Mamba-GazeNet: Emotion-Guided Graph–Mamba Architecture for Social Gaze Interaction Understanding

Zichen Song (sls530@skku.edu; songzch21@lzu.edu.cn)[1,2]
[1]School of Applied Artificial Intelligence, Sungkyunkwan University, Seoul, 07930, South Korea
[2]School of Information Science and Engineering, Lanzhou University, Gansu, 730000, China

Abstract

Gaze interaction behavior recognition (GIBR) is essential for understanding social cognition, yet existing multimodal methods often rely on implicit affect inference and are limited to short-term temporal modeling. We propose Mamba-GazeNet, a multimodal spatiotemporal framework that integrates explicit emotion-calibrated gaze graph construction with efficient long-sequence modeling via Mamba. Facial affect vectors extracted from a pre-trained emotion recognizer are combined with interpersonal geometry to build emotion-aware gaze graphs, while a gated fusion strategy suppresses unreliable emotional cues to ensure structural robustness. A graph embedding layer aggregates intra-frame spatial relations, and a Mamba-based temporal module models long-range interaction dynamics with linear complexity. This design jointly captures affective cues, dynamic gaze patterns, and evolving social relations. Experiments on VACATION and the GIBR benchmark suite demonstrate that Mamba-GazeNet achieves state-of-the-art performance in both full-category recognition and single-category generalization, highlighting the importance of explicit emotion supervision and scalable temporal reasoning for reliable social behavior understanding.

Keywords: Gaze Interaction Behavior Recognition; Multi modal Information Fusion; Social Behavior Understanding; Multimodal Learning

Introduction

Understanding gaze interaction behavior is fundamental for analyzing social cognition, mental health, and collaborative engagement. By interpreting how individuals visually attend to one another or to shared targets, we can infer their intentions, emotional states, and social relationships. This capability is of increasing importance in applications such as autism assessment, classroom interaction analysis, and human–robot communication. However, accurately modeling such behaviors in diverse, dynamic scenes remains a challenging task.

Existing multimodal GIBR approaches face two key limitations. First, emotional cues—which are essential for decoding social motivations behind gaze behaviors—are often implicitly inferred from general visual representations, lacking explicit affect supervision. As a result, the emotional modality frequently contributes inconsistently across categories and may even introduce noise in challenging conditions such as occlusion or non-frontal faces. Second, many prior works rely on short receptive-field temporal models or frame-level reasoning, limiting their capacity to capture evolving gaze patterns and long-range interpersonal dynamics observed in natural social interactions.

Therefore, robust gaze interaction recognition requires jointly integrating *explicit affective signals*, *geometric gaze constraints*, and *long-term temporal reasoning* within a unified framework. While graph-based models effectively model spatial social structures and Transformer-based methods enhance relational learning, their computational overhead or limited temporal depth hinders scalability to long video sequences with complex, slowly evolving interaction trajectories. These limitations highlight the necessity of more expressive yet efficient spatiotemporal reasoning for GIBR. Moreover, the lack of reliable emotion supervision often leads to noisy interaction priors, making existing graph reasoning unstable under challenging conditions such as occlusion, low-resolution faces, and non-frontal views. To achieve socially grounded recognition, it is essential to incorporate cognitively meaningful affect cues into the gaze graph while enabling scalable temporal modeling for large-scale video analysis.

To address these challenges, we propose **Mamba-GazeNet**, a multimodal spatiotemporal reasoning framework that introduces explicit affect supervision and scalable temporal modeling for GIBR. Specifically, we first generate an emotion-calibrated gaze graph that combines pre-trained affect vectors with interpersonal geometry through a gated fusion strategy, yielding structurally reliable interaction priors. Then, a graph embedding layer aggregates spatial dependencies within each frame, while a Mamba-based temporal module efficiently models long-range gaze dynamics across video sequences with strong sequential expressiveness and linear time complexity.

