer for EEG decoding and visualization, IEEE Trans. Neural Syst. Rehabil. Eng. 31 (2022) 710–719.
[59] B. Abibullaev, A. Keutayeva, A. Zollanvari, Deep learning in EEG-based BCIs: A comprehensive review of transformer models, advantages, challenges, and applications, IEEE Access 11 (2023) 127271–127301.
[60] T. Grootswagers, A.K. Robinson, S.M. Shatek, T.A. Carlson, The neural dynamics underlying prioritisation of task-relevant information, 2021, arXiv preprint arXiv: 2102.01303.
[61] T. Grootswagers, A.K. Robinson, T.A. Carlson, The representational dynamics of visual objects in rapid serial visual processing streams, NeuroImage 188 (2019) 668–679.
[62] T. Grootswagers, I. Zhou, A.K. Robinson, M.N. Hebart, T.A. Carlson, Human EEG recordings for 1,854 concepts presented in rapid serial visual presentation streams, Sci. Data 9 (1) (2022) 1–7.
[63] L.H. Chew, J. Teo, J. Mountstephens, Aesthetic preference recognition of 3D shapes using EEG, Cogn. Neurodynamics 10 (2) (2016) 165–173.
[64] B. Mildenhall, P.P. Srinivasan, M. Tancik, J.T. Barron, R. Ramamoorthi, R. Ng, Nerf: Representing scenes as neural radiance fields for view synthesis, Commun. ACM 65 (1) (2021) 99–106.
[65] S. Palazzo, C. Spampinato, I. Kavasidis, D. Giordano, J. Schmidt, M. Shah, Decoding brain representations by multimodal learning of neural activity and visual features, IEEE Trans. Pattern Anal. Mach. Intell. 43 (11) (2020) 3833–3849.
[66] F. Aoki, E. Fetz, L. Shupe, E. Lettich, G. Ojemann, Increased gamma-range activity in human sensorimotor cortex during performance of visuomotor tasks, Clin. Neurophysiol. 110 (3) (1999) 524–537.
[67] M. Siegel, T.H. Donner, A.K. Engel, Spectral fingerprints of large-scale neuronal interactions, Nature Rev. Neurosci. 13 (2) (2012) 121–134.
[68] J. Fell, N. Axmacher, The role of phase synchronization in memory processes, Nature Rev. Neurosci. 12 (2) (2011) 105–118.
[69] T. Harmony, The functional significance of delta oscillations in cognitive processing, Front. Integr. Neurosci. 7 (2013) 83.
[70] M.J. Kahana, D. Seelig, J.R. Madsen, Theta returns, Curr. Opin. Neurobiol. 11 (6) (2001) 739–744.
[71] G. Buzsáki, Theta rhythm of navigation: link between path integration and landmark navigation, episodic and semantic memory, Hippocampus 15 (7) (2005) 827–840.
[72] J.J. Foster, E. Awh, The role of alpha oscillations in spatial attention: limited evidence for a suppression account, Curr. Opin. Psychol. 29 (2019) 34–40.
[73] M.S. Worden, J.J. Foxe, N. Wang, G.V. Simpson, Anticipatory biasing of visuospatial attention indexed by retinotopically specific alpha-band electroencephalography increases over occipital cortex, J. Neurosci.: Off. J. Soc. Neurosci. 20 (6) (2000) RC63–RC63.
[74] G. Thut, A. Nietzel, S.A. Brandt, A. Pascual-Leone, α-Band electroencephalographic activity over occipital cortex indexes visuospatial attention bias and predicts visual target detection, J. Neurosci. 26 (37) (2006) 9494–9502.
[75] S.P. Kelly, E.C. Lalor, R.B. Reilly, J.J. Foxe, Increases in alpha oscillatory power reflect an active retinotopic mechanism for distracter suppression during sustained visuospatial attention, J. Neurophysiol. 95 (6) (2006) 3844–3851.
[76] M.S. Clayton, N. Yeung, R.C. Kadosh, The roles of cortical oscillations in sustained attention, Trends Cogn. Sci. 19 (4) (2015) 188–195.
[77] Y. Zhou, F. Li, Y. Li, Y. Ji, G. Shi, W. Zheng, L. Zhang, Y. Chen, R. Cheng, Progressive graph convolution network for EEG emotion recognition, Neurocomputing 544 (2023) 126262.
[78] J.M.F. Montenegro, V. Argyriou, Gaze estimation using EEG signals for HCI in augmented and virtual reality headsets, in: 2016 23rd International Conference on Pattern Recognition, ICPR, IEEE, 2016, pp. 1159–1164.
[79] A. Kastrati, M.B. Plomecka, J. Küchler, N. Langer, R. Wattenhofer, Electrode clustering and bandpass analysis of eeg data for gaze estimation, in: Annual Conference on Neural Information Processing Systems, PMLR, 2023, pp. 50–65.
[80] P. Thuwajit, P. Rangpong, P. Sawangjai, P. Autthasan, R. Chaisaen, N. Banluesombatkul, P. Boonchit, N. Tatsaringkansakul, T. Sudhawiyangkul, T. Wilaiprasitporn, EEGWaveNet: Multiscale CNN-based spatiotemporal feature extraction for EEG seizure detection, IEEE Trans. Ind. Inform. 18 (8) (2021) 5547–5557.