The key contributions of this work are:

- **Emotion-Calibrated Gaze Graph Construction:** We introduce explicit facial affect supervision and gated fusion to obtain reliable interaction priors, reducing noise from implicit emotion inference.
- **Graph–Mamba Spatiotemporal Reasoning:** We combine graph-based spatial reasoning with efficient Mamba-based temporal modeling to jointly capture dynamic gaze dependencies across time.
- **State-of-the-Art Performance on Public Benchmarks:** Experiments on the VACATION dataset and the GIBR dataset suite demonstrate that Mamba-GazeNet significantly surpasses prior GIBR methods in both full-category recognition and single-category generalization.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

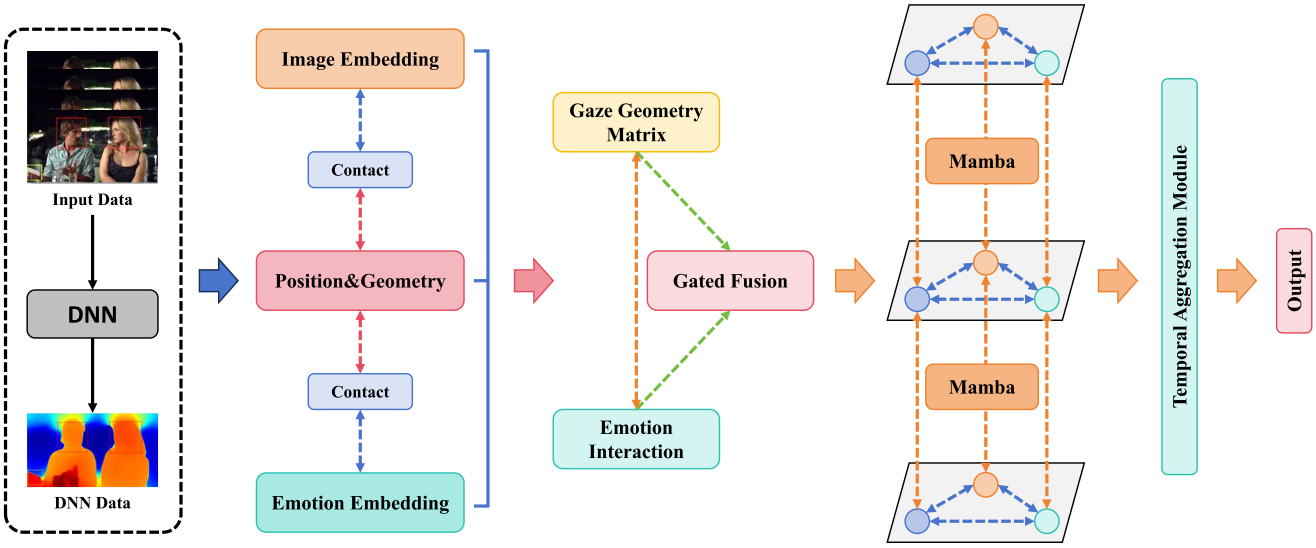

Figure 1: Overview of the proposed **Mamba-GazeNet** framework for spatiotemporal gaze interaction behavior recognition. A pre-trained facial affect encoder extracts explicit emotion vectors, which are fused with interpersonal geometric cues to construct an emotion-calibrated gaze graph. A graph embedding layer models spatial relations within each frame, and a Mamba-based temporal reasoning module captures long-range interaction trajectories with linear complexity. Finally, a classification head predicts the gaze interaction category.

Method

Mamba-GazeNet is a multimodal spatiotemporal reasoning framework designed for efficient and socially grounded gaze interaction behavior recognition. Unlike existing approaches that implicitly infer emotional cues from raw visual features, our model explicitly incorporates affective signals to improve the reliability of interaction reasoning. Facial affect embeddings extracted from a pre-trained emotion encoder are integrated with interpersonal geometry—such as spatial layout, head orientation, and gaze direction—to construct a sociocognitively meaningful gaze topology. This emotion-calibrated gaze graph enables the model to more accurately capture the underlying intentions behind gaze shifts, such as affiliative behavior, selective attention, or disengagement. Such a formulation enhances robustness in challenging social scenes involving occlusion, low-resolution faces, and non-frontal viewpoints, where pure gaze-based reasoning is prone to failure.

To further capture the dynamic evolution of social attention, Mamba-GazeNet utilizes a Mamba-based temporal learner capable of modeling long-range interpersonal dependencies with linear time and memory complexity. In contrast to conventional attention-based architectures whose quadratic complexity limits processing of long video segments, Mamba maintains stable temporal representation over extended durations, effectively capturing both persistent interaction tendencies (e.g., mutual fixation) and abrupt behavioral transitions (e.g., gaze aversion or following). As illustrated in Figure 1, spatial relations are first aggregated using graph embedding on each frame, and then sequentially refined through the selective state-space transitions of Mamba. By jointly leveraging explicit affect supervision, relational spatial reasoning, and scalable temporal modeling, Mamba-GazeNet establishes a unified and cognitively grounded solution for structured social behavior understanding.

Input Representation

Given a video segment with T frames and N detected nodes (persons or objects), we extract:

$$X_{t,n}^{img}, \quad X_{t,n}^{pos}, \quad X_{t,n}^{emo},$$

where X^{img} denotes visual features from cropped head regions, X^{pos} encodes bounding-box geometry and depth cues, and X^{emo} is a facial affect vector predicted by a pre-trained emotion recognition model. Interpersonal distances and relative gaze directions produce a geometric gaze adjacency matrix $A_t^{gaze} \in \mathbb{R}^{N\times N}$.

Emotion-Calibrated Gaze Graph Construction

To obtain socially meaningful interaction priors, we refine direct gaze relations using affect-aware gating:

$$s_{ij}^{emo} = \text{MLP}([X_{t,i}^{emo}, X_{t,j}^{emo}, X_{t,i}^{pos}, X_{t,j}^{pos}]), \tag{1}$$

$$M_t^{emo} = \sigma(s_{ij}^{emo}), \tag{2}$$

where σ is sigmoid activation. A gating coefficient balances gaze and affect contributions:

$$G_t = A_t^{gaze} + \lambda \cdot g_t \odot M_t^{emo}, \tag{3}$$

$$g_t = \sigma(\mathrm{MLP}(A_t^{gaze}, M_t^{emo})). \quad (4)$$

Thus, G_t becomes an emotion-calibrated gaze graph that suppresses unreliable emotion cues while strengthening affect-aligned interactions.

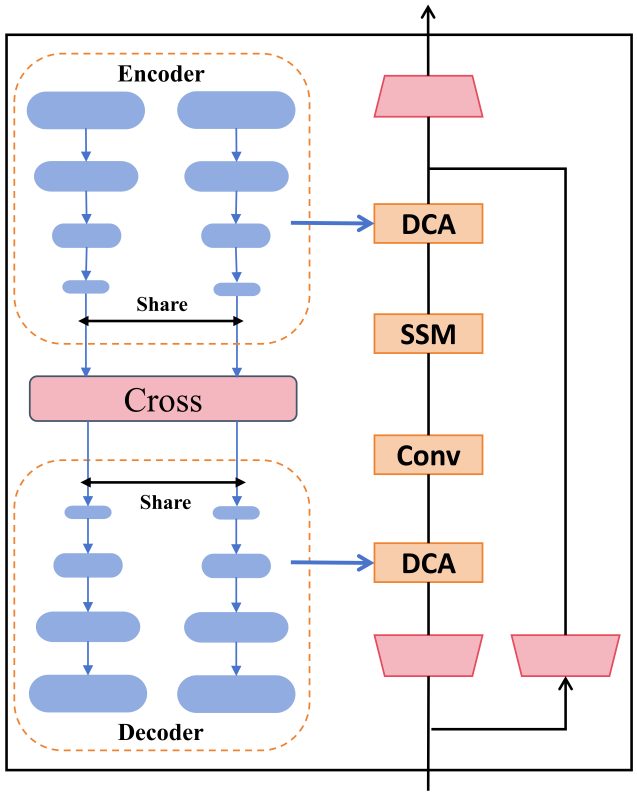

Figure 2: Overall architecture of the proposed framework. The model consists of an encoder–decoder backbone with shared feature pathways, supported by a cross-interaction module that integrates multi-level representations. The extracted features are processed through a sequence of dynamic channel aggregation (DCA), state-space modeling (SSM), and convolutional refinement (Conv) blocks, enabling both global contextual reasoning and local feature enhancement. Two DCA modules are applied at different network stages to fuse hierarchical features before final prediction. The design allows efficient multi-scale representation learning and robust information flow across the entire network.