Dantong Qin received her B.S. degree from Henan University in 2019 and her M.S. degree from Newcastle University in 2021. This work was conducted while she was a Ph.D. candidate in the Department of Computer Science at Durham University. In 2022, she was a visiting researcher at the School of Design, The Hong Kong Polytechnic University. Her research interests include brain signal decoding and developing vision models that assist and collaborate in visual content creation.

Yang Long is an Associate Professor in the Department of Computer Science, Durham University. He is also an IEEE Senior Member (SMIEEE) and MRC Innovation Fellow aiming to design scalable AI solutions for large-scale healthcare applications. His research background is in the highly interdisciplinary field of Computer Vision and Machine Learning. While he is passionate about unveiling the black-box of AI brain and transferring the knowledge to seek Scalable, Interactable, Interpretable, and sustainable solutions for other disciplinary researches, e.g. physical activity, mental health, design, education, security, and geoengineering. He has authored/co-authored 100+ top-tier papers in refereed journals/conferences such as IEEE TPAMI, TIP, CVPR, AAAI, and ACM MM.

Xun Zhang received his B.S. degree from Kunming University of Science and Technology in 2016 and his M.S. degree from The University of Manchester in 2018. He subsequently worked as a computer vision algorithm engineer at Digital Health Intelligence, focusing on medical image analysis. From 2021 to 2023, he was a research associate at the School of Design, The Hong Kong Polytechnic University. He is currently a Ph.D. candidate at the Faculty of Industrial Design Engineering, TU Delft. His research interests include computer vision, brain signal decoding, and human–computer interaction.

Zhibin Zhou is Research Assistant Professor in the School of Design at The Hong Kong Polytechnic University. Before joining PolyU, he was a temporary Research Faculty in the Hong Kong Center for Construction Robotics of the Hong Kong University of Science and Technology. He earned Ph.D. from the Zhejiang University School of Computer Science and Technology. With a background in human–computer interaction and artificial intelligence (AI), his prior research endeavors to capture the interaction between humans and AI in order to gain a greater understanding of AI as an emerging technology for empowering the user experience (UX). In addition, he was also a visiting Ph.D. candidate in the Politecnico di Milano, working on an AI-powered platform for facilitating the work of package designers.

Yuting Jin completed her undergraduate studies at Donghua University. She is currently pursuing a Master's degree in Integrated Product Design at the Faculty of Industrial Design Engineering, TU Delft. Her research focuses on medical product design and the integration of artificial intelligence in product design.

Pan Wang received the Ph.D. degree at the Data Science Institute and Dyson School of Design Engineering Imperial College London. She is currently an Assistant Professor in the Faculty of Industrial Design Engineering at Delft University of Technology. Her research is focused on Artificial Intelligence (AI) for design, especially Human–machine (AI) co-creativity. It addresses the intersection of human–computer interaction, brain–computer interface (BCI), creativity, AI design method and AI artworks.