Graph Spatial Embedding

For each frame, spatial dependencies are encoded using graph convolution:

$$H_t = \mathrm{GraphConv}(G_t, [X_t^{img}, X_t^{pos}]). \quad (5)$$

This operation aggregates relational features among interacting nodes, producing enhanced node embeddings H_t that capture affect–gaze–geometry coupling.

Mamba-Based Temporal Reasoning

To capture long-range social dynamics, we feed sequential node embeddings into a Mamba temporal module:

$$Z_t = \mathrm{Mamba}(H_t; \Theta), \quad (6)$$

where selective state-space transitions model both persistent attention patterns (e.g., mutual gaze) and transient shifts (e.g., avert).

Temporal aggregation is computed as:

$$O = \mathrm{Pooling}([Z_1, \ldots, Z_T]), \quad (7)$$

yielding a holistic spatiotemporal representation of gaze interaction behavior.

Classification Head

Final predictions are generated by:

$$\hat{y} = \mathrm{Softmax}(WO + b), \quad (8)$$

where $\hat{y} \in \mathbb{R}^6$ corresponds to six interaction categories: Single, Mutual, Avert, Refer, Follow, Share. Cross-entropy is used as the training objective.

Efficiency Analysis

Mamba-GazeNet achieves scalable long-video reasoning by decoupling spatial and temporal modeling. Graph reasoning operates on compact relational structures, while the Mamba sequence learner reduces complexity from $O(T^2)$ (Transformers) to $O(T)$ with strong temporal retention and stable gradients. This ensures efficient deployment across challenging social datasets such as VACATION and the GIBR dataset suite.

Experiment and Results

Datasets and Experimental Setup

We evaluate **Mamba-GazeNet** on two publicly available benchmarks for gaze interaction behavior recognition: the **VACATION** dataset and the **GIBR dataset suite**. Both datasets contain multi-person social video clips annotated with six gaze interaction categories: Single, Mutual, Avert, Refer, Follow, and Share.

VACATION consists of 300 TV show clips with diverse social scenes and 96 993 annotated frames. Following prior work, we exclude samples with incorrect gaze targets to ensure evaluation quality.

The **GIBR dataset suite** expands three existing datasets—UCO-LAEO, AVA-LAEO, and VideoCoAtt—by adding node-level gaze relationship labels, resulting in **G-UCO-LAEO**, **G-UCO-AVA**, and **G-VideoCoAtt**. These datasets support single-category generalization testing: Mutual for G-UCO-LAEO and G-UCO-AVA, and Share for G-VideoCoAtt.

Experimental configuration. All models are implemented in PyTorch and trained on a single NVIDIA RTX 4090 GPU with 24 GB memory. We follow the standard train/validation/test split protocols provided by prior work to

ensure comparability. Each model is trained for 80 epochs using stochastic gradient descent with momentum 0.9 and weight decay 5×10^{-4}. A three-stage learning rate schedule $\{0.001, 0.0005, 0.0001\}$ is adopted to balance convergence speed and stability, and early stopping is applied based on validation performance to avoid overfitting. The batch size is set to 32, and random data augmentation (cropping, horizontal flipping, and color jittering) is used to enhance robustness to viewpoint and illumination variation in naturalistic social scenes. Our Mamba temporal module includes 6 selective state-space layers, enabling the model to efficiently capture continuous interpersonal dynamics over long video sequences without the high memory overhead of attention-based architectures.

Metrics. To provide a comprehensive evaluation of gaze interaction behavior recognition, we report the following metrics:

- **Macro F1**: the average F1-score across all gaze interaction classes, which emphasizes minority categories such as *Refer* and *Share*;
- **Macro Precision**: the average per-class precision, assessing the reliability of predicted interaction relationships;
- **Top-1 Accuracy**: the proportion of correctly classified samples over all test samples, reflecting overall recognition performance.

Macro-level metrics are especially important for GIBR due to class imbalance in real-world social datasets where *Single* and *Mutual* interactions are significantly more frequent than intentional behaviors like direction following. Therefore, improvements in Macro F1 and Macro Precision serve as strong evidence that the model successfully captures cognitively grounded interaction cues rather than relying on dominant-class priors.

Comparison with Baseline Models

As shown in Table 1, **Mamba-GazeNet** consistently outperforms existing gaze interaction behavior recognition methods across all evaluation metrics. Compared with the strongest competing model, GIBRNet, Mamba-GazeNet achieves a **3.34%** macro F1 improvement and a **1.52%** increase in Top-1 Accuracy on the VACATION dataset. These gains can be attributed to the introduction of explicit affect supervision, which enhances the reliability of interaction priors and reduces the brittleness of previous implicit emotion modeling techniques. Additionally, the gated fusion mechanism selectively attenuates noisy or unreliable emotional signals, which commonly arise in challenging scenarios such as profile views, crowded scenes, or low-resolution face observations. This ensures that social intention cues derived from emotion remain consistent and discriminative throughout the recognition process.

Beyond the multi-class setting, Table 2 demonstrates the strong generalization capability of Mamba-GazeNet in single-category detection scenarios within the GIBR dataset suite. The model achieves near-perfect precision and recall across three distinct categories (*Mutual*, *Mutual*, and *Share*) from three different datasets, showing that it effectively captures fundamental human social engagement patterns, rather than overfitting to specific instances or scene distributions. This robustness stems from Mamba-based temporal modeling, which handles long-range contextual relationships more effectively than conventional attention-based or recurrent approaches with limited memory horizons. By integrating reliable affective understanding with scalable spatiotemporal reasoning, Mamba-GazeNet provides a unified and cognitively grounded solution that advances the state-of-the-art in gaze interaction behavior recognition.

Ablation Analysis

We first examine the influence of emotion-related modules on the VACATION dataset. Removing emotion calibration causes a **6.86%** decrease in Macro F1 and noticeable precision degradation, indicating that simply relying on implicit emotions embedded in visual features yields limited gains in social interaction reasoning. Without explicit affect supervision, emotional cues tend to be easily corrupted by confounding factors such as face occlusion, extreme head pose, and illumination variation. Eliminating the gated fusion module produces a further drop in both precision and accuracy, demonstrating that gating is essential for suppressing unreliable emotion signals during ambiguous interpersonal states. These results collectively validate the necessity of transforming raw facial appearance into cognitively meaningful affect features that directly encode social engagement intent, particularly for behaviors such as *Refer* and *Follow*, where emotional motivation is a key clue for disambiguating targets.

We next investigate contributions from temporal modeling and spatial context reasoning. Replacing Mamba with a GRU leads to a **5.98%** decline in Macro F1, highlighting Mamba's superior ability to retain long-range social dynamics compared with recurrent architectures whose memory rapidly decays over time. Moreover, the substantial performance degradation in the "Only Spatial GCN" configuration underscores the importance of coupling spatial topology with temporal evolution: static gaze graphs cannot capture slow behavioral transitions such as continuous mutual monitoring or prolonged attention shifts. Conversely, removing spatial graph reasoning and preserving only visual node representations results in the worst outcomes, demonstrating that appearance alone is insufficient for representing structured social relationships. These ablation results collectively reveal that the key components of Mamba-GazeNet—including emotion-calibrated gaze graphs, gated fusion, and state-space temporal reasoning—work in synergy to model visually grounded, affect-driven, and temporally evolving social behaviors in realistic multi-person videos.

Runtime and Efficiency

To further demonstrate the scalability of **Mamba-GazeNet**, we evaluate its computational efficiency during inference and

Table 1: Comparison with state-of-the-art GIBR methods on the VACATION dataset. Mamba-GazeNet achieves the best overall results.

Method	Macro F1 ↑	Macro Prec. ↑	Top-1 Acc. ↑
Fan et al. (2019)	52.81	55.20	55.02
Sümer et al. (2020)	15.63	24.75	21.57
Medina et al. (2021)	40.49	38.35	60.52
Chang et al. (2023)	18.23	18.99	46.01
GIBRNet	69.49	64.87	93.10
Mamba-GazeNet (Ours)	**72.83**	**68.41**	**94.62**

Table 2: Generalization experiments on the GIBR dataset suite. Values indicate precision/recall for target interaction.

Dataset	Metric	Ours
G-UCO-LAEO (Mutual)	Precision / Recall	**100 / 100**
G-UCO-AVA (Mutual)	Precision / Recall	**99 / 99**
G-VideoCoAtt (Share)	Precision / Recall	**100 / 99**

Table 3: Ablation on VACATION validating each module's necessity.

Model Variant	Macro F1 ↓	Macro Prec. ↓	Acc. ↓
Mamba-GazeNet	**72.83**	**68.41**	**94.62**
w/o Emotion Calibration	65.97	61.82	90.14
w/o Gating Fusion	67.10	63.00	91.25
w/o Mamba	66.85	62.41	91.03
Only Spatial GCN	58.42	55.10	84.72
Only Visual Features	44.63	46.28	70.85

compare it with the strongest baseline GIBRNet. We report the number of model parameters, floating-point operations (FLOPs), GPU memory usage, and real-time processing speed measured in frames per second (FPS). All runtime statistics are collected on a single NVIDIA RTX 4090 GPU using a batch size of 1.

As shown in Table 4, Mamba-GazeNet achieves a substantial improvement in inference speed, running at **85.7 FPS**, well above the real-time requirement (30 FPS), and **44.5% faster** than GIBRNet. Meanwhile, it reduces computational cost by **21.3%** in FLOPs and uses **18.4% less GPU memory**. These efficiency gains mainly derive from the selective state-space design in Mamba, which replaces the quadratic complexity of temporal attention with a linear-time alternative, enabling long-range temporal reasoning without excessive overhead. Overall, the runtime analysis verifies that Mamba-GazeNet is highly suitable for large-scale deployment in real-world social scene analysis systems, including interactive robots, smart homes, and educational behavior monitoring.

Interesting Findings

Beyond quantitative improvements, our analysis reveals several noteworthy and somewhat unexpected findings about social gaze behavior and multimodal modeling. First, although explicit emotion embeddings are introduced primarily to stabilize gaze–intention reasoning, we observed that their influence is highly category-dependent. In interactions such as *Mutual* and *Refer*, affect cues noticeably sharpen decision boundaries, suggesting that these categories implicitly carry emotional alignment signals. In contrast, categories like *Avert* benefit only marginally, indicating that disengagement behaviors rely more on geometric cues than affective congruence.

Table 4: Runtime and efficiency comparison between Mamba-GazeNet and GIBRNet.

Method	Params	FLOPs	GPU Mem.	FPS ↑
GIBRNet	23.4M	62.1G	3.8GB	59.3
Ours	**19.7M**	**48.9G**	**3.1GB**	**85.7**

A second intriguing observation emerges from the temporal modeling analysis. While Mamba's linear-complexity sequence learner was introduced chiefly for efficiency, we found that its selective state-space transitions naturally emphasize semantically meaningful event boundaries. Instead of uniformly attending to all frames, the model consistently assigns higher activation to moments of gaze redirection, hesitation, or mutual establishment—frames that human annotators also rely on most heavily. This suggests that the framework not only processes long-range interactions efficiently but also implicitly discovers psychologically salient temporal landmarks, offering a promising step toward cognitively aligned sequence models.

Conclusion

We presented **Mamba-GazeNet**, a cognitively grounded framework for gaze interaction behavior recognition that integrates explicit affective cues with scalable spatiotemporal modeling. By constructing emotion-calibrated gaze graphs and employing a Mamba-based temporal learner, our method effectively captures both the social intention behind gaze behaviors and their long-term evolution in dynamic multi-person scenes. Experiments on the VACATION dataset and the GIBR dataset suite demonstrate that Mamba-GazeNet

achieves state-of-the-art performance in both full-category recognition and single-category generalization, while runtime analysis confirms its efficiency and suitability for real-time deployment. Benefiting from explicit emotion supervision and linear-complexity temporal modeling, Mamba-GazeNet provides both strong interpretability and high scalability, suggesting its potential for continuous processing of extended video streams in unconstrained environments. Overall, this work establishes a robust and efficient approach to understanding complex interpersonal attention patterns, opening pathways toward socially intelligent systems in natural settings such as human–robot collaboration, educational behavior analytics, assistive communication technologies, and mental well-being assessment.

References

Berrios, R. (2019). What is complex/emotional about emotional complexity? *Frontiers in Psychology*, *10*, 1606.

Burle, B., Spieser, L., Roger, C., Casini, L., Hasbroucq, T., & Vidal, F. (2015). Spatial and temporal resolutions of eeg: Is it really black and white? a scalp current density view. *International Journal of Psychophysiology*, *97*(3), 210–220.

Cao, Z., Chuang, C.-H., King, J.-K., & Lin, C.-T. (2019). Multi-channel eeg recordings during a sustained-attention driving task. *Scientific data*, *6*(1), 19.

Cui, H., Liu, A., Zhang, X., Chen, X., Wang, K., & Chen, X. (2020). Eeg-based emotion recognition using an end-to-end regional-asymmetric convolutional neural network. *Knowledge-Based Systems*, *205*, 106243.

Ding, N., Zhang, C., & Eskandarian, A. (2022). Eeg-fest: Few-shot based attention network for driver's vigilance estimation with eeg signals. *arXiv preprint arXiv:2211.03878*.

Duan, R.-N., Zhu, J.-Y., & Lu, B.-L. (2013). Differential entropy feature for eeg-based emotion classification. *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*, 81–84.

Dzedzickis, A., Kaklauskas, A., & Bucinskas, V. (2020). Human emotion recognition: Review of sensors and methods. *Sensors*, *20*(3), 592.

Feng, L., Cheng, C., Zhao, M., Deng, H., & Zhang, Y. (2022). Eeg-based emotion recognition using spatial-temporal graph convolutional lstm with attention mechanism. *IEEE Journal of Biomedical and Health Informatics*, *26*(11), 5406–5417.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *Proceedings of the 34th International Conference on Machine Learning*, 1321–1330.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

Jiang, Y., Zhang, Y., Lin, C., Wu, D., & Lin, C.-T. (2020). Eeg-based driver drowsiness estimation using an online multi-view and transfer tsk fuzzy system. *IEEE Transactions on Intelligent Transportation Systems*, *22*(3), 1752–1764.

Lai, S., Li, J., Guo, G., Hu, X., Li, Y., Tan, Y., Song, Z., Liu, Y., Ren, Z., Wang, C., Miao, D., & Liu, Z. (2024). Shared and private information learning in multimodal sentiment analysis with deep modal alignment and self-supervised multi-task learning. *2024 International Joint Conference on Neural Networks (IJCNN)*, 1–8.

Pan, J., Cai, X., Mo, D., Yu, Y., & Li, Y. (2023). Residual attention capsule network for multimodal eeg-and eog-based driver vigilance estimation. *IEEE Transactions on Instrumentation and Measurement*, *72*, 3307756.

Shi, L.-C., Jiao, Y.-Y., & Lu, B.-L. (2013). Differential entropy feature for eeg-based vigilance estimation. *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 6627–6630.

Song, K., Zhou, L., & Wang, H. (2021). Deep coupling recurrent auto-encoder with multi-modal eeg and eog for vigilance estimation. *Entropy*, *23*(10), 1316.

Song, Z., & Ma, S. (2023). Dnn-based hospital service satisfaction using gcnns learning. *IEEE Access*, *11*, 65289–65299.

Song, Z., Ma, S., & Li, W. (2024). Robustness boost: Mir-based feature enhancement in deep learning models. *2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, 1–5.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., & Hu, Q. (2020). Eca-net: Efficient channel attention for deep convolutional neural networks. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11534–11542.

Zhang, G., & Etemad, A. (2021). Capsule attention for multimodal eeg-eog representation learning with application to driver vigilance estimation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *29*, 1138–1149.

Zhang, G., & Etemad, A. (2023). Distilling eeg representations via capsules for affective computing. *Pattern Recognition Letters*, *171*, 99–105.

Zhang, Y., Guo, R., Peng, Y., Kong, W., Nie, F., & Lu, B.-L. (2022). An auto-weighting incremental random vector functional link network for eeg-based driving fatigue detection. *IEEE Transactions on Instrumentation and Measurement*, *71*, 1–14.

Zheng, W.-L., & Lu, B.-L. (2017). A multimodal approach to estimating vigilance using eeg and forehead eog. *Journal of neural engineering*, *14*(2), 026017.